Essays on Panel and Network Modeling

Ming Li
Yale University Graduate School of Arts and Sciences, econmingli@gmail.com

Abstract

Essays on Panel and Network Modeling

Ming Li

2021

This dissertation studies identification and estimation in panel and network models. Panel models have long been a workhorse in empirical research. In the first two chapters, we analyze random coefficient linear panel model and panel multinomial choice model, respectively, where we incorporate features such as time-varying endogeneity and unobserved heterogeneity that are prevalent in real life into the models. We present new identification results and provide consistent estimators based on the identification strategy. Then, we apply the estimation procedures to panel data and obtain economically convincing results. The study of networks is a fast-growing area of economic research thanks to the increasing availability of network data and computing power. In the third chapter, we study network formation problems under non-transferrable utilities (NTU). We show how to identify the parameters of interest without additive separability based on "logical differencing" and provide consistent estimators.

In chapter 1, we propose a random coefficient linear panel model where the regressors can depend on the time-varying random coefficients in each period, a critical feature in many economic applications including production function estimation. The random coefficients are modeled as unknown functions of a fixed effect of arbitrary dimension and a random shock. The regressors may depend on the random coefficients due to agent's optimization behavior such as profit maximization, utility maximization, among others. We use a sufficiency argument to control for the fixed effect, which enables us to construct a feasible control function for the random shock and subsequently identify the moments of the random coefficients via a sequential argument. Based on the multi-step identification argument, we propose a series estimator and prove a new inference result. Monte Carlo simulations show that the proposed method can capture the distributional properties of the random coefficients. We then apply the procedure to panel data for Chinese manufacturing firms and find significant variation in the output elasticities both across firms and through time.

In chapter 2, we propose a simple yet robust method for semiparametric identification and estimation of panel multinomial choice models, where we allow infinite-dimensional fixed effects to enter consumer utilities in an additively nonseparable way, thus incorporating rich forms of unobserved heterogeneity. Such heterogeneity may take the form of, for example, brand loyalty or responsiveness to subtle flavor and packaging designs, which are hard to quantify but affect consumer choices in complex ways. Our identification strategy exploits the standard notion of multivariate monotonicity in its contrapositive form, which provides leverage for converting observable events into identifying restrictions on unknown parameters of interest. Based on our identification result, we construct consistent set (or point) estimators, together with a computational algorithm that adopts a machine learning algorithm and a new minimization procedure on the spherical-coordinate space. We demonstrate the practical advantages of our method with simulations and an empirical example using the Nielsen data. We find that special in-store displays boost sales not only through a direct promotion effect but also through the attenuation of consumers' price sensitivity.

In chapter 3, we consider a semiparametric model of dyadic network formation under NTU. NTU frequently arises in social interactions that require bilateral consent, such as Facebook friendship networks or informal risk-sharing networks in developing countries. However, NTU inherently induces additive non-separability, which makes identification challenging. Based on multivariate monotonicity, we identify structural parameters by constructing events involving the intersection of two mutually exclusive restrictions on the unobserved individual fixed effects to cancel them out. The constructive identification argument leads to a consistent estimator. We analyze the finite-sample performance of the

estimator via a simulation study. Then, we apply the method to the Nyakatoke risk-sharing network data. The results show that our approach can capture the essence of the network formation process. For instance, we find that the greater the difference in wealth between two households, the lower is the probability they are connected.

Essays on Panel and Network Modeling

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Ming Li

Dissertation Director: Donald W. K. Andrews and Yuichi Kitamura

June 2021

Contents

Abstract i

List of Tables viii

List of Figures ix

Acknowledgments xi

1 A Time-Varying Endogenous Random Coefficient Model with an Application to Production Functions 1

1.1 Introduction 1

1.1.1 Related Literature 6

1.2 Model 10

1.3 Identification 19

1.4 Estimation and Large Sample Theory 33

1.4.1 Estimation 33

1.4.2 Large Sample Theory 35

1.5 Simulation 45

1.5.1 DGP 46

1.5.2 Baseline Results 47

1.5.3 Robustness Checks 51

1.6 Production Function Application 53

1.6.1 Data and Methodology 54

1.6.2 Results 57

1.7 Conclusion 65

References 66

1.A Proofs in Section 1.3 74

1.B Proofs in Section 1.4 80

1.C Notation 94

2 Robust Semiparametric Estimation in Panel Multinomial Choice Models[1] 98

2.1 Introduction 98

2.2 Panel Multinomial Choice Model 105

2.2.1 Model Setup 105

2.2.2 Key Assumptions 109

2.3 Identification Strategy 113

2.4 Estimation and Computation 121

2.4.1 A Consistent Two-Step Estimator 121

2.4.2 Computation 124

2.5 Simulation 128

2.5.1 First-Stage Performance 129

2.5.2 Two-Stage Performance 130

2.6 Empirical Illustration 134

2.6.1 Data and Methodology 134

2.6.2 Results and Discussion 135

2.6.3 A Possible Explanation via Monte-Carlo Simulations 137

2.7 Monotone Multi-Index Models 139

2.8 Conclusion 142

References 143

2.A Proof of Theorem 2.2 148

2.B Pairwise Time Homogeneity of Errors 150

[1] Joint with Wayne Gao.

2.C Sufficient Conditions for Point Identification 151

2.D Additional Simulation Results 154

2.D.1 Adaptive-Grid Computation Algorithm 154

2.D.2 Results Varying D, J, T 156

3 Logical Differencing in Dyadic Network Formation Models with Nontransferable Utilities[2] 158

3.1 Introduction 158

3.2 A Nonseparable Dyadic Network Formation Model 165

3.3 Identification and Estimation 168

3.3.1 Identification via Logical Differencing 168

3.3.2 Sufficient Conditions for Point Identification 173

3.3.3 Tetrad Estimation and Consistency 175

3.4 Simulation 178

3.4.1 Setup of Simulation Study 178

3.4.2 Results under Symmetric Pairwise Observable Characteristics 180

3.4.3 Results under Asymmetric Pairwise Observable Characteristics 183

3.5 Empirical Illustration 185

3.5.1 Data Description 185

3.5.2 Methodology 188

3.5.3 Results and Discussion 188

3.6 Conclusion 189

References 190

3.A Proofs 196

3.A.1 Proof of Lemma 3.2 196

3.A.2 Proof of Theorem 3.1 197

3.A.3 Proof of Lemma 3.3 199

[2] Joint with Wayne Gao and Sheng Xu.

3.A.4 Proof of Theorem 3.2 202

3.B Asymmetry of Pairwise Observable Characteristics 204

List of Tables

1.1 Performance of $\widehat{\overline{\omega}}, \widehat{\overline{\beta}}^K$ and $\widehat{\overline{\beta}}^L$ . . . 47

1.2 Performance under Varying N and T . . . 52

1.3 Performance under Varying Orders of Basis Functions . . . 52

1.4 Performance with and without Ex-Post Shock . . . 53

1.5 Summary Statistics . . . 55

2.1 Performance of First Stage Estimator $G\left(\hat{\gamma}\right)$ . . . 130

2.2 Baseline Performance . . . 131

2.3 Performance under Varying N . . . 132

2.4 Performance with and without Point ID: Further Examination . . . 133

2.5 Empirical Application: Summary Statistics . . . 135

2.6 Empirical Application: Estimation Results . . . 136

2.7 Empirical Illustration: Comparison of Results . . . 136

2.8 Percentage of Correct Signs of Estimated Coefficients . . . 138

2.9 Performance Varying D, J, T . . . 157

3.1 Baseline Performance . . . 181

3.2 Results Varying N and $\overline{d}$ . . . 182

3.3 Results Varying $corr$ . . . 183

3.4 Results under Asymmetry . . . 184

3.5 Empirical Application: Summary Statistics . . . 187

3.6 Empirical Application: Estimation Results 188

List of Figures

1.1 Histogram of $\widehat{\overline{\omega}}_b$ and $\overline{\omega}_b$ 48

1.2 Histogram of $\widehat{\overline{\beta}}_b^K$ and $\overline{\beta}_b^K$ 49

1.3 Histogram of $\widehat{\overline{\beta}}_b^L$ and $\overline{\beta}_b^L$ 49

1.4 Histogram of $\widehat{\omega}_{it}$ versus ω_{it} 50

1.5 Histogram of $\widehat{\beta}_{it}^K$ versus β_{it}^K 50

1.6 Histogram of $\widehat{\beta}_{it}^L$ versus β_{it}^L 51

1.7 Comparison of Average Capital Elasticities 58

1.8 Comparison of Average Labor Elasticities 59

1.9 Dispersions of Elasticities across Firms 60

1.10 Dispersions of Elasticities through Time 61

1.11 Comparison of Dispersion of the Random Intercept 63

1.12 Estimated Correlation between the Random Intercept and Elasticities . . . 64

2.1 An Adaptive-Grid Algorithm 127

2.2 The Argmin Set in Θ 155

2.3 The Argmin Set in $\mathbb{S}^2$ 156

3.1 A Graphical Illustration of the Insurance Network of Nyakatoke 187

Acknowledgments

First and foremost, I am immensely grateful to Donald W. K. Andrews and Yuichi Kitamura, the co-chairs of my dissertation committee. Their continual guidance, support, and encouragement since my first year at Yale have made my Ph.D. a pleasant journey. Their wisdom and insight will continue to guide me for many years to come. I truly appreciate all those comments and suggestions Don hand-wrote for me. For several times the length of his comments exceeds that of my work itself. I learned from him to think differently, write properly, and present clearly. His expertise was invaluable in formulating the research questions and methodology. His standard on the quality of work is always the goal I aim to achieve. Yuichi taught me how to find good research topics and carry them out rigorously. Whether it was before the pandemic when I wrote research ideas on the whiteboard in his office or during the pandemic where I sat behind the screen, I always got detailed advice and helpful comments. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. The conversations I had with both of them were among the most enjoyable during my time at Yale. I cannot list the other one thousand things that they did for me, but I will always remember at the bottom of my heart. I feel extremely lucky and blessed to have both as my advisor not only at Yale, but also for the rest of my life.

Besides, I felt very honored to have Steve Berry on my committee. I benefited greatly from numerous discussions with him and learned how to do empirical research properly, among many other things. I also thank Fabrizio Zilibotti for commenting on my research and providing a teaching reference for me.

I would also like to express my gratitude to the many faculty members at Yale University. Since my first year, I have enjoyed interacting with Xiaohong Chen both academically and in life and received many advice and suggestions. Aside from time series econometrics, I learned the importance of perseverance in research from Peter Phillips. Edward Vytlacil and Tim Armstrong have always been attentive and accommodating, providing helpful comments on my projects. Giuseppe Moscarini and Joseph Altonji offered tremendous help on my job market preparation. Truman Bewley, Philip Haile, Zhen Huo, Mitsuru Igami, Kaivan Munshi, and Harrison Zhou have advised, taught, or encouraged me on repeated occasions to develop my academic career. Mitsuru Igami gave me helpful advice on my research and shared his career experience without reservation. I am also grateful to Kerry DeDomenico and Pamela O'Donnell for tirelessly helping me with countless administrative tasks. Beyond Yale, I thank Hao Wang, Miaojie Yu, and Li-an Zhou for their guidance and encouragement in my undergraduate and graduate studies.

I thank my coauthors, classmates, and friends who accompanied, helped, or encouraged me throughout the journey. I had the pleasure of working with Wayne Gao on several projects and interacting with him personally. The comradery built during those days and nights at CSSSI and other places will always be treasured. I also enjoyed discussing research with Yi Chen, Koohyun Kwon, Xiaosheng Mu, Guangjun Shen, Xinyang Wang, and Weijie Zhong. I thank Wayne Gao, Xiaosheng Mu, and Xinyang Wang for their full support on my job market preparation. I am very fortunate to have many supportive and caring friends, including Daisuke Adachi, Shuosong Chen, Chuan Du, Xiangliang Li, Oscar Soler Sanchez, Zhe Wang, Ge Zhang, Kevin Zhao, Ling Zhong, and Zihan Zhuo.

Last but not least, I genuinely appreciate the support and care my family provided throughout the journey. In 2015, they supported me in giving up a very decent job in the finance industry and pursuing my dream to become a scholar. They have always been sympathetic and stood by my side. Their companion means the world to me.

To my wife Shennan Song

For her love and support

Chapter 1

A Time-Varying Endogenous Random Coefficient Model with an Application to Production Functions

1.1 Introduction

Linear panel models with fixed coefficients have been a workhorse in empirical research. A leading example concerns production function estimation, where the output elasticities with respect to each input are assumed to be the same both across firms and through time (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015). But it is neither theoretically proven nor empirically verified that the coefficients should be fixed. For example, why would Apple have the same capital elasticity as Sony? Moreover, why would Apple in 2019 have the same labor elasticity as in 2020 when almost everyone is working from home? Restricting the coefficients to be constant can lead to biased estimates of important model parameters such as output elasticity with respect to capital or labor (León-Ledesma, McAdam, and Willman, 2010), and consequently misguided policy recommendations, e.g., income distribution policy, tax policy, among others. Therefore, it is crucial to properly account for the unobserved heterogeneity both across individuals and through time in panel models.

To accommodate the rich forms of unobserved heterogeneity in the economy, one may consider linear panel models with random coefficients that are either independent of the regressors or satisfy certain distributional assumptions joint with or given the regressors (Mundlak, 1978; Chamberlain, 1984; Wooldridge, 2005a). However, because of the agent's optimization behavior, it is rarely the case that one can justify any ex-ante distributional assumptions on the joint distribution of the random coefficients and the regressors. To see this, consider a firm with individually unique and time-varying output elasticities with respect to each input. Then, in each period, the firm chooses inputs by maximizing its expected profits *after* taking those heterogeneous elasticities into account. Consequently, the firm's heterogeneous elasticities enter its input choice decisions for each period in a potentially very complicated way, making it extremely difficult, if not impossible, to put any distributional assumption on the joint distribution of the random coefficients and the regressors.

The combination of unobserved heterogeneity and correlation between the regressors and the time-varying random coefficients in each period poses significant challenges for the analyst. The fact that the time-varying random coefficients are known to the agent when she optimally chooses the regressors but unobservable to the analyst gives rise to the classic simultaneity problem (Marschak and Andrews, 1944). Allowing the regressors to depend on the unobserved (to the econometrician) time-varying random coefficients in each period in an unknown and potentially complicated way makes traditional approaches inapplicable (Chamberlain, 1992; Arellano and Bonhomme, 2012; Graham and Powell, 2012; Laage, 2020). Therefore, a new method is needed to deal with the challenges discussed so far to identify and estimate the parameters of interest, e.g., the average partial effects (APE) (Chamberlain, 1984; Wooldridge, 2005b).

This paper proposes a time-varying endogenous random coefficient panel model where the regressors are allowed to depend on the random coefficients in each period, a feature called *time-varying endogeneity through the random coefficients.* The model is motivated by production function estimation, but can be applied to other important applications, e.g., consumer demand analysis, labor supply estimation, Engel curve analysis, among many others (Blundell, MaCurdy, and Meghir, 2007b; Blundell, Chen, and Kristensen, 2007a; Chernozhukov, Hausman, and Newey, 2019b). More specifically, the random coefficients in this paper are modeled as unknown and possibly nonlinear functions of a fixed effect of arbitrary dimension and a random shock that captures per-period shocks to the agent. In production function applications, one may interpret the fixed effect as managerial capability and the random shock as the R&D outcome. The modeling technique is based on the seminal paper of Graham and Powell (2012), with a major difference that will be discussed in detail in the model section. Then, the regressors are determined by the agent's optimization behavior and expressed as unknown and possibly complicated functions of the fixed effect, random shock, and exogenous instruments. For example, it can be the solution to a profit maximization problem with the fixed effect and random shock in the firm's information set. As a result, the firm's choices of inputs are functions of managerial capability, R&D outcome, and exogenous instruments.

For identification analysis, we use a sufficiency argument to control for the fixed effect without parametric assumptions, which enables one to construct a feasible control variable for the random shock given the sufficient statistic and the fixed effect, and subsequently to identify the moments of the random coefficients. More precisely, we use an exchangeability assumption on the conditional density of the vector of random shocks for all periods given the fixed effect to obtain a sufficient statistic that summarizes all of the time-invariant information about the individual fixed effect. Given this sufficient statistic, the agent's choice of regressors for a specific period is shown to not contain any additional information about the fixed effect. Thus, the density of the regressors for a specific period does not depend on the fixed effect given the sufficient statistic, allowing one to create a feasible control variable for the random shock given the sufficient statistic and the fixed effect. Finally, a sequential argument based on the independence result obtained in the first step, the feasible control variable constructed in the second step, and the law of iterated expectations (LIE), is adopted to identify the moments of the random coefficients. The intuition of the last step is after conditioning on the sufficient statistic and the feasible control variable, the residual variations in the regressors are exogenous. We further discuss how to extend the flexible identification argument to identify higher-order moments of the random coefficients, include vector-valued random shocks, incorporate group fixed effects, and allow exogenous shocks to the random coefficients.

It is worthwhile mentioning that the construction of the feasible control variable for the random shock in the presence of the fixed effect is not straightforward. Classical control function literature (Blundell and Powell, 2003) assumes one scalar-valued unobservable term in the first-step equation that determines the regressors. In this paper, however, there are two unobserved heterogeneity terms – the fixed effect of arbitrary dimension and the scalar-valued idiosyncratic shock – that both appear in the first-step equation. The inclusion of the fixed effect is crucial in applications such as production function estimation (Dhyne, Petrin, Smeets, and Warzynski, 2020). Therefore, one cannot directly apply the standard control function analysis (Newey, Powell, and Vella, 1999; Imbens and Newey, 2009). This paper shows how to exploit the sufficiency argument to construct a feasible control variable for the random shock in the presence of the unknown fixed effect.

The constructive identification analysis leads to multi-step series estimators for both conditional and unconditional moments of the random coefficients. We derive convergence rates and prove asymptotic normality for the proposed estimators. The new inference results build on existing ones for multi-step series estimators (Andrews, 1991; Newey, 1997; Imbens and Newey, 2009; Hahn and Ridder, 2013; Lee, 2018; Hahn and Ridder, 2019). The main deviations from the literature include that the object of interest is a partial mean process (Newey, 1994) of the derivative of the second-step estimator with a nonseparable first step, and that the last step of the three-step estimation is an unknown but only estimable functional of the conditional expectation of the outcome variable. Thus, one needs to take the estimation error from each of the three steps into consideration to obtain correct large sample properties.

Simulation results show that the proposed method can accurately estimate both the mean and the dispersion of the random coefficients. The mean of the random coefficients has long been the central object of interest in empirical research as it measures how responsive the outcome is to changes in regressors. The dispersion of the random coefficients may also be useful to answering policy-related questions. For example, to what extent is a new labor augmenting technology being diffused across firms? Such question can be answered based on the dispersion of labor elasticities estimated using the method of this paper. The results remain robust under various configurations of the data generating processes, including when one has different number of agents or periods in the data or use different orders of basis functions for estimation, and when an ex-post shock is added to the model.

Finally, the procedure is applied to comprehensive panel data on the production process for Chinese manufacturing firms. Specifically, we estimate the conditional means of the output elasticities with respect to capital and labor as well as the random intercept, all of which are allowed to be varying both across firms and through time. Three main findings emerge. First, larger capital, but smaller labor, elasticities on average than previous methods are obtained, which is more consistent with literature on the measurement of factor income shares (Bai, Qian, and Wu, 2008; Jia and Shen, 2016). Second, contrary to what fixed coefficients models imply, there are substantial variations in the elasticities of output with respect to capital and labor both across firms within each sector and for each firm through time. The results lead to a different interpretation of the data and policy implications than in the misallocation literature pioneered by Hsieh and Klenow (2009), who attribute all of the observed variation in input cost shares to output and input market distortions that drive wedges between the marginal products of capital and labor across firms. Third, we find the dispersion of the random intercept among firms is consistently larger than that obtained using the "proxy variable" based method of Olley and Pakes (1996), and show it is caused by negative correlations between the random intercept and output elasticities.

1.1.1 Related Literature

We review the three lines of literature that this paper is connected to. The first line concerns random coefficient models. See Hsiao (2014b) for a comprehensive survey. The closest paper to ours is Graham and Powell (2012), who also consider the identification of the APE in a linear panel model with time-varying random coefficients. Compared with the celebrated paper by Chamberlain (1992) who considers regular identification and derives the semiparametric variance bound of the APE, Graham and Powell (2012) show that the APE is irregularly identified when the number of periods equals the dimension of the regressors. However, as will be seen more clearly in the Section 1.2, their time stationarity assumption on the conditional distribution of idiosyncratic shocks given the whole vector of regressors effectively rules out time-varying endogeneity through the random coefficients. Therefore, their method does not directly apply here. Instead, we propose a different method for identification based on an exchangeability assumption and the control function approach.

Another closely related paper is Laage (2020), who also considers a correlated random coefficient linear panel model. Laage (2020) proposes a novel method for identification based on first differencing and the control function approach to identify APE when the number of periods is strictly larger than the dimension of the regressors. She allows for time-varying endogeneity through the residual term, but requires the random coefficient associated with each regressor to be time-invariant such that one can use first-differencing to cancel out the scalar fixed effect in the first step. As a result, her method does not apply to the setting considered in this paper. Similarly to Laage (2020), Arellano and Bonhomme (2012) also consider a time-invariant random coefficient model. They exploit information on the time dependence of the residuals to obtain identification of variances and distribution functions of the random coefficients. Their model assumptions and analysis are very different from ours. In addition to linear models, random coefficients are also widely used in discrete choice models (Berry, Levinsohn, and Pakes, 1995b; Bajari, Fox, and Ryan, 2007; Dubé, Fox, and Su, 2012; Gautier and Kitamura, 2013).

The second line of research concerns identifiability of models with unobserved heterogeneity. The concept of exchangeable sequences dates back to Jonnson (1924), and has been used in many papers in economics (McCall, 1991; Kyriazidou, 1997; Altonji and Matzkin, 2005). The closest paper in this aspect to our work is Altonji and Matzkin (2005), who assume the conditional density of the fixed effect and random shock given the regressors for all periods is a symmetric function of the regressors. This assumption is not applicable to our model, and we propose an arguably more primitive exchangeability condition on the conditional density of the random shocks for all periods given the individual fixed effect. We show how to obtain a sufficient statistic for the fixed effect, and subsequently identify moments of the random coefficients using the new exchangeability condition.

Another method used in this paper is related to the control function approach in triangular models (Newey, Powell, and Vella, 1999; Florens, Heckman, Meghir, and Vytlacil, 2008; Imbens and Newey, 2009; Torgovitsky, 2015; D'Haultfœuille and Février, 2015). The construction of the feasible control variable for the random shock in the identification analysis is built upon Imbens and Newey (2009), who assume a nonseparable first-step equation that determines the regressors and suggest a conditional cumulative distribution function (CDF) based approach for identification. The main difference between our model and theirs is in the first-step equation of the model considered in this paper, there are two unobserved heterogeneity terms comprised of a fixed effect of arbitrary dimension and a idiosyncratic shock, whereas Imbens and Newey (2009) assume one scalar-valued unobserved shock in their first-step equation. Therefore, one cannot directly apply their method to the problem considered in this paper because the control variable constructed using their method is infeasible. Instead, we use the implied conditional independence result from the sufficiency argument to construct a feasible control variable for the random shock given the fixed effect and the sufficient statistic. More recently, Kitamura and Stoye (2018) propose and implement a control function approach to account for endogenous expenditure in a nonparametric analysis of random utility models.

The third line of research concerns production function estimation. Production functions are one of the most fundamental components of economic analysis. Classical literature (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015) use a fixed coefficient linear model while allowing for a scalar-valued time-varying productivity shock. The endogeneity problem is caused by the fact that the productivity shock is unobserved by the econometrician but known to the firm when making input choice decisions. The key identification idea in this literature is to use some choice variable of the firm to uncover the productivity. Specifically, they suggest a "proxy variable" approach where investment (Olley and Pakes, 1996) or material (Levinsohn and Petrin, 2003) is assumed to be an invertible function of the productivity shock given other observables. Based on the invertibility condition, one can uncover the productivity as a nonparametric function of observables. Then, under the assumption that the innovation in productivity follows a first-order Markov process, an orthogonality condition between the innovation in productivity and lagged input choices can be formed to identify the output elasticities with respect to each input. The main difference between our paper and theirs is that we allow for time-varying endogeneity through not only the random intercept, but also output elasticities modeled as random coefficients. We also include a fixed effect of arbitrary dimension and propose a different identification strategy. Ackerberg, Chen, Hahn, and Liao (2014) study the asymptotic efficiency of semiparametric two-step GMM estimators and apply their method to production function estimation with fixed coefficients. Bang, Gao, Postlewaite, and Sieg (2020) develop a new method for estimating production functions when the inputs are partially latent.

There is some recent work trying to include a fixed effect into the fixed coefficient linear production model (Lee, Stoyanov, and Zubanov, 2019; Abito, 2020).

A couple of innovations have been made recently to relax the assumption of fixed output elasticities with respect to each input. Kasahara, Schrimpf, and Suzuki (2015) analyze Cobb-Douglas (C-D) production function with heterogeneous but time-invariant output elasticities modeled as finite mixtures. Li and Sasaki (2017) analyze C-D production function with heterogeneous output elasticities modeled as unknown functions of a latent technology term. Their analysis hinges on a key assumption that there is a one-to-one mapping between the latent technology term and the ratio of the two intermediate goods. The model assumptions and technique are very different from ours. Doraszelski and Jaumandreu (2018) propose an empirical strategy to analyze constant elasticity of substitution production function with labor augmenting productivity, which allows for multi-dimensional heterogeneity and non-neutral productivity. Fox, Haddad, Hoderlein, Petrin, and Sherman (2016) model the output elasticities as random walk processes and assume the input choice decisions are made in period one. They apply their method to the data for Indian manufacturing firms and find that there is significant variation in the elasticities both across firms and through time. The method proposed in this paper is different from theirs as we do not assume random walk for the innovation of the random coefficients and the firms are allowed to choose their inputs in each period.

In their influential paper, Gandhi, Navarro, and Rivers (2020) (GNR20) argue that the proxy variable based method is not sufficient for identification without functional form restrictions. They show how to use the first-order conditions from a firm's profit maximization problem to achieve nonparametric identification of the production function. Similarly, Demirer (2020) models the production function non-parametrically and assumes it satisfy a homothetic separability condition. He also assumes that the material per capital is a strictly monotonic function of labor augmenting productivity only, but not the Hicks neutral productivity. He shows that while the functional form of the production function and output elasticity with respect to capital are not identified, output elasticities with respect to labor and material are identified via cost minimization. Chen, Igami, Sawada, and Xiao (2020) study how ownership affects productivity by extending GNR20's framework. The assumptions and method of this paper are very different from those mentioned above.

The rest of this paper is organized as follows. Section 1.2 introduces the main model specification and assumptions. Section 1.3 presents the key identification strategy. Series estimators are provided in Section 1.4, together with their asymptotic properties. Section 1.5 contains a simulation study. In Section 1.6, we apply our method to panel data for the Chinese manufacturing firms to estimate their production functions. Finally, Section 1.7 concludes. All the proofs and an index of notation are presented in the Appendix.

1.2 Model

In this section, we present a time-varying endogenous random coefficient (TERC) model where the regressors can depend on the time-varying random coefficients in each period, a critical feature that appears in many important applications in economics. We provide three applications that share this feature, followed by assumptions on model primitives.

Consider the following triangular simultaneous equations model with time-varying random coefficients:

$$Y_{it} = X_{it}^{'}\beta_{it} + \varepsilon_{it}, \tag{1.1}$$

$$\beta_{it} = \beta\left(A_i, U_{it}\right), \tag{1.2}$$

$$X_{it} = g\left(Z_{it}, A_i, U_{it}\right), \tag{1.3}$$

where:

- $i \in \{1, ..., n\}$ denotes n decision makers and $t \in \{1, ..., T\}$ denotes $T \geq 2$ time periods.

- $Y_{it} \in \mathbb{R}$ represents the scalar-valued outcome variable for agent i in period t. One may interpret it as total output for firm i in year t in production function applications.

- $X_{it} \in \mathbb{R}^{d_X}$ is a vector of choice variables of the i^{th} decision maker in period t with the constant 1 as its last coordinate. It can include, for example, capital, labor and the constant 1, in the context of production function estimation.

- $Z_{it} \in \mathbb{R}^{d_Z}$ is a vector of exogenous instruments that affects the choice of X_{it} and is independent of (A_i, U_{it}). E.g., Z_{it} can include input prices in the context of production function estimation.

- A_i represents a fixed effect of arbitrary dimension. The fixed effect A_i can be interpreted, for example, as the managerial capability of firm i in production function applications.

- $U_{it} \in \mathbb{R}$ is a scalar-valued continuously distributed it-specific random shock term, which captures idiosyncratic shock that is correlated with input choices in each period such as an R&D shock to firm i in period t.

- $\beta_{it} \in \mathbb{R}^{d_X}$ is a vector of random coefficients, the central object of interest. They are modeled as unknown and possibly nonlinear functions of A_i and U_{it}. In production function applications, β_{it}'s are the output elasticities with respect to each input of X_{it}. A key feature here is each coordinate of β_{it} varies both across i and through t.

- $\varepsilon_{it} \in \mathbb{R}$ is a scalar-valued error term with mean zero. It can be considered as the measurement error or ex-post shock.

- $g(\cdot)$ is a vector-valued function of (Z_{it}, A_i, U_{it}) that determines each coordinate of the choice variables X_{it}. For example, capital input K_{it} may be determined by its first coordinate, $g^{(1)}(Z_{it}, A_i, U_{it})$, while labor input L_{it} equals $g^{(2)}(Z_{it}, A_i, U_{it})$, the second coordinate of $g(\cdot)$.

To clarify the information structure of the model, (Y_{it}, X_{it}, Z_{it}) are data and observable to both the econometrician and the firm, whereas (A_i, U_{it}) are only observable to the firm, but not to the econometrician. The functional form of $g(\cdot)$ and $\beta(\cdot)$ are only known to the firm, but not to the econometrician. The ex-post shock ε_{it} is unobservable to the firm when it makes input choice decisions in each period.

Model (1.1)–(1.3) naturally arises in many economic applications. We mention a few in this section.

Example 1.1. The leading example is production function estimation. Suppose firm i in period t observes its production function (1.1) in the classic C-D form, which is the workhorse model in the literature and is employed by Olley and Pakes (1996); Levinsohn and Petrin (2003); Ackerberg, Caves, and Frazer (2015), among many other papers. The firm also observes its input prices Z_{it} and input elasticities β_{it}, the latter of which is a function of the managerial capability A_i and the random R&D outcome U_{it}, both known to the firm. Then, the firm chooses capital, labor and materials by solving a profit maximization problem using the information of (Z_{it}, A_i, U_{it}), obtaining (1.3) as a consequence.

Example 1.2. Another example is Engel curve estimation. Suppose the budget share of gasoline Y_{it} for household i at time t is a function of gas price and total expenditure in (1.1). Here β_{it} is modeled as a function of the household fixed effect and an idiosyncratic wealth shock, and captures how elastic gasoline demand is with respect to total expenditure and gas price, respectively. Given the fixed effect, random wealth shock, and an instrument of gross income of the head of household Z_{it}, household i optimally chooses its gas price and total expenditure budget by solving a utility maximization problem, leading to (1.3) as a result. See Blundell, Chen, and Kristensen (2007a) for more details of the endogeneity issue in Engel curve estimation.

Example 1.3. The third example concerns labor supply estimation. Suppose individual i has a linear labor supply function in the form of (1.1), where Y_{it} is the number of annual hours worked and X_{it} includes the endogenous hourly wage and other exogenous demographics. The coordinate of β_{it} that corresponds to wage is the key object of interest which quantifies how labor supply responds to wage rate variations over time. Then, given exogenous instruments Z_{it} such as the minimum wage in the county or non-labor income, individual capability A_i, and random health shocks U_{it} to the individual, agent i chooses the job that provides a wage that is the solution to her utility maximization problem, leading to (1.3). See Blundell, MaCurdy, and Meghir (2007b) for more details on labor supply estimation.

The time-varying correlation between X_{it} and β_{it} in these examples highlights the prevalence and importance of *time-varying endogeneity through the random coefficients.* Nonetheless, models in this literature do not allow for this feature. Graham and Powell (2012) propose a panel model with time-varying random coefficients. Using their notation, they model $\beta_{it} = b^*(A_i, U_{it}) + d_t(U_{i,2t})$ and assume $U_{i,2t} \perp (\mathbf{X}_{\boldsymbol{i}}, A_i)$ where $\mathbf{X}_{\boldsymbol{i}} = (X_{i1}, .., X_{iT})^{'}$. Thus, the random coefficient β_{it} is time-varying and correlated with $\mathbf{X}_{\boldsymbol{i}}$ via (A_i, U_{it}). However, they impose a time stationarity assumption on the conditional distribution of U_{it} given $(\mathbf{X}_{\boldsymbol{i}}, A_i)$:

$$U_{it} | \, \mathbf{X}_{\boldsymbol{i}}, A_i \sim_d U_{is} | \, \mathbf{X}_{\boldsymbol{i}}, A_i, \text{ for } t \neq s, \tag{1.4}$$

which effectively rule out time-varying endogeneity through the random coefficients. To see why, omit $U_{i,2t}$ for now since it is exogenous. Consider a simple example where the number of periods $T = 2$ and the true data generating processes of β_{it} and X_{it} are

$$\beta_{it} = A_i + U_{it}, \; X_{it} = \beta_{it} \tag{1.5}$$

Then, suppose one observes $X_{i2} > X_{i1}$ in the data, which implies

$$\mathbb{E}\left[U_{i2} | \, \mathbf{X}_{\boldsymbol{i}}, A_i\right] > \mathbb{E}\left[U_{i1} | \, \mathbf{X}_{\boldsymbol{i}}, A_i\right], \tag{1.6}$$

thus violating (1.4). From this simple example, it is clear that under (1.4) one cannot allow X_{it} to depend on β_{it} in each period such that one may infer distributional characteristics about U_{it} given $\mathbf{X}_{\boldsymbol{i}}$, a feature that is important to applications such as production function estimation. As can be seen from (1.3), we allow such dependence between X_{it} and U_{it} in each period. Similarly, Chernozhukov, Hausman, and Newey (2019b) impose a time stationarity assumption on the conditional mean of the random coefficients given $\mathbf{X}_{\boldsymbol{i}}$, again ruling out time-varying endogeneity through the random coefficients. Arellano and Bonhomme (2012) consider time-invariant random coefficients that are correlated with X_{it}. Similarly to Arellano and Bonhomme (2012), Laage (2020) also models the random coefficients to be time-invariant and allows time-varying endogeneity only through the residual term.

In addition to the time-varying endogeneity of the regressors through the random coefficients, model (1.1)–(1.3) also features a nonseparable first step that determines X_{it} and a fixed effect A_i that enters both the first step (1.3) and the second step (1.1) nonlinearly. The nonseparability of $g(\cdot)$ in the instrument Z_{it}, fixed effect A_i, and random shock U_{it} appears naturally due to the agent's optimization behavior. For example, in C-D production functions firms choose their inputs by maximizing their expected profits without the knowledge of ε_{it}, leading to a nonseparable input choice function $g(\cdot)$. The nonlinearity of the fixed effect A_i appears in two places: (1) the unknown random coefficients $\beta(A_i, U_{it})$ could be nonlinear in A_i and (2) the first-step equation $g(\cdot)$ could be nonlinear in β_{it}. Allowing a nonseparable first step $g(\cdot)$ and a nonlinear fixed effect A_i significantly improves the flexibility and thus widens the applicability of the model, however at the cost of greater analytical challenges for identification. For example, the usual demeaning or first differencing techniques no longer apply to the model (1.1)–(1.3). Nonetheless, we show how to achieve identification via a sufficiency argument in the next section.

It is worthwhile mentioning that A_i and U_{it} appear in both the first-step equation (1.3) that determines X_{it} and the second-step equation (1.1) that determines Y_{it}. This is again a feature motivated by economic applications, because agents choose X_{it} optimally based on the *complete information* of (A_i, U_{it}), both of which affect the outcome Y_{it}. It is different from traditional triangular simultaneous equations models (Newey, Powell, and Vella, 1999; Imbens and Newey, 2009) which assume in (1.3) there is only one unknown scalar that is arbitrarily correlated with (A_i, U_{it}), which effectively assumes the agent has *incomplete information* of (A_i, U_{it}) when choosing X_{it}. The complete information assumption is arguably more realistic based on agent's optimization behavior, however makes identification challenging because now one has two unknown terms A_i and U_{it} in both (1.1) and (1.3). Thus, the control function approach suggested in Imbens and Newey (2009) does not directly apply. Instead, we show how to deal with both unobserved heterogeneity terms via a sequential argument in the identification section.

It should be pointed out that the fixed effect A_i, modeled as an arbitrary dimensional object, effectively incorporates unobserved variations in the distributions of the idiosyncratic shocks U_{it}. For example, if the joint distribution of $(U_{i1}, .., U_{iT})$ is F_i which does not depend on time, then the whole function F_i can be incorporated as part of the fixed effect A_i, which may lie in a vector of infinite-dimensional functions. F_i captures a form of heteroskedasticity specific to each agent, and our method is robust to such forms of heterogeneity in error distributions without the need to specify F_i.

Before proceeding to the assumptions, we briefly discuss some extensions to the model (1.1)–(1.3). First, suppose $U_{it} = \left(U_{it}^{(1)}, U_{it}^{(2)}\right)$ and X_{it} is two-dimensional. Then, we can allow β_{it} to depend on both A_i and $\left(U_{it}^{(1)}, U_{it}^{(2)}\right)$ and let each of the two coordinates of X_{it} depend on A_i and a different coordinate of U_{it}. For example, let $X_{it}^{(1)}$ depend on $\left(A_i, U_{it}^{(1)}\right)$ and $X_{it}^{(2)}$ depend on $\left(A_i, U_{it}^{(2)}\right)$. The modification is allowed and the identification argument can go through as given. Second, it is possible to follow Graham and Powell (2012) and include an exogenous $U_{2,it}$ in β_{it} to capture exogenous shocks to agents i at period t. Third, similarly to Arellano and Bonhomme (2012), both exogenous and endogenous regressors X_{it} can be included in the model (1.1) that are associated with constant coefficients β.

Next, we provide a list of assumptions on model primitives required for the subsequent identification argument, and discuss them in relation to the model (1.1)–(1.3).

Assumption 1.1 (**Monotonicity of $\boldsymbol{g}(\cdot)$**). *At least one coordinate of $g(Z, A, U)$ is known to be strictly monotonic and continuously differentiable in U, for every realization of $(Z, A) \in \mathcal{Z} \times \mathcal{A}$.*

Assumption 1.1 requires at least one coordinate of the unknown function $g(Z, A, U)$ defined in (1.3) that determines one element of X, say labor choice in production function applications, to be strictly monotonic in U on its support for every realization of (Z, A). Without loss of generality (wlog), assume the first coordinate of g, denoted by $g^{(1)}$, satisfies Assumption 1.1. Then, the assumption implies that there is a one-to-one mapping between the first coordinate of X and U given (Z, A), which is used to establish an exchangeability property and subsequently construct a feasible control variable for U.

It is worthwhile mentioning that strict monotonicity in U for all coordinates of g is not needed because a single U appears in both (1.1) and (1.3). We show in (1.67) that Assumption 1.1 suffices to prove the exchangeability condition (1.60), an essential step for the analysis. If one has a model with a multi-dimensional U in (1.1) and each coordinate of U appearing in one equation of (1.3), then for the proposed method to work, all of the coordinates of g are required to be strictly monotonic in U to properly control for the unobserved heterogeneity in the model.

Assumption 1.1 is mild in the sense that it is satisfied in many applications and models. For example, in production function applications one may interpret U as R&D outcome. Then, the firm takes advantage of a better R&D outcome (larger U) by purchasing more machines and hiring more workers, leading to a larger choice of each coordinate of X_{it} defined as the vector of capital and labor. Thus, Assumption 1.1 is satisfied. As in Newey, Powell, and Vella (1999), the assumption is automatically satisfied if $g(\cdot)$ is linear in U, but allows for more general forms of non-additive relations. An assumption similar to Assumption 1.1 is also imposed in Imbens and Newey (2009).

Assumption 1.2 (Exchangeability). *The conditional probability density function of $U_{i1}, ..., U_{iT}$ given A_i wrt Lebesgue measure is continuous in $(u_{i1}, .., u_{iT})$ and exchangeable across t, i.e.*

$$f_{U_{i1},...,U_{iT}|A_i}\left(u_{i1}, ..., u_{iT}|\, a_i\right) = f_{U_{i1},...,U_{iT}|A_i}\left(u_{it_1}, ..., u_{it_T}|\, a_i\right), \tag{1.7}$$

where $(t_1, ..., t_T)$ is any permutation of $(1, ..., T)$.

Assumption 1.2 requires that the conditional density of $(U_{i1}, ..., U_{iT})$ given A_i is invariant to any permutation of time. To provide a simple example when it holds, suppose $T = 2$, $U_{it} = A_i + \kappa_{it}$ for $t = 1, 2$ where κ_{it} are iid through time and independent of A_i. Then, Assumption 1.2 is satisfied and U_{i1} and U_{i2} are correlated. In this sense, Assumption 1.2 is milder than requiring U_{it} to be iid through time. Note that the simple example corresponds to the standard equicorrelated random effects specification due to Balestra and Nerlove (1966) from the panel analysis literature. Another attractive feature of Assumption 1.2 is that it does not rely on parametric assumptions on the joint density of $(\mathbf{U_i}, A_i)$.

It is worthwhile emphasizing that Assumption 1.2 requires exchangeability in the conditional density of U_{it}'s *given* A_i, thus allowing arbitrary correlation between A_i and U_{it} which is an important feature in many economic applications. For example, in production function estimation, one may expect that the better managerial capability a firm has, the greater chance a positive R&D outcome shall occur. Such correlation is allowed under Assumption 1.2.

Altonji and Matzkin (2005) also impose an exchangeability assumption (Assumption 2.3 in their paper) to achieve identification in a nonparametric regression setting. Compared with their exchangeability condition, Assumption 1.2 avoids directly imposing distributional assumptions on the conditional density of U_{it} given $\mathbf{X_i}$ and is arguably more primitive. More precisely, Altonji and Matzkin (2005) denote $\Phi_{it} \coloneqq (A_i, U_{it})$ and assumes

$$f_{\Phi_{it}|X_{i1},..,X_{iT}}\left(\varphi_{it}|\, x_{i1},..,x_{iT}\right) = f_{\Phi_{it}|X_{i1},..,X_{iT}}\left(\varphi_{it}|\, x_{it_1},..,x_{it_T}\right), \tag{1.8}$$

where $(t_1, ..., t_T)$ is any permutation of $(1, ..., T)$. There are two main differences between (1.7) and (1.8). First, Altonji and Matzkin (2005) do not distinguish A_i from U_{it} in the definition of Φ_{it}, whereas A_i and U_{it} play different roles in this paper. The difference between A_i and U_{it} could be important in applications such as production function estimation because they have different economic interpretations and implications. Second, and more importantly, the exchangeability assumption (1.8) requires the value of the conditional density function of Φ_{it} given regressors $(X_{i1}, .., X_{iT})$ does not depend on the order in which the regressors are entered into the function. In (1.7), the requirement is that the conditional density of $(U_{i1}, .., U_{iT})$ given A_i is exchangeable in $(U_{it}, .., U_{iT})$, which is on the model primitives (A, U) rather than on (X, A, U) as in (1.8). Moreover, it could be challenging to justify (1.8) since Φ_{it} includes U_{it} which determines X_{it} by (1.3), but not X_{is} for $s \neq t$, which creates asymmetry between X_{it} and X_{is} in (1.8).

In light of these differences and observations, we distinguish A_i from U_{it} in this paper and impose the exchangeability assumption on the conditional probability density function (pdf) of U_{it} given A_i in (1.7). In the next section, we use (1.7) to prove an exchangeability condition (1.15) on the conditional pdf of A_i wrt the elements (X_{it}, Z_{it}). We show that the new exchangeability condition (1.15) guarantees the existence of a vector-valued function W_i symmetric in the elements of $(\mathbf{X_i}, \mathbf{Z_i})$, such that conditioning on W_i, the fixed effect A_i is independent of (X_{it}, Z_{it}) for any fixed t.

Assumption 1.3 (**Random Sampling, Compact Support, and Exogeneity of $\boldsymbol{Z}$**)**.** $(\mathbf{X_i}, \mathbf{Z_i}, Y_i, A_i, U_i, \varepsilon_i)$ *is iid across* $i \in \{1, ..., n\}$ *with* $n \to \infty$ *and fixed* $T \geq 2$*. The support of* (X_{it}, Z_{it}) *is compact.* $Z_{it} \perp (A_i, U_{it})$*.*

The first part of Assumption 1.3 is a standard assumption on random sampling. Notice that only a short panel is required. We focus on cross-sectional asymptotics with the number of agents getting larger ($n \to \infty$), while the number of time periods T is held fixed. After obtaining W_i for each individual, which requires $T \geq 2$, one can treat each t-specific subsample across individuals separately in the identification analysis and one does not need to do inter-temporal differencing as in Graham and Powell (2012) or Laage (2020).

Assumption 1.3 can be relaxed to allow exogenous macro shocks in the model. One can still obtain consistency and normality results by using *conditional* law of large numbers and central limit theorems by conditioning on the sigma algebra generated by all of the random variables common to each individual i but specific to period t. This methodological convenience brings about significant computational advantages because parallel computing can be used to deal with each t-specific subsample simultaneously.

The second part of Assumption 1.3 requires the support of (X_{it}, Z_{it}) to be compact, which is required for the Weierstrass approximation theorem in the proof to show that W_i is a sufficient statistic for A_i. The last part of Assumption 1.3 requires the exogenous instrument Z_{it} to be independent of (A_i, U_{it}) unconditionally. In production function applications, it is satisfied when Z_{it} is chosen to be, for example, input prices. It is worthwhile mentioning that in the identification section, we impose another conditional independence assumption between Z_{it} and (A_i, U_{it}) conditioning on a sufficient statistic for A_i. The reason for deferring the conditional independence assumption is because we need first obtain the sufficient statistic, which is summarized in Lemma 1.1.

1.3 Identification

In this section, we show how to identify the first-order moments of the random coefficient β_{it}. To motivate the method, consider the classical linear regression model *without* random coefficients

$$Y_{it} = X_{it}^{'}\beta + \varepsilon_{it}. \tag{1.9}$$

Under the mean independence assumption that $\mathbb{E}\left[\left.\varepsilon_{it}\right| X_{it}\right] = 0$, one may take the conditional expectation on both sides of (1.9) given X_{it} to obtain

$$\mathbb{E}\left[\left.Y_{it}\right| X_{it}\right] = X_{it}^{'}\beta, \tag{1.10}$$

and subsequently exploit the exogenous variation in X_{it} to identify β. For example, taking the partial derivative on both sides of (1.10) wrt X_{it} identifies

$$\beta = \partial\mathbb{E}\left[\left.Y_{it}\right| X_{it}\right] / \partial X_{it} \tag{1.11}$$

provided there is enough variation in X_{it}. Since $\mathbb{E}\left[\left.Y_{it}\right| X_{it}\right]$ is an identifiable object from the data, β is thus identified.

But the identification argument (1.9)–(1.11) does not go through when β is random and X_{it} depends on β_{it} in each period. To see this, since β_{it} is now random and correlated with X_{it}, if one follows the analysis (1.9)–(1.11), instead of (1.10) she obtains

$$\mathbb{E}\left[\left.Y_{it}\right| X_{it}\right] = X_{it}^{'}\mathbb{E}\left[\left.\beta_{it}\right| X_{it}\right]. \tag{1.12}$$

If one follows (1.11) to take partial derivative wrt X_{it}, it will simultaneously change the conditional expectation $\mathbb{E}\left[\left.\beta_{it}\right| X_{it}\right]$ because the conditional pdf of β_{it} given X_{it} is changed. In this sense, the variation in X_{it} is no longer exogenous even though ε_{it} is still exogenous and satisfies $\mathbb{E}\left[\left.\varepsilon_{it}\right| X_{it}\right] = 0$, exactly because β_{it} is correlated with X_{it}.

Therefore, for identification the goal here is to find a set of feasible random variables that can control for the time-varying endogeneity through the random coefficients, such that after conditioning on these variables the residual variation in X_{it} is exogenous and can identify the moments of the random coefficients. More precisely, we show how to construct control variables in

$$\mathbb{E}\left[Y_{it} \middle| X_{it}, \mathrm{cv} \right] = X_{it}' \mathbb{E}\left[\beta_{it} \middle| \mathrm{cv} \right] \tag{1.13}$$

labeled as “cv” (control variable), such that conditioning on these variables, the residual variation in X_{it} is exogenous and can be used to identify the first-order moments of β_{it} as in (1.11).

The analysis is divided into four steps. First, we obtain a key sufficient statistic W_i for the fixed effect A_i via the exchangeability condition (1.7). Second, we construct a feasible variable V_{it} based on the sufficient statistic W_i and show that V_{it} is a control variable for U_{it} given (A_i, W_i). Third, if A_i is known, we prove the residual variation in X_{it} conditioning on (A_i, V_{it}, W_i) is exogenous and can be used to identify the first-order moments of β_{it}. Lastly, we deal with the unknown A_i via a LIE argument and show the “cv” vector in (1.13) to be the feasible (V_{it}, W_i) .

Step 1: Sufficient Statistic for $\boldsymbol{A_i}$

To construct a sufficient statistic for A_i, we exploit the exchangeability condition (1.7) and prove the following lemma.

Lemma 1.1 (**Sufficient Statistic for $\boldsymbol{A_i}$**). *Suppose that Assumptions 1.1–1.3 are satisfied. Then, one can construct a feasible vector-valued function $W_i := W\left(\mathbf{X_i}, \mathbf{Z_i}\right)$ that is symmetric in the elements of $\left(\mathbf{X_i}, \mathbf{Z_i}\right)$ and satisfies*

$$f_{A_i|X_{it},Z_{it},W_i}\left(a_i \middle| x_{it}, z_{it}, w_i \right) = f_{A_i|W_i}\left(a_i \middle| w_i \right) \tag{1.14}$$

for any fixed $t \in \{1, ..., T\}$.

Lemma 1.1 exemplifies that one can exploit the panel data structure to control for complicated unobserved individual heterogeneity terms. The intuition of Lemma 1.1 is that

W_i "absorbs" all the time-invariant information in the observable variables $\mathbf{X}_{\boldsymbol{i}}$ and $\mathbf{Z}_{\boldsymbol{i}}$. Given W_i, any t-specific X_{it} or Z_{it}, e.g., X_{i1}, Z_{i1}, does not contain any additional information about A_i. Therefore, one can exclude them from the conditioning set in (1.14) following the sufficiency argument. It is also worth emphasizing that Lemma 1.1 only concerns the density of the fixed effectA_i, not the random shock U_{it}, whereas Assumption 2.1 of Altonji and Matzkin (2005) concerns the joint distribution of $\mathit{\Phi}_{it} := (A_i, U_{it})$.

To see an example of W_i, suppose $T = 2$ and both X_{it} and Z_{it} are scalars. Then, one example of such W_i is $T^{-1} \sum_t (X_{it}, Z_{it}, X_{it}^2, Z_{it}^2, X_{it}Z_{it})$. See Weyl (1939) for a detailed illustration on how to construct W_i. Notice that we do not impose any distributional assumption on the conditional density of A_i given (X_{it}, Z_{it}) in Lemma 1.1. With that said, ex-ante information about A_i can be incorporated to reduce the number of elements appearing in W_i. For example, when one knows the probability distribution of A_i belongs to exponential family, such information can greatly simplify W_i. See Altonji and Matzkin (2005) for a more detailed discussion.

We prove Lemma 1.1 in Appendix 1.A. The key to the proof involves a change of variables step that uses the exchangeability condition (1.7) to establish that the conditional density of A_i given $(\mathbf{X}_{\boldsymbol{i}}, \mathbf{Z}_{\boldsymbol{i}})$ is exchangeable through time, i.e.,

$$\begin{aligned} & f_{A_i|X_{i1},Z_{i1},..,X_{iT},Z_{iT}} \left(a_i \middle| x_{i1}, z_{i1}, .., x_{iT}, z_{iT} \right) \\ = & f_{A_i|X_{i1},Z_{i1},..,X_{iT},Z_{iT}} \left(a_i \middle| x_{it_1}, z_{it_1}, .., x_{it_T}, z_{it_T} \right), \end{aligned} \tag{1.15}$$

where $(t_1, ..., t_T)$ is any permutation of $(1, ..., T)$. It is worth noting that the inclusion of Z_{it}'s in the conditioning set in (1.15) is necessary for the change of variable argument to go through. The exogeneity of Z_{it} is also crucial for the argument. Then, following Altonji and Matzkin (2005) one can construct a vector-valued function W_i symmetric in the elements of $(\mathbf{X}_{\boldsymbol{i}}, \mathbf{Z}_{\boldsymbol{i}})$, using the Weierstrass approximation theorem and the fundamental theorem of symmetric functions, such that (1.14) hold.

Lemma 1.1 serves as the key device in obtaining the identification of moments of the random coefficients β_{it}. In the following analysis, we first construct a *feasible* control variable for U_{it} given A_i in Step 2. Then, we exploit the exogenous variation in X_{it} using the exclusion condition (1.14) to identify moments of β_{it} in Step 3.

Step 2: Feasible Control Variable for $\boldsymbol{U}_{it}$

Given the nonseparable feature of the first-step $g(\cdot)$ function in (1.3), one may wish to use the method proposed in Imbens and Newey (2009) to construct a control variable for U_{it} and subsequently identify moments of β_{it} by exploiting the residual variation in X_{it} given the control variable. However, one cannot directly apply their technique in the current setting because the model considered in this paper has two unobserved heterogeneity terms A_i and U_{it}, whereas in their setting there is only one.

To see this more clearly, for brevity of exposition let X_{it} be a scalar that satisfies Assumption 1.1. Suppose one naively follows Imbens and Newey (2009) to exploit the strict monotonicity of $g(\cdot)$ in U given (Z, A) and constructs a conditional CDF $F_{X_{it}|Z_{it},A_i}(X_{it}|\, Z_{it}, A_i)$, which under Assumption 1.1 equals $F_{U_{it}|A_i}(U_{it}|\, A_i)$, as the control variable for U_{it}. Then, two issues arise. First, $F_{X_{it}|Z_{it},A_i}(X_{it}|\, Z_{it}, A_i)$ is not feasible because A_i is unknown. Thus, one cannot consistently estimate it from data. Second, unlike the unconditional CDF $F_{U_{it}}(U_{it})$ in their setting which is a one-to-one mapping of U_{it}, the conditional CDF $F_{U_{it}|A_i}(U_{it}|\, A_i)$ is a function of both A_i and U_{it}. Therefore, one can not uniquely pin down U_{it} using $F_{U_{it}|A_i}(U_{it}|\, A_i)$ if A_i is unknown. For example, given a fixed value c that $F_{U_{it}|A_i}(U_{it}|\, A_i)$ takes, there can be many U_{it}'s that satisfies $F_{U_{it}|A_i}(U_{it}|\, A_i) = c$, exactly because A_i is not fixed. Therefore, one needs to explicitly deal with unknown A_i when constructing a control variable for U_{it}.

In this step, we deal with the first issue that $F_{X_{it}|Z_{it},A_i}(X_{it}|\, Z_{it}, A_i)$ is infeasible and show how to construct a feasible variable that can be used later on to form a one-to-one mapping of U_{it}. The idea is to use the sufficient statistic W_i in Lemma 1.1 to get rid of A_i from the conditioning set of the conditional CDF $F_{X_{it}|Z_{it},A_i}\left(X_{it}|\, Z_{it}, A_i\right)$. More specifically, the sufficiency condition (1.14) implies $A_i \perp \left(X_{it}, Z_{it}\right)|\, W_i$, which further implies $X_{it} \perp A_i|\, \left(Z_{it}, W_i\right)$, i.e.,

$$f_{X_{it}|Z_{it},A_i,W_i}\left(x_{it}|\, z_{it}, a_i, w_i\right) = f_{X_{it}|Z_{it},W_i}\left(x_{it}|\, z_{it}, w_i\right). \tag{1.16}$$

The key observation here is the right hand side (rhs) of (1.16) is feasible since it only involves known or estimable objects from data. Suppose the first coordinate of X_{it} denoted by $X_{it}^{(1)}$ satisfies Assumption 1.1. Then, one can construct

$$V_{it} := F_{X_{it}^{(1)}\big|Z_{it},W_i}\left(X_{it}^{(1)}\Big|\, Z_{it}, W_i\right) \tag{1.17}$$

and use (1.16) to deduce that

$$V_{it} = F_{X_{it}^{(1)}\big|Z_{it},A_i,W_i}\left(X_{it}^{(1)}\Big|\, Z_{it}, A_i, W_i\right). \tag{1.18}$$

Next, we use Assumption 1.1 and the next assumption to prove

$$F_{X_{it}^{(1)}\big|Z_{it},A_i,W_i}\left(X_{it}^{(1)}\Big|\, Z_{it}, A_i, W_i\right) = F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right), \tag{1.19}$$

the rhs of which plays an essential role to the subsequent identification analysis.

Assumption 1.4 (**Conditional Independence**). $Z_{it} \perp U_{it}|\, A_i, W_i$.

Assumption 1.4 requires that the exogenous instrument Z_{it} is independent of U_{it} given A_i and W_i. Since one may view W_i as summarizing all the time-invariant information about A_i in the data, the assumption is, loosely speaking, requiring Z_{it} to be independent of U_{it} given A_i by the rules of conditional independence, which is already implied by the unconditional exogeneity assumption of $Z_{it} \perp \left(A_i, U_{it}\right)$ in Assumption 1.3. When Assumption 1.4 is satisfied depends on which W_i is used in practice. For example, if X and Z are both scalars and one uses $W_i = T^{-1} \sum_t (X_{it}, Z_{it})$, then Assumption 1.4 is satisfied when $g(Z_{it}, A_i, U_{it})$ is separable in Z_{it}. Assumption 1.4 is used to ensure that the residual variation in X_{it} given V_{it} and W_i is exogenous to (A_i, U_{it}).

Lemma 1.2 (Feasible Control Variable $\boldsymbol{V_{it}}$). *Suppose Assumptions 1.1–1.4 hold. Then, the random variable V_{it} satisfies*

$$V_{it} := F_{X_{it}^{(1)} \middle| Z_{it}, W_i} \left(X_{it}^{(1)} \middle| Z_{it}, W_i \right) = F_{U_{it}|A_i, W_i} \left(U_{it} \middle| A_i, W_i \right), \tag{1.20}$$

where $X_{it}^{(1)}$ denotes the first coordinate of X_{it} that is known to satisfy Assumption 1.1.

The important part of Lemma 1.2 is that V_{it} is feasible. As a result, it can be consistently estimated from data. The feasibility of V_{it} solves the first issue discussed at the beginning of this identification step. Note that one coordinate of X_{it} that satisfies Assumption 1.1 is sufficient to construct V_{it}. When there are multiple coordinates of X_{it} that are known to satisfy Assumption 1.1, one can choose whichever coordinate of X_{it} to construct V_{it} because by (1.20), a single variable V_{it} suffices to control for U_{it} given (A_i, W_i). We provide an extension when U_{it} is a vector towards the end of the identification section.

However, the conditional CDF $F_{U_{it}|A_i, W_i} \left(U_{it} \middle| A_i, W_i \right)$ on the rhs of (1.20) is not a one-to-one function of U_{it} because A_i is unknown. If A_i is known, then one can condition on (A_i, V_{it}, W_i), which by (1.20) is equivalent to conditioning on (A_i, U_{it}, W_i), and use the residual variation in X_{it} to identify moments of β_{it} as in (1.13). In the next step, we deal with unknown A_i using the sufficiency argument from the first step and the law of iterated expectations (LIE).

Step 3: Identify the First-Order Moments of $\boldsymbol{\beta_{it}}$

We impose the next two regularity assumptions on $F_{U_{it}|A_i, W_i} \left(U_{it} \middle| A_i, W_i \right)$ and the support of X_{it} given (V_{it}, W_i), respectively.

Assumption 1.5 (**Strict Monotonicity of CDF of $\boldsymbol{U_{it}}$**)**.** *The conditional CDF $F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right)$ is strictly increasing in U_{it} for all (A_i, W_i).*

Assumption 1.6 (**Residual Variation in $\boldsymbol{X_{it}}$**)**.** *The support of X_{it} given V_{it} and W_i contains some ball of positive radius a.s. wrt (V_{it}, W_i).*

Assumption 1.5 requires that the conditional CDF of U_{it} given (A_i, W_i) cannot have flat areas, i.e., for each possible realization $c \in [0,1]$ of $F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right)$ and fixed (A_i, W_i), there is one and only one value of U_{it} such that $F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right) = c$. Consequently, fixing the level of $F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right)$ as well as (A_i, W_i) is equivalent to fixing the level of U_{it}. Assumption 1.6 is like the rank condition that is familiar from the linear simultaneous equations model. It requires that conditional on V_{it} and W_i, there is residual variation in X_{it} to identify moments of β_{it}. Assumption 1.6 is imposed to facilitate a partial derivative based identification argument and thus rules out discrete X_{it}'s. One can include discrete X_{it}'s by using the within group variation among X_{it}'s given V_{it} and W_i. Then, the required support condition is there are at least d_X linearly independent points in the support of X_{it} given V_{it} and W_i.

It is worth mentioning that we do not require the conditional support of the control variable V_{it} given X_{it} is equal to the unconditional support of V_{it}, i.e., Assumption 2 of Imbens and Newey (2009), because we take advantage of the linear structure of the model and separately identify the unconditional mean of β_{it} *without* integrating over the marginal distribution of V_{it}, which identifies the average structural function.

Suppose A_i is known for now, we have

$$
\begin{aligned}
& \mathbb{E}\left[\beta_{it}|\, X_{it}, A_i, V_{it}, W_i\right] \\
&= \mathbb{E}\left[\beta\left(A_i, U_{it}\right)|\, g\left(Z_{it}, A_i, U_{it}\right), A_i, F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right), W_i\right] \\
&= \mathbb{E}\left[\beta\left(A_i, U_{it}\right)|\, A_i, V_{it}, W_i\right] =: \widetilde{\beta}\left(A_i, V_{it}, W_i\right), \qquad (1.21)
\end{aligned}
$$

where the first equality holds by the definition of V_{it} and (1.3), and the second equality is true because the sigma algebra generated by $\left(A_i, F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right), W_i\right)$ is equal to that generated by (A_i, U_{it}, W_i) by Assumption 1.5, which contains all the information necessary to calculate the first-order moment of β_{it} as a function of A_i and U_{it}. As a consequence, the variation in X_{it} does not contain any additional information given (A_i, V_{it}, W_i).

Next, to deal with unknown A_i appearing in (1.21), we use the LIE together with the sufficiency condition of (1.14). More specifically, taking the conditional expectation of $\widetilde{\beta}\left(A_i, V_{it}, W_i\right)$ wrt A_i given (X_{it}, V_{it}, W_i) gives

$$\begin{aligned} & \mathbb{E}\left[\left.\widetilde{\beta}\left(A_i, V_{it}, W_i\right)\right| X_{it}, V_{it}, W_i\right] \\ = {} & \mathbb{E}\left[\left.\mathbb{E}\left[\left.\widetilde{\beta}\left(A_i, V_{it}, W_i\right)\right| X_{it}, Z_{it}, W_i\right]\right| X_{it}, V_{it}, W_i\right] \\ = {} & \mathbb{E}\left[\left.\int \widetilde{\beta}\left(a, V_{it}, W_i\right) f_{A_i|W_i}\left(a|\, W_i\right)\mu\left(da\right)\right| X_{it}, V_{it}, W_i\right] =: \beta\left(V_{it}, W_i\right), \end{aligned} \tag{1.22}$$

where the first equality holds by the LIE and the fact that V_{it} is a function of (X_{it}, Z_{it}, W_i) and the second equality holds by (1.14). The measure $\mu\left(\cdot\right)$ in the third line of (1.22) represents the Lebesgue measure.

Given (1.22), taking the conditional expectation of both sides of (1.1) given (X_{it}, V_{it}, W_i) leads to

$$\mathbb{E}\left[Y_{it}|\, X_{it}, V_{it}, W_i\right] = X_{it}'\beta\left(V_{it}, W_i\right). \tag{1.23}$$

From (1.23), the “cv” appearing in (1.13) are (V_{it}, W_i). The result is intuitive because V_{it} is a feasible control variable for U_{it} given (A_i, W_i) and W_i is a sufficient statistic for A_i. Therefore, fixing (V_{it}, W_i) effectively controls for (A_i, U_{it}), thus the residual variation in X_{it} is exogenous.

When Assumption 1.6 holds, one can identify $\beta\left(V_{it}, W_i\right)$ by

$$\beta\left(V_{it}, W_i\right) = \partial\mathbb{E}\left[Y_{it}|\, X_{it}, V_{it}, W_i\right]/\partial X_{it}. \tag{1.24}$$

With $\beta\left(V_{it}, W_i\right)$ identified, one can then identify $\mathbb{E}\left[\beta_{it} \middle| X_{it}\right]$ and $\mathbb{E}\beta_{it}$ via the LIE. For example,

$$\mathbb{E}\beta_{it} = \mathbb{E}\beta\left(V_{it}, W_i\right) = \mathbb{E}\left(\partial \mathbb{E}\left[Y_{it} \middle| X_{it}, V_{it}, W_i\right] / \partial X_{it}\right), \tag{1.25}$$

where the expectation is taken wrt the joint distribution of (V_{it}, W_i), an identifiable object from data.

Theorem 1.1 (**Identification**). *If Assumptions 1.1–1.6 are satisfied, then* $\mathbb{E}\left[\beta_{it} \middle| V_{it}, W_i\right]$, $\mathbb{E}\left[\beta_{it} \middle| X_{it}\right]$, *and* $\mathbb{E}\beta_{it}$ *are identified.*

Theorem 1.1 presents the main identification result following the steps above. The idea is simple: find the feasible variables denoted by "cv" in (1.13) such that conditioning on these variables, the residual variation in X_{it} is exogenous to that in β_{it}. We have shown that the feasible variables are (V_{it}, W_i). The sufficient statistic W_i for A_i constructed in the first step plays an important role. It not only enables the construction of the feasible control variable V_{it} for U_{it} given (A_i, W_i) in the second step, but also manages to control for A_i in the last step. By exploiting the panel data structure, the proposed method extends the traditional control function approach where only one unknown scalar affects the regressors to the setting with a fixed effect of arbitrary dimension and a random shock, both of which affect the choice of X_{it} in a nonseparable way as in (1.3).

Higher-Order Moments of $\boldsymbol{\beta_{it}}$

We have shown the identification of the first-order expectation of the vector of the random coefficients. Higher-order moments such as variance of the random coefficients can also be of interest to researchers to answer policy-related questions. For example, policy makers may be interested in how fast labor-augmenting technology is being diffused among firms. In this section, we briefly discuss how to identify the second-order moments under regularity conditions.

For simplicity of exposition, we consider the case when the vector of regressors $(X_{it}, 1)$ is two-dimensional. With a slight abuse of notation, let $(\beta_{it}, \omega_{it}) \in \mathbb{R}^2$ where β_{it} is the random coefficient corresponding to the scalar X_{it} and ω_{it} is the random coefficient associated with the constant 1. The ex-post shock ε_{it} is omitted from the analysis for brevity of exposition. If ε_{it} is present, one may follow the approach proposed in Arellano and Bonhomme (2012) and impose a structure such as ARMA on the inter-temporal dependence among ε_{it}'s to identify the second-order moments of β_{it} and ω_{it}.

Since (1.22) holds with β_{it}^2 or ω_{it}^2 in place of β_{it}, one has

$$\begin{aligned}
&\mathbb{E}\left[\beta_{it}^2\middle| X_{it}, V_{it}, W_i\right] = \mathbb{E}\left[\beta_{it}^2\middle| V_{it}, W_i\right], \\
&\mathbb{E}\left[\omega_{it}^2\middle| X_{it}, V_{it}, W_i\right] = \mathbb{E}\left[\omega_{it}^2\middle| V_{it}, W_i\right], \\
&\mathbb{E}\left[\omega_{it}\beta_{it}\middle| X_{it}, V_{it}, W_i\right] = \mathbb{E}\left[\omega_{it}\beta_{it}\middle| V_{it}, W_i\right]. \qquad (1.26)
\end{aligned}$$

Thus, taking the conditional expectation of the squares of both sides of (1.1) given (X_{it}, V_{it}, W_i) gives

$$\mathbb{E}\left[Y_{it}^2\middle| X_{it}, V_{it}, W_i\right] = X_{it}^2\mathbb{E}\left[\beta_{it}^2\middle| V_{it}, W_i\right] + 2X_{it}\mathbb{E}\left[\beta_{it}\omega_{it}\middle| V_{it}, W_i\right] + \mathbb{E}\left[\omega_{it}^2\middle| V_{it}, W_i\right]. \qquad (1.27)$$

Then, the identification of $\mathbb{E}\left[\beta_{it}^2\middle| V_{it}, W_i\right]$, $\mathbb{E}\left[\omega_{it}^2\middle| V_{it}, W_i\right]$, and $\mathbb{E}\left[\omega_{it}\beta_{it}\middle| V_{it}, W_i\right]$ follows similarly to (1.24). More precisely, one can identify $\mathbb{E}\left[\beta_{it}^2\middle| V_{it}, W_i\right]$ by exploiting the second-order derivative of $\mathbb{E}\left[Y_{it}^2\middle| X_{it}, V_{it}, W_i\right]$ wrt X_{it}:

$$\mathbb{E}\left[\beta_{it}^2\middle| V_{it}, W_i\right] = \left(\partial^2\mathbb{E}\left[Y_{it}^2\middle| X_{it}, V_{it}, W_i\right]/\partial X_{it}^2\right)/2. \qquad (1.28)$$

Then, one can identify $\mathbb{E}\left[\beta_{it}\omega_{it}\middle| V_{it}, W_i\right]$ by

$$\mathbb{E}\left[\beta_{it}\omega_{it}\middle| V_{it}, W_i\right] = \left(\partial\mathbb{E}\left[Y_{it}^2\middle| X_{it}, V_{it}, W_i\right]/\partial X_{it} - 2X_{it}\mathbb{E}\left[\beta_{it}^2\middle| V_{it}, W_i\right]\right)/2 \qquad (1.29)$$

and finally identify $\mathbb{E}\left[\omega_{it}^{2}\middle| V_{it},W_i\right]$ by

$$\begin{aligned}&\mathbb{E}\left[\omega_{it}^{2}\middle| V_{it},W_i\right]\\&=\mathbb{E}\left[Y_{it}^{2}\middle| X_{it},V_{it},W_i\right]-X_{it}^{2}\mathbb{E}\left[\beta_{it}^{2}\middle| V_{it},W_i\right]-2X_{it}\mathbb{E}\left[\beta_{it}\omega_{it}\middle| V_{it},W_i\right].\end{aligned}\tag{1.30}$$

By induction, the analysis can be extended to identify any order of moments of β_{it}, which under regularity conditions (Stoyanov, 2000) uniquely determines the distribution function of β_{it}.

The flexible identification argument can also be used to identify intertemporal correlations of the random coefficients. For example, one can identify $\mathbb{E}\left[\beta_{it}\beta_{is}\middle| X_{it},X_{is},V_{it},V_{is},W_i\right]$ from $\mathbb{E}\left[Y_{it}Y_{is}\middle| X_{it},X_{is},V_{it},V_{is},W_i\right]$ for any $t,s\in\{1,..,T\}$ following an almost identical argument as in (1.26)–(1.30).

Other Extensions

The identification argument is flexible and can adapt to several extensions. First, when there is a vector of U_{it} (say two dimensional) in (1.1) while each coordinate of U_{it} appears in only one of (1.3), i.e.,

$$\begin{aligned}Y_{it}&=X_{it}^{'}\beta\left(A_i,U_{it}^{(1)},U_{it}^{(2)}\right)+\varepsilon_{it}\\X_{it}^{(1)}&=g^{(1)}\left(Z_{it},A_i,U_{it}^{(1)}\right)\\X_{it}^{(2)}&=g^{(2)}\left(Z_{it},A_i,U_{it}^{(2)}\right),\end{aligned}$$

one can construct

$$V_{it}^{(1)}:=F_{X_{it}^{(1)}\middle|Z_{it},W_i}\left(X_{it}^{(1)}\middle| Z_{it},W_i\right)\text{ and }V_{it}^{(2)}:=F_{X_{it}^{(2)}\middle|Z_{it},W_i}\left(X_{it}^{(2)}\middle| Z_{it},W_i\right),$$

and follow Step 1–3 to obtain

$$\mathbb{E}\left[Y_{it}\middle|\, X_{it}, V_{it}^{(1)}, V_{it}^{(2)}, W_i\right] = X_{it}'\beta\left(V_{it}^{(1)}, V_{it}^{(2)}, W_i\right).$$

Then, the identification follows identically to (1.24).

Second, to allow more flexible or even arbitrary inter-temporal correlation than (1.7) among the U_{it}'s, one may replace the individual fixed effect A_i with a group fixed effect A_j when i belongs to group j (Cameron, Gelbach, and Miller, 2012; Cameron and Miller, 2015). More precisely, we modify the model (1.1)–(1.3) to be

$$\begin{aligned} Y_{ijt} &= X_{ijt}'\beta\left(A_j, U_{ijt}\right) + \varepsilon_{ijt}, \\ X_{ijt} &= g\left(Z_{ijt}, A_j, U_{ijt}\right), \end{aligned} \tag{1.31}$$

where i is individual, j is group, and t is time. One may want to use this model instead of (1.1)–(1.3) if she desires to relax the restriction on the inter-temporal correlations between U_{it}'s and finds the evidence of a group fixed effect, e.g., location or sector or age fixed effect. Let $U_{ij} = \left(U_{ij1}, ..., U_{ijT}\right)'$. Then, one can use a "group" version of the exchangeability condition

$$f_{U_{1j},...,U_{Ij}|A_j}\left(u_{1j}, ..., u_{Ij}\middle|\, a_j\right) = f_{U_{1j},...,U_{Ij}|A_j}\left(u_{i_1 j}, ..., u_{i_I j}\middle|\, a_j\right), \tag{1.32}$$

where $(i_1, ..., i_I)$ is any permutation of $(1, ..., I)$, to construct a sufficient statistic W_j for A_j and proceed as in Step 2–3 to identify moments of the random coefficients.

Third, to deal with persistent shocks to X_{it} or deterministic time trend in X_{it}, one may model the inter-temporal change in X_{it}, or $\Delta X_{it} \coloneqq X_{it} - X_{it-1}$, as a function g of (Z, A, U) instead of modeling X_{it} as a function g of (Z, A, U). The identification is mostly the same as before, except that W_i is now a symmetric function in the elements of ΔX_{it} rather than X_{it}

and $V_{it} := F_{\Delta X_{it}|Z_{it},W_i}$. Then, one can identify the moment of β_{it} by taking partial derivative wrt ΔX_{it} on both sides of

$$\mathbb{E}\left[Y_{it}\right| X_{it-1}, \Delta X_{it}, V_{it}, W_i] = \left(X_{it-1} + \Delta X_{it}\right)' \beta\left(X_{it-1}, V_{it}, W_i\right). \tag{1.33}$$

The last extension concerns exogenous shocks. We maintain model (1.1) and (1.3) and follow Graham and Powell (2012) to replace (1.2) with $\beta_{it} = \beta\left(A_i, U_{it}\right) + d_t\left(U_{2,it}\right)$, where d_t is an unknown time-varying vector-valued function and $U_{2,it}$ is an exogenous shock independent of all other variables in the system. For example, $U_{2,it}$ can capture the effect of the pandemic on the mental/physical health of the employees of firm i in period t after the employees have been hired. Then, following the argument as before, we have

$$\mathbb{E}\left[\beta_{it}\right| X_{it}, V_{it}, W_i] = \mathbb{E}\left[\beta\left(A_i, U_{it}\right)\right| V_{it}, W_i] + \mathbb{E}\left[d_t\left(U_{2,it}\right)\right] =: \beta\left(V_{it}, W_i\right) + \delta_{0t}, \tag{1.34}$$

which implies

$$\mathbb{E}\left[Y_{it}\right| X_{it}, V_{it}, W_i] = X_{it}'\left[\beta\left(V_{it}, W_i\right) + \delta_{0t}\right]. \tag{1.35}$$

Taking the partial derivative wrt X_{it} on both sides of (1.35) gives

$$\partial\mathbb{E}\left[Y_{it}\right| X_{it}, V_{it}, W_i] / \partial X_{it} = \beta\left(V_{it}, W_i\right) + \delta_{0t}. \tag{1.36}$$

Repeating the same process for a different period $s \neq t$ leads to

$$\partial\mathbb{E}\left[Y_{is}\right| X_{is}, V_{is}, W_i] / \partial X_{is} = \beta\left(V_{is}, W_i\right) + \delta_{0s}. \tag{1.37}$$

Then, one identifies $\delta_{0t} - \delta_{0s}$ for any $t \neq s$ by

$$\delta_{0t} - \delta_{0s} = \left\{\partial\mathbb{E}\left[Y_{it}\right| X_{it}, V_{it}, W_i] / \partial X_{it} - \partial\mathbb{E}\left[Y_{is}\right| X_{is}, V_{is}, W_i] / \partial X_{is}\right\}\big|_{V_{it}=V_{is}}. \tag{1.38}$$

Using the same normalization of $\delta_{01} = 0$ as in Graham and Powell (2012), one identifies δ_{0t} for all t. Finally, the identification of $\beta(V_{it}, W_i)$ follows from (1.36).

1.4 Estimation and Large Sample Theory

The identification argument is constructive and leads to a feasible estimator for the first-order moment of β_{it}. In this section, we first estimate the conditional and unconditional moments of the random coefficients using multi-step series estimators. Then, we obtain the convergence rates and asymptotic normality results for the proposed estimators.

1.4.1 Estimation

The parameters of interest in this paper are

$$\beta(v,w) \coloneqq \mathbb{E}\left[\beta_{it} \middle| V_{it} = v, W_i = w\right], \ \beta(x) \coloneqq \mathbb{E}\left[\beta_{it} \middle| X_{it} = x\right], \text{ and } \overline{\beta} \coloneqq \mathbb{E}\beta_{it}. \tag{1.39}$$

We propose to estimate them using three-step series estimators. In the first step, we estimate $V(x,z,w) = F_{X_{it}|Z_{it},W_i}(x|\, z,w)$ and denote $V_{it} \coloneqq V(X_{it}, Z_{it}, W_i)$. Then, for $s = (x,v,w)$ we estimate $G(s) \coloneqq \mathbb{E}\left[Y_{it} \middle| X_{it} = x, V_{it} = v, W_i = w\right]$ using $\widehat{V}$ obtained in the first step and denote $G_{it} \coloneqq G(S_{it}) = G(X_{it}, V_{it}, W_i)$. Finally, we estimate $\beta(v,w)$, $\beta(X_{it})$ and $\overline{\beta}$, all of which are identifiable functionals of $G(s)$. For brevity of exposition, we provide definitions of all of the symbols appearing in this section in Appendix 1.C.

More specifically, we first estimate $V(x,z,w)$ by regressing $\mathbb{1}\{X_{it} \leq x\}$ on the basis functions $q^L(\cdot)$ of (Z_{it}, W_i) with trimming function $\tau(\cdot)$:

$$\begin{aligned}\widehat{V}(x,z,w) &= \tau\left(\widehat{F}_{X_{it}|Z_{it},W_i}(x|\, z,w)\right)\\ &= \tau\left(q^L(z,w)'\, \widehat{Q}^{-1}\sum_{j=1}^{n} q_j \mathbb{1}\{X_{jt} \leq x\}/n\right)\\ &=: \tau\left(q^L(z,w)'\, \widehat{\gamma}^L(x)\right). \end{aligned} \tag{1.40}$$

We highlight two properties of $\widehat{V}(x,z,w)$. First, unlike traditional series estimators, the regression coefficient $\widehat{\gamma}^L(x)$ in (1.40) depends on x because the dependent variable in V is a function of x. This fact makes the convergence rate of $\widehat{V}$ slower than the standard rates for series estimators (Imbens and Newey, 2009). Second, a trimming function τ is applied to $q^L(z,w)' \widehat{\gamma}^L(x)$ because we estimate a conditional CDF which by definition lies between zero and one. One example of τ is $\tau(x) = \mathbb{1}\{x \geq 0\} \times \min(x,1)$.

Next, we estimate $G(s)$ by regressing Y_{it} on the basis functions $p^K(\cdot)$ of $\left(X_{it}, \widehat{V}_{it}, W_i\right)$:

$$\widehat{G}(s) = p^K(s)' \widehat{P}^{-1}\widehat{p}' y/n =: p^K(s)' \widehat{\alpha}^K. \tag{1.41}$$

Following Newey, Powell, and Vella (1999), we construct the basis function $p^K(s) = x \otimes p^{K_1}(v,w)$ by exploiting the index structure of the model (1.1). The index structure enables a faster convergence rate for $\widehat{G}(s)$. Note that in (1.41) $\widehat{V}_{it}$ from the first-step is plugged in wherever V_{it} appears.

Finally, we estimate $\beta(v,w)$ by exploiting the index structure of the model (1.1) and calculate it as

$$\widehat{\beta}(v,w) = \partial \widehat{G}(s)/\partial x = \left(I_{d_X} \otimes p^{K_1}(v,w)\right)' \widehat{\alpha}^K =: \overline{p}(s)' \widehat{\alpha}^K, \tag{1.42}$$

where the second equality holds by the chain rule. To estimate $\beta(x)$ and $\overline{\beta}$, we use the LIE and regress $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ on the basis function $r^M(\cdot)$ of X_{it} and the constant 1, respectively :

$$\begin{aligned} \widehat{\beta}(x) &= r^M(x)' \widehat{R}^{-1} r' \widehat{B}/n =: r^M(x)' \widehat{\eta}^M, \\ \widehat{\overline{\beta}} &= n^{-1} \sum_{i=1}^{n} \widehat{\beta}\left(\widehat{V}_{it}, W_i\right). \end{aligned} \tag{1.43}$$

One may consider $\widehat{\overline{\beta}}$ as a “special case” of $\widehat{\beta}(x)$ by letting $r^M(\cdot) \equiv 1$, which simplifies the asymptotic analysis in the next section.

The objects of interest in this paper are $\beta(v,w)$, $\beta(x)$, and $\overline{\beta}$. $\beta(v,w)$ is the conditional expectation of β_{it} given $(V_{it}, W_i) = (v,w)$, and can be interpreted as the average of the partial effects of X_{it} on Y_{it} among the individuals with the same $(V_{it}, W_i) = (v,w)$. If one loosely considers V_{it} to be U_{it} and W_i to be A_i, then $\beta(v,w)$ is the same as β_{it}. In this sense, $\beta(v,w)$ provides the "finest" approximation of β_{it} among the three objects in (1.39). $\beta(x)$ measures the average partial effect averaged over the conditional distribution of the unobserved heterogeneity (A_i, U_{it}) when X_{it} equals x. It provides useful information about the partial effects of X_{it} on Y_{it} for a subpopulation characterized by $X_{it} = x$. For example, if one asks about the average output elasticity with respect to labor for firms with a certain level of capital and labor, then $\beta(x)$ contains relevant information to answer such questions. $\overline{\beta}$ is the APE that has been studied extensively in the literature (Chamberlain, 1984, 1992; Wooldridge, 2005b; Graham and Powell, 2012; Laage, 2020). It is interpreted as the average of the partial effect of X_{it} on Y_{it} over the unconditional distribution of (A_i, U_{it}). Depending on the scenario and application, all three objects can be useful to answer policy-related questions.

The multi-step series estimators proposed in this section cause challenges for inference due to their multi-layered nature. To obtain large sample properties of $\widehat{\beta}(v,w)$, $\widehat{\beta}(x)$ and $\widehat{\overline{\beta}}$, one needs to analyze the estimators step by step as the estimator from each step is plugged in and thus affects all subsequent ones. For asymptotic analysis, there is a key difference between $\beta(v,w)$ and $\beta(x)$ or $\overline{\beta}$: $\beta(v,w)$ is a *known* functional of $G(s)$, whereas both $\beta(x)$ and $\overline{\beta}$ are *unknown* but identifiable functionals of $G(s)$. We present in the next session how to deal with these challenges for the purpose of inference.

1.4.2 Large Sample Theory

Before proving convergence rates and asymptotic normality results for the three-step series estimators defined in (1.42)–(1.43), we first briefly review the related literature. Andrews (1991) analyzes the asymptotic properties of series estimators for nonparametric and semiparametric regression models. His results are applicable to a wide variety of estimands, including derivatives and integrals of the regression function. This paper builds on his results and shows asymptotic normality for a vector-valued functional of regression functions. Newey (1997) also studies series estimators and give conditions for obtaining convergence rates and asymptotic normality for the estimators of conditional expectations. Newey, Powell, and Vella (1999) present a two-step nonparametric estimator for a triangular simultaneous equation model with a separable first-step equation. They derive asymptotic normality for their two-step estimator with the first-step plugged in. Imbens and Newey (2009) also analyze a triangular simultaneous equation model, but with a nonseparable first-step equation. They show mean-squared convergence rates for the first-step estimator, and prove asymptotic normality for known functionals of the conditional expectation of the outcome variable given regressors and control variables. We build on and extend their asymptotic results to unknown but estimable functionals of the conditional expectations.

More recently, Hahn and Ridder (2013) derive a general formula of the asymptotic variance of the multi-step estimators using the pathwise derivative method by Newey (1994). They only consider the case that the first-stage model is a regression model with a separable error. Hahn and Ridder (2019) consider a setting with a nonseparable first step similar to the one in this paper. They focus on the full mean process instead of the partial mean process and show how to obtain influence functions for known functionals of the average structural functions rather than unknown functionals of the conditional expectation functions. Thus, their results do not directly apply to our setting. Mammen, Rothe, and Schienle (2012, 2016) study the statistical properties of nonparametric regression estimators using generated covariates. They focus on kernel estimators in these two papers. Lee (2018) considers partial mean process with generated regressors, where the average is over the generated regressors while fixing the treatment variable at a certain level. She proposes a nonparametric estimator where the second step consists of a kernel regression on regressors that are estimated in the first step. Her assumptions and method are quite different from those considered in this paper.

Alternatively to these papers, one may use sieve methods to establish large sample properties for the multi-step estimators considered in this paper. Ai and Chen (2007) consider the estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables, which include many control variable models similar to the one discussed in this paper. See Ackerberg, Chen, and Hahn (2012) for more details on how to apply the methods proposed in Ai and Chen (2007). Chen and Liao (2014) derive point-wise normality for slower than root-n functionals for general sieve M estimation. Chen and Liao (2015) consider semiparametric multi-step estimation and inference with weakly dependent data, where unknown nuisance functions are estimated via sieve extremum estimation in the first step. They show that the asymptotic variance of the multi-step estimator can be well approximated by sieve variances that have simple closed-form expressions. We refer interested readers to these papers for more details.

We now derive convergence rates and asymptotic normality results for the proposed estimators. Since we let $n \to \infty$ for each t in the asymptotic analysis, the t-subscript is suppressed for notational simplicity. First, we obtain convergence rates for $\widehat{\beta}\left(v,w\right)$, $\widehat{\beta}\left(x\right)$, and $\widehat{\widehat{\beta}}$, respectively. For $\widehat{\beta}\left(v,w\right)$, we adapt the results of Imbens and Newey (2009) to the TERC model considered in this paper. For $\widehat{\beta}\left(x\right)$ and $\widehat{\widehat{\beta}}$, the effects from first- and second-step estimations need to be taken into consideration. We present both mean squared and uniform rates for all three estimators.

Then, we prove asymptotic normality for the estimators, and show that the corresponding variances can be consistently estimated to construct valid confidence intervals. Asymptotic normality for $\widehat{\beta}\left(v,w\right)$ is established by applying the results of Andrews (1991) and Imbens and Newey (2002) to cover vector-valued functionals. For $\widehat{\beta}\left(x\right)$ and $\widehat{\widehat{\beta}}$, the main difference from the existing literature is that both estimators are *unknown* functionals of $G\left(\cdot\right)$ that are only estimable from the data. Therefore, one needs to correctly account for the additional estimation error and adjust the asymptotic variance.

Convergence Rates

Recall that the conditional and unconditional moments of the random coefficients are estimated via the three-step estimators (1.42)–(1.43). The convergence rates for the first- and second- step estimators $\widehat{V}$ and $\widehat{G}$ have been obtained in Imbens and Newey (2009). We adapt their results to our TERC model and impose the following regularity assumption.

Assumption 1.7. *Suppose the following conditions hold:*

1. *There exist d_1, $C > 0$ such that for every L there is a $L \times 1$ vector $\gamma^L(x)$ satisfying*

$$\sup_{x\in\mathcal{X},z\in\mathcal{Z},w\in\mathcal{W}} \left|F_{X|Z,W}\left(x|\, z,w\right) - q^L\left(z,w\right)' \gamma^L\left(x\right)\right| \leq CL^{-d_1}.$$

2. *The joint density of (X, V, W) is bounded above and below by constant multiples of its marginal densities.*

3. *There exist $C > 0$, $\zeta(K_1)$, and $\zeta_1(K_1)$ such that $\zeta(K_1) \leq C\zeta_1(K_1)$ and for each K_1 there exists a normalization matrix B such that $\widetilde{p}^{K_1}(v,w) = Bp^{K_1}(v,w)$ satisfies $\lambda_{\min}\left(\mathbb{E}\widetilde{p}^{K_1}(V_i,W_i)\,\widetilde{p}^{K_1}(V_i,W_i)'\right) \geq C$, $\sup_{v\in\mathcal{V},w\in\mathcal{W}}\left\|\widetilde{p}^{K_1}(v,w)\right\| \leq C\zeta(K_1)$, and $\sup_{v\in\mathcal{V},w\in\mathcal{W}}\left\|\partial\widetilde{p}^{K_1}(v,w)/\partial v\right\| \leq C\zeta_1(K_1)$. Furthermore, $K_1\zeta_1(K_1)^2\left(L/n + L^{1-2d_1}\right)$ is $o(1)$.*

4. *$G(s)$ is Lipschitz in v. There exist d_2, $C > 0$ such that for every $K = d_X \times K_1$ there is a $K \times 1$ vector α^K satisfying*

$$\sup_{s\in\mathcal{S}}\left|G\left(s\right) - p^K\left(s\right)' \alpha^K\right| \leq CK^{-d_2}.$$

5. *$Var\left(Y_i|\, X_i, Z_i, W_i\right)$ is bounded uniformly over the support of (X_i, Z_i, W_i).*

Assumption 1.7(1) and (4) specify the approximation rates for the series estimators. It is well-known that such rates exist when $F_{X|Z,W}(x|z,w)$ and $G(s)$ satisfy mild smoothness conditions and regular basis functions like splines are used. See Imbens and Newey (2009) for a detailed discussion.

Assumption 1.7(2) is imposed to guarantee that the smallest eigenvalue of $\mathbb{E}p^K(S_i)\,p^K(S_i)'$ is strictly larger than some positive constant C. It is imposed because in the analysis we exploit the index structure of our TERC model by choosing $p^K(s) = x \otimes p^{K_1}(v,w)$. The usual normalization (Newey, 1997) on the second moment of basis functions can only be done on x and $p^{K_1}(v,w)$ separately. Thus, we need Assumption 1.7(2) to make sure the second moment of $p^K(s)$ is well-behaved. A similar assumption is imposed in Imbens and Newey (2002) as well.

Assumption 1.7(3) is a normalization on the basis function $p^{K_1}(\cdot)$, which ensures that one can normalize $\mathbb{E}p^{K_1}(V_i,W_i)\,p^{K_1}(V_i,W_i)'$ to be the identity matrix I as in Newey (1997). Finally, the conditional variance of Y given (X,V,W) is assumed to be bounded in Assumption 1.7(5), which is common in the series estimation literature.

With Assumption 1.7 in position, we prove the following lemma.

Lemma 1.3 (First- and Second-Step Convergence Rates). *Suppose the conditions of Theorem 1.1 and Assumption 1.7 are satisfied. Then, we have*

$$n^{-1}\sum_i \left(\widehat{V}_i - V_i\right)^2 = O_P\left(L/n + L^{1-2d_1}\right) =: O_P\left(\Delta_{1n}^2\right)$$
$$\int \left[\widehat{G}(s) - G(s)\right]^2 dF(s) = O_P\left(K_1/n + K_1^{-2d_2} + \Delta_{1n}^2\right) =: O_P\left(\Delta_{2n}^2\right)$$
$$\sup_{s\in\mathcal{S}}\left|\widehat{G}(s) - G(s)\right| = O_P\left(\zeta(K_1)\,\Delta_{2n}\right).$$

Lemma 1.3 states that the mean squared convergence rate for $\widehat{G}$ is the sum of the first-step rate Δ_{1n}^2, the variance term K_1/n, and the squared bias term $K_1^{-2d_2}$. Both d_1 and d_2 are the uniform approximation rates that govern how well one is able to approximate the unknown functions V and G with $q^L(\cdot)$ and $p^K(\cdot)$, respectively. Note that even though the order of the basis function for the second-step estimation is K, by the TERC structure $K = d_X \times K_1$ and d_X is a finite constant. Thus, the effective order that matters for the convergence rate results is K_1.

We now obtain the convergence rates for $\widehat{\beta}(v, w)$, $\widehat{\beta}(x)$ and $\widehat{\widehat{\beta}}$. We impose the following assumption.

Assumption 1.8. *Suppose the following conditions hold:*

1. *There exist $d_3, C > 0$ such that for every M there is a $M \times d_X$ matrix η^M satisfying*

$$\sup_{x\in\mathcal{X}} \left\| \beta(x) - r^M(x)' \eta^M \right\| \leq CM^{-d_3}.$$

2. *There exist $C > 0$ and $\zeta(M)$ such that for each M there exists a normalization matrix B such that $\widetilde{r}^M(x) = Br^M(x)$ satisfies $\lambda_{\min}\left(\mathbb{E}\widetilde{r}^M(X_i)\,\widetilde{r}^M(X_i)'\right) \geq C$ and $\sup_{x\in\mathcal{X}} \left\|\widetilde{r}^M(x)\right\| \leq C\zeta(M)$.*

3. *Let $\xi_i = \beta(V_i, W_i) - \beta(X_i)$ and $\xi = (\xi_1, ..., \xi_n)'$. Then, $\mathbb{E}\left[\xi\xi' \middle| \mathbf{X}\right] \leq CI$ in the positive definite sense.*

4. *$\beta(v, w)$ is Lipschitz in v, with the Lipschitz constant bounded from above.*

Assumption 1.8 imposes conditions on the approximation rate of $\beta(x)$, the normalization of basis functions $r^M(x)$, and the boundedness of the second moment of ξ_i, similarly to those in Assumption 1.7.

Theorem 1.2 (Third-Step Convergence Rates). *Suppose the conditions of Lemma 1.3 and Assumption 1.8 are satisfied. Then, we have*

$$\int \left\|\widehat{\beta}(v, w) - \beta(v, w)\right\|^2 dF(v, w) = O_P\left(\Delta_{2n}^2\right),$$
$$\int \left\|\widehat{\beta}(x) - \beta(x)\right\|^2 dF(x) = O_P\left(\Delta_{2n}^2 + M/n + M^{-2d_3}\right) =: O_p\left(\Delta_{3n}^2\right),$$

$$\left\|\widehat{\overline{\beta}}-\overline{\beta}\right\|^2=O_P\left(\Delta_{2n}^2\right),$$

$$\sup_{v\in\mathcal{V},w\in\mathcal{W}}\left\|\widehat{\beta}\left(v,w\right)-\beta\left(v,w\right)\right\|=O_P\left(\zeta\left(K_1\right)\Delta_{2n}\right),\ \ \textit{and}$$

$$\sup_{x\in\mathcal{X}}\left\|\widehat{\beta}\left(x\right)-\beta\left(x\right)\right\|=O_P\left(\zeta\left(M\right)\Delta_{3n}\right).$$

The first three equations in Theorem 1.2 give mean squared convergence rates, while the last two show uniform ones. For $\widehat{\beta}\left(v,w\right)$, the convergence rate is the same as $\widehat{G}$ because they share the same regression coefficient $\widehat{\alpha}^K$ and only differ in the basis functions used. More precisely, for $\widehat{\beta}\left(v,w\right)$ we use $I_{d_X}\otimes p^{K_1}\left(v,w\right)$, while for $\widehat{G}\left(s\right)$ we use $x\otimes p^{K_1}\left(v,w\right)$. Meanwhile, the same regression coefficient $\widehat{\alpha}^K$ is used for both estimators. Therefore, under Assumption 1.7 and 1.8, the convergence rate result on $\widehat{G}\left(s\right)$ applies directly to $\widehat{\beta}\left(v,w\right)$.

For $\widehat{\beta}\left(x\right)$ and $\widehat{\overline{\beta}}$, further analysis is required because both estimators involve an additional estimation step. Specifically, for $\widehat{\beta}\left(x\right)$, we estimate it with

$$\widehat{\beta}\left(x\right)=r^M\left(x\right)'\left(\widehat{R}^{-1}r'\widehat{B}/n\right)=:r^M\left(x\right)'\widehat{\eta}^M. \tag{1.44}$$

To obtain the convergence rate for $\widehat{\beta}\left(x\right)$, the key steps include expanding

$$\widehat{\eta}^M-\eta^M=\widehat{R}^{-1}r'\left[\left(\widehat{B}-\widetilde{B}\right)+\left(\widetilde{B}-B\right)+\left(B-B^X\right)+\left(B^X-r\eta^M\right)\right]/n, \tag{1.45}$$

where η^M is defined in Assumption 1.8(1), and deriving the rate for each component. We show the proof in the Appendix 1.B.

For $\widehat{\overline{\beta}}$, we estimate it with

$$\widehat{\overline{\beta}}=n^{-1}\sum_i\widehat{\beta}\left(\widehat{V}_i,W_i\right). \tag{1.46}$$

It is possible to analyze $\widehat{\overline{\beta}}$ in a similar way as $\widehat{\beta}\left(x\right)$ by expanding $\widehat{\beta}\left(\widehat{V}_i,W_i\right)-\overline{\beta}$ stochastically and deriving the convergence rate component by component. However, with the convergence results established for $\widehat{\beta}\left(x\right)$, one can let $r^M\left(\cdot\right)\equiv 1$ in (1.44) and directly obtain the rate for $\widehat{\overline{\beta}}$. We follow this simpler approach in the proof.

Asymptotic Normality

In this section, we prove asymptotic normality for the estimators of $\beta(v,w)$, $\beta(x)$ and $\overline{\beta}$, and show that the corresponding covariance matrices can be consistently estimated for use in confidence intervals. Imbens and Newey (2002) have obtained asymptotic normality for estimators of known and scalar-valued linear functionals of $G(s)$. However, $\beta(v,w)$ is a known but vector-valued functional of $G(s)$. To apply their results, we use Assumption J(iii) of Andrews (1991) together with a Cramér–Wold device to show asymptotic normality for $\widehat{\beta}(v,w)$.

Assumption 1.9. *Suppose the following conditions hold:*

1. *There exist $C>0$ and $\zeta(L)$ such that for each L there exists a normalization matrix B such that $\widetilde{q}^L(z,w) = Bq^L(z,w)$ satisfies $\lambda_{\min}\left(\mathbb{E}\widetilde{q}^L(Z_i,W_i)\,\widetilde{q}^L(Z_i,W_i)'\right) \geq C$ and $\sup_{z\in\mathcal{Z},w\in\mathcal{W}}\left\|\widetilde{q}^L(z,w)\right\| \leq C\zeta(L)$.*

2. *$G(s)$ is twice continuously differentiable with bounded first and second derivatives. For functional $a(\cdot)$ of G and some constant $C>0$, it is true that $|a(G)| \leq C\sup_s|G(s)|$ and either (i) there is $\delta(s)$ and $\widetilde{\alpha}^K$ such that $\mathbb{E}\delta(S_i)^2<\infty$, $a\left(p_k^K\right) = \mathbb{E}\delta(S_i)\,p_k^K(S_i)$ for all $k=1,...,K$, $a(G)=\mathbb{E}\delta(S_i)\,G(S_i)$, and $\mathbb{E}\left(\delta(S_i)-p^K(S_i)'\,\widetilde{\alpha}^K\right)^2 \to 0$; or (ii) for some $\widetilde{\alpha}^K$, $\mathbb{E}\left[p^K(S_i)'\,\widetilde{\alpha}^K\right]^2 \to 0$ and $a\left(p^K(\cdot)'\,\widetilde{\alpha}^K\right)$ is bounded away from zero as $K\to\infty$.*

3. *$\mathbb{E}\left[\left.(Y-G(s))^4\right| X,Z,W\right]<\infty$ and $Var\left(\left.Y\right|X,Z,W\right)>0$.*

4. *nL^{1-2d_1}, nK^{-2d_2}, $K\zeta_1(K)^2L^2/n$, $\zeta(K)^6L^4/n$, $\zeta_1(K)^2LK^{-2d_2}$, and $\zeta(K)^4\zeta(L)^4L/n$ are $o(1)$.*

5. *There exist d_4 and $\overline{\alpha}^K$ such that for each element s_j of $s=(x,v,w)'$:*

$$\max\left\{\sup_{s\in\mathcal{S}}\left|G(s)-p^K(s)'\,\overline{\alpha}^K\right|,\sup_{s\in\mathcal{S}}\left|\partial\left(G(s)-p^K(s)'\,\overline{\alpha}^K\right)/\partial s_j\right|\right\} = O\left(K^{-d_4}\right).$$

6. *(As' J(iii) of Andrews (1991)) For a bounded sequence of constants* $\{b_{1n} : n \geq 1\}$ *and constant pd matrix* $\overline{\Omega}_1$*, it is true that* $b_{1n}\Omega_1 \xrightarrow{p} \overline{\Omega}_1$.

Assumptions 1.9(1)–(5) are imposed in Imbens and Newey (2002) and are regularity conditions required for the asymptotic normality of $\widehat{\beta}\left(v, w\right)$. See Newey (1997) for a detailed discussion of these assumptions. Assumption 1.9(6) is used in Andrews (1991) and guarantees that the normality result of Imbens and Newey (2002) applies to vector-valued functionals of $G\left(s\right)$. Essentially, it requires all the coordinates of $\widehat{\beta}\left(v, w\right)$ to converge at the same speed, which is a mild assumption under our settings because ex-ante we do not distinguish one coordinate of β_{it} from the others.

Theorem 1.3 (**Asymptotic Normality for $\widehat{\boldsymbol{\beta}}\left(\boldsymbol{v}, \boldsymbol{w}\right)$**)**.** *Suppose the conditions of Theorem 1.2 and Assumption 1.9 are satisfied. Then, we have*

$$\sqrt{n}\widehat{\Omega}_1^{-1/2}\left(\widehat{\beta}\left(v, w\right) - \beta\left(v, w\right)\right) \xrightarrow{d} N\left(0, I\right).$$

It is worth noting that $\widehat{\Omega}_1$ in Theorem 1.3 is a function of (v, w), which is omitted for simplicity of exposition. Theorem 1.3 concerns $\beta\left(v, w\right)$, a *known* functional of $G\left(s\right)$. However, the result does not directly apply to $\beta\left(x\right)$ and $\overline{\beta}$, because they are *unknown* functionals of $G\left(s\right)$ and both require an additional estimation step. More specifically, by the LIE one has

$$\beta\left(x\right) = \mathbb{E}\left[\partial G\left(S_i\right)/\partial X \middle| X_i = x\right], \ \overline{\beta} = \mathbb{E}\left[\partial G\left(S_i\right)/\partial X\right], \tag{1.47}$$

both of which involve integrating over the unknown but estimable distribution of (V_i, W_i). Therefore, one need estimate these unknown functionals and correctly account for the bias arising from this additional estimation step in asymptotic analysis.

Assumption 1.10. *Suppose the following conditions hold:*

1. *There exists* $C > 0$ *such that for each* M *and* K *there exist normalization matrices* B_1 *and* B_2 *such that* $\widetilde{r}^M(x) = B_1 r^M(x)$ *and* $\widetilde{\overline{p}}^K(s) = B_2 \overline{p}^K(s)$ *satisfy* $\lambda_{\min}\left(\mathbb{E}\widetilde{r}^M(X_i)\,\widetilde{r}^M(X_i)'\right) \geq C$, $\lambda_{\min}\left(\mathbb{E}\widetilde{\overline{p}}^K(S_i)\,\widetilde{\overline{p}}^K(S_i)'\right) \geq C$, $\lambda_{\min}\left(\mathbb{E}\widetilde{r}^M(X_i)\,\widetilde{\overline{p}}^K(S_i)'\left(\mathbb{E}p^K(S_i)\,p^K(S_i)'\right)^{-1}\mathbb{E}\widetilde{\overline{p}}^K(S_i)\,\widetilde{r}^M(X_i)'\right) \geq C$, $\sup_{x\in\mathcal{X}}\left\|\widetilde{r}^M(x)\right\| \leq C\zeta(M)$, *and* $\sup_{s\in\mathcal{S}}\left\|\widetilde{\overline{p}}^K(s)\right\| \leq C\zeta(K)$.

2. *The fourth order moment of* $\xi_i := \beta(V_i, W_i) - \beta(X_i)$ *satisfies* $\mathbb{E}\left[\left.\xi_i^4\right| X_i\right] < \infty$.

3. *For a sequence of bounded constants* $\{b_{2n} : n \geq 1\}$ *and some constant pd matrix* $\overline{\Omega}_2$, $b_{2n}\Omega_2 \xrightarrow{p} \overline{\Omega}_2$ *holds.*

Assumption 1.10(1) is a normalization on basis functions $r^M(\cdot)$ and $\overline{p}^K(\cdot)$. The substantial part is

$$\lambda_{\min}\left(\mathbb{E}\widetilde{r}^M(X_i)\,\widetilde{\overline{p}}^K(S_i)'\left(\mathbb{E}p^K(S_i)\,p^K(S_i)'\right)^{-1}\mathbb{E}\widetilde{\overline{p}}^K(S_i)\,\widetilde{r}^M(X_i)'\right) \geq C, \tag{1.48}$$

which is needed to show that the asymptotic covariance matrix Ω_2 of $\sqrt{n}\left(\widehat{\beta}(x) - \beta(x)\right)$ is positive definite. Assumption 1.10(2) is a regularity condition imposed for the Lindeberg–Feller Central Limit Theorem (CLT). Assumption 1.10(3) is similar to Assumption 1.9(6) and is needed to show the asymptotic normality result holds for vector-valued functionals of $G(s)$.

Theorem 1.4 (**Asymptotic Normality for $\widehat{\boldsymbol{\beta}}(\boldsymbol{x})$ and $\widehat{\overline{\boldsymbol{\beta}}}$**)**.** *Suppose the conditions of Theorem 1.3 and Assumption 1.10 are satisfied. Then, we have*

$$\sqrt{n}\widehat{\Omega}_2^{-1/2}\left(\widehat{\beta}(x) - \beta(x)\right) \xrightarrow{d} N(0, I).$$

Furthermore, if $\mathbb{E}\left\|\beta(v,w) - \overline{\beta}\right\|^4 < \infty$ *, we have*

$$\sqrt{n}\widehat{\Omega}_3^{-1/2}\left(\widehat{\overline{\beta}} - \overline{\beta}\right) \xrightarrow{d} N(0, I).$$

Theorem 1.4 gives the asymptotic normality results that can be used to construct confidence intervals and test statistics for both $\beta\left(x\right)$ and $\overline{\beta}$. To see why the results of Imbens and Newey (2002) are not directly applicable, suppose β is a scalar and let $\widehat{a}\left(\widehat{\beta},\widehat{V}\right) \coloneqq \widehat{\beta}\left(x\right)$ and $a\left(\beta,V\right) \coloneqq \beta\left(x\right)$. Then, we have

$$\begin{aligned} & \widehat{a}\left(\widehat{\beta},\widehat{V}\right) - a\left(\beta,V\right) \\ = & \underbrace{\widehat{a}\left(\widehat{\beta},\widehat{V}\right) - \widehat{a}\left(\beta,\widehat{V}\right)}_{\text{known functional of } G(s)} + \underbrace{\widehat{a}\left(\beta,\widehat{V}\right) - \widehat{a}\left(\beta,V\right)}_{\text{estimation of } V} + \underbrace{\widehat{a}\left(\beta,V\right) - a\left(\beta,V\right)}_{\text{estimation of } a}. \end{aligned} \tag{1.49}$$

From (1.49), it is clear that because one needs to estimate both unknown functional a and unknown random variable V, in addition to the first term in (1.49) that concerns a known functional of $G\left(s\right)$, there are two more terms that affects the asymptotic normality of $\beta\left(x\right)$. In Appendix 1.B, we show how to correctly account for the effects from both estimation steps on influence functions. It is worth mentioning that for $\widehat{\overline{\beta}}$ one can significantly simplify the analysis by observing that $\widehat{\overline{\beta}}$ can be viewed as a "special case" of $\widehat{\beta}\left(x\right)$, that is, choosing $r^{M}\left(\cdot\right) \equiv 1$ in the definition of $\widehat{\beta}\left(x\right)$ gives $\widehat{\overline{\beta}}$. Therefore, with slight modifications to the proof for $\widehat{\beta}\left(x\right)$ one proves normality for $\widehat{\overline{\beta}}$.

1.5 Simulation

In this section, we examine the finite-sample performance of the method via a Monte Carlo simulation study. A discussion of the data generating process (DGP) motivated by production function applications is first provided. Then, we show the baseline results and compare the distribution of the estimated random coefficients with the simulated ones. Finally, several robustness checks are conducted to investigate how the proposed method performs when one varies the number of periods and firms, as well as orders of basis functions used for series estimation, and when one includes ex-post shocks to the DGP.

1.5.1 DGP

The baseline DGP we consider is

$$Y_{it} = \omega_{it} + X_{it}^K \beta_{it}^K + X_{it}^L \beta_{it}^L, \tag{1.50}$$

where the random coefficients $\left(\omega_{it}, \beta_{it}^K, \beta_{it}^L\right)$ are functions of (A_i, U_{it}), X_{it}^K and X_{it}^L are input choices of (natural log of) capital and labor, and Y_{it} is the (log of) output. Following the functional form of C-D production functions, $\left(X_{it}^K, X_{it}^L, Y_{it}\right)$ can be thought of already taking natural log. To allow correlation between A_i and U_{it}, an important feature in empirical applications, we draw $A_i \sim \mathcal{U}[1,2]$ and let $U_{it} = A_i \times \eta_{it}^I + \eta_t^{II}$ where $\eta_{it}^I \sim \mathcal{U}[1,3/2]$ and $\eta_t^{II} \sim \mathcal{U}[1,3/2]$ capture idiosyncratic and macro shocks, respectively. Then, we construct the random coefficients as $\omega_{it} = U_{it}$, $\beta_{it}^K = A_i + U_{it}$, and $\beta_{it}^L = A_i \times U_{it}$ and let $\beta_{it} = \left(\omega_{it}, \beta_{it}^K, \beta_{it}^L\right)'$. Thus, we have a total of $N \times T \times B$ β_{it}'s where N, T and B are total number of firms, periods, and simulations, respectively. Based on the DGP, the true $\overline{\omega} := \mathbb{E}\omega_{it} = 25/8$ and APEs of $\overline{\beta}^K := \mathbb{E}\beta_{it}^K = 37/8$ and $\overline{\beta}^L := \mathbb{E}\beta_{it}^L = 115/24$ are calculated and define $\overline{\beta} := \left(\overline{\omega}, \overline{\beta}^K, \overline{\beta}^L\right)'$. Finally, we draw each element of the instrument $Z_{it} = (R_{it}, W_{it}, P_{it})'$ independently from $\mathcal{U}[1,3]$, and calculate capital X_{it}^K and labor X_{it}^L by solving a representative firm's profit maximization problem

$$\begin{aligned} X_{it}^K &= \left[\left(1 - \beta_{it}^L\right) \ln\left(R_{it}/\beta_{it}^K\right) + \beta_{it}^L \ln\left(W_{it}/\beta_{it}^L\right) - \ln\left(\omega_{it} P_{it}\right)\right] / \left(\beta_{it}^K + \beta_{it}^L - 1\right), \\ X_{it}^L &= \left[\left(1 - \beta_{it}^K\right) \ln\left(W_{it}/\beta_{it}^L\right) + \beta_{it}^K \ln\left(R_{it}/\beta_{it}^K\right) - \ln\left(\omega_{it} P_{it}\right)\right] / \left(\beta_{it}^K + \beta_{it}^L - 1\right). \end{aligned}$$

Note that we do not include the ex-post shocks ε_{it} for the baseline scenario, but will add it later on to investigate how it affects the performance.

In the simulations, the observable data are (X, Y, Z). We use these data to estimate $\beta(v, w)$, $\beta(x)$, and $\overline{\beta}$ via the three-step estimation outlined in Section 1.4.1. Then, the performance of the estimated $\widehat{\beta}(v, w)$, $\widehat{\beta}(x)$, and $\widehat{\overline{\beta}}$ is evaluated against the truth.

1.5.2 Baseline Results

For the baseline configuration, we set $N = 1000$ and $T = 3$, and use basis functions of degree two splines with knot at the median. We run $B = 100$ simulations and summarize the performance of $\widehat{\overline{\omega}}, \widehat{\overline{\beta}}^K$ and $\widehat{\overline{\beta}}^L$ in Table 2.1.

Table 1.1: Performance of $\widehat{\overline{\omega}}, \widehat{\overline{\beta}}^K$ and $\widehat{\overline{\beta}}^L$

	Formula	$\widehat{\overline{\omega}}$	$\widehat{\overline{\beta}}^K$	$\widehat{\overline{\beta}}^L$
Bias	$B^{-1}\sum_b \left(\widehat{\overline{\beta}}_b^{(d)} - \overline{\beta}^{(d)}\right) / \left\lvert\overline{\beta}^{(d)}\right\rvert$	0.0119	0.0144	0.0066
rMSE	$\sqrt{B^{-1}\sum_b \left(\widehat{\overline{\beta}}_b^{(d)} - \overline{\beta}^{(d)}\right)^2 / \left\lvert\overline{\beta}^{(d)}\right\rvert}$	0.0318	0.0257	0.0323

Table 2.1 shows that the proposed method can accurately estimate the APE $\overline{\beta}$. Specifically, the first row evaluates the performance based on the normalized average bias for each coordinate of $\overline{\beta}$ across B rounds of simulations. The bias is small for all three coordinates, with a magnitude between 0.66% and 1.44% of the length of corresponding $\overline{\beta}^{(d)}$. The second row measures the normalized rMSE of $\widehat{\overline{\beta}}$ against true $\overline{\beta}$ for each coordinate, and shows that the method is able to achieve a low rMSE between 2.57% and 3.23% of the length of corresponding $\overline{\beta}^{(d)}$. By the standard bias-variance decomposition of MSE, the results in Table 2.1 show that the bias of the estimator for the APE is dominated by its variance.

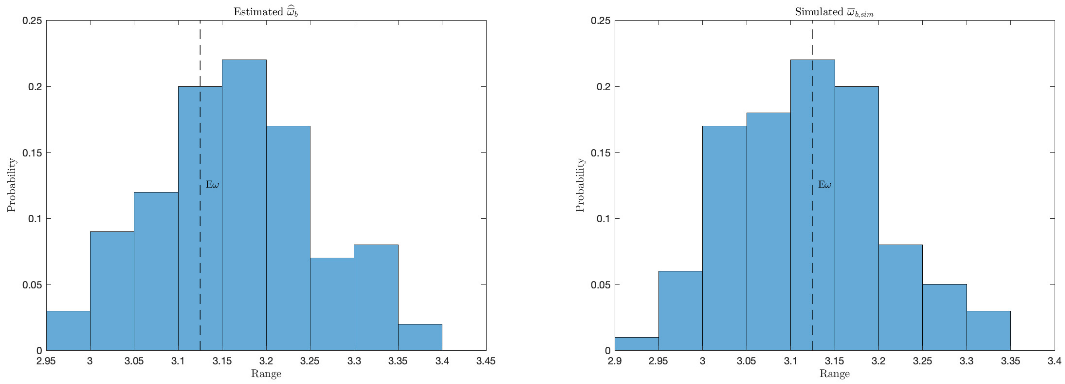


Figure 1.1: Histogram of $\widehat{\overline{\omega}}_b$ and $\overline{\omega}_b$

To provide more granular evidence on how well the proposed method can estimate the APE $\overline{\beta}$, we compare the histogram of the estimated $\widehat{\overline{\beta}}_b^{(d)}$ against the simulated APE $\overline{\beta}_b^{(d)} = (NT)^{-1}\sum_{i,t}\beta_{it,b}^{(d)}$, where $\beta_{it,b}^{(d)}$ is the d^{th} dimension of the it-specific β_{it} for the b^{th} round of simulation, across all B simulations. Figure 1.1 compares the distribution of $\widehat{\overline{\omega}}_b$ with $\overline{\omega}_b$ across those B simulations. It shows that the proposed method can capture the dispersion of the true $\overline{\omega}_b$ reasonably well. The distribution of $\widehat{\overline{\omega}}_b$ centers around $\mathbb{E}\omega_{it} = 25/8$, echoing the findings in Table 2.1. It is also worthwhile mentioning that the majority of $\widehat{\overline{\omega}}_b$ lies in $[2.95, 3.4]$, a short interval relative to the size of $\mathbb{E}\omega_{it}$. Note that the distribution of $\widehat{\overline{\omega}}_b$ appears to be slightly right-skewed across B simulations.

We conduct the same comparison for β^K and β^L and present the results in Figure 1.2 and 1.3, respectively. The results are similar to that obtained for ω. Once again, the method can capture the distributional characteristics of the true APE well, with the estimated coefficients located in a tight interval centered around the true APE.

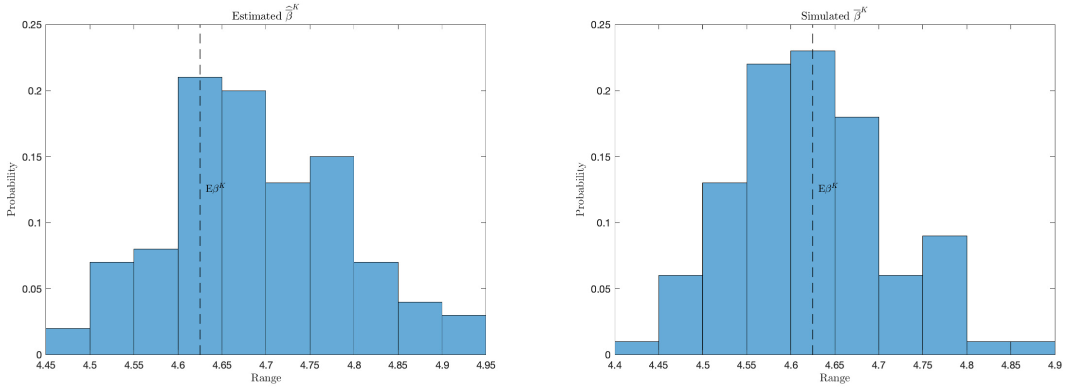


Figure 1.2: Histogram of $\widehat{\overline{\beta}}_b^K$ and $\overline{\beta}_b^K$

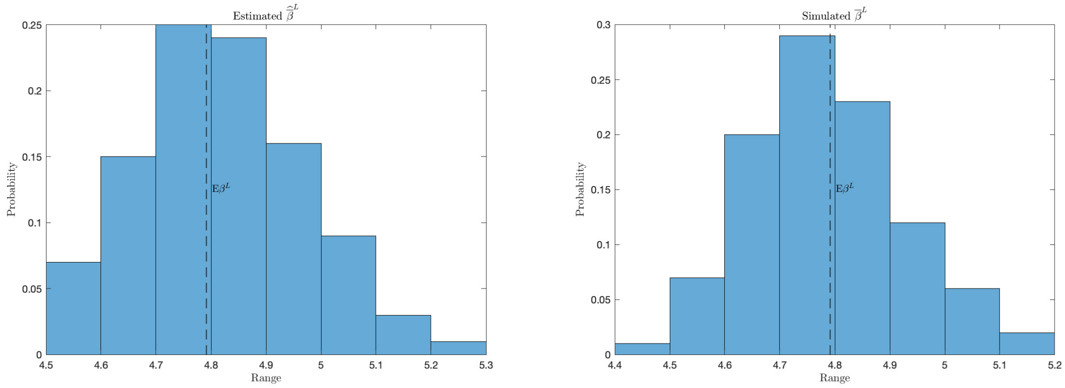


Figure 1.3: Histogram of $\widehat{\overline{\beta}}_b^L$ and $\overline{\beta}_b^L$

Finally, since $\beta\left(V_{it}, W_i\right)$ can be thought of as the "finest" approximation of β_{it}, one may wonder how closely the distribution of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ mimics that of true β_{it}. The distributional characteristics such as the variance of β_{it}^L can be important to answering policy-related questions. For example, policymakers may want to know the extent to which new labor augmenting technology is being diffused among firms. In the following analysis, we compare the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ with that of true β_{it} to show how accurately the method can capture the distributional properties of the random coefficients.

Figure 1.4–1.6 show the histogram of each coordinate of the estimated (brown) $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ versus that of true (blue) β_{it}. In all three figures, the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ centers around the corresponding population mean. It is worth mentioning that the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ seems more centered around its mean with slightly thinner tails than the corresponding coordinate of the simulated β_{it}, which is possibly caused by the fact that $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ is an estimator of $\mathbb{E}\left[\beta_{it}\middle| V_{it}, W_i\right]$ and thus already involves averaging across individuals with the same (V_{it}, W_i). Nonetheless, it is evident in Figure 1.4–1.6 that there is significant overlap between the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ and that of β_{it}, implying that the proposed method can accurately estimate both the mean and the dispersion of the random coefficients.

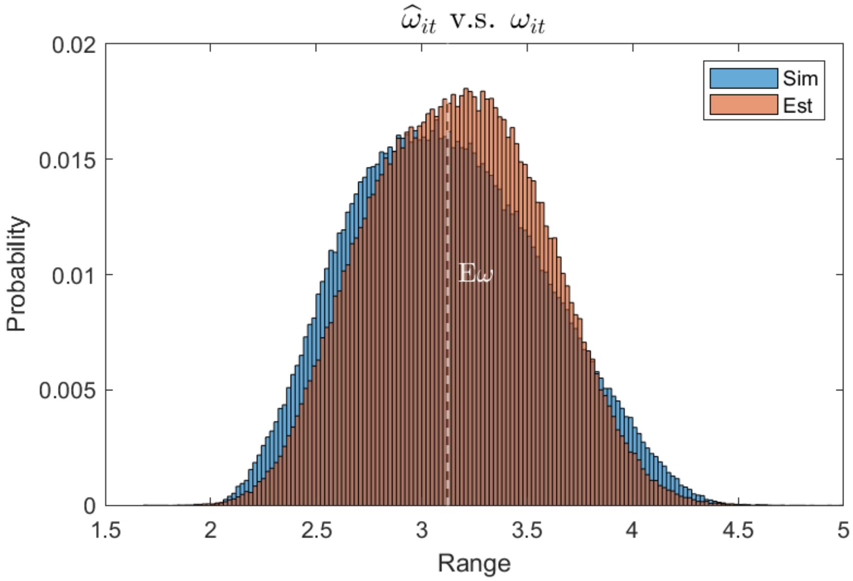


Figure 1.4: Histogram of $\widehat{\omega}_{it}$ versus ω_{it}

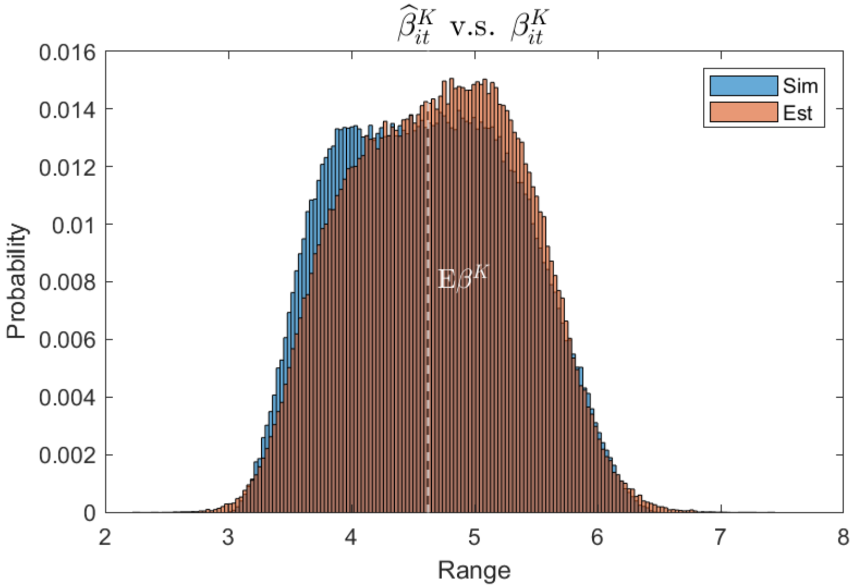


Figure 1.5: Histogram of $\widehat{\beta}_{it}^{K}$ versus β_{it}^{K}

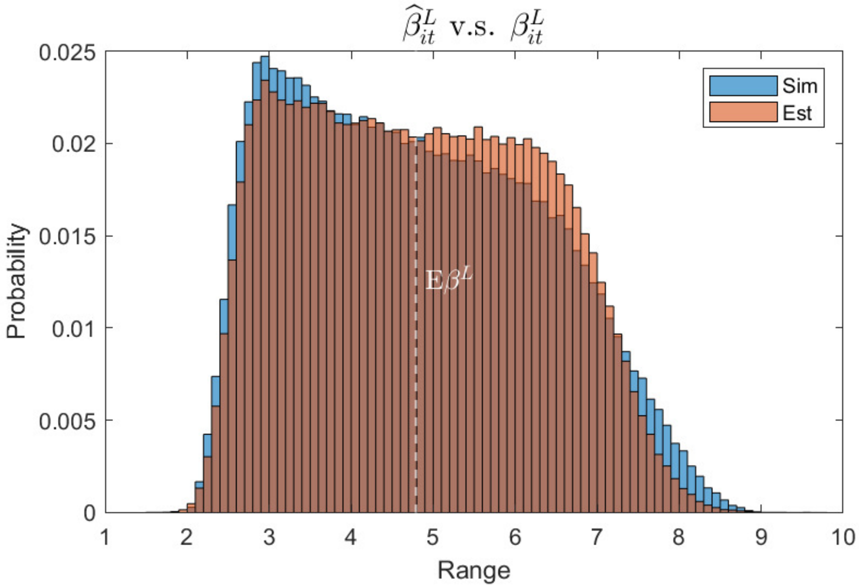


Figure 1.6: Histogram of $\widehat{\beta}_{it}^{L}$ versus β_{it}^{L}

Figure 1.6 is especially interesting because the true β_{it}^{L} follows a non-standard distribution that is right-skewed. Nonetheless, the histogram of $\widehat{\beta}_{it}^{L}$ looks very similar to the non-standard distribution of β_{it}^{L}, providing further evidence that the method works well even under irregular DGPs.

1.5.3 Robustness Checks

To show how robust the method is in estimating the APE, we conduct another set of exercises in this section. We evaluate the performance of the proposed method using both rMSE defined as $\sqrt{B^{-1}\sum_b \left\|\widehat{\overline{\beta}}_b - \overline{\beta}\right\|^2 / \left\|\overline{\beta}\right\|^2}$, and mean normed deviation (MND) defined as $B^{-1}\sum_b \left\|\widehat{\overline{\beta}}_b - \overline{\beta}\right\| / \left\|\overline{\beta}\right\|$.

First, we vary the size of N and T, and summarize the results in Table 1.2. As expected, a larger N is good for overall performance. We also find the proposed method benefit from the increase in T for each fixed N, possibly due to better controlling for the fixed effect A_i with more periods of data available for each individual.

Table 1.2: Performance under Varying N and T

	rMSE		MND	
	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
$T = 3$	0.0305	0.0298	0.0251	0.0242
$T = 5$	0.0241	0.0223	0.0206	0.0191

Second, we vary the order of the basis functions used to construct the series estimators, and present the results in Table 1.3. We find that increasing the orders of basis functions generally improves estimation accuracy. With that said, by using higher-order basis functions, one puts more pressure on the data because there are more regressors in each step of estimation, which may explain why the improvement in performance from increasing the order of basis functions from two to three is significantly smaller than that from going from one to two. Motivated by the simulation result, we use a basis function with an order of two in the empirical illustration in the next section.

Table 1.3: Performance under Varying Orders of Basis Functions

Order of Basis Functions	rMSE	MND
1	0.0607	0.0562
2	0.0298	0.0242
3	0.0290	0.0237

Lastly, we examine how including ε_{it}, interpreted as measurement error or ex-post shock, into the model affects finite sample performance. Specifically, $\varepsilon_{it} \sim \mathcal{U}\left[-1/2, 1/2\right]$ is drawn independently from all other variables. Results are presented in Table 1.4. It is clear that adding ε_{it} negatively affects the performance of the proposed estimator, however the impact is mild. When ε_{it} is included, rMSE increases from 0.0298 to 0.0391 and MND rises from 0.0242 to 0.0318. The magnitude in the change in performance is small, showing that the proposed method is robust to the inclusion of measurement error.

Table 1.4: Performance with and without Ex-Post Shock

Ex-Post Shock?	rMSE	MND
No	0.0298	0.0242
Yes	0.0391	0.0318

1.6 Production Function Application

In this section, we apply the procedure to comprehensive production data for Chinese manufacturing firms. Specifically, for each firm in the data we estimate a valued-added production function, where output elasticities and the intercept are allowed to vary across firms and periods, and, more importantly, input choices are allowed to depend on time-varying output elasticities and the random intercept in each period in a nonseparable way.

Output elasticity is an essential object of interest in the study of production functions as it quantifies how output responds to variations of each input, e.g., labor, capital, or material. It also helps answer important policy-related questions such as what returns to scale faced by a firm are, how the adoption of a new technology affects production, how the allocation of firm inputs relates to productivity, among others. Using the estimation method proposed in this paper, we find larger capital, but smaller labor, elasticities on average within each sector than those obtained by applying Olley and Pakes (1996)'s method (henceforth OP96) to the same data. The new estimates of average output elasticities in this paper are consistent with the literature on the measurement of factor income shares among manufacturing firms in China (Bai, Qian, and Wu, 2008; Jia and Shen, 2016). Then, a summary of the dispersions of the estimated output elasticities both across firms and through time is provided. Results show that there is substantial variation in the output elasticities in both dimensions, leading to a different interpretation of the data than in the misallocation literature pioneered by Hsieh and Klenow (2009).

The random intercept, usually considered as TFP in the C-D production function estimation literature, is another object of primary interest in the literature of firm innovation, R&D, trade openness, among others. We investigate the dispersion of the random intercept within each sector and compare them with those derived using OP96's method. Echoing recent results reported by Fox, Haddad, Hoderlein, Petrin, and Sherman (2016), we find larger dispersion in the random intercept among firms than those obtained using OP96's method. We provide an economic justification and investigate it empirically. Results show that the larger dispersion in the random intercept may be caused by its negative correlations with each of the output elasticities.

1.6.1 Data and Methodology

We use China Annual Survey of Industrial Firms, a comprehensive longitudinal micro-level data for the period of 1998–2007 that include information for all state-owned industrial firms and non-state-owned firms with annual sales above 5 million RMB. The data provide detailed information on ownership, production, and balance sheet of the firms surveyed. It is collected by National Bureau of Statistics of China and discussed in detail in Brandt, Van Biesebroeck, and Zhang (2014). Containing over 2 million observations, the data are representative of the industrial activity in China. According to Brandt, Van Biesebroeck, Wang, and Zhang (2017), they account for 91 percent of the gross output, 71 percent of employment, 97 percent of exports, and 91 percent of total fixed assets for the sampled periods. Many research on topics such as firm behavior, international trade, foreign direct investment, and growth theory use this data. See, for example, Hsieh and Klenow (2009), Song, Storesletten, and Zilibotti (2011), Brandt, Van Biesebroeck, Wang, and Zhang (2017), and Roberts, Yi Xu, Fan, and Zhang (2018).

This paper focuses on the manufacturing sector and follows Brandt, Van Biesebroeck, Wang, and Zhang (2017) to deal with the change in the Chinese Industry Classification codes occurred in 2003, which results in a total of 27 two-digit sectors. We choose to focus on two-digit sectors to ensure a large enough sample size for the robustness of the estimation results. The simulation results in Section 1.5 suggest the method can benefit from a larger T. Thus, firms that appear in the data for at least 6 years, with strictly positive amount of capital, employment, value-added output, wage expense and real interests are used for estimation. There are other sanity checks such as total assets should be no smaller than current assets. See Nie, Jiang, and Yang (2012) for a detailed discussion.

The final data is an unbalanced panel with the total number of firms increasing from 160K in 1998 to 330K in 2007. Only around 40K firms appear throughout the whole period, indicating a large amount of entry and exit behaviors in the data. The main variables include year, firm id, industry code, value-added output, capital, labor, and interest payments. Following Brandt, Van Biesebroeck, and Zhang (2014), appropriate price deflators for inputs and outputs are applied separately. The summary statistics are presented in Table 1.5.

Table 1.5: Summary Statistics

Variables	N	mean	sd	min	max
ln(value-added output)	415,333	9.155	1.441	-6.163	16.960
ln(capital)	415,215	9.352	1.644	0.077	18.560
ln(labor)	415,336	5.306	1.131	2.079	12.050
ln(interest)	415,336	5.960	1.741	0.012	14.350
Year	10	-	-	1998	2007
Firm ID	55,093	-	-	-	-
Industry Code	27	-	-	-	-

The value-added production function under consideration is

$$\begin{aligned} &Y_{it} = \omega_{it} + \beta_{it}^{K} K_{it} + \beta_{it}^{L} L_{it}, \\ &\beta_{it}^{K} = \beta^{K}\left(A_i, U_{it}\right), \ \beta_{it}^{L} = \beta^{L}\left(A_i, U_{it}\right), \ \omega_{it} = \omega\left(A_i, U_{it}\right), \\ &K_{it} = g^{K}\left(Z_{it}, A_i, U_{it}\right), \ L_{it} = g^{L}\left(Z_{it}, A_i, U_{it}\right), \ Z_{it} = \ln\left(\text{interest}\right), \end{aligned} \tag{1.51}$$

where Y_{it} and K_{it} are the natural log of inflation-adjusted real value-added output and capital measured in dollars as in Brandt, Van Biesebroeck, Wang, and Zhang (2017), respectively. There are two key features in the production function (1.51). First, the output elasticities wrt to capital β_{it}^K and labor β_{it}^L are both allowed to be time-varying and different across firms. Traditional methods (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015) do not allow for such heterogeneity. Second, and more importantly, the choices of capital K and labor L are modeled as nonparametric functions of fixed effect A_i interpreted as manager ability and idiosyncratic shock U_{it} interpreted as R&D outcome, both of which determine β^K and β^L. Therefore, model (1.51) allows input choices to depend on time-varying output elasticities in each period, a feature that naturally arises due to firm's profit maximization behavior.

It is worth noting that the output measure is the total revenue in dollars, not physical quantities in pieces due to lack of individual output prices in the data. When firms operate in distinct imperfectly competitive output markets, this may cause issues as pointed out by Klette and Griliches (1996). To allow for unobserved labor quality heterogeneity, we measure labor input in dollars. As a consequence, firm level average wages cannot be used as an instrument because it is already included in the labor input in the baseline case. The instrument Z_{it} is the log of real interests, which is likely to be exogenous because its fluctuation is mostly driven by exogenous policy in China. For robustness purposes, we use the inter-temporal difference in log of real interests and both interests and wages as instruments, and find the results are quite similar. There are other possible choices of instruments including local minimum wage, lagged inputs (De Loecker and Warzynski, 2012; Shenoy, 2020), demand instruments (Goldberg, Khandelwal, Pavcnik, and Topalova, 2010), and product/firm characteristics of direct competitors within the same sector and location (Berry, Levinsohn, and Pakes, 1995b).

We estimate conditional and unconditional expectations of the individually unique and time-varying output elasticities $\beta_{it} := \left(\beta_{it}^K, \beta_{it}^L\right)$ as well as random intercept ω_{it} within each two-digit sector. More specifically, first we construct $W_i := \left(\overline{K}_i, \overline{L}_i, \overline{Z}_i, \overline{K}_i^2, \overline{L}_i^2, \overline{Z}_i^2\right)$, where the means are through time. Then, we estimate $V_{it} := F_{K_{it}|Z_{it},W_i}\left(K_{it}|\, Z_{it}, W_i\right)$ using second-order polynomial basis functions. The choice of the order of basis functions is motivated by simulation results in Section 1.5. Next, the conditional expectation of Y_{it} given $(K_{it}, L_{it}, V_{it}, W_i)$, defined as G_{it}, is estimated with a series estimator where $\widehat{V}_{it}$ from the previous step is plugged in. Finally, we estimate $\beta\left(V_{it}, W_i\right) := \mathbb{E}\left[\beta_{it}|\, V_{it}, W_i\right]$ by taking the partial derivative of G_{it} with respect to (K_{it}, L_{it}). With moments of β_{it} obtained, we estimate the moments of ω_{it} by exploiting the index structure in (1.51).

1.6.2 Results

Applying the proposed method on the data for Chinese manufacturing firms, we obtain estimates of the conditional expectation of output elasticities $\beta\left(V_{it}, W_i\right)$ and random intercept $\omega\left(V_{it}, W_i\right)$ for each firm in each year. Yang (2015) applies OP96's method to the same data used in this paper to estimate a value-added production function. Therefore, the results are directly comparable. First, we compare the mean of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ within each sector through time with that obtained using OP96's method. Second, the dispersions of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ both across firms and through time are presented. Lastly, we compare the dispersion of $\widehat{\omega}\left(\widehat{V}_{it}, W_i\right)$ across firms within each sector with that derived using OP96's method.

Average Output Elasticities

In this section, we compare the mean of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ within each sector through time with that obtained using OP96's method. Output elasticity is an essential object of interest in economics because it quantifies how responsive output is to variations of each input. Moreover, by the solution to the canonical firm's profit maximization problem (PMP) given C-D production functions in a perfectly competitive market, the output elasticities equal the input cost share of total outputs, i.e., $\beta^K = rK/pY$ and $\beta^L = wL/pY$ where (w, r, p) stand for wage, interest rate and output price, respectively. If firms maximize their profits when choosing inputs, the estimated output elasticities should in theory be close to input income shares. Therefore, one may be interested in comparing the estimated elasticities with input income shares measured from the data. Note that the result that the output elasticity equals the corresponding input income share obtained by solving the PMP holds for C-D production functions regardless of whether the inputs and output are measured using quantities or dollars.

First, we average $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ across firms and through time within each sector, and compare it with those obtained using OP96's method on the same data. Results are summarized in Figure 1.7. Our estimates of the average capital elasticities are larger than that obtained using OP96's method for all but one sectors. The average capital elasticity across all sectors is 49% using our method, whereas the number is 35% by applying OP96's method to the same data. We repeat the same analysis for $\widehat{\beta}^L\left(\widehat{V}_{it}, W_i\right)$ and find that the pattern is reversed for labor elasticities. Figure 1.8 shows that our estimates of the average labor elasticities are consistently smaller than that obtained by applying OP96's method to the same data for each of the 27 sectors. Our estimate of average labor elasticities across all sectors is 43%, which is significantly smaller than 62% obtained using OP96's method.

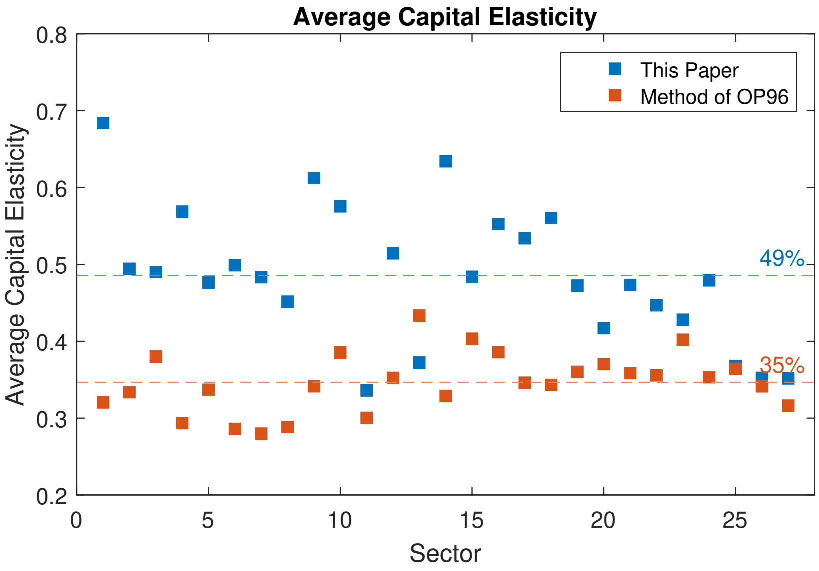


Figure 1.7: Comparison of Average Capital Elasticities

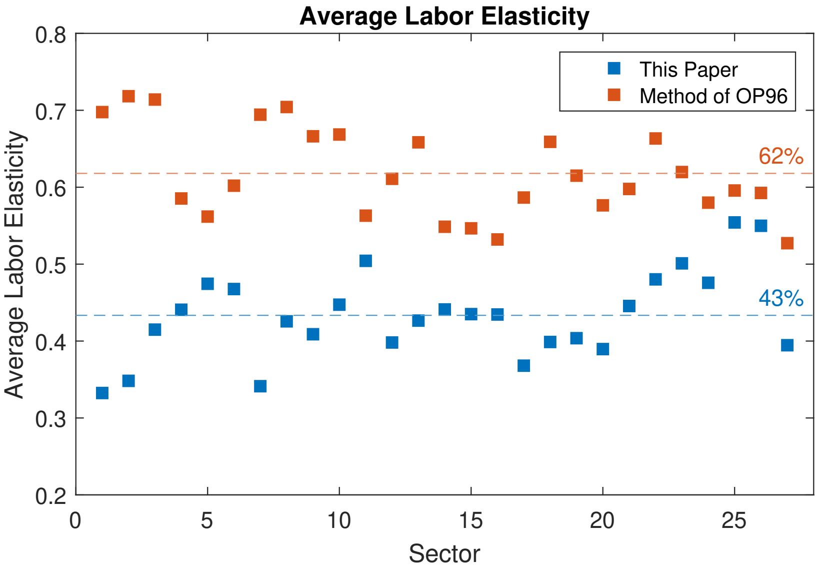


Figure 1.8: Comparison of Average Labor Elasticities

Based on the theoretical result that output elasticities equal corresponding factor income shares, we compare the estimated elasticities with the factor income shares measured in the literature. Bai, Qian, and Wu (2008) estimates the average capital income shares to be 55–65% for manufacturing sectors between 1998–2005 in China. A more recent result by Jia and Shen (2016) shows that on average 50–60% of total output is distributed to capital. Hsieh and Klenow (2009) briefly mentioned that roughly half of output is distributed to capital according to the Chinese input-output tables and the national accounts. As can be seen from Figure 1.7, the average estimated capital elasticity is 49%, which by the solution to firm's PMP means about half of total output is distributed to capital. Therefore, our estimates are consistent with the factor income shares documented in the literature. In contrast, the average capital elasticity using OP96's method for Chinese manufacturing firms is only 35%.

The results show that the proposed method in this paper is able to obtain estimates of elasticities that are closer to those found in the factor income share literature. One possible explanation for the results is that it is firm's optimization behavior that leads to the first-order condition of $\beta^K = rK/pY$ and $\beta^L = wL/pY$. When β_{it}'s are random, it is natural that the elasticities affect the choice of each input in each period, leading to time-varying endogeneity through the random coefficients. Our TERC model explicitly takes firm's optimization behavior into account, whereas traditional fixed coefficients models do not allow for this feature. As a consequence, the correlations between β_{it} and X_{it} are not captured in traditional fixed coefficients models, leading to a potential omitted variable bias.

Dispersions of the Output Elasticities

Next, we examine the variations of the output elasticities with respect to each input. More specifically, because the elasticities are not comparable across sectors, we calculate the standard deviation of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ within each sector for each year, excluding top and bottom 1% extreme values for robustness purposes. These standard deviations are then normalized by the absolute value of the mean of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ within each sector for each year. The dispersion of the normalized standard deviations across sectors is summarized in Figure 1.9.

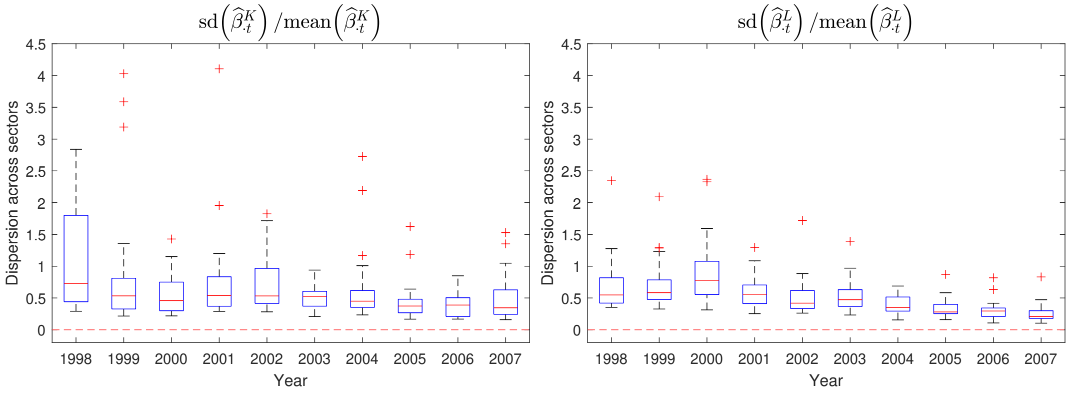
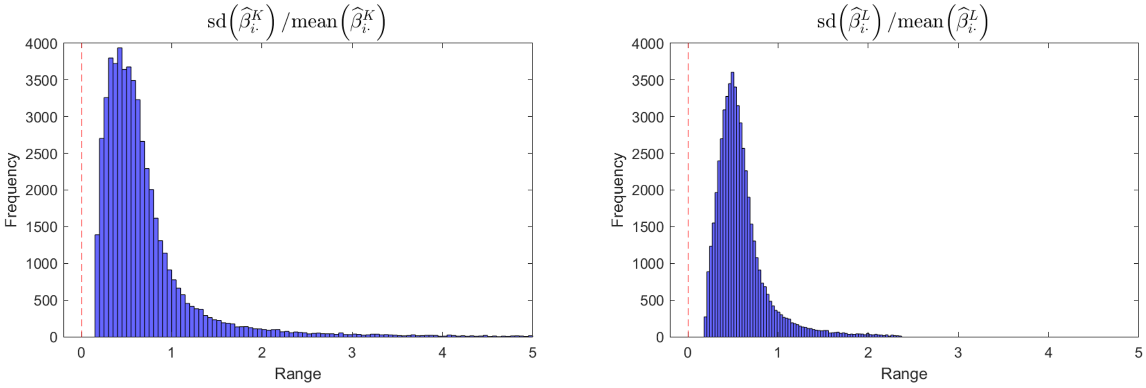


Figure 1.9: Dispersions of Elasticities across Firms

Results show that there are substantial variations in each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ among firms within each sector for each year. More precisely, the normalized standard deviation of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ in 1998 has a median of around 0.7 and a maximum of about 2.9, which implies that the median sector and the maximum sector have a standard deviation that is about 70% and 2.9 times of the absolute value of their means of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$, respectively. A similar pattern is also found for $\widehat{\beta}^L\left(\widehat{V}_{it}, W_i\right)$, with the magnitude of the standard deviations slightly smaller than that of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$.

Another important feature of the model in this paper is that the random coefficients are allowed to be time-varying. To show how dispersed the elasticities are through time, we first calculate the standard deviation of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ through time for each firm. Then, the standard deviations are normalized by the absolute value of the means of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ for the same firm through time. As a consequence, the normalized standard deviations are directly comparable across firms. We pool the normalized standard deviations together and summarize the results in Figure 1.10.

According to Figure 1.10, there are significant variations in output elasticities with respect to both capital and labor through time. The majority of the normalized standard deviations of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ lies around 0.5, implying that for these firms the standard deviation of the output elasticity with respect to capital through time is about 50% of its mean through time. The normalized standard deviation of the output elasticity with respect to labor through time also centers around 0.5, however with a smaller maximum of about 2 times compared to that of 5.5 times for capital. Note that if one uses fixed coefficient linear models, the standard deviations of the elasticities both across firms and through time will be constant zero by definition.


$\text{sd}\left(\widehat{\beta}^K_{i\cdot}\right)/\text{mean}\left(\widehat{\beta}^K_{i\cdot}\right)$

$\text{sd}\left(\widehat{\beta}^L_{i\cdot}\right)/\text{mean}\left(\widehat{\beta}^L_{i\cdot}\right)$

Figure 1.10: Dispersions of Elasticities through Time

The dispersions of the output elasticities across firms and periods provide an explanation to the observed variation in input cost shares across firms that is different from the misallocation theory pioneered by Hsieh and Klenow (2009). Hsieh and Klenow (2009) model the elasticities as constants and attribute the variation the marginal revenue product of inputs to external distortions that the firm faces. They further identify the distortions using firm's first-order condition shown as equation (17)–(18) in their paper, assuming the elasticities are constant across firms and periods. However, there is no obvious reason why the output elasticities should be the same for intrinsically heterogeneous firms. In addition to distortions, the firms may also have different elasticities driven by their fixed effect and idiosyncratic shocks in each period. Therefore, the dispersions shown in Figure 1.9–1.10 provide an alternative explanation to the observed variation in input cost shares across firms than the misallocation theory.

Dispersion of the Random Intercept

Lastly, we compare the estimated dispersion of the random intercept within each sector with that obtained by applying OP96's method on the same data. OP96 allow the intercept to be both time-varying and correlated with input choices, but require the output elasticities to be constants. Using OP96's method, Yang (2015) obtains estimates of intercepts for each firm and year. We compare the estimated $\widehat{\omega}\left(\widehat{V}_{it}, W_i\right)$ with his results. For robustness purposes, we exclude the top and bottom 1% of the estimated $\widehat{\omega}\left(\widehat{V}_{it}, W_i\right)$ within each sector for each year. Then, we compute the standard deviations of $\widehat{\omega}\left(\widehat{V}_{it}, W_i\right)$ for each sector and year, normalized by the absolute value of the mean of $\widehat{\omega}\left(\widehat{V}_{it}, W_i\right)$ for the corresponding sector and year. We do the same trimming and normalization for the estimates based on OP96's method. Results for all years and sectors are pooled together and summarized in Figure 1.11.

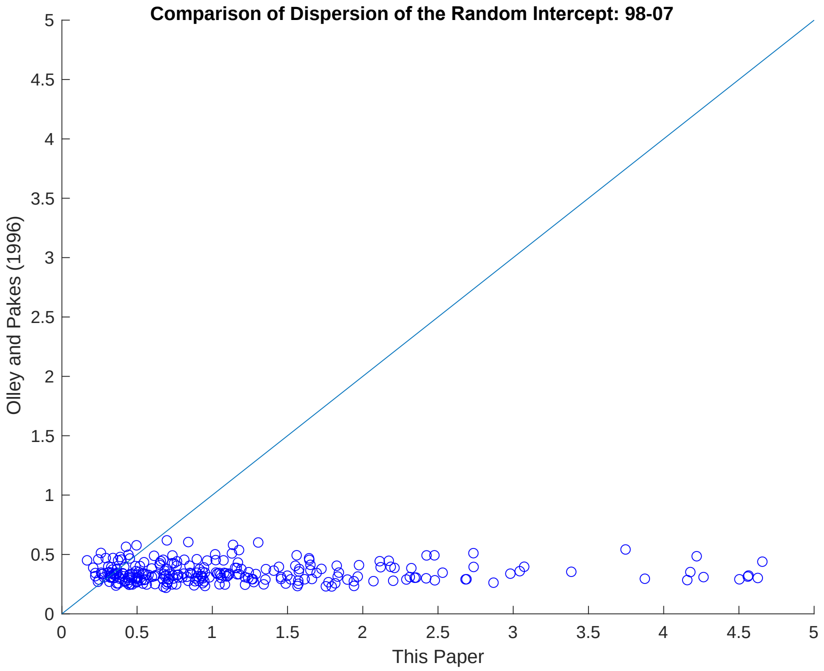


Figure 1.11: Comparison of Dispersion of the Random Intercept

In Figure 1.11, the horizontal axis represents the normalized standard deviation of the random intercept within each sector obtained using this paper's method while the vertical axis stands for the normalized standard deviation derived using OP96's method. Each blue circle corresponds to a sector and year. When the circle is located to the right of the 45 degree line, the normalized standard deviation of the random intercept using our method is larger than that obtained using OP96's method. As is evident from Figure 1.11, the majority of the dispersions of the random intercept calculated using our method are larger than that obtained using OP96's method. The results of this paper echo the findings of Fox, Haddad, Hoderlein, Petrin, and Sherman (2016), who model the output elasticities as random walk processes and apply their model to Indian production data. They find a larger dispersion of random intercept than that derived using OLS regression with fixed coefficients.

One of the possible explanations to why making the coefficients random *increases* the dispersion of the random intercept is that it is negatively correlated with output elasticities. In a linear production function, the random intercept contains all the latent factors used in the production process that are not explicitly included as regressors in the model. When, for example, the output elasticity with respect to labor is large for a certain period due to a positive shock, the firm can take advantage of it and hire more workers, reducing the contribution to output from the latent factors because the firm may have a limited budget to spend on all factors. Therefore, it can be the substitution effect between the observed and latent inputs that causes the negative correlation between the random intercept and output elasticities.

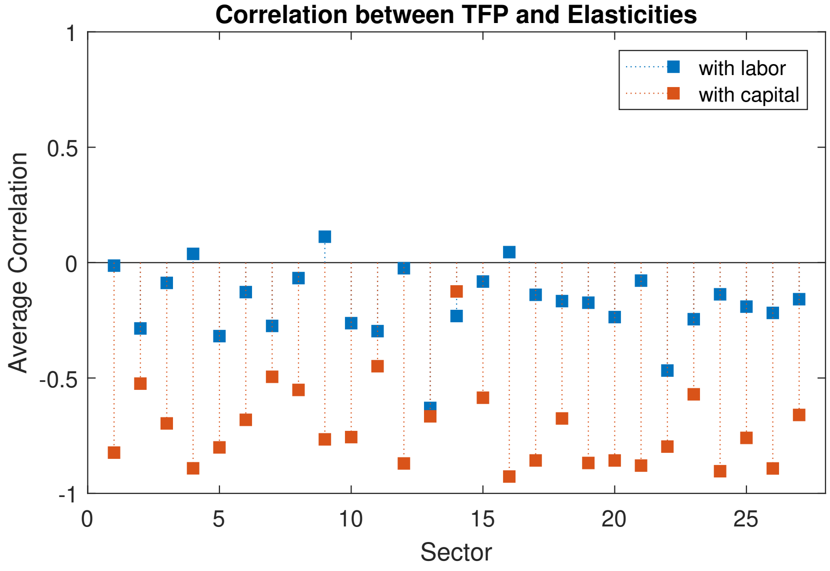


Figure 1.12: Estimated Correlation between the Random Intercept and Elasticities

We take this idea to the data, and run estimation based on the identification of second-order moments of the random coefficients in (1.29). More specifically, we estimate $\widehat{Corr}\left(\omega_{it}, \beta_{it}^{L}\right)$ and $\widehat{Corr}\left(\omega_{it}, \beta_{it}^{K}\right)$ for each sector, and summarize the results in Figure 1.12. The estimated correlation coefficients between the random intercept and capital elasticity are negative consistently across all sectors. A similar pattern is found for labor elasticity with only three sectors reporting small positive correlation coefficient around zero. The results provide empirical evidence that the larger dispersion of the random intercept is likely to be caused by a negative correlation between the random intercept and the output elasticities.

1.7 Conclusion

This paper proposes a flexible random coefficients panel model where the regressors are allowed to depend on the time-varying random coefficients in each period, a critical feature in many economic applications such as production function estimation. The model allows for a nonseparable first-step equation, a nonlinear fixed effect of arbitrary dimension, and an idiosyncratic shock that can be arbitrarily correlated with the fixed effect and that affects the choice of the regressors in a nonlinear way. A sufficiency argument is used to control for the fixed effect, which enables one to construct a feasible control function for the random shock and subsequently identify the moments of the random coefficients. We provide consistent series estimators for the moments of the random coefficients and prove a new asymptotic normality result. Applying the estimation procedure to panel data for Chinese manufacturing firms, we obtain three main findings. First, larger capital, but smaller labor, elasticities are derived than those obtained using traditional methods. Our estimates are consistent with the findings in the factor income share literature. Second, there are substantial variations in the output elasticities across firms and periods, providing a different explanation to the observed variation in input cost shares from the well-known misallocation theory. Third, the dispersion of the random intercept is larger than that obtained using classical methods, caused by negative correlations between the random intercept and each of the output elasticities.

We mention several extensions to this paper for future research. First, although we have briefly discussed how to identify second-order moments of the random coefficients in Section 1.3, it remains an open question how to separate the variance of the exogenous ex-post shocks from that of the random intercept. One may follow Arellano and Bonhomme (2012) to impose time-dependence assumptions such as moving average process on the ex-post shock. Second, one may prefer to include lagged regressors in the first-step equation (1.3). We have provided a group exchangeability condition (1.32) that can allow first-step function $g\left(Z, A, U\right)$ in (1.3) to also depend on lagged regressors X_{it-1}. Nonetheless, it can be challenging to obtain asymptotic properties for the estimators with group fixed effects. Another related question is whether one can incorporate the timing assumptions widely used in the proxy variable based approaches to make lagged inputs valid instruments. Third, it can be useful to construct a test of whether the coefficients vary across individuals and/or through time.

References

Abito, J. M. (2020): "Estimating Production Functions with Fixed Effects," *Available at SSRN 3510068*.

Ackerberg, D., X. Chen, and J. Hahn (2012): "A practical asymptotic variance estimator for two-step semiparametric estimators," *Review of Economics and Statistics*, 94, 481–498.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81, 919–943.

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification properties of recent production function estimators," *Econometrica*, 83, 2411–2451.

——— (2007): "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables," *Journal of Econometrics*, 141, 5–43.

Altonji, J. G. and R. L. Matzkin (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73, 1053–1102.

Andrews, D. W. K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307–45.

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

Bai, C.-E., Z. Qian, and K. Wu (2008): "Determinants of Factor Shares in China's Industrial Sector," *Economic Research Journal*, 16–28.

Bajari, P., J. T. Fox, and S. P. Ryan (2007): "Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients," *American Economic Review*, 97, 459–463.

Balestra, P. and M. Nerlove (1966): "Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas," *Econometrica: Journal of the econometric society*, 585–612.

Bang, M., W. Gao, A. Postlewaite, and H. Sieg (2020): "Estimating Production Functions with Partially Latent Inputs," .

——— (1995b): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, 841–890.

Blundell, R., X. Chen, and D. Kristensen (2007a): "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 75, 1613–1669.

Blundell, R., T. MaCurdy, and C. Meghir (2007b): "Labor supply models: Unobserved heterogeneity, nonparticipation and dynamics," *Handbook of econometrics*, 6, 4667–4775.

Blundell, R. and J. L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," *Econometric society monographs*, 36, 312–357.

Brandt, L., J. Van Biesebroeck, L. Wang, and Y. Zhang (2017): "WTO accession and performance of Chinese manufacturing firms," *American Economic Review*, 107, 2784–2820.

——— (2014): "Challenges of working with the Chinese NBS firm-level data," *China Economic Review*, 30, 339–352.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2012): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*.

Cameron, A. C. and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50, 317–372.

Chamberlain, G. (1984): "Panel data," *Handbook of econometrics*, 2, 1247–1318.

——— (1992): "Efficiency bounds for semiparametric regression," *Econometrica: Journal of the Econometric Society*, 567–596.

Chen, X. and Z. Liao (2014): "Sieve M inference on irregular parameters," *Journal of Econometrics*, 182, 70–86.

——— (2015): "Sieve semiparametric two-step GMM under weak dependence," *Journal of Econometrics*, 189, 163–186.

Chen, Y., M. Igami, M. Sawada, and M. Xiao (2020): "Privatization and productivity in china," *Available at SSRN 2695933*.

Chernozhukov, V., J. A. Hausman, and W. K. Newey (2019b): "Demand analysis with many prices," Tech. rep., National Bureau of Economic Research.

De Loecker, J. and F. Warzynski (2012): "Markups and firm-level export status," *American economic review*, 102, 2437–71.

Demirer, M. (2020): "Production function estimation with factor-augmenting technology: An application to markups," Tech. rep., MIT working paper.

D'Haultfœuille, X. and P. Février (2015): "Identification of nonseparable triangular models with discrete instruments," *Econometrica*, 83, 1199–1210.

Dhyne, E., A. Petrin, V. Smeets, and F. Warzynski (2020): "Theory for Extending Single-Product Production Function Estimation to Multi-Product Settings," *Working Paper.*

——— (2018): "Measuring the bias of technological change," *Journal of Political Economy*, 126, 1027–1084.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012): "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation," *Econometrica*, 80, 2231–2267.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76, 1191–1206.

Fox, J. T., V. Haddad, S. Hoderlein, A. K. Petrin, and R. P. Sherman (2016): "Heterogeneous production functions, panel data, and productivity dispersion," *Slide deck, Rice Univ.*

Gandhi, A., S. Navarro, and D. A. Rivers (2020): "On the identification of gross output production functions," *Journal of Political Economy*, 128, 2973–3016.

Gautier, E. and Y. Kitamura (2013): "Nonparametric estimation in random coefficients binary choice models," *Econometrica*, 81, 581–607.

Goldberg, P. K., A. K. Khandelwal, N. Pavcnik, and P. Topalova (2010): "Imported intermediate inputs and domestic product growth: Evidence from India," *The Quarterly journal of economics*, 125, 1727–1767.

Graham, B. S. and J. L. Powell (2012): "Identification and estimation of average partial effects in irregular correlated random coefficient panel data models," *Econometrica*, 80, 2105–2152.

HAHN, J. AND G. RIDDER (2013): "Asymptotic variance of semiparametric estimators with generated regressors," *Econometrica*, 81, 315–340.

——— (2019): "Three-stage semi-parametric inference: Control variables and differentiability," *Journal of econometrics*, 211, 262–293.

——— (2014b): *Variable-Coefficient Models*, Cambridge University Press, 167–229, Econometric Society Monographs, 3 ed.

HSIEH, C.-T. AND P. J. KLENOW (2009): "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 124, 1403–1448.

IMBENS, G. W. AND W. K. NEWEY (2002): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *NBER Working Paper*.

——— (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

JIA, S. AND G. SHEN (2016): "Corporate Risk and Labor Income Share: Evidence from China's Industrial Sector," *Economic Research Journal*, 116–129.

JONNSON, W. (1924): *Logic, Part III: Logical Foundation of Science*, Cambridge University Press.

KASAHARA, H., P. SCHRIMPF, AND M. SUZUKI (2015): "Identification and estimation of production function with unobserved heterogeneity," *University of British Columbia mimeo*.

KITAMURA, Y. AND J. STOYE (2018): "Nonparametric analysis of random utility models," *Econometrica*, 86, 1883–1909.

KLETTE, T. J. AND Z. GRILICHES (1996): "The inconsistency of common scale estimators when output prices are unobserved and endogenous," *Journal of applied econometrics*, 11, 343–361.

Kyriazidou, E. (1997): "Estimation of a panel data sample selection model," *Econometrica: Journal of the Econometric Society*, 1335–1364.

Laage, L. (2020): "A correlated random coefficient panel model with time-varying endogeneity," *arXiv preprint arXiv:2003.09367*.

Lee, Y., A. Stoyanov, and N. Zubanov (2019): "Olley and Pakes-style Production Function Estimators with Firm Fixed Effects," *Oxford Bulletin of Economics and Statistics*, 81, 79–97.

Lee, Y.-Y. (2018): "Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models," *Available at SSRN 3250485*.

León-Ledesma, M. A., P. McAdam, and A. Willman (2010): "Identifying the elasticity of substitution with biased technical change," *American Economic Review*, 100, 1330–57.

Levinsohn, J. and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70, 317–341.

Li, T. and Y. Sasaki (2017): "Constructive Identification of Heterogeneous Elasticities in the Cobb-Douglas Production Function," *arXiv preprint arXiv:1711.10031*.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric regression with nonparametrically generated covariates," *The Annals of Statistics*, 40, 1132–1170.

——— (2016): "Semiparametric estimation with generated covariates," *Econometric Theory*, 32, 1140–1177.

Marschak, J. and W. H. Andrews (1944): "Random simultaneous equations and the theory of production," *Econometrica, Journal of the Econometric Society*, 143–205.

McCall, J. J. (1991): "Exchangeability and its economic applications," *Journal of Economic Dynamics and Control*, 15, 549 – 568.

Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Econometrica: journal of the Econometric Society*, 69–85.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, 79, 147–168.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67, 565–603.

Nie, H., T. Jiang, and R. Yang (2012): "The Current Status and Potential Issues of Chinese Industrial Enterprises Database," *The Journal of World Economy*, 5.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Roberts, M. J., D. Yi Xu, X. Fan, and S. Zhang (2018): "The role of firm factors in demand, cost, and export market selection for Chinese footwear producers," *The Review of Economic Studies*, 85, 2429–2461.

Shenoy, A. (2020): "Estimating the Production Function Under Input Market Frictions," *Review of Economics and Statistics*, 1–45.

Song, Z., K. Storesletten, and F. Zilibotti (2011): "Growing like china," *American economic review*, 101, 196–233.

Stoyanov, J. (2000): "Krein condition in probabilistic moment problems," *Bernoulli*, 6, 939–949.

Torgovitsky, A. (2015): "Identification of nonseparable models using instruments with small support," *Econometrica*, 83, 1185–1197.

Weyl, H. (1939): *The Classical Groups: Their Invariants and Representations.*, Princeton University Press.

Wooldridge, J. M. (2005a): “Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models,” *Review of Economics and Statistics*, 87, 385–390.

——— (2005b): “Unobserved heterogeneity and estimation of average partial effects,” *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 27–55.

Yang, R. (2015): “Study on the Total Factor Productivity of Chinese Manufacturing Enterprises,” *Economic Research Journal*, 2, 61–74.

Appendix

1.A Proofs in Section 1.3

Proof of Lemma 1.1. The proof is divided into two parts. First, we establish the exchangeability condition (1.15) using Assumption 1.2. Then, we show that there exist W_i such that (1.14) holds. For simplicity of notations, we assume X_{it} and Z_{it} are both scalars. The proof goes through when X_{it} and Z_{it} are vectors. We prove (1.15) for $T = 2$, which is wlog because T is finite and thus any permutation of $(1, ..., T)$ can be achieved by switching pairs of (t_i, t_j) finite number of times. For example, one can obtain (t_3, t_1, t_2) from (t_1, t_2, t_3) by $(t_1, t_2, t_3) \to (t_1, t_3, t_2) \to (t_3, t_1, t_2)$. We suppress i subscripts in all variables in this proof.

By Assumption 1.2, we have

$$f_{U_1,U_2|A}\left(u_1, u_2 | \, a\right) = f_{U_1,U_2|A}\left(u_2, u_1 | \, a\right), \tag{1.52}$$

which implies

$$f_{A,U_1,U_2}\left(a, u_1, u_2\right) = f_{A,U_1,U_2}\left(a, u_2, u_1\right). \tag{1.53}$$

Let $g^{-1}\left(X, Z, A\right)$ denote the inverse function of $g\left(Z, A, U\right)$ with respect to U. Define $u_1 = g^{-1}\left(x_1, z_1, a\right)$ and $u_2 = g^{-1}\left(x_2, z_2, a\right)$. Calculate the determinants of the Jacobians as

$$J_1$$

$$
\begin{aligned}
&:= \left| \begin{matrix} \dfrac{\partial A}{\partial X_1} & \dfrac{\partial A}{\partial X_2} & \dfrac{\partial A}{\partial A} \\ \dfrac{\partial g^{-1}(X_1, Z_1, A)}{\partial X_1} & \dfrac{\partial g^{-1}(X_1, Z_1, A)}{\partial X_2} & \dfrac{\partial g^{-1}(X_1, Z_1, A)}{\partial A} \\ \dfrac{\partial g^{-1}(X_2, Z_2, A)}{\partial X_1} & \dfrac{\partial g^{-1}(X_2, Z_2, A)}{\partial X_2} & \dfrac{\partial g^{-1}(X_2, Z_2, A)}{\partial A} \end{matrix} \right|_{\substack{(X_1, X_2, Z_1, Z_2, A) \\ = (x_1, x_2, z_1, z_2, a)}} \\
&= \left| \begin{matrix} 0 & 0 & 1 \\ \dfrac{\partial g^{-1}(X_1, Z_1, A)}{\partial X_1} & 0 & \dfrac{\partial g^{-1}(X_1, Z_1, A)}{\partial A} \\ 0 & \dfrac{\partial g^{-1}(X_2, Z_2, A)}{\partial X_2} & \dfrac{\partial g^{-1}(X_2, Z_2, A)}{\partial A} \end{matrix} \right|_{\substack{(X_1, X_2, Z_1, Z_2, A) \\ = (x_1, x_2, z_1, z_2, a)}} \\
&= \partial g^{-1}(X, Z, A) / \partial X\Big|_{(X,Z,A)=(x_1,z_1,a)} \times \partial g^{-1}(X, Z, A) / \partial X\Big|_{(X,Z,A)=(x_2,z_2,a)}, \qquad (1.54)
\end{aligned}
$$

and

$$
\begin{aligned}
&J_2 \\
&:= \left| \begin{matrix} \dfrac{\partial g(Z_1, A, U_1)}{\partial A} & \dfrac{\partial g(Z_1, A, U_1)}{\partial U_1} & \dfrac{\partial g(Z_1, A, U_1)}{\partial U_2} \\ \dfrac{\partial g(Z_2, A, U_2)}{\partial A} & \dfrac{\partial g(Z_2, A, U_2)}{\partial U_1} & \dfrac{\partial g(Z_2, A, U_2)}{\partial U_2} \\ \dfrac{\partial A}{\partial A} & \dfrac{\partial A}{\partial U_1} & \dfrac{\partial A}{\partial U_2} \end{matrix} \right|_{\substack{(Z_1, Z_2, A, U_1, U_2) \\ = (z_2, z_1, a, u_2, u_1)}} \\
&= \left| \begin{matrix} \dfrac{\partial g(Z_1, A, U_1)}{\partial A} & \dfrac{\partial g(Z_1, A, U_1)}{\partial U_1} & 0 \\ \dfrac{\partial g(Z_2, A, U_2)}{\partial A} & 0 & \dfrac{\partial g(Z_2, A, U_2)}{\partial U_2} \\ 1 & 0 & 0 \end{matrix} \right|_{\substack{(Z_1, Z_2, A, U_1, U_2) \\ = (z_2, z_1, a, u_2, u_1)}} \\
&= \partial g(Z, A, U) / \partial U|_{(Z,A,U)=(z_2,a,u_2)} \times \partial g(Z, A, U) / \partial U|_{(Z,A,U)=(z_1,a,u_1)}. \qquad (1.55)
\end{aligned}
$$

Then, we have

$$
f_{X_1,X_2,A|Z_1,Z_2}(x_1, x_2, a|\, z_1, z_2)
$$

$$
\begin{aligned}
&= f_{A,U_1,U_2|Z_1,Z_2}\left(a, g^{-1}\left(x_1, z_1, a\right), g^{-1}\left(x_2, z_2, a\right)\Big| z_1, z_2\right)\left|J_1\right| \\
&= f_{A,U_1,U_2|Z_1,Z_2}\left(a, g^{-1}\left(x_2, z_2, a\right), g^{-1}\left(x_1, z_1, a\right)\Big| z_2, z_1\right)\left|J_1\right| \\
&= f_{X_1,X_2,A|Z_1,Z_2}\left(x_2, x_1, a\right| z_2, z_1\left)\right|J_2 J_1| \\
&= f_{X_1,X_2,A|Z_1,Z_2}\left(x_2, x_1, a\right| z_2, z_1\left),\right. \qquad (1.56)
\end{aligned}
$$

where the first equality holds by change of variables, the second equality uses (1.53) and $Z \perp (A, U)$, the latter of which enables one to switch the order of (z_1, z_2) in the conditioned set, the third equality holds again by change of variables and

$$
\begin{aligned}
X_1 &= g\left(z_2, a, g^{-1}\left(x_2, z_2, a\right)\right) = x_2 \\
X_2 &= g\left(z_1, a, g^{-1}\left(x_1, z_1, a\right)\right) = x_1, \qquad (1.57)
\end{aligned}
$$

and the last equality uses the fact that the product of derivatives of inverse functions is 1, i.e.,

$$
\begin{aligned}
& J_1 J_2 \\
&= \partial g^{-1}\left(X, Z, A\right)/\partial X\Big|_{(X,Z,A)=(x_1,z_1,a)} \times \partial g^{-1}\left(X, Z, A\right)/\partial X\Big|_{(X,Z,A)=(x_2,z_2,a)} \\
&\quad \times \partial g\left(Z, A, U\right)/\partial U|_{(Z,A,U)=(z_2,a,u_2)} \times \partial g\left(Z, A, U\right)/\partial U|_{(Z,A,U)=(z_1,a,u_1)} \\
&= \left[\partial g^{-1}\left(X, Z, A\right)/\partial X\Big|_{(X,Z,A)=(x_1,z_1,a)} \times \partial g\left(Z, A, U\right)/\partial U|_{(Z,A,U)=(z_1,a,u_1)}\right] \\
&\quad \times \left[\partial g^{-1}\left(X, Z, A\right)/\partial X\Big|_{(X,Z,A)=(x_2,z_2,a)} \times \partial g\left(Z, A, U\right)/\partial U|_{(Z,A,U)=(z_2,a,u_2)}\right] \\
&= 1 \times 1 = 1. \qquad (1.58)
\end{aligned}
$$

Given (1.56), we have

$$
\begin{aligned}
f_{X_1,X_2|Z_1,Z_2}\left(x_1, x_2\right| z_1, z_2\left)\right. &= \int f_{X_1,X_2,A|Z_1,Z_2}\left(x_1, x_2, a\right| z_1, z_2\left)\mu\left(da\right)\right. \\
&= \int f_{X_1,X_2,A|Z_1,Z_2}\left(x_2, x_1, a\right| z_2, z_1\left)\mu\left(da\right)\right.
\end{aligned}
$$

$$= f_{X_1,X_2|Z_1,Z_2}\left(x_2,x_1\right| z_2,z_1) . \tag{1.59}$$

which implies

$$\begin{aligned}
&f_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2) \\
&= f_{X_1,X_2,A|Z_1,Z_2}\left(x_1,x_2,a\right| z_1,z_2) / f_{X_1,X_2|Z_1,Z_2}\left(x_1,x_2\right| z_1,z_2) \\
&= f_{X_1,X_2,A|Z_1,Z_2}\left(x_2,x_1,a\right| z_2,z_1) / f_{X_1,X_2|Z_1,Z_2}\left(x_2,x_1\right| z_2,z_1) \\
&= f_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_2,x_1,z_2,z_1) ,
\end{aligned} \tag{1.60}$$

where the second equality holds by (1.56) and (1.59).

Next, we follow Altonji and Matzkin (2005) to show that the conditional density $f_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2)$ can be approximated arbitrarily closely by a function of the form $f_{A|W}\left(a\right| W)$, where W is a vector-valued function symmetric in the elements of X and Z. By Assumption 1.3, the supports of X and Z are compact. By Assumption 1.1–1.3, $f_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2)$ is continuous in (X_1,X_2,Z_1,Z_2). Therefore, from the Stone-Weierstrass Theorem one can find a function $f^w_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2)$ that is a polynomial in (X_1,X_2,Z_1,Z_2) over a compact set with the property that for any fixed δ that is arbitrarily close to 0,

$$\max_{x_t\in\mathcal{X},z_t\in\mathcal{Z},\forall t}\left| f_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2) - f^w_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2)\right| \leq \delta. \tag{1.61}$$

Let

$$\begin{aligned}
&\overline{f}_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2) \\
&:= \left[f_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_1,x_2,z_1,z_2) + f_{A|X_1,X_2,Z_1,Z_2}\left(a\right| x_2,x_1,z_2,z_1)\right]/2!
\end{aligned} \tag{1.62}$$

denote the simple averages of $f_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2)$ over all $T!$ (here $T=2$) unique permutations of (x_t, z_t), and similarly for $\overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2)$. By (1.60), we have

$$\overline{f}_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2) = f_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2). \tag{1.63}$$

Also note that by construction, we have

$$\overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2) = \overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\,x_2,x_1,z_2,z_1). \tag{1.64}$$

By (1.60) and T, it is true that

$$\begin{aligned} & \left|f_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2) - \overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2)\right| \\ = & \left|\overline{f}_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2) - \overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2)\right| \\ \leq & \; T! \times (\delta/T!) = \delta. \end{aligned} \tag{1.65}$$

Since f^w can be chosen to make δ arbitrarily small, (1.65) implies that $f_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2)$ can be approximated arbitrarily closely by a polynomial $\overline{f}^w$ that is symmetric in (x_t, z_t) for $t = 1, 2$. Thus, by the fundamental theorem of symmetric functions, $\overline{f}^w$ can be written as a polynomial function of the elementary symmetric functions of $((x_1, z_1), (x_2, z_2))$. We denote this function by W and obtain that $f_{A|X_1,X_2,Z_1,Z_2}(a|\,x_1,x_2,z_1,z_2)$ can be approximated arbitrarily closely by $f_{A|W}(a|\,W)$. Let $\delta \to 0$ in (1.61). Then, for any $t \in \{1, .., T\}$ and (X_t, Z_t, A, W) on its support we have

$$f_{A|X_t,Z_t,W}(a|\,x_t, z_t, w) = f_{A|W}(a|\,w). \tag{1.66}$$

To see why Assumption 1.1 only requires one coordinate of X_t to be strictly monotonic in U_t, suppose $X_t = (K_t, L_t)' = (g_K(Z_t, A, U_t), g_L(Z_t, A, U_t))'$ and only g_K is strictly monotonic in U_t. Then, to establish a similar result as (1.56), for $(k_1, l_1, k_2, l_2, z_1, z_2, a)$ on the support of $(K_1, L_1, K_2, L_2, Z_1, Z_2, A)$ we have

$$\begin{aligned}
& f_{K_1,L_1,K_2,L_2,A|Z_1,Z_2}\left(k_1, l_1, k_2, l_2, a \middle| z_1, z_2\right) \\
= {} & f_{U_1,L_1,U_2,L_2,A|Z_1,Z_2}\left(g_K^{-1}\left(k_1, z_1, a\right), l_1, g_K^{-1}\left(k_2, z_2, a\right), l_2, a \middle| z_1, z_2\right)\left|\widetilde{J}_1\right| \\
= {} & f_{A,U_1,U_2|Z_1,Z_2}\left(a, g_K^{-1}\left(k_1, z_1, a\right), g_K^{-1}\left(k_2, z_2, a\right) \middle| z_1, z_2\right)\left|\widetilde{J}_1\right| \\
= {} & f_{A,U_1,U_2|Z_1,Z_2}\left(a, g_K^{-1}\left(k_2, z_2, a\right), g_K^{-1}\left(k_1, z_1, a\right) \middle| z_2, z_1\right)\left|\widetilde{J}_1\right| \\
= {} & f_{U_1,L_1,U_2,L_2,A|Z_1,Z_2}\left(g_K^{-1}\left(k_2, z_2, a\right), l_2, g_K^{-1}\left(k_1, z_1, a\right), l_1, a \middle| z_2, z_1\right)\left|\widetilde{J}_1\right| \\
= {} & f_{K_1,L_1,K_2,L_2,A|Z_1,Z_2}\left(k_2, l_2, k_1, l_1, a \middle| z_2, z_1\right)\left|\widetilde{J}_2\right|\left|\widetilde{J}_1\right| \\
= {} & f_{K_1,L_1,K_2,L_2,A|Z_1,Z_2}\left(k_2, l_2, k_1, l_1, a \middle| z_2, z_1\right), \qquad (1.67)
\end{aligned}$$

where the first and second to last equality holds by change of variables, the second and fourth equality holds because L is a function of (Z, A, U), the third equality holds by (1.53) and the exogeneity of $Z \perp (A, U)$, and the last equality holds by $\left|\widetilde{J}_2\right|\left|\widetilde{J}_1\right| = 1$ which is derived similarly to (1.58). The rest of the proof follows similarly as in the scalar X case above. □

Proof of Lemma 1.2. Let $g^{-1}(x, z, a)$ denote the inverse function for $g(z, a, u)$ in its first argument, which exists by Assumption 1.1. Assume X_{it} is a scalar for brevity of exposition. For any (x, z, a, w) in the support of (X, Z, A, W), we have

$$\begin{aligned}
& F_{X_{it}|Z_{it},W_i}\left(x \middle| z, w\right) \\
= {} & F_{X_{it}|Z_{it},A_i,W_i}\left(x \middle| z, a, w\right) \\
= {} & \mathbb{P}\left(X_{it} \leq x \middle| Z_{it} = z, A_i = a, W_i = w\right) \\
= {} & \mathbb{P}\left(g\left(z, a, U_{it}\right) \leq x \middle| Z_{it} = z, A_i = a, W_i = w\right) \\
= {} & \mathbb{P}\left(U_{it} \leq g^{-1}\left(x, z, a\right) \middle| A_i = a, W_i = w\right) \\
= {} & F_{U_{it}|A_i,W_i}\left(g^{-1}\left(x, z, a\right) \middle| a, w\right), \qquad (1.68)
\end{aligned}$$

where the first equality holds by (1.16), the third uses (1.3), the fourth holds by Assumption 1.1 and 1.4, and the last equality holds by definition of the conditional CDF of U_{it} given (A_i, W_i).

By (1.3), $U_{it} = g^{-1}(X_{it}, Z_{it}, A_i)$, so that plugging in gives

$$V_{it} \coloneqq F_{X_{it}|Z_{it},W_i}\left(X_{it}|\, Z_{it}, W_i\right) = F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right). \tag{1.69}$$

□

1.B Proofs in Section 1.4

The proof of Lemma 1.3 follows directly from that of Theorem 12 in Imbens and Newey (2009). Thus, it is omitted for brevity. First, we prove Theorem 1.2. Note that by T, we obtain the mean squared and uniform convergence results if we can prove it for each coordinate of β. Therefore, wlog we assume β is a scalar throughout the proof. Then, we prove Theorem 1.3 and 1.4. The proof of Theorem 1.3 follows from Imbens and Newey (2002), Andrews (1991), and a Cramér–Wold device. The proof of Theorem 1.4 requires more efforts. As discussed before, for $\overline{\beta}$ one can obtain its normality by choosing the basis function $r^M(\cdot) \equiv 1$ and applying the results for $\beta(x)$.

Proof of Theorem 1.2. As discussed before, the convergence rate for $\widehat{\beta}(v, w)$ is the same as $\widehat{G}(s)$ because they share the same series regression coefficients $\widehat{\alpha}^K$. Under Assumption 1.7 and 1.8, the convergence rate result on $\widehat{G}(s)$ applies directly to $\widehat{\beta}(v, w)$ and the proof is thus omitted.

We focus on $\widehat{\beta}(x)$, since the result for $\widehat{\overline{\beta}}$ follows by setting $r^M(\cdot) \equiv 1$. Following Newey (1997), we normalize $\mathbb{E}r_i r_i' = I$ and have $\lambda_{\min}\left(\widehat{R}\right) \geq C > 0$. By (1.45), we have

$$\left\| \widehat{R}^{1/2}\left(\widehat{\eta}^M - \eta^M\right) \right\|^2$$

$$\begin{aligned}&\leq \left(\widehat{B}-\widetilde{B}\right)^{'} r\widehat{R}^{-1}r^{'}\left(\widehat{B}-\widetilde{B}\right)/n^2+\left(\widetilde{B}-B\right)^{'} r\widehat{R}^{-1}r^{'}\left(\widetilde{B}-B\right)/n^2\\&\quad+\left(B-B^X\right)^{'} r\widehat{R}^{-1}r^{'}\left(B-B^X\right)/n^2+\left(B^X-r\eta^M\right)^{'} r\widehat{R}^{-1}r^{'}\left(B^X-r\eta^M\right)/n^2. \qquad (1.70)\end{aligned}$$

Following the proof for Theorem 1 of Newey (1997), Lemma A1 and Lemma A3 of Imbens and Newey (2002), under Assumption 1.7 we have

$$\left\|n^{-1}\sum_i \widehat{\overline{p}}_i\widehat{\overline{p}}_i^{'}-\mathbb{E}\overline{p}_i\overline{p}_i^{'}\right\|=o_P(1)\ \text{ and } \mathbb{E}\overline{p}_i\overline{p}_i^{'}\leq CI. \qquad (1.71)$$

Then, we have

$$\begin{aligned}&\left(\widehat{B}-\widetilde{B}\right)^{'} r\widehat{R}^{-1}r^{'}\left(\widehat{B}-\widetilde{B}\right)/n^2\\&\leq C\left(\widehat{B}-\widetilde{B}\right)^{'}\left(\widehat{B}-\widetilde{B}\right)/n\\&=Cn^{-1}\sum_i\left(\widehat{\beta}\left(\widehat{V}_i,W_i\right)-\beta\left(\widehat{V}_i,W_i\right)\right)^2\\&=Cn^{-1}\sum_i\left(\widehat{\overline{p}}_i^{'}\left(\widehat{\alpha}^K-\alpha^K\right)+\left(\widehat{\overline{p}}_i^{'}\alpha^K-\beta\left(\widehat{V}_i,W_i\right)\right)\right)^2\\&\leq C\left\|\widehat{\alpha}^K-\alpha^K\right\|^2+\sup_{s\in\mathcal{S}}\left\|\overline{p}^K(s)^{'}\alpha^K-\beta(v,w)\right\|^2=O_P\left(\Delta_{2n}^2\right) \qquad (1.72)\end{aligned}$$

where the first inequality holds because $r\widehat{R}^{-1}r^{'}/n$ is idempotent, the last inequality holds by (1.71), and the last equality uses Lemma 1.3.

Next, we have

$$\begin{aligned}&\left(\widetilde{B}-B\right)^{'} r\widehat{R}^{-1}r^{'}\left(\widetilde{B}-B\right)/n^2\\&\leq Cn^{-1}\sum_i\left(\beta\left(\widehat{V}_i,W_i\right)-\beta\left(V_i,W_i\right)\right)^2\\&\leq Cn^{-1}\sum_i\left(\widehat{V}_i-V_i\right)^2=O_P\left(\Delta_{1n}^2\right), \qquad (1.73)\end{aligned}$$

where the last inequality holds by Assumption 1.8(4) and the equality holds by Lemma 1.3.

Finally, for the last two terms in (1.70), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(B-B^{X}\right)^{'} r\widehat{R}^{-1}r'\left(B-B^{X}\right)/n^{2}\middle|\mathbf{X}\right]\\
&=tr\left\{\mathbb{E}\left[\xi' r\widehat{R}^{-1}r'\xi\middle|\mathbf{X}\right]\right\}/n^{2}\\
&=tr\left\{\mathbb{E}\left[\xi\xi'\middle|\mathbf{X}\right]r\widehat{R}^{-1}r'\right\}/n^{2}\\
&\leq tr\left\{CIr\widehat{R}^{-1}r'\right\}/n^{2}=Ctr\left\{\widehat{R}^{-1}\widehat{R}\right\}/n=CM/n. \qquad (1.74)
\end{aligned}$$

and

$$\begin{aligned}
&\left(B^{X}-R\eta^{M}\right)^{'} r\widehat{R}^{-1}r'\left(B^{X}-R\eta^{M}\right)/n^{2}\\
&\leq\left(B^{X}-R\eta^{M}\right)^{'}\left(B^{X}-R\eta^{M}\right)/n=O_{P}\left(M^{-2d_{3}}\right). \qquad (1.75)
\end{aligned}$$

Collecting terms and using $\lambda_{\min}\left(\widehat{R}\right)\geq C$, we have

$$\left\|\widehat{\eta}^{M}-\eta^{M}\right\|^{2}=O_{P}\left(\Delta_{2n}^{2}+M/n+M^{-2d_{3}}\right)=:O_{P}\left(\Delta_{3n}^{2}\right), \qquad (1.76)$$

which implies

$$\begin{aligned}
&\int\left\|\widehat{\beta}\left(x\right)-\beta\left(x\right)\right\|^{2}dF\left(x\right)\\
&\leq\int\left(r^{M}\left(x\right)'\left(\widehat{\eta}^{M}-\eta^{M}\right)+\left(r^{M}\left(x\right)'\eta^{M}-\beta\left(x\right)\right)\right)^{2}dF\left(x\right)\\
&\leq C\left\|\widehat{\eta}^{M}-\eta^{M}\right\|^{2}+\sup_{x\in\mathcal{X}}\left|\beta\left(x\right)-r^{M}\left(x\right)'\eta^{M}\right|^{2}=O_{P}\left(\Delta_{3n}^{2}\right), \qquad (1.77)
\end{aligned}$$

and

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left\|\widehat{\beta}\left(x\right)-\beta\left(x\right)\right\|&\leq\sup_{x\in\mathcal{X}}\left\|r^{M}\left(x\right)\right\|\left\|\widehat{\eta}^{M}-\eta^{M}\right\|+\sup_{x\in\mathcal{X}}\left|\beta\left(x\right)-r^{M}\left(x\right)'\eta^{M}\right|\\
&=O_{P}\left(\zeta\left(M\right)\Delta_{3n}\right).
\end{aligned}$$

□

Proof of Theorem 1.3. Recall that the analysis of Imbens and Newey (2002) applies to scalar functionals of $G(s)$. By Cramér–Wold device and Imbens and Newey (2002), for any constant vector c with $c'c = 1$ we have

$$\begin{aligned} c'\sqrt{n}\Omega_1^{-1/2}\left(\widehat{\beta}(v,w)-\beta(v,w)\right) &\to_d N(0,1) \text{ and} \\ \left(c'\Omega_1 c\right)^{-1}\left[c'\left(\widehat{\Omega}_1-\Omega_1\right)c\right] &\xrightarrow{p} 0. \end{aligned} \tag{1.78}$$

By (1.78) and Assumption 1.9(6), it is true that

$$c'\left(b_{1n}\widehat{\Omega}_1 - b_{1n}\Omega_1\right)c \xrightarrow{p} 0, \tag{1.79}$$

which implies

$$b_{1n}\widehat{\Omega}_1 \xrightarrow{p} \overline{\Omega}_1. \tag{1.80}$$

Combining (1.78) – (1.80), we have

$$\begin{aligned} &\sqrt{n}\widehat{\Omega}_1^{-1/2}\left(\widehat{\beta}(v,w)-\beta(v,w)\right) \\ &= \left(b_{1n}\widehat{\Omega}_1\right)^{-1/2}\left(b_{1n}\Omega_1\right)^{1/2}\sqrt{n}\Omega_1^{-1/2}\left(\widehat{\beta}(v,w)-\beta(v,w)\right) \\ &\xrightarrow{d} \overline{\Omega}_1^{-1/2}\overline{\Omega}_1^{1/2}\mathcal{N}(0,I) = \mathcal{N}(0,I), \end{aligned} \tag{1.81}$$

where the convergence holds by (1.78), (1.80), and Assumption 1.9(6). □

Proof of Theorem 1.4. Following the proof of Theorem 1.3, one can extend the results to vector-valued functionals using Cramér–Wold device and the proofs of Andrews (1991). Therefore, wlog we assume $\beta(x)$ is a scalar in this proof. First, we derive the influence functions that correctly account for the effects from estimating $\beta(x)$ and prove asymptotic normality using Lindeberg–Feller CLT. Then, we show consistency for the estimator of the variance, which can be used to construct feasible confidence intervals. We write $r^M(x)$ as $r(x)$ and suppress t subscript when there is no confusion.

By Assumption 1.10(1), we normalize $\mathbb{E}r_i r_i' = I$ and obtain $\left\| \widehat{R} - I \right\| = o_P(1)$ using a similar argument as in the proof of Theorem 1 of Newey (1997). Recall that $\widehat{\beta}(x) = r^M(x)' \widehat{R}^{-1} r' \widehat{B}/n$. Let

$$\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) = r^M(x)' \widehat{R}^{-1} r' \widehat{B}/n, \text{ and } a(\beta, V) = \mathbb{E}\left[\beta_i | X = x\right] \tag{1.82}$$

and define

$$\begin{aligned} \Omega_{21} &= \mathbb{E}\left(A_1 P^{-1} p_i u_i\right)\left(A_1 P^{-1} p_i u_i\right)' \\ \Omega_{22} &= \mathbb{E}\left[\begin{array}{c} \left(A_1 P^{-1} \overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i\left(\beta\left(V_i, W_i\right) - \beta\left(X_i\right)\right)\right)\right) \\ \times\left(A_1 P^{-1} \overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i\left(\beta\left(V_i, W_i\right) - \beta\left(X_i\right)\right)\right)\right)' \end{array}\right]. \end{aligned} \tag{1.83}$$

Then, we have $\Omega_2 = \Omega_{21} + \Omega_{22}$.

Let $F = \Omega_2^{-1/2}$, which is well-defined because

$$\begin{aligned} \Omega_{21} &= A_1 P^{-1}\left(\mathbb{E} p_i p_i' u_i^2\right) P^{-1} A_1' \\ &= A_1 P^{-1}\left(\mathbb{E} p_i p_i' \mathbb{E}\left(u_i^2 \middle| X_i, V_i, W_i\right)\right) P^{-1} A_1' \\ &\geq C A_1 P^{-1} A_1' = C r(x)'\left(\mathbb{E} r_i \overline{p}_i'\right)\left(\mathbb{E} p_i p_i'\right)^{-1}\left(\mathbb{E} \overline{p}_i r_i'\right) r(x) > 0, \end{aligned} \tag{1.84}$$

where the first inequality holds by Assumption 1.9(3) and the last inequality holds by Assumption 1.10(1).

We expand

$$\begin{aligned} &\sqrt{n} F\left(\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) - a(\beta, V)\right) \\ = &\sqrt{n} F\left(\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) - \widehat{a}\left(\beta, \widehat{V}\right) + \widehat{a}\left(\beta, \widehat{V}\right) - \widehat{a}(\beta, V) + \widehat{a}(\beta, V) - a(\beta, V)\right) \\ = &n^{-1/2} \sum_i \left(\psi_{1i} + \psi_{2i} + \psi_{3i}\right) + o_P(1) \end{aligned} \tag{1.85}$$

and show that

$$\psi_{1i} = H_1\left(p_i u_i - \overline{\mu}_i^I\right), \ \psi_{2i} = H_2 \overline{\mu}_i^{II}, \text{ and } \psi_{3i} = H_2 r_i \xi_i. \tag{1.86}$$

First, for ψ_{1i} we have

$$\begin{aligned}
&\sqrt{n}F\left(\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) - \widehat{a}\left(\beta, \widehat{V}\right)\right) \\
=\ & \sqrt{n}Fr\left(x\right)' \widehat{R}^{-1} r'\left(\widehat{B} - \widetilde{B}\right)/n \\
=\ & \sqrt{n}Fr\left(x\right)' \widehat{R}^{-1} r'\left(\widehat{\overline{p}}\widehat{P}^{-1}\widehat{p}' Y/n - \widetilde{B}\right)/n \\
=\ & n^{-1/2}Fr\left(x\right)' \widehat{R}^{-1} r'\left[n^{-1}\widehat{\overline{p}}\widehat{P}^{-1}\widehat{p}'\left(Y - G + G - \widetilde{G} + \widetilde{G} - \widehat{p}\alpha^K\right) + \left(\widehat{\overline{p}}\alpha^K - \widetilde{B}\right)\right] \\
=\ & n^{-1/2}\sum_i \widehat{H}_1 \widehat{p}_i\left[u_i - \left(G\left(\widehat{s}_i\right) - G\left(s_i\right)\right)\right] + n^{-1/2}\widehat{H}_1\widehat{p}'\left(\widetilde{G} - \widehat{p}\alpha^K\right) \\
& + n^{-1/2}\widehat{H}_2 r'\left(\widehat{\overline{p}}\alpha^K - \widetilde{B}\right) =: D_{11} + D_{12} + D_{13}.
\end{aligned} \tag{1.87}$$

We show $D_{11} = n^{-1/2}\sum_i \psi_{1i} + o_P\left(1\right)$, $D_{12} = o_P\left(1\right)$, and $D_{13} = o_P\left(1\right)$.

The proof of

$$D_{11} = n^{-1/2}\sum_i \psi_{1i} + o_P\left(1\right) \tag{1.88}$$

is analogous to that of Lemma B7 and B8 of Imbens and Newey (2002), except that we need to establish $\left\|\widehat{H}_1 - H_1\right\| = o_P\left(1\right)$. To prove this claim, first we have

$$\left\|H_1\right\| = O\left(1\right) \text{ and } \left\|H_2\right\| = O\left(1\right), \tag{1.89}$$

because $\left\|H_1\right\|^2 \leq CA_1A_1'/\Omega_2 \leq C$ and $\left\|H_2\right\|^2 = A_2A_2'/\Omega_2 \leq CA_1A_1'/\Omega_2 \leq C$. In addition, we have $\left\|\widehat{P} - P\right\| = o_P\left(1\right)$, $\left\|\widehat{R} - I\right\| = o_P\left(1\right)$, and $\left\|n^{-1}\sum_i r_i\overline{p}_i - \mathbb{E}r_i\overline{p}_i'\right\| = o_P\left(1\right)$ as in the proof of Theorem 1 of Newey (1997). By Slutsky Theorem, $\left\|\widehat{R}^{-1} - I\right\| = o_P\left(1\right)$. Using CS and Lemma A3 of Imbens and Newey (2002), we have

$$\left\|n^{-1}\sum_i r_i\left(\overline{p}_i - \widehat{\overline{p}}_i\right)'\right\|^2 \leq n^{-1}\sum_i \left\|r_i\right\|^2 \times n^{-1}\sum_i \left\|\widehat{\overline{p}}_i - \overline{p}_i\right\|^2$$

$$= O_P\left(M\zeta_1(K)^2\Delta_n^2\right) = o_P(1). \tag{1.90}$$

Therefore, by T we have with probability approaching 1

$$\begin{aligned}
&\left\|\widehat{H}_1 - H_1\right\|^2 \\
=&\left\|F\widehat{A}_1\widehat{P}^{-1} - FA_1P^{-1}\right\|^2 \\
\leq& 2\left\|F\left(\widehat{A}_1 - A_1\right)\widehat{P}^{-1}\right\|^2 + 2\left\|FA_1\left(\widehat{P}^{-1} - P^{-1}\right)\right\|^2 \\
=& 2\left\|F\left(r(x)'(I + o_P(1))\left(\mathbb{E}r_i\overline{p}_i' + o_P(1)\right) - r(x)'\mathbb{E}r_i\overline{p}_i'\right)\widehat{P}^{-1}\right\|^2 \\
&+ 2\left\|FA_1P^{-1}\left(P - \widehat{P}\right)\widehat{P}^{-1}\right\|^2 \\
\leq& \|H_2\|^2 o_P(1) + \|H_1\|^2 o_P(1) = o_P(1).
\end{aligned} \tag{1.91}$$

and similarly $\left\|\widehat{H}_2 - H_2\right\| = o_P(1)$. The result follows as in the proof of Lemma B7 and B8 of Imbens and Newey (2002).

Next, recall that

$$\left(\widetilde{G} - \widehat{p}\alpha^K\right)'\left(\widetilde{G} - \widehat{p}\alpha^K\right)/n = O_P\left(K^{-2d_2}\right) \tag{1.92}$$

by Assumption 1.7(4). Therefore,

$$\begin{aligned}
\left|n^{-1/2}\widehat{H}_1\widehat{p}'\left(\widetilde{G} - \widehat{p}\alpha^K\right)\right|^2 &\leq n\left[\widehat{H}_1\widehat{P}\widehat{H}_1'\right]\left[\left(\widetilde{G} - \widehat{p}\alpha^K\right)'\left(\widetilde{G} - \widehat{p}\alpha^K\right)/n\right] \\
&\leq \left\|\widehat{H}_1\right\|^2 O_P\left(nK^{-2d_2}\right) = o_P(1).
\end{aligned} \tag{1.93}$$

For D_{13}, similarly to (1.93) we have

$$\begin{aligned}
\left|n^{-1/2}\widehat{H}_2 r'\left(\widehat{p}\alpha^K - \widetilde{B}\right)\right|^2 &\leq n\left[\widehat{H}_2\widehat{R}\widehat{H}_2\right]\left[\left(\widetilde{B} - \widehat{p}\alpha^K\right)'\left(\widetilde{B} - \widehat{p}\alpha^K\right)/n\right] \\
&= O_P\left(nK^{-2d}\right) = o_P(1).
\end{aligned} \tag{1.94}$$

Summarizing (1.88)–(1.94), we obtain

$$\psi_{1i} = H_1 \left(p_i u_i - \overline{\mu}_i^I \right). \tag{1.95}$$

To obtain ψ_{2i}, we have

$$\begin{aligned}
&\sqrt{n} F \left(\widehat{a} \left(\beta, \widehat{V} \right) - \widehat{a} \left(\beta, V \right) \right) \\
&= \sqrt{n} F r \left(x \right)^{'} \widehat{R}^{-1} r^{'} \left(\widetilde{B} - B \right) / n \\
&= \widehat{H}_2 n^{-1/2} \sum_i r_i \left(\widetilde{\beta}_i - \beta_i \right) \\
&= \widehat{H}_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\widehat{V}_i - V_i \right) + \widehat{H}_2 n^{-1/2} \sum_i r_i \beta_{vv} \left(\widetilde{V}_i, W_i \right) \left(\widehat{V}_i - V_i \right)^2 / 2 \\
&=: D_{21} + D_{22}.
\end{aligned} \tag{1.96}$$

We prove $D_{21} = n^{-1/2} \sum_i H_2 \overline{\mu}_i^{II} + o_P \left(1 \right)$ and $D_{22} = o_P \left(1 \right)$. For D_{21}, we have

$$\begin{aligned}
D_{21} &= \widehat{H}_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\widehat{V}_i - V_i \right) \\
&= H_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \Delta_i^I + \left(\widehat{H}_2 - H_2 \right) n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\widehat{V}_i - V_i \right) \\
&\quad + H_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\Delta_i^{II} + \Delta_i^{III} \right) \\
&=: D_{211} + D_{212} + D_{213},
\end{aligned} \tag{1.97}$$

where

$$\begin{aligned}
&\delta_{ij} = F \left(X_i | \, Z_j, W_j \right) - q_j^{'} \gamma^L \left(X_i \right), \ \Delta_i^I = q_i^{'} \widehat{Q}^{-1} \sum_j q_j v_{ij} / n, \\
&\Delta_i^{II} = q_i^{'} \widehat{Q}^{-1} \sum_j q_j \delta_{ij} / n, \text{ and } \Delta_i^{III} = -\delta_{ii}.
\end{aligned} \tag{1.98}$$

Following the proof of Lemma B7 of Imbens and Newey (2002), we obtain

$$D_{211} = n^{-1/2} \sum_i H_2 \overline{\mu}_i^{II} + o_P(1) . \tag{1.99}$$

For D_{212}, we have

$$\begin{aligned} \left| D_{212} \right|^2 &\leq Cn \left[\left(\widehat{H}_2 - H_2 \right) \hat{R} \left(\widehat{H}_2 - H_2 \right)' \right] \left[n^{-1} \sum_i \left(\widehat{V}_i - V_i \right)^2 \right] \\ &= O_P \left\{ n \left(\zeta(M)^2 M/n \right) \Delta_{1n}^2 \right\} = o_P(1) . \end{aligned} \tag{1.100}$$

For D_{213}, we have

$$\left| D_{213} \right|^2 \leq Cn \left[H_2 \widehat{R} H_2' \right] \left[\sum_i \left(\left(\Delta_i^{II} \right)^2 + \left(\Delta_i^{III} \right)^2 \right) /n \right] = O_P \left(nL^{1-2d_1} \right) = o_P(1) , \tag{1.101}$$

where the first equality is established in the proof of Theorem 4 of Imbens and Newey (2002).

Next, for D_{22}, we have

$$\begin{aligned} \left| D_{22} \right| &\leq C\sqrt{n} \left\| \widehat{H}_2 \right\| \sup_{x \in \mathcal{X}} \left\| r(x) \right\| \left| n^{-1} \sum_i \left(\widehat{V}_i - V_i \right)^2 \right| \\ &= O_P \left(\sqrt{n} \zeta(M) \Delta_n^2 \right) = o_P(1) . \end{aligned} \tag{1.102}$$

Combining the results for D_{21} and D_{22}, we obtain

$$\sqrt{n} F \left(\widehat{a} \left(\beta, \widehat{V} \right) - \widehat{a}(\beta, V) \right) = n^{-1/2} \sum_i H_2 \overline{\mu}_i^{II} + o_P(1) . \tag{1.103}$$

To obtain ψ_{3i}, first we expand

$$\begin{aligned} &\sqrt{n} F \left(\widehat{a}(\beta, V) - a(\beta, V) \right) \\ &= n^{-1/2} \sum_i \widehat{H}_2 r_i \beta_i - \sqrt{n} F \beta(x) \end{aligned}$$

$$
\begin{aligned}
&= n^{-1/2}\sum_i H_2 r_i \left(\beta\left(V_i, W_i\right) - \beta\left(X_i\right)\right) + n^{-1/2}\sum_i \left(\widehat{H}_2 - H_2\right) r_i \left(\beta\left(V_i, W_i\right) - \beta\left(X_i\right)\right) \\
&\quad + n^{-1/2}\sum_i \widehat{H}_2 r_i \left(\beta\left(X_i\right) - r_i'\eta^M\right) - \sqrt{n}F\left(\beta\left(x\right) - r\left(x\right)'\eta^M\right) \\
&=: D_{31} + D_{32} + D_{33} + D_{34}. \qquad (1.104)
\end{aligned}
$$

Recall that $D_{31} = n^{-1/2}\sum_i H_2 r_i \xi_i$ by definition of ξ_i. Thus, we show D_{32}, D_{33}, and D_{34} are all $o_P\left(1\right)$.

For D_{32}, we have

$$
\begin{aligned}
\mathbb{E}\left[\left|D_{32}\right|^2 \middle| \mathbf{X}\right] &= \left(\widehat{H}_2 - H_2\right) r'\mathbb{E}\left[\xi\xi'\middle|\mathbf{X}\right] r\left(\widehat{H}_2 - H_2\right)'/n \\
&\le C\left(\widehat{H}_2 - H_2\right)\widehat{R}\left(\widehat{H}_2 - H_2\right)' \\
&\le C\left\|\widehat{H}_2 - H_2\right\|^2\left(1 + \left\|\widehat{R} - I\right\|\right) \\
&= O_P\left\{\left\|\widehat{H}_2 - H_2\right\|^2\right\} = O_P\left(\zeta\left(M\right)^2 M/n\right) = o_P\left(1\right), \qquad (1.105)
\end{aligned}
$$

where the first inequality holds by Assumption 1.8(3) and the fact that $\widehat{H}_2$ and r are functions of X_i only, the second equality holds by $\left\|\widehat{R} - I\right\| = o_P\left(1\right)$, and the third equality follows similarly as in equation (A.1) and (A.6) of Newey (1997). Therefore, $D_{32} = o_P\left(1\right)$ by CM.

For D_{33}, by CS we have

$$
\begin{aligned}
\left|D_{33}\right|^2 &\le n\left(\widehat{H}_2\widehat{R}\widehat{H}_2'\right)\sum_i\left(\beta\left(X_i\right) - r_i'\eta^M\right)^2/n \\
&= O_P\left(nM^{-2d_3}\right) = o_P\left(1\right), \qquad (1.106)
\end{aligned}
$$

where the first equality holds by Assumption 1.8(1).

For D_{34}, we have

$$
\left|D_{34}\right|^2 = nF^2\left(\beta\left(x\right) - r\left(x\right)'\eta^M\right)^2 = O_P\left(nM^{-2d_3}\right) = o_P\left(1\right). \qquad (1.107)
$$

Summarizing (1.104)–(1.107), we obtain

$$\sqrt{n}F\left(\widehat{a}\left(\beta,V\right)-a\left(\beta,V\right)\right)=n^{-1/2}\sum_{i}H_{2}r_{i}\xi_{i}+o_{P}\left(1\right). \tag{1.108}$$

In sum, we have shown

$$\sqrt{n}F\left(\widehat{a}\left(\widehat{\beta},\widehat{V}\right)-a\left(\beta,V\right)\right)=n^{-1/2}\sum_{i}\left(\psi_{1i}+\psi_{2i}+\psi_{3i}\right)+o_{P}\left(1\right), \tag{1.109}$$

where

$$\psi_{1i}=H_{1}\left(p_{i}u_{i}-\overline{\mu}_{i}^{I}\right),\ \psi_{2i}=H_{2}\overline{\mu}_{i}^{II},\ \text{and}\ \psi_{3i}=H_{2}r_{i}\xi_{i} \tag{1.110}$$

and

$$H_{1}p_{i}u_{i}\perp\left(H_{1}\overline{\mu}_{i}^{I},H_{2}\overline{\mu}_{i}^{II},H_{2}r_{i}\xi_{i}\right) \tag{1.111}$$

because $\mathbb{E}\left(\left.u_{i}\right|X_{i},V_{i},W_{i}\right)=0$ by construction.

Let $\Psi_{in}=n^{-1/2}\left(\psi_{1i}+\psi_{2i}+\psi_{3i}\right)$. We have $\mathbb{E}\Psi_{in}=0$ and $Var\left(\Psi_{in}\right)=1/n$. For any $\varepsilon>0$, under Assumption 1.9 and 1.10, we have

$$\begin{aligned}
&n\mathbb{E}\left[\mathbb{1}\left\{\left|\Psi_{in}\right|>\varepsilon\right\}\Psi_{in}^{2}\right]\\
&\leq n\varepsilon^{2}\mathbb{E}\left[\mathbb{1}\left\{\left|\Psi_{in}\right|>\varepsilon\right\}\left(\Psi_{in}/\varepsilon\right)^{4}\right]\leq n\varepsilon^{-2}\mathbb{E}\Psi_{in}^{4}\\
&\leq C\mathbb{E}\left[\left(H_{1}p_{i}u_{i}\right)^{4}+\left(H_{1}\overline{\mu}_{i}^{I}\right)^{4}+\left(H_{2}\overline{\mu}_{i}^{II}\right)^{4}+\left(H_{2}r_{i}\xi_{i}\right)^{4}\right]/n\\
&\leq C\left(\zeta\left(K\right)^{2}K+\zeta\left(K\right)^{4}\zeta\left(L\right)^{4}L+\zeta\left(M\right)^{4}\zeta\left(L\right)^{4}L+\zeta\left(M\right)^{2}M\right)/n\rightarrow0,
\end{aligned} \tag{1.112}$$

where the last inequality follows a similar argument as in the proof of Lemma B5 of Imbens and Newey (2002). Then, by Lindeberg–Feller CLT we obtain

$$\sqrt{n}\Omega_{2}^{-1/2}\left(\widehat{a}\left(\widehat{\beta},\widehat{V}\right)-a\left(\beta,V\right)\right)\xrightarrow{d}N\left(0,1\right). \tag{1.113}$$

To construct a feasible confidence interval, one needs a consistent estimator of the covariance matrix. Thus, we show $\widehat{\Omega}_2/\Omega_2 - 1 \xrightarrow{p} 0$. Recall that

$$\Omega_2 = \mathbb{E}\left(A_1 P^{-1} p_i u_i\right)^2 + \mathbb{E}\left(A_1 P^{-1} \overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i \xi_i\right)\right)^2 = \Omega_{21} + \Omega_{22} \tag{1.114}$$

and

$$\widehat{\Omega}_2 = n^{-1} \sum_i \left(\widehat{A}_1 \widehat{P}^{-1} \widehat{p}_i \widehat{u}_i\right)^2 + n^{-1} \sum_i \left(\widehat{A}_1 \widehat{P}^{-1} \widehat{\mu}_i^I - \widehat{A}_2 \widehat{R}^{-1}\left(\widehat{\mu}_i^{II} + r_i \widehat{\xi}_i\right)\right)^2 =: \widehat{\Omega}_{21} + \widehat{\Omega}_{22}. \tag{1.115}$$

The proof of $\widehat{\Omega}_{21}/\Omega_2 - \Omega_{21}/\Omega_2 \xrightarrow{p} 0$ follows the proof of Lemma B10 of Imbens and Newey (2009), with the $\widehat{A}_1$ instead of A_1 appearing in the definition of $\widehat{H}_1$. Nonetheless, we have shown that $\left\|\widehat{H}_1 - H_1\right\| = o_P(1)$. Thus, the proof for $\widehat{\Omega}_{21}$ follows similarly and is omitted for brevity.

For $\widehat{\Omega}_{22}$, we first show

$$\begin{aligned}
&n^{-1} \sum_i \left(\widehat{H}_1 \widehat{\overline{\mu}}_i^I - H_1 \overline{\mu}_i^I\right)^2 = o_P(1) \\
&n^{-1} \sum_i \left(\widehat{H}_2 \widehat{\overline{\mu}}_i^{II} - H_2 \overline{\mu}_i^{II}\right)^2 = o_P(1) \\
&n^{-1} \sum_i \left(\widehat{H}_2 r_i \widehat{\xi}_i - H_2 r_i \xi_i\right)^2 = o_P(1).
\end{aligned} \tag{1.116}$$

The first two convergence results hold by following the argument of the proof of Lemma B9 in Imbens and Newey (2002). For the last one, we have

$$\begin{aligned}
&\widehat{H}_2 r_i \widehat{\xi}_i - H_2 r_i \xi_i \\
&= \widehat{H}_2 r_i \left(\widehat{\xi}_i - \xi_i\right) + \left(\widehat{H}_2 - H_2\right) r_i \xi_i \\
&= \widehat{H}_2 r_i \left(\widehat{\beta}\left(\widehat{V}_i, W_i\right) - \widehat{\beta}(X_i) - \beta(V_i, W_i) + \beta(X_i)\right) + \left(\widehat{H}_2 - H_2\right) r_i \xi_i \\
&= \widehat{H}_2 r_i \left(\widehat{\beta}\left(\widehat{V}_i, W_i\right) - \beta\left(\widehat{V}_i, W_i\right)\right) + \widehat{H}_2 r_i \left(\beta\left(\widehat{V}_i, W_i\right) - \beta(V_i, W_i)\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \widehat{H}_2 r_i \left(\beta\left(X_i\right) - \widehat{\beta}\left(X_i\right)\right) + \left(\widehat{H}_2 - H_2\right) r_i \xi_i \\
&=: D_{41i} + D_{42i} + D_{43i} + D_{44i}. \qquad (1.117)
\end{aligned}
$$

For D_{41}, we have

$$
\begin{aligned}
n^{-1}\sum_i D_{41i}^2 &\leq \left\|\widehat{H}_2\right\|^2 \sup_{x\in\mathcal{X}} \left\|r\left(x\right)\right\|^2 n^{-1}\sum_i \left(\widehat{\beta}\left(\widehat{V}_i, W_i\right) - \beta\left(\widehat{V}_i, W_i\right)\right)^2 \\
&\leq C\zeta\left(M\right)^2 n^{-1}\sum_i \left[\left(\widehat{\overline{p}}_i^{\prime}\left(\widehat{\alpha}^K - \alpha^K\right)\right)^2 + \left(\widehat{\overline{p}}_i^{\prime}\alpha^K - \beta\left(\widehat{v}_i, w_i\right)\right)^2\right] \\
&= O_P\left(\zeta\left(M\right)^2 \Delta_{2n}^2\right) = o_P\left(1\right), \qquad (1.118)
\end{aligned}
$$

where the second inequality holds by $\left\|\widehat{H}_2\right\| = O_P\left(1\right)$ and Assumption 1.10(1) and the first equality holds by (1.72).

For D_{42}, we have

$$
\begin{aligned}
n^{-1}\sum_i D_{42i}^2 &\leq \left\|\widehat{H}_2\right\|^2 \sup_{x\in\mathcal{X}} \left\|r\left(x\right)\right\|^2 n^{-1}\sum_i \left(\beta\left(\widehat{V}_i, W_i\right) - \beta\left(V_i, W_i\right)\right)^2 \\
&\leq C\zeta\left(M\right)^2 n^{-1}\sum_i \left(\widehat{V}_i - V_i\right)^2 = O_P\left(\zeta\left(M\right)^2 \Delta_{1n}^2\right) = o_P\left(1\right), \qquad (1.119)
\end{aligned}
$$

where the first equality holds by Lemma 1.3.

The proof of $n^{-1}\sum_i D_{43i}^2 = o_P\left(1\right)$ is completely analogous to (1.118) and is thus omitted.

For D_{44}, we have

$$
\begin{aligned}
\mathbb{E}\left[n^{-1}\sum_i D_{44i}^2 \middle| \mathbf{X}\right] &= \left(\widehat{H}_2 - H_2\right) n^{-1}\sum_i r_i r_i^{\prime} \mathbb{E}\left(\xi_i^2 \middle| X_i\right)\left(\widehat{H}_2 - H_2\right)^{\prime} \\
&\leq C\left(\widehat{H}_2 - H_2\right)\widehat{R}\left(\widehat{H}_2 - H_2\right)^{\prime} \\
&\leq C\left\|\widehat{H}_2 - H_2\right\|^2 = o_P\left(1\right), \qquad (1.120)
\end{aligned}
$$

where the first equality holds by $\widehat{H}_2$ and r_i are both functions of $\mathbf{X}$, the first inequality holds by Assumption 1.8(3), and the last inequality uses $\left\|\widehat{R} - I\right\| = o_P\left(1\right)$. Then, by CM, we have

$$n^{-1} \sum_i D_{44i}^2 = o_P\left(1\right). \tag{1.121}$$

Combining results for D_{41}–D_{44}, we have

$$n^{-1} \sum_i \left(\widehat{H}_2 r_i \widehat{\xi}_i - H_2 r_i \xi_i\right)^2 = o_P\left(1\right). \tag{1.122}$$

Therefore, we have proven (1.116), which implies

$$\begin{aligned} & n^{-1} \sum_i \left(\left(\widehat{H}_1 \widehat{\overline{\mu}}_i^I - \widehat{H}_2 \widehat{\overline{\mu}}_i^{II} - \widehat{H}_2 r_i \widehat{\xi}_i\right) - \left(H_1 \overline{\mu}_i^I - H_2 \overline{\mu}_i^{II} - H_2 r_i \xi_i\right)\right)^2 \\ \leq\ & C n^{-1} \sum_i \left(\widehat{H}_1 \widehat{\overline{\mu}}_i^I - H_1 \overline{\mu}_i^I\right)^2 + C n^{-1} \sum_i \left(\widehat{H}_2 \widehat{\overline{\mu}}_i^{II} - H_2 \overline{\mu}_i^{II}\right)^2 \\ & + C n^{-1} \sum_i \left(\widehat{H}_2 r_i \widehat{\xi}_i - H_2 r_i \xi_i\right)^2 = o_P\left(1\right). \end{aligned} \tag{1.123}$$

Since $\mathbb{E}\left(H_1 \overline{\mu}_i^I - H_2 \overline{\mu}_i^{II} - H_2 r_i \xi_i\right)^2 = \Omega_{22}/\Omega_2 \leq 1$, by M and Lemma B6 of Imbens and Newey (2002), we have

$$\left|\widehat{\Omega}_{22}/\Omega_2 - n^{-1} \sum_i \left(H_1 \overline{\mu}_i^I - H_2 \overline{\mu}_i^{II} - H_2 r_i \xi_i\right)^2\right| = o_P\left(1\right). \tag{1.124}$$

By LLN, we have

$$\left|n^{-1} \sum_i \left(H_1 \overline{\mu}_i^I - H_2 \overline{\mu}_i^{II} - H_2 r_i \xi_i\right)^2 - \Omega_{22}/\Omega_2\right| = o_P\left(1\right). \tag{1.125}$$

Therefore, by T, we obtain

$$\widehat{\Omega}_{22}/\Omega_2 - \Omega_{22}/\Omega_2 = o_P\left(1\right). \tag{1.126}$$

Combining results for $\widehat{\Omega}_{21}$ and $\widehat{\Omega}_{22}$, we have

$$\widehat{\Omega}_2/\Omega_2 - 1 \xrightarrow{p} 0. \tag{1.127}$$

$\square$

1.C Notation

A_i	: individual fixed effect
$A_1, \widehat{A}_1, A_2$	: $A_1 = r^M(x)' \, \mathbb{E} r_i \overline{p}_i', \widehat{A}_1 = r^M(x)' \, \widehat{R}^{-1} \left(n^{-1} \sum_i r_i \widehat{\overline{p}}_i' \right), A_2 = r^M(x)$
$B, \widetilde{B}, \widehat{B}, B^X$	: $(\beta_1, ..., \beta_n)', \left(\widetilde{\beta}_1, ..., \widetilde{\beta}_n\right)', \left(\widehat{\beta}_1, ..., \widehat{\beta}_n\right)', (\beta(X_1), ..., \beta(X_n))'$
d_X	: dimension of X_{it}
d_1	: series approx rate for $V(x, z, w)$
d_2	: series approx rate for $G(s)$
d_3	: series approx rate for $\beta(x)$
F	: $\Omega_2^{-1/2}$
$G(S), \widehat{G}(S)$	: $\mathbb{E}[Y \vert X, V, W], \ p^K(S)' \, \widehat{\alpha}^K$
$H_1, \widehat{H}_1, H_2, \widehat{H}_2$	: $H_1 = FA_1P^{-1}, \widehat{H}_1 = F\widehat{A}_1\widehat{P}^{-1}, H_2 = FA_2, \widehat{H}_2 = FA_2\widehat{R}^{-1}$
K	: degree of basis functions $p^K(\cdot)$ used to estimate G
K_1	: degree of $p^{K_1}(\cdot)$, a component of $p^K(\cdot)$ and $\overline{p}^K(\cdot)$
L	: degree of basis functions $q(\cdot)$ used to estimate V
M	: degree of basis functions $r(\cdot)$ used to estimate $\beta(x)$
$p^K(s)$	: $x \otimes p^{K_1}(v, w)$ for $s = (x, v, w)$, a $DK_1 \times 1$ vector
$\overline{p}^K(s)$	: $I_D \otimes p^{K_1}(v, w)$, a $DK_1 \times D$ matrix

$p^{K_1}(v,w)$: component basis function of (v,w)

$q_i, p_i, \widehat{p}_i, \overline{p}_i, \widehat{\overline{p}}_i, r_i : q^L(X_i, Z_i, W_i), p^K(s_i), p^K(\widehat{s}_i), \overline{p}^K(s_i), \overline{p}^K(\widehat{s}_i), r^M(X_i)$

$p, \overline{p}, \widehat{p}, \widehat{\overline{p}} : (p_1, ..., p_n)', (\overline{p}_1, ..., \overline{p}_n)', (\widehat{p}_1, ..., \widehat{p}_n)', \left(\widehat{\overline{p}}_1, ..., \widehat{\overline{p}}_n\right)'$

$q, r : (q_1, ..., q_n)', (r_1, ..., r_n)'$

$P, \widetilde{P}, \widehat{P} : \mathbb{E}p_i p_i', \ n^{-1}\sum p_i p_i', \ n^{-1}\sum \widehat{p}_i \widehat{p}_i'$

$Q, \widehat{Q} : \mathbb{E}q_i q_i', \ n^{-1}\sum q_i q_i'$

$R, \widehat{R} : \mathbb{E}r_i r_i', \ n^{-1}\sum r_i r_i'$

$s, S : (x, v, w), (X, V, W)$

U : random shock per period

V : $F_{X|Z,W}$ control function for U

W : sufficient statistic for A

X : regressors for Y, e.g. labor, capital

Y_{it}, y : outcome variable e.g. value-added output, $y = (Y_1, ..., Y_n)'$

Z : instruments for X, e.g. interest rate

$\mathcal{X}, \mathcal{Z}, \mathcal{W}, \mathcal{V}, \mathcal{S}$: the support of X, Z, W, V, S

s, x, z, w : realization of random variables

X_{it}, Z_{it} : random vectors

$\mathbf{X}_i, \mathbf{Z}_i$: random matrix $(X_{i1}, ..., X_{iT})', (Z_{i1}, ..., Z_{iT})'$

$\alpha^K, \widehat{\alpha}^K$: series approx coefficient for $G(s)$, $\widehat{P}^{-1}\widehat{p}' y/n$

β_{it} : random coefficients

$\overline{\beta} : \mathbb{E}\beta_{it}$

$\beta(x) : \mathbb{E}[\beta_{it} | X_{it} = x]$

$\beta(v,w) : \mathbb{E}[\beta_{it} | V_{it} = v, W_i = w]$

$\beta_v(v,w) : \partial\beta(v,w)/\partial v$

$$
\begin{aligned}
\beta_i, \widetilde{\beta}_i, \widehat{\beta}_i &: \beta\left(V_i, W_i\right), \beta\left(\widehat{V}_i, W_i\right), \widehat{\beta}\left(\widehat{V}_i, W_i\right) \\
\delta_{0t} &: \mathbb{E}\left[d_t\left(U_{2,it}\right)\right] \\
\gamma^L(\cdot) &: \text{series approx coefficient for } V(x, z, w) \\
\eta^M &: \text{series approx coefficient for } \beta(x) \\
\lambda &: \text{eigenvalue of a matrix} \\
\text{psd, pd} &: \text{positive semi-definite, positive definite} \\
\overline{\mu}_i^I, \overline{\mu}_i^{II} &: \mathbb{E}\left[G_v\left(S_j\right)\tau'\left(V_j\right)p_j q_j' q_i v_{ji}\middle| \mathcal{I}_i\right], \mathbb{E}\left[\beta_v\left(V_j, W_j\right) r_j q_j' q_i v_{ji}\middle| \mathcal{I}_i\right] \\
\Omega_1 &: \overline{p}^K(s)' P^{-1}\left(\Sigma + \Sigma_1\right) P^{-1}\overline{p}^K(s) \\
\Sigma, \Sigma_1 &: \mathbb{E}p_i p_i' u_i^2,\ \mathbb{E}\overline{\mu}_i^I \overline{\mu}_i^{I\prime} \\
u_i, \widehat{u}_i &: Y_i - G\left(S_i\right),\ Y_i - \widehat{G}\left(\widehat{S}_i\right) \\
v_{ji} &: \mathbb{1}\left\{x_i \leq x_j\right\} - F\left(x_j \middle| z_i, w_i\right) \\
\widehat{\Omega}_1 &: \overline{p}^K(s)' \widehat{P}^{-1}\left(\widehat{\Sigma} + \widehat{\Sigma}_1\right)\widehat{P}^{-1}\overline{p}^K(s) \\
\widehat{\Sigma}, \widehat{\Sigma}_1 &: n^{-1}\sum_i \widehat{p}_i \widehat{p}_i' \widehat{u}_i^2, n^{-1}\sum_i \widehat{\mu}_i^I \widehat{\mu}_i^{I\prime} \\
\widehat{\overline{\mu}}_i^I, \widehat{\overline{\mu}}_i^{II} &: n^{-1}\sum_j \widehat{G}_v\left(\widehat{S}_j\right)\widehat{p}_j q_j' \widehat{Q}^- q_i \widehat{v}_{ji},\ n^{-1}\sum_j \widehat{\beta}_v\left(\widehat{S}_j\right) r_j q_j' \widehat{Q}^- q_i \widehat{v}_{ji} \\
\widehat{v}_{ji} &: \mathbb{1}\left\{x_i \leq x_j\right\} - \widehat{F}\left(x_j \middle| z_i, w_i\right). \\
\Omega_{21} &: \mathbb{E}\left(A_1 P^{-1} p_i u_i\right)\left(A_1 P^{-1} p_i u_i\right)' \\
\Omega_{22} &: \mathbb{E}\left[\left(A_1 P^{-1}\overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i \xi_i\right)\right)\left(A_1 P^{-1}\overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i \xi_i\right)\right)'\right] \\
\xi_i, \widehat{\xi}_i &: \beta\left(V_i, W_i\right) - \beta\left(X_i\right),\ \widehat{\beta}\left(\widehat{V}_i, W_i\right) - \widehat{\beta}\left(X_i\right) \\
\Omega_2 &: \Omega_{21} + \Omega_{22} \\
\widehat{\Omega}_{21} &: \widehat{A}_1 \widehat{P}^{-1}\left(n^{-1}\sum_i \widehat{p}_i \widehat{p}_i' \widehat{u}_i^2\right)\widehat{P}^{-1}\widehat{A}_1' \\
\widehat{\Omega}_{22} &: n^{-1}\sum_i \left(\widehat{A}_1 \widehat{P}^{-1}\widehat{\overline{\mu}}_i^I - \widehat{A}_2\left(\widehat{\overline{\mu}}_i^{II} + r_i \widehat{\xi}_i\right)\right)\left(\widehat{A}_1 \widehat{P}^{-1}\widehat{\overline{\mu}}_i^I - \widehat{A}_2\left(\widehat{\overline{\mu}}_i^{II} + r_i \widehat{\xi}_i\right)\right)'
\end{aligned}
$$

In the proofs :

CM : Conditional Markov Inequality

CS : Cauchy–Schwarz Inequality

LLN : Law of Large Numbers

M : Markov Inequality

T : Triangle Inequality

Chapter 2

Robust Semiparametric Estimation in Panel Multinomial Choice Models[1]

2.1 Introduction

The prevalence of heterogeneity and its importance in economic research are now well recognized. As pointed out by Heckman (2001), one of the most important discoveries in microeconometrics is the pervasiveness of diversity in economic behavior, which in turn has profound theoretical and practical implications. Browning and Carro (2007) survey the treatment of heterogeneity in applied microeconometrics, and find that "there is usually much more heterogeneity than researchers allow for", arguing that it is important yet difficult to accommodate heterogeneity in satisfactory ways. Moreover, the increasing availability of vast digital databases in this so-called "Big Data Era" brings about new challenges as well as opportunities for the treatment and understanding of heterogeneity (Fan, Han, and Liu, 2014).

More concretely, in analyzing consumer choices, a topic of wide theoretical and practical interest in microeconometrics, there might be rich forms of unobserved heterogeneity in

[1] Joint with Wayne Gao.

consumer and product characteristics that influence choice behavior in significant yet complex ways. For example, it has long been recognized that brand loyalty is an important factor in determining choices of consumer products (Howard and Sheth, 1969), and research by Reichheld and Schefter (2000) along with their colleagues from Bain & Company, a leading management consulting firm, finds that brand loyalty is becoming even more important for online businesses. However, in modeling of consumer behavior it is very difficult (Luarn and Lin, 2003) to incorporate brand loyalty, a potentially complicated object that is clearly heterogeneous, hard to measure, and often unobserved in data. Besides brand loyalty, there may also be other forms of unobserved heterogeneity, such as subtle flavors and packaging designs, that may influence our choices of consumer products in everyday life. It is neither theoretically nor empirically clear whether all such complicated forms of unobserved heterogeneity can be fully captured by scalar-valued fixed effects in fully additive models, as often found in the literature.

Given these motivations, this paper proposes a simple and robust method for semiparametric identification and estimation in a panel multinomial choice model, where we allow for infinite-dimensional (functional) fixed effects that enter into consumer utilities in an additively nonseparable and thus fully flexible way, incorporating rich forms of unobserved heterogeneity. Our identification strategy exploits multivariate monotonicity in its contrapositive form, which provides powerful leverage for converting observable events into identifying restrictions under lack of additive separability. We provide consistent estimators based on our identification strategy, together with a computational algorithm implemented in a spherical-coordinate reparameterization that brings about a combination of topological, geometric and arithmetic advantages. A simulation study and an empirical illustration using the Nielsen data on popcorn sales are conducted to analyze the finite-sample performance of our estimation method and demonstrate the adequacy of our computational procedure for practical implementation.

We consider the following panel multinomial choice model in a short-panel setting:

$$y_{ijt} = \mathbb{1}\left\{u\left(X_{ijt}^{'}\beta_0,\, A_{ij},\, \epsilon_{ijt}\right) \geq \max_{k\in\{1,\ldots,J\}} u\left(X_{ikt}^{'}\beta_0,\, A_{ik},\, \epsilon_{ikt}\right)\right\}$$

where agent i's utility from a candidate product j at time t, represented by $u\left(X_{ijt}^{'}\beta_0, A_{ij}, \epsilon_{ijt}\right)$, is taken to be a function of three components. The first is a linear index $X_{ijt}^{'}\beta_0$ of observable characteristics X_{ijt}, which contains a finite-dimensional parameter of interest β_0 we will identify and estimate. The second term A_{ij} is an infinite-dimensional fixed effect matrix that can be heterogeneous across each agent-product combination. The last term ϵ_{ijt} is an idiosyncratic time-varying error term of arbitrary dimensions. The three components are then aggregated by an unknown utility function u in an additively nonseparable way, with the only restriction being that each agent's utility $u\left(X_{ijt}^{'}\beta_0, A_{ij}, \epsilon_{ijt}\right)$ is *increasing* in its first argument, i.e., the linear index of observable characteristics $X_{ijt}^{'}\beta_0$. Each agent then chooses a certain product in a given time period, represented by $y_{ijt} = 1$, if and only if this product gives him the highest utility among all available products.

The infinite-dimensionality of the terms u, A_{ij} and ϵ_{ij} and the additive nonseparability in their interactions jointly produce rich forms of unobserved heterogeneity. Across each agent-product combination ij, we are effectively allowing for flexible variations in agent utilities as functions of the index $X_{ijt}^{'}\beta_0$, which serve as nonparametric proxies for the effects of complicated unobserved factors that influence choice behavior, including brand loyalty, subtle flavors and packaging designs as discussed earlier. Moreover, unrestricted heterogeneity in the distribution of the error term ϵ_{ijt} is accommodated, allowing for in particular heteroskedasticity in agent random utilities .

The generality of our setup encompasses many semiparametric (or parametric) panel multinomial choice models with scalar-valued fixed effects, scalar-valued error terms and various degrees of additive separability in the previous literature, including the following standard formulation:

$$y_{ijt} = \mathbb{1}\left\{X_{ijt}^{'}\beta_0 + A_{ij} + \epsilon_{ijt} \geq \max_{k\in\{1,\ldots,J\}}\left(X_{ikt}^{'}\beta_0 + A_{ik} + \epsilon_{ikt}\right)\right\}.$$

Relatively speaking, in this paper we are able to accommodate the infinite dimensionality of unobserved heterogeneity and the lack of additive separability in agent utility functions, under a standard time homogeneity assumption on the idiosyncratic error term that is widely adopted in the related literature.

Our key identification strategy exploits the standard notion of multivariate monotonicity in its contrapositive form. The idea is very simple and intuitive, and can be loosely described as the following: whenever we observe a *strict increase* in the choice probabilities of a specific product from one period to another, by logical contraposition it *cannot* be possible that this product becomes *worse* while all other products become *better* over the two periods. More formally, we show that a certain configuration of conditional choice probabilities satisfies the standard notion of weak multivariate monotonicity in all product indexes, which is naturally induced by the multinomial nature of our model and the monotonicity of each agent's utility function in each product's index. Then, we construct a collection of observable inequalities on conditional choice probabilities based on intertemporal comparison and cross-sectional aggregation, which preserves weak monotonicity in the index structure. Finally, we simply take a logical contraposition of the inequality on conditional choice probabilities, and obtain an identifying restriction on the index values free of all infinite-dimensional nuisance parameters, with which we construct a population criterion function that is guaranteed to be minimized at the true parameter value. The validity of this idea relies only on monotonicity in an index structure, and therefore it may have wider applicability beyond multinomial choice models.

Based on our identification result, we provide consistent set (or point) estimators, together with a computational algorithm adapted to the technical niceties and challenges of our framework. Specifically, our estimator can be computed through a two-stage procedure. The first stage takes the form of a standard nonparametric regression, where we nonparametrically estimate a collection of intertemporal differences in conditional choice probabilities, using a machine learning algorithm based on artificial neural networks. In the second stage, we numerically minimize our sample criterion function, constructed as the sample analog of our population criterion function with the first-stage nonparametric estimates plugged in. A highlight of our estimation and computation procedure is the adoption of a spherical-coordinate reparameterization of our criterion functions in terms of *angles*, which enables us to exploit a combination of topological, geometric and computational advantages.

A simulation study is conducted to analyze the finite-sample performance of our method and the adequacy of our computational procedure for practical implementation. We investigate the performances of the first-stage and the final estimators under different model configurations, and show how the results vary with the sizes and dimensions of data. We also compare the performances of our estimator under set identification and point identification, and demonstrate the informativeness of our set estimator under the lack of point identification.

An empirical illustration of our procedure is also provided, where we use the Nielsen data [2] on popcorn sales in the United States to explore the effects of marketing promotion effects. The results show that our procedure produces estimates that conform well with economic intuition. For example, we find that special in-store displays boost sales not only through a direct promotion effect but also through the attenuation of consumer price sensitivity, a result that cannot be produced by other methods based on additive separability. Intuitively, marketing managers are more likely to promote products that they know consumers are more

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

price and promotion sensitive to. Hence, the average effective price sensitivity of promoted products tend to be larger than those not promoted due to the selection effect. Given the nonadditive nature of such selection effects, estimators based on additive separability will be biased. In contrast, our method is robust to such confounding effects, thus producing more economically sensible estimates.

As a further generalization, we discuss the wider applicability of our identification strategy beyond panel multinomial choice models, using an umbrella framework called *monotone multi-index* models. This framework captures the key ingredients of a large class of models, such as sample selection models and network formation models. In particular, we provide a specific illustration of a dyadic network formation model under the setting of nontransferable utility, which naturally induces lack of additive separability in a micro-founded manner. The applicability of our current method, though with some nontrivial adaptions to the additional complications in network settings, is investigated in a companion paper by Gao, Li, and Xu (2020).

This paper builds upon and contributes to a large literature in econometrics on semiparametric (and parametric) discrete choice models, dating back to McFadden (1974a) and Manski (1975), and more specifically a recent branch of research that focuses on panel multinomial choice models.

Our work is most closely related to the work by Pakes and Porter (2016), who also exploit weak monotonicity and time homogeneity. Our current paper adopts a similar approach that heavily exploits monotonicity, but does not restrict the effect of unobserved heterogeneity as a scalar index that is additively separable from the scalar index of observable characteristics. Hence, it is no longer feasible in our model to directly calculate the differences between the indexes of observable characteristics as in Pakes and Porter (2016).

Another related paper is Shi, Shum, and Song (2018), who propose a novel approach that exploits cyclical monotonicity of *vector*-valued functions in a fully additive panel multinomial choice model, where scalar-valued fixed effects are differenced out through "cyclical summation". Khan, Ouyang, and Tamer (2019) consider a similar additive multinomial choice model, but utilize the subsample of observations with time-invariant covariates along *all products but one* so as to leverage monotonicity in a single linear index for the construction of a rank-based estimator a la Manski (1987). Relatedly, the earlier work by Honoré and Kyriazidou (2000) also exploits monotonicity in a single index when certain covariates across two periods are equal in a dynamic panel setting. Another recent paper by Chernozhukov, Fernández-Val, and Newey (2019) studies a nonseparable multinomial choice model with bounded derivatives, and demonstrates semiparametric identification in a specialized panel setting with an additive effect under an "on-the-diagonal" restriction (i.e., when covariates at two different time periods coincide). Our method is significantly different from and thus complementary to those proposed in these afore-cited papers.

At a more general level, our work can be related to and compared to semiparametric methods of identification and estimation in *monotone single-index models*. A related class of estimators that leverage univariate monotonicity, known as *maximum score* or *rank-order estimator*s, date back to a series of important contributions by Manski (1975, 1985, 1987), and are further investigated in Han (1987), Horowitz (1992), Abrevaya (2000), Honoré and Lewbel (2002) and Fox (2007). Despite the similarity in the reliance on monotonicity, the multinomial or *multi-index* nature of our current model induces a key difference from the single-index setting, leading to a significantly different method of estimation relative to rank-order estimators.

Finally, our model and method are complementary to another class of models that fall into the framework of *invertible multi-index models*. The celebrated paper by Berry, Levinsohn, and Pakes (1995) first utilizes the invertibility of the market share function to obtain a vector of unknown indexes, which is investigated more generally by Berry, Gandhi, and Haile (2013) and Berry and Haile (2014). Outside the context of demand estimation, a recent paper by Ahn, Ichimura, Powell, and Ruud (2018) provides a high-level treatment of multi-index models based on invertibility. In comparison, our paper does not involve invertibility, but relies on monotonicity.

The rest of this paper is organized as follows. Section 2.2 introduces our main model specifications and assumptions. Section 2.3 presents our key identification strategy. In Section 2.4 we provide consistent estimators along with a computational procedure to implement it. Section 2.5 and Section 3.5 contain a simulation study and an empirical illustration with the Nielsen data. Section 2.7 discusses the generalization of our method to monotone multi-index models, and finally we conclude with Section 2.8.

2.2 Panel Multinomial Choice Model

2.2.1 Model Setup

In this section we present a semiparametric panel multinomial choice model featured by infinite-dimensional unobserved heterogeneity and flexible forms of nonseparability, which we will use as the main model to illustrate our identification and estimation method. See Section 2.7 for a more general discussion about the wide applicability of our proposed methods.

Specifically, we consider the following discrete choice model, which states that agent i chooses product j at time t if and only if i prefers product j to all other alternatives at time t:

$$y_{ijt} = \mathbb{1}\left\{u\left(X_{ijt}'\beta_0,\, A_{ij},\, \epsilon_{ijt}\right) \geq \max_{k\in\{0,1,\ldots,J\}} u\left(X_{ikt}'\beta_0,\, A_{ik},\, \epsilon_{ikt}\right)\right\} \tag{2.1}$$

where:

- $i \in \{1, ...N\}$ denotes N decision makers, or simply *agents*.
- $j \in \{0, 1..., J\}$ denotes $J + 1$ choice alternatives, with J *products* indexed by $1, ..., J$ and an *outside option* denoted by 0.

- $t \in \{1, ..., T\}$ denotes $T \geq 2$ different time periods.
- X_{ijt} is $\mathbb{R}^D$-valued vector of observable characteristics specific to each agent-product-time tuple ijt. This could include, for example, buyer characteristics such as income level, product characteristics such as price and promotion status, as well as interaction and higher-order terms of those characteristics.
- y_{ijt} is an observable binary variable, with $y_{ijt} = 1$ indicating that buyer i chooses products j at time t and $y_{ijt} = 0$ indicating otherwise.
- $\beta_0 \in \mathbb{R}^D$ is a finite-dimensional unknown parameter of interest. We will repeatedly refer to the term $\delta_{ijt} := X_{ijt}^{'}\beta_0$ as the (ijt-specific) *index* throughout this paper, which is intended to capture how the observable characteristics X_{ijt} influence agent i's choice of j at t, *ceteris paribus*. Further discussion on the index is offered later.
- A_{ij} represents an ij-specific time-invariant unobserved heterogeneity term of arbitrary dimensions, which we will refer to as the (ij-specific) *fixed effect*.
- ϵ_{ijt} is an ijt-specific unobserved error term of arbitrary dimensions, which captures time-idiosyncratic utility shocks to product j for agent i at time t.
- u is an unknown function, interpreted as a *utility function* that aggregates the parametric index $X_{ijt}^{'}\beta_0$, the fixed effect A_{ij} and the error term ϵ_{ijt} into a scalar representing agent i's utility from choosing product j at time t.

We now provide some further clarifications and explanations for model (2.1).

We begin with a brief comparison that highlights the differences between our current model (2.1) to other models studied in several closely related papers on panel multinomial choice models. Notice first that model (2.1) includes as a special case the standard panel multinomial choice model under full additivity and scalar-valued unobserved heterogeneity:

$$y_{ijt} = \mathbb{1}\left\{X_{ijt}^{'}\beta_0 + A_{ij} + \epsilon_{ijt} \geq \max_{k \in \{1,...,J\}} X_{ikt}^{'}\beta_0 + A_{ik} + \epsilon_{ikt}\right\}. \tag{2.2}$$

Such models have been studied in recent work by Khan, Ouyang, and Tamer (2019) and Shi, Shum, and Song (2018) with different methods of identification and estimation. In another recent paper by Pakes and Porter (2016), they investigate a generalized version of (2.2) in the following form:

$$y_{ijt} = \mathbb{1}\left\{g_j\left(X_{ijt}, \beta_0\right) + f_j\left(A_{ij}, \epsilon_{ijt}\right) \geq \max_{k \in \{1, \ldots, J\}} g_k\left(X_{ikt}, \beta_0\right) + f_k\left(A_{ik}, \epsilon_{ikt}\right)\right\}, \tag{2.3}$$

where the function g_j produces a potentially nonlinear parametric index and f_j aggregates fixed effects and idiosyncratic errors into a scalar value in a nonseparable way, while additive separability between the observable covariate index $g_j\left(X_{ijt}, \beta_0\right)$ and the unobserved heterogeneity index $f_j\left(A_{ij}, \epsilon_{ijt}\right)$ is still maintained. Moreover, although the dimensions of $\left(A_{ij}, \epsilon_{ijt}\right)$ are not restricted in Pakes and Porter (2016), their overall effect is taken to be represented by a scalar value, $f_j\left(A_{ij}, \epsilon_{ijt}\right)$. We reiterate that our model (2.1) not only incorporates infinite-dimensionality in unobserved heterogeneity as captured by A_{ij} and ϵ_{ijt}, but also allows such heterogeneity to enter into agent utility functions in a fully *nonseparable* way.

The combination of infinite dimensionality and nonseparability jointly produces rich forms of heterogeneity in agent utility functions. Particularly, nonseparability translates into unrestricted flexibility regarding the ways in which the nonparametric fixed effect A_{ij} may enter into the utility function $u\left(X'_{ijt}\beta_0, A_{ij}, \epsilon_{ijt}\right)$. In fact, we could equivalently suppress the notation A_{ij} and instead write the utility function u to be ij-specific,[3] i.e., $u_{ij}\left(X'_{ijt}\beta_0, \epsilon_{ijt}\right) \equiv u\left(X'_{ijt}\beta_0, A_{ij}, \epsilon_{ijt}\right)$. Written in this form, our formulation allows for flexible time-invariant heterogeneity in how the index $X'_{ijt}\beta_0$ affects agent i's utility from product j. In other words, given a fixed value of the index $\overline{\delta}$, the utility $u_{ij}\left(\overline{\delta}, \epsilon_{ijt}\right)$ can vary across each agent-product pair in totally unrestricted ways. Such heterogeneity can

[3] This reformulation, however, will introduce randomness to the utility function u_{ij} when we consider the sampling process and assume cross-sectional random sampling later. Hence, to fully separate random elements from nonrandom ones, and to explicitly emphasize the dependence on A_{ij}, we will retain the notations of model (2.1) unless explicitly stated otherwise.

be induced by a plethora of complicated factors, such as subtle flavors, styles of design and social perceptions, the effects of which may be highly subjective on an individual basis. Some people may have a strong preference for Coca Cola over Pepsi or vice versa, while there might not exist any objective measure of flavor to assess, or even to describe, the subtle differences between the two popular soft drinks. Car shoppers may have heterogeneous tastes over engineering and design features in terms of safety, reliability, comfort, sportiness or luxury, while leading car manufacturers are often famous for their unique blends of features along these various dimensions, therefore appealing to different groups of customers to different extents. Beyond these examples, our formulation nests in itself arbitrary dimensions of agent-product specific heterogeneity that are time invariant.

It should be pointed out in particular that the fixed effect A_{ij} effectively incorporates unobserved variations in the distributions of error terms ϵ_{ijt}. For example, if we assume that ϵ_{ijt} is real-valued and follows a time-invariant distribution with a cumulative distribution function (CDF) F_{ij}, then the whole function F_{ij} can be readily incorporated as part of the fixed effect A_{ij}, which may lie in a vector of infinite-dimensional functions. The CDF F_{ij} absorbs a form of *heteroskedasticity* specific to each agent-product pair, and our method will be robust against such forms of heterogeneity in error distributions without the need to explicitly specify F_{ij}.

On a technical note, we now briefly discuss how the potential concern of tie-breaking can be handled in our framework. In cases where ties occur with nonzero probabilities, one popular approach in the literature is to incorporate a random tie-breaking process, modeled as a (potentially unknown) selection probability distribution among ties. The conceptual idea underlying this approach is to recognize the incompleteness of the model with respect to the determination of choice behaviors, and use an ad hoc selection probability to capture the effects of all unmodeled randomness. When we move from the scalar additive model (2.2) to model (2.1), rich forms of unmodeled randomness under (2.2) are automatically absorbed into the infinite-dimensional error term ϵ_{ijt}, which nests in itself all possible latent variables that affect utilities in some appropriate yet unspecified ways.[4] As a result, the assumption that ties occur with zero probabilities is effectively a much weaker restriction under our current model (2.1) than under model (2.2).

The flexibility induced by nonseparability and infinite-dimensionality comes with the consequent analytical challenges to handle them. Various traditional techniques in the style of *differencing* based on additivity no longer work in our current model. For example, the recent method based on cyclical monotonicity proposed by Shi, Shum, and Song (2018) requires additivity to sum along a cycle of comparisons and cancel out the scalar-valued fixed effects via this summation, which becomes infeasible under nonseparability in our model (2.1). To confront the challenges induced by such nonseparability, we instead exploit a standard shape restriction, or more specifically, *monotonicity*, which captures a general commonality shared by many additive models but on its own does not involve additivity at all.

2.2.2 Key Assumptions

We now continue with a list of key assumptions required for our subsequent analysis, and discuss these assumptions in relation to model (2.1). To economize on notation, we will from now on frequently refer to the collection of variables concatenated along product and time dimensions: $\mathbf{X}_{it} := (X_{ijt})_{j=1}^{J}$, $\mathbf{X}_i = (\mathbf{X}_{it})_{t=1}^{T}$, $\mathbf{A}_i := (A_{ij})_{j=1}^{J}$, $\boldsymbol{\epsilon}_{it} = (\epsilon_{ijt})_{j=1}^{J}$ and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{it})_{t=1}^{T}$. The first assumption below imposes a monotonicity restriction on the utility function.

Assumption 2.1 (Monotonicity in the Index). *$u(\delta_{ijt}, A_{ij}, \epsilon_{ijt})$ is weakly increasing in the index δ_{ijt}, for every realization of (A_{ij}, ϵ_{ijt}).*

[4] It should be pointed out that the standard ad hoc approach, using selection probabilities among ties, and our current approach, where latent variables are explicitly modeled by the infinite-dimensional error ϵ_{ijt}, are two distinct approaches, neither of which includes the other as a special case. The key distinction comes from the *lexicographic* nature of the selection-probability approach, which cannot be fully represented by utility functions. It might be debatable whether the lexicographic structure is more conceptually justifiable or practically relevant, but we refrain from further discussion on this topic, as it is tangential to the main focus of this paper.

It should first be clarified that the substantive part of Assumption 2.1 is the restriction of monotonicity in the index, while increasingness is without loss of generality given that the index $\delta_{ijt} = X_{ijt}'\beta_0$ contains an unknown parameter with unrestricted signs. Moreover, the monotonicity restriction is imposed on the index δ_{ijt}, but not directly on any specific observable characteristics in X_{ijt}: quadratic or higher-order polynomial terms as well as other nonlinear or non-monotone functions of observable characteristics may be included in X_{ijt} whenever appropriate.

Assumption 2.1 not only serves as a key restriction that will be heavily leveraged upon by our subsequent identification and estimation method, but may also be regarded as an integral part of our semiparametric model: monotonicity endows the index δ_{ijt} with an interpretation as an objective summary statistic for the direct effect of observable covariates on agent utilities. In other words, δ_{ijt} may be considered as a quality measure of the match between agent i and product j based on their observable characteristics at time t, inducing a consequent interpretation of the parameter β_0 as representing how a certain change in a linear combination of observable characteristics may increase utilities for *all* agents from a certain product j, *ceteris paribus*.

Given the parametric index structure $\delta_{ijt} = X_{ijt}'\beta_0$, monotonicity itself seems a rather weak assumption widely satisfied in a large class of models. In many additive models where a parametric index in the style of $X_{ijt}'\beta_0$ is added to other components of the model, Assumption 2.1 could be trivially satisfied by construction, such as the standard panel multinomial choice model (2.2). In Section 2.7, we provide more examples of parametric and semiparametric models featured by monotonicity in an index structure beyond the multinomial choice setting.

Assumption 2.2 (Cross-Sectional Random Sampling). $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{A}_i, \boldsymbol{\epsilon}_i)$ *is i.i.d. across* $i \in \{1, ..., N\}$ *with* $N \to \infty$.

Assumption 2.2 is a standard assumption on random sampling.[5] In particular, we only require a *short panel*, where we focus on cross-sectional asymptotics with the number of agents getting large ($N \to \infty$) but the number of time periods T held fixed.

Assumption 2.3 (**Conditional Time Homogeneity of Errors**). *The conditional distribution of* $\boldsymbol{\epsilon}_{it}$ *given* $(\mathbf{X}_i, \mathbf{A}_i)$ *is stationary over time* t*, i.e.,* $\boldsymbol{\epsilon}_{it} | (\mathbf{X}_i, \mathbf{A}_i) \sim \mathbb{P}(\cdot | \mathbf{A}_i)$.

Finally, we impose a conditional time homogeneity assumption on the idiosyncratic shocks. Assumption 2.3 is strictly stronger than necessary for our purpose, but leads to easier notations afterwards for clearer illustration of our key method. Alternatively, we could impose the following weaker version:

Assumption 2.3' (**Pairwise Time Homogeneity of Errors**). *The marginal distributions of* $\boldsymbol{\epsilon}_{it}$ *and* $\boldsymbol{\epsilon}_{is}$ *conditional on* $(\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i)$ *are the same across any pair of periods* $t \neq s \in \{1, ..., T\}$*, i.e.,* $\boldsymbol{\epsilon}_{it} | (\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i) \sim \boldsymbol{\epsilon}_{is} | (\mathbf{X}_{it}, \mathbf{X}_{is}, \mathbf{A}_i)$.

Assumption 2.3', a multinomial extension of the group homogeneity assumption in Manski (1987), is also imposed in Pakes and Porter (2016) and Shi, Shum, and Song (2018), both containing further discussions about the interpretation, flexibility and restrictions associated with this assumption. Assumption 2.3' suffices for our subsequent analysis based on pairwise intertemporal comparisons, while allowing for some dependence of $\boldsymbol{\epsilon}_{it}$ on time-varying component of observable covariates $(\mathbf{X}_{it}, \mathbf{X}_{is})$. We demonstrate in Appendix 2.B that our identification and estimation results carry over under Assumption 2.3', but until then we will work with the stronger Assumption 2.3 for notational simplicity.

It might be worth noting that Assumption 2.3 (or 2.3'), a statement conditioned on the arbitrarily dimensional fixed effect $\mathbf{A}_i$ in a fully flexible manner, automatically absorbs all possible *time-invariant* components in $\mathbf{X}_{it} = (X_{ijt})_{j=1}^{J}$ and $\boldsymbol{\epsilon}_{it} = (\epsilon_{ijt})_{j=1}^{J}$. As discussed earlier, long-term brand loyalty, potentially produced by a mixture of complicated factors

[5] It is worth noting that so far we have not made any explicit restriction on the structure of the spaces on which the arbitrary dimensional random elements $\mathbf{A}_i$ and $\boldsymbol{\epsilon}_i$ are defined, but implicit in our specification as well as Assumption 2.2 is the requirement that $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{A}_i, \epsilon_i)$ be well-defined as random elements (measurable functions) on a large enough probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

such as design, style, flavor, consumer personality or social perception, is just one example that applied researchers have found to be important since long ago (Howard and Sheth, 1969) yet conceptually difficult to incorporate empirically (Luarn and Lin, 2003). Such factors are often hard, if not impossible, to measure quantitatively and therefore are largely unobserved, and it is neither theoretically nor empirically clear whether a single-dimensional scalar term is sufficient to capture the effects from such factors. In the meanwhile, completely ignoring these factors will likely create endogeneity issues in econometric analysis of consumer behaviors, and it might be hard to find proper instruments for every potentially relevant latent factor. Therefore, we believe that our main model along with the assumptions above, admittedly with its own restriction to the fixed-effect specification, constitutes a step forward in the direction of accommodating more complex unobserved heterogeneity.

A noteworthy restriction of Assumption 2.3 lies in that it rules out random coefficients, a widely adopted modeling device proposed by Berry, Levinsohn, and Pakes (1995) to induce sophisticated substitution patterns among products with multi-dimensional characteristics space. However, the flexibility afforded by our general fixed effect specification can incorporate arbitrarily complicated substitution patterns with respect to *time-invariant* components of observed and unobserved product characteristics, by exploiting the panel structure of observable data along with the time homogeneity assumption (Assumption 2.3). It is thus worth pointing out that our current fixed-effect approach and the random-coefficient approach are two rather different methods: neither nests the other as a special case, and the two approaches may be more suitable for different sets of empirical applications. The random-coefficient approach using market share inversion, as developed by Berry, Levinsohn, and Pakes (1995), Berry, Gandhi, and Haile (2013) and Berry and Haile (2014), has already been widely used in various settings of demand analysis where time-varying (or market-varying) endogeneity is a major concern. Our infinite-dimensional fixed-effect approach based on weak monotonicity might be more suitable to panel-data settings where researchers are more interested in incorporating an arbitrarily complicated form of time-invariant heterogeneity across agent-product pairs.

Finally, as briefly discussed in Section 2.2.1 and formally stated in Assumption 2.3, the whole distribution of $\boldsymbol{\epsilon}_{it}$ can be indexed by the fixed effect $\mathbf{A}_i$. Furthermore, serial autocorrelation in $\boldsymbol{\epsilon}_{it}$ is not ruled out either, as Assumption 2.3 concerns only the marginal distributions of $\boldsymbol{\epsilon}_{it}$ in different periods.

We may now proceed to provide identification arguments for the leading parameter of interest, β_0, in Section 2.3 and construct estimators of β_0 in Section 2.4.

2.3 Identification Strategy

In this section, we present semiparametric identification results for model (2.2) under Assumptions 2.1-2.3. However, as will become clear later in this section, the underlying idea of our identification strategy applies more widely beyond panel multinomial choice models. See Section 2.7 for more details.

Our key identification strategy exploits the standard notion of multivariate monotonicity in its contrapositive form. As a reminder, we start with a standard definition of multivariate monotonicity, followed by a statement of its logical contraposition.

Definition 2.1 (**Multivariate Monotonicity**)**.** A real-valued function $\psi : \mathbb{R}^J \to \mathbb{R}$ is said to be *weakly increasing* if, for any pair of vectors $\overline{\boldsymbol{\delta}}$ and $\underline{\boldsymbol{\delta}}$ in $\mathbb{R}^J$, if $\overline{\boldsymbol{\delta}}_j \leq \underline{\boldsymbol{\delta}}_j$ for every $j = 1, ..., J$, then $\psi\left(\overline{\boldsymbol{\delta}}\right) \leq \psi\left(\underline{\boldsymbol{\delta}}\right)$.

Remark 2.1 (**Logical Contraposition**). The following is equivalent to Definition 2.1:

$$\psi\left(\overline{\boldsymbol{\delta}}\right) > \psi\left(\underline{\boldsymbol{\delta}}\right) \quad \Rightarrow \quad \text{NOT } \left\{\overline{\boldsymbol{\delta}}_j \leq \underline{\boldsymbol{\delta}}_j \text{ for all } j = 1, ..., J\right\}. \tag{2.4}$$

for any $\left(\overline{\boldsymbol{\delta}}, \underline{\boldsymbol{\delta}}\right)$, where "NOT" denotes the logical negation operator.

Our subsequent identification strategy will leverage heavily the simple contraposition of monotonicity (2.4), and our arguments proceed in three major steps. First, we define a multivariate monotone function in the form of conditional choice probabilities. Second, we construct an observable inequality based on the monotone function we define, effectively producing the left-hand side of (2.4). Finally, we use the contraposition of monotonicity to obtain the right-hand side of (2.4), which will translate into identifying restrictions on the parameter β_0 via the indexes $\boldsymbol{\delta}_{it} := (\delta_{ijt})_{j=1}^{J}$.

We now present our key identification strategy step by step. For the moment, we fix a particular product $j \in \{1, ..., J\}$, a pair of time periods $t \neq s \in \{1, ..., T\}$ and condition on a generic realization of the observable covariates in the two periods t and s, i.e., $(\mathbf{X}_{it}, \mathbf{X}_{is}) = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \in Supp\,(\mathbf{X}_{it}, \mathbf{X}_{is})$.

Step 1: Construction of a monotone function

For each individual i, consider i's choice probability of j given $(\mathbf{X}_{it}, \mathbf{A}_i)$:

$$\begin{aligned}\mathbb{E}\left[y_{ijt}\middle|\,\mathbf{X}_{it}, \mathbf{A}_i\right] &= \int \mathbb{1}\left\{u\left(X'_{ijt}\beta_0, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k\neq j} u\left(X'_{ikt}\beta_0, A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\epsilon_{ijt}\middle|\,\mathbf{X}_{it}, \mathbf{A}_i\right)\\ &= \int \mathbb{1}\left\{u\left(\delta_{ijt}, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k\neq j} u\left(\delta_{ikt}, A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\epsilon_{ijt}\middle|\,\mathbf{A}_i\right)\\ &=: \psi_j\left(\delta_{ijt}, \left(-\delta_{ikt}\right)_{k\neq j}, \mathbf{A}_i\right) \qquad (2.5)\end{aligned}$$

where the second equality follows from the index definition $\delta_{ijt} = X'_{ijt}\beta_0$ and Assumption 2.3 (Conditional Time Homogeneity of Errors), which enables us to write ψ_j without the time subscript t. Clearly, the monotonicity of the utility function u in the index argument δ_{ijt} (Assumption 2.1) translates into the multivariate monotonicity of the function ψ_j in the vector of indexes $\left(\delta_{ijt}, (-\delta_{ikt})_{k\neq j}\right)$[6]:

Lemma 2.1. $\psi_j\left(\,\cdot\,, \mathbf{A}_i\right) : \mathbb{R}^J \to \mathbb{R}$ *is weakly increasing, for any realized* $\mathbf{A}_i$.

[6] We flip the signs of $(\delta_{ikt})_{k\neq j}$ purely for the ease of exposition: as discussed earlier, it is the monotonicity, not the exact direction of monotonicity, that matters in our analysis.

In terms of economic interpretation, $\psi_j\left(\boldsymbol{\delta}_{it}, \mathbf{A}_i\right)$ summarizes each agent i's conditional choice probability of product j given i's fixed effect $\mathbf{A}_i$ as a function of the index vector $\boldsymbol{\delta}_{it}$. Lemma 2.1 admits a simple interpretation: if a product j becomes weakly better for agent i (in terms of the index δ_{ijt}), while all other products $k \neq j$ becomes weakly worse, then agent i's choice probability of product j should weakly increase.

However, as the realization of $\mathbf{A}_i$ is not observable, the conditional choice probability function $\psi_j\left(\,\cdot\,, \mathbf{A}_i\right)$ is not directly identified from data in the short-panel setting under consideration here. In the next step, we construct an observable quantity based on ψ_j by averaging out $\mathbf{A}_i$.

Step 2: Construction of an observable inequality

Consider the following intertemporal difference in conditional choice probabilities:

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) := \mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}\right] \tag{2.6}$$

which is by construction directly identified from data.

Write $\overline{\boldsymbol{\delta}} := \overline{\mathbf{X}}\beta_0 \equiv \left(\overline{X}_j^{\prime}\beta_0\right)_{j=1}^{J}$ and similarly for $\underline{\boldsymbol{\delta}}$, and $\mathbf{X}_{i,ts} := \left(\mathbf{X}_{it}, \mathbf{X}_{is}\right)$. The following lemma translates the monotonicity of $\psi_j\left(\overline{\boldsymbol{\delta}}, \mathbf{A}_i\right)$ in the index vector $\overline{\boldsymbol{\delta}}$ into a restriction on the sign of the observable quantity $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$, effectively corresponding to an observable scalar inequality.

Lemma 2.2. $\overline{\delta}_j \leq \underline{\delta}_j$ and $\overline{\delta}_k \geq \underline{\delta}_k$ for all $k \neq j \implies \gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \leq 0$.

To see why Lemma 2.2 is true, rewrite $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ as

$$\begin{aligned}
\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) &= \mathbb{E}\left[\mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right] \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{A}_i\right] - \mathbb{E}\left[y_{ijs} \middle| \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right] \\
&= \int \left[\psi_j\left(\overline{\delta}_j, \left(-\overline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) - \psi_j\left(\underline{\delta}_j, \left(-\underline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right)\right] d\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right).
\end{aligned}$$

Whenever $\overline{\delta}_j \leq \underline{\delta}_j$ and $\overline{\delta}_k \geq \underline{\delta}_k$ for all $k \neq j$, by Lemma 1 we have

$$\psi_j \left(\overline{\delta}_j, \left(-\overline{\delta}_k \right)_{k \neq j}, \mathbf{A}_i \right) - \psi_j \left(\underline{\delta}_j, \left(-\underline{\delta}_k \right)_{k \neq j}, \mathbf{A}_i \right) \leq 0$$

for *every* possible realization of $\mathbf{A}_i$. Consequently, the inequality will be preserved after integrating over the fixed effect $\mathbf{A}_i$ *cross-sectionally* with respect to the conditional distribution $\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}} \right)$, a potentially hugely complicated probability measure that we leave unspecified.

Step 3: Derivation of the key identifying restriction

We now take the logical contraposition of Lemma 2.2:

Proposition 2.1 (**Key Identifying Restriction**). *Under Assumptions 2.1, 2.2 and 2.3,*

$$\gamma_{j,t,s} \left(\overline{\mathbf{X}}, \underline{\mathbf{X}} \right) > 0 \;\Rightarrow\; \text{NOT } \left\{ \left(\overline{X}_j - \underline{X}_j \right)' \beta_0 \leq 0 \text{ and } \left(\overline{X}_k - \underline{X}_k \right)' \beta_0 \geq 0 \; \forall k \neq j \right\} \tag{2.7}$$

Recall that $\delta_{ijt} = X_{ijt}' \beta_0$, so Proposition 2.1 follows immediately from Lemma 2.2 and defines an identifying restriction on β_0 that is free of all unknown nonparametric heterogeneity terms u, $\mathbf{A}$ and $\boldsymbol{\epsilon}$. Proposition 2.1 is also very intuitive: if we observe an intertemporal increase in the conditional choice probability of product j from one period to another, it is impossible that product j's index becomes worse, while all other products' indexes become better.

The simple idea behind Proposition 2.1 is to leverage the contraposition of monotonicity in the index vector, which, apart from its simplicity, brings about robustness against the rich built-in forms of unobserved heterogeneity along with nonseparability. As the validity of this idea relies only on monotonicity in an index structure, it is applicable more widely beyond the panel multinomial choice settings we are currently considering. See Section 2.7 for a general framework under which the contraposition of monotonicity may be utilized. In particular, in a companion paper (Gao, Li, and Xu, 2020), we adapt this idea to the additional complications induced in a network formation setting, where nonseparability arises naturally from nontransferable utilities.

We also note that the same idea can be readily extended to any nonempty subset of products, as summarized in the following corollary:

Corollary 2.1. *If* $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right) > 0$ *for all* $j \in J_1 \subseteq \{0,1,...,J\}$*, it must NOT be that* $\left(\overline{X}_j - \underline{X}_j\right)'\beta_0 \leq 0$ *for all* $j \in J_1$ *while* $\left(\overline{X}_k - \underline{X}_k\right)'\beta_0 \geq 0$ *for all* $k \in J\backslash J_1$*.*

Intuitively, if we observe that the conditional choice probabilities of all products in J_1 strictly increase across two periods of time, it cannot be the case that the indexes of all products in J_1 have weakly worsened while the indices of all products outside J_1 have weakly improved. Li (2019) shows that, at least in the case of $T = 2$, the collection of all identifying restrictions in Corollary 2.1 lead to *sharp* identification of β_0. That said, for the rest of the paper we will focus on the identifying restrictions in Proposition 2.1, while noting that all the analysis below can be readily adapted to incorporate the additional restrictions in Corollary 2.1.

Formulation of Population Criterion Functions

We now formulate a population criterion function based on Proposition 2.1. For every candidate parameter $\beta \in \mathbb{R}^D$, we represent in Boolean algebra the right hand side of (2.7) in Proposition 2.1 by

$$\lambda_j\left(\overline{\mathbf{X}},\underline{\mathbf{X}};\beta\right) := \prod_{k=1}^{J} \mathbb{1}\left\{(-1)^{\mathbb{1}\{k\neq j\}}\left(\overline{X}_k - \underline{X}_k\right)'\beta \leq 0\right\}, \tag{2.8}$$

where $(-1)^{\mathbb{1}\{k\neq j\}}$ takes the value -1 for $k \neq j$ and 1 for $k = j$. Therefore, Proposition 2.1 can be written algebraically as: $\gamma_{j,t,s}\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right) > 0$ implies $\lambda_j\left(\overline{\mathbf{X}},\underline{\mathbf{X}};\beta_0\right) \equiv 0$ for any $\left(\overline{\mathbf{X}},\underline{\mathbf{X}}\right)$. We now define the following criterion function by taking a cross-sectional expectation over the random realization of $(\mathbf{X}_{it},\mathbf{X}_{is})$:

$$Q_{j,t,s}(\beta) := \mathbb{E}\left[\mathbb{1}\left\{\gamma_{j,t,s}\left(\mathbf{X}_{it},\mathbf{X}_{is}\right) > 0\right\}\lambda_j\left(\mathbf{X}_{it},\mathbf{X}_{is};\beta\right)\right], \tag{2.9}$$

which is clearly nonnegative and minimized to zero at the true parameter value β_0. Without normalization and further assumptions for point identification, there might be multiple values of β_0 that minimize $Q_{j,t,s}$ to zero.

More generally, fix any function $G : \mathbb{R} \to \mathbb{R}$ that is *one-sided sign preserving*, i.e., $G(z) > 0$ for $z > 0$ and $G(z) = 0$ for $z \leq 0$. For example, we can choose $G(z) = [z]_+$ where $[z]_+$ is the positive part function. Then, we define $Q^G_{j,t,s}$ as

$$Q^G_{j,t,s}(\beta) := \mathbb{E}\left[G\left(\gamma_{j,t,s}\left(\mathbf{X}_{it}, \mathbf{X}_{is}\right)\right) \lambda_j\left(\mathbf{X}_{it}, \mathbf{X}_{is}; \beta\right)\right], \tag{2.10}$$

which is also minimized to zero at the true parameter value β_0. The sign-preserving function G, if also set to be monotone, continuous or bounded, serves as a *smoothing* function that helps with the finite-performance of our estimators. We will provide more discussions on function G in the next section, when we construct estimators based on the sample analog of the population criterion function defined here. It is worth pointing out that this smoothing function G is built into the *population* criterion function as in (2.10), which is different from the usual technique where smoothing is only done in finite samples but not in the population. For notational simplicity, we suppress G in $Q^G_{j,t,s}$ and simply write $Q_{j,t,s}$ throughout this paper.

So far we have focused on a fixed product j and a fixed pair of periods (t, s), but in practice we may utilize the information across all products and all pairs of periods by defining the aggregated criterion function:

$$Q(\beta) := \sum_{j=1}^{J} \sum_{t \neq s}^{T} Q_{j,t,s}(\beta), \quad \text{for any } \beta \in \mathbb{R}^D. \tag{2.11}$$

We summarize our main identification result in the following theorem.

Theorem 2.1 (Set Identification). *Under model* (2.1) *and Assumptions 2.1-2.3,*

$$\beta_0 \in B_0 := \left\{\beta \in \mathbb{R}^D : Q(\beta) = 0\right\}. \tag{2.12}$$

We will refer to B_0 as the *identified set*. In Appendix 2.C, we provide sufficient conditions for point identification of β_0 up to scale normalization, with similar styles of assumptions imposed for point identification in the literature on maximum-score or rank-order estimation, dating back to Manski (1985), as well as in related work on panel multinomial choice models, such as Shi, Shum, and Song (2018) and Khan, Ouyang, and Tamer (2019).[7] However, since point identification, or lack thereof, is conceptually irrelevant to our key methodology, and as set identification and set estimation are becoming increasingly relevant in econometric theory as well as applied research, we will focus on set identification and estimation results in the main text, following a similar approach adopted by Manski (1975). Of course, whenever the additional assumptions for point identification are satisfied in data, the set estimator will shrink to a point asymptotically.

Our criterion function is constructed to be an aggregation of the identifying restrictions on β_0 in the form of Boolean variables across all (j, t, s) in the data, obtained via the logical contraposition of weak multivariate monotonicity whenever $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is}) > 0$ occurs. As $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is}) = -\gamma_{j,s,t}(\mathbf{X}_{is}, \mathbf{X}_{it})$, either $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is}) > 0$ or $\gamma_{j,s,t}(\mathbf{X}_{is}, \mathbf{X}_{it}) > 0$ occurs for each unordered pair of periods $\{t, s\}$, provided that there is nonzero intertemporal variation in the relevant conditional choice probabilities.

It is important to note that the stochastic relationship between the outcome variable $\mathbf{y}_i$ and the observable covariates $\mathbf{X}_i$ enters into our criterion function Q only through the intertemporal differences in conditional choice probabilities as represented by the term $\gamma_{j,t,s}(\mathbf{X}_{it}, \mathbf{X}_{is})$. As the randomness of $\mathbf{y}$ conditional on $\mathbf{X}$ is completely averaged out in $\gamma_{j,t,s}$, the only remaining form of randomness in our population criterion function is the random sampling of observable covariates $\mathbf{X}_i$, which no longer involves the outcome variable $\mathbf{y}_i$.

[7] It might be worth pointing out that the identification arguments in Shi, Shum, and Song (2018) and Khan, Ouyang, and Tamer (2019) feature conditioning on *equality* events in the form of $\left\{\overline{X}_k - \underline{X}_k = \mathbf{0}, \text{ for all } k \neq j\right\}$, which essentially utilizes subsamples where observable covariates stay unchanged except for a single product j across two periods. In contrast, our point identification argument, available in Appendix 2.C, does not involve conditioning on *equalities*, but only *inequalities* that define (intersections of) half-spaces in the parameter space $\mathbb{R}^D$.

As a result, the systematic component of our population criterion function $Q_{j,t,s}$, as defined in (2.9) and (2.10), is nonstandard relative to usual forms of moment conditions as studied in the literature on extremum estimation. Specifically, in our criterion function the expectation (moment) operators show up twice, the first time in the definition of the conditional expectation $\gamma_{j,t,s}$ and the second time in the expectation over observable covariates $(\mathbf{X}_{it}, \mathbf{X}_{is})$. Moreover, the two expectation operators are separated by the *nonlinear* one-sided sign-preserving function G, so it is impossible to push inside the expectation operators via the law of iterated expectations.

Relative to the well-known maximum-score or rank-order criterion function as studied by Manski (1985, 1987) utilizing univariate monotonicity, the nonstandardness of our criterion function arises from a key difference of multivariate monotonicity from univariate monotonicity. To see this more clearly, consider the special case of a *single-index* setting $(J = 1)$[8], in which our population criterion function degenerates to the maximum-score or rank-order criterion function if we choose G to be $G(z) = [z]_+$, suppress the product subscript j, and denote X_t as the *vector* of observable covariates:

$$\begin{aligned} Q_{t,s}(\beta) + Q_{s,t}(\beta) =& \mathbb{E}\left[[\gamma(X_t, X_s)]_+ \mathbb{1}\{(X_t - X_s)\beta \geq 0\}\right] \\ &+ \mathbb{E}\left[[\gamma(X_s, X_t)]_+ \mathbb{1}\{(X_s - X_t)\beta \geq 0\}\right] \\ =& \mathbb{E}\left[(y_t - y_s)\operatorname{sgn}((X_t - X_s)\beta)\right]. \end{aligned} \tag{2.13}$$

The last line of (2.13) is the familiar maximum-score criterion function, constructed based on the following equivalence relationship induced by univariate monotonicity:

$$\gamma(X_t, X_s) > 0 \;\Leftrightarrow\; (X_t - X_s)\beta > 0, \tag{2.14}$$

[8] This arises naturally in binomial choice models with the characteristics of the outside option set to be zero. In this case, even though there are nominally two choice alternatives, choice behavior is completely determined by a single index based on the characteristics of the non-default option.

Such an equivalence relationship is a unique feature of the univariate setting, which can be derived as a special case of Proposition 2.1:

$$\gamma\left(X_t, X_s\right) > 0 \Rightarrow \text{NOT}\left\{\left(X_t - X_s\right)\beta \leq 0\right\} \Leftrightarrow \left(X_t - X_s\right)\beta > 0 \Rightarrow \gamma\left(X_t, X_s\right) \geq 0,$$

which becomes (2.14) if the monotonicity of γ is strict.

However, such equivalence relationships *cannot* be generalized to the multivariate setting with $J \geq 2$, as the right hand side of (2.7),

$$\text{NOT }\left\{\left(\overline{X}_j - \underline{X}_j\right)' \beta_0 \leq 0 \text{ and } \left(\overline{X}_k - \underline{X}_k\right)' \beta_0 \geq 0 \text{ for all } k \neq j\right\},$$

does not imply $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \geq 0$ in the converse direction. This breaks the equivalence built into the maximum-score criterion function. As a result, we can no longer aggregate $Q_{j,t,s}$ and $Q_{j,s,t}$ into a unified representation as in (2.13).

Hence, our population criterion function is a generalization of the maximum-score criterion functions to multi-index settings, where the lack of equivalence as described above leads to a key difference in the criterion functions, and consequently a different approach of estimation, which will be discussed in the next section.

2.4 Estimation and Computation

2.4.1 A Consistent Two-Step Estimator

We construct our estimator as a semiparametric two-step M-estimator.

The first stage of our procedure concerns with nonparametrically estimating the intertemporal differences in conditional choice probabilities of the following form

$$\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) = \mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right]$$

for all on-support realizations $\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$, all pairs of periods (t, s) and all products j.[9]

Given the first-stage estimators $\hat{\gamma}_{j,t,s}$ and the smoothing function G, in the second stage we numerically compute minimizers of the sample criterion function,

$$\begin{aligned}\hat{Q}\left(\beta\right) &:= \sum_{j=1}^{J} \sum_{t \neq s}^{T} \hat{Q}_{j,t,s}\left(\beta\right), \\ \hat{Q}_{j,t,s}\left(\beta\right) &:= \frac{1}{N} \sum_{i=1}^{N} G\left(\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right) \lambda_j\left(\mathbf{X}_{i,ts}; \beta\right).\end{aligned}$$

Observing that the scale of β_0 cannot be identified given that $\lambda_j\left(\mathbf{X}_{i,ts}; \beta\right)$ consists of indicator functions of the the form $\mathbb{1}\left\{\left(X_{ijt} - X_{ijs}\right)' \beta \geq 0\right\}$, we imposes the following scale normalization $\beta_0 \in \mathbb{S}^{D-1} := \left\{v \in \mathbb{R}^D : \|v\| = 1\right\}$. Following Chernozhukov, Hong, and Tamer (2007), we define the set estimator by

$$\hat{B}_{\hat{c}} := \left\{\beta \in \mathbb{S}^{D-1} : \ \hat{Q}\left(\beta\right) \leq \min_{\tilde{\beta} \in \mathbb{S}^{D-1}} \hat{Q}\left(\tilde{\beta}\right) + \hat{c}\right\} \tag{2.15}$$

with $\hat{c} := O_p\left(c_N \log N\right)$. We now introduce assumptions for the consistency of $\hat{B}_{\hat{c}}$.

Assumption 2.4 (**First-Stage Estimation**). *For any* (j, t, s)*:*

(i) $\gamma_{j,t,s} \in \Gamma$*, and* $\mathbb{P}\left(\hat{\gamma}_{j,t,s} \in \Gamma\right) \to 1$*, with* Γ *being a* $\mathbb{P}$*-Donsker class of functions in* $L_2\left(\mathbf{X}\right)$ *s.t.* $\sup_{\gamma_{j,t,s} \in \Gamma} \mathbb{E}\left|\gamma_{j,t,s}\right| < \infty$*;*

(ii) $\left\|\hat{\gamma}_{j,t,s} - \gamma_{j,t,s}\right\|_2 := \sqrt{\int \left(\hat{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right) - \gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)^2 \mathrm{d}\mathbb{P}\left(\mathbf{X}_{i,ts}\right)} = O_p\left(c_N\right)$ with $c_N \searrow 0$ as $N \to \infty$.

Through Assumption 2.4 we take as given the large set of theoretical results on nonparametric regression in the literature. Many kernel-based and sieve-based methods have been developed

[9] In practice, we only need to estimate $\gamma_{j,t,s}$ for $(J-1)$ products and $\frac{1}{2}T\left(T-1\right)$ *ordered* pairs of periods. The former is because conditional choice probabilities must sum to one across all J products, so we may easily compute the estimator for the last product from the other $(J-1)$ estimates: $\gamma_{J,t,s} = 1 - \sum_{j=1}^{J-1} \gamma_{j,t,s}$. The latter is because $\gamma_{j,t,s} = -\gamma_{j,s,t}$ by construction, so we may estimate it for either (t, s) or (s, t). Notice, however, that each ordered pair (t, s) or (s, t) provides complementary identifying information, as $\lambda\left(\mathbf{X}_{i,ts}; \beta\right)$ and $\lambda\left(\mathbf{X}_{i,st}; \beta\right)$ do not admit such kind of deterministic relationship.

with different properties demonstrated under various sets of conditions. See Wasserman (2006) and Chen (2007) for more comprehensive surveys.

Assumption 2.5 (**Nice Smoothing Function**). *The one-sided sign-preserving function* $G : \mathbb{R} \to \mathbb{R}_+$ *is Lipschitz continuous with a finite Lipschitz constant.*

Assumption 2.5 is not necessary for consistency per se given that our identification result is valid with any choice of the one-sided sign-preserving function G, nevertheless we take G to be Lipschitz so as to simplify the proof.

To state the next assumption, we decompose each row (product) of $\overline{\mathbf{X}} - \underline{\mathbf{X}}$ as the product of its norm and its *direction*, i.e., $\overline{X}_k - \underline{X}_k \equiv r_k\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) \cdot v_k\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right)$, where $r_k\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) := \left\|\overline{X}_k - \underline{X}_k\right\|$, and $v_k\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) := \left(\overline{X}_k - \underline{X}_k\right) / \left\|\overline{X}_k - \underline{X}_k\right\|$ if $\overline{X}_k \neq \underline{X}_k$ while $v_k\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) := 0$ if $\overline{X}_k = \underline{X}_k$.

Assumption 2.6 (**Continuous Distribution of Directions**). *The marginal distribution of* $v_k\left(\mathbf{X}_{it} - \mathbf{X}_{is}\right)$ *has no mass point except possibly at* $\mathbf{0}$ *for each* (k, t, s).

Assumption 2.6 is a technical assumption that ensures the continuity of the population criterion function $Q(\theta)$. It is likely to be *not* necessary for consistency, but we impose it for simplicity. We note that Assumption 2.6 is fairly weak: it essentially requires that the *directions* of intertemporal differences in observable characteristics are continuously distributed on their own supports. In particular, this allows all but one dimensions of observable characteristics to be discrete.

With the above assumptions, we now establish the consistency of the set estimator $\hat{B}_{\hat{c}}$ based on Chernozhukov, Hong, and Tamer (2007).

Theorem 2.2 (**Consistency**). *Under Assumptions 2.1-2.6, the set estimator* $\hat{B}_{\hat{c}}$ *is consistent in Hausdorff distance:* $d_H\left(\hat{B}_{\hat{c}}, B_0\right) = o_p(1)$, *where*

$$d_H\left(\hat{B}_{\hat{c}}, B_0\right) := \max\left\{\sup_{\beta \in \hat{B}_{\hat{c}}} \inf_{\tilde{\beta} \in B_0} \left\|\beta - \tilde{\beta}\right\|, \sup_{\beta \in B_0} \inf_{\tilde{\beta} \in \hat{B}_{\hat{c}}} \left\|\beta - \tilde{\beta}\right\|\right\}.$$

Furthermore, if β_0 is point-identified on $\mathbb{S}^{D-1}$, $\left\|\hat{\beta}-\beta_0\right\| = o_p(1)$ for any $\hat{\beta} \in \hat{B} := \arg\min_{\tilde{\beta}\in\mathbb{S}^{D-1}} \hat{Q}\left(\tilde{\beta}\right)$.

2.4.2 Computation

We now provide more details on how we practically implement our estimator.

First-Stage Nonparametric Regression

For the first-stage nonparametric estimation of γ, we adopt a machine learning estimator based on single-layer artificial neural networks, which has been widely adopted in many disciplines due to its theoretical and numerical advantages in estimating nonlinear and high dimensional functions. Clearly, model (2.1) naturally induces nonlinearity through the complex inequalities inside the multinomial choice model (2.1) with unknown forms of utility functions. Also, given that the estimation of $\gamma_{j,t,s}$ includes (time-varying) all observable product characteristics from two periods, the potentially high dimensionality of covariates also makes machine learning algorithm a suitable choice. For single-layer neural network estimators, Chen and White (1999) provides theoretical results on the convergence rates, establishing that $c_N = \left(\frac{\log N}{N}\right)^{\frac{1+2/(d+1)}{4(1+1/(d+1))}}$. On the computational side, there are also many readily usable computational packages to implement neural-network estimators. For example, in our simulation study and empirical illustration, we use the `R` package “`mlr`” by Bischl et al. (2016), which provides a front end for cross validation and hyperparameter tuning.

Choice of the Smoothing Function G

Besides the requirement of Lipschitz continuity in Assumption 2.5, in practice we take G to be bounded from above by setting $G(z) = 2\Phi\left([z]_+\right) - 1$, where Φ is the standard normal CDF. We now motivate our choice of G.

Recall that our identification strategy is based on the logical implication of the event $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0$, so for identification purposes we are only interested in $\mathbb{1}\left\{\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0\right\}$, i.e., whether the event $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0$ occurs, but not in the exact magnitude of $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$. However, in finite-sample, when $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is close to zero, the estimator $\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is relatively more likely to have the wrong sign, so that the plug-in estimator $\mathbb{1}\left\{\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0\right\}$ may induce an error of the size 1. Hence the smoothing by $G(\cdot)$ helps down-weight the observations when $\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is close to zero and shrinks the magnitude of possible errors.

On the other hand, when $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ is positive and large so that $\mathbb{1}\left\{\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0\right\}$ can be estimated well, we do not care much about the magnitude of $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$, which does not provide additional identifying information per se. By setting G to be bounded from above, we dampen the effects of large $\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ at the same time, so that the numerical maximization of $\hat{Q}$ is not too sensitive to potential large but redundant variations in $\hat{\gamma}_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$.

Angle-Space Reparameterization of $\mathbb{S}^{D-1}$

In the second stage optimization of $\hat{Q}(\beta)$ over $\beta \in \mathbb{S}^{D-1}$, we work with a reparameterization of $\mathbb{S}^{D-1}$ with $(D-1)$ angles in spherical coordinates[10]. Specifically, define the angle space Θ by

$$\Theta := [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{D-2}, \tag{2.16}$$

[10] The idea and the motivations for using the angle-space reparameterization were also found in Manski and Thompson (1986), who however used only one angle parameter, given two pre-chosen orthogonal unit vectors on $\mathbb{S}^{D-1}$.

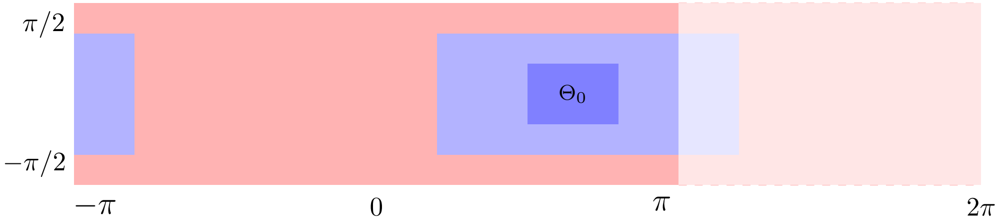
and the transformation $\theta \longmapsto \beta(\theta)$ by

$$\beta(\theta) = \begin{cases} \beta_1(\theta) & := \cos\theta_{D-1}\ldots\cos\theta_2\cos\theta_1, \\ \beta_2(\theta) & := \cos\theta_{D-1}\ldots\cos\theta_2\sin\theta_1, \\ \vdots & \vdots \\ \beta_{D-1}(\theta) & := \cos\theta_{D-1}\sin\theta_{D-2}, \\ \beta_D(\theta) & := \sin\theta_{D-1}, \end{cases}$$

we now instead solves the optimization of $\hat{Q}(\beta(\theta))$ over Θ, which we further equip with its natural geodesic metric $\rho_\Theta\left(\theta,\tilde{\theta}\right) := \arccos\left(\beta(\theta)'\,\beta\left(\tilde{\theta}\right)\right)$, which is strongly equivalent to the (imported) Euclidean distance $\left\|\beta(\theta) - \beta\left(\tilde{\theta}\right)\right\|$.

This reparameterization (Θ, ρ_Θ) enables us to exploit the compactness and convexity of the parameter space $\Theta = [-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{D-2}$, which takes the form of a hyper-rectangle. First, (Θ, ρ_Θ) preserves all topological structure of the unit sphere, and particularly inherits the compactness of $\left(\mathbb{S}^{D-1}, \|\cdot\|\right)$, automatically satisfying the compactness condition usually imposed for extremum estimation and making it numerically feasible to initiate a grid on the whole parameter space. Second, while the unit sphere $\mathbb{S}^{D-1}$ is not convex, the new parameter space Θ becomes convex algebraically, making it computationally easy to define bisection points in the parameter space. Third, it also preserves the geometric structures of the sphere, including for instance the obvious observation that $-\pi$ and π in the first coordinate of Θ should be treated as exactly the same point, or more rigorously, $\rho_\Theta((\pi-\epsilon, \theta_2, ..., \theta_{D-1}), (-\pi, \theta_2, ..., \theta_{D-1})) \to 0$ as $\epsilon \to 0$. This seemingly trivial property is nevertheless important in defining and interpreting whether certain parameter estimates converge asymptotically or not, and provides conceptual foundations for subsequent asymptotic theories.

Figure 2.1: An Adaptive-Grid Algorithm



An Adaptive-Grid Algorithm

With the angle reparameterization, we seek to numerically compute a conservative rectangular enclosure of $\arg\min \hat{Q}(\theta)$, deploying a bisection-style grid-search algorithm that recursively shrinks and refines an *adaptive grid* to any pre-chosen precision (as defined by ρ_Θ). Unlike gradient-based local optimization algorithms, our adaptive grid algorithm handles well the built-in discreteness in our sample criterion function, which has zero derivative almost everywhere, while maintains global initial coverage over the whole parameter space. While a brute-force global search algorithm is the safest choice if the dimension of product characteristics D is relatively small, our adaptive-grid algorithm performs significantly faster. The essential structure of our algorithm is laid out as follows, with a corresponding illustration in Figure 2.1.

Step 1: Initialize a global grid $\Theta^{(1)}$ of some chosen size M_0^{D-1} on Θ.

Step 2: Compute $\hat{Q}(\theta)$ for each $\theta \in \Theta^{(1)}$, and select all points in $\Theta^{(1)}$ with a criterion value below the αth-quantile in $\hat{Q}\left(\Theta^{(1)}\right) := \left\{\hat{Q}(\theta) : \theta \in \Theta^{(1)}\right\}$ into

$$\underline{\Theta}^{(1)} := \left\{\theta \in \Theta^{(1)} : \ \hat{Q}(\theta) \leq \text{quantile}_\alpha \left(\hat{Q}\left(\Theta^{(1)}\right)\right)\right\}.$$

Step 3: Take the enclosing rectangle of $\underline{\Theta}^{(1)}$, by defining $\underline{\theta}_d^{(1)} := \min^* \underline{\Theta}_d^{(1)}$ and $\overline{\theta}_d^{(1)} := \max^* \underline{\Theta}_d^{(1)}$, where $\underline{\Theta}_d^{(1)} := \left\{\theta_d : \theta \in \underline{\Theta}^{(1)}\right\}$ for each $d = 1, ..., D-1$ and the operator $\min^*$ and $\max^*$ have standard definitions of min and max except for the first dimension $d = 1$. For the first dimension, it is necessary to account for the underlying spherical geometry and the periodicity of angles, i.e. $\theta_1 + 2\pi \equiv \theta_1$ and in particular $-\pi \equiv \pi$. This, however, is largely a programming nuisance: whenever $\underline{\Theta}_1^{(1)} \subsetneq \Theta_1^{(1)}$ crosses over at $-\pi$ and π, we can add 2π to every $\theta_1 \in \underline{\Theta}_1^{(1)}$ and obtain lower and upper bounds of $\underline{\Theta}_1^{(1)} + 2\pi$, as illustrated in Figure 2.1.

Step 4: We initialize a refined grid $\Theta^{(2)}$ on $\overline{\underline{\Theta}}^{(1)} := \times_{d=1}^{D-1} \left[\underline{\theta}_d^{(1)}, \overline{\theta}_d^{(1)}\right]$ of size M_0^{D-1}.

Step 5: Reiterate until refinement stops (falls below a certain numerical precision).

Note that the above is simply a sketch of our algorithm.[11] To be conservative, we add in buffers at each step of refinement, keep track of both outer and inner boundaries of the lower-quantile set $\underline{\Theta}^{(m)}$, and make sure that the minimizers of the criterion functions at all computed points are indeed enclosed by the set returned in the end. We find the current algorithm to be conservative and perform reasonably well in our simulation study and empirical illustration.

2.5 Simulation

In this section, we examine the finite-sample performance of our estimation method via a Monte Carlo simulation study. We start by studying the performance of the first-stage nonparametric estimator $\hat{\gamma}$ or $G\left(\hat{\gamma}\right)$. Then, we show how the two-stage estimator $\hat{\beta}$ performs under various configurations of the data generating process (DGP). Finally, we investigate how our estimator performs without point identification.

Setup of Simulation Study

For each DGP configuration, we run $M = 100$ simulations of model (2.1) with the following utility specification for each agent-product-time tuple ijt:

[11] Our algorithm relies heavily on the compactness and convexity of the angle space Θ. Compactness allows us to start with a global grid over the whole parameter space for initial evaluations of the sample criterion function. At each step of recursion, the convexity of Θ enables us to conveniently refine the grid by separately cutting each coordinate of $\overline{\underline{\Theta}}^{(m)}$ into smaller pieces through simple division.

$$u\left(X_{ijt}'\beta_0,\, A_{ij},\, \epsilon_{ijt}\right) = A_{i0}\left(X_{ijt}'\beta_0 + A_{ij}\right) + \epsilon_{ijt},$$

where A_{i0} is an unobserved scale fixed effect that captures agent-level heteroskedasticity in utilities, and A_{ij} is an unobserved location shifter specific to each agent-product pair. The ability to deal with nonlinear dependence caused by the unobservable fixed effects A in a robust way differentiates our method from others. To allow for such dependence, we generate correlation between the observable characteristics $\mathbf{X}_i$ and the fixed effects $\mathbf{A}_i$ via a latent variable Z[12]. Furthermore, we set $\beta_0 = (2, 1, ..., 1)' \in \mathbb{R}^D$ and draw $\epsilon_{ijt} \sim TIEV(0, 1)$. To summarize, for each of the $M = 100$ simulations we first generate $(\beta_0, \mathbf{X}_{it}, \mathbf{A}_i, \boldsymbol{\epsilon}_{it})$ for all it combinations. Then we calculate the binary individual choice $\mathbf{Y}$ matrix according to model (2.1). Lastly, we compute $\hat{\beta}$ from the simulated observable data of $(\mathbf{X}, \mathbf{Y})$, and finally compare our estimator $\hat{\beta}$ with the true parameter value β_0 normalized to $\mathbb{S}^{D-1}$.

2.5.1 First-Stage Performance

We examine the performance of our first stage estimator $\hat{\gamma}$ or $G(\hat{\gamma})$. First, we calculate the true γ or $G(\gamma)$ using the knowledge of DGP which serves as the benchmark for comparison later on. Next, we estimate γ with only the observable data $(\mathbf{X}, \mathbf{Y})$ using single-layered neural networks and calculate the plugged-in functional $G\left(\hat{\gamma}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right)$ at each realized $\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$. Finally, we evaluate the performance of our estimated $G(\hat{\gamma})$ by comparing it against the true $G(\gamma)$.

We report in Table 2.1 both the means and the maximums of the mean squared errors (MSE) across M simulations to evaluate the performance of our first stage estimator $G(\hat{\gamma})$. The header of Table 2.1 lists the three choices of the one-sided sign preserving function G. The first row, “mean MSE”, reports the average MSE of $G(\hat{\gamma})$ against the true $G(\gamma)$, i.e.

[12] We draw $Z_i \sim \mathcal{N}(0, 1)$ and let $A_{i2} = [Z_i]_+$. Then, we construct $X_{ijt}^{(2)} = W_{ijt} + Z_i$ with $W_{ijt} \sim \mathcal{N}(0, 2J)$. The DGP for the rest of $\mathbf{A}$ and $\mathbf{X}$ are: $A_{i0} \sim \mathcal{U}[2, 2.5]$, $A_{i1} \equiv 0$, $A_{ij} \sim \mathcal{U}[-0.25, 0.25]$ for $j \geq 3$, $X_{ijt}^{(1)} \sim \mathcal{U}[-1, 1]$, $X_{ijt}^{(d)} \sim \mathcal{N}(0, 1)$ for $d \geq 3$.

Table 2.1: Performance of First Stage Estimator $G\left(\hat{\gamma}\right)$

	$\mathbb{1}\left\{\hat{\gamma} > 0\right\}$	$\left[\hat{\gamma}\right]_+$	$2\Phi\left(\left[\hat{\gamma}\right]_+\right) - 1$
mean MSE	0.1290	0.0221	0.0109
max MSE	0.1578	0.0254	0.0124

$\frac{1}{M}\sum_{m=1}^{M} \text{MSE}^{(m)}$ where $\text{MSE}^{(m)}$ is the MSE of $G\left(\hat{\gamma}\right)$ in the m^{th} simulation. The second row reports the maximum MSE of $G\left(\hat{\gamma}\right)$.

From Table 2.1, we see that the adjusted normal CDF $2\Phi\left(\left[\hat{\gamma}\right]_+\right) - 1$ performs the best in terms of both mean MSE and max MSE, while the indicator function gives the worst results and that the performance of the positive part function lies somewhere in between. This is expected because when the true γ is close to zero, it is more likely to have the estimated sign of $\hat{\gamma}$ to be different from γ. The discontinuity of the indicator function $\mathbb{1}\left\{\hat{\gamma} > 0\right\}$ at 0 magnifies this uncertainty around zero and leads to a higher MSE. When the true γ is positive and large, it actually does not matter for our method whether the exact value of γ is estimated well by $\hat{\gamma}$. All we need is the sign of $\hat{\gamma}$ coincides with the sign of γ so as to obtain identifying restrictions on β_0. The adjusted normal CDF $2\Phi\left(\left[\hat{\gamma}\right]_+\right) - 1$ performs the best, because it not only dampens the uncertainty in the estimated sign of $\hat{\gamma}$ near zero, but also attenuates the sensitivity to the exact value of $\hat{\gamma}_+$ relative to γ_+ when γ is positive and large. For this reason, we will use the adjusted normal CDF function in our second stage.

2.5.2 Two-Stage Performance

We present the performance of our second stage estimator $\hat{\beta}$. First, we show the simulation results under the baseline DGP configuration, where β_0 is point-identified. Next, we study the performance of our algorithm under different numbers of individuals N.[13] Finally, we inspect how our estimator performs without point identification.

[13] We also vary dimensions of observable characteristics D, numbers of products available J, and numbers of time periods T and present the results in Appendix 2.D.

Table 2.2: Baseline Performance

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
bias	$\frac{1}{M}\sum_m \left(\hat{\beta}_d^m - \beta_{0,d}\right)$	-0.0050	0.0021	0.0006
upper bias	$\frac{1}{M}\sum_m \left(\hat{\beta}_d^u - \beta_{0,d}\right)$	0.0015	0.0084	0.0108
lower bias	$\frac{1}{M}\sum_m \left(\hat{\beta}_d^l - \beta_{0,d}\right)$	-0.0115	-0.0042	-0.0096
mean(u−l)	$\frac{1}{M}\sum_m \left(\hat{\beta}_d^u - \hat{\beta}_d^l\right)$	0.0130	0.0126	0.0205
root MSE	$\left(\frac{1}{M}\sum_m \left\Vert\hat{\beta}^m - \beta_0\right\Vert^2\right)^{1/2}$		0.0745	
mean norm deviations	$\frac{1}{M}\sum_m \left\Vert\hat{\beta}^m - \beta_0\right\Vert$		0.0648	

Baseline Results

For the baseline configuration we set $N = 10,000, D = 3, J = 3, T = 2$. Since the sufficient conditions for point identification are satisfied under the baseline configuration, any point from the argmin set $\hat{B} := \arg\min_{\beta\in\mathbb{S}^{D-1}} \hat{Q}(\beta)$, is a consistent estimator of β_0. Specifically, we define

$$\hat{\beta}_d^u := \max \hat{B}_d, \quad \hat{\beta}_d^l := \min \hat{B}_d, \text{and} \quad \hat{\beta}_d^m := \frac{1}{2}\left(\hat{\beta}_d^u + \hat{\beta}_d^l\right)$$

for each dimension of product characteristics $d = 1, ..., D$, where $\hat{\beta}_d^u$ is the maximum value along dimension d of the argmin set $\hat{B}$, $\hat{\beta}_d^l$ is the minimum value along dimension d of $\hat{B}$, and $\hat{\beta}_d^m$ is the middle point along dimension d of $\hat{B}$.

Table 2.2 summarizes the main results for the simulations under our baseline configuration. In the first row of Table 2.2 we use the middle value $\hat{\beta}^m$ along each dimension of set estimator $\hat{B}$ to calculate the average bias against the true β_0 across all $M = 100$ simulations. The bias is very small across all three dimensions with a magnitude between -0.0050 and 0.0021. The next two rows show the biases in estimating $\beta_{0,d}$ using $\hat{\beta}_d^u$ and $\hat{\beta}_d^l$ respectively and the biases are again close to zero. The fourth row of Table 2.2 measures the average width of the set estimator $\hat{B}$ along each dimension. It is relatively tight compared to the magnitude of β_0. In the second part of Table 2.2 we report the root MSE (rMSE) and mean

Table 2.3: Performance under Varying N

	$\sum_d \lvert\text{bias}_d\rvert$	$\sum_d \text{mean(u-l)}_d$	rMSE	MND
$N = 10,000$	0.0077	0.0461	0.0745	0.0648
$N = \ 4,000$	0.0174	0.0715	0.1006	0.0884
$N = \ 1,000$	0.0694	0.1076	0.1690	0.1405
	$\left(\frac{N}{1,000}\right)^{1/2}$	$\left(\frac{N}{1,000}\right)^{1/3}$	$\frac{\text{rMSE}_{1000}}{\text{rMSE}_N}$	$\frac{\text{MND}_{1000}}{\text{MND}_N}$
$N = 10,000$	3.16	2.15	$\frac{0.1690}{0.0745} \approx 2.27$	$\frac{0.1405}{0.0648} \approx 2.17$
$N = \ 4,000$	2.00	1.59	$\frac{0.1690}{0.1006} \approx 1.68$	$\frac{0.1405}{0.0884} \approx 1.59$

norm deviations (MND) using $\hat{\beta}^m$. Our proposed algorithm is able to achieve a low rMSE and MND.

Results Varying N

We vary N while maintaining $D = 3, J = 3, T = 2$ to show how our method performs under different sample sizes. In addition to our baseline setup with $N = 10,000$, we calculate mean absolute deviation (MAD), average size of the estimated set, rMSE and MND for $N = 4,000$ and $N = 1,000$. Results are summarized in Table 2.3.

From Table 2.3, it is clear that a larger N helps with overall performance. MAD decreases from 0.0694 to 0.0077 when N increases from $1,000$ to $10,000$. The average size of the estimated sets, the rMSE, and the MND show a similar pattern. However, even with a relatively small $N = 1,000$ the result from our method is still quite informative and accurate, with the average size of the estimated set and the MND being equal to 0.1076 and 0.1405, respectively. We emphasize that here the total number of time periods T is set to a minimum of 2. Our method can extract information from each of the $T\,(T-1)$ ordered pairs of time periods, which increase quadratically with T. See Appendix 2.D for results with larger T.

Next, we numerically investigate the speed of convergence of our method when we increase sample size N from $1,000$ to $4,000$ and $10,000$ in the second part of Table (2.3). Compared

Table 2.4: Performance with and without Point ID: Further Examination

point ID ?	$\hat{c}$	rMSE			MND		
		$\hat{\beta}^m$	$\hat{\beta}^u$	$\hat{\beta}^l$	$\hat{\beta}^m$	$\hat{\beta}^u$	$\hat{\beta}^l$
(i) yes	-	0.0770	0.0789	0.0795	0.0661	0.0685	0.0697
(ii) no	0.01	0.0872	0.0880	0.0894	0.0753	0.0767	0.0775
	0.1	0.0860	0.0929	0.0939	0.0737	0.0833	0.0832
	1	0.0790	0.1268	0.1447	0.0668	0.1207	0.1295

with the case of $N_0 = 1,000$, the relative ratios of rMSE are 1.68 for $N = 4,000$ and 2.27 for $N = 10,000$, both of which lie between $(N/N_0)^{1/3}$ and $(N/N_0)^{1/2}$. A similar pattern is also found for calculations based on MND. These results indicate that our method achieves a convergence rate slower than the $N^{-1/2}$ but slightly faster than the $N^{-1/3}$ rate.

Estimation without Point Identification

We now investigate the performance of our estimator under specifications where point identification fails. To make things comparable, we fix (N, D, J, T) as in the baseline case, but we modify the configuration in two different ways. We maintain the point identification of β_0 in one setting but lose the point identification in the other[14]. We deliberately control the location and scale of each variable to be comparable across the two configurations, with the only differences being the presence of discreteness and boundedness of supports. When point identification fails, we compute the set estimator $\hat{B}_{\hat{c}}$of (2.15) with $\hat{c} > 0$. Table 2.4 contains simulation results under the two configurations, with different choices of $\hat{c}$ when point identification fails. [15]

[14] Specifically, we set $Z_i \sim \mathcal{U}\left[-\sqrt{3}, \sqrt{3}\right]$, $X_{ijt}^{(1)} \sim \mathcal{U}\left[-1, 1\right]$, $X_{ijt}^{(2)} = Z_i + \mathcal{N}(0, 6)$, and $X_{ijt}^{(3)} \sim \mathcal{N}(0, 1)$ for the point identified case. For the DGP without point identification, we let $Z_i \sim \mathcal{U}\left[-\sqrt{3}, \sqrt{3}\right]$, $X_{ijt}^{(1)} \sim \mathcal{U}\left\{-1, 1\right\}$, $X_{ijt}^{(2)} = Z_i + \mathcal{U}\left(-\sqrt{6}, \sqrt{6}\right)$, and $X_{ijt}^{(3)} \sim \mathcal{U}\left[-1, 1\right]$.

[15] Specifically, noting that $c_N \log N \leq N^{-1/4} \log N \approx 0.92 \leq 1$ for $N = 10,000$, we set $\hat{c} = 0.01$, 0.1 and 1, respectively.

In Table 2.4 , we calculate the rMSE and MND of the upper bound $\hat{\beta}^u$, the lower bound $\hat{\beta}^l$ and the middle point $\hat{\beta}^m$ of the (approximate) argmin sets$\hat{B}_{\hat{c}}$ (with $\hat{c} = 0$ under point identification and three choices of $\hat{c}$ under partial identification) with respect to the true normalized parameter β_0. Across rows in (i) and (ii), we see that the lack of point identification does negatively affect the performance of our estimates, but the impact is limited to a moderate degree. Within rows in (ii), we observe that, as expected, a more conservative choice of the constant $\hat{c}$ worsens performances of the upper and lower bounds by enlarging the estimated sets; in the meanwhile, it appears that the size (and the performance) of our estimator based on $\hat{\beta}^m$ is not terribly sensitive to the choice of $\hat{c}$.

2.6 Empirical Illustration

2.6.1 Data and Methodology

As an empirical illustration, we apply our method to the Nielsen Retail Scanner Data on popcorn sales to explore the effects of display promotion effects. The Nielsen Retail Scanner Data contains weekly information on store-level price, sales and display promotion status generated by about 35,000 participating retail store with point-of-sale systems across the United States. Among a huge variety of products covered by the Nielsen data, we choose to focus on popcorn for two reasons. First, purchases of popcorn are more likely to be driven by temporary urges of consumption without too much dynamic planning. Second, there is good variation in the display promotion status of popcorn, which enables us to estimate how important special in-store displays affect consumer's purchase decisions.

We aggregate the store level data to the $N = 205$ designated market area (DMA) level for year 2015. We focus on the top 3 brands ranked by market share, aggregate the rest into a fourth product "all other products", and allow an outside option of "no purchase". We calculate the dependent variable "market share" for each of the $J = 5$ brands. The observed

Table 2.5: Empirical Application: Summary Statistics

	mean	s.d.	min	max
DMA-level Market Share s_{ijt}	25.00%	21.59%	0.07%	96.69%
Price_{ijt}	0.4924	0.1803	0.1094	1.3587
Promo_{ijt}	0.0282	0.0377	0.0000	0.5000
$\text{Price}_{ijt} \times \text{Promo}_{ijt}$	0.0136	0.0203	0.0000	0.4505

product characteristics $\mathbf{X}$ include price, promotion status and their interaction term[16]. The summary statistics of the variables discussed above are provided in Table 3.5.

To describe the methodology, we use the observed DMA-level market shares as an estimate of $s_{ijt} = \mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i \right]$. Under the strong stationarity assumption, we run the first-stage estimation of

$$\mathbb{E}\left[s_{ijt} - s_{ijs} \middle| \mathbf{X}_{i,ts} \right] = \int \left(\mathbb{E}\left[y_{ijt} \middle| \mathbf{X}_{it}, \mathbf{A}_i \right] - \mathbb{E}\left[y_{ijs} \middle| \mathbf{X}_{is}, \mathbf{A}_i \right] \right) d\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_{i,ts} \right).$$

Specifically, we nonparametrically regress $(s_{ijt} - s_{ijs})$ on $\mathbf{X}_{i,ts}$ using single-layered neural networks from the `mlr` package in `R`, and obtain an estimator $\hat{\gamma}_j$ of $\gamma_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) := \mathbb{E}\left[s_{ijt} - s_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) \right]$. Then, we plug $\hat{\gamma}$ into our second-stage algorithm and compute the (approximate) argmin set $\hat{B}_{\hat{c}}$.

2.6.2 Results and Discussion

We report our estimation results in Table 3.6. $\left[\hat{\beta}^l, \hat{\beta}^u\right]_{\hat{c}}$ corresponds to the lower and upper bounds of the (approximate) argmin set $\hat{B}_{\hat{c}}$, while $\hat{\beta}^m_{\hat{c}} := \frac{1}{2}\left(\hat{\beta}^l_{\hat{c}} + \hat{\beta}^u_{\hat{c}}\right)$ corresponds to the middle point. We show both the exact argmin set ($\hat{c} = 0$) and the approximate argmin set

[16] We calculate Price_{ijt} as the weighted average unit price of all UPCs of the brand j in DMA i during week t. In the Nielsen data we find two variables related to promotion: display and feature. Due to their similarity, we calculate Promo_{ijt} as $(\text{feature}\vee\text{display})_{ijt}$. The interaction term $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ is included in X to show the effect of promotion on the price elasticity of consumers.

Table 2.6: Empirical Application: Estimation Results

	$\hat{\beta}^m_{\hat{c}=0}$	$\left[\hat{\beta}^l, \hat{\beta}^u\right]_{\hat{c}=0}$	$\hat{\beta}^m_{\hat{c}=0.014}$	$\left[\hat{\beta}^l, \hat{\beta}^u\right]_{\hat{c}=0.014}$
Price_{ijt}	-0.9681	[-0.9687, -0.9677]	-0.9236	[-0.9711, -0.8761]
Promo_{ijt}	0.1970	[0.1861, 0.2078]	0.1565	[0.0662, 0.2469]
$\text{Price}_{ijt} \times \text{Promo}_{ijt}$	0.1550	[0.1399, 0.1700]	0.2731	[0.0687, 0.4776]

Table 2.7: Empirical Illustration: Comparison of Results

	$\hat{\beta}^m$	$\hat{\beta}^{CyclicMono}$	$\hat{\beta}^{OLS}$	$\hat{\beta}^{OLS-FE}$	$\hat{\beta}^{MLogit-FE}$
Price_{ijt}	-0.9236	-0.3781	0.0240	-0.3803	-0.8511
Promo_{ijt}	0.1565	-0.0567	0.5760	0.5978	0.4589
$\text{Price}_{ijt} \times \text{Promo}_{ijt}$	0.2731	0.9240	-0.8171	-0.7057	-0.2552

with $\hat{c} = 0.01 \times N^{-\frac{1}{4}} \log(N) \approx 0.014$ for $N = 205$. The estimated coefficients for Price (negative) and Promo (positive) are clearly consistent with economic intuitions.

The most interesting result is the positive estimated coefficient on the interaction term $\text{Price}_{ijt} \times \text{Promo}_{ijt}$. An intuitive explanation for the positive sign is that by displaying certain products in front rows, consumers no longer see the price tags of these products adjacent to those of their competitors, and consequently become less price-sensitive for these specially promoted products.

To further illustrate the advantages of our method, we compare our $\hat{\beta}^m$ with the estimates obtained through four other different popular methods, i.e. Cyclic Monotonicity (CM) based on Shi, Shum, and Song (2018)[17], classic OLS, OLS with scalar-valued fixed effects (OLS-FE) and the multinomial logit with fixed effects (MLogit-FE). Results (normalized to $\mathbb{S}^{D-1}$) are summarized in Table 2.7.

The OLS regression result shows that the estimated coefficient on Price_{ijt} is 0.0240, which is counterintuitive and unreasonable. Moreover, as explained before, displaying the product

[17] We used 2-week cycles for all available weeks in the data for the CM method.

at the front row of the store will likely make consumers less price sensitive, implying a positive coefficient for $\text{Price}_{ijt}\times\text{Promo}_{ijt}$. However, the estimated coefficients for the interaction term using OLS, OLS-FE and MLogit-FE are all negative, contrary to that intuition. Finally, the CM-based method reports a small but negative coefficient of -0.0567 for Promo_{ijt}, which could be hard to rationalize.

We regard the contrast between our result and the results obtained in these alternative methods as an empirical illustration that by accommodating more flexible forms of unobserved heterogeneity, through the arbitrary dimensional fixed effects that are allowed to enter into consumers' utility functions in an additively nonseparable way, our method is able to produce economically more reasonable results.

2.6.3 A Possible Explanation via Monte-Carlo Simulations

In this section, we propose a possible explanation to the empirical findings in Table 2.7 via a Monte Carlo simulation. Recall that "Promo" captures whether a product gains increased exposure by being highlighted by stores. We argue that the negative estimated coefficients obtained in traditional methods in Table 2.7 for $\text{Price}_{ijt} \times \text{Promo}_{ijt}$ may be caused by a positive correlation between display promotion and unobserved index sensitivity, the latter of which enters the utility function nonlinearly.

Specifically, suppose the utility function can be written as

$$u_{ijt} = A_{ij} \times \left(X_{ijt}'\beta_0\right) + \epsilon_{ijt}, \tag{2.17}$$

where X_{ijt} contains Price, Promo, and Price×Promo, A_{ij} is the $ij-$specific fixed effect which may capture index sensitivity (which can be thought as inversely related to unobserved brand loyalty), and ϵ_{ijt} is the exogenous random shock. Suppose A_{ij} and Promo_{ijt} is positively correlated, which is reasonable because marketing managers with their expertise are more likely to promote products to which consumers are more price and promotion sensitive. Thus,

Table 2.8: Percentage of Correct Signs of Estimated Coefficients

α	$\hat{\beta}^{m}$	$\hat{\beta}^{CyclicMono}$	$\hat{\beta}^{OLS}$	$\hat{\beta}^{OLS-FE}$	$\hat{\beta}^{MLogit-FE}$
0.15	96%	0%	0%	0%	6%
0.30	97%	0%	0%	0%	0%
0.50	82%	0%	0%	0%	0%

traditional estimation methods that base on linearity would be unable to detect such pattern and wrongly attribute the effect on price elasticities from A_{ij} to Promo.

To provide some numerical evidence of the claim, we run the following Monte Carlo simulation. We let $\beta_0 = (-4, 2, 2)^{'}$, $Z \sim \mathcal{U}[0,1]$, $A_{ij} = Z + 1$, and $\epsilon_{ijt} \sim TIEV(0,1)$. For X_{ijt} vector, we draw $X_{ijt}^{(1)} \sim \mathcal{U}[0,4]$ and $W \sim \mathcal{U}[0,1]$, and let $X_{ijt}^{(2)} = (1-\alpha) \times W + \alpha \times Z$ and $X_{ijt}^{(3)} = X_{ijt}^{(1)} \times X_{ijt}^{(2)}$. We emphasize that $X_{ijt}^{(2)}$(Promo) is positively correlated with A_{ij} through Z, with α measuring the strength of the correlation. We consider three values of α: $0.15, 0.3$ and 0.5.

We run 100 simulations for each of the five methods in Table 2.7 to estimate β_0. To replicate the data structure of the empirical exercise, we set $N = 205$, $D = 3$, $J = 4$, and $T = 52$. We report in Table 2.8 the percentage of simulations that the corresponding method is able to generate correct signs for all coordinates of X_{ijt}.

The percentages that our proposed method is able to generate correct signs for all coordinates of X_{ijt} for $\alpha = 0.15,\ 0.3$, and 0.5 are 96%, 97%, and 82%, respectively. The accuracy of the estimator is negatively affected by the correlation between $X_{ijt}^{(2)}$ (Promo) and A_{ij} (multiplicative fixed effect). None of the other methods in Table 2.8 generates estimates of β_0 with correct signs. It is worth mentioning that the CM-based method requires A_{ij} entering the utility function linearly, which is violated in our DGP in (2.17). Apparently, all these other models than ours, due to their additive separable structure, completely ignore the positive dependence between the observable covariate $X_{ijt}^{(2)}$ (promotion) and the multiplicative fixed effect A_{ij}, thus producing biases in their estimates.

Intuitively, since products with larger A_{ij} are more likely to be promoted $\left(X_{ijt}^{(2)}=1\right)$ by the selection of marketing managers, the average effective price sensitivity of promoted products tend to be larger than those products not promoted. This drives those estimators that ignore such confounding selection effects to produce a negative coefficient on the interaction term $X_{ijt}^{(1)} \times X_{ijt}^{(2)}$ (Price × Promo), as found in the empirical illustration (Table 2.7). In contrast, our method handles such *non-additive* dependence between observable characteristics and unobserved fixed effects reasonably well, illustrating the robustness of our methods.

2.7 Monotone Multi-Index Models

We now present a general framework under which our identification strategy is applicable, using the notation of Ahn, Ichimura, Powell, and Ruud (2018, AIPR thereafter):

$$\gamma\left(\mathbf{X}_i\right)=\phi\left(\mathbf{X}_i\beta_0\right) \tag{2.18}$$

in which: $\left(y_i,\mathbf{X}_i\right)_{i=1}^N$ constitutes a random sample of N observations on a scalar[18] random variable y_i and a $J\times D$ random matrix $\mathbf{X}_i$. $\gamma\left(\overline{\mathbf{X}}\right)=\mathcal{T}\left(F_{y_i|\mathbf{X}_i=\overline{\mathbf{X}}}\left(\cdot\right)\right)$ is a real variable defined as a known functional $\mathcal{T}$ of the conditional distribution of y_i given $\mathbf{X}_i=\overline{\mathbf{X}}$. A leading example is to set $\gamma\left(\mathbf{X}_i\right):=\mathbb{E}\left[y_i|\,\mathbf{X}_i\right]$, so that model (2.18) becomes a conditional moment condition; however, this is not necessary. $\phi:\mathbb{R}^J\to\mathbb{R}$ is an unknown real-valued function. $\beta_0\in\mathbb{R}^D\backslash\left\{\mathbf{0}\right\}$ is the unknown finite-dimensional parameter of interest. Again, we normalize $\beta_0\in\mathbb{S}^{D-1}$, as β_0 is at best identified up to scale given that ϕ is an unknown function. As in Lee (1995), Powell and Ruud (2008) and AIPR, model (2.18) restricts the dependence of $\gamma\left(\mathbf{X}_i\right)$ on the matrix $\mathbf{X}_i$ to the J linear parametric indexes $\mathbf{X}_i\beta_0\equiv\left(X_{ij}'\beta_0\right)_{j=1}^J$.[19]

[18] Similar to AIPR, the dimension of y_i is largely irrelevant to the analysis of model (2.18): it is the dimension of γ that matters. Nevertheless, for the clarity of presentation, we take y_i to be a scalar.

[19] Note that model (2.18) is WLOG relative to the following seemingly more general formulation, in which β_0 is explicitly allowed to be heterogeneous across the J rows of $\mathbf{X}_i$: $\gamma\left(\mathbf{X}_i\right)=\phi\left(\left(X_{ij}'\beta_{0j}\right)_{j=1}^J\right)$, where

A noteworthy difference of model (2.18) from the setup in AIPR is that we take $\gamma\left(\mathbf{X}_i\right)$ here to be scalar-valued, while AIPR require their $\gamma\left(\mathbf{X}_i\right)$ to have dimension, using their notation R, no smaller than J. This "order condition" $R \geq J$ is necessary for their vector-valued function ϕ to admit a left-inverse ϕ^{-1} such that $\phi^{-1}\left(\gamma\left(\mathbf{X}_i\right)\right) = \mathbf{X}_i\beta_0$, which constitutes the foundation for their subsequent analysis. In contrast, we impose no such order condition for the sake of invertibility, as we will not rely on invertibility at all. Instead, we impose the following monotonicity assumption.

Assumption 2.7 (**Weak Monotonicity**). *ϕ is nondegenerate and nondecreasing in each of its J arguments on Supp$\left(\mathbf{X}_i\beta_0\right) \subseteq \mathbb{R}^J$.*

With no other restrictions besides Assumption 2.7 on the unknown function ϕ, model (2.18) builds in the fundamental lack of additive separability across the parametric indexes. As demonstrated in Section 2.2, the key idea developed below for the general multi-index model (2.18) naturally applies to the analysis of the panel multinomial choice model under complete lack of additive separability.

We now provide a few illustrative examples for model (2.18) that satisfy Assumption 2.7 beyond multinomial choice settings.

Example 2.1 (**Sample Selection Model**). Consider the sample selection model studied by Heckman (1979), where $y_i = y_i^* \cdot d_i$ with $y_i^* = W_i^{'}\mu_0 + u_i$ and $d_i = \mathbb{1}\left\{Z_i^{'}\lambda_0 + v_i \geq 0\right\}$. We observe (y_i, W_i, Z_i) but not y_i^*. Suppose $(u_i, v_i) \perp (X_i, Z_i)$ and the joint distribution of (u_i, v_i) is bivariate normal with a positive correlation. Then we have

$$\mathbb{E}\left[y_i \middle| W_i, d_i = 1\right] = X_i^{'}\mu_0 + \mathbb{E}\left[u_i \middle| v_i \geq -Z_i^{'}\lambda_0\right] =: \phi\left(W_i^{'}\mu_0, -Z_i^{'}\lambda_0\right).$$

By taking $X_i := (W_i, Z_i, d_i)$ and $\beta_0 := (\mu_0, \lambda_0)$, we may easily rewrite the model in the formulation of model (2.18) with Assumption 2.7 satisfied.

$\beta_0 := \left(\beta_{01}^{'}, ..., \beta_{0J}^{'}\right)^{'}$ is a $\sum_{j=1}^{J} D_j$-dimensional vector. This, however, could be readily incorporated in model (2.18) by appropriately redefining $\tilde{\mathbf{X}}_i$ to obtain the representation $\gamma\left(\tilde{\mathbf{X}}_i\right) = \phi\left(\tilde{\mathbf{X}}_i\beta_0\right)$ as in model (2.18).

Example 2.2 (Dyadic Network Formation Model under Nontransferable Utilities). Consider the following simple dyadic network formation model under nontransferable utilities (NTU):

$$D_{ij} = \mathbb{1}\left\{W_{ij}^{'}\mu_0 + Z_{ij}^{'}\gamma_0 \geq \epsilon_{ij}\right\}\mathbb{1}\left\{W_{ij}^{'}\mu_0 + Z_{ji}^{'}\gamma_0 \geq \epsilon_{ji}\right\}, \tag{2.19}$$

where $W_{ij} \equiv W_{ji}$ denotes some symmetric observable characteristics between a pair of individuals ij, (Z_{ij}, Z_{ji}) represent some asymmetric observable characteristics between ij, and $(\epsilon_{ij}, \epsilon_{ji})$ denote some potentially asymmetric idiosyncratic shocks to i's and j's utilities from linking with each other. The observed binary variable $D_{ij} \equiv D_{ji}$ of an undirected link between ij is determined jointly by two threshold-crossing conditions, interpreted as the requirement of mutual consent in the establishment of a link between ij. Clearly, we have

$$\mathbb{E}\left[D_{ij}\middle| W_{ij}, Z_{ij}, Z_{ji}\right] = \phi\left(W_{ij}^{'}\mu_0, Z_{ij}^{'}\gamma_0, Z_{ji}^{'}\gamma_0\right),$$

which falls under model (2.18) with Assumption 2.7 satisfied. It is worth noting that the NTU setting, which is a highly plausible feature in the formation of social networks, naturally induces lack of additive separability via the multiplication of two threshold-crossing conditions, even if we have a fully additive specification inside each threshold-crossing condition as in (2.19). Hence, the NTU setting provides a micro-founded motivation for confronting nonseparability, which our key method is well suited to deal with.

In a companion paper (Gao, Li, and Xu, 2020), we study a related but more complicated model of dyadic link formation with unobserved degree heterogeneity:

$$D_{ij} = \mathbb{1}\left\{u\left(W_{ij}^{'}\beta_0, A_i, A_j\right) \geq \epsilon_{ij}\right\}\mathbb{1}\left\{u\left(W_{ij}^{'}\beta_0, A_j, A_i\right) \geq \epsilon_{ji}\right\},$$

where A_i and A_j are scalar-valued individual "fixed effects" that represent each individual's unobserved heterogeneity in sociability. The involvement of the two-way fixed effects in this network formation setting adds further complications relative to the panel multinomial choice model considered in this paper, and we propose a new method, called *logical differencing*, to cancel out the two-way fixed effects, by constructing an observable event that contains the intersection of two mutually exclusive restrictions on the fixed effects. Nevertheless, the logical contraposition of multivariate monotonicity remains a convenient device for our identification arguments.

Proposition 2.2 (**General Identifying Restriction**). *Under model* (2.18) *with Assumption 2.7, for any* $\overline{\mathbf{X}}, \underline{\mathbf{X}} \in Supp(\mathbf{X}_i)$, $\gamma\left(\overline{\mathbf{X}}\right) > \gamma\left(\underline{\mathbf{X}}\right)$ *implies NOT* $\left(\left(\overline{X}_j - \underline{X}_j\right)\beta_0 \leq 0,\ \forall j = 1, ..., J\right)$.

Proposition 2.2 generalizes our key identification result (Theorem 2.1). Notice that Proposition 2.2 applies to all functionals γ on the conditional distribution $y_i|\,\mathbf{X}_i$ that satisfy the monotonicity assumption. Besides conditional expectations, there are many models where conditional quantiles or higher-order conditional moments are more natural choices of γ. In some cases where the whole conditional distribution $y_i|\,\mathbf{X}_i$ can be ranked by first-order or second-order stochastic dominance, we may aggregate the identifying information from many choices of γ into a joint restriction on β_0. We leave a further analysis of this topic to future research.

2.8 Conclusion

This paper proposes a simple and robust method for semiparametric identification and estimation in a panel multinomial choice model, exploiting the standard notion of multivariate monotonicity in an index vector of observable characteristics.

Our key identification strategy using logical contraposition of multivariate monotonicity is very simple, but it is exactly this conceptual simplicity that lends us the ability to accommodate infinite dimensionality of unobserved heterogeneity and lack of additive separability in consumer preferences. As the validity of this methodology essentially relies on nothing but monotonicity in a parametric index structure, it should be more widely applicable beyond the multinomial choice settings we consider.

However, a more comprehensive or in-depth investigation of whether and how this strategy can be adapted to the peculiarities of specific economic problems still requires a substantial amount of future work to be done. For applications in industrial organization, it might be worthwhile to inspect whether certain form of monotonicity can be preserved, at least approximately, in the presence of additional features, such as random coefficients and time-varying endogeneity, under certain conditions. In connection to microeconomic theory, it might also be interesting to investigate whether theoretical results on monotone comparative statics can be combined with our monotonicity-based method to provide a venue of identification and estimation in endogenous economic systems.

Furthermore, the asymptotic theory of the semiparametric estimator considered in this paper turns out to be interesting even in a binary choice model with point identification, as it features a nonstandard interplay between the nonsmooth sample criterion and the effective smoothing asymptotically provided by the first-stage estimator. Given that the asymptotic theory of such estimators is of independent interest and is better studied under different settings and notations, we refer interested readers to Gao and Xu (2020) for more details.

References

——— (2000): “Rank Estimation of a Generalized Fixed-Effects Regression Model,” *Journal of Econometrics*, 95, 1–23.

Ahn, H., H. Ichimura, J. L. Powell, and P. A. Ruud (2018): “Simple Estimators for Invertible Index Models,” *Journal of Business & Economic Statistics*, 36.

Berry, S., A. Gandhi, and P. Haile (2013): “Connected Substitutes and Invertibility of Demand,” *Econometrica*, 81, 2087–2111.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones (2016): "mlr: Machine Learning in R," *The Journal of Machine Learning Research*, 17, 5938–5942.

Browning, M. and J. Carro (2007): *Heterogeneity and Microeconometrics Modeling*, Cambridge University Press, vol. 3 of *Econometric Society Monographs*, 47–74.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Elsevier B.V., vol. 6B.

Chen, X. and H. White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45, 682–691.

Chernozhukov, V., I. Fernández-Val, and W. K. Newey (2019): "Nonseparable Multinomial Choice Models in Cross-Section and Panel Data," *Journal of econometrics*, 211, 104–116.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Fan, J., F. Han, and H. Liu (2014): "Challenges of Big Data Analysis," *National Science Review*, 1, 293–314.

Fox, J. T. (2007): "Semiparametric Estimation of Multinomial Discrete-Choice Models Using a Subset of Choices," *The RAND Journal of Economics*, 38, 1002–1019.

Gao, W. Y., M. Li, and S. Xu (2020): "Logical Differencing in Dyadic Network Formation Models with Nontransferable Utilities," Working Paper.

Gao, W. Y. and S. Xu (2020): "Two-Stage Maximum Score Estimator in Monotone Index Models," Working Paper.

Han, A. K. (1987): "Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303–316.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

——— (2001): "Micro data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748.

Honoré, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Honoré, B. E. and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogeneous Regressors," *Econometrica*, 70, 2053–2063.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for The Binary Response Model," *Econometrica: journal of the Econometric Society*, 505–531.

Howard, J. A. and J. N. Sheth (1969): *The Theory of Buyer Behavior*, Wiley.

Khan, S., F. Ouyang, and E. Tamer (2019): "Inference on Semiparametric Multinomial Response Models," Working Paper.

Lee, L.-F. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65, 381–428.

Li, L. (2019): "Identification of Structural and Counterfactual Parameters in a Large Class of Structural Econometric Models," Working Paper.

Luarn, P. and H.-H. Lin (2003): “A Customer Loyalty Model for E-Service Context,” *Journal of Electronic Commerce Research*, 4, 156–167.

Manski, C. F. (1975): “Maximum Score Estimation of the Stochastic Utility Model of Choice,” *Journal of Econometrics*, 3, 205–228.

——— (1985): “Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator,” *Journal of Econometrics*, 27, 313–333.

——— (1987): “Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data,” *Econometrica*, 55, 357–362.

Manski, C. F. and T. S. Thompson (1986): “Operational Characteristics of Maximum Score Estimation,” *Journal of Econometrics*, 32, 85–108.

McFadden, D. (1974a): “Conditional Logit Analysis of Qualitative Choice Behavior,” *Frontiers in Econometrics*, 105–142.

Newey, W. and D. McFadden (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, Elsevier, vol. IV, chap. 36.

Pakes, A. and J. Porter (2016): “Moment Inequalities for Multinomial Choice with Fixed Effects,” Working Paper, National Bureau of Economic Research.

Powell, J. L. and P. A. Ruud (2008): “Simple Estimators for Semiparametric Multinomial Choice Models,” *University of California, Berkeley.*

Reichheld, F. F. and P. Schefter (2000): “E-loyalty: Your Secret Weapon on The Web,” *Harvard Business Review*, 78, 105–113.

Shi, X., M. Shum, and W. Song (2018): “Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity,” *Econometrica*, 86, 737–761.

Van Der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Wasserman, L. (2006): *All of Nonparametric Statistics*, Springer Science & Business Media.

Appendix

2.A Proof of Theorem 2.2

We first prove two lemmas before formally proving Theorem 2.2.

Lemma 2.3. $Q : \mathbb{S}^{d-1} \to \mathbb{R}_+$ *is continuous.*

Proof. Recalling that $\cdot v_k\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) = \overline{X}_k - \underline{X}_k / \left\|\overline{X}_k - \underline{X}_k\right\|$ whenever $\overline{X}_k \neq \underline{X}_k$ while $v_k\left(\overline{\mathbf{X}} - \underline{\mathbf{X}}\right) = 0$ when $\overline{X}_k = \underline{X}_k$, we have

$$\begin{aligned} G\left(\gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\lambda_j\left(\mathbf{X}_{i,ts};\beta\right) =& G\left(\gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\prod_{k=1}^{J}\mathbb{1}\left\{(-1)^{\mathbb{1}\{k=j\}}\left(X_{ikt} - X_{iks}\right)'\beta \geq 0\right\} \\ =& G\left(\gamma_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\prod_{k=1}^{J}\mathbb{1}\left\{(-1)^{\mathbb{1}\{k=j\}}v_k\left(\mathbf{X}_{it} - \mathbf{X}_{is}\right)'\beta \geq 0\right\} \end{aligned}$$

which is continuous in β with probability one, since $v_k\left(\mathbf{X}_{it} - \mathbf{X}_{is}\right)$ has no mass point except possibly at $\mathbf{0}$, in which case the indicator degenerates to a constant over $\beta \in \mathbb{S}^{d-1}$. Since $\mathbf{X}_{i,ts}$ is i.i.d. across i, $\mathbb{S}^{d-1}$ is compact, and the indicator function is bounded, all conditions for Lemma 2.4 in Newey and McFadden (1994) are satisfied, by which we conclude that $Q = \sum_{j,t,s} Q_{j,t,s}$ is continuous on $\mathbb{S}^{d-1}$. □

Lemma 2.4. *Under Assumptions 2.2, 2.5 and 2.6,* $\sup_{\beta\in\mathbb{S}^{d-1}}\left|\hat{Q}\left(\beta\right) - Q\left(\beta\right)\right| = O_p\left(c_N\right).$

Proof. We first prove the convergence of $\hat{Q}_{j,t,s}(\beta)$ to $Q_{j,t,s}(\beta)$ for each (j,t,s). For each generic deterministic function $\tilde{\gamma}_{j,t,s}$, define

$$\begin{aligned}
Q_{j,t,s}(\beta,\tilde{\gamma}) &:= \mathbb{E}\left[G\left(\tilde{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\lambda_j\left(\mathbf{X}_{i,ts};\beta\right)\right], \\
\hat{Q}_{j,t,s}(\beta,\tilde{\gamma}) &:= \frac{1}{n}\sum_{i=1}^{n} G\left(\tilde{\gamma}_{j,t,s}\left(\mathbf{X}_{i,ts}\right)\right)\lambda_j\left(\mathbf{X}_{i,ts};\beta\right).
\end{aligned}$$

so that $\hat{Q}_{j,t,s}(\beta) = \hat{Q}_{j,t,s}(\beta,\tilde{\gamma}_{j,t,s})$ and $Q_{j,t,s}(\beta) = Q_{j,t,s}(\beta,\gamma)$. For notational simplicity we suppress the subscript (j,t,s) for the moment.

Defining $\mathcal{Q} := \left\{G\left(\tilde{\gamma}\left(\overline{\mathbf{X}}\right)\right)\lambda\left(\mathbf{X}_{i,ts};\beta\right) : \tilde{\gamma}\in\Gamma, \beta\in\mathbb{S}^{d-1}\right\}$, we first argue that $\mathcal{Q}$ is a $\mathbb{P}$-Donsker class based on Van Der Vaart and Wellner (1996). First, it is easy to show by Assumption 2.5 that $G(0)=0$, which together with the Lipschitz continuity of G, we have $\mathbb{E}\left[G^2\left(\tilde{\gamma}\left(\mathbf{X}_i\right)\right)\right] \leq M\mathbb{E}\left[\tilde{\gamma}^2\left(\mathbf{X}_i\right)\right] < \infty$ and $\mathbb{E}\left|G\left(\tilde{\gamma}\left(\mathbf{X}_i\right)\right)\right| \leq \mathbb{E}\left|\tilde{\gamma}\left(\mathbf{X}_i\right)\right| \leq \sup_{\tilde{\gamma}\in\Gamma}\mathbb{E}\left|\tilde{\gamma}\left(\mathbf{X}_i\right)\right| < \infty$. Then, as Γ is $\mathbb{P}$-Donsker, $G\circ\tilde{\gamma}$ must also be $\mathbb{P}$-Donsker. Second, recall that $\lambda\left(\mathbf{X}_{i,ts};\beta\right)$ is the product of indicators of half planes, while the collection of $\mathbb{1}\left\{\left(\overline{X}_k - \underline{X}_k\right)'\beta \geq 0\right\}$ over $\beta\in\mathbb{S}^{d-1}$ is a well-known VC Class of functions (sets) and is thus $\mathbb{P}$-Donsker. Finally, since the indicator function is uniformly bounded and $\sup_{\tilde{\gamma}\in\Gamma}\mathbb{E}\left|G\left(\tilde{\gamma}\left(\mathbf{X}_i\right)\right)\right| < \infty$, we conclude that $\mathcal{Q}$ is also $\mathbb{P}$-Donsker:

$$\sup_{\beta\in\mathbb{S}^{d-1}}\sup_{\tilde{\gamma}\in\Gamma}\left|\hat{Q}(\beta,\tilde{\gamma}) - Q(\beta,\tilde{\gamma})\right| = O_p\left(N^{-\frac{1}{2}}\right). \tag{2.20}$$

Next, by Assumption 2.4, we have

$$\begin{aligned}
\sup_{\beta\in\mathbb{S}^{d-1}}\left|Q(\beta,\hat{\gamma}) - Q(\beta,\gamma)\right| \leq& \sup_{\beta\in\mathbb{S}^{d-1}}\int\left|G\left(\hat{\gamma}\left(\overline{\mathbf{X}}\right)\right) - G\left(\gamma\left(\overline{\mathbf{X}}\right)\right)\right|\lambda_j\left(\overline{\mathbf{X}};\beta\right)d\mathbb{P}\left(\overline{\mathbf{X}}\right) \\
\leq& M\sqrt{\int\left(\hat{\gamma}\left(\overline{\mathbf{X}}\right) - \gamma\left(\overline{\mathbf{X}}\right)\right)^2 d\mathbb{P}\left(\overline{\mathbf{X}}\right)} = O_p\left(c_N\right) \qquad (2.21)
\end{aligned}$$

by Lipschitz continuity of G, $|\lambda_j| \leq 1$ and Cauchy-Schwarz. Combining (2.20) and (2.21), we have

$$\begin{aligned}\sup_{\beta \in \mathbb{S}^{d-1}} \left| \hat{Q}\left(\beta, \hat{\gamma}\right) - Q\left(\beta, \gamma\right) \right| \leq & \sup_{\beta \in \mathbb{S}^{d-1}} \sup_{\tilde{\gamma} \in \Gamma} \left| \hat{Q}\left(\beta, \tilde{\gamma}\right) - Q\left(\beta, \tilde{\gamma}\right) \right| + \sup_{\beta \in \mathbb{S}^{d-1}} \left| \hat{Q}\left(\beta, \hat{\gamma}\right) - Q\left(\beta, \hat{\gamma}\right) \right| \\ = & O_p\left(N^{-\frac{1}{2}}\right) + O_p\left(c_N\right) = O_p\left(c_N\right)\end{aligned}$$

since $N^{-\frac{1}{2}} = O_p\left(c_N\right)$ for nonparametric estimators. Summing over all (j, t, s), we have $\sup_{\beta \in \mathbb{S}^{d-1}} \left| \hat{Q}\left(\beta\right) - Q\left(\beta\right) \right| = O_p\left(c_N\right)$. □

Main Proof of Theorem 2.2

Proof. We verify Condition C.1 in Chernozhukov, Hong, and Tamer (2007, CHT thereafter) so as to apply their Theorem 3.1. Condition C.1(a) on the nonemptiness and compactness of parameter space is satisfied given Theorem 2.1. Condition C.1(b) on the continuity of the population criterion function Q is proved by Lemma 2.3. Condition C.1(c) on measurability of the sample criterion function is satisfied by its construction. Condition C.1(d)(e) regarding the uniform convergence of Q_n are satisfied by Lemma 2.4. Hence Theorem 3.1.(1) in CHT implies the Hausdorff consistency of $\hat{B}$. The consistency of the point estimator under the additional assumption of point identification (i.e., B_0 is a singleton) follows from Theorem 3.2 of CHT. □

2.B Pairwise Time Homogeneity of Errors

As mentioned in Section 2.2.2, Assumption 2.3 is stronger than necessary, and our identification strategy carries over under the weaker Assumption 2.3', which requires that $\boldsymbol{\epsilon}_{it} \sim \boldsymbol{\epsilon}_{is} | \left(\mathbf{X}_{i,ts}, \mathbf{A}_i\right)$. To see why Proposition 2.1 still holds, consider:

$$\mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i \right]$$

$$
\begin{aligned}
&= \int \mathbb{1}\left\{u\left(\overline{\delta}_j, A_{ij}, \epsilon_{ijt}\right) \geq \max_{k\neq j} u\left(\overline{\delta}_k,\ A_{ik}, \epsilon_{ikt}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{it} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&\quad - \int \mathbb{1}\left\{u\left(\underline{\delta}_j, A_{ij}, \epsilon_{ijs}\right) \geq \max_{k\neq j} u\left(\underline{\delta}_k,\ A_{ik}, \epsilon_{iks}\right)\right\} \mathrm{d}\mathbb{P}\left(\boldsymbol{\epsilon}_{is} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&= \int \mathbb{1}\left\{u\left(\overline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k\neq j} u\left(\overline{\delta}_k,\ A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \mathrm{d}\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}}_{i} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&\quad - \int \mathbb{1}\left\{u\left(\underline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k\neq j} u\left(\underline{\delta}_k,\ A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \mathrm{d}\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}}_{i} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right) \\
&= \int \left[\begin{array}{c} \mathbb{1}\left\{u\left(\overline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k\neq j} u\left(\overline{\delta}_k,\ A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \\ -\mathbb{1}\left\{u\left(\underline{\delta}_j, A_{ij}, \tilde{\epsilon}_{ij}\right) \geq \max_{k\neq j} u\left(\underline{\delta}_k,\ A_{ik}, \tilde{\epsilon}_{ik}\right)\right\} \end{array}\right] \mathrm{d}\mathbb{P}\left(\tilde{\boldsymbol{\epsilon}}_{i} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right)
\end{aligned}
$$

where $\overline{\boldsymbol{\delta}} = \overline{\mathbf{X}}\beta_0$, $\underline{\boldsymbol{\delta}} = \underline{\mathbf{X}}\beta_0$, and $\tilde{\boldsymbol{\epsilon}}_i$ denotes generic realizations of $\boldsymbol{\epsilon}_{it}$ and $\boldsymbol{\epsilon}_{is}$ conditional on $\mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)$ and $\mathbf{A}_i$. Notice that the second equality follows from the assumption that $\boldsymbol{\epsilon}_{it} \sim \boldsymbol{\epsilon}_{is} | \left(\mathbf{X}_{i,ts}, \mathbf{A}_i\right)$.

Again, if $\overline{\delta}_j \leq \underline{\delta}_j$ and $\overline{\delta}_k \geq \underline{\delta}_k$ for all $k \neq j$, the bracketed term in the last line of the displayed equation above must be nonpositive for all realizations of $\mathbf{A}_i$ and $\tilde{\boldsymbol{\epsilon}}_i$, so that $\mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{it} = \overline{\mathbf{X}}, \mathbf{X}_{is} = \underline{\mathbf{X}}, \mathbf{A}_i\right] \leq 0$ for all realizations of $\mathbf{A}_i$, which further implies that

$$
\gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) = \mathbb{E}\left[\mathbb{E}\left[y_{ijt} - y_{ijs} \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right), \mathbf{A}_i\right] \middle| \mathbf{X}_{i,ts} = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right] \leq 0.
$$

Taking the logical contraposition again gives Proposition 2.1.

2.C Sufficient Conditions for Point Identification

In this section, we prove sufficient conditions for the point identification of β_0. For simplicity of notation, we fix $T = 2$.We first need to impose an assumption of strict multivariate monotonicity on the function ψ_j defined in (2.5).

Assumption 2.8 (**Strict Monotonicity of $\boldsymbol{\psi_j}$**). *For any realized $\mathbf{A}_i$, the function $\psi_j\left(\,\cdot\,, \mathbf{A}_i\right) : \mathbb{R}^J \to \mathbb{R}$ is strictly increasing, i.e., if $\overline{\delta}_j > \underline{\delta}_j$ for all j, then $\psi\left(\overline{\boldsymbol{\delta}}, \mathbf{A}_i\right) > \psi\left(\underline{\boldsymbol{\delta}}, \mathbf{A}_i\right)$.*

We note that Assumption 2.8 is implied by a stronger version of Assumption 2.1 together with an additional condition on the support of u given $(\mathbf{X}_i, \mathbf{A}_i)$.

Assumption 2.8' (Strict Monotonicity of $\boldsymbol{u}$). *$u(\delta_{ijt}, A_{ij}, \epsilon_{ijt})$ is strictly increasing in the index δ_{ijt}, for every realization of (A_{ij}, ϵ_{ijt}).*

Assumption 2.8" (Overlapping Supports). *Conditional on any realization of $\mathbf{X}_i$ and $\mathbf{A}_i$, we have $\bigcap_{j=1}^{J} int\left(Supp\left(u\left(X'_{ijt}\beta_0, A_{ij}, \epsilon_{ijt}\right)\right)\right) \neq \emptyset$.*

In particular, Assumption 2.8" is directly implied by the assumption of $Supp\left(u\left(X'_{ijt}\beta_0, A_{ij}, \epsilon_{ijt}\right)\right) = \mathbb{R}$ conditional on any realization of $\mathbf{X}_i$ and $\mathbf{A}_i$, which is again satisfied in additive panel multinomial choice models with scalar fixed effects a la $u\left(X'_{ijt}\beta_0, A_{ij}, \epsilon_{ijt}\right) = X'_{ijt}\beta_0 + A_{ij} + \epsilon_{ijt}$ under the assumption of $Supp\left(\epsilon_{ijt} \middle| \mathbf{X}_i, \mathbf{A}_i\right) = \mathbb{R}$ as commonly imposed in the literature.

Lemma 2.5. *Assumptions 2.8' and 2.8" imply Assumption 2.8.*

Finally, we impose the following assumption on $\Delta\mathbf{X}_i$, with $\Delta X_{ij} := X_{ij1} - X_{ij2}$ for all individual i and product j across period 1 and period 2.

Assumption 2.9 (Full-Directional Support of $\Delta\mathbf{X}_i$). *Suppose either (a) or (b) is true:*

(a) $\mathbf{0} \in int\left(Supp\left(\Delta\mathbf{X}_i\right)\right)$.

(b) There exists some $k \in \{1, ..., d_x\}$ such that $\beta_0^k \neq 0$ and $Supp\left(\Delta X_{ij}^k \middle| \Delta X_{il}, \, l \neq j\right) = \mathbb{R}$ for all $j \in \{1, ..., J\}$. Furthermore, for all $j \in \{1, ..., J\}$, $Supp\left(\Delta X_{ij} \middle| \Delta X_{il}, \, l \neq j\right)$ is not contained in a proper linear subspace of $\mathbb{R}^{d_x}$.

Assumption 2.9(a) is satisfied when (X_{ij}) is continuous random vector. On the other hand, Assumption 2.9(b) can accommodate discrete regressors generally, but requires one continuous covariate with large support. Assumption 2.9 ensures that $\Delta X'_{ij}\beta_0 > 0$ and $\Delta X'_{ik}\beta_0 < 0$ for all $k \neq j$ hold simultaneously with strictly positive probability.

Theorem 2.3 (Point Identification). *Under Assumptions 2.2, 2.3, 2.8 and 2.9, β_0 is point identified on $\mathbb{S}^{D-1}$.*

Proof. Recall first that

$$\gamma_j\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) = \int \left[\psi_j\left(\overline{\delta}_j, \left(-\overline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) - \psi_j\left(\underline{\delta}_j, \left(-\underline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right)\right] \mathrm{d}\mathbb{P}\left(\mathbf{A}_i \middle| \mathbf{X}_i = \left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right)\right).$$

Hence, under Assumption 2.8, we have

$$\overline{\delta}_j < \underline{\delta}_j \text{ and } \overline{\delta}_k > \underline{\delta}_k \text{ for all } k \neq j \quad \Rightarrow \quad \gamma_{j,t,s}\left(\overline{\mathbf{X}}, \underline{\mathbf{X}}\right) > 0, \tag{2.22}$$

since $\psi_j\left(\overline{\delta}_j, \left(-\overline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right) < \psi_j\left(\underline{\delta}_j, \left(-\underline{\delta}_k\right)_{k\neq j}, \mathbf{A}_i\right)$ for every realization of $\mathbf{A}_i$. Together with Assumption 2.9, we deduce that

$$\mathbb{P}\left\{\gamma_{j,t,s}\left(\mathbf{X}_i\right) > 0\right\} \geq \mathbb{P}\left\{\Delta X_{ij}^{'}\beta_0 > 0 \ \wedge \ \Delta X_{ik}^{'}\beta_0 < 0, \forall k \neq j\right\} > 0.$$

Now for any $\beta \in \mathbb{S}^{D-1} \backslash \left\{\beta_0\right\}$, define for any product j,

$$H_j\left(\beta\right) := \left\{\mathbf{v} \in Supp\left(\Delta\mathbf{X}_i\right) : \, v_j^{'}\beta < 0 < v_j^{'}\beta_0, \ \wedge \ v_k^{'}\beta_0 < 0 < v_k^{'}\beta, \forall k \neq j\right\}.$$

As $\beta \neq \beta_0$, by Assumption 2.9 we know that

$$\mathbb{P}\left(\Delta\mathbf{X}_i \in H_j\left(\beta\right)\right) > 0. \tag{2.23}$$

Moreover, for any realization of $\mathbf{X}_i$ s.t. $\Delta\mathbf{X}_i \in H_j\left(\beta\right)$, we must have: (i) $\gamma_{j,t,s}\left(\mathbf{X}_i\right) > 0$ by (2.22), and (ii):

$$\lambda_j\left(\Delta\mathbf{X}_i, \beta\right) = \prod_{k=1}^{J} \mathbb{1}\left\{(-1)^{\mathbb{1}\{k=j\}} \Delta X_{ik}^{'}\beta \geq 0\right\} = 1$$

so that $G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\lambda_j\left(\Delta\mathbf{X}_i, \beta\right) = G\left(\gamma_j\left(\mathbf{X}_i\right)\right) > 0$ for all such $\mathbf{X}_i$. Hence,

$$\mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right) \middle| \Delta\mathbf{X}_i \in H_j\left(\beta\right)\right] > 0. \tag{2.24}$$

Combining (2.23) and (2.24), we have:

$$\begin{aligned}
Q_j(\beta) &= \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\lambda_j\left(\Delta\mathbf{X}_i,\beta\right)\right] \\
&\geq \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\lambda_j\left(\Delta\mathbf{X}_i,\beta\right)\mathbb{1}\left\{\Delta\mathbf{X}_i \in H_j(\beta)\right\}\right] \\
&= \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\mathbb{1}\left\{\Delta\mathbf{X}_i \in H_j(\beta)\right\}\right] \\
&= \mathbb{E}\left[G\left(\gamma_j\left(\mathbf{X}_i\right)\right)\middle|\,\Delta\mathbf{X}_i \in H_j(\beta)\right]\mathbb{P}\left(\Delta\mathbf{X}_i \in H_j(\beta)\right) \\
&> 0 = Q_j\left(\beta_0\right).
\end{aligned}$$

□

2.D Additional Simulation Results

2.D.1 Adaptive-Grid Computation Algorithm

In this section, we illustrate a typical output of our second-step computation algorithm based on the adaptive-grid search over the angle space, and show that the algorithm works well. For this purpose we consider a simplified DGP without fixed effect A_{ij}. We draw each of $X_{ijt}^{(d)}$ independently across each dimension $d \in \{1, ..., D\}$ from the standard normal distribution, and set the distribution of the idiosyncratic shock to be $\epsilon_{ijt} \sim TIEV\,(0,1)$, so that we can skip the first-step estimation and directly calculate the true conditional choice probability conditioned on each $\mathbf{X}_i$. Note that the conditions for point identification of β_0 are satisfied. Because we are only seeking to illustrate the validity of the algorithm itself, we set N to be large with $N = 10^7$ and $D = 3, J = 3, T = 2$. Then we apply our adaptive-grid algorithm to search for β_0.

Figure 2.2 shows how our computational algorithm works in finding the true unknown θ_0, the angle representation of the true β_0 in the Θ space. The horizontal and vertical axes correspond to the two polar coordinates that are associated with $\mathbb{S}^2$. The blue dots represent the points that our algorithm searches over but find *not* to be minimizers of the

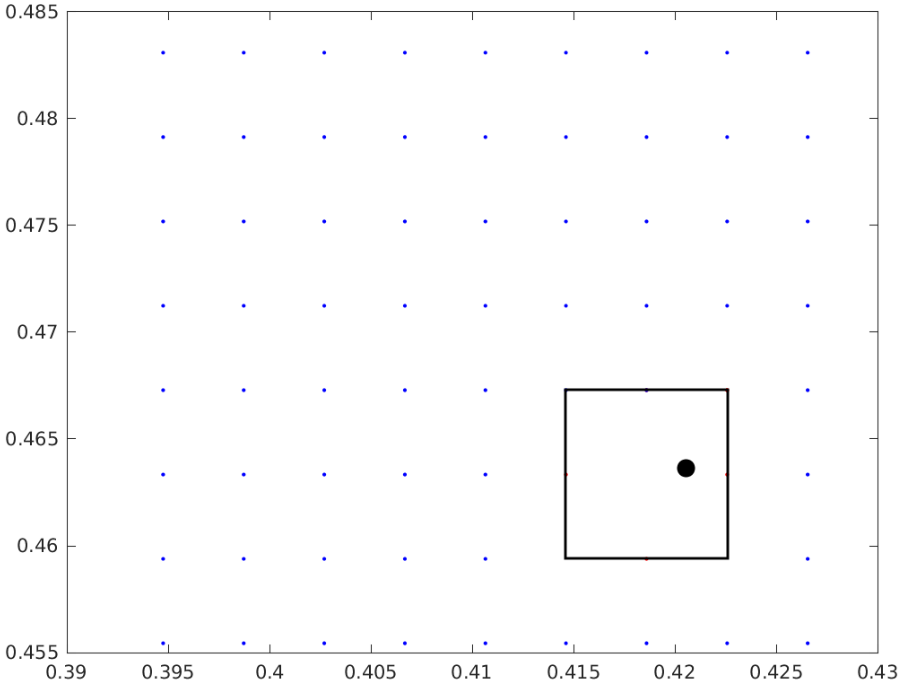


Figure 2.2: The Argmin Set in Θ

sample criterion $\hat{Q}$. The black box indicates the area that the minimizers for the sample criterion $\hat{Q}$ lie within, or more precisely, a rectangular enclosure of the numerical argmin set. The big black dot stands for the true parameter value $\theta_0 = (0.4205, 0.4636)'$.

It is evident from Figure 2.2 that our adaptive-grid algorithm is able to correctly locate an area that covers the true θ_0, which lies within the small black box representing the estimated set of $\hat{\theta}$, demonstrating the efficacy of the algorithm. Besides, it is worth mentioning that our algorithm computes reasonably fast, as it first performs a rough search on the whole unit sphere $\mathbb{S}^2$, then focuses on the area where the minimizers are most likely to lie. In the last few rounds of search, the algorithm evaluates the criterion function $\hat{Q}$ on a relatively small area of points shown by those blue and red dots in Figure 2.2 until the desired level of accuracy is achieved.

For a more transparent representation, we translate the angles θ in the polar coordinates into unit vectors β on the unit sphere $\mathbb{S}^2$ and show it in Figure 2.3.

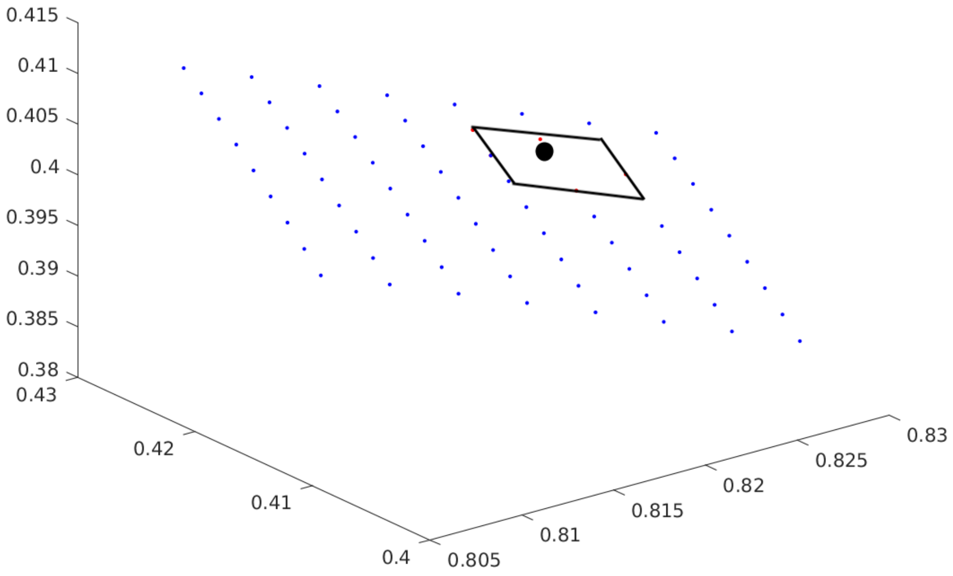


Figure 2.3: The Argmin Set in $\mathbb{S}^2$

Figure 2.3 is now plotted on $\mathbb{S}^2 \subseteq \mathbb{R}^3$. Again the blue dots represent the points that do not achieve the minimum of $\hat{Q}$; the black box shows an enclosing set of the minimizers of $\hat{Q}$. The big black dot represents the true parameter value β_0, which resides inside the black box of the minimizers of $\hat{Q}$. Figure 2.3 illustrates that our computation algorithm is able to locate a tight area around β_0.

2.D.2 Results Varying D, J, T

In this section, we show how our estimator performs under different (D, J, T). We maintain $N = 10,000$ as in the baseline configuration. We draw $Z_i \sim \mathcal{N}(0,1)$ and construct A and X according to the following specifications:

$$A_{ij} \sim \begin{cases} 0, & j = 1, \\ [Z_i]_+, & j = 2, \\ \mathcal{U}[-0.25, 0.25], & j = 3, ..., J, \end{cases} \quad X_{ijt}^{(d)} \sim \begin{cases} U[-1,1], & d = 1, \\ Z_i + \mathcal{N}(0,6), & d = 2, \\ \mathcal{N}(0,1), & d = 3, ..., D, \end{cases}$$

which coincides with the baseline model at $D = 3, J = 3$. We emphasize that in all configurations we allow for nonlinear dependence between A and X via the latent variable Z_i.

We report in Table 2.9 the performance of our estimators for each of the corresponding configurations across all $M = 100$ simulations.

Table 2.9: Performance Varying D, J, T

rMSE	$J = 3$		$J = 4$	
	$T = 2$	$T = 4$	$T = 2$	$T = 4$
$D = 3$	0.0745	0.0397	0.1137	0.0722
$D = 4$	0.0945	0.0580	0.1357	0.0807
MND	$J = 3$		$J = 4$	
	$T = 2$	$T = 4$	$T = 2$	$T = 4$
$D = 3$	0.0648	0.0348	0.1005	0.0639
$D = 4$	0.0864	0.0539	0.1233	0.0750

From Table 2.9 we find a larger T improves the performance of our estimator, which is arguably more practically relevant given the increasing availability of long panel data nowadays. The improvement in performance with larger T is because our method can extract more information from $T \times (T - 1)$ ordered pairs of time periods which effectively increase the total number of observations. We also find that increase in D or J adversely affects the performance of our estimator, which is expected because more information is required to estimate more covariates (D) or deal with more alternatives (J). However, as can be seen from Table 2.9, the magnitude of such decline in performance is mild. For example, when J is 4 and T is 4, an increase in the dimension of product characteristics D from 3 to 4 will increase the rMSE from 0.0722 to 0.0807. Likewise, when $D = 4$ and $T = 4$, an increase in J from 3 to 4 will increase the rMSE from 0.0580 to 0.0807.

Chapter 3

Logical Differencing in Dyadic Network Formation Models with Nontransferable Utilities[1]

3.1 Introduction

This paper considers a semiparametric model of dyadic network formation under *nontransferable utilities* (NTU), which arise naturally in the modeling of real-world social interactions that require bilateral consent. For instance, friendship is usually formed only when both individuals in question are willing to accept each other as a friend, or in other words, when both individuals derive sufficiently high utilities from establishing the friendship. It is often plausible that the two individuals may derive very different utilities from the friendship for a variety of reasons: for example, one of them may simply be more introvert than the other and derive lower utilities from the friendship. In addition, there may not be a feasible way to perfectly transfer utilities between the two individuals. Monetary payments may not be customary in many social contexts, and even in the presence of monetary or in-kind

[1] Joint with Wayne Gao and Sheng Xu.

transfers, *utilities* may not be perfectly transferable through these feasible forms of transfers, say, when individuals have different marginal utilities with respect to these transfers.[2] Given the considerable academic and policy interest in understanding the underlying drivers of network formation,[3] it is not only theoretically interesting but also empirically relevant to incorporate NTU in the modeling of network formation.

This paper contributes to the line of econometric literature on network formation by introducing and incorporating *nontransferable utilities* into dyadic network formation models.[4] Previous work in this line of literature focuses primarily on case of *transferable utilities*, as represented in Graham (2017a), which considers a parametric model with homophily effects and individual unobserved heterogeneity of the following form:

$$D_{ij} = \mathbb{1}\left\{w\left(X_i, X_j\right)^{\prime} \beta_0 + A_i + A_j \geq U_{ij}\right\} \tag{3.1}$$

where D_{ij} is an observable binary variable that denotes the presence or absence of a link between individual i and j, $w\left(X_i, X_j\right)$ represents a (symmetric) vector of pairwise observable characteristics specific to ij generated by a known function w of the individual observable characteristics X_i and X_j of i and j, while A_i and A_j stand for unobserved individual-specific degree heterogeneity and U_{ij} is some idiosyncratic utility shock. Model (3.1) essentially says that, if the (stochastic) *joint surplus* generated by a bilateral link $s_{ij} := w\left(X_i, X_j\right)^{\prime} \beta_0 + A_i + A_j - U_{ij}$ exceeds the threshold zero, then the link between i and j is formed. The model implicitly assumes that the link surplus can be freely distributed among the two individuals i

[2] See surveys by Aumann (1967), Hart (1985) and McLean (2002) for discussions on the implications of NTU on link (bilateral relationship) and group formation from a micro-theoretical perspective.

[3] For example, the formation of friendship among U.S. high-school students has been studied by a long line of literature, such as Moody (2001), Currarini, Jackson, and Pin (2009, 2010), Boucher (2015), Currarini et al. (2016), Xu and Fan (2018) among others.

[4] It should be pointed out that the line of econometric literature on *strategic network formation* models, which primarily uses pairwise stability (Jackson and Wolinsky, 1996a) as the solution concept for network formation, often builds NTU (along with link interdependence) into the econometric specification from scratch. See, for example, De Paula, Richards-Shubik, and Tamer (2018b), Graham (2016a), Leung (2015a), Menzel (2015b), Mele (2017a), Mele (2017c) and Ridder and Sheng (2017a). This paper does not belong to that line of literature but instead contributes to the line of econometric literature on dyadic network formation models, which abstracts away from link interdependence but usually incorporates more flexible forms of unobserved individual heterogeneity.

and j, and that bargaining efficiency is always achieved, so that the undirected link is formed if and only if the link surplus is positive. Given this specification, Graham (2017a) provides consistent and asymptotically normal maximum-likelihood estimates for the homophily effect parameters β_0, assuming that the exogenous idiosyncratic pairwise shocks U_{ij} are independently and identically distributed with a logistic distribution. Recently, Candelaria (2016) and Toth (2017) provide semiparametric generalizations of Graham (2017a), while Gao (2020) established nonparametric identification of a class of index models that further generalize (3.1).

This paper, however, generalizes Graham (2017a) along a different direction, and seeks to incorporate the natural micro-theoretical feature of NTU into this class of network formation models. To illustrate[5], consider the following simple adaption of model (3.1) with two threshold-crossing conditions:

$$D_{ij} = \mathbb{1}\left\{w\left(X_i, X_j\right)^{'}\beta_0 + A_i \geq U_{ij}\right\} \cdot \mathbb{1}\left\{w\left(X_i, X_j\right)^{'}\beta_0 + A_j \geq U_{ji}\right\}, \tag{3.2}$$

where the unobserved individual heterogeneity A_i and A_j *separately* enter into two different threshold-crossing conditions. This formulation could be relevant to scenarios where A_i represents individual i's own intrinsic valuation of a generic friend: for a relatively shy or introvert person i, a lowerA_i implies that i is less willing to establish a friendship link, regardless of how sociable the counterparty is. For simplicity, suppose for now that $w\left(X_i, X_j\right) \equiv \mathbf{0}$ and $U_{ij} \sim_{iid} U_{ji} \sim F$. Focusing completely on the effects of A_i and A_j, it is clear that the TU model (3.1) implies that only the sum of "sociability", $A_i + A_j$, matters: the linking probability among pairs with $A_i = A_j = 1$ (two moderately social persons) should be exactly the same as the linking probability among pairs with $A_i = 2$ and $A_j = 0$ (one very social person and one very shy person), which might not be reasonable or realistic in social scenarios. In comparison, the linking probability among pairs with $A_i = 2$ and $A_j = 0$

[5] Starting from Section 3.2, we consider a more general specification than the illustrative model (3.1) introduced here.

is lower than the linking probability among pairs with $A_i = A_j = 1$ under the NTU model (3.2) with i.i.d. U_{ij} and U_{ji} that follow any log-concave distribution[6]:

$$\begin{aligned} & \mathbb{E}\left[D_{ij} \middle| w\left(X_i, X_j\right) \equiv \mathbf{0}, A_i = 2, A_j = 0 \right] \\ = & F\left(0\right) F\left(2\right) \\ < & F\left(1\right) F\left(1\right) \\ = & \mathbb{E}\left[D_{ij} \middle| w\left(X_i, X_j\right) \equiv \mathbf{0}, A_i = A_j = 1 \right] \end{aligned}$$

This is intuitive given the observation that, under bilateral consent, the party with relatively lower utility is the pivotal one in link formation. Moreover, even though we maintain strict monotonicity in the unobservable characteristics A_i and A_j, the NTU setting can still effectively incorporate homophily effects on unobserved heterogeneity: given that $w\left(X_i, X_j\right) \equiv \mathbf{0}$ and $A_i + A_j = 2$, the linking probability is effectively decreasing in $\left|A_i - A_j\right|$ under log-concave F. Hence, by explicitly modeling NTU in dyadic network formation, we can accommodate more flexible or realistic patterns of conditional linking probabilities and homophily effects that are not present under the TU setting.

However, the NTU setting immediately induces a key technical complication: as can be seen explicitly in model (3.2), the observable indexes ($W_{ij}'\beta_0$ and $W_{ji}'\beta_0$) and the unobserved heterogeneity terms (A_i and A_j) are no longer additively separable from each other. In particular, notice that, even though the utility specification for each individual inside each of the two threshold-crossing conditions in model (3.2) remains completely linear and additive, the multiplication of the two (nonlinear) indicator functions directly destroys both linearity and additive separability, rendering inapplicable most previously developed econometric

[6] A distribution is log-concave if $F\left(x\right)^{\lambda} F\left(y\right)^{1-\lambda} \leq F\left(\lambda x + \left(1-\lambda\right) y\right)$. Many commonly used distributions, such as uniform, normal, exponential, logistic, chi-squared distributions, are log-concave. See Bagnoli and Bergstrom (2005) for more details on log-concave distributions from a microeconomic theoretical perspective.

techniques that arithmetically "difference out" the "two-way fixed effects" A_i and A_j based on additive separability.[7]

Given this technical challenge, this paper proposes a new identification strategy termed *logical differencing*, which helps cancel out the unobserved heterogeneity terms, A_i and A_j, without requiring additive separability but leveraging the logical implications of *multivariate monotonicity* in model (3.2). The key idea is to construct an observable event involving the intersection of two mutually exclusive restrictions on the fixed effects A_i and A_j, which logically imply an event that can be represented without A_i or A_j. Specifically, in the context of the illustrative model (3.2) above, we start by considering the event where a given individual $\bar{i}$ is *more popular* than another individual $\bar{j}$ among a group of individuals k with observable characteristics $X_k = \overline{x}$ while $\bar{i}$ is simultaneously *less popular* than another individual $\bar{j}$ among a group of individuals with a certain realization of observable characteristics $\underline{x}$. This is the same as the conditioning event in Toth (2017) and analogous to the tetrad comparisons made in Candelaria (2016). However, instead of using arithmetic differencing to cancel out the unobserved heterogeneity $A_{\bar{i}}$ and $A_{\bar{j}}$ as in Candelaria (2016) and Toth (2017), we make the following logical deductions based on the monotonicity of the conditional popularity of $\bar{i}$ in $w\left(X_{\bar{i}}, \overline{x}\right)' \beta_0$ and $A_{\bar{i}}$. First, the event that $\bar{i}$ is *more popular* than another individual $\bar{j}$ among the group of individuals with $X_k = \overline{x}$ implies that either $w\left(X_{\bar{i}}, \overline{x}\right)' \beta_0 > w\left(X_{\bar{j}}, \overline{x}\right)' \beta_0$ or $A_{\bar{i}} > A_{\bar{j}}$, while the event that $\bar{i}$ is *less popular* than another individual $\bar{j}$ among a different group of individuals with $X_l = \underline{x}$ implies that either $w\left(X_{\bar{i}}, \underline{x}\right)' \beta_0 < w\left(X_{\bar{j}}, \underline{x}\right)' \beta_0$ or $A_{\bar{i}} < A_{\bar{j}}$. Second, when both events occur simultaneously, we can logically deduce that either $w\left(X_{\bar{i}}, \overline{x}\right)' \beta_0 > w\left(X_{\bar{j}}, \overline{x}\right)' \beta_0$ or $w\left(X_{\bar{i}}, \underline{x}\right)' \beta_0 < w\left(X_{\bar{j}}, \underline{x}\right)' \beta_0$ must have occurred, because $A_{\bar{i}} > A_{\bar{j}}$ and $A_{\bar{i}} < A_{\bar{j}}$ cannot simultaneously occur. Intuitively, the "switch" in the relative popularity of $\bar{i}$ and $\bar{j}$ among the two groups of individuals with

[7] Equivalently, one could write model (3.2) in an alternative form as a "single" *composite* threshold-crossing condition:

$$D_{ij} = \mathbb{1}\left\{\min\left\{W_{ij}^{'}\beta_0 + A_i - U_{ij}, W_{ji}^{'}\beta_0 + A_j - U_{ji}\right\} \geq 0\right\},$$

where additive separability is again lost in this alternative formulation.

characteristics $\overline{x}$ and $\underline{x}$ cannot be driven by individual unobserved heterogeneity $A_{\bar{i}}$ and $A_{\bar{j}}$, and hence when we indeed observe such a "switch", we obtain a restriction on the parametric indices $w\left(X_{\bar{i}},\overline{x}\right)'\beta_0$, $w\left(X_{\bar{i}},\overline{x}\right)'\beta_0$, $w\left(X_{\bar{i}},\underline{x}\right)'\beta_0$, and $w\left(X_{\bar{j}},\underline{x}\right)'\beta_0$, which helps identify β_0.

Based on this identification strategy we provide sufficient conditions for point identification of the parameter β_0 up to scale normalization as well as a consistent estimator for β_0. Our estimator has a two-step structure, with the first step being a standard nonparametric estimator of conditional linking probabilities, which we use to assert the occurrence of the conditioning event, while in the second step we use the identifying restriction on β_0 when the conditioning event occurs. The computation of the estimator essentially follows the same method proposed in (Gao and Li, 2020), with some adaptions to the network data setting. We analyze the finite-sample performance in a simulation study, and present an empirical illustration of our method using data from Nyakatoke on risk-sharing network collected by Joachim De Weerdt.

This paper belongs to the line of literature that studies dyadic network formation in a single large network setting, including Blitzstein and Diaconis (2011a), Chatterjee, Diaconis, and Sly (2011a), Yan and Xu (2013a), Yan, Leng, and Zhu (2016), Graham (2017a), Charbonneau (2017), Dzemski (2017a), Jochmans (2017a), Yan, Jiang, Fienberg, and Leng (2018b), Candelaria (2016), Toth (2017) and Gao (2020). According to our knowledge Shi and Chen (2016) is the only previous paper that explicitly incorporates NTU into dyadic network formation models, but Shi and Chen (2016) considers a fully parametric model and establishes the consistency and asymptotic normality of the maximum likelihood estimators. In contrast, we consider a semiparametric model here where the functional form of the idiosyncratic shock distribution is left unrestricted.

This paper is also related to a line of research that utilizes the form of link formation models considered here in order to study structural social interaction models: for instance, Arduini et al. (2015a), Auerbach (2016a), Goldsmith-Pinkham and Imbens (2013a), Hsieh and Lee (2016a) and Johnsson and Moon (2017). In these papers, the social interaction models are the main focus of identification and estimation, while the link formation models are used mainly as a tool (a control function) to deal with network endogeneity or unobserved heterogeneity problems in the social interaction model. Even though some of the network formation models considered in this line of literature is consistent with the NTU setting, this line of literature is usually not primarily concerned with the full identification and estimation of the network formation model itself.

It should be pointed out that in this paper we do not consider link interdependence in network formation. See Graham (2015a), Chandrasekhar (2016a) and de Paula (2016a) for reviews on the econometric literature on strategic network formation with link interdependence.

This paper is also a companion paper to Gao and Li (2020), which similarly leverages multivariate monotonicity in a multi-index structure under a panel multinomial choice setting, which incorporate rich individual-product specific unobserved heterogeneity in the form of an infinite-dimensional fixed effect that enters into individual's utility functions in an additively nonseparable way. The structural similarity between network data and panel data has long been noted in the econometric literature, but it should also be pointed out that the network structure considered in this paper is technically more complicated than the panel structure, as there are no direct ways in the network setting to make "intertemporal comparison" as in the panel setting that holds the fixed effects unchanged across two observable periods of time. It is precisely this additional complication induced by the network setting that requires the technique of logical differencing proposed in this paper.

The rest of the paper is organized as follows. In Section 3.2, we describe the general specifications of the dyadic network formation model we consider. Section 3.3 establishes identification of the parameter of interests in our model, and also provides a consistent tetrad estimator. Simulation results are reported in Section 3.4. We present an empirical illustration of our method using the risk-sharing data of Nyakatoke in Section 3.5. We conclude with Section 3.6. Proofs are available in the Appendix.

3.2 A Nonseparable Dyadic Network Formation Model

We consider the following dyadic network formation model:

$$\mathbb{E}\left[D_{ij}\middle| X_i, X_j, A_i, A_j\right] = \phi\left(w\left(X_i, X_j\right)' \beta_0, A_i, A_j\right) \tag{3.3}$$

where:

- $i \in \{1, ..., n\}$ denote a generic individual in a group of n individuals.
- X_i is a $\mathbb{R}^{d_x}$-valued vector of observable characteristics for individual i. This could include wealth, age, education and ethnicity of individual i.
- D_{ij} denotes a binary observable variable that indicates the presence or absence of an undirected and unweighted link link between two distinct individuals i and j: $D_{ij} = D_{ji}$ for all pairs of individuals ij, with $D_{ij} = 1$ indicating that ij are linked while $D_{ij} = 0$ indicating that ij are not linked.
- $w : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}^{d_w}$ is a known function that is *symmetric*[8] with respect to its two vector arguments.
- $\beta_0 \in \mathbb{R}^{d_\beta}$ is an unknown finite-dimensional parameter of interest. Assume $\beta_0 \neq \mathbf{0}$ so that we may normalize $\|\beta_0\| = 1$, i.e., $\beta_0 \in \mathbb{S}^{d_\beta - 1}$.
- A_i is an unobserved scalar-valued variable that represents unobserved individual heterogeneity.

[8] Our method can also be adapted to the case with *asymmetric* w. See Remark 3.1.

- $\phi : \mathbb{R}^3 \to \mathbb{R}$ is an unknown measurable function that is symmetric with respect to its second and third arguments.

In addition, we impose the following two assumptions:

Assumption 3.1 (**Monotonicity**). *ϕ is weakly increasing in each of its arguments.*

Assumption 3.1 is the key assumption on which our identification analysis is based, which requires that the conditional linking probability between individuals with characteristics (X_i, A_i) and (X_j, A_j) be monotone in a parametric index $\delta_{ij} := w\left(X_i, X_j\right)^{'} \beta_0$ as well as the unobserved individual heterogeneity terms A_i and A_j. It should be noted that, given monotonicity, increasingness is without loss of generality as ϕ, β_0 and A_i, A_j are all unknown or unobservable. Also, Assumption 3.1 is only requiring that ϕ is monotonic in the index $w\left(X_i, X_j\right)^{'} \beta_0$ as a whole, not individual components of $w\left(X_i, X_j\right)$. Therefore, we may include nonlinear or non-monotone functions $w\left(\cdot, \cdot\right)$ on the observable characteristics as long as Assumption 3.1 is maintained.

Next, we also impose a standard random sampling assumption:

Assumption 3.2 (**Random Sampling**). *(X_i, A_i) is i.i.d. across $i \in \{1, ..., n\}$.*

In particular, Assumption 3.2 allows arbitrary dependence structures between the observable characteristics X_i and the unobservable characteristic A_i.

Model (3.3) along with the specifications and the two assumptions introduced above encompass a large class of dyadic network formation models in the literature. For example, the standard dyadic network formation model (3.1) studied by Graham (2017a) can be written as

$$\mathbb{E}\left[D_{ij} \middle| X_i, X_j, A_i, A_j\right] = F\left(w\left(X_i, X_j\right)^{'} \beta_0 + A_i + A_j\right)$$

where F is the CDF of the standard logistic distribution. For the semiparametric version considered by Candelaria (2016); Toth (2017); Gao (2020), we can simply take F to be some unknown CDF. In either case, the monotonicity of the CDF F and the additive structure of $w\left(X_i, X_j\right)^{\prime} \beta_0 + A_i + A_j$ immediately imply Assumption 3.1.

However, our current model specification and assumptions further incorporate a larger class of dyadic network formation models with potentially nontransferable utilities. Specifically, consider the joint requirement of two threshold-crossing conditions,

$$D_{ij} = \mathbb{1}\left\{u\left(w\left(X_i, X_j\right)^{\prime} \beta_0, A_i, A_j, \epsilon_{ij}\right) \geq 0\right\} \cdot \mathbb{1}\left\{u\left(w\left(X_j, X_i\right)^{\prime} \beta_0, A_j, A_i, \epsilon_{ji}\right) \geq 0\right\} \tag{3.4}$$

where u is an unknown function that is not necessarily symmetric with respect to its second and third arguments (A_i, A_j), and $(\epsilon_{ij}, \epsilon_{ji})$ are idiosyncratic pairwise shocks that are i.i.d. across the each unordered ij pair with some unknown distribution. In particular, notice that model (3.2) is a special case of (3.4). Suppose we further impose the following two lower-level assumptions Assumption 3.1a and 3.1b:

Assumption 3.1a. $(\epsilon_{ij}, \epsilon_{ji})$ *are independent from* (X_i, A_i, X_j, A_j).

Assumption 3.1b. u *is weakly increasing in its first three arguments.*

Then, the conditional linking probability

$$\begin{aligned}
&\mathbb{E}\left[D_{ij} \mid X_i, X_j, A_i, A_j\right] \\
&= \int \mathbb{1}\left\{u\left(w\left(X_i, X_j\right)^{\prime} \beta_0, A_i, A_j, \epsilon_{ij}\right) \geq 0\right\} \\
&\quad \times \mathbb{1}\left\{u\left(w\left(X_j, X_i\right)^{\prime} \beta_0, A_j, A_i, \epsilon_{ji}\right) \geq 0\right\} d\mathbb{P}\left(\epsilon_{ij}, \epsilon_{ji}\right) \\
&=: \phi\left(w\left(X_i, X_j\right)^{\prime} \beta_0, A_i, A_j\right).
\end{aligned} \tag{3.5}$$

can be represented by model (3.3) with Assumption 3.1 satisfied.

In particular, notice that we do not require $\epsilon_{ij} \perp \epsilon_{ji}$. In fact, $\epsilon_{ij} \equiv \epsilon_{ji}$ is readily incorporated in our model. If u is furthermore assumed to be symmetric with respect to its second and third arguments (A_i and A_j), then our model degenerates to the case of transferable utilities,

$$D_{ij} = \mathbb{1}\left\{u\left(w\left(X_i, X_j\right)'\beta_0, A_i, A_j, \epsilon_{ij}\right) \geq 0\right\},$$

where effectively only one threshold crossing condition will be determining the establishment of a given network link.

Remark 3.1 (**Symmetry of $\boldsymbol{w\left(X_i, X_j\right)}$**). To explain the key idea of our identification strategy in a notation-economical way, we will be focusing on the case of symmetric w in the most of the following sections. However, it should be pointed out that our method can also be applied to the case where w is allowed to be asymmetric in (3.4), so that individual utilities based on observable characteristics can also be made asymmetric (nontransferable). In that case, model (3.4) need to be modified as

$$\mathbb{E}\left[D_{ij}\middle| X_i, X_j, A_i, A_j\right] = \phi\left(w\left(X_i, X_j\right)'\beta_0, w\left(X_j, X_i\right)'\beta_0, A_i, A_j\right), \tag{3.6}$$

where $w\left(X_i, X_j\right)'\beta_0$ may be different from $w\left(X_j, X_i\right)'\beta_0$, but ϕ is symmetric with respect to its first two arguments $w\left(X_i, X_j\right)'\beta_0, w\left(X_j, X_i\right)'\beta_0$ *whenever* $A_i = A_j$. Moreover, Assumption 3.1 should also be understood as monotonicity with respect to all *four* arguments of ϕ. See Appendix (3.B) for more discussion on how our identification strategy can be adapted to accommodate asymmetric w under appropriate conditions.

3.3 Identification and Estimation

3.3.1 Identification via Logical Differencing

In this section, we explain the key idea of our identification strategy. We construct a mutually exclusive event to cancel out the unobservable heterogeneity A_i and A_j, which leads to an identifying restriction on β_0. We call this technique "*logical differencing*".

For each fixed individual $\bar{i}$, and each possible $\bar{x} \in \mathbb{R}^{d_x}$, define

$$\rho_{\bar{i}}\left(\bar{x}\right) := \mathbb{E}\left[\left.D_{\bar{i}k}\right| X_k = \bar{x}\right] \tag{3.7}$$

as the linking probability of this specific individual $\bar{i}$ with a group of individuals, individually indexed by k, with the same observable characteristics $X_k = \bar{x}$ (but potentially different fixed effects A_k). Clearly, $\rho_i\left(\bar{x}\right)$ is directly identified from data in a single large network.

Suppose that individual $\bar{i}$ has observed characteristics $X_{\bar{i}} = x_{\bar{i}}$ and unobserved characteristics $A_{\bar{i}} = a_{\bar{i}}$. Then, by model (3.3) we have

$$\begin{aligned}
\rho_{\bar{i}}\left(\bar{x}\right) &= \mathbb{E}\left[\left.\mathbb{E}\left[\left.D_{\bar{i}k}\right| X_k = \bar{x}, A_k, X_{\bar{i}} = x_{\bar{i}}, A_{\bar{i}} = a_{\bar{i}}\right]\right| X_k = \bar{x}\right] \\
&= \mathbb{E}\left[\left.\phi\left(w\left(x_{\bar{i}}, \bar{x}\right)' \beta_0, a_{\bar{i}}, A_k\right)\right| X_k = \bar{x}\right] \\
&=: \psi_{\bar{x}}\left(w\left(x_{\bar{i}}, \bar{x}\right)' \beta_0, a_{\bar{i}}\right),
\end{aligned} \tag{3.8}$$

where the expectation in the second to last line is taken over A_k conditioning on $X_k = \bar{x}$. As we allow A_k and X_k to be arbitrarily correlated, the $\psi_{\bar{x}}$ function defined in the last line of (3.8) is dependent on $\bar{x}$. In the same time, notice that $\psi_{\bar{x}}$ does not depend on the identity of $\bar{i}$ beyond the values of $w\left(x_{\bar{i}}, \bar{x}\right)' \beta_0$ and $a_{\bar{i}}$. By Assumption 3.1, $\psi_{\bar{x}}\left(w\left(x_{\bar{i}}, \bar{x}\right)' \beta_0, a_{\bar{i}}\right)$ must be bivariate weakly increasing in the index $w\left(x_{\bar{i}}, \bar{x}\right)' \beta_0$ and the unobserved heterogeneity scalar $a_{\bar{i}}$. We now show how to use bivariate monotonicity to obtain identifying restrictions on β_0.

Fixing two distinct individuals $\bar{i}$ and $\bar{j}$ in the population, we first consider the event that *individual* $\bar{i}$ is *strictly more popular than individual* $\bar{j}$ among the *group* of individuals with observed characteristics $X_k = \bar{x}$:

$$\rho_{\bar{i}}\left(\bar{x}\right) > \rho_{\bar{j}}\left(\bar{x}\right), \tag{3.9}$$

which is an event directly identifiable from observable data given (3.7). Even though event (3.9) is the same conditioning event as considered in Toth (2017) and analogous to the tetrad comparisons made in Candelaria (2016), we now exploit the following logical deduction based on the bivariate monotonicity of the conditional popularity of $\bar{\imath}$ in $w\left(X_{\bar{\imath}}, \overline{x}\right)^{'} \beta_0$ and $A_{\bar{\imath}}$ without the assumption of additivity between them. Specifically, writing $\left(x_{\bar{\imath}}, a_{\bar{\imath}}\right)$ and $\left(x_{\bar{\jmath}}, a_{\bar{\jmath}}\right)$ as the observable and unobservable characteristics of individuals $\bar{\imath}$ and $\bar{\jmath}$, by (3.8) we have

$$\begin{aligned} & \rho_{\bar{\imath}}\left(\overline{x}\right) > \rho_{\bar{\jmath}}\left(\overline{x}\right). \\ \Leftrightarrow \quad & \psi_{\overline{x}}\left(w\left(x_{\bar{\imath}}, \overline{x}\right)^{'} \beta_0, a_{\bar{\imath}}\right) > \psi_{\overline{x}}\left(w\left(x_{\bar{\jmath}}, \overline{x}\right)^{'} \beta_0, a_{\bar{\jmath}}\right) \\ \Rightarrow \quad & \left\{w\left(x_{\bar{\imath}}, \overline{x}\right)^{'} \beta_0 > w\left(x_{\bar{\jmath}}, \overline{x}\right)^{'} \beta_0\right\} \text{ OR } \left\{a_{\bar{\imath}} > a_{\bar{\jmath}}\right\}, \end{aligned} \tag{3.10}$$

Note that the last line of equation (3.10) is a natural necessary (but not sufficient) condition for $\rho_{\bar{\imath}}\left(\overline{x}\right) > \rho_{\bar{\jmath}}\left(\overline{x}\right)$ under bivariate monotonicity.

Now, consider the event that *individual* $\bar{\imath}$ is *strictly less popular than individual* $\bar{\jmath}$ among the *group* of individuals with observed characteristics $X_h = \underline{x}$, i.e.,

$$\rho_{\bar{\imath}}\left(\underline{x}\right) < \rho_{\bar{\jmath}}\left(\underline{x}\right). \tag{3.11}$$

Then, by a similar argument to (3.10), we deduce

$$\rho_i\left(\underline{x}\right) < \rho_j\left(\underline{x}\right) \quad \Rightarrow \quad \left\{w\left(x_{\bar{\imath}}, \underline{x}\right)^{'} \beta_0 < w\left(x_{\bar{\jmath}}, \underline{x}\right)^{'} \beta_0\right\} \text{ OR } \left\{a_{\bar{\imath}} < a_{\bar{\jmath}}\right\}. \tag{3.12}$$

Notice that the event $\left\{a_{\bar{\imath}} < a_{\bar{\jmath}}\right\}$ in (3.12) is mutually exclusive with the event $\left\{a_{\bar{\imath}} > a_{\bar{\jmath}}\right\}$ that shows up in (3.10).

Next, consider the event that the two events (3.9) and (3.11) described above *simultaneously happen.* Then, by (3.10), (3.12) and basic logical operations, we have

$$\left\{\rho_{\bar{\imath}}\left(\overline{x}\right) > \rho_{\bar{\jmath}}\left(\overline{x}\right)\right\} \text{ AND } \left\{\rho_{\bar{\imath}}\left(\underline{x}\right) < \rho_{\bar{\jmath}}\left(\underline{x}\right)\right\}$$

$$\begin{aligned}
\Rightarrow \quad & \left(\left\{w\left(x_{\bar{i}},\overline{x}\right)'\beta_0 > w\left(x_{\bar{j}},\overline{x}\right)'\beta_0\right\} \text{ OR } \left\{a_{\bar{i}} > a_{\bar{j}}\right\}\right) \\
& \text{AND } \left(\left\{w\left(x_{\bar{i}},\underline{x}\right)'\beta_0 < w\left(x_{\bar{j}},\underline{x}\right)'\beta_0\right\} \text{ OR } \left\{a_{\bar{i}} < a_{\bar{j}}\right\}\right) \\
\Leftrightarrow \quad & \left(\left\{w\left(x_{\bar{i}},\overline{x}\right)'\beta_0 > w\left(x_{\bar{j}},\overline{x}\right)'\beta_0\right\} \text{ AND } \left\{w\left(x_{\bar{i}},\underline{x}\right)'\beta_0 < w\left(x_{\bar{j}},\underline{x}\right)'\beta_0\right\}\right) \\
& \text{OR } \left(\left\{w\left(x_{\bar{i}},\overline{x}\right)'\beta_0 > w\left(x_{\bar{j}},\overline{x}\right)'\beta_0\right\} \text{ AND } \left\{a_{\bar{i}} < a_{\bar{j}}\right\}\right) \\
& \text{OR } \left(\left\{a_{\bar{i}} > a_{\bar{j}}\right\} \text{ AND } \left\{w\left(x_{\bar{i}},\underline{x}\right)'\beta_0 < w\left(x_{\bar{j}},\underline{x}\right)'\beta_0\right\}\right) \\
& \text{OR } \left(\left\{a_{\bar{i}} > a_{\bar{j}}\right\} \text{ AND } \left\{a_{\bar{i}} < a_{\bar{j}}\right\}\right) \\
\Rightarrow \quad & \left(\left\{w\left(x_{\bar{i}},\overline{x}\right)'\beta_0 > w\left(x_{\bar{j}},\overline{x}\right)'\beta_0\right\} \text{ AND } \left\{w\left(x_{\bar{i}},\underline{x}\right)'\beta_0 < w\left(x_{\bar{j}},\underline{x}\right)'\beta_0\right\}\right) \\
& \text{OR } \left\{w\left(x_{\bar{i}},\overline{x}\right)'\beta_0 > w\left(x_{\bar{j}},\overline{x}\right)'\beta_0\right\} \\
& \text{OR } \left\{w\left(x_{\bar{i}},\underline{x}\right)'\beta_0 < w\left(x_{\bar{j}},\underline{x}\right)'\beta_0\right\} \\
\Leftrightarrow \quad & \left\{\left(w\left(x_{\bar{i}},\overline{x}\right) - w\left(x_{\bar{j}},\overline{x}\right)\right)'\beta_0 > 0\right\} \text{ OR } \left\{\left(w\left(x_{\bar{i}},\underline{x}\right) - w\left(x_{\bar{j}},\underline{x}\right)\right)'\beta_0 < 0\right\}, \qquad (3.13)
\end{aligned}$$

The derivations above exploit two simple logical properties: first,

$$\left\{a_{\bar{i}} > a_{\bar{j}}\right\} \text{ AND } \left\{a_{\bar{i}} < a_{\bar{j}}\right\} \quad = \quad \text{FALSE},$$

and second,

$$\left\{w\left(x_{\bar{i}},\overline{x}\right)'\beta_0 > w\left(x_{\bar{j}},\overline{x}\right)'\beta_0\right\} \text{ AND } \left\{a_{\bar{i}} < a_{\bar{j}}\right\} \quad \Rightarrow \quad \left\{w\left(x_{\bar{i}},\overline{x}\right)'\beta_0 > w\left(x_{\bar{j}},\overline{x}\right)'\beta_0\right\},$$

which uses only necessary but not sufficient condition, so that we can obtain an identifying restriction (3.13) on β_0 that does not involve $a_{\bar{i}}$ nor $a_{\bar{j}}$. These two forms of logical operations together enable us to "difference out" (or "cancel out") the unobserved heterogeneity terms $a_{\bar{i}}$ and $a_{\bar{j}}$.

In contrast with various forms of "*arithmetic differencing*" techniques proposed in the econometric literature (including Candelaria, 2016 and Toth, 2017 specific to the dyadic network formation literature), our proposed technique does *not* rely on additive separability between the parametric index $w\left(x_{\bar{i}}, \overline{x}\right)' \beta_0$ and the unobserved heterogeneity term $a_{\bar{i}}$. Instead, our identification strategy is based on multivariate monotonicity and utilizes logical operations rather than standard arithmetic to cancel out the unobserved heterogeneity terms. Hence we term our method "*logical differencing*".

The identifying arguments above are derived for a fixed pair of individuals $\bar{i}$ and $\bar{j}$, but clearly the arguments can be applied for any pair of individuals ij with observable characteristics x_i and x_j. Writing

$$\begin{aligned}\tau_{ij}\left(\overline{x}, \underline{x}\right) &:= \mathbb{1}\left\{\rho_i\left(\overline{x}\right) > \rho_j\left(\overline{x}\right)\right\} \cdot \mathbb{1}\left\{\rho_i\left(\underline{x}\right) < \rho_j\left(\underline{x}\right)\right\}, \\ \lambda\left(\overline{x}, \underline{x}; x_i, x_j; \beta\right) &:= \mathbb{1}\left\{\left(w\left(x_i, \overline{x}\right) - w\left(x_j, \overline{x}\right)\right)' \beta_0 \leq 0\right\} \cdot \mathbb{1}\left\{\left(w\left(x_i, \underline{x}\right) < w\left(x_j, \underline{x}\right)\right)' \beta_0 \geq 0\right\},\end{aligned}$$

for each $\beta \in \mathbb{S}^{m-1}$, we summarize the identifying arguments above by the following lemma.

Lemma 3.1 (Identifying Restriction). *Under model* (3.3) *and Assumptions 3.1 and 3.2, we have.*

$$\tau_{ij}\left(\overline{x}, \underline{x}\right) = 1 \quad \Rightarrow \quad \lambda\left(\overline{x}, \underline{x}; x_i, x_j; \beta_0\right) = 0.$$

A simple (but clearly not unique) way to build a criterion function based on the above lemma is to define

$$Q\left(\beta\right) := \mathbb{E}_{ij,kl}\left[\tau_{ij}\left(X_k, X_l\right) \lambda_{ij}\left(X_k, X_l; X_i, X_j; \beta\right)\right], \tag{3.14}$$

where the expectation is $\mathbb{E}_{ij,kl}$ taken over random samples of ordered tetrads (i, j, k, l) from the population, and (X_i, X_j, X_k, X_l) denote the random variables corresponding to the observable characteristics of (i, j, k, l). According to Lemma 3.1, $Q\left(\beta_0\right) = 0$, which is always smaller than or equal to $Q(\beta) \geq 0 = Q(\beta_0)$ for any $\beta \neq \beta_0$ because $\tau_{ij} \geq 0$ and $\lambda_{ij} \geq 0$ by construction.

Observing that the scale of β_0 is never identified, we write

$$B_0 := \left\{\beta \in \mathbb{S}^{d_\beta - 1} : Q(\beta) = 0\right\}$$

to represent the normalized "identified set" relative to the criterion Q defined above. Lemma 3.1 implies that $\beta_0 \in B_0$, but in general there is no guarantee that B_0 is a singleton. The next subsection contains a set of sufficient conditions that guarantees $B_0 = \{\beta_0\}$.

3.3.2 Sufficient Conditions for Point Identification

We now present a set of sufficient conditions that guarantee point identification of β_0 on the unit sphere $\mathbb{S}^{m-1}$, where for notational simplicity $m := d_\beta$.

Assumption 3.3 (**Full Directional Support**). *There exist a pair of $\overline{x}, \underline{x}$, both of which lie in the support of $Supp(X_i)$, such that $Supp(w(\overline{x}, X_i) - w(\underline{x}, X_i))$ contains all directions in $\mathbb{R}^m$.*

When $w(\overline{x}, \underline{x})$ is a component-wise Euclidean distance function, i.e., $w_h(\overline{x}, \underline{x}) = |\overline{x}_h - \underline{x}_h|$ where h indexes each coordinate of possibly vector valued $w(\cdot, \cdot)$ function, Assumption 3.3 is satisfied if $Supp(X_i)$ has nonempty interior[9], which is analogous to the standard assumption imposed for point identification on the unit sphere. When some components of $w(\overline{x}, X_j)$ have discrete range space, we need to require that at least one component of $w(\overline{x}, X_i) - w(\underline{x}, X_i)$ have full support on $\mathbb{R}$ and the coordinate of β_0 it corresponds to is nonzero, such that it creates enough variation in $w(\overline{x}, X_i) - w(\underline{x}, X_i)$ to guarantee Assumption 3.3 is satisfied.

[9] When $Supp(X_i)$ has nonempty interior, there exist $\overline{x}, \underline{x} \in Supp(X_i)$ such that $\overline{x} >> \underline{x}$ in the point-wise sense and $\times_{h=1}^{d_x} [\underline{x}_h, \overline{x}_h] \subseteq int(Supp(X_i))$. In particular, $\frac{1}{2}(\overline{x} + \underline{x}) \in int(Supp(X_i))$ and thus $0 \in int(Supp(w(\overline{x}, X_i) - w(\underline{x}, X_i)))$. Consequently, one can construct a $\varepsilon-$ball around origin for $Supp(w(\overline{x}, X_i) - w(\underline{x}, X_i))$ by choosing X_i from $\times_{h=1}^{d_x} [\underline{x}_h, \overline{x}_h]$ and Assumption 3.3 is satisfied.

Assumption 3.4 (**Conditional Support of $\boldsymbol{A_i}$**). *A_i is continuously distributed on the same support, conditional on $X_i = x_i$ for any $x_i \in \operatorname{supp}(X_i)$.*

Assumption 3.5 (**Continuity of $\boldsymbol{\phi}$**). *ϕ is continuous with respect to the second and third arguments.*

Assumption 3.4 together with Assumption 3.2 implies that conditional on X_i and X_j, for two randomly sampled agents i, j, 0 is in the support of $|A_i - A_j|$. Assumption 3.5 then ensures that $\tau_{ij}(\overline{x}, \underline{x}) = 1$ occurs with strictly positive probability, which is required for the point identification result.

Next, we lay out the lemma that will be used in the proof of point identification of β_0.

Lemma 3.2 (**Tools for Point Identification**). *Under model* (3.3)*, Assumptions 3.1, 3.2, 3.3, 3.4, and 3.5, for each $\beta \in \mathbb{S}^{m-1} \backslash \beta_0$, there exist x_i, x_j, $\overline{x}$, and $\underline{x}$ all in the support of X_i such that*

$$\tau_{ij}(\overline{x}, \underline{x}) = 1, \tag{3.15}$$

$$\lambda_{ij}(\overline{x}, \underline{x}; x_i, x_j; \beta_0) = 0, \tag{3.16}$$

$$\lambda_{ij}(\overline{x}, \underline{x}; x_i, x_j; \beta) = 1. \tag{3.17}$$

We are now ready to present the point identification result.

Theorem 3.1 (**Point Identification of $\boldsymbol{\beta_0}$**). *Under model* (3.3) *and Assumptions 3.1, 3.2, 3.3, 3.4, and 3.5. Then β_0 is the unique minimizer of $Q(\beta)$ defined in 3.14 over the unit sphere $\mathbb{S}^{d_\beta - 1}$. Furthermore, for any $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\inf_{\beta \in \mathbb{S}^{d_\beta - 1} \backslash B(\beta_0, \epsilon)} Q(\beta) \geq Q(\beta_0) + \delta,$$

where $B(\beta_0, \epsilon) = \left\{\beta \in \mathbb{S}^{d_\beta - 1} : \|\beta - \beta_0\| \leq \epsilon\right\}$.

Remark 3.2 (**Asymmetry of $\boldsymbol{w}$, Continued**). In Appendix (3.B), we show how the identification arguments and assumptions above can be adapted to accommodate asymmetry of w. In short, the technique of logical differencing applies without changes, but the identifying restriction we obtained become weaker. In particular, when w is *antisymmetric* in the sense that $w(\overline{x}, \underline{x}) + w(\underline{x}, \overline{x}) \equiv 0$, the identifying restriction we obtained through logical differencing becomes trivial, and $B_0 = \mathbb{S}^{d_\beta - 1}$. However, with asymmetric but not antisymmetric w, it is still feasible to strengthen Assumption 3 so as to obtain point identification. See more discussions in Appendix (3.B).

3.3.3 Tetrad Estimation and Consistency

We now proceed to present a consistent estimator of β_0 in the framework of extremum estimation. We will first construct the sample criterion function and show how to estimate β_0 via a two-step estimation procedure. Then we will list one additional assumption before presenting the consistency result.

Define the sample analog of the population criterion $Q(\beta)$ in (3.14) by

$$\begin{aligned}\widehat{Q}_n(\beta) := \frac{(n-4)!}{n!} \sum_{1 \le i \ne j \ne k \ne l \le n} & \mathbb{1}\{\hat{\rho}_i(X_k) > \hat{\rho}_j(X_k)\} \cdot \mathbb{1}\{\hat{\rho}_i(X_l) < \hat{\rho}_j(X_l)\} \\ & \cdot \left[\begin{array}{l} \mathbb{1}\left\{w(X_i, X_k)' \beta \le w(X_j, X_k)' \beta\right\} \\ \cdot \mathbb{1}\left\{w(X_i, X_l)' \beta \ge w(X_j, X_l)' \beta\right\} \end{array}\right],\end{aligned} \tag{3.18}$$

where $\hat{\rho}_i(x)$ is a first-step nonparametric estimator of $\rho_i(x)$. The two-step tetrad estimator for β_0 is defined as

$$\widehat{\beta}_n := \arg\min_{\beta \in \mathbb{S}^{d_\beta - 1}} \widehat{Q}_n(\beta). \tag{3.19}$$

The first-step estimation of $\rho_i(x) := \mathbb{E}\left[D_{ik} \middle| i, X_k = x\right]$ function is a standard nonparametric regression problem. Computationally, one can fix each i in the sample, and regress

D_{ik}, the indicator function for the link between i and k, on the basis functions chosen by the researcher evaluated at observable characteristics X_k for all $k \neq i$. There are many tools readily available to nonparametrically estimate $\rho_i(x)$ in the first stage. For example, one can use kernel, sieve, or neural networks. In Section 3.4, we use second order sieves with knot at the median to estimate $\rho_i(x)$ for the simulation study. Theoretical properties of our sieve estimator $\hat{\rho}_i(x)$ can be found in Chen (2007).

It is worth mentioning that we can smooth each component of $\tau_{ij}(\overline{x}, \underline{x})$ to achieve better numerical performance as long as the sign of the differences between $\rho_i(x)$ and $\rho_j(x)$ is preserved. Recall that

$$\tau_{ij}(\overline{x}, \underline{x}) := \mathbb{1}\left\{\rho_i(\overline{x}) > \rho_j(\overline{x})\right\} \cdot \mathbb{1}\left\{\rho_i(\underline{x}) < \rho_j(\underline{x})\right\}. \tag{3.20}$$

When $\rho_i(x)$ is close to $\rho_j(x)$, the estimation of $\tau_{ij}(\overline{x}, \underline{x})$ may be imprecise and sensitive to errors during data collection and analysis procedure. Therefore, we may wish to smooth both $\mathbb{1}\left\{\rho_i(\overline{x}) > \rho_j(\overline{x})\right\}$ and $\mathbb{1}\left\{\rho_i(\underline{x}) < \rho_j(\underline{x})\right\}$ such that the potential bias caused by the imprecise estimation at the boundary point of 0 is smaller. In practice, we can do so by applying a known smooth one-directional function H on $\rho_i(x) - \rho_j(x)$. A concrete example of H is the standard normal CDF, i.e. replace $\mathbb{1}\left\{\rho_i(\overline{x}) > \rho_j(\overline{x})\right\}$ with $2 \times \Phi\left[\left(\rho_i(\overline{x}) - \rho_j(\overline{x})\right)_+\right] - 1$ and replace $\mathbb{1}\left\{\rho_i(\underline{x}) < \rho_j(\underline{x})\right\}$ with $2 \times \Phi\left[\left(\rho_j(\underline{x}) - \rho_i(\underline{x})\right)_+\right] - 1$ in $\tau_{ij}(\overline{x}, \underline{x})$, where $(c)_+$ is the positive part of c, otherwise 0, and Φ is the CDF of standard normal $\mathcal{N}(0,1)$. We use smoothed $\tau_{ij}(\overline{x}, \underline{x})$ in the simulation part. See Section 3.4 for details.

For the second step, we estimate β_0 by minimizing the sample criterion function $\widehat{Q}_n(\beta)$ over the unit sphere $\mathbb{S}^{d_\beta - 1}$ after plugging in the first stage estimator $\hat{\tau}_{ij}(\overline{x}, \underline{x})$. To exploit the topological characteristics of the parameter space $\mathbb{S}^{d_\beta - 1}$, i.e. compactness and convexity, we develop a new bisection-style nested rectangle algorithm that recursively shrinks and refines an adaptive grid on the angle space. The key novelty of the algorithm is that instead of working with the edges of the Euclidean parameter space $\mathbb{R}^{d_\beta}$, we deterministically "cut" the angle space in each dimension of $\mathbb{S}^{d_\beta - 1}$ to search for the area that minimizes $\widehat{Q}_n(\beta)$. Additional measures are taken to ensure the search algorithm is conservative. Simulation and empirical results show that our algorithm performs reasonably well with a relatively small sample size. Gao and Li (2020) provides more details regarding the implementation in a panel multinomial choice setting.

For consistency, we impose the following assumption regarding the first-step nonparametric estimator $\hat{\rho}_i(\cdot)$ of the $\rho_i(\cdot)$ function.

Assumption 3.6 (**Uniform Consistency for $\boldsymbol{\rho_i(\cdot)}$**). *$\hat{\rho}_i(\cdot)$ is a uniformly consistent estimator of $\rho_i(\cdot)$ for each agent i.*

The usual kernel and sieve methods we mentioned above to estimate $\rho_i(x)$ have been proved to satisfy Assumption 3.6: see Bierens (1983) for results on kernel estimators and Chen (2007) on sieve estimators.

Lemma 3.3 (**Uniform Convergence of $\boldsymbol{\widehat{Q}_n(\beta)}$**). *Under model* (3.3) *and Assumptions 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6, we have*

$$\sup_{\beta \in \mathbb{S}^{d_\beta - 1}} \left| \widehat{Q}_n(\beta) - Q_n(\beta) \right| \xrightarrow{p} 0.$$

Finally, we state the consistency result of the tetrad estimator $\widehat{\beta}_n$.

Theorem 3.2 (**Consistency**). *Under model* (3.3) *and Assumptions 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6, $\hat{\beta}_n$ is consistent for β_0, i.e.,*

$$\widehat{\beta}_n \xrightarrow{p} \beta_0.$$

3.4 Simulation

In this section, we conduct a simulation study to analyze the finite-sample performance of our two-step tetrad estimator. We start by specifying the data generating process (DGP) of the Monte Carlo simulations. Next, we show and discuss the performance of our 2-step estimation method under the baseline setup. Then, we vary the number of individuals N, the dimension of the pairwise observable characteristics $\overline{d}$, and the degree of correlation between X and A to further examine the robustness of our method. Finally, we show how the method performs when $w(X_i, X_j)$ is an asymmetric function of X_i and X_j, i.e. $w(X_i, X_j) \neq w(X_j, X_i)$.

3.4.1 Setup of Simulation Study

For each DGP configuration, we run $B = 100$ independent simulations of model 3.3 with the following network formation rule unknown to the econometrician for each agent pair (i, j)

$$D_{ij} = \mathbb{1}\left\{w(X_i, X_j)' \beta_0 + A_i > \epsilon_{ij}\right\} \cdot \mathbb{1}\left\{w(X_j, X_i)' \beta_0 + A_j > \epsilon_{ji}\right\}, \tag{3.21}$$

where the usual linear additivity is excluded by construction that D_{ij} equals the product of two indicator functions. In (3.21), D_{ij} equals one if i and j are connected, zero otherwise. X_i and X_j are $d_x \times 1$ vectors of observable characteristics of individual i and j, respectively. $w(X_i, X_j)$ is a known vector-valued function mapping (X_i, X_j) pairs to a $d_w \times 1$ vector. β_0 is a $d_\beta \times 1$ vector of structural parameter of interest. We maintain $d_x = d_w = d_\beta = \overline{d}$ in all our configurations. A_i represents the unobservable scalar valued fixed effect that is possibly correlated with X_i. ϵ_{ij} is the scalar valued iid random shocks independent of X and A.

In our baseline DGP configuration where we fix $N = 100$ and $\overline{d} = 3$, each coordinate of X_i is drawn independently across both individuals i and dimensions d from a uniform distribution on $[-0.5, 0.5]$. Then we compute $W_{ij}^{(d)}$, the d^{th} coordinate of $w(X_i, X_j)$ vector, as $W_{ij}^{(d)} = \left|X_i^{(d)} - X_j^{(d)}\right|$. Note that for the baseline setup we maintain the symmetry of W_{ij}

in (X_i, X_j) pairs, i.e., $W_{ij} = W_{ji}$. Later on, we will relax this restriction and investigate the asymmetric case where $W_{ij} \neq W_{ji}$.

Next, we construct the unobserved heterogeneity A_i. To allow for the correlation between A_i and X_i, we draw iid sequence ξ_i independently from X_i from a uniform distribution on $[-0.5, 0.5]$ and let $A_i = corr \times X_i^{(1)} + (1 - corr) \times \xi_i$, where $corr$ controls the degree of correlation between X_i and A_i and is set to be 0.2. Later on, we will vary the correlation to see how robust our estimator is against correlation between A and X. As for the random utility shock ϵ_{ij}, we draw them from a uniform distribution on $[0, 1]$. Note that our estimation method does not require the knowledge of the distribution of A_i or ϵ_{ij}. We set the true $\beta_0 \in \mathbb{R}^{d_\beta}$ to be $(1, ..., 1)'$, and estimate the direction of β_0, represented by the normalized vector $\overline{\beta}_0 := \beta_0 / \|\beta_0\|$ on the unit sphere $\mathbb{S}^{d_\beta - 1}$ because the scale of β_0 is not identified. We shall maintain the specification of (ϵ, A, β_0) and the network formation rule (3.21) to be the same across all simulations.

Our method allows for asymmetry of W_{ij} in (X_i, X_j) pairs. To numerically show this, for the last coordinate $d = \overline{d}$ we compute $W_{ij}^{(\overline{d})}$ as $\left|2X_i^{(\overline{d})} - X_j^{(\overline{d})}\right| \times (2/3)$. The reason for multiplying 2/3 is to make the size of $W_{ij}^{(\overline{d})}$ similar to other coordinates of W_{ij}. This way we generate asymmetry because $W_{ij}^{(\overline{d})} \neq W_{ji}^{(\overline{d})}$ unless $\left|X_i^{(d)}\right| = \left|X_j^{(d)}\right|$, which is a probability zero event under our DGP setting. For other dimensions $d = 1, ..., \overline{d} - 1$, we maintain the baseline assumption. As a robustness check, we also vary N and $\overline{d}$ under asymmetry to show how our method works.

To summarize, for each of the $B = 100$ simulations we randomly generate data on the characteristics of and the network structure among individuals. Then based on the observable $(X_i, W_{ij}, D_{ij})_{i,j \in \{1,...,N\}}$ matrix we construct our two-step estimator $\widehat{\beta}$ for the true parameter of interest $\overline{\beta}_0$. Specifically, we use a sieve estimator with 2nd-order spline with its knot at median for the first-stage nonparametric estimation of $\rho_i(\cdot)$. The spline is chosen to ensure a relatively small number of regressors in the nonparametric regression considering the small size of N. In the second stage, we adapt to the adaptive-gird search on the unit sphere algorithm developed in Gao and Li (2020) to calculate $\widehat{\beta}$ that minimizes the sample criterion function $\widehat{Q}(\beta)$ over the unit sphere. We refer interested readers to that paper for more technical details. It should be noted that constrained by computational power, when calculating the sample criterion $\widehat{Q}(\beta)$ for each $\beta \in \mathbb{S}^{\overline{d}-1}$ we randomly draw $M = 1000$ (i, j) pairs of individuals and vary across all possible (k, l) pairs excluding i or j. One can improve those results by increasing M when computational constraint is not present. Lastly, we compare our estimator $\widehat{\beta}$ with the true parameter value $\overline{\beta}_0$ based on several performance metrics including rMSE, mean norm deviations (MND), and maximum absolute deviation (MAB).

3.4.2 Results under Symmetric Pairwise Observable Characteristics

Baseline Results

For the baseline configuration, we fix number of individuals $N = 100$, dimension of W_{ij} $\overline{d} = 3$, number of (i, j) pairs used in evaluating $\widehat{Q}(\beta)$ $M = 1000$, and number of simulations $B = 100$. We define for each simulation round b the argmin set estimator $\widehat{B}_b$ as the set of points that achieve the minimum of $\widehat{Q}(\beta)$ over the unit sphere $\mathbb{S}^{\overline{d}-1}$. Under point identification, any element from $\widehat{B}_b$ is a consistent point estimator for $\overline{\beta}_0$. In particular, we further define ,for each simulation $b = 1, ..., B$ and each dimension $d = 1, ..., \overline{d}$ of β

$$\widehat{\beta}^l_{b,d} := \min \widehat{B}_{b,d}, \quad \widehat{\beta}^u_{b,d} := \max \widehat{B}_{b,d}, \quad \widehat{\beta}^m_{b,d} := \frac{1}{2}\left(\widehat{\beta}^l_{b,d} + \widehat{\beta}^u_{b,d}\right),$$

where $\widehat{\beta}^l_{b,d}$ is the minimum value along dimension d for simulation round b of the argmin set $\widehat{B}_b$, $\widehat{\beta}^u_{b,d}$ is the maximum value along dimension d for simulation round b of the argmin set $\widehat{B}_b$, and $\widehat{\beta}^m_{b,d}$ is the middle point along dimension d for simulation round b of the argmin set $\widehat{B}_b$. Note by construction for each simulation round b, the argmin set $\widehat{B}_b$ is a subset of

Table 3.1: Baseline Performance

		β_1	β_2	β_3
bias	$\frac{1}{B}\sum_b \left(\hat{\beta}^m_{b,d} - \overline{\beta}_{0,d}\right)$	-0.0021	0.0052	-0.0053
upper bias	$\frac{1}{B}\sum_b \left(\hat{\beta}^u_{b,d} - \overline{\beta}_{0,d}\right)$	0.0048	0.0118	-0.0002
lower bias	$\frac{1}{B}\sum_b \left(\hat{\beta}^l_{b,d} - \overline{\beta}_{0,d}\right)$	-0.0091	-0.0015	-0.0105
mean$(u-l)$	$\frac{1}{B}\sum_b \left(\hat{\beta}^u_{b,d} - \hat{\beta}^l_{b,d}\right)$	0.0138	0.0132	0.0103
root MSE	$\sqrt{\frac{1}{B}\sum_b \left\lVert \hat{\beta}^m_b - \overline{\beta}_0 \right\rVert^2}$		0.0488	
mean norm deviations	$\frac{1}{B}\sum_b \left\lVert \hat{\beta}^m_b - \overline{\beta}_0 \right\rVert$		0.0417	
max absolute deviations	$\max_d \left\lvert \frac{1}{B}\sum_b \left(\hat{\beta}^m_{b,d} - \overline{\beta}_{0,d}\right) \right\rvert$		0.0053	

the rectangle $\widehat{\Xi}_b := \times_{d=1}^{\overline{d}} \left[\widehat{\beta}^l_{b,d},\ \widehat{\beta}^u_{b,d}\right]$. We calculate the baseline performance using $\widehat{\beta}^l, \widehat{\beta}^u, \widehat{\beta}^m$ respectively.

Below in Table 3.1 we report the performance of our estimators. In the first row of Table 3.1 we calculate the mean bias across $B = 100$ simulations using $\widehat{\beta}^m$ along each dimension $d = 1, ..., \overline{d}$. The result shows the estimation bias is very small across all dimensions with a magnitude between -0.0053 and 0.0052. Similar performance is observed using $\widehat{\beta}^u$ and $\widehat{\beta}^l$ as shown in row 2 and 3. We do not find any sign of persistent over/under- estimation of $\overline{\beta}_0$ across each dimension. Row 4 measures the average width of the rectangle $\widehat{\Xi}$ along each dimension. The size of $\widehat{\Xi}$ is very small, indicating a very tight area for the estimated set. In the second part of Table 3.1 we report rMSE, MND, and MAB, all of which are small in magnitude and provide evidence that our estimator work well in finite sample.

Results Varying N and $\overline{d}$

In this section we vary the number of individuals N and dimension of W_{ij} $\overline{d}$ to examine how robust our method is against these variations. We investigate the performance when $N = 50, 100, 200$ and $\overline{d} = 3, 4$, respectively. We maintain the symmetry in W_{ij} and other

Table 3.2: Results Varying N and $\overline{d}$

$\overline{d}=3$	rMSE	MND	MAB	$\overline{d}=4$	rMSE	MND	MAB
$N=50$	0.0839	0.0724	0.0051	$N=50$	0.1119	0.1030	0.0091
$N=100$	0.0488	0.0417	0.0053	$N=100$	0.0692	0.0647	0.0038
$N=200$	0.0334	0.029	0.0043	$N=200$	0.0543	0.0523	0.0038

distributional assumptions as in baseline setup. M, the number of (i,j) pairs used to evaluate objective function, is set to be 1000 in all simulations. Note that one could make M larger for larger N to better capture the more information available from the increase in N. In this sense, our results are conservative below. Results are summarized in Table 3.2.

The left part of Table 3.2 shows the performance of our estimator when N changes and $\overline{d}$ is fixed at 3. When N increases, rMSE, MND and sum of absolute bias all show moderate decline in magnitude, indicating the performance is improving. Similar pattern is also observed for $\overline{d}=4$. This is intuitive because with more individuals in the sample, one can achieve more accurate estimation of $\rho_i(\cdot)$ and calculation of $\widehat{Q}(\beta)$. Moreover, we can see even with a relatively small sample size of $N=50$, the rMSE is 0.0839 when $\overline{d}=3$ and 0.1119 when $\overline{d}=4$, showing that our method is informative and accurate. When $N=200$, the performance is very good, with rMSE being as small as 0.0334 and 0.0543 for $\overline{d}=3$ and $\overline{d}=4$, respectively. When we fix N and compare between $\overline{d}=3$ and $\overline{d}=4$, it is clear the increase in $\overline{d}$ adversely affects the performance of our estimator, with rMSE and MND increasing for each N. Overall, Table 3.2 provides evidence that our method is able to estimate $\overline{\beta}_0$ accurately even with a small sample size.

Results Varying *corr*

Correlation between observable characteristics X and unobservable fixed effect A is important in network formation models. We show how our estimator performs when the correlation

Table 3.3: Results Varying *corr*

corr	rMSE	MND	MAB
0.20	0.0488	0.0417	0.0053
0.50	0.0489	0.0435	0.0186
0.75	0.0763	0.0690	0.0506
0.90	0.1010	0.0951	0.0743

between X and A varies. Recall that we construct A_i as

$$A_i = corr \times X_i^{(1)} + (1 - corr) \times \xi_i, \tag{3.22}$$

where ξ_i is iid uniform on $[-0.5, 0.5]$ and is independent of X_i. We set *corr* to be 0.2 in the baseline configuration. In Table 3.3, we vary *corr* from 0.20 to 0.90 while fixing $N = 100$, $\overline{d} = 3$, $M = 1000$ and obtain the performance of our estimator among $B = 100$ simulations when W_{ij} is symmetric.

It can be seen from Table 3.3 that even though increase in *corr* adversely affects the performance of our estimator, the magnitude of the impact is relatively small. For example, rMSE only increases from 0.0488 to 0.1010 when *corr* increase dramatically from 0.2 to 0.9. Similar pattern is also observed using other performance metrics. Therefore, our estimator is robust against correlation between X and A.

3.4.3 Results under Asymmetric Pairwise Observable Characteristics

In this section, we investigate how our method works when W_{ij} is asymmetric. To introduce asymmetry, we construct $W_{ij}^{(\overline{d})} = \left|2X_i^{(\overline{d})} - X_j^{(\overline{d})}\right| \times (2/3)$ for each i, j pair. The reason for multiplying 2/3 is to make the size of $W_{ij}^{(\overline{d})}$ similar to other coordinates of W_{ij}. As discussed

Table 3.4: Results under Asymmetry

$\overline{d}=3$	rMSE	MND	MAB	$\overline{d}=4$	rMSE	MND	MAB
$N=50$	0.1498	0.1403	0.0936	$N=50$	0.2225	0.2124	0.1521
$N=100$	0.1096	0.1028	0.0741	$N=100$	0.1751	0.1695	0.1301
$N=200$	0.0943	0.0893	0.0672	$N=200$	0.1595	0.1555	0.1222

before, under our DGP $W_{ij}^{(\overline{d})} \neq W_{ji}^{(\overline{d})}$ unless $\left|X_i^{(\overline{d})}\right| = \left|X_j^{(\overline{d})}\right|$, which is a probability zero event. For $d = 1, ..., \overline{d}-1$, we follow the configuration for $W_{ij}^{(d)}$ mentioned in section 3.4.1 for the asymmetric case. We maintain other distributional assumptions for X, A, ϵ and fix the number of (i, j) pairs M at 1000 for evaluation of $\widehat{Q}(\beta)$. Finally, we vary N and D under the asymmetric setting to show how our estimator performs. Table 3.4 summarizes the results.

From Table 3.4 one can see our method performs reasonably well when W_{ij} is asymmetric. First, when the number of individuals N increases, the overall performance is improved, with rMSE decreasing from 0.1498 to 0.0943 for $\overline{d} = 3$ and from 0.2225 to 0.1595 for $\overline{d} = 4$ when N increases from 50 to 200. This result is caused by the more information available in the sample and echos what we have seen for the symmetric W_{ij} case. When the dimension of W_{ij} $\overline{d}$ increases from 3 to 4, the performance is worse, with rMSE increasing from 0.0943 to 0.1595 for $N = 200$. It shows that more data (information) is required for accurate estimation when the dimension of $\overline{\beta}_0$ is larger. Second, when compared with the symmetric W_{ij} case, the overall performance under asymmetry in W_{ij} is worse, with rMSE being 0.1498 for asymmetric W_{ij} versus 0.0839 for symmetric W_{ij} when $N = 50$ and $\overline{d} = 3$. In Appendix 3.B we discuss the implications of asymmetric W_{ij}. It is shown there the identifying power of the objective function is in general “less restrictive” than the corresponding identifying restriction in Lemma 3.1. Therefore, one would expect larger bias than symmetric W_{ij} case, which is exactly what one observes in Table 3.4. Recall that we set total number of (i, j) pairs for the evaluation of objective function M to be 1000 for all simulations. Based on results in Table 3.4, when W_{ij} is asymmetric and computational power allows, we suggest one increases M to improve performance.

3.5 Empirical Illustration

As an empirical illustration, we estimate a network formation model under NTU with data of a small village network called Nyakatoke in Tanzania. Nyakatoke is a small Haya community of 119 households in 2000 located in the Kagera Region of Tanzania. We are interested in how important factors, such as wealth, distance, and blood or religious ties, are relative to each other in deciding the formation of risk-sharing links among local residents. The estimation results demonstrate that our proposed method produces estimates that are consistent with economic intuition.

3.5.1 Data Description

The risk-sharing data of Nyakatoke, collected by Joachim De Weerdt in 2000, cover all of the 119 households in the community. It includes the information about whether or not two households are linked in the insurance network. It also provides detailed information on total USD assets and religion of each household, as well as kinship and distance between households. See De Weerdt (2004); De Weerdt and Dercon (2006); De Weerdt and Fafchamps (2011) for more details of this dataset.

To define the dependent variable *link*, the interviewer asks each household the following question:

> "*Can you give a list of people from inside or outside of Nyakatoke, who you can personally rely on for help and/or that can rely on you for help in cash, kind or labor?*"

The data contains three answers of “bilaterally mentioned”, “unilaterally mentioned”, and “not mentioned” between each pair of households. Considering the question is about whether one can rely on the other for help, we interpret both “bilaterally mentioned” and “unilaterally mentioned” as they are connected in this undirected network, meaning that *link* equals 1. We also run a robustness check by constructing a weighted network based on the answers, i.e. “bilaterally mentioned” means *link* equals 2, “unilaterally mentioned” means *link* equals 1, and “not mentioned” means *link* equals 0, and found that results are very similar.

We estimate the coefficients for wealth difference, distance and tie between households with our two-step estimator. *Wealth* is defined as the total assets in USD owned by each household in 2000, including livestocks, durables and land. *Distance* measures how far away two households are located in kilometers. *Tie* is a discrete variable, defined to be 3 if members of one household are parents, children and/or siblings of members of the other household, 2 if nephews, nieces, aunts, cousins, grandparents and grandchildren, 1 if any other blood relation applies or if two households share the same religion, and 0 if no blood religious tie exists. Following the literature we take natural log on *wealth* and *distance*, and we construct the *wealth difference* variable as the absolute difference in *wealth*.

Figure 3.1 illustrates the structure of the insurance network in Nyakatoke. Each node in the graph represents a household. The solid line between two nodes indicates they are connected, i.e., *link* equals 1. The size of each node is proportional to the USD wealth of each household. Each node is colored according to its rank in *wealth*: green for the top quartile, red for the second, yellow for the third and purple for the fourth quartile.

In the dataset there are 5 households that lack information on *wealth* and/or *distance*. We drop these observations, resulting in a sample size N of 114. The total number of ordered household pairs is 12,882. Summary statistics for the dependent and explanatory variables used in our analysis are presented in Table 3.5.

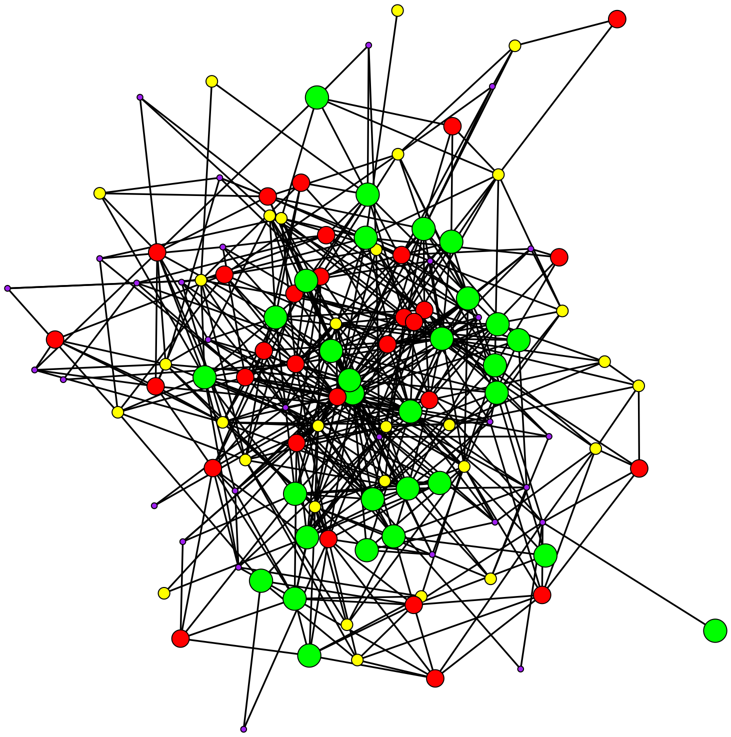


Figure 3.1: A Graphical Illustration of the Insurance Network of Nyakatoke

Table 3.5: Empirical Application: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
link	12,882	0.0732	0.2606	0	1
|(ln) wealth difference|	12,882	1.0365	0.8228	0.0004	5.8898
(ln) distance	12,882	6.0553	0.7092	2.6672	7.4603
tie	12,882	0.4260	0.6123	0.0000	3.0000

Table 3.6: Empirical Application: Estimation Results

Variable	$\hat{\beta}^m$	$\left[\hat{\beta}^l,\ \hat{\beta}^u\right]$
|(ln) wealth difference|	-0.1948	$[-0.1964,\ -0.1932]$
(ln) distance	-0.8036	$[-0.8043,\ -0.8029]$
tie	0.5619	$[0.5608,\ 0.5630]$

3.5.2 Methodology

To estimate β_0, we need to first estimate $\rho_i(x) := \mathbb{E}\left[D_{ik}\middle|i, W_{ik} = w\right]$ in order to construct $\tau_{ij}(\cdot)$. We use the second degree spline sieve with its knot at the median to estimate $\rho_i(w)$. Specifically, for each household i in the data, we regress dependent variable *link* D_{ik} on each dimension of W_{ik}, W_{ik}^2, and $\left[\left(W_{ik} - \text{median}\left(W_{ik}\right)\right)_+\right]^2$ including constant for $k \neq i$. The reason why we could regress on basis functions constructed with W instead of X is because X affects D only through W. We obtain an estimator $\hat{\rho}_i(\cdot)$ evaluated at each realized $W_{ik} = w$ in the data for each household i. We also smooth each component of $\tau_{ij}(\cdot)$, i.e. $\mathbb{1}\{\rho_i(\overline{w}) > \rho_j(\overline{w})\}$ and $\mathbb{1}\{\rho_i(\underline{w}) < \rho_j(\underline{w})\}$ with normal CDF to improve the performance. In the second stage, we estimate β_0 with $\hat{\beta}$ that minimizes the sample criterion $\widehat{Q}(\beta)$ by adapting to the adaptive-grid search on the unit sphere algorithm developed in Gao and Li (2020). As shown in finite sample simulations, the method is able to converge fast to the area that contains true β_0.

3.5.3 Results and Discussion

Table 3.6 summarizes our estimation results. The column of $\hat{\beta}^m$ corresponds to the center of the estimated rectangle $\hat{\Xi}$. We will use it as the point estimator of the coefficients for each variable of $\mathbf{W}$ vector. $\left[\hat{\beta}^l,\ \hat{\beta}^u\right]$ corresponds to the upper and lower bound of $\hat{\Xi}$. While the scale of β_0 is unidentified, we can still compare the estimated coefficients with each other to obtain an idea about which variable affects the formation of the link more than the other.

The estimated coefficients for each variable conform well with economic intuition. Our method estimate the coefficient for *absolute wealth difference* to be negative in the range of $[-0.1964,\ -0.1932]$, which implies the more absolute difference in wealth between two households, the lower likelihood they are connected. The estimated set for *distance* is $[-0.8043,\ -0.8029]$. It is natural households rely more on neighbors for help than ones that live farther away. The estimated coefficient for *tie* falls in the positive range of $[0.5608,\ 0.5630]$, which is also consistent with economic intuition that one would depend on support from family when negative shock occurs.

It is worth mentioning the estimated set $\hat{\Xi}$ is very tight in each dimension, with a maximum width of 0.0032 for *tie.* Usually the discreteness could make the estimated set wide, but our algorithm is able to circumvent this issue by leveraging the large support in the two other continuous variables, i.e., wealth difference and distance. The relative magnitude and sign of coefficient for *tie* are estimated in line with expectation. The empirical results show that our proposed estimator is able to generate economically intuitive estimates under NTU.

3.6 Conclusion

This paper considers a semiparametric model of dyadic network formation under nontransferable utilities, a natural and realistic micro-theoretical feature that translates into the lack of additive separability in econometric modeling. We show how a new methodology called *logical differencing* can be leveraged to cancel out the two-way fixed effects, which correspond to unobserved individual heterogeneity, without relying on arithmetic additivity. The key idea is to exploit the logical implication of weak multivariate monotonicity and use the intersection of mutually exclusive events on the unobserved fixed effects. It would be interesting to explore whether and how the idea of *logical differencing*, or more generally the use of fundamental logical operations, can be applied to other econometric settings.

Simulation results show that our method performs reasonably well with a relatively small sample size, and robust to various configurations. The empirical illustration using the real network data of Nyakatoke reveals that our method is able to capture the essence of the network formation process by generating estimates that conform well with economic intuition.

This paper also reveals several further research questions regarding dyadic network formation models under the NTU setting. First, given the observation that the NTU setting can capture "homophily effects" with respect to the unobserved heterogeneity (under log-concave error distributions) while imposing monotonicity in the unobserved heterogeneity in the same time, it is interesting to investigate whether we can differentiate homophily effects generated by "intrinsic preference" from homophily effects generated by bilateral consent, NTU and log-concave errors. Second, admittedly the identifying restriction obtained in this paper becomes uninformative when we have *antisymmetric* pairwise observable characteristics. However, preliminary analysis based on an adaption of Gao (2020) to the NTU setting suggests that individual unobserved heterogeneity can be nonparametrically identified up to location and inter-quantile range normalizations. After the identification of individual unobserved heterogeneity terms (A_i), it becomes straightforward to identify the index parameter β_0 based on the observable characteristics, even in the presence of antisymmetric pairwise characteristics. However, consistent estimators of A_i and β_0 in a semiparametric framework based on identification strategy in Gao (2020) are still being developed. We thus leave these research questions to future work.

References

Arduini, T., E. Patacchini, and E. Rainone (2015a): "Parametric and Semiparametric IV Estimation of Network Models with Selectivity," Working paper, Einaudi Institute for Economics and Finance (EIEF).

Auerbach, E. (2016a): "Identification and Estimation of Models with Endogenous Network Formation," Tech. rep., Working Paper.

Aumann, R. J. (1967): "A survey of cooperative games without side payments," *Essays in mathematical economics*, 3–27.

Bagnoli, M. and T. Bergstrom (2005): "Log-concave probability and its applications," *Economic theory*, 26, 445–469.

Bierens, H. J. (1983): "Uniform Consistency of Kernel Estimators of a Regression Function under Generalized Conditions," *Journal of the American Statistical Association*, 78, 699–707.

Blitzstein, J. and P. Diaconis (2011a): "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet mathematics*, 6, 489–522.

Boucher, V. (2015): "Structural homophily," *International Economic Review*, 56, 235–264.

Candelaria, L. E. (2016): "A Semiparametric Network Formation Model with Multiple Linear Fixed Effects," Working paper, Duke University.

Chandrasekhar, A. (2016a): "Econometrics of network formation," *The Oxford Handbook of the Economics of Networks*, 303–357.

——— (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20, S1–S13.

Chatterjee, S., P. Diaconis, and A. Sly (2011a): "Random graphs with a given degree sequence," *The Annals of Applied Probability*, 1400–1435.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Elsevier B.V., vol. 6B.

Currarini, S., M. O. Jackson, and P. Pin (2009): "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 77, 1003–1045.

——— (2010): "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences*, 107, 4857–4861.

Currarini, S., J. Matheson, and F. Vega-Redondo (2016): "A simple model of homophily in social networks," *European Economic Review*, 90, 18–39.

de Paula, A. (2016a): "Econometrics of network models," Tech. rep., cemmap working paper, Centre for Microdata Methods and Practice.

De Paula, Á., S. Richards-Shubik, and E. Tamer (2018b): "Identifying preferences in networks with bounded degree," *Econometrica*, 86, 263–288.

De Weerdt, J. (2004): "Risk-sharing and Endogenous Group Formation ?, chapter 10 in Dercon, S.(ed.), Insurance against Poverty, Oxford University Press," .

De Weerdt, J. and S. Dercon (2006): "Risk-sharing Networks and Insurance against Illness," *Journal of Development Economics*, 81, 337–356.

De Weerdt, J. and M. Fafchamps (2011): "Social identity and the formation of health insurance networks," *Journal of Development Studies*, 47, 1152–1177.

Dzemski, A. (2017a): "An empirical model of dyadic link formation in a network with unobserved heterogeneity," Working paper.

Gao, W. Y. (2020): "Nonparametric Identification in Index Models of Link Formation," *Journal of Econometrics*, 215, 399–413.

Gao, W. Y. and M. Li (2020): "Robust Semiparametric Estimation in Panel Multinomial Choice Models," *Working Paper*.

Goldsmith-Pinkham, P. and G. W. Imbens (2013a): "Social networks and the identification of peer effects," *Journal of Business & Economic Statistics*, 31, 253–264.

Graham, B. S. (2015a): "Methods of identification in social networks," *Annual Review of Economics*, 7, 465–485.

——— (2016a): "Homophily and transitivity in dynamic network formation," Tech. rep., National Bureau of Economic Research.

——— (2017a): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063.

Hart, S. (1985): "Nontransferable utility games and markets: some examples and the Harsanyi solution," *Econometrica: Journal of the Econometric Society*, 1445–1450.

Hsieh, C.-S. and L. F. Lee (2016a): "A social interactions model with endogenous friendship formation and selectivity," *Journal of Applied Econometrics*, 31, 301–319.

Jackson, M. O. and A. Wolinsky (1996a): "A strategic model of social and economic networks," *Journal of Economic Theory*, 71, 44–74.

Jochmans, K. (2017a): "Semiparametric analysis of network formation," *Journal of Business & Economic Statistics*, 1–9.

——— (2017): "Estimation of peer effects in endogenous social networks: control function approach," Working paper.

Leung, M. (2015a): "A Random-Field Approach to Inference in Large Models of Network Formation," Working paper.

McLean, R. P. (2002): "Values of non-transferable utility games," *Handbook of Game Theory with Economic Applications*, 3, 2077–2120.

——— (2017a): "A structural model of dense network formation," *Econometrica*, 85, 825–850.

——— (2017c): "A Structural Model of Homophily and Clustering in Social Networks," Working paper.

——— (2015b): "Strategic network formation with many agents," Tech. rep., New York University.

Moody, J. (2001): "Race, school integration, and friendship segregation in America," *American journal of Sociology*, 107, 679–716.

Newey, W. K. and D. McFadden (1994b): "Chapter 36 large sample estimation and hypothesis testing. volume 4 of Handbook of Econometrics," .

Nolan, D. and D. Pollard (1987): "U-processes: rates of convergence," *The Annals of Statistics*, 780–799.

Ridder, G. and S. Sheng (2017a): "Estimation of large network formation games," Tech. rep., Working papers, UCLA.

Serfling, R. J. (2009): *Approximation theorems of mathematical statistics*, vol. 162, John Wiley & Sons.

Sherman, R. P. (1994): "Maximal inequalities for degenerate U-processes with applications to optimization estimators," *The Annals of Statistics*, 439–459.

Shi, Z. and X. Chen (2016): "A Structural Network Pairwise Regression Model with Individual Heterogeneity," *CUHK Working Paper.*

Toth, P. (2017): "Semiparametric estimation in network formation models with homophily and degree heterogeneity," SSRN 2988698.

Xu, Y. and L. Fan (2018): “Diverse friendship networks and heterogeneous peer effects on adolescent misbehaviors,” *Education Economics*, 26, 233–252.

——— (2018b): “Statistical inference in a directed network model with covariates,” *arXiv preprint arXiv:1609.04558v4*.

Yan, T., C. Leng, and J. Zhu (2016): “Asymptotics in directed exponential random graph models with an increasing bi-degree sequence,” *The Annals of Statistics*, 44, 31–57.

Yan, T. and J. Xu (2013a): “A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices,” *Biometrika*, 100, 519–524.

Appendix

3.A Proofs

3.A.1 Proof of Lemma 3.2

Proof. For notational simplicity, we denote $\Delta(x; x_i, x_j)$ to be $w(x_i, x) - w(x_j, x)$ and d_β by m. It follows that

$$\lambda_{ij}\left(\overline{x}, \underline{x}; x_i, x_j; \beta\right) = \mathbb{1}\left\{\Delta(\overline{x}; x_i, x_j)'\beta \leq 0\right\}\mathbb{1}\left\{\Delta(\underline{x}; x_i, x_j)'\beta \geq 0\right\}. \tag{3.23}$$

Therefore, the event (3.16) is equivalent to $\left\{\Delta(\overline{x}; x_i, x_j)'\beta_0 > 0\right\} \cup \left\{\Delta(\underline{x}; x_i, x_j)'\beta_0 < 0\right\}$ and the event (3.17) is equivalent to $\left\{\Delta(\overline{x}; x_i, x_j)'\beta \leq 0\right\} \cap \left\{\Delta(\underline{x}; x_i, x_j)'\beta \geq 0\right\}$. By Assumption 3.3, there exist x_i and x_j in $Supp\left(X_i\right)$ such that $\Delta\left(X_k; x_i, x_j\right)$ has full directional support. Hence, given any β_0 and $\beta \neq \beta_0$ in $\mathbb{S}^{m-1}$, there exists some $\overline{x} \in Supp\left(X_i\right)$ such that

$$\Delta\left(\overline{x}; x_i, x_j\right)'\beta_0 > 0 \text{ AND } \Delta\left(\overline{x}; x_i, x_j\right)'\beta \leq 0,$$

and some $\underline{x} \in Supp\left(X_i\right)$ such that

$$\Delta\left(\underline{x}; x_i, x_j\right)'\beta_0 < 0 \text{ AND } \Delta\left(\underline{x}; x_i, x_j\right)'\beta \geq 0.$$

Hence, (3.16) and (3.17) hold simultaneously with strictly positive probability. Denote the set of $(x_i, x_j, \overline{x}, \underline{x})$ satisfying these restrictions by

$$\Xi := \left\{ (x_i, x_j, \overline{x}, \underline{x}) \;\middle|\; \begin{array}{c} \Delta(\overline{x}; x_i, x_j)'\beta_0 > 0,\ \Delta(\underline{x}; x_i, x_j)'\beta_0 < 0, \\ \Delta(\overline{x}; x_i, x_j)'\beta \leq 0,\ \text{and } \Delta(\underline{x}; x_i, x_j)'\beta \geq 0. \end{array} \right\}. \tag{3.24}$$

Note that Ξ_{ij} occurs with strictly positive probability.

For such a combination of x_i, x_j, $\overline{x}$, and $\underline{x}$, we show next the event (3.15) holds with strictly positive probability. According to the fact that $\left\{\Delta(\overline{x}; x_i, x_j)'\beta_0 > 0\right\}$ holds for x_i, x_j, $\overline{x}$, and $\underline{x}$, under Assumption 3.5 there exists some $\epsilon_1 > 0$ such that $\rho_i(\overline{x}) > \rho_j(\overline{x})$ whenever $|A_i - A_j| \leq \epsilon_1$. This is true because when the difference between A_i and A_j is small enough, the relative magnitude of $\rho_i(\overline{x})$ compared to $\rho_j(\overline{x})$ will be solely determined by whether $\Delta(\overline{x}; x_i, x_j)'\beta_0 > 0$ or not according to (3.7). Similarly, there exists some $\epsilon_2 > 0$ such that $\rho_i(\underline{x}) < \rho_j(\underline{x})$ whenever $|A_i - A_j| \leq \epsilon_2$. Thus, there exists some $\epsilon := \min\{\epsilon_1, \epsilon_2\}$ such that

$$\begin{aligned} \mathbb{P}\{\tau_{ij}(\overline{x}, \underline{x}) = 1\} &\geq \mathbb{P}\{|A_i - A_j| \leq \epsilon, (x_i, x_j, \overline{x}, \underline{x}) \in \Xi\} \\ &= \mathbb{P}\{|A_i - A_j| \leq \epsilon | (x_i, x_j, \overline{x}, \underline{x}) \in \Xi\}\, \mathbb{P}\{(x_i, x_j, \overline{x}, \underline{x}) \in \Xi\} \\ &> 0, \end{aligned} \tag{3.25}$$

where the first inequality holds by $\{|A_i - A_j| \leq \epsilon, (x_i, x_j, \overline{x}, \underline{x}) \in \Xi\}$ is sufficient for $\{\tau_{ij}(\overline{x}, \underline{x}) = 1\}$ and the last inequality holds by Assumption 3.4.

Therefore, we conclude the three events (3.15), (3.16), and (3.17), hold simultaneously with strictly positive probability for some x_i, x_j, $\overline{x}$, and $\underline{x}$ all in the support of X. □

3.A.2 Proof of Theorem 3.1

Proof. By Lemma 3.1, we have $\beta_0 \in \arg\min_{\beta \in \mathbb{S}^{m-1}} Q(\beta)$ because $Q(\beta_0) = 0 \leq Q(\beta)$ by the construction of the population criterion $Q(\cdot)$. Furthermore, we have β_0 is the unique minimizer of $Q(\beta)$ because for any $\beta \neq \beta_0$, we have

$$\begin{aligned} Q(\beta) &= \mathbb{E}\left[\lambda_{ij}\left(\overline{x}, \underline{x}; x_i, x_j; \beta\right) \tau_{ij}\left(\overline{x}, \underline{x}\right)\right] \\ &= \mathbb{P}\left\{\left\{\lambda_{ij}\left(\overline{x}, \underline{x}; x_i, x_j; \beta\right) = 1\right\} \cap \left\{\tau_{ij}\left(\overline{x}, \underline{x}\right) = 1\right\}\right\} > 0, \end{aligned} \tag{3.26}$$

where the first equality holds by (3.14) and the last inequality holds by Lemma 3.2.

Next, we show that $\mathbb{S}^{m-1}$ is a compact set and $Q(\beta)$ is continuous on $\mathbb{S}^{m-1}$, which together with the uniqueness of β_0 shown in (3.26) guarantee the identification result holds by Newey and McFadden (1994b). The former claim is true by the definition of $\mathbb{S}^{m-1}$. To prove the continuity of $Q(\beta)$, define

$$g_{ij}(z, \beta) := \lambda_{ij}\left(\overline{x}, \underline{x}; x_i, x_j; \beta\right) \tau_{ij}\left(\overline{x}, \underline{x}\right) \tag{3.27}$$

and let z denote $(\overline{x}, \underline{x}; x_i, x_j)$. Following Newey and McFadden (1994b), the sufficient condition for the continuity of $Q(\beta)$ is

(i) $\mathbb{P}\left\{g_{ij}(z, \beta) \text{ is continuous at } \beta = \beta^*\right\} = 1$ for every $\beta^* \in \mathbb{S}^{m-1}$, and

(ii) $\mathbb{E} \sup_{\beta \in \mathbb{S}^{m-1}} \left|g_{ij}(z, \beta)\right| < \infty$.

Part (i) is true because $\lambda_{ij}(\overline{x}, \underline{x}; x_i, x_j; \beta)$ is a binary function of $z = (\overline{x}, \underline{x}; x_i, x_j)$ and the change in value from 0 to 1 or from 1 to 0 only occurs when $d(\overline{x}; x_i, x_j)'\beta = 0$ or $d(\underline{x}; x_i, x_j)'\beta = 0$. Under Assumption 3.3, these two events have zero probability of happening. Thus, part (i) is verified. For part (ii), note that by construction $g_{ij}(z, \beta) \in \{0, 1\}$ is a bounded function of β for all z. Therefore,

$$\mathbb{E} \sup_{\beta \in \mathbb{S}^{m-1}} \left|g_{ij}(z, \beta)\right| \leq 1 < \infty. \tag{3.28}$$

Hence we have for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\inf_{\beta \in \mathbb{S}^{m-1} \backslash B(\beta_0, \epsilon)} Q(\beta) \geq Q(\beta_0) + \delta, \tag{3.29}$$

where $B(\beta_0, \epsilon) := \left\{\beta \in \mathbb{S}^{m-1} : \|\beta - \beta_0\| \leq \epsilon\right\}$. □

3.A.3 Proof of Lemma 3.3

Proof. Define the infeasible criterion $\widetilde{Q}_n(\beta)$ as

$$\widetilde{Q}_n(\beta) := \frac{(n-4)!}{n!} \sum_{1 \leq i \neq j \neq k \neq l \leq n} \mathbb{1}\{\rho_i(X_k) > \rho_j(X_k)\} \cdot \mathbb{1}\{\rho_i(X_l) < \rho_j(X_l)\} \times \left[\begin{array}{c} \mathbb{1}\left\{d(X_k; X_i, X_j)'\beta \leq 0\right\} \\ \times \mathbb{1}\left\{d(X_l; X_i, X_j)'\beta \geq 0\right\} \end{array} \right]. \tag{3.30}$$

By triangular inequality, we have

$$\sup_{\beta \in \mathbb{S}^{m-1}} \left|\widehat{Q}_n(\beta) - Q(\beta)\right| \leq \sup_{\beta \in \mathbb{S}^{m-1}} \left|\widehat{Q}_n(\beta) - \widetilde{Q}(\beta)\right| + \sup_{\beta \in \mathbb{S}^{m-1}} \left|\widetilde{Q}(\beta) - Q(\beta)\right|. \tag{3.31}$$

According to the decomposition (3.31), we divide our proof into two steps.

Step 1. $\sup_{\beta \in \mathbb{S}^{m-1}} \left|\widehat{Q}_n(\beta) - \widetilde{Q}(\beta)\right| \xrightarrow{p} 0$.

By the fact that $\lambda_{ij}(X_k, X_l; X_i, X_j; \beta)$ is either 0 or 1 for any $\beta \in \mathbb{S}^{m-1}$, we have

$$
\begin{aligned}
&\sup_{\beta\in\mathbb{S}^{m-1}}\left|\widehat{Q}_n(\beta)-\widetilde{Q}(\beta)\right| \\
&=\frac{(n-4)!}{n!}\sum_{1\le i\neq j\neq k\neq l\le n}\sup_{\beta\in\mathbb{S}^{m-1}}\left|\lambda_{ij}(X_k,X_l;X_i,X_j;\beta)\right| \\
&\quad\times\left|\begin{array}{c}\mathbb{1}\{\rho_i(X_k)>\rho_j(X_k)\}\cdot\mathbb{1}\{\rho_i(X_l)<\rho_j(X_l)\}\\ -\mathbb{1}\{\hat\rho_i(X_k)>\hat\rho_j(X_k)\}\cdot\mathbb{1}\{\hat\rho_i(X_l)<\hat\rho_j(X_l)\}\end{array}\right| \qquad (3.32)\\
&\le\frac{(n-4)!}{n!}\sum_{1\le i\neq j\neq k\neq l\le n}\left|\begin{array}{c}\mathbb{1}\{\rho_i(X_k)>\rho_j(X_k)\}\cdot\mathbb{1}\{\rho_i(X_l)<\rho_j(X_l)\}\\ -\mathbb{1}\{\hat\rho_i(X_k)>\hat\rho_j(X_k)\}\cdot\mathbb{1}\{\hat\rho_i(X_l)<\hat\rho_j(X_l)\}\end{array}\right| \\
&\le\frac{(n-4)!}{n!}\sum_{1\le i\neq j\neq k\neq l\le n}\left[\begin{array}{c}\left|\mathbb{1}\{\rho_i(X_k)>\rho_j(X_k)\}-\mathbb{1}\{\hat\rho_i(X_k)>\hat\rho_j(X_k)\}\right|\\ +\left|\mathbb{1}\{\rho_i(X_l)<\rho_j(X_l)\}-\mathbb{1}\{\hat\rho_i(X_l)<\hat\rho_j(X_l)\}\right|\end{array}\right],
\end{aligned}
$$

where the first inequality uses $|\lambda_{ij}(X_k,X_l;X_i,X_j;\beta)|$ is bounded from above by 1 and the last inequality uses the fact that whenever the LHS of the last inequality equals 1, the RHS must always equals 1.

It follows that

$$
\begin{aligned}
&\mathbb{E}\sup_{\beta\in\mathbb{S}^{m-1}}\left|\widehat{Q}_n(\beta)-\widetilde{Q}(\beta)\right| \\
&\le\mathbb{E}\left|\mathbb{1}\{\rho_i(X_k)>\rho_j(X_k)\}-\mathbb{1}\{\hat\rho_i(X_k)>\hat\rho_j(X_k)\}\right| \\
&\quad+\mathbb{E}\left|\mathbb{1}\{\rho_i(X_l)<\rho_j(X_l)\}-\mathbb{1}\{\hat\rho_i(X_l)<\hat\rho_j(X_l)\}\right| \qquad (3.33)
\end{aligned}
$$

By Assumption 3.6, we obtain

$$
\mathbb{E}\sup_{\beta\in\mathbb{S}^{m-1}}\left|\widehat{Q}_n(\beta)-\widetilde{Q}(\beta)\right|\to 0 \qquad (3.34)
$$

using Dominated Convergence Theorem.

Finally, by Markov inequality, we have

$$\sup_{\beta\in\mathbb{S}^{m-1}}\left|\widehat{Q}_n\left(\beta\right)-\widetilde{Q}\left(\beta\right)\right|\xrightarrow{p}0. \tag{3.35}$$

Step 2. $\sup_{\beta\in\mathbb{S}^{m-1}}\left|\widetilde{Q}_n\left(\beta\right)-Q\left(\beta\right)\right|\xrightarrow{p}0.$

For this part of the proof, we adapt to section 9.5 of Toth (2017) and use existing results from the U-process literature. We have $\left\{\widetilde{Q}_n\left(\beta\right)-Q\left(\beta\right):\beta\in\mathbb{S}^{m-1}\right\}$ is a centered U-process of order 4. We follow the arguments from the seminal papers Nolan and Pollard (1987) and Sherman (1994). For a systematic understanding of U-statistics, we refer the readers to Serfling (2009).

First, we show $\left\{\widetilde{Q}_n\left(\beta\right)-Q\left(\beta\right):\beta\in\mathbb{S}^{m-1}\right\}$ is Euclidean for the constant envelope of 1 (See Definition 8 in Nolan and Pollard (1987)). To see why, first note that the unsymmetrized kernel of $\widetilde{Q}_n\left(\beta\right)-Q\left(\beta\right)$ for any $\beta\in\mathbb{S}^{m-1}$ is defined to be

$$\begin{aligned} kernel :=\ & \lambda_{ij}\left(X_k,X_l;X_i,X_j;\beta\right)\mathbb{1}\left\{\rho_i(X_k)>\rho_j(X_k)\right\}\\ & \times\cdot\mathbb{1}\left\{\rho_i(X_l)<\rho_j(X_l)\right\}\\ & -\mathbb{E}\Big[\tau_{ij}\left(X_k,X_l\right)\cdot\lambda_{ij}\left(X_k,X_l;X_i,X_j;\beta\right)\Big]. \end{aligned} \tag{3.36}$$

The kernel defined in (3.36) belongs to a Euclidean class if and only if the function class of $\lambda_{ij}\left(X_k,X_l;X_i,X_j;\beta\right)$ indexed by β is Euclidean because the property is closed under finite addition, multiplication and linear operations, see Nolan and Pollard (1987). By (3.23), we have

$$\lambda_{ij}\left(X_k,X_l;X_i,X_j;\beta\right)=\mathbb{1}\left\{d(X_k;X_i,X_j)^{'}\beta\leq 0\right\}\mathbb{1}\left\{d(X_l;X_i,X_j)^{'}\beta\geq 0\right\}. \tag{3.37}$$

Note that the function class of $\lambda_{ij}\left(X_k, X_l; X_i, X_j; \beta\right)$ indexed by β is Euclidean if and only if the function class of $d(X_k; X_i, X_j)^{'}\beta$ is Euclidean, again by closure under finite multiplication and indicator functions.

Define the function class $\mathscr{G}$ of $g\left(X; Y, Z\right) := d(X; Y, Z)^{'}\beta$ to be

$$\mathscr{G} := \left\{ d(X; Y, Z)^{'}\beta \middle| \beta \in \mathbb{S}^{m-1} \right\}. \tag{3.38}$$

We have $\mathscr{G}$ forms a finite dimensional vector space of functions as long as $m < \infty$. By Lemma 18 of Nolan and Pollard (1987), the collection of all sets of the form $\{g \geq 0\}$ or $\{g \leq 0\}$ or $\{g > 0\}$ or $\{g < 0\}$ for any $g \in \mathscr{G}$ is a polynomial class, which implies $\{\text{graph}\left(g\right) : g \in \mathscr{G}\}$ is a polynomial class of sets because any class of subsets of $\mathbb{R}$ is a polynomial class. From this result and Lemma 19 of Nolan and Pollard (1987), we have $\mathscr{G}$ is Euclidean. Therefore, the kernel defined in (3.36) indeed belongs to a Euclidean class, and according to Corollary 7 in Sherman (1994), we have

$$\sup_{\beta \in \mathbb{S}^{m-1}} \left| \widetilde{Q}_n\left(\beta\right) - Q\left(\beta\right) \right| \xrightarrow{p} 0. \tag{3.39}$$

Combining (3.35) and (3.39), we have

$$\sup_{\beta \in \mathbb{S}^{m-1}} \left| \widehat{Q}_n\left(\beta\right) - Q\left(\beta\right) \right| \xrightarrow{p} 0. \tag{3.40}$$

□

3.A.4 Proof of Theorem 3.2

Proof. We aim to prove, for any $\epsilon > 0$, $\mathbb{P}\left(\|\widehat{\beta}_n - \beta_0\|\right) > \epsilon \to 0$. According to the proof in Theorem 3.1, we have for any $\epsilon > 0$, there exists $\delta > 0$ such that $\inf_{\beta \in \mathbb{S}^{m-1} \setminus B_m(\beta_0, \epsilon)} Q\left(\beta\right) \geq Q\left(\beta_0\right) + \delta$, where $B_m(\beta_0, \epsilon) = \left\{\beta \in \mathbb{S}^{m-1} : \|\beta - \beta_0\| \leq \epsilon\right\}$. It follows that there exist $\delta > 0$ such that

$$\mathbb{P}\left(\|\widehat{\beta}_n - \beta_0\| > \epsilon\right) = \mathbb{P}\left(\widehat{\beta}_n \in \mathbb{S}^{m-1} \backslash B_m\left(\beta_0, \epsilon\right)\right) \leq \mathbb{P}\left(Q\left(\widehat{\beta}_n\right) \geq Q\left(\beta_0\right) + \delta\right). \tag{3.41}$$

By construction of $\widehat{\beta}_n$, we have $\widehat{Q}_n(\widehat{\beta}_n) - \widehat{Q}_n(\beta_0) \leq 0$. Therefore,

$$\begin{aligned} &\mathbb{P}\left(Q(\widehat{\beta}_n) \geq Q(\beta_0) + \delta\right) \\ &= \mathbb{P}\Big(Q(\widehat{\beta}_n) - \widehat{Q}_n(\widehat{\beta}_n) + \widehat{Q}_n(\widehat{\beta}_n) - \widehat{Q}_n(\beta_0) + \widehat{Q}_n(\beta_0) - Q(\beta_0) \geq \delta\Big) \\ &\leq \mathbb{P}\Big(Q(\widehat{\beta}_n) - \widehat{Q}_n(\widehat{\beta}_n) + 0 + \widehat{Q}_n(\beta_0) - Q(\beta_0) \geq \delta\Big). \end{aligned} \tag{3.42}$$

It follows that

$$\mathbb{P}\left(Q(\widehat{\beta}_n) \geq Q(\beta_0) + \delta\right) \leq \mathbb{P}\left(\sup_{\beta \in \mathbb{S}^{m-1}} \left|\widehat{Q}_n\left(\beta\right) - Q\left(\beta\right)\right| \geq \delta/2\right). \tag{3.43}$$

By Lemma 3.3, we have for any $\delta > 0$

$$\mathbb{P}\left(\sup_{\beta \in \mathbb{S}^{m-1}} \left|\widehat{Q}_n\left(\beta\right) - Q\left(\beta\right)\right| \geq \delta/2\right)\Big) \to 0 \text{ as } n \to \infty. \tag{3.44}$$

Therefore, we have for any $\epsilon > 0$

$$\mathbb{P}\left(\|\widehat{\beta}_n - \beta_0\| > \epsilon\right) \leq \mathbb{P}\left(\sup_{\beta \in \mathbb{S}^{m-1}} \left|\widehat{Q}_n\left(\beta\right) - Q\left(\beta\right)\right| \geq \delta/2\right) \to 0 \text{ as } n \to \infty. \tag{3.45}$$

□

3.B Asymmetry of Pairwise Observable Characteristics

So far we have been focusing on the case with symmetric pairwise observable characteristics, i.e.,

$$w\left(X_i, X_j\right) \equiv w\left(X_j, X_i\right).$$

In this section, we briefly discuss how our method can be adapted to accommodate asymmetric pairwise observable characteristics.

As in Remark 3.1, consider the adapted model (3.6):

$$\mathbb{E}\left[D_{ij}\middle| X_i, X_j, A_i, A_j\right] = \phi\left(w\left(X_i, X_j\right)' \beta_0, w\left(X_j, X_i\right)' \beta_0, A_i, A_j\right) \tag{3.46}$$

where w needs not be symmetric with respect to its two vector arguments and $\phi : \mathbb{R}^4 \to \mathbb{R}$ is required to be monotone in all its four arguments.

The technique of logical differencing still applies in the exactly same way as before. Specifically, the event $\left\{\rho_{\bar{i}}\left(\overline{x}\right) > \rho_{\bar{j}}\left(\overline{x}\right)\right\}$ implies that

$$\left\{w\left(X_{\bar{i}}, \overline{x}\right)' \beta_0 > w\left(X_{\bar{j}}, \overline{x}\right)' \beta_0\right\} \text{ OR } \left\{w\left(\overline{x}, X_{\bar{i}}\right)' \beta_0 > w\left(\overline{x}, X_{\bar{j}}\right)' \beta_0\right\} \text{ OR } \left\{A_{\bar{i}} > A_{\bar{j}}\right\},$$

while the event $\left\{\rho_{\bar{i}}\left(\underline{x}\right) < \rho_{\bar{j}}\left(\underline{x}\right)\right\}$ implies that

$$\left\{w\left(X_{\bar{i}}, \underline{x}\right)' \beta_0 < w\left(X_{\bar{j}}, \underline{x}\right)' \beta_0\right\} \text{ OR } \left\{w\left(\underline{x}, X_{\bar{i}}\right)' \beta_0 < w\left(\underline{x}, X_{\bar{j}}\right)' \beta_0\right\} \text{ OR } \left\{A_{\bar{i}} < A_{\bar{j}}\right\}.$$

The joint occurrence of $\left\{\rho_{\bar{i}}\left(\overline{x}\right) > \rho_{\bar{j}}\left(\overline{x}\right)\right\}$ and $\left\{\rho_{\bar{i}}\left(\underline{x}\right) < \rho_{\bar{j}}\left(\underline{x}\right)\right\}$ now implies that

$$\begin{aligned} &\left\{w\left(X_{\bar{i}}, \overline{x}\right)' \beta_0 > w\left(X_{\bar{j}}, \overline{x}\right)' \beta_0\right\} \text{ OR } \left\{w\left(\overline{x}, X_{\bar{i}}\right)' \beta_0 > w\left(\overline{x}, X_{\bar{j}}\right)' \beta_0\right\} \\ \text{OR } &\left\{w\left(X_{\bar{i}}, \underline{x}\right)' \beta_0 < w\left(X_{\bar{j}}, \underline{x}\right)' \beta_0\right\} \text{ OR } \left\{w\left(\underline{x}, X_{\bar{i}}\right)' \beta_0 < w\left(\underline{x}, X_{\bar{j}}\right)' \beta_0\right\}, \end{aligned} \tag{3.47}$$

which is in general “less restrictive” than the corresponding identifying restriction in Lemma 3.1.

In particular, in the extreme case where w is *antisymmetric* in the sense of

$$w\left(X_i, X_j\right) \equiv -w\left(X_j, X_i\right),$$

the identifying restriction on the RHS of

$$\left\{w\left(X_{\bar{i}}, \overline{x}\right)' \beta_0 > w\left(X_{\bar{j}}, \overline{x}\right)' \beta_0\right\} \text{ OR } \left\{w\left(\overline{x}, X_{\bar{i}}\right)' \beta_0 > w\left(\overline{x}, X_{\bar{j}}\right)' \beta_0\right\}$$

becomes

$$\left\{w\left(X_{\bar{i}}, \overline{x}\right)' \beta_0 \neq w\left(X_{\bar{j}}, \overline{x}\right)' \beta_0\right\},$$

which can be generically true and thus becomes (almost) trivial.

Correspondingly, Assumption 3.3 needs to be strengthened for point identification:

Assumption 3.3’. *There exist a pair of $\overline{x}, \underline{x}$, both of which lie in the support of $Supp\left(X_i\right)$, such that*

$$Supp\left(w\left(\overline{x}, X_i\right) - w\left(\underline{x}, X_i\right)\right) \cap Supp\left(w\left(X_i, \overline{x}\right) - w\left(X_i, \underline{x}\right)\right)$$

contains all directions in $\mathbb{R}^m$.

Clearly, the case of antisymmetric w is ruled out by Assumption 3.3’. Assumption 3.3’ ensures that, for any $\beta \neq \beta_0$, there exist in-support x_i and x_j such that

$$\begin{aligned} &\left\{w\left(x_i, X_k\right)' \beta_0 > w\left(x_j, X_k\right)' \beta_0\right\} \text{ AND } \left\{w\left(x_i, X_l\right)' \beta_0 < w\left(x_j, X_l\right)' \beta_0\right\} \\ \text{AND} &\left\{w\left(X_k, x_i\right)' \beta_0 > w\left(X_k, x_j\right)' \beta_0\right\} \text{ AND } \left\{w\left(X_l, x_i\right)' \beta_0 < w\left(X_l, x_j\right)' \beta_0\right\} \end{aligned} \tag{3.48}$$

and

$$\left\{w\left(x_i, X_k\right)' \beta \leq w\left(x_j, X_k\right)' \beta\right\} \text{ AND } \left\{w\left(x_i, X_l\right)' \beta \geq w\left(x_j, X_l\right)' \beta\right\}$$

$$\text{AND}\left\{w\left(X_k, x_i\right)^{'}\beta \leq w\left(X_k, x_j\right)^{'}\beta\right\}\ \text{AND}\ \left\{w\left(X_l, x_i\right)^{'}\beta \geq w\left(X_l, x_j\right)^{'}\beta\right\} \tag{3.49}$$

occur simultaneously with strictly positive probability. (3.48) and (3.49) are sufficient for $\left\{\rho_{\bar{i}}\left(\overline{x}\right) > \rho_{\bar{j}}\left(\overline{x}\right)\right\}$ and $\left\{\rho_{\bar{i}}\left(\underline{x}\right) < \rho_{\bar{j}}\left(\underline{x}\right)\right\}$ to occur simultaneously under the maintained assumption on the support of A_i. It thus can guarantee point identification of β_0.

The estimator can be correspondingly adapted in an obvious manner.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA