UNIVERSITY OF CALIFORNIA

Los Angeles

Identification and Estimation in Semiparametric Social Interaction Models

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

by

Nan Liu

2020

ABSTRACT OF THE DISSERTATION

Identification and Estimation in Semiparametric
Social Interaction Models

by

Nan Liu
Doctor of Philosophy in Economics
University of California, Los Angeles, 2020
Professor Jinyong Hahn, Co-Chair
Professor Zhipeng Liao, Co-Chair

This dissertation investigates semiparametric social interaction models. The goal is to identify and estimate the endogenous social interactions using a flexible semiparametric model to control for confounding factors. The rationale for considering nonparametric controls is that, if the groups or networks are not randomly assigned, or if the contextual effects are heterogeneous, identifying the endogenous social interaction effect is difficult without adequate controls. This thesis contains two chapters.

Chapter 1 first studies the identification of the endogenous social interaction effect in the semiparametric models. The identification is attained by using the instrumental variable (IV) approach after partialling out the nonparametric controls. To estimate the endogenous social interaction effect, I propose a semiparametric two-step generalized method of moments (GMM) estimator with the optimal weight matrix clustered at the group or network level. This chapter focuses on the semiparametric estimators that use the first step series method. The primitive regularity conditions are provided for the consistency and asymptotic normality of the semiparametric series GMM estimators.

In Chapter 2, I apply more flexible machine learning methods in the first step nonparametric estimation to detect severe nonlinearities and higher-order interactions, including LASSO, Random Forest, and Neural Nets. Monte Carlo simulations are conducted to investigate the finite sample performance of semiparametric estimators using different first-step Machine Learning methods. The results suggest that no estimation method dominates across all the Data Generating Processes (DGPs) considered. It is also reflected in the simulation results that the debiased estimators using first step post-LASSO or Neural Nets methods are more reliable and performs relatively well across the settings considered. For this reason, these two debiased estimators are recommended for use in empirical studies.

The dissertation of Nan Liu is approved.

Denis Nikolaye Chetverikov

Rosa Liliana Matzkin

Shuyang Sheng

Ying Nian Wu

Jinyong Hahn, Committee Co-Chair

Zhipeng Liao, Committee Co-Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my PhD advisors, Prof. Jinyong Hahn and Zhipeng Liao, for their guidance and support throughout my doctoral study, for their patience, motivation and immense knowledge. I am grateful for their valuable feedback, advice, and encouragement when writing this thesis.

My sincere thanks also goes to Prof. Shuyang Sheng, Rosa Matzkin, Denis Chetverikov, Andres Santos, and Yingnian Wu, for giving the encouragement and sharing insightful suggestions. I would also like to express my special thanks to Whitney K. Newey for his helpful suggestions.

I would like to thank all my PhD colleagues, for the great moments that we have shared at UCLA. I would also like to express special thanks to my friends Lina Zhang, Yarong Chen, and Yanbo Liu for their emotional support in those unsettled times.

I am deeply thankful to my parents and bother, for their unconditional love, encouragement, and support. I would not have made it this far without them.

VITA

Education

University of California, Los Angeles	Los Angeles, USA
- C.Phil. Economics, Department of Economics	*2016*
Peking University	Beijing, China
- M.A. Economics, National School of Development	*2014*
Renmin University of China	Beijing, China
-B.A. Agricultural Economics	*2010*
-B.S. Mathematics	*2010*

Fellowships and Awards

UCLA Graduate Fellowship.	*2014-2019*
Academic Excellence Award of Peking University.	*2012-2013*
The First Prize Scholarship of Peking University.	*2011-2012*
The First Prize Scholarship of Renmin University.	*2008-2009*

Publications

"Significance test in nonstationary logit panel model with serially correlated dependent variable", with Chia-Shang J. Chu and Lina Zhang. *Economics Letters*, Volume 159, 2017, PP. 37-41.

"Significance test in nonstationary multinomial logit model", with Chia-Shang J. Chu and Lina Zhang. *Economics Letters*, Volume 143, 2016, PP. 94-98.

Research Assistant

Human Capital, Economic Growth and Mid-income Trap.	*2014*
China Health and Retirement Longitudinal Study (CHARLS).	*2011*

Teaching Assistant

Teaching Assistant, Department of Economics, UCLA.

-Econometrics III (Graduate Course), *Spring 2017*

-Introduction to Econometrics, *Spring 2019; Fall 2019; Winter 2018, 2019*

-Microeconomic Theory, *Spring 2016; Fall 2016*

-Statistics for Economists, *Fall 2015, 2017, 2018; Winter 2016, 2017*

-Principles of Economics, *Spring 2018*

Teaching Assistant, National School of Development, Peking University.

-Advanced Econometrics II (Graduate Course), *Spring 2014*

-Introduction to Econometrics, *Spring 2013*

CHAPTER 1

Identification and Estimation in Semiparametric Social Interaction Models

1.1 Introduction

It is commonly observed that individual outcomes (e.g. GPA) are highly correlated with the outcomes of their peers (e.g. class average). Empirical researchers are interested in identifying the social interaction effects (or peer effects) out of these group correlations. For instance, many studies have investigated peer influence in education (Sacerdote, 2001; Duflo et al., 2011), criminal activity (Glaeser et al., 1996), welfare participation (Bertrand et al., 2000), and the job search (Marmaros and Sacerdote, 2002).

The main challenge in these empirical studies arises from the fact that a variety of reasons can lead to the outcome correlation among individuals and their peers. This problem might lead to spurious or misleading measure of social interactions. A canonical example is classroom peer effects. A student's test score might depend on the class average, which Manski (1993) terms the endogenous social interaction effect. Other influential factors could be the student and class characteristics such as parental income, and also the average parental income in what Manski (1993) terms the contextual effect (or exogenous social interaction effect). Both the endogenous and contextual effects tend to cause students' performance to be correlated. Manski (1993) proposes the reflection problem which states the impossibility of separately identify the endogenous and contextual effects in his setup.

Subsequent researchers have extended Manski (1993)'s study in various ways to resolve the reflection problem. The commonly used approach relies on the parametric assumption or functional form restrictions to distinguish the endogenous social interaction and contextual effects. Refer to Brock and Durlauf (2001), Graham (2008), Bramoullé, Djebbari, and Fortin (2009), and a detailed literature review in Blume et al. (2011). However, a fully parametric model is often too restrictive to capture the structure of social interactions. Estimators based on this identification strategy can be severely biased if there are modest departures from the parametric assumptions (Goldberger, 1983).

To address the above consideration, a more flexible semiparametric model can be applied to capture the social interactions between individuals and their peer groups. This paper focuses on the identification and estimation of the endogenous social interaction effect in this semiparametric model setup. The endogenous social interaction effect is assumed to be parametric and also linear-in-means. This setup is applied in much of the literature on peer effects since Manski (1993), especially in the empirical studies of the education field (Angrist and Lang, 2004). But unlike the existing literature, this paper uses a more robust nonparametric function to capture the effects of confounding factors, such as the contextual effects. By introducing the nonparametric function for covariates, the social interaction model can control the interactions and other transformations such as higher-orders of the individual and group characteristics, making the identification of endogenous social interaction more credible. For instance, the average parental income might have a heterogeneous effect on student's performance. Omitting the interaction terms in the parametric model could cause bias in the endogenous social interaction effect. This problem can be addressed by applying the semiparametric model of social interactions.

This paper then presents a semiparametric instrumental variable (IV) approach to identify the endogenous social interaction effect. The models of social interactions involve simultaneity in determining the outcomes. Consequently, the semiparametric identification of endogenous peer effects is different from classical models of treatment effects (Abadie, 2003). Also, in contrast with the parametric social interaction models, the identification approach taken in this paper accommodates the nonparametric controls. Concretely, the semiparametric identification in this paper is achieved by partialling out the nonparametric nuisance function in the first step and then adopting an IV approach to identify the endogenous social interaction effect.

The IV methods have been proposed to recover the endogenous social interaction effect in the parametric models, such as in Brock and Durlauf (2001), Graham and Hahn (2005), and among others[1]. Brock and Durlauf (2001) introduces a particular nonlinear in mean function and establishes the identification for those functions with nonzero second derivatives. Graham and Hahn (2005) constitute exclusion restrictions by assuming not all the exogenous individual variables appear as the contextual effect. This paper does not assume nonlinearity in average group outcome and follows the approach of Graham and Hahn (2005), which impose exclusion restrictions based on prior information. In the example of classmates' peer effects, the excluded contextual factors could be the childbearing age of the parents and whether the students attend kindergarten or not. These variables might determine the student's IQ level, and thus can affect their school performance. In contrast, childhood characteristics of a student's classmates are unlikely to have a direct effect on that student's high school scores. Thus, excluded contextual factors can be used as instrumental variables for the endogenous variable of the average class score.

Based on the identification strategy, this paper proposes a semiparametric two-step procedure to estimate the endogenous social interaction effect. The two-step procedure builds upon the well established semiparametric estimation literature, see Newey (1994), Newey and McFadden (1994), Ai and Chen (2003, 2007), Chen, Linton, and Van Keilegom (2003), Chen (2007), and Ackerberg et al. (2012) among others. The general semiparametric two-step estimation methods are widely studied in the treatment effects literature (Belloni, Chernozhukov, and Hansen, 2014; Athey, Imbens, and Wager, 2016). To the best of my knowledge, no paper discusses the estimation and inference for the semiparametric social interaction model.

This paper fills in the gap between the general semiparametric theory and the specific semiparametric model of social interactions with clustered data. In the first step, conditional expectations can be estimated using any nonparametric methods. In the second step,

[1] A more detailed literature review for the identification of social interaction models is presented in Appendix 1.B.

the endogenous social interaction effect is then estimated using a semiparametric GMM by plugging the first step nonparametric estimators. The orthogonal moment condition is also constructed for second-step estimation to reduce the bias induced by the first step nonparametric estimators.

To establish the asymptotic normality of the semiparametric two-step GMM estimator, this paper specializes in first step series estimator and makes use of the idea in Newey (1994, 1997). Primitive conditions are provided for the consistency and asymptotic normality of the semiparametric two-step estimators for endogenous social interaction effect. The results show that the application of series estimators are limited to low-dimensional settings and relatively smooth nonparametric functions, which is the general limitations for the traditional nonparametric methods.

This paper uses a Monte Carlo simulation approach to investigate the finite sample performance of semiparametric two-step estimators for endogenous social interaction effects. I apply both the parametric linear and nonparametric series methods in the first step to obtain the second step estimators for β_0. The results show that if the true model is nonlinear but the linear method is applied in the first step, then the estimator for β_0 suffers from an over-rejection problem and more likely a significant social interaction effect is obtained which is actually spurious. However, the semiparametric series estimators do not have such an issue and have the correct size across all the nonlinear models considered in the simulation study.

The remainder of this paper is organized as follows. In section 1.2, the semiparametric social interaction model is introduced. Section 1.3 shows semiparametric identification conditions for the endogenous social interaction effect. Section 1.4 utilizes the identification results to construct the moment condition and also the orthogonal moment condition. Section 1.5 proposes a semiparametric two-step GMM estimator for endogenous effect. Primitive conditions for asymptotic normality of the semiparametric estimator are also provided with first step series estimator. Section 1.6 uses cross-fitting to construct a semiparametric estimator for endogenous social interaction effect and shows the asymptotic normality of the cross-fitting estimator under weaker conditions. In section 1.7, Monte Carlo simulations are conducted to investigate the finite sample performance of semiparametric GMM estimators.

Section 1.8 concludes and discusses the directions for future research. Proofs are provided in Appendix 1.A.

1.2 The Semiparametric Social Interaction Model

This paper investigates a semiparametric model of group social interactions. Assume econometricians can observe a random sequence of G non-overlapping groups with the number of groups $G \to \infty$. For the gth group, assume the group size is n_g, which is bounded but allowed to differ across groups. Let $n = \sum_{g=1}^{G} n_g$ denote the total number of individuals.

Assume that the outcome variable of individual i in group g, $Y_{g,i}$ $(g = 1, ..., G,\ i = 1, ..., n_g)$, is determined according to the following semiparametric model of social interactions:

$$Y_{g,i} = \beta_0 \overline{Y}_{g,-i} + h_0(X_{g,i}, W_{g,i}, \overline{W}_{g,-i}, \Upsilon_g) + U_{g,i}; \quad (g = 1, ..., G;\ i = 1, ..., n_g), \tag{1.2.1}$$

where $Y_{g,i}$ is the outcome variable of interest for individual i in group g. $\overline{Y}_{g,-i} = \frac{1}{n_g-1} \sum_{j \in I_g, j \neq i} Y_j$ denotes the leave-i-out average outcome within the gth group. Parametric linear-in-means form for average group outcome is considered in this model. Following the terminology of Manski (1993), the one-dimensional coefficient for $\overline{Y}_{g,-i}$, β_0, captures the endogenous social interaction effect. β_0 is also the parameter of interest in this paper.

Econometricians can also observe vectors of exogenous individual characteristics $X_{g,i}$ and $W_{g,i}$ for i in group g, and also the group characteristics Υ_g. The difference between the individual characteristics $X_{g,i}$ and $W_{g,i}$ comes from whether they can generate contextual effects on the outcome $Y_{g,i}$. To be specific, $X_{g,i}$ denotes the d dimensional ($\dim(X_{g,i}) = d$) individual-specific characteristics that only affect the outcome $Y_{g,i}$ through individual level, which means the leaving-i-out group average $\overline{X}_{g,-i}$ does not affect the outcome directly, while $W_{g,i}$ denotes the d_W dimensional ($\dim(W_{g,i}) = d_W$) individual characteristics which also induce the contextual effect. That is, the leaving-i-out group average for $W_{g,i}$, $\overline{W}_{g,-i}$, is allowed to affect the outcome $Y_{g,i}$. Another contextual factor comes from the d_Υ dimensional observed group characteristics Υ_g ($\dim(\Upsilon_g) = d_\Upsilon$) which can also influence the individual outcome.

This paper relaxes the fully parametric model setup but allows the individual and contextual effects to be nonparametric. Let $h_0(\cdot)$ denote an unknown function that summarizes the effect of $X_{g,i}$, $W_{g,i}, \overline{W}_{g,-i}$ and Υ_g on the outcome $Y_{g,i}$. Then the effects of $\overline{W}_{g,-i}$ and Υ_g on $Y_{g,i}$ through $h_0(\cdot)$ capture the contextual effect following the terminology in Manski (1993). The last term in model (1.2.1), $U_{g,i}$, is the disturbance term that is unobserved to econometricians. This paper will work with the exogeneity condition, which requires $U_{g,i}$ to be independent of all the controls.

Let $\vartheta_0 = (\beta_0, h_0(\cdot))'$ be the true parameter vector. The parameter of interest β_0 is the 1-dimensional parametric part which captures the endogenous social interaction effect. To have a stable equilibrium social interaction model, it is reasonable to assume that $|\beta_0| < 1$. The unknown function $h_0(\cdot)$ is the nonparametric part of the model which captures the individual and contextual effects. This nonparametric function is also a nuisance parameter in this paper.

The goal of this paper is to obtain a root-n consistent estimator for β_0, in the presence of the possibly complex nuisance function $h_0(.)$ and also the endogenous effect of group social interactions. Model (1.2.1) belongs to the semiparametric partially linear model studied in seminal paper of Robinson (1988). The nonparametric function can be concentrated out by using the generalized residual regression. And then the parameter of interest β_0 can be estimated using the GMM approach with properly choose instrumental variables for the endogenous effect.

Motivating Example

To illustrate the specification of semiparametric model (1.2.1) of social interactions, I will use a concrete example of classmates' peer effects on a student's test scores. The object of interest is the effect on a student's test score by the average class score of other students in the same class. Suppose we can observe a random sample of G classrooms with n_g students in the g th class. The test score of student i in class $g, Y_{g,i}$ is subject to the average score of other students in class $g, \overline{Y}_{g,-i}$. The reason might be that a student with poor classmates (lower $\overline{Y}_{g,i}$) tend to make less effort in studies thus could perform worse. This is the endogenous peer effect that this paper focuses on.

The student's test score also depends on individual characteristics, such as parents' income, education level, childbearing age, or whether the student attended kindergarten or not. Model (1.2.1) distinguishes two sources of individual effects, $X_{g,i}$ and $W_{g,i}$, by whether they can generate contextual effect on student i 's test score or not. In this specific example, $W_{g,i}$ can be parents' income or education level since the class average of these variables for other students could also account for a student's test score. For instance, the average parental income for other students in class g could affect the investment in school or class, and thus influence a student's performance, which generates the contextual effect. In contrast, childhood characteristics of a student's classmates, such as parents' childbearing age and whether the student attended kindergarten or not, are unlikely to have a direct effect on a student's high school test scores. These two variables are defined as the individual-only characteristics, $X_{g,i}$, which provide an exclusion restriction for the identification of the endogenous social interaction effect.

Anther key feature of social interaction model (1.2.1) is to introduce a nonparametric function to control the confounding factors that could affect a student's performance. This nonparametric specification is important for the identification of parametric part of endogenous peer effect. One reason is that students might not be randomly assigned to different classes. For instance, students with high entrance scores might be assigned to one class with the teacher has rich experience, then the interaction of students' score and teacher's experience also matters. If not fully controlled, the omitted factors become the correlated effects termed in Manski (1993), which could then contaminate the endogenous social interaction effect.

Nonparametric controls can also be used to better capture the heterogeneous individual or contextual effects. For example, the average parental income might has a heterogeneous effect on student's performance, which might be negative for students from high income family but positive for low income's. Then the interaction term should also be included. By introducing the nonparametric function for covariates, the social interaction model is able to control the interactions and other transformations such as higher orders of the individual and group characteristics, making the identification of endogenous classmates' peer effect more credible.

1.3 Identification of Semiparametric Social Interaction Effect

In this section, I will consider the identification condition for the parameter of interest β_0. Section 1.3.1 restates the assumptions on the data structural of semiparametric social interaction model (1.2.1). To illustrate the main idea of identification in models (1.2.1), I will initially work with a simplified version of model (1.2.1) by excluding the contextual effect (i.e. $W_{g,i}, \overline{W}_{g,-i}$ and Υ_g are not included in $h_0(\cdot)$) in Section 1.3.2. Thus, the semiparametric model of social interactions becomes

$$Y_{g,i} = \beta_0 \overline{Y}_{g,-i} + h_0(X_{g,i}) + U_{g,i}; \quad (g = 1, ..., G; i = 1, ..., n_g). \tag{1.3.1}$$

The identification results can be easily extended to the general model (1.2.1) with contextual effect if certain exclusion restrictions hold. This paper assumes that there exists individual-only characteristics, $X_{g,i}$, that does not generate contextual effect on the outcome $Y_{g,i}$. Then the identification of endogenous social interaction effect can go through using the same strategy as the simplified case. The identification for the general model (1.2.1) with contextual effect will be discussed in Section 1.3.3.

1.3.1 Data Structure

Let Y_g, X_g, W_g, and U_g denote the collection of observations for the corresponding variables in the gth group. For example, the outcome variable for group g is defined by $Y_g = (Y_{g,1}, \cdots Y_{g,n_g})$. The following assumption restates the data requirements discussed in the preceding section.

Assumption 1.3.1: *(i) Econometricians can observe a random sequence of* G *non overlapping groups indexed by* g*,* $(g = 1, \ldots, G)$ *with the size of group* g *equal to* n_g*, the output* Y_g*, and controls* (X_g, W_g, Υ_g) *for group* g*. (ii) The number of groups* $G \to \infty$*, with group size fixed* $(\max_{g=1,\cdots G} n_g < M)$ *but allowed to be different across groups. (iii) Observations* $\{Y_{g,i}, X_{g,i}, W_{g,i}\}$ $(g = 1, \cdots G, i = 1, \cdots n_g)$ *have identical marginal distribution with finite second moments.*

Assumption 1.3.1(i) implies that the observations $\{Y_g, X_g, W_g, \Upsilon_g\}$ $(g = 1, \cdots G)$ are independent across groups while the dependence within each group is unrestricted. For simplicity, the identical marginal distribution for $\{Y_{g,i}, X_{g,i}, W_{g,i}\}$ $(g = 1, \cdots G, i = 1, \cdots n_g)$ is imposed in Assumption 1.3.1(iii) but can be easily relaxed. Assumption 1.3.1(ii) allows for heterogeneity in group sizes but with bounded group size, M, and large numbers of groups, G. Concretely, econometricians are assumed to have available data for repeated data on small groups. The bounded group size can be relaxed by allowing n_g to diverge but at a rate slower than $\sqrt{n}$, which also implies large number of groups with $G \to \infty$. Identification with dataset from a single large network or finite number of networks is complicated and beyond the scope of this paper. All the subsequent discussions on identification of β_0 are based on the data requirements given in Assumption 1.3.1.

1.3.2 Identification without Contextual Effect

To illustrate the identification condition for the endogenous social interaction effect, β_0, first I consider the semiparametric model of social interactions defined in (1.3.1) where no contextual effect is included. I will work with the following exogenous assumption:

Assumption 1.3.2: *The controls* X_g *are assumed to be strictly exogenous and the group size* n_g *is also exogenously formed.*

$$E[U_{g,i}|X_{g,i}, X_{g,-i}, n_g] = 0; \quad (g = 1, ..., G; i = 1, ..., n_g). \tag{1.3.2}$$

The restriction of Assumption 1.3.2 depends on the group selection process. If the groups are randomly assigned, then the Assumption 1.3.2 only needs to hold without controling $X_{g,i}$, i.e. $E[U_{g,i}|X_{g,-i}, n_g] = 0$. The reason is because $X_{g,i}$ and $X_{g,-i}$ are independent under random assignment. Thus, omitting the controls $X_{g,i}$ does not generate the model misspecification problem with random peers.

However, empirical studies seldom have randomly assigned data available which easily make the assumption to fail without adequate controls. For example, a parametric linear model of social interactions can not capture the higher-order or interaction terms of the controls that could affect both the individual and average group outcomes. Thus, applying a more flexible semiparametric model of social interactions is desirable from the perspective of making this strong exogeneity assumption more plausible.

Given Assumptions 1.3.1 and 1.3.2, the identification for the semiparametric social interaction model (1.3.1) can be discussed. The identification strategy adopted is to partial out the nuisance nonparametric function $h_0(\cdot)$ in the first step, and then apply a semiparametric IV approach to identify the endogenous social interaction effect β_0. Based on the strong exogeneity Assumption 1.3.2, any function $t(X_{g,-i}) : R^{d(n_g-1)} \to R^q$ with $E[\|t(x)\|^2] < \infty$ can be used as the instrumental variables for $\overline{Y}_{g,-i}$. To obtain the identification results, full rank condition for $t(X_{g,i})$ is imposed as in the following assumption.

Assumption 1.3.3: *Assume there exists some function $t(X_{g,-i}) : R^{d(n_g-1)} \to R^q$ with $E[\|t(x)\|^2] < \infty$, such that*

$$E\left[t(X_{g,-i})\big(Y_{g,-i} - E[Y_{g,-i}|X_{g,i}]\big)\right] \quad \textit{has full column rank.}$$

The full rank condition in Assumption 1.3.3 requires additional information on $X_{g,-i}$ that can be used as IV for the endogenous variable $\overline{Y}_{g,-i}$. This assumption is commonly imposed in the instrumental variable identification literature. Specifically, if $t(X_{g,-i}) : R^{d(n_g-1)} \to R$ is a one dimensional function, then the Assumption 1.3.3 degenerates to

$$E\left[t(X_{g,-i})\big(Y_{g,-i} - E[Y_{g,-i}|X_{g,i}]\big)\right] \neq 0$$

.

The optimal choice of instrumental variables is beyond the scope of this paper. Instead I provide some idea for the choice of the instrumental variables, $t(X_{g,-i})$, based on the reduced form of the model (1.3.1):

$$\overline{Y}_{g,-i} = \frac{n_g - 1}{(1-\beta_0)(n_g - 1 + \beta_0)}\overline{h}_0(X_{g,-i}) + \frac{\beta_0}{(1-\beta_0)(n_g - 1 + \beta_0)} h_0(X_{g,i}) + V(U_{g,i}, U_{g,-i}), \tag{1.3.3}$$

where $\overline{h}_0(X_{g,-i}) = \frac{1}{n_g-1}\sum_{j\in I_g, j\neq i} h_0(X_j)$ and $V(U_{g,i}, U_{g,-i})$ is the error term which depends on the linear combination of $U_{g,i}$ and $U_{g,-i}$. From the reduced form model (1.3.3), $\overline{h}_0(X_{g,-i})$ can be used as IVs for $\overline{Y}_{g,-i}$ given $h_0(\cdot)$ is known. However, in semiparametric model defined in (1.3.1), $h_0(\cdot)$ is relaxed to be unknown nonparametric function. Thus, $\overline{h}_0(X_{g,-i})$ is infeasible, but the series expansions of $\overline{h}_0(X_{g,-i})$ can be used as IV for $\overline{Y}_{g,-i}$. For simplicity, only the linear approximation $\overline{X}_{g,-i} = \frac{1}{n_g-1}\sum_{j\in I_g, j\neq i} X_j$ is used as IV in the following discussions. Then Assumption 1.3.3(iii) becomes

$$E\left[\overline{X}_{g,-i}\big(Y_{g,-i} - E[Y_{g,-i}|X_{g,i}, n_g]\big)\right] \quad \text{has full column rank.}$$

Based on the above assumptions, the identification result for the semiparametric social interaction model (1.3.1) without contextual effect is established in the following theorem.

Theorem 1.3.1: *Under assumption 1.3.1, 1.3.2, and 1.3.3, the endogenous social interaction effect β_0 and the nonparametric control function h_0 in model (1.3.1) are identified. Specifically,*

(i) β_0 is identified by

$$E\left[t(X_{g,-i})\left(Y_{g,i} - E[Y_{g,i}|X_{g,i}, n_g]\right)\right] = \beta_0 E\left[t(X_{g,-i})\left(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|X_{g,i}, n_g]\right)\right].$$

(ii) h_0 is identified by $h_0(X_{g,i}) = E[Y_{g,i} - \beta_0\overline{Y}_{g,-i}|X_{g,i}]$ given β_0 identified.

Remark: *If $t(\overline{X}_{g,-i}) = \overline{X}_{g,-i}$, i.e., if $\overline{X}_{g,-i}$ is used as an IV for $\overline{Y}_{g,-i}$, then β_0 can be* *identified by*

$$E\left[\overline{X}_{g,-i}\big(Y_{g,i} - E[Y_{g,i}|X_{g,i}, n_g]\big)\right] = \beta_0 E\left[\overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|X_{g,i}, n_g]\big)\right]. \tag{1.3.4}$$

Furthermore, if $\dim(\overline{X}_{g,-i}) = 1$*, then* β_0 *can be simply calculated by*

$$\beta_0 = \frac{E\left[\overline{X}_{g,-i}\big(Y_{g,i} - E[Y_{g,i}|X_{g,i}, n_g]\big)\right]}{E\left[\overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|X_{g,i}, n_g]\big)\right]}.$$

The proof of Theorem 1.3.1 is presented in Appendix 1.A. The basic idea is to partial out the nonparametric nuisance function $h_0(\cdot)$ by subtracting the conditional expectations on both side of model (1.3.1), $Y_{g,i} - E[Y_{g,i}|X_{g,i}, n_g] = \beta_0\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|X_{g,i}, n_g]\big) + U_{g,i}$, and then apply the IV approach to address the endogeneity issue caused by simultaneous equations on the residulized model to identify β_0.

The identification results can be extended to related social interaction models, which are discussed as follows.

Discussion 1.3.1: *Identification with large numbers of same size groups:*
The identification results in Theorem 1.3.1 covers the same group size model as a special case. It is well known that Manski's reflection problem shows up for the same group size social interaction model since there is no extra between-group variation to disentangle the endogenous social interaction effect and contextual effect. Here we assume that there are individual-only characteristics, $X_{g,i}$*, that affect* $Y_{g,i}$ *without generating contextual effects. Thus, social interaction model with same group size can still be identified using the same strategy as discussed above.*

Discussion 1.3.2: *Identification with large numbers of networks:*
The identification results in Theorem 1.3.1 can also be applied to a semiparametric model of social interactions with large numbers of networks. Network structures have exclusion restrictions due to the incomplete overlap among individuals and their peers. Under this case, the average characteristics of friends of friends can drive variation in endogenous effects without having a direct effect on outcomes, and can be used as IV for average group outcomes.

Thus, excluded factors of individual-only characteristics $X_{g,i}$ is not required to identify β_0. This strategy is applied in Bramoullé, Djebbari, and Fortin (2009) for the parametric social interaction models. Based on the results in Theorem 1.3.1, the semiparametric model of network social interactions can also be identified using the same strategy.

1.3.3 Identification with Contextual Effect

Here we discuss the identification condition for a more general semiparametric model (1.2.1) which also controls the contextual effects $\overline{W}_{g,-i}$ and Υ_g in the nonparametric function $h_0(\cdot)$.

Let $\mathcal{X}_{g,i} = (X_{g,i}, W_{g,i}, \overline{W}_{g,-i}, \Upsilon_g, n_g)$ denotes the collection of control variables at both the individual and group levels. Including the contextual effect in model (1.2.1) brings challenge for identifying the endogenous social interaction effect, which is the well-known reflection problem (Manski, 1993). To resolve the reflection problem, this paper assumes that not all exogenous variables appear as contextual variables, i.e. $d = \dim(X_{g,i}) > 0$ and the leaving-i-out group average $\overline{X}_{g,-i}$ does not affect individual's outcome directly, which can provide an exclusion restriction. Similar as Assumption 1.3.2, for model (1.2.1), the following exogenous assumption is imposed:

Assumption 1.3.4: *The controls $\mathcal{X}_g = (X_g, W_g, \Upsilon_g, n_g)$ are assumed to be strictly exogenous:*

$$E[U_{gi} \mid \mathcal{X}_g \ ] = E[U_{gi}|X_g, W_g, \Upsilon_g, n_g] = 0 \tag{1.3.5}$$

Assumption 1.3.4 relaxed the strong exogeneity in Assumption 1.3.2 by including additional individual and contextual effects $W_{g,i}, \overline{W}_{g,-i}, \Upsilon_g$ in the nonparametric function. After conditioning on a set of additional group-level covariates, Assumption 1.3.4 becomes more plausible and credible.

Given the exclusion restriction and exogenous Assumption 1.3.4, the excluded contextual factors $\overline{X}_{g,-i}$ can be used as IV for $\overline{Y}_{g,-i}$ to identify β_0 if the standard full rank condition holds:

Assumption 1.3.5: *Assume $E\left[\overline{X}_{g,-i}\big(Y_{g,-i} - E[Y_{g,-i}|\mathcal{X}_{g,i}]\big)\right]$ has full column rank.*

Based on the above assumptions, the identification result for the semiparametric social interaction model (1.2.1) with contextual effect is established in the following theorem:

Theorem 1.3.2: *Under assumption 1.3.1, 1.3.4, and 1.3.5, the endogenous social interaction effect β_0 and the nonparametric control function h_0 in model (1.2.1) are identified. Specifically,*

(i) β_0 is then identified by

$$E\left[\overline{X}_{g,-i}\big(Y_{g,i} - E[Y_{g,i}|\mathcal{X}_{g,i}]\big)\right] = \beta_0 E\left[\overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|\mathcal{X}_{g,i}]\big)\right] \tag{1.3.6}$$

(ii) h_0 is identified by $h_0(X_{g,i}) = E[Y_{g,i} - \beta_0\overline{Y}_{g,-i}|\mathcal{X}_{g,i}]$ given β_0 identified.

The proof is similar as Theorem 1.3.1 and is presented in Appendix 1.A.

Comparing the identification condition for social interaction model with contextual effect in (1.3.6) and the condition (1.3.4) for model without contextual effect, I find that the only difference comes from the partialled out conditional expectations, which will not affect the discussion for the semiparametric estimation for β_0. Thus, in the subsequent sections, I just use the notation in the social interaction model (1.3.1) with no contextual effect to illustrate the semiparametric estimation procedure and asymptotic properties of the estimators.

This paper will work under the strong instrumental variable case to derive theoretical results in the subsequent sections. However, by partialling out additional control variable, $\overline{X}_{g,-i}$ will be more likely to suffer from weak instrumental variable problem in the finite samples. The weak identification issue will be addressed in the Monte Carlo simulations.

1.4 Moment and Orthogonal Moment Conditions

In this section, the moment condition and orthogonal moment condition are constructed based on the identification results in the preceding section. To simplify the notation, I consider the model (1.3.1) without contextual effect in the subsequent sections. As discussed above, the results can be easily extended to the general model (1.2.1) with the contextual effect.

First, this section restates the identification conditions for β_0 in model (1.3.1) using the moment conditions, and then discusses the construction of Neyman orthogonal moment conditions which are robust to the first step nonparametric nuisance parameters.

Let $\mu_0(X_{g,i}) \triangleq E[Y_{g,i}|X_{g,i}]$; $\nu_0(X_{g,i}) \triangleq E[\overline{Y}_{g,-i}|X_{g,i}]$; $\phi_0(X_{g,i}) \triangleq E[\overline{X}_{g,-i}|X_{g,i}]$ denote the conditional expectations of $Y_{g,i}, \overline{Y}_{g,-i}, \overline{X}_{g,-i}$ on $X_{g,i}$, respectively. The corresponding residualized variables after subtracting the conditional expectations are denoted by $\eta_{g,i} \equiv Y_{g,i} - \mu(X_{g,i})$; $\zeta_{g,i} \equiv \overline{Y}_{g,-i} - \nu(X_{g,i})$; $\varepsilon_{g,i} \equiv \overline{X}_{g,-i} - \phi(X_{g,i})$. Then $\mu_0(\cdot)$, $\nu_0(\cdot)$ and $\phi_0(\cdot)$ can be estimated by the following conditional moments:

$$E[Y_{g,i} - \mu_0(X_{g,i})|X_{g,i}] = 0, \tag{1.4.1}$$

$$E[\overline{Y}_{g,-i} - \nu_0(X_{g,i})|X_{g,i}] = 0, \tag{1.4.2}$$

$$E[\overline{X}_{g,-i} - \phi_0(X_{g,i})|X_{g,i}] = 0. \tag{1.4.3}$$

The identification condition for β_0 in Theorem 1.3.1 can be restated using the following moment condition:

$$E\Big[\overline{X}_{g,-i}\Big(\big(Y_{g,i} - \mu_0(X_{g,i})\big) - \beta_0\big(\overline{Y}_{g,-i} - \nu_0(X_{g,i})\big)\Big)\Big] = 0, \tag{1.4.4}$$

where nonparametric nuisance parameters $\mu_0(\cdot)$ and $\nu_0(\cdot)$ can be estimated using conditional moments (1.4.1) and (1.4.2) in the first step. The parameter of interest, β_0, then can be estimated using (1.4.4) by plugging in the first step nonparametric estimators.

However, the estimator of β_0 based on the moment condition (1.4.4) is sensitive to the estimation bias of nonparametric parameters $\mu_0(\cdot)$ and $\nu_0(\cdot)$ (Chernozhukov et al., 2018a). To be specific, let $m(\cdot)$ denote the moment function in condition (1.4.4),

$$m(Z_{g,i}; \beta, \mu, \nu) = \overline{X}_{g,-i}(Y_{g,i} - \mu(X_{g,i})) - \beta \overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - \nu(X_{g,i})\big), \tag{1.4.5}$$

where $Z_{g,i} = (X_{g,i}, Y_{g,i}, \overline{X}_{g,-i}, \overline{Y}_{g,-i})$. Let $\frac{\partial E[m(Z_{g,i};\beta_0,\mu_0,\nu_0)]}{\partial \mu}[v_\mu]$ denote the $d_m \times 1$ vector of pathwise derivative of $E\left[m\left(Z_{g,i};\beta_0,\mu_0,\nu_0\right)\right]$ with respect to μ in the direction v_μ evaluated at true parameters (β_0, μ_0, ν_0). Similarly, let $\frac{\partial E[m(Z_{g,i};\beta_0,\mu_0,\nu_0)]}{\partial \nu}[v_\nu]$ denote the path-wise derivative over ν in the direction v_ν. It is showed in Appendix 1.A that these two pairwise derivatives are nonzero. Thus, the moment condition defined in (1.4.4) is not orthogonal to the first step nonparametric parameters $\mu(\cdot)$ and $\nu(\cdot)$. This might lead to a bias in the second step estimator for β_0 based on the moment condition (1.4.4).

The robust strategy for estimating β_0 is to use orthogonal moment condition instead. Following the strategy in Chernozhukov et al. (2018a), the orthogonal moment condition can be constructed by adding an adjustment term,

$$E\Big[\Big(\overline{X}_{g,-i} - \phi_0(X_{g,i})\Big)\Big((Y_{g,i} - \mu_0(X_{g,i})) - \beta_0\big(\overline{Y}_{g,-i} - \nu_0(X_{g,i})\big)\Big)\Big] = 0. \tag{1.4.6}$$

An additional nuisance function $\phi_0(X_{g,i}) = E[\overline{X}_{g,-i}|X_{g,i}]$ is introduced which can be estimated by conditional moment (1.4.3). Let $\psi(\cdot)$ denote the orthogonal moment function in (1.4.6),

$$\psi(Z_{g,i}, \beta, \mu, \nu, \phi) = \Big(\overline{X}_{g,-i} - \phi(X_{g,i})\Big)\Big(\big(Y_{g,i} - \mu(X_{g,i})\big) - \beta\big(\overline{Y}_{g,-i} - \nu(X_{g,i})\big)\Big). \tag{1.4.7}$$

It is verified in in Appendix 1.A that the pairwise derivatives of $E\left[\psi\left(Z_{g,i};\beta_0,\mu_0,\nu_0,\phi_0\right)\right]$ with respect to $\mu(\cdot), \nu(\cdot)$ and $\phi(\cdot)$ vanish.

$$\begin{aligned} &\frac{\partial E\left[\psi\left(Z;\beta_0,\mu_0,\nu_0,\phi_0\right)\right]}{\partial \mu}[v_\mu] = 0; \quad \frac{\partial E\left[\psi\left(Z;\beta_0,\mu_0,\nu_0,\phi_0\right)\right]}{\partial \nu}[v_\nu] = 0 \\ &\frac{\partial E\left[\psi\left(Z;\beta_0,\mu_0,\nu_0,\phi_0\right)\right]}{\partial \phi}[v_\phi] = 0 \end{aligned} \tag{1.4.8}$$

That is, the moment is close to zero as the nuisance functions μ, ν, ϕ deviate from their true value. Thus, the estimator for β_0 based on the orthogonal moment condition (2.6) is locally robust to the first step nonparametric estimators for $\mu_0(\cdot), \nu_0(\cdot)$ and $\phi_0(\cdot)$

Also, the orthogonal moment function $\psi\left(Z_{g,i}, \beta, \mu, \nu, \phi\right)$ is affine in each of μ, ν and

ϕ holding others fixed.[2] By Theorem 5 in Chernozhukov et al. (2018a), the orthogonal moment (1.4.6) is also a doubly robust (DB) moment condition. It can be verified that $\psi(Z_{g,i}, \beta, \mu, \nu, \phi)$ satisfies the definition of DB moment condition:

$$E\left[\psi\left(Z_{g,i}, \beta_0, \mu, \nu, \phi_0\right)\right] = 0; \quad E\left[\psi\left(Z_{g,i}, \beta_0, \mu_0, \nu_0, \phi\right)\right] = 0 \tag{1.4.9}$$

Thus, the orthogonal moment condition still holds if part of the first stage nuisance functions (μ, ν) or ϕ to be incorrect. This will lead to a simpler conditions for the asymptotic normality in general (Chernozhukov et al., 2018a). The asymptotic property for the estimator of semiparametric social interaction model will be discussion in Section 1.5.

1.5 Semiparametric Estimation of Endogenous Social Interactions

In this section, a semiparametric two-step GMM estimator is proposed to estimate the endogenous social interaction effect, β_0, in model (1.3.1). This section then shows the asymptotic normality of the semiparametric GMM estimator with first step series estimators under primitive regularity conditions.

1.5.1 Construction of Semiparametric Estimators

Let $\vartheta_0(x) = (\mu_0(x), \nu_0(x), \phi_0(x))$ denote the first step conditional expectations. Then, $\vartheta_0(x)$ can be estimated by any nonparametric methods. In the second step, the endogenous social interaction effect β_0 is then estimated using GMM method with sample moment functions of (1.4.5) and (1.4.7) after plugging the first step nonparametric estimators $\widehat{\vartheta}(x)$.

In the following discussion, I will call the estimator for β_0 based on moment function (1.4.5) as the plug-in (PI) estimator, and the estimator based on the orthogonal moment function (1.4.7) as the debiasing (DB) estimator.

[2]An affine function is a linear function followed by a translation. It is verified in Appendix 1.A that $\psi(Z_{g,i}, \beta, \mu, \nu, \phi)$ defined in is affine in each of μ, ν and ϕ holding others fixed.

1.5.1.1 First-step Nonparametric Estimation

In the first step, the conditional expectations

$$\mu_0\left(X_{g,i}\right)=E\left[Y_{g,i}|X_{g,i}\right],\quad \nu_0\left(X_{g,i}\right)=E\left[\overline{Y}_{g,-i}|X_{g,i}\right],\quad \phi_0\left(X_{g,i}\right)=E\left[\overline{X}_{g,-i}|X_{g,i}\right]$$

are nonparametrically estimated. There is a huge body of literature that addresses the estimation of conditional expectations. Parametric method usually uses a linear function to approximate the conditional expectation, and then estimates it using least squares. Nonparametric methods including kernel or series estimators are more flexible in the choice of function forms.

In this paper, I will study a semiparametric GMM estimator of β_0, which applies nonparametric power series in the first step to estimate the conditional expectations $\vartheta_0(x)=(\mu_0(x),\nu_0(x),\phi_0(x))$. Assume we use the power series expansion with basis function $P^K(x)=(P_1(x),\ldots,P_K(x))$ to approximate the first step conditional expectations. For $K=K(n)\rightarrow\infty$, the first step series nonparametric estimators for conditional expectations are:

$$\begin{aligned}\widehat{\mu}\left(X_{g,i}\right) &= P^K\left(X_{g,i}\right)'\left(P'P\right)^{-1}P'Y\\ \widehat{\nu}\left(X_{g,i}\right) &= P^K\left(X_{g,i}\right)'\left(P'P\right)^{-1}P'\overline{Y}_-\\ \widehat{\phi}\left(X_{g,i}\right) &= P^K\left(X_{g,i}\right)'\left(P'P\right)^{-1}P'\overline{X}_-\end{aligned}\tag{1.5.1}$$

where $P=\left(P\left(X_{g1}\right),\cdot,P\left(X_{G,n_G}\right)\right)',Y=\left(Y_{g1},\cdots,Y_{G,n_G}\right),\overline{Y}_-=\left(\overline{Y}_{g,-1},\cdots,\overline{Y}_{G,-n_G}\right)$, and $\overline{X}_-=\left(\overline{X}_{g,-1},\cdots,\overline{X}_{G,-n_G}\right)$.

The motivation for considering this particular nonparametric method (power series) is to provide the primitive regularity conditions for showing the consistency and asymptotic normality of the semiparametric GMM estimator for β_0.

1.5.1.2 Second-step Estimation for β_0

In the second step, the parameter of interest, β_0, can be estimated by plugging in the first step nonparametric estimators $\widehat{\vartheta}(x) = (\widehat{\mu}(x), \widehat{\nu}(x), \widehat{\phi}(x))$ into the moment function (1.4.5) or orthogonal moment function (1.4.7). This paper proposes a semiparametric two-step GMM method (Chen, 2007; Ackerberg, Chen, and Hahn, 2012) to estimate the parameter of interest, β_0. The idea is to make the sample analog of moment condition to be as close to zero as possible.

To address the clustered structural data of the grouped social interaction model, it will be convenient to define the clustered sum of moment functions:

$$m_g\left(Z_g, \beta, \vartheta\right) = \sum_{i=1}^{n_g} m\left(Z_{g,i}; \beta, \vartheta\right); \quad \psi_g\left(Z_g, \beta, \vartheta\right) = \sum_{i=1}^{n_g} \psi\left(Z_{g,i}; \beta, \vartheta\right) \tag{1.5.2}$$

where $Z_g = (Z_{g,1}, \cdots, Z_{g,n_g})$. Z_g includes the group-level variables $(X_g, Y_g, \overline{X}_{g,-}, \overline{Y}_{g,-})$, where $\overline{X}_{g,-} = (\overline{X}_{g,-1}, \cdots, \overline{X}_{g,-n_g})$ and $\overline{Y}_{g,-} = (\overline{Y}_{g,-1}, \cdots, \overline{Y}_{g,-n_g})$ are the collections of leave-i-out group average variables for group g.

The semiparametric plug-in GMM estimator, $\widehat{\beta}^{gmm}$, is constructed by making the sample analogy of moment function, $\frac{1}{n}\sum_{g=1}^{G} m_g\left(Z_g, \beta, \widehat{\vartheta}\right)$, as close to zero as possible. It solves the following minimization problems:

$$\widehat{\beta}^{gmm} = \arg\min_{\beta} \left(\frac{1}{n}\sum_{g=1}^{G} m_g\left(Z_g, \beta, \widehat{\vartheta}\right)\right)' \widehat{\Omega}^{-1} \left(\frac{1}{n}\sum_{g=1}^{G} m\left(Z_g, \beta, \widehat{\vartheta}\right)\right), \tag{1.5.3}$$

where $\widehat{\Omega}^{-1}$ denotes an $d \times d$ positive definite weight matrix. Because of the linearity of moment function, $m_g(\cdot)$, with respect to β, the semiparametric GMM estimator $\widehat{\beta}^{gmm}$ defined in (1.5.3) has a closed form solution:

$$\widehat{\beta}^{gmm} = \Big(\sum_{g=1}^{G} \widehat{\zeta}_g' \overline{X}_{g,-} \widehat{\Omega}^{-1} \sum_{g=1}^{G} \overline{X}_{g,-}' \widehat{\zeta}_g\Big)^{-1} \Big(\sum_{g=1}^{G} \widehat{\zeta}_g' \overline{X}_{g,-} \widehat{\Omega}^{-1} \sum_{g=1}^{G} \overline{X}_{g,-}' \widehat{\eta}_g\Big). \tag{1.5.4}$$

where $\widehat{\eta}_g$ and $\widehat{\zeta}_g$ are group level residuals for Y_g and $\overline{Y}_{g,-}$, respectively. As usual, $\widehat{\Omega}$ is chosen to be the consistent estimator of the variance $\Omega = \text{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} m_g\left(Z_g, \beta_0, \widehat{\vartheta}\right)\right)$.

The semiparametric debiased GMM estimator, $\widehat{\beta}_{db}^{gmm}$, is constructed by making the sample analogy of orthogonal moment function $\psi_g\left(Z_g, \beta, \vartheta\right)$, as close to zero as possible.

$$\widehat{\beta}_{db}^{gmm} = \arg\min_{\beta}\left(\frac{1}{n}\sum_{g=1}^{G}\psi_g\left(Z_g, \beta, \widehat{\vartheta}\right)\right)'\widehat{\Omega}_{db}^{-1}\left(\frac{1}{n}\sum_{g=1}^{G}\psi_g\left(Z_g, \beta, \widehat{\vartheta}\right)\right), \tag{1.5.5}$$

where $\widehat{\Omega}_{db}^{-1}$ denotes an $d \times d$ positive definite weight matrix. Similarly, the closed form solution for the debiased GMM estimator $\widehat{\beta}_{db}^{gmm}$ is:

$$\widehat{\beta}_{db}^{gmm} = \left(\sum_{g=1}^{G}\widehat{\zeta}_g'\widehat{\varepsilon}_g\widehat{\Omega}_{db}^{-1}\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\zeta}_g\right)^{-1}\left(\sum_{g=1}^{G}\widehat{\zeta}_g'\widehat{\varepsilon}_g\widehat{\Omega}_{db}^{-1}\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\eta}_g\right) \tag{1.5.6}$$

where $\widehat{\eta}_g$, $\widehat{\zeta}_g$, and $\widehat{\phi}$ are group level residuals for Y_g, $\overline{Y}_{g,-}$, and $\overline{X}_{g,-}$, respectively. $\widehat{\Omega}_{db}$ is chosen to be the consistent estimator of the variance $\Omega_{db} = \text{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\psi_g\left(Z_g, \beta_0, \widehat{\vartheta}\right)\right)$. Since $\psi_g\left(Z_g, \beta, \vartheta\right)$ is doubly robust to the nonparametric estimation of ϑ_0, it can be shown that $\Omega_{db} = \Omega_0 + o_p(1)$, where $\Omega_0 = Var\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\psi_g\left(Z_g, \beta_0, \vartheta_0\right)\right)$.

Remark: *Just identified second step:*

If $\dim(m) = \dim(\psi) = \dim(X_{g,i}) = 1$*, then the second step estimation for* β_0 *is exactly identified by moment condition (1.4.4) or (1.4.6). The semiparametric two-step Plug-in and debiased estimators for* β_0*,* $\widehat{\beta}_{db}$*, can be simplified by directly solving the sample analogy of moment and orthogonal moment conditions. The closed form solution for* $\widehat{\beta}$ *and* $\widehat{\beta}_{db}$ *are defined as follows:*

$$\widehat{\beta} = \left(\sum_{g=1}^{G}\overline{X}_{g,-}'\widehat{\zeta}_g\right)^{-1}\left(\sum_{g=1}^{G}\overline{X}_{g,-}'\widehat{\eta}_g\right) \tag{1.5.7}$$

$$\widehat{\beta}_{db} = \left(\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\zeta}_g\right)^{-1}\left(\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\eta}_g\right) \tag{1.5.8}$$

where $\widehat{\eta}_g$, $\widehat{\zeta}_g$, and $\widehat{\phi}$ are group level residuals for Y_g, $\overline{Y}_{g,-}$, and $\overline{X}_{g,-}$, respectively.

1.5.1.3 Relation of Plug-in and Debiased Estimators for β_0

This subsection will discuss the relationship between the plug-in estimator $\widehat{\beta}^{gmm}$ and the debiased estimator $\widehat{\beta}^{gmm}_{db}$ if the nonparametric parameters ϑ_0 are estimated using series method in the first step.

First, consider weight matrices for two semiparametric GMM estimators. The following proposition shows the relation of the inverse weight matrix for the plug-in estimator $\Omega = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} m_g\left(Z_g, \beta_0, \widehat{\vartheta}\right)\right)$, the debiased estimator $\Omega_{db} = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} \psi_g\left(Z_g, \beta_0, \widehat{\vartheta}\right)\right)$, and also $\Omega_0 = Var\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} \psi_g\left(Z_g, \beta_0, \vartheta_0\right)\right)$.

Proposition 1.5.1: *Suppose*

$$\frac{1}{n}\left\|\sum_{g=1}^{G} \psi_g\left(Z_g, \beta_0, \widehat{\vartheta}\right) - \sum_{g=1}^{G} \psi_g\left(Z_g, \beta_0, \vartheta_0\right)\right\|^2 \xrightarrow{p} 0, \tag{1.5.9}$$

then

$$\Omega = \Omega_{db} = \Omega_0 + o_p(1)$$

That is

$$Var\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} m_g\left(\beta_0, \widehat{\vartheta}\right)\right) = Var\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{n} \psi_g\left(\beta_0, \widehat{\vartheta}\right)\right) = Var\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} \psi_g\left(\beta_0, \vartheta_0\right)\right) + o_p(1) \tag{1.5.10}$$

where $\widehat{\vartheta}$ is any consistent nonparametric estimator of $\vartheta_0 = (\mu_0, \nu_0, \phi_0)$.

The influence function of $\frac{1}{\sqrt{n}}\sum_{g=1}^{G} m_g\left(\beta_0, \widehat{\vartheta}\right)$ is $\frac{1}{\sqrt{n}}\sum_{g=1}^{G} \psi_g\left(\beta_0, \vartheta_0\right)$. Following Newey (1994), it is sufficient show $\Omega = \Omega_{db} = \Omega_0 + o_p(1)$ if (1.5.9) holds and $\widehat{\vartheta}$ can be any consistent first step nonparametric estimators. I will specify the conditions for (1.5.9) holding if the first step nonparametric parameters are estimated using series methods in the following section. The proof of Proposition 1.5.1 is shown in Appendix 1.A.

Based on Proposition 1.5.1, the estimators for the optimal weight matrices of plug-in and debiased estimators can be constructed by

$$\begin{aligned}
\widehat{\Omega} &= \frac{1}{n}\sum_{g=1}^{G}\psi_g\left(\widetilde{\beta},\widehat{\vartheta}\right)\psi_g\left(\widetilde{\beta},\widehat{\vartheta}\right)'; \\
\widehat{\Omega}_{db} &= \frac{1}{n}\sum_{g=1}^{G}\psi_g\left(\widetilde{\beta}_{db},\widehat{\vartheta}\right)\psi_g\left(\widetilde{\beta}_{db},\widehat{\vartheta}\right)', \qquad (1.5.11)
\end{aligned}$$

where $\widetilde{\beta}$ and $\widetilde{\beta}_{db}$ are initial estimators for β_0 using plug-in or debiased methods with identity weight matrix, respectively. It can be shown that $\widetilde{\beta} = \widetilde{\beta}_{db}$ if the first nonparametric parameters are estimated using series. Then, it follows that $\widehat{\Omega} = \widehat{\Omega}_{db}$.

Finally, the following proposition states the relationships of plug-in estimator, $\widehat{\beta}^{gmm}$, and debiased estimator $\widehat{\beta}^{gmm}_{db}$ if the series estimator is applied in the first step.

Proposition 1.5.2: *Assume (i) the first step nonparametric parameters are estimated using series as in (1.5.1); (ii) the weight matrices for plug-in and debiased estimators are defined in 1.5.11; where (iii) the initial estimators $\widetilde{\beta}$ and $\widetilde{\beta}_{db}$ are estimated with identity weight matrices for both the plug-in and debiased estimators. Then the semiparametric plug-in and debiased estimators are exactly the same:*

$$\widehat{\beta}^{gmm} = \widehat{\beta}^{gmm}_{db}$$

The proof of Proposition 1.5.2 are shown in Appendix 1.A. Our result verifies Newey, Hsieh, and Robins (1998)'s statement that series estimators of conditional expectations belong to the idempotent estimator class which has this smoothing correction built-in.

In the following discussion of the semiparametric series estimators for β_0, I use the formula of debiased estimators, $\widehat{\beta}^{gmm}_{db}(\widehat{\beta}_{db})$, to facilitate the proofs.

1.5.2 Asymptotic Normality of Semiparametric Series Estimators

This section discusses the consistency and asymptotic normality of the semiparametric two-step estimator for β_0 by plugging the series estimators in the first step. Theoretical results for semiparametric estimators with the first step series have been studied by Newey (1994, 1997) among others. This section applies results from Newey (1997) to derive regularity conditions of asymptotic normality for semiparametric estimators, $\widehat{\beta}_{db}^{gmm}$, using first step series method.

The following regularity conditions are imposed to control the first step series approximation error and estimation error of the conditional expectations $\vartheta_0 = (\mu_0, \nu_0, \phi_0)$.

Assumption 1.5.1:

(i) Assume $\frac{1}{n}\sum_{g=1}^{G} E(\eta_g'\eta_g|X)$, $\frac{1}{n}\sum_{g=1}^{G} E(\zeta_g'\zeta_g|X)$ *and* $\frac{1}{n}\sum_{g=1}^{G} E(\epsilon_g'\epsilon_g|X)$ *are all bounded.*

(ii) The smallest eigenvalue $\lambda_{min}\left(E[P^K(x)P^K(x)']\right)$ *is bounded away from zero uniformly in* K*.*

(iii) Assume the approximation error to the nonparametric functions μ_0, ν_0, ϕ_0 *satisfies the uniform convergence rate as follows:*

$$|\mu_0 - P^{K'}\pi_\mu| = O(K^{-\alpha_\mu}),\ |\nu_0 - P^{K'}\pi_\nu| = O(K^{-\alpha_\nu}),\ |\phi_0 - P^{K'}\pi_\phi| = O(K^{-\alpha_\phi})$$

Assumption 1.5.1 (i)-(iii) are all standard conditions for the series estimators of the nonparametric conditional expectations. The bounded conditional variance in Assumption 1.5.1(i) helps bound the variance of series estimators for the conditional expectations. Assumption 1.5.1(ii) impose restrictions on the approximation basis functions to have nonsingular second moment. Assumption 1.5.1 (iii) helps to control the approximation bias of series expansions. From Newey (1997), for power series, Assumption 1.5.1(iii) is satisfied with $\alpha_\mu = s_\mu/d,\ \alpha_\nu = s_\nu/d,\ \alpha_\phi = s_\phi/d$ where $\alpha_\mu, \alpha_\nu, \alpha_\phi$ is the order of continuous derivative of $\mu, \nu,\ \phi$, respectively and $d = \dim(X)$.

To illustrate the Assumption 1.5.1 (iii) for the semiparametric social interaction model, consider the special model which has same group size $n_g = 2$. The reduced form of the model is as follows,

$$\begin{aligned} Y_{g,i} &= \frac{1}{1-\beta_0^2}(h_0(X_{g,i})+\beta_0 h_0(\overline{X}_{g,-i})+V_i \\ \overline{Y}_{g,-i} &= \frac{1}{1-\beta_0^2}(h_0(\overline{X}_{g,-i})+\beta_0 h_0(X_{g,i}))+V_{-i}, \end{aligned}$$

Then the reduced form model implies that:

$$\begin{aligned} \mu(X_{g,i}) &= E[Y_{g,i}|X_{g,i}] = \frac{1}{1-\beta_0^2}\left(h_0(X_{g,i})+\beta_0 E[h_0(\overline{X}_{g,-i})|X_{g,i}]\right) \\ \nu(X_{g,i}) &= E[\overline{Y}_{g,-i}|X_{g,i}] = \frac{1}{1-\beta_0^2}\left(\beta_0 h_0(X_{g,i})+E[h_0(\overline{X}_{g,-i})|X_{g,i}]\right) \end{aligned}$$

Thus $\mu(X_{g,i})$ and $\nu(X_{g,i})$ have the same continuous differentiable order which depends on the smoothness of $h_0(X_{g,i})$ and $E[h_0(\overline{X}_{g,-i})|X_{g,i}]$.

To provide the asymptotic distribution of $\widehat{\beta}_{db}^{gmm}$, define

$$\begin{aligned} &M_n = \frac{1}{n}\sum_{g=1}^{G} E\left[\varepsilon_g'\zeta_g\right], \\ &\Omega_n = \frac{1}{n}\sum_{g=1}^{G} E\left[\varepsilon_g' U_g U_g'\varepsilon_g\right], \\ &V_n = M_n^{-1}\Omega_n M_n^{-1}. \end{aligned}$$

The next theorem show the result on the asymptotic normality of semiparametric GMM estimator $\widehat{\beta}_{db}^{gmm}$ when the first step conditional expectations $\vartheta_0(x) = (\mu_0(x), \nu_0(x), \phi_0(x))$ are estimated by series estimators as defined in (1.5.1).

Theorem 1.5.1: *If Assumptions 1.3.1, 1.3.2, 1.3.3, and 1.5.1 are satisfied and assume the following condition holds*

$$\sqrt{n}K^{-\left(\alpha_\phi+\alpha_\mu\right)}+\sqrt{n}K^{-\left(\alpha_\phi+\alpha_\nu\right)}+K^{-\alpha_\phi}+K^{-\alpha_\mu}+K^{-\alpha_\nu}+K/\sqrt{n}\to 0 \qquad (1.5.12)$$

then it follows that

$$\sqrt{n}V_n^{-1/2}(\widehat{\beta}_{db}^{gmm} - \beta_0) \xrightarrow{d} N(0,1). \tag{1.5.13}$$

Condition (1.5.12) controls the approximation and estimation errors induced by the first step series estimators, $\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$. If $K/\sqrt{n} \to 0$ is satisfied, then the remainder rate is dominated by order $\sqrt{n}K^{-\left(\alpha_\phi + \max\{\alpha_\mu, \alpha_\nu\}\right)}$. Furthermore, there exists $K = K_n$ satisfying $K/\sqrt{n} \to 0$, such that condition (1.5.12) holds if and only if $\alpha_\phi + \max\{\alpha_\mu, \alpha_\nu\} > 1$.

The general requirement for asymptotic normality of semiparametric estimator need to control each of $\sqrt{n}K^{-\alpha_\phi}$, $\sqrt{n}K^{-\alpha_\mu}$, and $\sqrt{n}K^{-\alpha_\nu}$ converges to zero (Newey, 1994). Here the condition (1.5.12) only requires the product of them goes to zero, which simplifies the requirement for the asymptotic normality of the semiparametric estimator. This is because the estimator $\widehat{\beta}_{db}^{gmm}$ is constructed based on the doubly robust moment condition (1.4.6), which still holds if one of the first step nonparametric estimators is not correct. The asymptotic normality of doubly robust estimators have simpler conditions in general which are discussed in Chernozhukov et al. (2018b) and Newey and Robins (2018).

Next, consider the construction of consistent estimator for the variance of $\widehat{\beta}_{db}^{gmm}$. As usual the estimator for the variance of $\widehat{\beta}_{db}$ can be constructed by

$$\widehat{V}_n = \widehat{M}_n^{-1}\widehat{\Omega}_n\widehat{M}_n^{-1} \tag{1.5.14}$$

with $\widehat{M}_n = \frac{1}{n}\sum_{g=1}^{G}\left[\widehat{\varepsilon}_g'\widehat{\zeta}_g\right]$ and $\widehat{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G}\left[\widehat{\varepsilon}_g'\widehat{U}_g\widehat{U}_g'\widehat{\varepsilon}_g\right]$, where $\widehat{\eta}_g$, $\widehat{\zeta}_g$, and $\widehat{\varepsilon}_g$ denote the first step residualized terms for $Y_g, \overline{Y}_{g,-}, \overline{X}_{g,-}$respectively, and $\widehat{U}_g = \widehat{\eta}_g - \widehat{\beta}_{db}^{gmm}\widehat{\zeta}_g$.

The variance estimator, $\widehat{V}_n$ is constructed "as if" the nonparametric parameters μ_0, ν_0, and ϕ_0 were known. This works because the moment condition (1.4.6) is locally/doubly robust to the first step estimators (Ackerberg, Chen, and Hahn, 2012).

The following result gives results for the consistency of the variance estimator, $\widehat{V}_n$.

Theorem 1.5.2: *If the assumptions in Theorem 1.5.1 hold, then*

$$\left\|\widehat{M}_n - M_n\right\| \xrightarrow{p} 0, \left\|\widehat{\Omega}_n - \Omega_n\right\| \xrightarrow{p} 0, \text{ and } \left\|\widehat{V}_n - V_n\right\| \xrightarrow{p} 0$$

The proofs for Theorem 1.5.1 and Theorem 1.5.2 are in Appendix 1.A.

1.6 Semiparametric Cross-fitting Estimation of Endogenous Social Interactions

The plug-in and debiased semiparametric estimators for β_0 considered in Section 1.5 have the own-observation bias problem. With first step power series of K terms, the bias has an order of $K/\sqrt{n}$. Chernozhukov et al. (2018a) has considered to apply the single cross-fitting method to eliminate the own observation bias. The single cross-fitting method uses the same subsamples to estimate the first step nonparametric functions. Thus, it will still suffer from the nonlinearity bias. Newey and Robins (2018) proposes a double cross-fitting procedure which uses different subsamples in the first step nonparametric estimation to further remove the nonlinearity bias.

In this paper, I consider to apply the cross-fitting methods for the semiparametric social interaction model to further eliminate the bias from first step nonparametric estimation.

1.6.1 Construction of Cross-fitting estimators

I will focus on the cross-fitting semiparametric estimator which is based on the doubly robust moment function (1.4.7). The main idea of cross-fitting method is to estimate the nonparametric parameters (μ_0, ν_0, ϕ_0) and the second-step parameter β_0 with different subsets in order to remove the own observation bias.

For the social interaction models, researchers can obtain a clustered sample $\{Z_g\}_{g=1}^{G}$. This paper assumes the observations $\{Z_g\}$ $(g = 1, \cdots G)$ are independent across groups but the dependence within each group is unrestricted. Thus, the subsamples should be generated at the group level instead of the individual level. Assume the group indexes $\{1, ...G\}$ are randomly partitioned into L distinct subsets I_ℓ ($\ell = 1, ...L$).

For notation simplicity, this section considers the just identified second step with $\dim(m) = \dim(\psi) = \dim(X_{g,i}) = 1$.

A single cross-fitting (CF) Estimator is defined as

$$\begin{aligned}\widehat{\beta}_{cf} &= \sum_{\ell=1}^{L}\Big(\sum_{g\in I_\ell}\sum_{i=1}^{n_g}\big(\overline{X}_{g,-i} - \widetilde{\phi}_\ell(X_{g,i})\big)'\big(\overline{Y}_{g,-i} - \widetilde{\nu}_\ell(X_{g,i})\big)\Big)^{-1} \\ &\quad \sum_{\ell=1}^{L}\Big(\sum_{g\in I_\ell}\sum_{i=1}^{n_g}\big(\overline{X}_{g,-i} - \widetilde{\phi}_\ell(X_{g,i})\big)'\big(Y_{g,i} - \widetilde{\mu}_\ell(X_{g,i})\big)\Big) \qquad (1.6.1)\end{aligned}$$

where $\widetilde{\mu}_\ell$, $\widetilde{\nu}_\ell$ and $\widetilde{\phi}_\ell$ are estimated using the subset $\widetilde{I}_\ell$. The set $\widetilde{I}_\ell$ belongs to the partitioned L subsets and is disjoint with I_ℓ ($\widetilde{I}_\ell \cap I_\ell = \emptyset$). For example, the two-folded cross-fitting estimator uses one subset to estimate the first step nonparametric parameters and then plug into another subset to obtain the second step estimator of β_0. The two subsets are then flipped to obtain another estimator of β_0. By taking average of these two estimators of β_0, the efficiency of the CF estimator is improved.

However, the single cross-fitting estimator still have a nonlinearity bias which is induced by the product of the estimation bias of $\widetilde{\phi}_\ell$ with $(\widetilde{\mu}_\ell, \widetilde{\nu}_\ell)$. The bias also has the order of $K/\sqrt{n}$ with first step series estimator (Newey and Robins, 2018). To eliminate the nonlinearity bias, I apply the doubly cross-fitting estimator for the social interaction effect β_0 by following Newey and Robins (2018).

To eliminate the nonlinearity bias, the doubly cross-fitting estimator further uses two different subsamples to estimate the first step nonparametric parameters. $\widetilde{\mu}_\ell$ and $\widetilde{\nu}_\ell$ are estimated using the subset $\widetilde{I}_\ell$ which is disjoint with subset I_ℓ, and $\check{\phi}_\ell$ is estimated using subset $\check{I}_\ell$ which is disjoint with both $\widetilde{I}_\ell$ and I_ℓ. Then, a doubly cross-fitting estimator for β_0

is obtained by plugging the nonparametric first step estimators $(\widetilde{\mu}_\ell, \widetilde{\nu}_\ell, \check{\phi}_\ell)$ into the subset I_ℓ.

$$\begin{aligned} \widehat{\beta}_{dcf} &= \sum_{\ell=1}^{L}\Big(\sum_{g\in I_\ell}\sum_{i=1}^{n_g}\big(\overline{X}_{g,-i} - \check{\phi}_\ell(X_{g,i})\big)'\big(\overline{Y}_{g,-i} - \widetilde{\nu}_\ell(X_{g,i})\big)\Big)^{-1} \\ & \quad \sum_{\ell=1}^{L}\Big(\sum_{g\in I_\ell}\sum_{i=1}^{n_g}\big(\overline{X}_{g,-i} - \check{\phi}_\ell(X_{g,i})\big)'\big(Y_{g,i} - \widetilde{\mu}_\ell(X_{g,i})\big)\Big) \end{aligned} \tag{1.6.2}$$

The doubly cross-fitting also use the subsets flipping to obtain an averaging estimator which helps to improve the efficient of the estimator for β_0.

This section will focus on the doubly cross-fitting estimator defined in (1.6.2) by plugging in the first step series estimation of $\vartheta_0 = (\mu_0, \nu_0, \phi_0)$. I continue to apply the power series expansion with basis function $P^K(x) = (P_1(x), \ldots, P_K(x))$. The nonparametric series estimators $\widetilde{\mu}(x), \widetilde{\nu}(x), \check{\phi}(x)$ are constructed by

$$\begin{aligned} \widetilde{\mu}(x) &= P^K(x)\left(\frac{1}{\widetilde{G}_\ell}\sum_{g\in\widetilde{I}_\ell} P^K(X_g)P^K(X_g)'\right)^{-1}\left(\frac{1}{\widetilde{G}_\ell}\sum_{g\in\widetilde{I}_\ell} P^K(X_g)Y_g\right) \\ \widetilde{\nu}(x) &= P^K(x)\left(\frac{1}{\widetilde{G}_\ell}\sum_{g\in\widetilde{I}_\ell} P^K(X_g)P^K(X_g)'\right)^{-1}\left(\frac{1}{\widetilde{G}_\ell}\sum_{g\in\widetilde{I}_\ell} P^K(X_g)\overline{Y}_{g,-}\right) \\ \check{\phi}(x) &= P^K(x)\left(\frac{1}{\check{G}_\ell}\sum_{g\in\check{I}_\ell} P^K(X_g)P^K(X_g)'\right)^{-1}\left(\frac{1}{\check{G}_\ell}\sum_{g\in\check{I}_\ell} P^K(X_g)\overline{X}_{g,-}\right) \end{aligned} \tag{1.6.3}$$

where $\widetilde{I}_\ell$ is the index set for estimating μ_0 and ν_0, $\check{I}_\ell$ for estimating ϕ_0. $\widetilde{G}_\ell$ and $\check{G}_\ell$ denote the number of groups for these two subsamples.

In the following subsection, I will discuss the asymptotic property of the doubly cross-fitting estimator defined in 1.6.2 by plugging in the first step series estimators defined in (1.6.3).

1.6.2 Asymptotic Normality of Cross-fitting Estimators

The asymptotic considered in this paper is to let the number of groups $G \to \infty$. The cross-fitting estimator needs to split the groups into L subgroups ($L \geq 3$ for doubly cross-fitting estimator). Let G_ℓ denote the number of groups in the subset indexed by I_ℓ. The following Assumption 1.6.1 imposes restrictions on the sample spitting.

Assumption 1.6.1: *(i) Assume $\{I_\ell\}(\ell = 1, ...L)$ are mutually exclusive and exhaustive subsets of the group index set $\{1, ..., G\}$. That is, $\sum_{\ell=1}^{L} G_\ell = G$ and $I_\ell \cap I_{\ell'} = \emptyset$ ($\ell \neq \ell' \in \{1, \cdots, L\}$). (ii) Assume the number of groups in each subset is of the same order as G. That is, $\frac{G_\ell}{G} = c$ ($\ell = 1, ...L$), where $0 < c < \infty$.*

The assumption $\sum_{\ell=1}^{L} G_\ell = G$ guarantees that each grouped data Z_g can be used to estimate β_0 in the second step if we flipping the subsets. It is also assumed that G_ℓ has the same rate with G. Thus, $G \to \infty$ implies that each subset has $G_\ell \to \infty$ ($\ell = 1, ..., L$). Practically, the groups can be partitioned into nearly equal sized subsets.

Next, I will discuss the conditions for the asymptotic normality of the doubly cross-fitting estimator $\widehat{\beta}_{dcf}$ by plugging in the first step series estimation of ϑ_0 in (1.6.3). The conditions for bound the approximation error of the series expansions are the same with Assumption 1.5.1.

The conditions for the asymptotic normality of $\widehat{\beta}_{dcf}$ requires the remainder terms to be $o_p(1)$. The doubly cross-fitting estimator removes the own observation and nonlinearity bias by estimating different nonparametric and parametric parameters using disjoint subsamples. In general, Newey and Robins (2018) shows that the doubly cross-fitting estimator can obtain the fastest remainder rate with first step series. For the semiparametric social interaction model with clustered data, the following result gives the conditions for the asymptotic normality of $\widehat{\beta}_{dcf}$ defined in (1.6.2).

Theorem 1.6.1: *If Assumptions 1.3.1, 1.3.2, 1.3.3, and 1.5.1 are satisfied and assume the following condition holds*

$$\sqrt{n}K^{-\left(\alpha_\phi+\alpha_\mu\right)} + \sqrt{n}K^{-\left(\alpha_\phi+\alpha_\nu\right)} + K^{-\alpha_\phi} + K^{-\alpha_\mu} + K^{-\alpha_\nu} + \sqrt{K}/\sqrt{n} \to 0 \tag{1.6.4}$$

then it follows that

$$\sqrt{n}V_n^{-1/2}(\widehat{\beta}_{dcf} - \beta_0) \xrightarrow{d} N(0, 1). \tag{1.6.5}$$

The proof of Theorem 1.6.1 will be shown in Appendix 1.A.

Comparing the condition (1.5.12) for doubly robust estimator, the condition (1.6.4) for the doubly cross-fitting estimator improves the remainder rate by changing the estimation bias from order $K/\sqrt{n}$ to $\sqrt{K}/\sqrt{n}$. If $K/n \to 0$ is satisfied, then the remainder term is dominated by the order of $\sqrt{n}K^{-\left(\alpha_\phi+\max\{\alpha_\mu,\alpha_\nu\}\right)}$. Furthermore, there exists $K = K_n$ satisfying $K/n \to 0$, such that condition (1.6.4) holds if and only if $\alpha_\phi + \max\{\alpha_\mu, \alpha_\nu\} > 1/2$.

1.7 Monte Carlo Simulation

This section investigates the finite sample performance of semiparametric two-step estimators for endogenous social interaction effects, β_0, using Monte Carlo simulations. The goal is to show that parametric linear estimator might lead the spurious social interaction effect if the true model is nonlinear, under which the semiparametric model should be applied.

To show the possible spurious social interaction effect when the linear estimator is applied, I set the true value of endogenous social interaction effect $\beta_0 = 0.5$. Different data generating processes (DGPs)are considered with different nonparametric nuisance functions, including the linear form and also nonlinear cases. For each DGP, I apply both the parametric linear and semiparametric series estimators to obtain the estimator for β_0. The results show that if the true model is nonlinear, then the linear estimator suffers from an over-rejection problem and more likely to obtain a significant social interaction effect which is actually spurious. However, the semiparametric series estimators do not have such an issue and have the correct size across all the nonlinear models considered in the simulation study.

1.7.1 Simulation Set-up

In the Monte Carlo Simulation studies, I consider the semiparametric social interaction model with group size $M_g = 2$ across all the groups.

$$\begin{bmatrix} Y_{g,1} \\ Y_{g,2} \end{bmatrix} = \beta_0 \begin{bmatrix} Y_{g,2} \\ Y_{g,1} \end{bmatrix} + \begin{bmatrix} h_0(X_{g,1}) \\ h_0(X_{g,2}) \end{bmatrix} + \begin{bmatrix} U_{g,1} \\ U_{g,2} \end{bmatrix}, \quad g = 1, \cdots, G. \tag{1.7.1}$$

The following describes the DGPs considered in the Monte Carlo simulation studies.

(i) The dimensionality of control variables X.

This simulation study considers the DGPs with multivariate case, $\dim(X) = 3$.

(ii) The function forms of nonparametric nuisance function $h_0(\cdot)$.

Both the linear and nonlinear function forms are considered in the simulation study, including $h_0(X) = X\gamma$, $h_0(X) = exp(X\gamma)$, and $h_0(X) = sin(X\gamma) + cos(X\gamma)$.

(iii) The distribution of the control variables X.

Assume the group pair regressors $(X_{g,1}, X_{g,2})$ are independently draw across groups. I wil consider four cases in the simulation study: (1) $(X_{g,1}, X_{g,2})$ follows a independent normal distribution; (2) $(X_{g,1}, X_{g,2})$ follows joint normal distribution with correlation $\rho = 0.5$; (3) $(X_{g,1}, X_{g,2})$ follows a dependent bivariate logistic distribution; (4) $(X_{g,1}, X_{g,2})$ follows a uncorrelated but dependent distribution.

(iv) The distribution of the disturbance U_i.

Assume $U_i \overset{iid}{\sim} N(0, \sigma_u^2)$ where σ_u^2 measures the noise/signal level. For all the DGPs, $\sigma_u^2 = 0.5$ is considered.

(v) The parameter of interest β_0.

β_0 should belong to $(-1, 1)$. For discussing the possible spurious social interaction effect, let the true $\beta_0 = 0.5$.

For each simulated dataset, the endogenous social interaction effect, β_0, is estimated by plugging in the first step parametric linear estimator or the nonparametric series estimator. Linear estimators simply apply the linear regression with X as explanatory variables. Series estimators consider using the polynomial expansion of X as explanatory variables which should be more robust under the nonlinear case. The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

To evaluate the performance of different estimators, this paper uses the following measures: Bias, Variance, mean squared error (MSE), mean absolute error (MAE) for estimators $\widehat{\beta}$ and empirical size for the corresponding t test.

1.7.2 Simulation Results

The simulation results for the four cases with different DGP are reported in Table 1.1 - Table 1.4.

Case I: Uncorrelated Bivariate-normal Distribution

Table 1.1 reports the simulation results that assumes $(X_{g,i}, X_{g,-i})$ follows a joint normal distribution with correlation $\rho = 0$. Since $X_{g,i}$ and $X_{g,-i}$ are independent, $X_{g,-i}$ is a valid IV even if the we omitted the nonparametric function $h_0(x)$. Thus, both the parametric linear estimator and semiparametric series estimator should be consistent. This property is reflected in the simulation results of Table 1.1. It shows that the estimator for β_0 by applying both the parametric linear estimator and nonparametric series estimator have similar performance in terms of MSE and MAE, and have correct size across all the linear and nonlinear DGPs considered.

Case II: Correlated Bivariate-normal Distribution

In Table 1.2, the results for a joint normal with correlation $\rho = 0.5$ are reported. One feature of the joint normal distribution is that $X_{g,-i}$ is uncorrelated with all the even order terms of $X_{g,i}$, $(cor(X_{g,-i}, X_{g,i}^2) = 0$, et.al.). Thus, omitting all the even order terms $(X_{g,i}^2, X_{g,i}^4, ...)$ does not generate bias in the estimation of β_0. But it might still suffer from model misspecification bias for omitting the odd order terms such as $(X_{g,i}^3, X_{g,i}^5, ...)$. From the simulation results in Table 1.2 we learn that, the parametric linear estimator is slightly biased and oversized if the true model is nonlinear.

Case III: Correlated Bi-logistic Distribution

In the simulation study, I also consider the case that $(X_{g,i}, X_{g,-i})$ is generated from bivariate-logistic distribution with dependence parameter $r = 0.5$. Under this DGP, $X_{g,-i}$ is correlated with each term of the polynomials of $X_{g,i}$. Thus, model misspecification might lead to nonnegligible bias for the parametric linear estimator of β_0. From the simulation results in Table 1.3, it can be found that linear estimators are severely biased and have over-rejection problem under the nonlinear model. However, the semiparametric series estimator proposed in this paper is robust to model specification, which has small bias and correct size across the DGPs considered.

Case IV: Uncorrelated but Dependent Distribution

Finally, a special case of uncorrelated but dependent distribution for $(X_{g,i}, X_{g,-i})$ is considered. To be specific, $X_{g,i} \sim U[-1, 1]$, and $X_{g,-i} = 1\{X_{g,i} > 0\}X_{g,i} - 1\{X_{g,i} \leq 0\}X_{g,i}$[3]. It can be shown that $X_{g,i}$ and $X_{g,-i}$ are linearly uncorrelated but are dependent in higher order terms. The simulation results are reported in Table 1.4 which have similar conclusion as case III.

Simulation Summary

From the simulation results in Table 1.1-Table 1.4, it can learned that the parametric linear estimator has the correct size if the true model is linear or groups are randomly assigned. But it suffers from an over-rejection problem if applied to the nonlinear model. The results indicate that using linear estimators is likely to obtain a significant social interaction effect which is spurious. It can also be learned from the simulation result that the distortion is more server if the true function becomes more curvature. However, the semiparametric series estimators do not have such an issue and have the correct size for the nonlinear models considered in the simulation study.

1.8 Conclusion and Future Research

The existing literature on the social interaction model focus on the identification issue of the parametric setup. However, the parametric model is often restrictive and might lead to a

[3] $1(.)$ denote the indicator function.

spurious or misleading social interaction effect in the empirical studies. This paper studies a semiparametric social interaction model with a parametric linear-in-means endogenous social interaction part and a nonparametric control variables part. To highlight the semiparametric feature of the model, this paper excludes the correlated effect and also imposes restrictions on the contextual effect to avoid the complexity of identification issues.

This paper adopts a semiparametric instrumental variable (IV) approach to identify the endogenous social interaction effect. Based on the identification condition, this paper proposes a two-step semiparametric estimator in which the first step nuisance functions could be estimated by any nonparametric methods. This paper focused on the series estimation in the first step. The second step parametric components are then estimated by method of moment (MM) or generalized method of moments (GMM). I also consider using the orthogonal moment condition in the second step estimation to reduce the bias induced by the first step nonparametric estimators. The result shows that the two-step semiparametric estimator with first step series is root-n consistent and asymptotically normally distributed under regularity conditions.

The strong exogenous assumption and restricted contextual effect imposed on our model could be a problem. It is meaningful to study how to disentangle the endogenous social interaction effect from the contextual and correlated effect in the semiparametric model setup. The identification strategy used in the parametric social interaction literature could be applied to the semiparametric model. Based on the identification condition, in principle, it is straightforward to generalize our semiparametric two-step estimation methods to this case. I will consider relaxing these assumptions in future research.

Besides, the discussed group social interaction model can also be extended to a more flexible network social interaction model. The network depicts the connections between individuals and does not need to have the group structural which can be applied to a much richer social structures. Also, the network structural impose certain exclusion restrictions on the model and make the identification to be easier (Bramoullé, Djebbari, and Fortin (2009)). Based on the identification condition, the semiparametric two-step estimation methods can also be applied to the social interaction model with networks. I will consider extending our results for the semiparametric group social interaction model to the semiparametric network social interaction model in future research.

Appendix

1.A Proofs

1.A.1 Proofs in Section 1.3

Proof for Theorem 1.3.1

Proof: *First, consider the identification of $h_0(\cdot)$ given β_0 identified. Under assumption 1.3.1, $U_{g,i}$ and $X_{g,i}$ are independent, then $h_0(\cdot)$ is nonparametrically identified by*

$$h_0(X_{g,i}) = E[Y_{g,i} - \beta_0 \overline{Y}_{g,-i} | X_{g,i}].$$

Next, consider the identification of endogenous social interaction effect β_0. Based on Assumptions 1.3.1 and 1.3.2, $E[U_{g,i}|X_{g,i}, n_g] = 0$, the nuisance function $h_0(\cdot)$ can be partialled out from the original model (1.3.1).

$$Y_{g,i} - E[Y_{g,i}|X_{g,i}, n_g] = \beta_0 \big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|X_{g,i}, n_g]\big) + U_{g,i}. \tag{1.A.1}$$

Due to the simultaneous structural of the social interaction model, the term $\overline{Y}_{g,-i}$ is endogenous, i.e. $E\big[\overline{Y}_{g,-i} U_i\big] \neq 0$. In order to identify β_0, we adopt an IV approach by using the additional exogenous assumption $E[U_{g,i}|X_{g,-i}] = 0$ in Assumption 1.3.2 and impose the condition on equation (1.A.1),

$$E\left[(Y_{g,i} - E[Y_{g,i}|X_{g,i}])\,\big|X_{g,-i}\right] = \beta_0 E\left[\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|X_{g,i}]\big)\,\big|X_{g,-i}\right]. \tag{1.A.2}$$

Then by law of iterated expectation, it follows the unconditional moment condition,

$$E\left[t(X_{g,-i})\left(Y_{g,i} - E[Y_{g,i}|X_{g,i}, n_g]\right)\right] = \beta_0 E\left[t(X_{g,-i})\left(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|X_{g,i}, n_g]\right)\right], \tag{1.A.3}$$

holds for any function $t(x) : R^{d(n_g-1)} \to R^q$ with $E[\|t(x)\|^2] < \infty$. $t(X_{g,-i})$ can be taken as the instrumental variables for the endogenous outcome variable $\overline{Y}_{g,-i}$. Then the identification *of β_0 can be achieved under full rank condition in Assumption 1.3.3.* $\square$

Proof for Theorem 1.3.2

Proof: *The proof for 1.3.2 is similar as Theorem 1.3.1. First consider the identification of $h_0(\cdot)$ given β_0 identified. Under Assumption 1.3.4, $U_{g,i}$ and $\mathcal{X}_{g,i}$ are independent, then $h_0(\cdot)$ is nonparametrically identified by*

$$E[Y_{g,i} - \beta_0 \overline{Y}_{g,-i} | \mathcal{X}_{g,i}] = h_0(\mathcal{X}_{g,i}) \tag{1.A.4}$$

Next, consider the identification of endogenous social interaction effect β_0. Based on Assumption 1.3.4, $E[U_{g,i}|\mathcal{X}_{g,i}] = 0$, we have

$$E[Y_{g,i}|\mathcal{X}_{g,i}] = \beta_0 E[\overline{Y}_{g,-i}|\mathcal{X}_{g,i}] + h_0(\mathcal{X}_{g,i}) \tag{1.A.5}$$

We can partial out the identified nuisance function $h_0(\cdot)$ by subtracting (1.A.5) form the original model (1.2.1) for both side:.

$$Y_{g,i} - E[Y_{g,i}|\mathcal{X}_{g,i}] = \beta_0 \big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|\mathcal{X}_{g,i}]\big) + U_{g,i} \tag{1.A.6}$$

Combing the exclusive restriction, the exogenous assumption 1.3.4, $\overline{X}_{g,-i}$ that is not included in the model can be used as IV for $\overline{Y}_{g,-i}$ to identify β_0.

$$E\left[\overline{X}_{g,-i}\big(Y_{g,i} - E[Y_{g,i}|\mathcal{X}_{g,i}]\big)\right] = \beta_0 E\left[\overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|\mathcal{X}_{g,i}]\big)\right]. \tag{1.A.7}$$

Then β_0 is identified given the standard full rank condition in Assumption 1.3.5. $\square$

1.A.2 Proofs in Section 1.4

The original moment and orthogonal moment functions are defined as

$$\begin{aligned}
m(Z_{g,i}, \beta, \mu, \nu) &= \overline{X}_{g,-i}\Big(\big(Y_{g,i} - \mu(X_{g,i})\big) - \beta\big(\overline{Y}_{g,-i} - \nu(X_{g,i})\big)\Big), \\
\psi(Z_{g,i}, \beta, \mu, \nu, \phi) &= \big(\overline{X}_{g,-i} - \phi(X_{g,i})\big)\Big(\big(Y_{g,i} - \mu(X_{g,i})\big) - \beta\big(\overline{Y}_{g,-i} - \nu(X_{g,i})\big)\Big).
\end{aligned}$$

Verify the orthogonality of $m(\cdot)$ and $\psi(\cdot)$

Proof: *First, we can show that the pairwise derivative of original moment function with respect to μ and ν are not zero.*

$$\begin{aligned}
\frac{\partial E\left[m\left(Z_{g,i};\beta_0,\mu_0,\nu_0\right)\right]}{\partial \mu}[v_\mu] &= \frac{\partial E\left[m\left(Z_{g,i};\beta_0,\mu_0+\tau v_\mu,\nu_0\right)\right]}{\partial \tau}\Big|_{\tau=0} = E\left[X_{g,-i}\right] \neq 0 \\
\frac{\partial E\left[m\left(Z_{g,i};\beta_0,\mu_0,\nu_0\right)\right]}{\partial \nu}[v_\nu] &= \frac{\partial E\left[m\left(Z_{g,i};\beta_0,\mu_0,\nu_0+\tau v_\nu\right)\right]}{\partial \tau}\Big|_{\tau=0} = -\beta_0 E\left[X_{g,-i}\right] \neq 0
\end{aligned}$$

Next, we can show that the pairwise derivatives for orthogonal moment function with respect to μ, ν and ϕ vanish.

$$\begin{aligned}
\frac{\partial E\left[\psi\left(Z_{g,i};\beta_0,\mu_0,\nu_0,\phi_0\right)\right]}{\partial \mu}[v_\mu] &= \frac{\partial E\left[m\left(Z_{g,i};\beta_0,\mu_0+\tau v_\mu,\nu_0,\phi_0\right)\right]}{\partial \tau}\Big|_{\tau=0} \\
&= -E\left[X_{g,-i} - \phi(X_{g,i})\right] = E\left[X_{g,-i} - E[X_{g,-i}|X_{g,i}]\right] = 0 \\
\frac{\partial E\left[\psi\left(Z_{g,i};\beta_0,\mu_0,\nu_0,\phi_0\right)\right]}{\partial \nu}[v_\nu] &= \frac{\partial E\left[\psi\left(Z_{g,i};\beta_0,\mu_0,\nu_0+\tau v_\nu,\phi_0\right)\right]}{\partial \tau}\Big|_{\tau=0} \\
&= -\beta_0 E\left[X_{g,-i} - \phi(X_{g,i})\right] = 0 \\
\frac{\partial E\left[\psi\left(Z_{g,i};\beta_0,\mu_0,\nu_0,\phi_0\right)\right]}{\partial \phi}[v_\phi] &= \frac{\partial E\left[\psi\left(Z_{g,i};\beta_0,\mu_0,\nu_0,\phi_0+\tau v_\phi\right)\right]}{\partial \tau}\Big|_{\tau=0} \\
&= -E\big[\big(Y_{g,i} - \mu(X_{g,i})\big) - \beta_0\big(Y_{g,-i} - \nu(X_{g,i})\big)\big] = 0. \quad \square
\end{aligned}$$

Verify the doubly robustness of $\psi(\cdot)$

First, it can be verified that $\psi(Z_{g,i}, \beta, \mu, \nu, \phi)$ is affine in each of μ, ν, and ϕ holding others fixed. For any $0 < \lambda < 1$, $\mu_1(\cdot)$, and $\mu_2(\cdot)$,

$$
\begin{aligned}
& \psi(Z_{g,i}, \beta, \lambda\mu_1 + (1-\lambda)\mu_2, \nu, \phi) \\
= \; & \left(\overline{X}_{g,-i} - \phi(X_{g,i})\right)\left(\left(Y_{g,i} - \lambda\mu_1(X_{g,i}) - (1-\lambda)\mu_2(X_{g,i})\right) - \beta\left(\overline{Y}_{g,-i} - \nu(X_{g,i})\right)\right) \\
= \; & \lambda\psi(Z_{g,i}, \beta, \mu_1, \nu, \phi) + (1-\lambda)\psi(Z_{g,i}, \beta, \mu_2, \nu, \phi)
\end{aligned}
$$

Thus, $\psi(Z_{g,i}, \beta, \mu, \nu, \phi)$ is affine in μ holding ν, ϕ fixed. Similarly, it can be shown that $\psi(Z_{g,i}, \beta, \mu, \nu, \phi)$ is affine in ν and ϕ.

Then the doubly robustness of $\psi(\cdot)$ follows by Theorem 5 in Chernozhukov et al. (2018a).

1.A.3 Proofs in Section 1.5

Proof for Proposition 1.5.1

The influence function of $\frac{1}{\sqrt{n}}\sum_{g=1}^{G} m_g\left(\beta_0, \widehat{\vartheta}\right)$ is $\frac{1}{\sqrt{n}}\sum_{g=1}^{G} \psi_g\left(\beta_0, \vartheta_0\right)$. Thus, the estimation of $\vartheta_0 = (\mu_0, \nu_0, \phi_0)$ will not affect $\Omega_{db} = Var\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} \psi_g\big(\beta_0, \widehat{\vartheta}\big)\right)$. Then, the results follows directly from Ackerberg, Chen, Hahn, and Liao (2014).

Proof for Proposition 1.5.2

Proof: *To proof Proposition 1.5.2, I use vector form of the estimators.*
First, we consider the just identified case for β_0. We can show that debiased estimator, $\widehat{\beta}_{db}$ is exactly the same with the original estimator, $\widehat{\beta}$, in (1.5.7) if we plugging in the series estimators in the first step, i.e.

$$
\widehat{\beta} = (X_-'\widehat{\zeta})^{-1}(X_-'\widehat{\eta}) \quad = \quad \widehat{\beta}_{db} = (\widehat{\varepsilon}'\widehat{\zeta})^{-1}(\widehat{\varepsilon}'\widehat{\eta}) \tag{1.A.8}
$$

Let $M_P = I - P(P'P)^{-1}P'$, the proof for (1.A.8) is as follows:

$$\begin{aligned}
\widehat{\beta}_{db} &= (\widehat{\varepsilon}'\widehat{\zeta})^{-1}(\widehat{\varepsilon}'\widehat{\eta}) \\
&= \left[(M_P X_-)'(M_P Y_-)\right]^{-1}\left[(M_P X_-)'(M_P Y)\right] \\
&= (X_- M_P Y_-)^{-1}(X_- M_P Y) \\
&= (X_-'\widehat{\zeta})^{-1}(X_-'\widehat{\eta}) \\
&= \widehat{\beta}
\end{aligned}$$

Next, we consider the over identified case, that is the more usual case which studies $\dim(X_i) = d > 1$. *We can also show that the debiased semiparametric GMM estimator,* $\widehat{\beta}^{gmm}_{db}$*, defined in (1.5.7) is the same with the original semiparametric GMM estimator,* $\widehat{\beta}^{gmm}$*, in (1.5.7) if we plugging in series estimator for the first step, i.e.*

$$\widehat{\beta}^{gmm} = \left(\widehat{\zeta}' X_- \widehat{\Omega}^{-1} X_-' \widehat{\zeta}\right)^{-1}\left(\widehat{\zeta}' X_- \widehat{\Omega}^{-1} X_-' \widehat{\eta}\right) \; = \; \widehat{\beta}^{gmm}_{db} = \left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\Omega}^{-1}_{db}\widehat{\varepsilon}'\widehat{\zeta}\right)^{-1}\left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\Omega}^{-1}_{db}\widehat{\varepsilon}'\widehat{\eta}\right) \quad (1.A.9)$$

Firstly, we can show that the weight matrix are the same. i.e.

$$\widehat{\Omega} = \widehat{\varepsilon}'\widehat{U}\widehat{U}'\widehat{\varepsilon} \; = \; \widehat{\Omega}_{db} = \widehat{\varepsilon}'\widetilde{U}\widetilde{U}'\widehat{\varepsilon} \qquad (1.A.10)$$

where $\widehat{U} = \widehat{\eta} - \widetilde{\beta}\widehat{\zeta}$ *and* $\widetilde{U} = \widehat{\eta} - \widetilde{\beta}_{db}\widehat{\zeta}$. *Since we plug in the same first step estimators* $(\widehat{\eta}, \widehat{\zeta}, \widehat{\varepsilon})$ *and also can show the preliminary estimators* $\widetilde{\beta}$ *and* $\widetilde{\beta}_{db}$ *are the same*

$$\begin{aligned}
\widetilde{\beta}_{db} &= \left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\varepsilon}'\widehat{\zeta}\right)^{-1}\left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\varepsilon}'\widehat{\eta}\right) \\
&= \left(Y_-' M_P M_P X_- X_-' M_P M_P Y_-\right)^{-1}\left(Y_-' M_P M_P X_- X_-' M_P M_P Y\right) \\
&= \left(\widehat{\zeta}' X_- X_-'\widehat{\zeta}\right)^{-1}\left(\widehat{\zeta}' X_- X_-'\widehat{\eta}\right) = \widetilde{\beta},
\end{aligned}$$

it follows that $\widehat{\Omega} = \widehat{\Omega}_{db}$. *Then we can show that the semiparametric debiased GMM estimator* *equals to the original estimator as follows:*

$$
\begin{aligned}
\widehat{\beta}_{db}^{gmm} &= \left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\Omega}_{db}^{-1}\widehat{\varepsilon}'\widehat{\zeta}\right)^{-1}\left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\Omega}_{db}^{-1}\widehat{\varepsilon}'\widehat{\eta}\right) \\
&= \left(Y_-'M_PM_PX_-\widehat{\Omega}^{-1}X_-'M_PM_PY_-\right)^{-1}\left(Y_-'M_PM_PX_-\widehat{\Omega}^{-1}X_-'M_PM_PY\right) \\
&= \left(\widehat{\zeta}'X_-\widehat{\Omega}^{-1}X_-'\widehat{\zeta}\right)^{-1}\left(\widehat{\zeta}'X_-\widehat{\Omega}^{-1}X_-'\widehat{\eta}\right) \\
&= \widehat{\beta}^{gmm}. \quad \square
\end{aligned}
$$

Lemmas for proof of Theorem 1.5.1 and 1.5.2

(I) WLLN and CLT for grouped data

Assumption 1.A.1: *Assume observations X_g are independently draw across groups, $X_{g,i}$ has identical marginal distribution, and the group size n_g is fixed but are allowed to be different across groups.*

Let the sample mean $\overline{X}_n = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{g,i}$, and group sums $\overline{X}_g = \sum_{i=1}^{n_g} X_{g,i}$ which are mutually independent given random sampling for groups. Then $\overline{X}_n = \frac{1}{n}\sum_{g=1}^{G}\overline{X}_g$. The variance of $\sqrt{N}\overline{X}_n$ is

$$\Omega_n = E\left[n\left(\overline{X}_n - E\overline{X}_n\right)\left(\overline{X}_n - E\overline{X}_n\right)'\right] = \frac{1}{n}\sum_{g=1}^{G} E\left[\left(\overline{X}_g - E[\overline{X}_g]\right)\left(\overline{X}_g - E[\overline{X}_g]\right)'\right].$$

Lemma 1.A.1 (WLLN for grouped data (Hansen and Lee 2018)):
If Assumption 1.A.1 hold and $E[|X_{g,i}|] < \infty$, then as $n \to \infty$,

$$\|\overline{X}_n - E[\overline{X}_n]\| \xrightarrow{p} 0$$

Remark 1: A sufficient condition allowing for distributional heterogeneity is

$$\sup_{g,i} E[|X_{g,i}|^r] < \infty \quad \textit{for some} \quad r > 1.$$

Remark 2: The fixed group size assumption can be relaxed to

$$\max_{g\leq G}\frac{n_g}{n}\rightarrow 0,\quad n\rightarrow\infty$$

Lemma 1.A.2 (CLT for grouped data (Hansen and Lee 2018)):
If Assumption 1.A.1 hold and $E\left\|X_{g,i}\right\|^2<\infty$, $\lambda_n=\lambda_{\min}\left(\Omega_n\right)\geq\lambda>0$. *Then*

$$\sqrt{n}\ \Omega_n^{-1/2}\left(\overline{X}_n-E\overline{X}_n\right)\xrightarrow{d}N\left(\mathbf{0},I_d\right)$$

Remark i: A sufficient condition allowing for distributional heterogeneity is

$$\sup_{g,i}E\left\|X_{g,i}\right\|^s<\infty\ \ \textit{for some}\ \ s>r\geq 2$$

Remark ii: The fixed group size assumption can be relaxed to

$$\max_{g\leq G}\frac{n_g^2}{n}\rightarrow 0,\quad n\rightarrow\infty$$

But then the convergence rate of $\Omega_n^{1/2}$ *might not be* $n^{-1/2}$. *Because we also have the approximation and estimation error to control, we just assume* n_g *is fixed.*

(II) Lemmas for series estimator

Lemma 1.A.3: *Let* $Q_n=P'P/n$ *Under assumption 1.5.1, we have,*

$$Q_n^{-1/2}\frac{P'\zeta}{n}=O_p\big(\sqrt{K/n}\big),\quad \textit{and}\ \ Q_n^{-1/2}\frac{P'\varepsilon}{n}=O_p\big(\sqrt{K/n}\big)\tag{1.A.11}$$

Proof (Lemma 1.A.3): *Following Newey (1997) Theorem 1, we can show that*

$$\begin{aligned}
E\big[\|Q_n^{-1/2}P'\zeta/n\|^2|X\big] &= E\big[\zeta'P(P'P)^{-1}P'\zeta\big|X\big]/n \\
&= E\big[tr\big(P(P'P)^{-1}P'\zeta\zeta'\big)|X\big]/n \\
&= tr\big(P(P'P)^{-1}P'E[\zeta\zeta'|X]\big)/n \\
&\lesssim K/n
\end{aligned}$$

By Markov inequality, $Q_n^{-1/2}\frac{P'\zeta}{n} = O_p\big(\sqrt{K/n}\big)$

Similarly, we can also show that $Q_n^{-1/2}\frac{P'\varepsilon}{n} = O_p\big(\sqrt{K/n}\big)$ □

Lemma 1.A.4: *Let* $Q_n = P'P/n$ *Under assumption 1.5.1, we have,*

$$Q_n^{-1/2}\frac{P'(\phi-\phi^K)}{n} = O_p\big(K^{-\alpha_\phi}\sqrt{K/n}\big), \quad \textit{and} \quad Q_n^{-1/2}\frac{P'(\nu-\nu^K)}{n} = O_p\big(K^{-\alpha_\nu}\sqrt{K/n}\big) \tag{1.A.12}$$

Proof (Lemma 1.A.3): *Following Newey (1997) Theorem 1, we can show that*

$$\begin{aligned}
E\big[\|Q_n^{-1/2}P'(\phi-\phi^K)/n\|^2\big] &= E\big[(\phi-\phi^K)'P(P'P)^{-1}P'(\phi-\phi^K)\big]/n \\
&= E\big[tr\big(P(P'P)^{-1}P'(\phi-\phi^K)(\phi-\phi^K)'\big)\big]/n \\
&= trE\big[P(P'P)^{-1}P'(\phi-\phi^K)(\phi-\phi^K)'\big]/n \\
&\lesssim O(K^{-2\alpha_\phi}K/n)
\end{aligned}$$

By Markov inequality, $Q_n^{-1/2}\frac{P'(\phi-\phi^K)}{n} = O_p\big(K^{-\alpha_\phi}\sqrt{K/n}\big)$.

Similarly, we can also show that $Q_n^{-1/2}\frac{P'(\nu-\nu^K)}{n} = O_p\big(K^{-\alpha_\nu}\sqrt{K/n}\big)$ □

Lemma 1.A.5:

$$\frac{\zeta'(\phi-\phi^K)}{n} = O_p(K^{-\alpha_\phi}/\sqrt{n}), \quad \frac{\varepsilon'(\nu-\nu^K)}{n} = O_p(K^{-\alpha_\nu}/\sqrt{n})$$

Proof (Lemma 1.A.5): *First consider* $\frac{1}{n}\zeta'(\phi-\phi^K)$,

$$
\begin{aligned}
E\big[\|\zeta'(\phi-\phi^K)/n\|^2|X\big] &= E\big[(\phi-\phi^K)'\zeta\zeta'(\phi-\phi^K)|X\big]/n^2 \\
&= E\big[tr\big((\phi-\phi^K)'\zeta\zeta'(\phi-\phi^K)\big)|X\big]/n^2 \\
&= tr\big((\phi-\phi^K)(\phi-\phi^K)'\big)E[\zeta\zeta'|X]/n^2 \\
&\lesssim tr\big((\phi-\phi^K)(\phi-\phi^K)'\big)/n^2 \\
&= \|\phi(X)-\phi^K(X)\|_2^2/n^2 \\
&= O(K^{-2\alpha}/n)
\end{aligned}
$$

where α_ϕ *is related to the smoothness of the function* $\phi(x)$*, the dimensionality of* x*,* d*. For power series,* $\alpha_\phi = s_\phi/d$ *where* s_ϕ *is the number of continuous derivatives of* $\phi(x)$ *that exist. By Markov inequality,*

$$\zeta'(\phi-\phi^K)/n = O_p(K^{-\alpha_\phi}/\sqrt{n}).$$

Similarly, $\varepsilon'(\nu-\nu^K)/n = O_p(K^{-\alpha_\nu}/\sqrt{n})$*, where* $\alpha_\nu = s_\nu/d$*.*

Proof for Theorem 1.5.1

Proof: *Show* $\sqrt{n}V_n^{-1/2}(\widehat{\beta}_{db}-\beta_0) \xrightarrow{d} N(0,1)$.

$$
\begin{aligned}
\sqrt{n}(\widehat{\beta}_{db}-\beta_0) &= \Big(\frac{1}{n}\sum_{i=1}^n \big(X_{-i}-\widehat{\phi}(X_i)\big)\big(Y_{-i}-\widehat{\nu}(X_i)\big)\Big)^{-1} \\
&\quad \Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(X_{-i}-\widehat{\phi}(X_i)\big)\big((Y_i-\widehat{\mu}(X_i))-\beta_0(Y_{-i}-\widehat{\nu}(X_i))\big)\Big)
\end{aligned}
$$

(i) First Consider the Jacobian term,

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\big(X_{-i}-\widehat{\phi}(X_i)\big)\big(Y_{-i}-\widehat{\nu}(X_i)\big) \\
= & \frac{1}{n}\sum_{i=1}^{n}\Big(X_{-i}-\phi(X_i)+\phi(X_i)-\widehat{\phi}(X_i)\Big)\Big(Y_{-i}-\nu(X_i)+\nu(X_i)-\widehat{\nu}(X_i)\Big) \\
= & \frac{1}{n}\sum_{i=1}^{n}\big(\varepsilon_i+\phi_i-\widehat{\phi}_i\big)\big(\zeta_i+\nu_i-\widehat{\nu}_i\big) \\
= & \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\zeta_i}_{\textit{LLN for group data}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big(\varepsilon_i(\nu_i-\widehat{\nu}_i)+\zeta_i(\phi_i-\widehat{\phi}_i)+(\phi_i-\widehat{\phi}_i)(\nu_i-\widehat{\nu}_i)\Big)}_{A\ =\ A1+A2+A3}
\end{aligned}
$$

By Assumption 1.5.1(i), $E\big[|\varepsilon_i\zeta_i|\big]<\infty$, *then by WLLN for grouped data for* $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\zeta_i = \frac{1}{n}\sum_{g=1}^{G}\varepsilon_g'\zeta_g$

$$
\frac{1}{n}\sum_{g=1}^{G}\varepsilon_g'\zeta_g-\frac{1}{n}\sum_{g=1}^{G}E\Big[\varepsilon_g'\zeta_g\Big]\xrightarrow{p}0 \tag{1.A.13}
$$

Consider $\nu_i-\widehat{\nu}_i;\ \mu_i-\widehat{\mu}_i;\ \phi_i-\widehat{\phi}_i,$

$$
\begin{aligned}
\nu_i-\widehat{\nu}_i &= (\nu_i-\nu_i^K)+\nu_i^K-\widehat{\nu}_i \\
&= (\nu_i-\nu_i^K)-P(X_i)(P'P)^{-1}P'(Y_{-}-\nu^K) \\
&= (\nu_i-\nu_i^K)-P(X_i)(P'P)^{-1}P'(\nu-\nu^K)-P(X_i)(P'P)^{-1}P'\zeta \\
\mu_i-\widehat{\mu}_i &= (\mu_i-\mu_i^K)-P(X_i)(P'P)^{-1}P'(\mu-\mu^K)-P(X_i)(P'P)^{-1}P'\eta \\
\phi_i-\widehat{\phi}_i &= (\phi_i-\phi_i^K)-P(X_i)(P'P)^{-1}P'(\phi-\phi^K)-P(X_i)(P'P)^{-1}P'\varepsilon
\end{aligned}
$$

$$\begin{aligned}
A1 &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\nu_i-\widehat{\nu}_i) \\
&= \frac{1}{n}\varepsilon'(\nu-\nu^K)-\frac{1}{n}\varepsilon'P(P'P)^{-1}P'(\nu-\nu^K)-\frac{1}{n}\varepsilon'P(P'P)^{-1}P'\zeta \\
&= \frac{1}{n}\varepsilon'(\nu-\nu^K)-\left(Q_n^{-1/2}\frac{P'\varepsilon}{n}\right)'\left(Q_n^{-1/2}\frac{P'(\nu-\nu^K)}{n}\right)-\left(Q_n^{-1/2}\frac{P'\varepsilon}{n}\right)'\left(Q_n^{-1/2}\frac{P'\zeta}{n}\right) \\
&= O_P\big(K^{-\alpha_\nu}/\sqrt{n}\big)+O_p(\sqrt{K/n}\sqrt{K/n}K^{-\alpha_\nu})+O_p(K/n)\big) \\
&= O_P\big(K^{-\alpha_\nu}/\sqrt{n}+K/n\big) \qquad (1.A.14)
\end{aligned}$$

The third equality follows from Lemma 1.A.3, 1.A.4, and 1.A.5. Similarly,

$$A2 = \frac{1}{n}\sum_{i=1}^{n}\zeta_i(\phi_i-\widehat{\phi}_i)=O_P\big(K^{-\alpha_\phi}/\sqrt{n}+K/n\big) \qquad (1.A.15)$$

$$\begin{aligned}
A3 &= \frac{1}{n}\sum_{i=1}^{n}\big(\phi_i-\widehat{\phi}_i\big)\big(\nu_i-\widehat{\nu}_i\big) \\
&= (\phi-\phi^K)'(\nu-\nu^K)/n+(\phi-\phi^K)'P(P'P)^{-1}P'(\nu-\nu^K)/n+\varepsilon'P(P'P)^{-1}P'\zeta/n \\
&= O_p\Big(K^{-(\alpha_\phi+\alpha_\nu)}\Big)+O_p\Big(K^{-(\alpha_\phi+\alpha_\nu)}K/n\Big)+O_p\Big(K/n\Big) \qquad (1.A.16) \\
&= O_p\Big(K^{-(\alpha_\phi+\alpha_\nu)}+K/n\Big) \qquad (1.A.17)
\end{aligned}$$

Adding up equation (1.A.14)-(1.A.17), we have,

$$A = O_p\Big(K^{-(\alpha_\phi+\alpha_\nu)}+K^{-\alpha_\phi}/\sqrt{n}+K^{-\alpha_\nu}/\sqrt{n}+K/n\Big) \qquad (1.A.18)$$

Under Assumption 1.5.1(iv), $K\to\infty$ and $K/n\to 0$, the error term $A=o_p(1)$. Thus, the Jacobian term,

$$\frac{1}{n}\sum_{i=1}^{n}\big(X_{-i}-\widehat{\phi}(X_i)\big)\big(Y_{-i}-\widehat{\nu}(X_i)\big)-\frac{1}{n}\sum_{g=1}^{G}E\Big[\varepsilon_g'\zeta_g\Big]\xrightarrow{p}0 \qquad (1.A.19)$$

(ii) Next, consider the score term:

$$
\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(X_{-i}-\widehat{\phi}(X_i)\Big)\Big((Y_i-\widehat{\mu}(X_i))-\beta_0(Y_{-i}-\widehat{\nu}(X_i))\Big) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\varepsilon_i+\phi_i-\widehat{\phi}_i\Big)\Big(U_i+(\mu_i-\widehat{\mu}_i)+\beta_0(\nu_i-\widehat{\nu}_i)\Big) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i U_i + B1 + B1 + (B21+B22) + (B31+B31)
\end{aligned}
$$

By assumption 1.5.1(i), $E\big[|\varepsilon_i\zeta_i|^2\big]<\infty$, then by CLT for grouped data for $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i U_i = \frac{1}{\sqrt{n}}\sum_{g=1}^{G}\varepsilon_g' U_g$,

$$
\Omega_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\varepsilon_g' U_g \xrightarrow{d} N(0,1) \tag{1.A.20}
$$

From Lemma 1.A.3, 1.A.4, and 1.A.5 we know

$$
\begin{aligned}
B1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\phi_i-\widehat{\phi}_i)U_i = O_P\big(K^{-\alpha_\phi}+K/\sqrt{n}\big) \\
B21 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon(\mu_i-\widehat{\mu}_i) = O_P\big(K^{-\alpha_\mu}+K/\sqrt{n}\big) \\
B22 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{\phi}_i-\phi_i)(\widehat{\mu}_i-\mu_i) = O_P\big(\sqrt{n}K^{-(\alpha_\phi+\alpha_\mu)}+K/\sqrt{n}\big) \\
B31 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon(\nu_i-\widehat{\nu}_i) = O_P\big(K^{-\alpha_\nu}+K/\sqrt{n}\big) \\
B32 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{\phi}_i-\phi_i)(\nu_i-\widehat{\nu}_i) = O_P\big(\sqrt{n}K^{-(\alpha_\phi+\alpha_\nu)}+K/\sqrt{n}\big)
\end{aligned} \tag{1.A.21}
$$

Adding up equations in (1.A.21), we have,

$$
B = O_p\Big(\sqrt{n}K^{-(\alpha_\phi+\alpha_\mu)}+\sqrt{n}K^{-(\alpha_\phi+\alpha_\nu)}+K^{-\alpha_\phi}+K^{-\alpha_\mu}+K^{-\alpha_\nu}+K/\sqrt{n}\Big) \tag{1.A.22}
$$

Under Assumption 1.5.1(iv), $K\to\infty$ and $\sqrt{n}K^{-\alpha_\phi-\min\{\alpha_\mu,\alpha_\nu\}}+K^{-\alpha_\phi}+K^{-\alpha_\mu}+K^{-\alpha_\nu}+$

$K/\sqrt{n} \to 0$, the error term $B = o_p(1)$. Thus, the Score term,

$$\Omega_n^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(X_{-i} - \widehat{\phi}(X_i) \Big) \Big((Y_i - \widehat{\mu}(X_i)) - \beta_0 (Y_{-i} - \widehat{\nu}(X_i)) \Big) \xrightarrow{d} N(0,1) \tag{1.A.23}$$

Combining (1.A.19) and (1.A.23) and applying Slutsky theorem, it can be concluded that under Assumptions 1.5.1:

$$\sqrt{n} V_n^{-1/2} (\widehat{\beta}_{db} - \beta_0) \xrightarrow{d} N(0,1).$$

Proof: *Show $\sqrt{n} V_n^{-1/2} (\widehat{\beta}_{db}^{gmm} - \beta_0) \xrightarrow{d} N(0,1)$.*

(i) First we show the root-n consistency and asymptotic normality of the preliminary estimator $\widetilde{\beta}_{db} = \big(\widehat{\zeta}' \widehat{\varepsilon} \widehat{\varepsilon}' \widehat{\zeta}\big)^{-1} \big(\widehat{\zeta}' \widehat{\varepsilon} \widehat{\varepsilon}' \widehat{\eta}\big)$.

$$\sqrt{n}(\widetilde{\beta} - \beta_0) \quad = \quad \left(\Big(\frac{\widehat{\varepsilon}' \widehat{\zeta}}{n}\Big)' \Big(\frac{\widehat{\varepsilon}' \widehat{\zeta}}{n}\Big) \right)^{-1} \left(\Big(\frac{\widehat{\varepsilon}' \widehat{\zeta}}{n}\Big)' \Big(\frac{\widehat{\varepsilon}' \widehat{\eta} - \beta_0 \widehat{\varepsilon}' \widehat{\zeta}}{\sqrt{n}}\Big) \right),$$

From the proof of Theorem 1.5.1 we know, for $M_n = \frac{1}{n} E[\varepsilon' \zeta]$ and $\Omega_n = \frac{1}{n} E[\varepsilon' U U' \varepsilon]$

$$\left\| \frac{\widehat{\varepsilon}' \widehat{\zeta}}{n} - M_n \right\| \xrightarrow{p} 0 \tag{1.A.24}$$

$$\Omega_n \frac{\widehat{\varepsilon}' \widehat{\eta} - \beta_0 \widehat{\varepsilon}' \widehat{\zeta}}{\sqrt{n}} \xrightarrow{d} N(0, I_d) \tag{1.A.25}$$

Let $\widetilde{V}_n = \big(M_n' M_n\big)^{-1} \big(M_n' \Omega_n M_n\big) \big(M_n' M_n\big)^{-1}$. By Slutsky theorem, we have

$$\sqrt{n} \widetilde{V}_n^{-1/2} (\widetilde{\beta} - \beta_0) \to N\,(0, 1)$$

(ii) Next we show the consistency of the estimator for inverse weight matrix Ω_n,

$$\widehat{\Omega}_{db} = \frac{1}{n} \widehat{\varepsilon}' \widetilde{U} \widetilde{U}' \widehat{\varepsilon} \tag{1.A.26}$$

where $\widetilde{U} = (Y - \widehat{\mu}) - \widetilde{\beta}_{db}(Y_{-} - \widehat{\nu})$*. Then,*

$$\widehat{\varepsilon}'\widetilde{U} = (\varepsilon - (\widehat{\phi} - \phi_0))'(U + \widetilde{U} - U) = \varepsilon' U - (\widehat{\phi} - \phi_0)'\widetilde{U} + \varepsilon'(\widetilde{U} - U) \tag{1.A.27}$$

By ordinary LLN,

$$\left\| \frac{1}{n}\varepsilon' U U' \varepsilon - \Omega_n \right\| \xrightarrow{p} 0, \tag{1.A.28}$$

We need to show all the other terms in $\widehat{\Omega}_{db}$ *equals to* $o_p(1)$*. Let* $\widehat{U} = (Y - \widehat{\mu}) - \beta_0(Y_{-} - \widehat{\nu})$*. Then,*

$$\frac{1}{\sqrt{n}}(\widehat{\phi} - \phi_0)'\widetilde{U} = \frac{1}{\sqrt{n}}(\widehat{\phi} - \phi_0)'\widehat{U} + \frac{1}{\sqrt{n}}(\widehat{\phi} - \phi_0)'(\widetilde{U} - \widehat{U}) \tag{1.A.29}$$

$$\frac{1}{\sqrt{n}}\varepsilon'(\widetilde{U} - U) = \frac{1}{\sqrt{n}}\varepsilon'(\widehat{U} - U) + \frac{1}{\sqrt{n}}\varepsilon'(\widetilde{U} - \widehat{U}) \tag{1.A.30}$$

From the proof in Theorem 1.5.1, we know under the assumption for Theorem 1.5.1,

$$\frac{1}{\sqrt{n}}(\widehat{\phi} - \phi_0)'\widehat{U} = o_p(1); \quad \frac{1}{\sqrt{n}}\varepsilon'(\widehat{U} - U) = o_p(1) \tag{1.A.31}$$

Consider $\frac{1}{\sqrt{n}}(\widehat{\phi} - \phi_0)'(\widetilde{U} - \widehat{U})$*. Since* $\widetilde{U} - \widehat{U} = (\widetilde{\beta}_{db} - \beta_0)\widehat{\eta}$*, then,*

$$\frac{1}{\sqrt{n}}(\widehat{\phi} - \phi_0)'(\widetilde{U} - \widehat{U}) = \sqrt{n}(\widetilde{\beta}_{db} - \beta_0)\frac{1}{n}(\widehat{\phi} - \phi_0)'\widehat{\eta}$$

where $\sqrt{n}(\widetilde{\beta}_{db} - \beta_0) = O_P(1)$ *is shown in the proof of part (i), and* $\frac{1}{n}(\widehat{\phi} - \phi_0)'\widehat{\eta} = o_p(1)$ *is showed in the proof of theorem 1.5.1. Combining these results, we have*

$$\frac{1}{\sqrt{n}}(\widehat{\phi} - \phi_0)'(\widetilde{U} - \widehat{U}) = o_p(1). \tag{1.A.32}$$

Next consider $\frac{1}{\sqrt{n}}\varepsilon'(\widetilde{U} - \widehat{U})$*,*

$$\frac{1}{\sqrt{n}}\varepsilon'(\widetilde{U} - \widehat{U}) = \sqrt{n}(\widetilde{\beta}_{db} - \beta_0)\frac{1}{n}\varepsilon'\widehat{\eta}$$

As showed in the previous proofs, $\sqrt{n}(\widetilde{\beta}_{db} - \beta_0) = O_p(1)$ *and* $\frac{1}{n}\varepsilon'\widehat{\eta} = o_p(1)$ *under the assump-*

tions for Theorem 1.5.1. Then

$$\frac{1}{\sqrt{n}}\varepsilon'(\widetilde{U}-\widehat{U}) = o_p(1) \tag{1.A.33}$$

Combing the above equations, we have

$$\left\|\widehat{\Omega}_{db}-\Omega_n\right\| \xrightarrow{p} 0 \tag{1.A.34}$$

(iii) Finally, we can show that the semiparametric GMM estimator

$$\widehat{\beta}_{db}^{gmm} = \left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\Omega}_{db}^{-1}\widehat{\varepsilon}'\widehat{\zeta}\right)^{-1}\left(\widehat{\zeta}'\widehat{\varepsilon}\widehat{\Omega}_{db}^{-1}\widehat{\varepsilon}'\widehat{\eta}\right) \tag{1.A.35}$$

is also root-n consistent and asymptotically distributed.
The proof follows directly from part (i) and (ii). And the asymptotic variance for $\widehat{\beta}_{db}^{gmm}$ is

$$V_n = \left(M_n'\Omega_n M_n\right)^{-1}\left(M_n'\Omega_n\Omega_n^{-1}\Omega_n M_n\right)\left(M_n'\Omega_n M_n\right)^{-1} = \left(M_n'\Omega_n M_n\right)^{-1} \tag{1.A.36}$$

Thus we have

$$\sqrt{n}V_n^{-1/2}(\widehat{\beta}_{db}^{gmm}-\beta_0) \to N\left(0,1\right)$$

□

Proof for Theorem 1.5.2

Proof: *The proof of Theorem 1.5.2 follows directly from the proof of Theorem 1.5.1 $\sqrt{n}V^{-1/2}(\widehat{\beta}_{db}^{gmm}-\beta_0) \xrightarrow{d} N(0,1)$ part (ii).* □

1.A.4 Proofs in Section 1.6

Proof for Theorem 1.6.1

Proof: *Define the group variable notations:*

$$Z_g = \big(Z_{g,1}, ...Z_{g,n_g}\big)\,; \quad \overline{Z}_{g-} = \big(\overline{Z}_{g,-1}, ...\overline{Z}_{g,-n_g}\big)\,; \quad f(X_g) = \big(f(X_{g,1}), ...X_{g,n_g}\big)$$

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_{cfdb} - \beta_0\right) \quad &= \quad \frac{1}{n}\Big(\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\big(\overline{X}_{g,-} - \check{\phi}_\ell(X_g)\big)'\big(\overline{Y}_{g,-} - \widetilde{\nu}_\ell(X_g)\big)\Big)^{-1} \\ &\quad\quad \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\Big(\overline{X}_{g,-} - \check{\phi}_\ell(X_g)\Big)'\Big(Y_g - \widetilde{\mu}_\ell(X_g) - \beta_0\big(\overline{Y}_{g,-} - \widetilde{\nu}_\ell(X_g)\big)\Big)\end{aligned}$$

(i) Show $\sqrt{n}V_n^{-1/2}(\widehat{\beta}_{cfdb} - \beta_0) \xrightarrow{d} N(0,1)$.

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\Big(\overline{X}_{g,-} - \check{\phi}_\ell(X_g)\Big)'\Big(Y_g - \widetilde{\mu}_\ell(X_g) - \beta_0\big(\overline{Y}_{g,-} - \widetilde{\nu}_\ell(X_g)\big)\Big) \\ = \quad &\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\varepsilon_g' U_g + \big(D_1 + D_{21} + D_{22} + D_{31} + D_{32}\big) \qquad (1.A.37)\end{aligned}$$

By assumption 1.5.1(i), $E\big[|\varepsilon_g U_g|^2\big] < \infty$, *then by CLT for grouped data for* $\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\varepsilon_g U_g$

$$\Omega_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\varepsilon_g' U_g \xrightarrow{d} N(0,1) \qquad (1.A.38)$$

Consider the remainder terms $D = D_1 = D_{21} + D_{22} + D_{31} + D_{32}$.

$$D_1 \quad = \quad \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\big(\check{\phi}_\ell(X_g) - \phi_0(X_g)\big)'U_g \qquad (1.A.39)$$

where $\check{\phi}_\ell(X_g) = P(X_g)\Big(\frac{1}{\check{n}_\ell}\sum_{g\in\check{I}_\ell}P(X_g)P(X_g)'\Big)^{-1}\Big(\frac{1}{\check{n}_\ell}\sum_{g\in\check{I}_\ell}P(X_g)\overline{X}_{g,-}\Big)$
Let $D_1(I_\ell) = \frac{1}{\sqrt{n}}\sum_{g\in I_\ell}D_1(I_\ell, g)$ *where* $D_1(I_\ell, g) = \big(\check{\phi}_\ell(X_g) - \phi_0(X_g)\big)'U_g$.

Denote $Z^c_{I_\ell} := \{Z_g : g \notin I_\ell\}$*. Then,*

$$
\begin{aligned}
&E\left[D_1(I_{\ell,g})\middle|Z^c_{I_\ell}\right] = 0, \quad \textit{for } g \in I_\ell;\\
&E\left[D_1(I_{\ell,g})'D_1(I_{\ell,g'})\middle|Z^c_{I_\ell}\right] = 0, \quad \textit{for } g, g' \in I_\ell;\\
&E\left[D_1(I_{\ell,g})'D_1(I_{\ell,g}) \;\middle|\; Z^c_{I_\ell}\right] = O_p(K^{-2\alpha_\phi} + K/n) \quad \textit{for } g \in I_\ell
\end{aligned}
$$

Then we have,

$$
E[D_1(I_\ell)^2|Z^c_{I_\ell}] \;\; = \;\; E\left[\Big(\frac{1}{\sqrt{n}}\sum_{g\in I_\ell}\big(\check{\phi}_\ell(X_g) - \phi_0(X_g)\big)'U_g\Big)^2 \;\Big|\; Z^c_{I_\ell}\right] = O_p(K^{-2\alpha_\phi} + K/n)
$$

Therefore

$$
D_1 = \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\big(\check{\phi}_\ell(X_g) - \phi_0(X_g)\big)'U_g = O_P\big(K^{-\alpha_\phi} + \sqrt{K}/\sqrt{n}\big)
$$

Similarly, it can be shown that,

$$
\begin{aligned}
D_{21} &= \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\varepsilon_g'\big(\widetilde{\mu}_\ell(X_g) - \mu_0(X_g)\big) = O_P\big(K^{-\alpha_\mu} + \sqrt{K}/\sqrt{n}\big)\\
D_{22} &= \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\big(\check{\phi}_\ell(X_g) - \phi_0(X_g)\big)'\big(\widetilde{\mu}_\ell(X_g) - \mu_0(X_g)\big)\\
&= O_P\big(\sqrt{n}K^{-(\alpha_\phi+\alpha_\mu)} + \sqrt{K}/\sqrt{n}\big)\\
D_{31} &= \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\varepsilon_g'\big(\widetilde{\nu}_\ell(X_g) - \nu_0(X_g)\big) = O_P\big(K^{-\alpha_\nu} + \sqrt{K}/\sqrt{n}\big)\\
D_{32} &= \frac{1}{\sqrt{n}}\sum_{\ell=1}^{L}\sum_{g\in I_\ell}\big(\check{\phi}_\ell(X_g) - \phi_0(X_g)\big)'\big(\widetilde{\nu}_\ell(X_g) - \nu_0(X_g)\big)\\
&= O_P\big(\sqrt{n}K^{-(\alpha_\phi+\alpha_\nu)} + \sqrt{K}/\sqrt{n}\big)
\end{aligned}
$$

Adding up $D_1, D_{21}, D_{22}, D_{31}$*, and* D_{32}*, it follows that,*

$$
D = O_p\Big(\sqrt{n}K^{-(\alpha_\phi+\alpha_\mu)} + \sqrt{n}K^{-(\alpha_\phi+\alpha_\nu)} + K^{-\alpha_\phi} + K^{-\alpha_\mu} + K^{-\alpha_\nu} + \sqrt{K}/\sqrt{n}\Big) \tag{1.A.40}
$$

Under Assumption $K \to \infty$ and $\sqrt{n}K^{-\alpha_\phi - \min\{\alpha_\mu, \alpha_\nu\}} + K^{-\alpha_\phi} + K^{-\alpha_\mu} + K^{-\alpha_\nu} + \sqrt{K}/\sqrt{n} \to 0$, the error term $B = o_p(1)$. Thus, the Score term,

$$\Omega_n^{-1/2} \frac{1}{\sqrt{n}} \sum_{\ell=1}^{L} \sum_{g \in I_\ell} \Big(\overline{X}_{g,-} - \check{\phi}_\ell(X_g)\Big)' \Big(Y_g - \widetilde{\mu}_\ell(X_g) - \beta_0\big(\overline{Y}_{g,-} - \widetilde{\nu}_\ell(X_g)\big)\Big) \xrightarrow{d} N(0,1) \quad (1.A.41)$$

(ii) We can also show that, under Assumption for theorem 1.5.2, the Jacobian term

$$\frac{1}{n}\Big(\sum_{\ell=1}^{L} \sum_{g \in I_\ell} \big(\overline{X}_{g,-} - \check{\phi}_\ell(X_g)\big)'\big(\overline{Y}_{g,-} - \widetilde{\nu}_\ell(X_g)\big)\Big) - \frac{1}{n}\sum_{g=1}^{G} E\Big[\varepsilon_g' \zeta_g\Big] \xrightarrow{p} 0 \quad (1.A.42)$$

Combining (i) and (ii), it follows that

$$\sqrt{n} V_n^{-1/2} (\widehat{\beta}_{dcf} - \beta_0) \xrightarrow{d} N(0,1). \quad (1.A.43)$$

1.B Related Literature on Identification for Social Interaction Models

Our identification and estimation strategy for the semiparametric social interaction model based on several branches of literature, including the identification for linear social interaction model, semiparametric two-step method and also the machine learning method. In this subsection, we will give a review of the related literature on identification for the social interaction model, the estimation for the semiparametric model and also the machine learning methods.

In the following subsection, we provide a brief review of the recent literature on the identification of the endogenous social interaction effect in the parametric linear-in-means model. In the pioneering work, Manski (1993) proposes the linear-in-means social interaction model in which the individual's outcome is determined according to

$$Y_{g,i} = \beta_0 \overline{Y}_{g,-i} + X_{g,i}\gamma + W_g\delta + \alpha_g + U_{g,i}, \quad (1.B.1)$$

where individual's outcome, $Y_{g,i}$, depends on the mean of all the linked friends' outcomes, $\overline{Y}_{g,-i}$, their own characteristics $X_{g,i}$ and average of group-level characteristics, W_g, the unobserved group characteristics α_g that might drive the self-selection into certain groups, and also the unobserved individual characteristics, $U_{g,i}$. Following the terminology in Manski (1993), β_0 denotes the endogenous peer effect which is the parameter of interest. δ captures the exogenous peer effect (or contextual effect). α_g denotes the correlated effect within groups.

Manski (1993) first points out the Reflection problem in the linear-in means social interaction model which contains two sources of identification issue. The first is social interaction effect (including endogenous and exogenous) can not be identified from the correlated effect driven by unobserved self-selection into certain groups. He further shows even if we rule out the self-selection issue, the endogenous social interaction effect still can not be distinguished from the exogenous social interaction effect.

We start by considering the identification problem by imposing a strong assumption that there is no contextual and correlated effect. Subsequently, we relax the assumption and allow for contextual effect and discuss how to solve the reflection problem. Finally, we consider the identification issue also with a correlated effect.

1.B.1 Endogenous Social Interaction Effect

Consider the linear-in-means group social interaction model with no contextual or correlated effect. Assume all the variables are demeaned. For notation simplicity, we omit the subscribe g in the following discussion.

$$Y_i = \beta_0 Y_{-i} + X_i \gamma + U_i, \;\; E[U_i|X] = 0 \tag{1.B.2}$$

where $-i$ denote the friend's index of individual i. Y_{-i} is the average friend's outcome which is endogenous and β_0 captures the endogenous social interaction effect. By excluding the W_g and α_g, we assume there is no contextual or correlated effect. Assume the model is in equilibrium condition, $\overline{Y}_g = \frac{1}{1-\beta_0}\overline{X}_g\gamma_0 + \overline{U}_g$, which implies the following reduced form

$$Y_{-i} = \frac{\beta_0}{1-\beta_0^2}X_i\gamma + \frac{1}{1-\beta_0^2}X_{-i}\gamma + V_{-i} \tag{1.B.3}$$

where $V_{-i} = \frac{\beta_0}{1-\beta_0^2}U_i + \frac{1}{1-\beta_0^2}U_{-i}$.

The most widely used method to identify β_0 in the literature is the instrumental variables approach, Lee (2003, 2007) Bramoullé, Djebbari, and Fortin (2009) to name a few. Researchers use friend's characteristics (X_{-i}) or leave out average of X to instrument the peer effect Y_{-i}. The validity of the IV X_{-i} comes from the strictly exogenous assumption for the regressors which implies $E[X_{-i}U_i] = 0$. Also, from the reduced form model, X_{-i} contributes to Y_{-i} as long as $\gamma \neq 0$ and X_{-i} is linearly independent with X_i. Then β_0 and γ_0 are identified by the following moment conditions:

$$\begin{aligned} E\left[X_i(Y_i - \beta_0 Y_{-i} - X_i\gamma_0)\right] &= 0 \\ E\left[X_{-i}(Y_i - \beta_0 Y_{-i} - X_i\gamma_0)\right] &= 0 \end{aligned}$$

Based on the identification condition, Kelejian and Prucha (1998) and Lee (2003) use a 2SLS procedure for estimating β_0. Denote the regressors $\widetilde{X}_i = (Y_{-i}, X_i)$ and IV $Z_i = (X_{-i}, X_i)$, then $\vartheta_0 = (\beta_0, \gamma_0)$ can be estimated by

$$\widehat{\vartheta} = \left(\widetilde{X}P_Z\widetilde{X}\right)^{-1}(\widetilde{X}P_Z Y) \tag{1.B.4}$$

where $\widetilde{X} = (\widetilde{X}_1, ..., \widetilde{X}_N)', Y = (Y_1, ..., Y_N)'$, $Z = (Z_1, ..., Z_N)'$ and $P_Z = Z(Z'Z)^{-1}Z'$

Instead of using the IV approach, Angrist (2014) utilizes the between group variation to identify β_0. Consider the reduced form equilibrium model

$$\overline{Y}_g = \frac{\gamma_0}{1-\beta_0}\overline{X}_g\gamma_0 + \frac{1}{1-\beta_0}\overline{U}_g$$

where the term $\frac{\gamma_0}{1-\beta_0}$ reflect the social multiplier of the endogenous peer effect which inflates the effect of individual characteristics by $\frac{1}{1-\beta_0}$. Denote $\phi_1 = \frac{\gamma_0}{1-\beta_0}$, then ϕ_1 can be identified by regressing the group average of outcome, Y_g, over group average of regressors, X_g.

$$\phi_1 = \frac{E[X_g Y_g]}{V[X_g]} \tag{1.B.5}$$

Angrist (2014) also utilizes the OLS estimator of Y_i on X_i which is defined as

$$\phi_0 = \frac{E[X_i Y_i]}{V[X_i]} \tag{1.B.6}$$

Based on the structural form model, $E[X_i Y_i] = \beta_0 E[X_i Y_{-i}] + V(X_i)\gamma_0$, it implies that

$$\phi_0 = \beta_0 \tau^2 \phi_1 + \gamma_0 \tag{1.B.7}$$

where $\tau^2 = \frac{V[X_g]}{V[X_i]}$. Then β_0 can be identified by

$$\beta_0 = \frac{\phi_1 - \phi_0}{\phi_1} \frac{1}{(1-\tau^2)} \tag{1.B.8}$$

And the estimator for β_0 just follows the sample analogy of the identification condition.

Angrist (2014) draws an interesting connection of the endogenous social interaction effect with the difference between 2SLS of Y_i on X_i with group dummies as IV and OLS estimators. He shows that numerically the estimator for ϕ_1 is exactly the same with the 2SLS estimator of Y_i on X_i with group dummies as IV. Due to this property, he also concerns that other sources of difference between 2SLS and OLS estimator such as measurement error of regressors might lead to spurious peer effects.

Another approach to identify β_0 is based on the variance structural of the error terms, Lee (2007); Graham (2008) to name a few. Consider the second moment for Y

$$\begin{aligned} E[Y'Y] &= \frac{1}{(1-\beta_0)^2} E\left[(X\gamma + X_{-}\gamma_0\beta_0)'(X\gamma + X_{-}\gamma_0\beta_0)\right] \\ &+ \frac{1}{(1-\beta_0)^2} E\left[(U + \beta_0 U_{-})'(U + \beta_0 U_{-})'\right] \end{aligned} \tag{1.B.9}$$

The covariance matrix of Y and X can be identified from the data. If we assume U_i's are i.i.d., $E[UU'] = \sigma_u^2 I$ (Lee (2007), then

$$E\left[(U + \beta_0 U_-)'(U + \beta_0 U_-)'\right] = (1 + \beta_0^2)\sigma_U^2. \tag{1.B.10}$$

Combining (1.B.9) and (1.B.10) can be used as the moment conditions to identify and estimate β_0. If we relax the i.i.d assumption for U_i and allow for within group correlations but still assume (U_i, U_{-i}) are i.i.d distributed across groups. Then $E[UU']$ becomes the block diagonal matrix with each block equals to

$$B = \begin{bmatrix} E\left[U_i^2\right] & E\left[U_i U_{-i}\right] \\ E\left[U_i U_{-i}\right] & E\left[U_{-i}^2\right] \end{bmatrix} \tag{1.B.11}$$

Equation (1.B.9), combined with (1.B.11), can be used as the additional moment condition to identify and estimate structural parameters β_0.

The variance approach for identifying β_0 does not need to find the IVs for Y_{-i} which requires some exclusive restrictions. However, it also comes with the cost that it requires the restrictions on the variance structural of error terms. In our study for the identification of the semiparametric social interaction models, we consider using the IV approach.

1.B.2 Reflection Problem

In our setup, we impose restrictions on the contextual and exclude the correlated effect, such that the reflection problem can be solved. However, it is still important to consider how to deal with these issues. In this section, we give a brief review of how to solve the reflection problem in the recent literature for the linear social interaction models.

First consider the identification issue of distinguishing endogenous and exogenous social interaction effect, assume no unobserved group characteristics that drive the self-selection into certain groups in the model. Brock and Durlauf (2001) relax the linear-in-means set up in Manski (1993) and consider the average linked friends' outcome affects one's outcome through a nonlinear function which is known. To be specific, they include $\phi(\overline{Y}_{g,-i})$ as the endogenous social interaction term and show that structural parameters are identified if $\frac{\partial^2\phi(\overline{Y}_{g,-i})}{\partial\overline{Y}_{g,-i}} \neq 0$. Graham (2008) propose another identification approach by making use of the covariance structure of the error terms. Bramoullé, Djebbari, and Fortin (2009) discussed the identification results of social interactions in a network framework. And they impose assumptions on the network structure D, which requires the identity matrix I, D and D^2 are linearly independent. If the assumption holds, then friends' friends' outcomes can be used to identify the endogenous social interaction effect. This condition rules out the group interaction case by Manski (1993) which induces the reflection problem.

For the identification issue with an unobserved group effect, α_g, one needs to provide extra information to distinguish the social interaction effect from the correlated effect. Graham and Hahn (2005) restrict the structural model by excluding the exogenous peer effect first. They reinterpret the linear-in-means models as a quasi-panel model because the reduced form model is exactly a panel with a time-invariant regressor. They use the generalization of the instrumental variables strategy of Hausman and Taylor (1981) to identify the time-invariant coefficient and then showed that the structural parameter of endogenous social interactions can also be identified. Goldsmith-Pinkham and Imbens (2013); Hsieh and Lee (2016); Qu and Lee (2015); Johnsson and Moon (2015), and Auerbach (2016) also discuss this identification issue by providing extra sources of information.

This paper focus on the identification of endogenous social interaction effects with our semiparametric model. In model (1.3.1), we also consider the endogenous peer effect is linear-in-means where the average friends' outcome affects the individual's output linearly. We first assume there is no unobserved group effect that drives self-selection into certain groups in our model,i.e. $\alpha_g = 0$, which is guaranteed by the strong exogeneity assumption.

But unlike the existing literature, we introduced a more flexible non-parametric part to capture the effect of an individual's characteristics. Thus we have a semiparametric model and we still focus on the identification of the endogenous social interaction effect which is the parametric part.

The identification issue such as the reflection problem is widely discussed in the literature. These issues are important because if we mixed the endogenous social interaction effect with the contextual effect or correlated effect, then we might obtain a spurious social interaction effect. However, another source of spurious effect might come from the misspecification in the parametric model setup. Thus in this paper, we focus on the semiparametric model setup. We discuss the identification condition in detail in Section 1.3.

For other identification issues, such as correlated effect, the identification results for the linear social interaction model can also be applied in our semiparametric model.

1.C Simulation Results

Table 1.1: Simulation Results for Uncorrelated Bi-Normal $\rho = 0$

$h_0(x)$	Est.	Bias	Vars	MSE	MAE	Size5	Size10	F-stat
x	Linear	0.0296	0.2691	0.2700	0.4144	0.0660	0.1210	210.8683
	Series	0.0291	0.2780	0.2789	0.4210	0.0720	0.1280	211.0296
$exp(x)$	Linear	0.0577	0.9433	0.9466	0.7709	0.0770	0.1270	64.7422
	Series	0.0727	0.8627	0.8680	0.7399	0.0760	0.1330	65.8090
$sin(x) + cos(x)$	Linear	0.1153	1.1738	1.1871	0.8648	0.0590	0.1090	51.4334
	Series	0.0728	0.8599	0.8652	0.7374	0.0720	0.1200	54.2721

*Bias $= \frac{1}{S}\sum_{s=1}^{S}\widehat{\beta}_s - \beta_0$; VAR $= \frac{1}{S}\sum_{s=1}^{S}\left(*\widehat{\beta}_s - \overline{\widehat{\beta}}_s\right)^2$

*MSE $= \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\beta}_s - \beta_0\right)^2 = \text{Bias}^2 + \text{VAR}$; *MAE $= \frac{1}{S}\sum_{s=1}^{S}|\widehat{\beta}_s - \beta_0|$

*Size5/10 $= \sum_{s=1}^{S}\mathbf{1}\left(|t| > z_\alpha\right)/S$. Two-sided empirical size with nominal size 5% and 10%;

*F-stat: F-stats of first stage regression in semiparametric two-step estimator of β_0:

*Linear: regression with X as explanatory variables.

*Series: regression polynomial expansion of order 3 of X as explanatory variables.

*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 1.2: Simulation Results for Correlated Bi-Normal $\rho = 0.5$

$h_0(x)$	Est.	Bias	Vars	MSE	MAE	Size5	Size10	F-stat
x	Linear	0.0231	0.2714	0.2719	0.4184	0.0680	0.1140	158.5402
	Series	0.0199	0.2891	0.2895	0.4322	0.0670	0.1200	158.6238
$exp(x)$	Linear	0.0302	1.0484	1.0493	0.8184	0.0700	0.1250	47.8174
	Series	0.0462	0.9021	0.9042	0.7628	0.0660	0.1260	49.3396
$sin(x) + cos(x)$	Linear	0.1179	1.6410	1.6549	1.0308	0.0490	0.0940	37.3101
	Series	0.0397	0.9066	0.9082	0.7611	0.0650	0.1280	42.0433

*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 1.3: Simulation Results for Correlated Bi-Logistic $r = 0.5$

$h_0(x)$	Est.	Bias	Vars	MSE	MAE	Size5	Size10	F-stat
x	Linear	0.0067	0.1394	0.1394	0.3003	0.0530	0.1000	531.5796
	Series	0.0050	0.1451	0.1452	0.3045	0.0540	0.0970	526.6395
$exp(x)$	Linear	1.7188	1.6513	4.6057	1.7744	0.4200	0.5300	62.2892
	Series	-0.0714	0.3733	0.3784	0.4855	0.0900	0.1700	117.9166
$sin(x) + cos(x)$	Linear	-2.7664	8.7366	16.3897	3.2982	0.1710	0.2630	10.0896
	Series	-0.2421	2.3107	2.3693	1.1986	0.0560	0.1110	25.8162

*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 1.4: Simulation Results for Uncorrelated but Dependent Distribution

$h_0(x)$	Est.	Bias	Vars	MSE	MAE	Size5	Size10	F-stat
x	Linear	-1.1763	1.9181	3.3019	1.4422	0.0920	0.1810	51.7247
	Series	0.0224	0.8122	0.8127	0.7113	0.0590	0.1190	91.2694
$exp(x)$	Linear	0.2985	1.0281	1.1172	0.8601	0.0750	0.1410	73.1534
	Series	0.0232	1.6581	1.6586	1.0110	0.0580	0.1110	49.1132
$sin(x) + cos(x)$	Linear	-1.1763	1.9181	3.3019	1.4422	0.0920	0.1810	51.7247
	Series	0.0224	0.8122	0.8127	0.7113	0.0590	0.1190	91.2694

*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

CHAPTER 2

Machine Learning Estimation of Semiparametric Social Interaction Models

2.1 Introduction

The social interaction model is widely discussed in the empirical research. A great many studies have investigated the existence and measure of the social interaction effect, for instance, the peer effect in education (Sacerdote, 2001; Duflo et al., 2011), criminal activity (Glaeser et al., 1996), welfare participation (Bertrand et al., 2000), and job search (Marmaros and Sacerdote, 2002). The commonly used approach in the above-mentioned literature relies on the parametric assumptions and functional form restrictions to identify the endogenous social interaction effect.

However, a fully parametric model is often too restrictive to capture the structure of the data. Estimators based on this approach can be severely biased if the parametric assumptions are violated (Goldberger, 1983). That is, the parametric model set-up could lead to a spurious or misleading endogenous social interaction effect. Consequence, it is desirable to consider a robust semiparametric model to control properly for confounding variables.

Motivated by the above consideration, this paper relaxes the parametric assumption and considers a more flexible semiparametric model of social interactions. This paper assumes the endogenous social interaction effect to be parametric and also linear-in-means where the average peer group's outcome affects the individual's outcome linearly. But unlike the existing literature, this paper introduces a more flexible non-parametric function to capture the individual-specific effect and also contextual effect. This paper focuses on the semiparametric feature of the model and imposes certain restrictions to avoid the complexity of identification issues. To be specific, I assume there is no correlated effect or self-selection that drives individual into certain groups which are guaranteed by the strong exogeneity assumption. This paper also assumes the contextual effect is restrictive, i.e. not all the exogenous individual variables appear as the contextual effect which can be taken as imposing exclusion restrictions.

The identification strategy for the endogenous social interaction effect in this semiparametric model is discussed in Section 1.3 of Chapter 1. The main idea is to partial out the nuisance nonparametric function in the first step and then adopts a semiparametric IV approach to identify the parametric endogenous social interaction effect. The moment condition and orthogonal moment conditions which depend on the first step conditional expectations are then constructed in Section 1.4 of Chapter 1.

Based on the moment and orthogonal moment conditions, this thesis proposes a semiparametric two-step procedure to estimate the endogenous social interaction effect. In the first step, conditional expectations can be estimated using any nonparametric method. In the second step, the endogenous social interaction effect is then estimated using a semiparametric MM/GMM by plugging the first step nonparametric estimators into the moment or orthogonal moment conditions.

In Chapter 1, the first step conditional expectations are estimated using the traditional nonparametric methods, such as series estimator. Chapter 1 also provides the primitive conditions for the consistency and asymptotic normality of the semiparametric two-step estimators for endogenous social interaction effect with first step series. The results show that the application of series estimators are limited to low-dimensional settings and relatively smooth nonparametric functions, which is the general limitations for the traditional nonparametric methods.

To overcome the limitations of the traditional nonparametric methods, this chapter utilizes the more recent Machine Learning methods in the first step conditional expectations estimation, such as LASSO (Tibshirani, 1996), Random Forest (Breiman, 2001) and Neural

Nets (Friedman and Stuetzle, 1981). The Machine Learning methods may be more flexible and able to handle severe nonlinearity functions, higher-order interactions and more covariates in the nonparametric functions.

The semiparametric two-step estimator with first step LASSO-type methods are widely studied in the recent literature (Belloni, Chen, Chernozhukov, and Hansen, 2012; Belloni, Chernozhukov, and Hansen, 2014; Zhang and Zhang, 2014). There are also some work on applying other Machine Learning methods (Chernozhukov, Hansen, and Spindler, 2015; Athey, Imbens, and Wager, 2016; Athey and Imbens, 2015). Asymptotic properties of semiparametric two-step estimators by plugging the first step LASSO or Random Forest have been shown in Belloni, Chen, Chernozhukov, and Hansen (2012) and Athey, Imbens, and Wager (2016). However, there are no theoretical results that compare the performance across different Machine Learning methods.

This paper uses a Monte Carlo simulation approach to investigate the performance of semiparametric estimators for endogenous social interaction effect across various first step nonparametric methods. The Monte Carlo simulation results suggest that no estimator outperforms the others across the data generating processes (DGPs) considered. However, it is reflected in the simulation results that the debiased estimators using first step post-LASSO or Neural Nets methods are more reliable and performs relatively well across the settings considered. For this reason, these two debiased estimators are recommended for use in empirical studies.

To illustrate the semiparametric two step estimator with first step Machine Learning methods, this paper considers an empirical example which investigates the endogenous classmates' peer effect on student's performance. The data used in this study are from the China Education Panel Survey (CEPS), which provides large-scale, nationally representative, longitudinal survey datasets. The results show that the classmates' peer effect is insignificant for student's cognitive test scores across all the first step nonparametric estimators applied. For student's level of self confidence, the results are not consistent with various first-step estimators applied. A significant endogenous classmates' peer effect is obtained if Random Forest or Neural Nets are utilized in the first step.

The remainder of this chapter is organized as follows. In Section 2.2, I briefly restate the framework, including the semiparametric model of social interactions, identification results and estimation procedure. A review of the related literature on Machine Learning methods for conditional expectations are given in Section 2.3. Section 2.4 utilizes the identification results to construct the semiparametric two-step estimators for the endogenous social interaction effect. Monte Carlo simulations are conducted to investigate the finite sample performance of different estimation methods in Section 2.5. Section 2.6 illustrates the estimation procedure by investing the classmates' peer effect using the CEPS data. Section 2.7 concludes and discusses the directions for future research.

2.2 The Framework

In this section, I will briefly restate the semiparametric model of social interactions, identification results and estimation procedure.

2.2.1 The Semiparametric Social Interaction Model

This subsection reviews the semiparametric social interaction models proposed in Chapter 1. Assume that the outcome variable of individual i in group g, $Y_{g,i}$ $(g = 1, ..., G,\ i = 1, ..., n_g)$, is determined according to the following semiparametric model of social interactions:

$$Y_{g,i} = \beta_0 \overline{Y}_{g,-i} + h_0(X_{g,i}, W_{g,i}, \overline{W}_{g,-i}, \Upsilon_g) + U_{g,i}; \quad (g = 1, ..., G;\ i = 1, ..., n_g), \tag{2.2.1}$$

where $Y_{g,i}$ is the outcome variable of interest for individual i in group g. $\overline{Y}_{g,-i} = \frac{1}{n_g-1} \sum_{j \in I_g, j \neq i} Y_j$ denotes the leave-i-out average outcome within the gth group. $X_{g,i}$ denotes the d dimensional ($\dim(X_{g,i}) = d$) individual-specific characteristics that only affect the outcome $Y_{g,i}$ through individual level, which means the leaving-i-out group average $\overline{X}_{g,-i}$ does not affect the outcome directly. $W_{g,i}$ denotes the d_W dimensional ($\dim(W_{g,i}) = d_W$) individual characteristics which also induce the contextual effect. That is, the leaving-i-out group average for $W_{g,i}$, $\overline{W}_{g,-i}$, is allowed to affect the outcome $Y_{g,i}$. Υ_g denotes the d_Υ dimensional

($\dim(\Upsilon_g) = d_\Upsilon$) observed group characteristics. The last term in model (2.2.1), $U_{g,i}$, is the disturbance term that is unobserved to econometricians. This paper will work with the exogeneity condition, which requires $U_{g,i}$ to be independent of all the controls.

Let $\vartheta_0 = (\beta_0, h_0(\cdot))'$ be the true parameter vector. The parameter of interest β_0 is the 1-dimensional parametric part which captures the endogenous social interaction effect. To have a stable equilibrium social interaction model, it is reasonable to require that $|\beta_0| < 1$. The unknown function $h_0(\cdot)$ is the nonparametric part of the model which captures the individual and contextual effects. This nonparametric function is also a nuisance parameter in this paper.

The goal of this paper is to obtain a estimator for β_0, in the presence of the possibly highly complex nuisance function $h_0(\cdot)$ and also the endogenous effect of group social interactions. To handle severe nonlinearity functions, higher-order interactions and more covariates in the nonparametric functions $h_0(\cdot)$, Machine Learning methods are applied to concentrated out $h_0(\cdot)$, such as LASSO, Random Forest, and Neural Nets. And then the parameter of interest β_0 can be estimated using the MM/GMM approach with properly choose instrumental variables for the endogenous social interaction effect.

2.2.2 Identification and Moment Conditions

Here I restate the identification results for the semiparametric social interaction model (2.2.1).

Let $\mathcal{X}_{g,i} = (X_{g,i}, W_{g,i}, \overline{W}_{g,-i}, \Upsilon_g, n_g)$ denote the collection of control variables at both the individual and group levels. The identification strategy is to partial out the nonparametric nuisance function $h_0(\cdot)$ by subtracting the conditional expectations on both side of model (2.2.1) in the first step,

$$Y_{g,i} - E[Y_{g,i}|\mathcal{X}_{g,i}, n_g] = \beta_0\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|\mathcal{X}_{g,i}, n_g]\big) + U_{g,i}. \tag{2.2.2}$$

In the second step, an IV approach is applied to address the endogeneity issue caused by simultaneous equations on the residulized model to identify β_0. Throughout this paper, I consider to use $\overline{X}_{g,-i}$ as the IV for the endogenous variable $\overline{Y}_{g,-i}$. Then, the parameter of interest, β_0, is identified by,

$$E\left[\overline{X}_{g,-i}\big(Y_{g,i} - E[Y_{g,i}|\mathcal{X}_{g,i}]\big)\right] = \beta_0 E\left[\overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - E[\overline{Y}_{g,-i}|\mathcal{X}_{g,i}]\big)\right]. \tag{2.2.3}$$

For notation simplicity, in the subsequent sections, I just use the notation $X_{g,i}$ instead of $\mathcal{X}_{g,i}$. The results for the semiparametric estimation for β_0 will not be affected.

Let $\mu_0(X_{g,i}) \triangleq E[Y_{g,i}|X_{g,i}]$; $\nu_0(X_{g,i}) \triangleq E[\overline{Y}_{g,-i}|X_{g,i}]$; $\phi_0(X_{g,i}) \triangleq E[\overline{X}_{g,-i}|X_{g,i}]$ denote the conditional expectations of $Y_{g,i}, \overline{Y}_{g,-i}, \overline{X}_{g,-i}$ on $X_{g,i}$, respectively. The identification condition for β_0 can be restated using the following moment condition:

$$E\Big[\overline{X}_{g,-i}\Big(\big(Y_{g,i} - \mu_0(X_{g,i})\big) - \beta_0\big(\overline{Y}_{g,-i} - \nu_0(X_{g,i})\big)\Big)\Big] = 0, \tag{2.2.4}$$

The parameter of interest, β_0, then can be estimated using (2.2.4) by plugging in the first step nonparametric estimators of μ_0 and ν_0. However, the moment function of (2.2.4),

$$m(Z_{g,i};\beta,\mu,\nu) = \overline{X}_{g,-i}\Big(\big(Y_{g,i} - \mu(X_{g,i})\big) - \beta\overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - \nu(X_{g,i})\big)\Big), \tag{2.2.5}$$

is not orthogonal to the first step nonparametric parameters, which might lead to severe bias for the semiparametric estimation of β_0 in the second step (Newey, 1994; Chernozhukov et al., 2018b).

The robust strategy for estimating β_0 is to use orthogonal moment condition instead. Following the strategy in Newey (1994); Chernozhukov et al. (2018a), the orthogonal moment condition can be constructed by adding an adjustment term,

$$E\Big[\Big(\overline{X}_{g,-i} - \phi_0(X_{g,i})\Big)\Big((Y_{g,i} - \mu_0(X_{g,i})) - \beta_0\big(\overline{Y}_{g,-i} - \nu_0(X_{g,i})\big)\Big)\Big] = 0. \tag{2.2.6}$$

An additional nuisance function $\phi_0(X_{g,i}) = E[\overline{X}_{g,-i}|X_{g,i}]$ is introduced. Let $\psi(\cdot)$ denote the orthogonal moment function in (2.2.6),

$$\psi(Z_{g,i},\beta,\mu,\nu,\phi)=\Big(\overline{X}_{g,-i}-\phi(X_{g,i})\Big)\Big(\big(Y_{g,i}-\mu(X_{g,i})\big)-\beta\big(\overline{Y}_{g,-i}-\nu(X_{g,i})\big)\Big). \tag{2.2.7}$$

It can be verified that the moment function $\psi(\cdot)$ is not only locally robust but also doubly robust to the first step nonparametric estimators.

In the following discussion, I will call the estimator for β_0 based on moment function (2.2.5) as the plug-in (PI) estimator, and the estimator based on the orthogonal moment function (2.2.7) as the debiasing (DB) estimator. For a detailed discussion about the identification results and moment conditions of the semiparametric social interaction model, please refer to Section 1.3 and 1.4 in Chapter 1.

2.2.3 Estimation Procedure

The goal of this paper is to obtain the estimator for β_0, in the presence of the possibly highly complex nuisance function $h_0(.)$ and also the endogenous effect of group social interactions. Based on the (orthogonal) moment conditions, I propose a semiparametric two-step estimation procedure for the endogenous social interaction effect β_0 for model (2.2.1).

The first step regression require the estimation of conditional expectations,

$$\mu_0(X_{g,i})=E[Y_{g,i}|X_{g,i}],\quad \nu_0(X_{g,i})=E[\overline{Y}_{g,-i}|X_{g,i}] \text{ and } \phi_0(X_{g,i})=E[\overline{X}_{g,-i}|X_{g,i}],$$

which are all nonparametric functions. This paper considers to apply the Machine Learning methods, such as LASSO, Random Forest and Neural Nets, in the first step estimation. The Machine Learning methods are widely used for estimating the conditional expectations and are able to handle highly complex function forms. Section 2.3 will give a brief review of the widely used Machine Learning methods for the estimation of conditional expectations and also discuss the application of these methods for our social interaction model.

In the second step, the parameter of interest,β_0, is obtained using semiparametric MM or GMM estimation by plugging in the first step Machine Learning estimators $\widehat{\mu},\widehat{\nu},\widehat{\phi}$ into moment condition (2.2.4) or orthogonal moment condition (2.2.6). Due to the regularization bias of the first step Machine Learning methods, the plug-in estimator for β_0 based on moment condition (2.2.4) could be severely biased and cannot obtain the root-n consistency. Following Chernozhukov et al. (2018a), this paper make use the idea of orthogonal moment condition (2.2.6) to remove the regularization bias. Section 2.4 will discuss the semiparametric MM/GMM estimation for β_0 in detail.

2.3 Machine Learning Methods for Conditional Expectations

This section will briefly review the widely used Machine Learning methods for the estimation of conditional expectations and discuss the application of these methods for our social interaction model. And based on our problem, the review will focus on the supervised Machine Learning methods for regression models, including (1) regularized linear regression models; (2) regularized basis function models; (3) regression trees and Random Forest; and (4) Neural Nets.

In general, consider estimating the conditional expectation of the outcome variable Y_i given the covariates or features X_i.

$$g(X_i) = E[Y_i|X_i],$$

where $g(\cdot)$ denotes the unknown function that is aimed to estimate. Y_i is a 1-dimensional outcome variable for individual i, and $X_i = (X_{i1}, \cdots, X_{ip})$ denote the p-dimensional vector of the covariates or features, $(i = 1, \ldots, n)$.

Consider the conditional mean $g(X) = E[Y|X]$ is a linear function

$$g(X) = X'\gamma = \sum_{j=1}^{p} X_{\cdot j}\gamma_j$$

where $X_{\cdot j} = (X_{1j}, \cdots, X_{nj})$ denotes the jth covariates, $j = 1, \cdots, p$. I omit the first subscript $(i = 1, \cdots, n)$ for notation simplicity. γ_j is the corresponding coefficient for covariates $X_{\cdot j}$.

Under high dimensional setup, the number of covariates p could be large relative to the number of observations n. Then the traditional linear regression estimation method Least Square $\widehat{\gamma}_{ls}$ has poor properties or even does not work for $(p > n)$.

$$\widehat{\gamma}_{\mathrm{ls}} = \arg\min_{\gamma} \sum_{i=1}^{N} \left(Y_i - X_i'\gamma\right)^2 . \tag{2.3.1}$$

The idea of the Machine Learning methods is to shrink the LS estimator $\widehat{\gamma}_{\mathrm{ls}}$ towards zero using regularization. The usual way of regularization is by adding a penalty term in the criterion function (2.3.1).

2.3.1 LASSO

Assume we have a high dimensional sparse linear model, that is the number of regressors, p, can be larger or even much large than the sample size n. But only a small number $s < n$ of the regressors are of substantial importance for carving the conditional expectation. (Belloni and Chernozhukov, 2011).

$$s = \left\|\gamma_0\right\|_0 := |\{j : \gamma_{0,j} \neq 0\}| \ll n.$$

The classic AIC/BIC estimator (Akaike, 1974; Schwarz, 1978) solves the following oracle problem:

$$\widehat{\gamma}_{\mathrm{o}} = \arg\min_{\gamma} \sum_{i=1}^{N} \left(Y_i - X_i'\gamma\right)^2 + \lambda\|\gamma\|_0, \tag{2.3.2}$$

where $\left\|\gamma_0\right\|_0 := |\{j : \gamma_{0,j} \neq 0\}|$ denotes the ℓ_0 norm and λ is the penalty parameter. Then $\widehat{\gamma}_{\mathrm{o}}$ defined in (2.3.2) can achieve the oracle convergence rate $O_P\left(\sqrt{s/n}\right)$. However, the criterion function (2.3.2) is a non-convex function and solving the minimization problem requires $\sum_{k\leqslant n}\binom{p}{k}$ least square estimations which is an NP-hard problem (Natarajan, 1995).

The LASSO estimator (Tibshirani, 1996) avoid the NP-hard problem by adding the convex function ℓ_1 norm as the penalty term in the criterion function

$$\widehat{\gamma}_{\text{lasso}} = \arg\min_{\gamma} \sum_{i=1}^{N} (Y_i - X_i'\gamma)^2 + \lambda\|\gamma\|_1, \tag{2.3.3}$$

where $\|\gamma\|_1 = \sum_{j=1}^{p} |\gamma_j|$ denotes for the l_1 norm. By adding the ℓ_1 penalty, the LASSO estimators for coefficients can be exactly driven to zero during the regularization process and can be used for variable selection. Besides, the criterion function (2.3.3) is convex thus the computation for LASSO estimator is efficient. The λ is a penalty parameter that controls the shrinkage of estimators and variable selection. We review the choice of λ both in theoretical and practical cross-validation methods.

Theoretically, λ should be large enough to dominate the noise with high probability, $\lambda > 2\left\|n^{-1}\sum_{i=1}^{n} X_i\varepsilon_i\right\|_\infty$. At the same time, λ should be as small as possible to reduce the bias induced by shrinkage. In practice, Bickel, Ritov, and Tsybakov (2009) suggest to set

$$\lambda = 2 \cdot c\sigma\sqrt{2n\log(2p/\alpha)}, \tag{2.3.4}$$

where $c > 1$ and $\alpha \in (0, 1)$ are some constants, σ is the standard deviation of residual ε. Typically σ is unknown and needs to be estimated from the data using iteration method. Belloni and Chernozhukov (2013) also propose a choice of λ which is

$$\lambda = 2c\sigma\sqrt{n}\Phi^{-1}(1 - \alpha/2p), \tag{2.3.5}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard Normal distribution. As showed in Bickel, Ritov, and Tsybakov (2009) and Belloni and Chernozhukov (2013), their choice of λ in (2.3.4) and (2.3.5) lead to a nearly oracle rates of convergence for the estimator $\widehat{\gamma}_{lasso}$ under general conditions.

$$\|\widehat{\gamma}(\lambda) - \gamma\|_2 = O_P\left(\sqrt{\frac{s\log p}{n}}\right).$$

With ℓ_1 penalty in the criterion function, the LASSO estimators of the coefficients γ can be driven exactly to zero during the regularization process. Hence this technique can be used for variable selection and generating more parsimonious model. But only under special cases LASSO can perfectly select the oracle model. In general, Belloni and Chernozhukov (2013) show that the LASSO estimator $\widehat{\gamma}(\lambda)$ with λ defined in (2.3.5) can obtain sparsity results. Specifically, let $T = \{j : \gamma_j \neq 0\}$ and $\widehat{T} = \{j : \widehat{\gamma}_j(\lambda) \neq 0\}$. Then $|\widehat{T} \backslash T| \leq Cs$ with high probability, where C is a constant, which indicates the number of irrelevant regressors selected by LASSO at most has the same order with the true sparsity. The result also implies that $\widehat{s} := |\widehat{T}| \leq s + |\widehat{T} \backslash T| \leq \widetilde{C}s$ with high probability. Thus, the LASSO estimator with penalty choice (2.3.5) has the sparsity property.

LASSO estimator can drive some parameters exactly to zero, but also shrinks all the non-zero parameters towards zero which lead to the estimation bias. In order to eliminate this bias, Belloni and Chernozhukov (2013) suggest to apply Post-LASSO estimator which minimizes the least squares criterion (1) over the non-zero components selected by the LASSO estimator.

$$\widetilde{\gamma} \in \arg\min_{\gamma \in \mathbb{R}^p} \left\{ \sum_{i=1}^{N} \left(Y_i - X_i'\gamma\right)^2 : \gamma_j = 0 \text{ for each } j \in \widehat{T}^c \right\} \tag{2.3.6}$$

where $\widehat{T}^c = \{j : \widehat{\gamma}_j = 0\}$. If the variables are perfectly selected, then the Post-LASSO estimator is exactly the oracle estimator for γ . But even if the model selection is not perfect, Belloni and Chernozhukov (2013) proves that the Post-LASSO estimator can achieves the same near-oracle convergence rate as LASSO and strictly faster under certain cases. Also, by construction, post-LASSO estimator has smaller shrinkage bias.

The LASSO estimator based on theoretical penalty choice (2.3.4) or (2.3.5) has good theoretical properties in both the convergence rate and variable selections. However, the choice of parameters c, α in (2.3.4) and (2.3.5) are arbitrary in practice and they might affect the performance of the estimators. In practice, researchers often prefer to use cross-validation to choose the penalty parameter for the Lasso estimator (Chetverikov, Liao, and Chernozhukov, 2019). Consider the K-folded cross-validation, the sample is partitioned into K subsample. Denote $\widehat{\gamma}^{-k}(\lambda)$ as the LASSO estimator computed with the kth sample removed given any penalty level λ.

$$\widehat{\gamma}_{-k}(\lambda) = \arg\min_{\gamma \in \mathbb{R}^p} \left(\frac{1}{n - n_k} \sum_{i \notin I_k} (Y_i - X_i'\gamma)^2 + \lambda \|\gamma\|_1 \right).$$

Then the cross-validated penalty parameter $\widehat{\lambda}$ is chosen by minimizing the summation of prediction errors on the validation sets,

$$\widehat{\lambda} = \arg\min_{\lambda} \sum_{k=1}^{K} \sum_{i \in I_k} (Y_i - X_i'\widehat{\gamma}_{-k}(\lambda))^2 .$$

Chetverikov, Liao, and Chernozhukov (2019) show that K-fold cross-validated Lasso estimator $\widehat{\gamma}(\widehat{\lambda})$ can attain optimal rate of convergence up to certain logarithmic factors.

$$\|\widehat{\gamma}(\widehat{\lambda}) - \gamma\|_2 = O_P\left(\sqrt{\frac{s \log p}{n}} \times \sqrt{\log(pn)} \right)$$

Their simulation results show that the cross-validation LASSO estimator have much smaller estimation error than the LASSO estimator with λ chosen by (2.3.5).

Chetverikov, Liao, and Chernozhukov (2019) also discuss the sparsity bound for cross-validation LASSO estimator. Theoretically, they show that the number of non-zero components in the cross-validated Lasso estimator $\widehat{\gamma}(\widehat{\lambda})$ may exceed s only by the small factor, $\left(\log^2 p\right)(\log n)(\log(pn) + s^{-1}\log^r)$. However, the simulation results suggest that cross-validation typically yields a small value of λ, thus tends to select too many covariates. Also, for the cross-validation LASSO, to the best of my knowledge, there are still no theoretical results for the performance of the post-LASSO estimator.

In practice, the choice of K for cross-validation is a bias-variance trade-off problem. If K is large, for example, $K = N$ (leave one out), the cross-validation estimator has small bias but high variance. And vice versa for K to be small. Overall, $K = 5$ or $K = 10$ is recommended in practice as a good balance between bias and variance trade-off (Hastie, Tibshirani, and Friedman, 2009).

In our Monte Carlo simulation, we use the cross-validation to choose the penalty parameter λ to avoid the arbitrary choice of parameters in theoretical results. We also consider the performance of the post-LASSO estimator based on the cross-validation variable selection although there are no theoretical results for that case.

2.3.2 Regularized Series Expansions

In regression problems, the conditional expectation $g(X) = \mathrm{E}(Y|X)$ is typically a nonlinear function. Under this case, non-parametric methods such as kernel regression or series estimation methods are more flexible and can be applied.

Series estimators approximate $g(\cdot)$ by a linear model with additional variables which are transformations of the covariates X. Denote $h_j(X) : \mathbb{R}^p \mapsto \mathbb{R}, (j = 1, \dots, d)$ to be the transformation of X. Then the conditional mean can be represented by

$$g(X_i) = \sum_{j=1}^{d} h_j(X_i)\gamma_j + \xi_i,$$

where ξ_i denotes for the approximation error. The choice of the transformations $h_0(\cdot)$ can be some basis functions just as the non-parametric methods. For example, we can use polynomial expansions with

$$h_j(X) = X_m,\ h_j(X) = X_m^2 \text{ or } h_j(X) = X_m X_k, \cdots$$

To capture the high-order Taylor expansions, we can include high-degree polynomials. But the number of the variables $h_0(\cdot)$ grows exponentially with the degree of the polynomials, $O\left(p^d\right)$ for degree-d, which might induce high-dimensional covariates in our regression.

The order-M spline with knots $\xi_j, j = 1, ..., K$ would include

$$h_j(X) = X^{j-1}, (j = 1, \dots, M);\ h_{M+\ell}(X) = (X - \xi_\ell)_+^{M-1}, \ell = 1, \dots, K,$$

where $(X - \xi_\ell)_+ = \max(0, X - \xi_\ell)$, and ξ_ℓ denotes a set of knots which needs to be chosen.

A simple way is to just use the observations X_i as the knots. In practice, the order of the spline M are typically chosen to be 2 (linear spline) or 4 (cubic spline).

The traditional non-parametric methods with series expansion restrict the model by determining the number of regressors before-hand, $p = O(n^{1/2})$. The choice of the basis functions could be arbitrary such that the approximation error ξ_i is not guaranteed to be small. We can relax that restriction by applying the regularization methods which can include high dimensional transformations and impose the penalty on the coefficients to let data determines the choice of basis functions.

Linear regularization methods can be applied to basis-function transformations of the variables and this considerably expands their scope. Suppose we fit a linear spline of the form $g(x) = \sum_{j=1}^{p} \gamma_j \max(0, x - \xi_j)$ and add a l_1 penalty for regularization, then the minimization problem

$$\min_{\gamma} \sum_{i=1}^{n} \left\| Y_i - \sum_{j=1}^{p} \gamma_j \max(0, X_i - \xi_j) \right\|^2 + \lambda \|\gamma\|_1 \tag{2.3.7}$$

Similarly for the polynomials expansion, we can just replace the basis function in equation (2.3.7).

Next, consider the Kernel estimation of the conditional mean. Suppose we want to learn a regression curve of the form

$$g(x) = \sum_{i=1}^{n} c_i K(x, x_i)$$

where $K(x, x_i)$ denotes for the kernel function, for example, $K(x, x') = e^{-\gamma\|x-x'\|^2}$ is the Gaussian kernel. Then the regularized kernel estimator is define as

$$\min_{c} \sum_{i=1}^{n} \left\| y_i - \sum_{j=1}^{n} c_j K(x_i, x_j) \right\|^2 + \lambda \sum_{i,j} c_i c_j K(x_i, x_j) \tag{2.3.8}$$

In both series and kernel regressions, we regularized a high-dimensional linear model with features to be the transformations of the original covariates.

In our Monte Carlo simulations, we use series expansions to approximate the nonlinear conditional expectation function. We include higher order polynomials to reduce the approximation bias and then apply the LASSO regularization to let the data select the important basic functions. We do not consider kernel estimation in our Monte Carlo simulation because kernel estimator is known to have the boundary bias problem (Gasser and Müller, 1979; Rice and Rosenblatt, 1981).

2.3.3 Regression Trees and Random Forest

Regression trees (Breiman et al., 1984), and its extension Random Forest (Breiman, 2001) have become very popular and can be used as a more flexibly estimating methods for the conditional expectations $g(X) = E[Y|X]$.

The basic idea for regression tree is to recursively spit the covariates space into several small regions. Then given a observation X_i, the conditional expectation is simply estimated by the average of the output variable Y within the region that contains X_i. Assume we finally split the sample into K regions based on certain splitting procedure and stopping rule. Then the conditional expectation of Y given X can be estimated by

$$\widehat{g}(X) = \widehat{\mathbb{E}}[Y|X] = \sum_{k=1}^{K} \overline{Y}_{R_k} I\left\{X \in R_k\right\}$$

where R_k denotes for the kth region, $\overline{Y}_{R_k}$ is the mean value of outcome variable in that region.

The splits process takes sequentially and based on a single covariate $X_{\cdot j}$ together with a threshold t at each step. Consider a simple example with two covariates $(X_{\cdot 1}, X_{\cdot 2})$. At the first step, $X_{\cdot 1}$ is compared with the threshold t_1. If smaller, continue to compare $X_{\cdot 2}$ with another threshold t_2. If smaller, we stop and take them as subgroup R_1. Similarly, the other subgroups can be reached.

In general, a tree is built by choosing the splitting variable $X_{\cdot j}$ and the threshold t in each step to minimize the quadratic loss function. Consider the selection problem of jth covariates and the corresponding threshold t in the first step

$$(j^*, t^*) = \arg\min_{j,t} \Big[\sum_{X_i \in R_1(j,t)} \big(Y_i - \overline{Y}_{R_1}\big)^2 + \sum_{X_i \in R_2(j,t)} \big(Y_i - \overline{Y}_{R_2}\big)^2 \Big]. \tag{2.3.9}$$

After finding the best split in the first step, we partition the covariates space into two regions. And for each region, we can repeat the procedure to further split the subspace. The question is to determine when to stop splitting the tree. If the tree is too deep then it might have an overfitting problem. While a shallow tree might not be enough to capture the structure. Hastie, Tibshirani, and Friedman (2009) discussed two ways to determine the depth of the tree.

One approach is to stop splitting the tree if the decrease in quadratic loss function (2.3.9) is smaller than some threshold. But the problem is that a good split might still happen after this stopping rule reached. Thus a tree built using such a stopping rule might be too shallow.

Hastie, Tibshirani, and Friedman (2009) suggest using another strategy which is to grow a large tree T_0 at first, stopping the splitting process only when some minimum node size (say 5) is reached. Then this large tree is pruned to small shallow trees using cost-complexity pruning. Let $|T|$ denotes the number of terminal nodes in any subtree $T \in T_0$. Then the cost-complexity criterion function is defined as

$$\sum_{k=1}^{|T|} \sum_{X_i \in R_k} \big(Y_i - \overline{Y}_{R_k}\big)^2 + \alpha |T| \tag{2.3.10}$$

where α is the tuning parameter that controls the trade-off between the size of trees and overfitting of the data. Given each α, there is a smallest subtree T_α that minimizes the criterion function (2.3.10). If $\alpha = 0$, then T is just the original tree T_0. As α increases, the tree T_α becomes smaller. And the penalty parameter α can be chosen via K-folded cross-validation just like the LASSO estimator. The regression tree estimator is straightforward and easy to implement. However, it suffers from a high variance problem.

The idea for Bagging (bootstrap aggregating) (Breiman, 1996) is to fit the regression tree to different subsamples and then taking averaging over many trees to reduce the variance.

The additional trees can be built by bootstrapping the original data (random draw with replacement). And then $g(X) = E[Y|X]$ is estimated by

$$\widehat{g}(X) = \sum_{m=1}^{M} \frac{1}{M} \widehat{g}_m(X) \tag{2.3.11}$$

with $\widehat{g}_m(X)$ estimated by regression trees using the mth bootstrap sample. Each regression tree can grow deep and is not pruned. Thus the variance for each tree if high and the bias is low. Taking an average of over M trees can reduce the variance. However, the bagged trees built using this method will look similar and the predictions based on these bagged trees will be highly correlated. Thus, the bagging estimator in (2.3.11) which is defined as averaging over highly correlated trees will not leads to a significant reduction in variance (Breiman, 2001).

Random Forest (Breiman, 2001) which is also based on the regression trees further reduces the variance by eliminating the correlation between trees. The idea is to randomly select some input variables instead of using all covariates in the tree-growing process. Specifically, when growing a tree on a bootstrapped sample, at each step, $s < p$ covariates are selected at random for splitting and they are changed for every split. Typically we choose $s = \sqrt{p}$. Then the Random Forest estimator is the average of the M de-correlated trees.

$$\widehat{g}_{rf}(X) = \sum_{m=1}^{M} \frac{1}{M} \widehat{g}_m^s(X) \tag{2.3.12}$$

where $\widehat{g}_m(X)$ is estimated by regression trees using the mth bootstrap sample and random selected $s = \sqrt{p}$ inputs in each split.

Due to the random selection of a small set of covariates for each split, Random Forest could perform poorly for very sparse models. For example, if $p = 100$ but only 3 of the covariates are relevant. Assume we choose $m = \sqrt{p} = 10$, then the probability of selecting the 3 relevant variables in each split is small. If that is the case, the Random Forest estimator will perform poorly. But as the number of relevant variables increases, then Random Forest becomes robust to the number of noisy variables because the splitting will generally ignore the irrelevant ones (Athey and Wager, 2017).

One merit of the Random Forest is that it does not suffer from the overfitting problem (Hastie, Tibshirani, and Friedman, 2009). The Random Forest is defined as the average of over M de-correlated trees. First increasing the number of trees M can not cause the overfitting problem. Second, for the depth of each tree in the Random Forest, Segal (2004) show that there are small improvement in the performance if we prune the individual trees in the forest. But Hastie, Tibshirani, and Friedman (2009) suggests using full-grown trees instead because it will not cost too much and we do not need to deal with the tuning parameter problem.

Another important feature of Random Forest is the use of out-of-bag (OOB) samples to do prediction. Each time we draw a bootstrap sample, we only use part of the original sample. Thus, to do a prediction for observation X_i, we only need to pick the bootstrap samples which do not contain X_i (OOB sample). And then the OOB predictor is constructed by averaging over the predictors for the OOB samples. Athey and Wager (2017) show that the performance of the target parameter will be improved if we use such techniques in the first step estimation of nuisance parameters. And Random Forest can achieve that goal without doing sample splitting which might cause efficiency loss.

Athey, Tibshirani, and Wager (2019) also discuss the connection between Random Forest and the traditional kernel estimation methods. For a regression tree, the prediction of observation X_i is simply the sample average of outcome over the region that contains X_i. The region can be taken as the set of nearest neighbors for the targeted observation X_i. Thus the regression tree is just a special case of matching estimator. And Random Forest is the average of several such matching estimators. Which can be interpreted as the weighted average estimator with weighting functions analogous to kernel estimators. Compared with the kernel regression, Random Forest is robust to increasing the noisy covariates since it can ignore them during the splitting.

2.3.4 Neural Nets

The basic idea of Neural Nets (Friedman and Stuetzle, 1981) is to generate new features using the linear combinations of the original covariates, and then model the outcome variable as a nonlinear function of these features. Consider the single hidden layer Neural Nets.

In the first stage, the hidden features $h_m, (m = 1, \ldots, M)$ are generated by the linear combinations of the covariates.

$$h_m = h_0(X'\gamma_m)$$

where γ_m is a p dimensional coefficients vector for each $m = 1, \cdots, M$.
In the second stage, the output is then generated by a linear combinations of the $h_m, (m = 1, \cdots, M)$.

$$g(X) = \sum_{m=1}^{M} \alpha_m h_m = \sum_{m=1}^{M} \alpha_m h\left(X'\gamma_m\right) \tag{2.3.13}$$

$h_0(\cdot)$ is called the activation function and typically we use the nonlinear Hinge or Sigmoid function

$$\text{Hinge function : } \; h_0(v) = \max\{0, v\},$$

$$\text{Sigmoid function: } h_0(v) = 1/\left(1 + e^{-v}\right).$$

The nonlinear Hinge or Sigmoid function imposed on a linear combinations on the covariates can generate a large class of models, thus can handle complex function forms of the conditional expectation (Hastie, Tibshirani, and Friedman, 2009).

The Neural Nets is related to the basis expansion methods such as series or kernel regression. The hidden units h_m are similar to the basis expansions of the original covariates X. But these h_m are unobserved and the function parameters are learned from the data.

Usually, there are too many weight parameters α and γ such that overfitting is a problem. A simple and explicit method for regularization is to add a penalty term to the criterion function just as the LASSO/ridge regression in the linear model.

$$\min_{\alpha,\gamma}\left\{\sum_{i=1}^{n}\left(Y_i-\sum_{m=1}^{M}\alpha_m h\left(X_i'\gamma_m\right)\right)^2+\lambda\left(\sum_{m=1}^{M}\alpha_m^2+\sum_{m,\ell}\gamma_{m,\ell}^2\right)\right\}. \tag{2.3.14}$$

Neural Nets estimator defined in (2.3.14) is just an extension of the traditional non-linear least square (NLS) estimator with the ℓ_2 penalty term. The direct computation of the nonlinear regularization is a problem. Rumelhart et al. (1986) proposed the back-propagation algorithm to compute the weight coefficient α,γ defined in equation (2.3.14).

Step (i) Pick an initial value for α,γ

Step (ii) Compute the value of each $h_m,(m=1,\dots,M)$, and then estimate $g(\cdot)$ using ridge regression. –(feed-forward)

Step (iii) Compute the errors $Y-\widehat{g}(X)$

Step (iv) Update all parameters via gradient descent, and go back to Step (ii) to iterate. – (back propagation)

We can extend the single hidden layer network model to the Deep Neural Nets model by including more layers. The Deep Neural Nets can be even more flexible for fitting the models (Hastie, Tibshirani, and Friedman, 2009). However, choosing the number of hidden layers and the number of hidden units for each layer is still a problem. For our simulation, we just use the simple single hidden layer Neural Nets model to estimate the conditional expectations in the first step.

2.4 Semiparametric Estimation of Endogenous Social Interaction Effect

In this paper, the parameter of interest is the endogenous social interaction effect, β_0, in model (2.2.1). I apply a semiparametric instrumental variable strategy to identify the parameter of interest β_0. Based on the identification condition, a semiparametric two-step estimator is constructed in which the first step nonparametric functions could be estimated by Machine Learning methods as discussed in Section 2.3, such as LASSO, Random Forest, and Neural Nets. Then the second step parametric endogenous social interaction effect, β_0, is estimated based on the moment condition (2.2.4) or orthogonal moment condition (2.2.6).

There are several papers that discuss the semiparametric two step estimation for the parameters identified by moment or orthogonal moment conditions. Following the terminology in literature, they are named as "plug-in" estimator and "debiased" estimator, respectively. I will first give a review of the existing literature and then discuss how to apply the method in the semiparametric social interaction model (2.2.1).

2.4.1 Review of the Semiparametric Estimation Methods

There are several papers discuss the semiparametric plug-in estimation for the parameters identified by moment condition (2.2.4).

Andrews (1994) proposed a general semiparametric two-step M-estimators for β_0 and provide the conditions for the root-n consistency and asymptotic normality for these estimators. The estimators are named MINPIN because they minimize a criterion function that depends on a preliminary infinite-dimensional nuisance parameter estimators of ϑ_0 in the first step.

This method can be used to derive the asymptotic distribution of a wide variety of semiparametric estimators. For example, a simple sample analog estimator for β_0 is an M-estimator that minimizes

$$\left(\frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta,\widehat{\vartheta})\right)'\left(\frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta,\widehat{\vartheta})\right)$$

where $\widehat{\vartheta}$ is some preliminary nonparametric estimator for ϑ_0 in the first step.

Andrews (1994) provided a set of relatively high level assumptions to show the root-n consistency and asymptotic normality for the MINPIN estimators of β_0. The key condition employed is stochastic equicontinuity which can be verified using empirical process results.

Define the empirical process $\nu_n(\vartheta)$ by

$$\nu_n(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} \Big(m(Z_i, \beta_0, \vartheta) - E\left[m(Z_i, \beta_0, \vartheta)\right] \Big) \tag{2.4.1}$$

Then the definition for stochastic equicontinuity of $\{\nu_n(\cdot)\}$ is that

$$\left| \nu_n(\widehat{\vartheta}) - \nu_n\left(\vartheta_0\right) \right| \xrightarrow{p} 0 \tag{2.4.2}$$

holds for all $\widehat{\vartheta}$ that satisfies $\rho\left(\widehat{\vartheta}, \vartheta_0\right) \xrightarrow{p} 0$, where $\rho(\cdot)$ is pseudo-metric.

Andrews (1993, 1994) provide a set of sufficient conditions for stochastic equicontinuity. However, they are still high-level conditions and are difficult to verify especially for some highly nonlinear functions. Indeed, it does not hold for all classes of functions. Some restrictions on the functions are necessary. If the model is too complex, for example, high dimensional models, the stochastic equicontinuity condition might not hold.

Ichimura and Lee (2010) presented a set of relatively low-level conditions to ensure the root-n consistency and asymptotic normality of the semiparametric two-step M-estimator of β_0, which is relatively easier to verify and also weaker than the conditions given by Andrews (1994).

Newey (1994), Pakes and Olley (1995), Chen, Linton, and Van Keilegom (2003), and Chen (2007) studied the semiparametric two-step MM/GMM for β_0. If β_0 is just identified by moment condition, i.e. $\dim(m) = \dim(\beta_0)$, then β_0 can be directly estimated by solving the sample analogy of the moment function:

$$\frac{1}{n} \sum_{i=1}^{n} m(Z_i, \widehat{\beta}, \widehat{\vartheta}) = 0 \tag{2.4.3}$$

where $\widehat{\vartheta}$ denote an estimator of the true nuisance function ϑ_0 using any nonparametric methods.

If β_0 is over identified by moment condition, i.e. $\dim(m) > \dim(\beta_0)$, then β_0 can be estimated by making $\frac{1}{n} \sum_{i=1}^{n} m(Z_i, \widehat{\beta}, \widehat{\vartheta})$ as close to zero as possible, this is the idea of GMM estimator. As considered in Newey (1994), Pakes and Olley (1995), Chen, Linton, and Van Keilegom (2003), and Chen (2007), a semiparametric two-step GMM with a nonparametric first step estimator is defined

$$\widehat{\beta}^{gmm} = \arg\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} m(Z_i, \beta, \widehat{\vartheta}) \right)' \widehat{W} \left(\frac{1}{n} \sum_{i=1}^{n} m(Z_i, \beta, \widehat{\vartheta}) \right) \tag{2.4.4}$$

where $\widehat{W}$ is a positive-definite (random) weight matrix and $\widehat{W} \xrightarrow{p} W$.

Newey (1994), Pakes and Olley (1995), Chen, Linton, and Van Keilegom (2003), and Chen (2007) have also studied conditions to ensure the asymptotic properties of the semiparametric two-step MM/GMM estimator. Chen, Linton, and Van Keilegom (2003) 's results extend Newey (1994) and Pakes and Olley (1995) 's by considering where the criterion function does not obey standard smoothness conditions. The key condition for the asymptotic properties of β_0 is still to show the stochastic equicontinuity for $\nu_n(\vartheta)$. Besides, Newey (1994) showed that the limiting distribution of the second stage estimation of β_0 is invariant to the choice for the first step nonparametric estimator as long as certain convergence rate requirement is satisfied.

For specific first step estimators, such as kernel and series, more primitive conditions are employed in the literature to ensure the consistency and root-n asymptotic normality of β_0. For example, Newey (1994) derives the regularity conditions given the first step to be series estimators. Newey and Mcfadden (1994) take the first step kernel estimator as an example to illustrate the primitive conditions. Chen (2007, 2013) considers the series extremum estimation in the first step. Abadie (2003) studies the semiparametric IV estimation of treatment response models. He provides primitive conditions for asymptotic normality of the treatment effect that use power series in the first step.

Newey (1994) also provides a general formula to compute the asymptotic variance of the semiparametric second-step estimator for β_0 by taking into account the first step nonparametric estimates of nuisance functions. The asymptotic variance estimator based on the influence function constructed by adding an adjustment term in the usual formula for the

M-estimator if ϑ_0 is known.

$$\phi(z) = -M^{-1} \{m(z, \beta_0, \vartheta_0) + \alpha(z)\} \tag{2.4.5}$$

where $M \equiv \partial E[m(z, \beta, h_0)] / \partial\beta|_{\beta_0}$ and $\alpha(z)$ is the adjustment term such that $E[\alpha(z)] = 0$ and $\partial E[m(z, h_0(\theta))]/\partial\theta = E[\alpha(z)S(z)]$.

Newey (1994) showed that given the influence function, the estimator of ϑ_0 does not affect calculating the asymptotic variance, the same as if ϑ_0 is known. Adding the adjustment term $\alpha(z)$ is to make the second step moment condition to be orthogonal to the first step nuisance function which ensures the invariance to the first step nonparametric estimations.

The other papers mentioned above also derive the formula for the semiparametric asymptotic variance of the corresponding estimators they considered. Some of them also show how to consistently estimate the asymptotic variance. However, in practice, it is not straightforward for empirical researchers to implement. There are also papers consider to use the bootstrap to estimate asymptotic variances (Ellickson and Misra, 2008), but the computation burden is heavy and is difficult to verify theoretically (Ackerberg, Chen, and Hahn, 2012).

Ackerberg, Chen, and Hahn (2012) provide several equivalence results that can greatly simplify the semiparametric asymptotic variance estimation. They showed that for a large class of models, the semiparametric asymptotic variance with nonparametric methods in the first step is numerically identical to pretending to use a parametric first step. Thus, in the empirical studies, one can ignore the semiparametric setup and simply calculate the asymptotic variance as if the first step is parametrically estimated.

2.4.2 Plug-in Estimator for Endogenous Social Interaction Effect

In the semiparametric social interaction models, the parameter of interest, β_0, is identified by the moment condition (2.2.4). This paper adopts the semiparametric MM/GMM method to estimate β_0 by plugging in the first step nonparametric (Machine Learining) estimators

$\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$.

The endogenous social effect β_0 is a one-dimensional parameter ($\dim(\beta_0) = 1$) while the dimension of moment conditions (2.2.4) (denoted by $\dim(m)$) used to estimate β_0 has the same dimension as $X_{g,i}$. Denote $d = \dim(X_{g,i})$ and assume d is fixed, then $\dim(m) = d$. In the following discussion, I will consider the estimator for β_0 under two cases: $d = 1$ ('just identified') and $d > 1$ ('over identified'). Here I might abuse the concept for 'just' and 'over' identification. By using 'just' or 'over' identified, I do not refer to the model is just or over identified based on the strictly exogenous assumption, but assume we are restricted to use moment conditions (2.2.4) to identify β_0.

Just identified second step

Consider the just identified second step case for β_0 by moment condition (2.2.4), i.e. $\dim(m) = \dim(\beta_0) = 1$. Then the semiparametric plug-in estimator $\widehat{\beta}$ for β_0 can be computed by directly solving the sample analogy of moment conditions (2.2.4).

$$0 = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\overline{X}_{g,-i}\Big(\big(Y_{g,i} - \mu(X_{g,i})\big) - \beta\big(\overline{Y}_{g,-i} - \nu(X_{g,i})\big)\Big). \tag{2.4.6}$$

Thus the semiparametric plug-in estimators $\widehat{\beta}$ has the closed form solutions as follows:

$$\begin{aligned}\widehat{\beta} &= \Big(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\overline{X}_{g,-i}\big(\overline{Y}_{g,-i} - \widehat{\nu}(X_{g,i})\big)\Big)^{-1}\Big(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\overline{X}_{g,-i}\big(Y_{g,i} - \widehat{\mu}(X_{g,i})\big)\Big) \\ &= \Big(\sum_{g=1}^{G}\overline{X}'_{g,-}\widehat{\zeta}_g\Big)^{-1}\Big(\sum_{g=1}^{G}\overline{X}'_{g,-}\widehat{\eta}_g\Big) \end{aligned}\tag{2.4.7}$$

where $\widehat{\eta}_g$ and $\widehat{\zeta}_g$ are the group level residualized terms for Y_g and $\overline{Y}_{g,-}$, respectively.

Over identified second step

Next, I will discuss the over identified second step where $\dim(m) > \dim(\beta_0) = 1$. To address the clustered structural data of the grouped social interaction model, it will be convenient to define the clustered sum of moment functions, $m_g\left(Z_g,\beta,\vartheta\right)=\sum_{i=1}^{n_g} m\left(Z_{g,i};\beta,\vartheta\right)$ where $Z_g=\left(Z_{g,1},\cdots,Z_{g,n_g}\right)$. Then β_0 is estimated based on the sample analog of $\frac{1}{n}\sum_{g=1}^{G} m\left(z_g,\widehat{\beta},\widehat{\vartheta}\right)$ to be as close to zero as possible. This is the idea of the semiparametric two-step GMM procedure (Chen, 2007; Ackerberg, Chen, and Hahn, 2012).

The Semiparametric Plug-in GMM estimator, $\widehat{\beta}^{gmm}$, solves the following minimization problems:

$$\widehat{\beta}^{gmm}=\arg\min_{\beta}\left(\frac{1}{n}\sum_{g=1}^{G} m_g\left(z_g,\beta,\widehat{\vartheta}\right)\right)'\widehat{\Omega}^{-1}\left(\frac{1}{n}\sum_{g=1}^{G} m\left(z_g,\beta,\widehat{\vartheta}\right)\right) \tag{2.4.8}$$

where $\widehat{\Omega}^{-1}$ denotes an $d\times d$ positive definite weight matrix. Because of the linearity of moment function $m_g(\cdot)$ with respect to β, $\widehat{\beta}^{gmm}$ defined in (2.4.8) has a closed form solution:

$$\widehat{\beta}^{gmm} = \Big(\sum_{g=1}^{G}\widehat{\zeta}_g'\overline{X}_{g,-}\widehat{\Omega}^{-1}\sum_{g=1}^{G}\overline{X}_{g,-}'\widehat{\zeta}_g\Big)^{-1}\Big(\sum_{g=1}^{G}\widehat{\zeta}_g'\overline{X}_{g,-}\widehat{\Omega}^{-1}\sum_{g=1}^{G}\overline{X}_{g,-}'\widehat{\eta}_g\Big). \tag{2.4.9}$$

where $\widehat{\eta}_g$ and $\widehat{\zeta}_g$ are the group level residualized terms for Y_g and $\overline{Y}_{g,-}$, respectively.

Next, I will discuss the choice of the inverse of weight matrix Ω. Following Ackerberg, Chen, and Hahn (2012) and Ackerberg, Chen, Hahn, and Liao (2014), the optimal choice of the weight matrix that minimizes the asymptotic variance of $\sqrt{n}(\widehat{\beta}^{gmm}-\beta_0)$ should be the consistent estimator of the inverse of the variance defined as follows,

$$\Omega=\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} m_g\left(z_g,\beta_0,\widehat{\vartheta}\right)\right). \tag{2.4.10}$$

Due to the first step nonparametric estimator $\widehat{\vartheta}$, in general the variance Ω is different from $\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G} m_g\left(z_g,\beta_0,\vartheta_0\right)\right)$. Following Newey (1994), the variance Ω can be computed using the influence function by taking into account the first step nonparametric estimators. In Section 2.2, we have shown that the influence function or orthogonal moment function for $m(.)$ is $\psi(.)$ which is defined in 2.2.7. Then the inverse weight matrix Ω equals to

$$\Omega = \Omega_0 = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\psi_g\left(z_g, \beta_0, \vartheta_0\right)\right) \tag{2.4.11}$$

where $\psi_g(.)$ is the orthogonal moment function. It follows from (2.4.11) that the inverse weight matrix Ω can be estimated by

$$\widehat{\Omega} = \frac{1}{n}\sum_{g=1}^{G}\left(\widehat{\varepsilon}_g'\widehat{U}_g\widehat{U}_g'\widehat{\varepsilon}_g\right) \tag{2.4.12}$$

where $\widehat{\varepsilon}_g = \overline{X}_{g,-} - \widehat{\phi}(X_g)$ and $\widehat{U}_g = (Y_g - \widehat{\mu}(X_g)) - \widetilde{\beta}(\overline{Y}_{g,-} - \widehat{\nu}(X_g))$. $\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$ are any first step consistent nonparametric estimators for $\vartheta_0 = (\mu_0, \nu_0, \phi_0)$, and $\widetilde{\beta}$ can be any consistent estimator for β_0 that solves the minimization problem (2.4.8) with an arbitrary weight matrix.

Finally, the asymptotic variance of $\widehat{\beta}_{db}^{gmm}$ can be constructed by

$$V_n = M_n^{-1}\Omega_n M_n^{-1} \tag{2.4.13}$$

where $M_n = \frac{1}{n}\sum_{g=1}^{G} E\left[\varepsilon_g'\zeta_g\right]$ and $\Omega_n = \frac{1}{n}\sum_{g=1}^{G} E\left[\varepsilon_g'U_gU_g'\varepsilon_g\right]$.

As usual the estimator for the variance of $\widehat{\beta}_{db}$ can be constructed by

$$\widehat{V}_n = \widehat{M}_n^{-1}\widehat{\Omega}_n\widehat{M}_n^{-1} \tag{2.4.14}$$

with $\widehat{M}_n = \frac{1}{n}\sum_{g=1}^{G}\left[\widehat{\varepsilon}_g'\widehat{\zeta}_g\right]$ and $\widehat{\Omega}_n = \frac{1}{n}\sum_{g=1}^{G}\left[\widehat{\varepsilon}_g'\widehat{U}_g\widehat{U}_g'\widehat{\varepsilon}_g\right]$, where $\widehat{\eta}_g$, $\widehat{\zeta}_g$, and $\widehat{\varepsilon}_g$ denote the first step residualized terms for $Y_g, \overline{Y}_{g,-}, \overline{X}_{g,-}$, respectively, and $\widehat{U}_g = \widehat{\eta}_g - \widehat{\beta}^{gmm}\widehat{\zeta}_g$. The variance estimator, $\widehat{V}_n$ is constructed "as if" the nonparametric parameters ϕ, μ, and ν were known. This works because the moment condition (2.2.6) is locally/doubly robust to the first step estimators (Ackerberg, Chen, Hahn, and Liao, 2014).

The following is the algorithm for obtaining the plug-in semiparametric GMM estimator $\widehat{\beta}^{gmm}$ and the estimator for its variance.

(i) Estimate $\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$ by any nonparametric (Machine Learning) methods, such as LASSO, Random Forest, and Neural Nets.

(ii) Using an arbitrary inverse weight matrix $\widetilde{\Omega}$, plug into (2.4.9) to obtain a preliminary estimator $\widetilde{\beta}$ for β_0. For example, if we choose $\widetilde{\Omega}$ to be identity matrix, then $\widetilde{\beta} = \left(\widehat{\zeta}' X_- X_-' \widehat{\zeta}\right)^{-1} \left(\widehat{\zeta}' X_- X_-' \widehat{\eta}\right)$

(iii) Construct the consistent estimator for the optimal inverse weight matrix,

$$\widehat{\Omega} = \frac{1}{n} \sum_{g=1}^{G} \left(\widehat{\varepsilon}_g' \widetilde{U}_g \widetilde{U}_g' \widehat{\varepsilon}_g \right)$$

where $\widehat{\varepsilon}_g = \overline{X}_{g,-} - \widehat{\phi}(X_g)$ and $\widetilde{U}_g = (Y_g - \widehat{\mu}(X_g)) - \widetilde{\beta}(\overline{Y}_{g,-} - \widehat{\nu}(X_g))$. $\widehat{\mu}, \widehat{\nu}, \widehat{\phi}$ and $\widetilde{\beta}$ are defined in step (i) and (ii).

(iv) The semiparametric estimator $\widehat{\beta}^{gmm}$ for β_0 can be constructed by plugging in first step estimators $\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$ in step (i) and also the inverse weight matrix estimator $\widehat{\Omega}$ in step (iii) into (2.4.9).

(v) The consistent estimator for asymptotic variance of $\widehat{\beta}^{gmm}$ can be constructed by plugging in first step estimators $\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$ in step (i) and the second step semiparametric estimator $\widehat{\beta}^{gmm}$ in step (iv) into (2.4.14).

2.4.3 Debiased Estimator for Endogenous Social Interaction Effect

The semiparametric debiased estimator for the parameter of interest β_0 is constructed based on the orthogonal moment condition (2.2.6) by plugging in the first step nonparametric (Machine Learining) estimators $\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$. In the following discussion, I will still consider both the just identified case ($\dim(X_i) = 1$) and over identified case ($\dim(X_i) > 1$).

Just identified second step

Consider the just identified second step case for β_0 by moment condition (2.2.6), i.e. $\dim(m) = \dim(\beta_0) = 1$. Then the semiparametric debiasing estimator $\widehat{\beta}$ for β_0 can be computed by directly solving the sample analogy of moment conditions (2.2.6).

$$0 = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\overline{X}_{g,-i} - \widehat{\psi}(X_{g,i})\Big)\Big(\big(Y_{g,i} - \mu(X_{g,i})\big) - \beta\big(\overline{Y}_{g,-i} - \nu(X_{g,i})\big)\Big). \quad (2.4.15)$$

Thus the semiparametric debiased estimators $\widehat{\beta}_{db}$ has the closed form solutions as follows:

$$\widehat{\beta}_{db} = \Big(\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\zeta}_g\Big)^{-1}\Big(\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\eta}_g\Big) \quad (2.4.16)$$

where $\widehat{\eta}_g$, $\widehat{\zeta}_g$, and $\widehat{\phi}_g$ are group level residuals for Y_g, $\overline{Y}_{g,-}$, and $\overline{X}_{g,-}$, respectively.

Over identified second step

The semiparametric debiased GMM estimator, $\widehat{\beta}_{db}^{gmm}$, is constructed by making the sample analogy of orthogonal moment function $\psi_g\left(Z_g, \beta, \vartheta\right)$, as close to zero as possible.

$$\widehat{\beta}_{db}^{gmm} = \arg\min_{\beta}\left(\frac{1}{n}\sum_{g=1}^{G}\psi_g\left(Z_g, \beta, \widehat{\vartheta}\right)\right)'\widehat{\Omega}_{db}^{-1}\left(\frac{1}{n}\sum_{g=1}^{G}\psi_g\left(Z_g, \beta, \widehat{\vartheta}\right)\right), \quad (2.4.17)$$

where $\widehat{\Omega}_{db}^{-1}$ denotes an $d\times d$ positive definite weight matrix. Similarly, the closed form solution for the debiased GMM estimator $\widehat{\beta}_{db}^{gmm}$ is:

$$\widehat{\beta}_{db}^{gmm} = \Big(\sum_{g=1}^{G}\widehat{\zeta}_g'\widehat{\varepsilon}_g\widehat{\Omega}_{db}^{-1}\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\zeta}_g\Big)^{-1}\Big(\sum_{g=1}^{G}\widehat{\zeta}_g'\widehat{\varepsilon}_g\widehat{\Omega}_{db}^{-1}\sum_{g=1}^{G}\widehat{\varepsilon}_g'\widehat{\eta}_g\Big) \quad (2.4.18)$$

where $\widehat{\eta}_g$, $\widehat{\zeta}_g$, and $\widehat{\phi}$ are group level residuals for Y_g, $\overline{Y}_{g,-}$, and $\overline{X}_{g,-}$, respectively. $\widehat{\Omega}_{db}$ is chosen to be the consistent estimator of the variance $\Omega_{db} = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\psi_g\left(z_g, \beta_0, \widehat{\vartheta}\right)\right)$. Since $\psi_g\left(Z_g, \beta, \vartheta\right)$ is doubly robust to the estimation of ϑ_0, it can be shown that $\Omega_{db} = \Omega_0 + o_p(1)$, where $\Omega_0 = Var\left(\frac{1}{\sqrt{n}}\sum_{g=1}^{G}\psi_g\left(z_g, \beta_0, \vartheta_0\right)\right)$.

The algorithm for obtaining the debiased semiparametric GMM estimator $\widehat{\beta}_{db}^{gmm}$ is similar as the plug-in GMM estimator $\widehat{\beta}^{gmm}$ discussed in Section 2.4.2.

2.5 Monte Carlo Simulation

This section investigates the finite sample performance of semiparametric two-step estimators for endogenous social interaction effects, β_0, using Monte Carlo simulations. The goal is to evaluate the performance of different estimation methods for β_0 under the semiparametric model setup.

I will conduct a Monte Carlo study with different DGPs to compare the performance. I consider different DGPs with different function forms in the nonparametric nuisance function part, including linear or nonlinear functions. For each DGP, different semiparametric estimators will be applied with different first-step nonparametric estimation methods, including the traditional series estimators and the more recent Machine Learning methods, and also different second step methods, including the plug-in estimators and debiased estimators.

I will report five measures to evaluate the performance of different estimation methods: Bias, Variance, Mean square error (MSE), Mean absolute error (MAE) and empirical size. I find that there is no unique estimation method that dominates all the other estimators consistently across different DGPs. However, the results show that a certain estimator has a more stable performance for all the DGP considered which is recommended in the empirical studies for the semiparametric social interaction models.

2.5.1 Simulation Set-up

2.5.1.1 Data generating process (DGP)

In the Monte Carlo Simulation studies, I will consider the semiparametric social interaction model with group size $M_g = 2$ across all the groups.

$$\begin{bmatrix} Y_{g,1} \\ Y_{g,2} \end{bmatrix} = \beta_0 \begin{bmatrix} Y_{g,2} \\ Y_{g,1} \end{bmatrix} + \begin{bmatrix} h_0(X_{g,1}) \\ h_0(X_{g,2}) \end{bmatrix} + \begin{bmatrix} U_{g,1} \\ U_{g,2} \end{bmatrix}, \quad g = 1, \cdots, G \tag{2.5.1}$$

The following are the DGPs considered in the Monte Carlo simulation studies.

(i) The dimensionality of control variables X.

This simulation study considers the DGPs with both univariate control variable case, $\dim(X) = 1$, and multivariate case, $\dim(X) = 3$.

(ii) The function forms of nonparametric parameter $h_0(\cdot)$.

Both the linear and nonlinear function forms are considered in the simulation study, including $h_0(X) = X\gamma$, $h_0(X) = 1/(1 + exp(-X\gamma))$, $h_0(X) = exp(X\gamma)$, and $h_0(X) = sin(X\gamma) + cos(X\gamma)$.

(iii) The distribution of the control variables X.

Assume the group pair regressors $(X_{g,1}, X_{g,2})$ are independently draw across groups.

First, consider $(X_{g,1}, X_{g,2})$ follows multivariate Gaussian distribution:

$$\begin{bmatrix} X_{g,1} \\ X_{g,2} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

where ρ measures the within group correlations of $X_{g,1}$ and $X_{g,2}$. I consider two cases which are $\rho = 0$ and $\rho = 0.5$. $\rho = 0$ denotes the case that groups are randomly assigned and $\rho = 0.5$ represents the case that groups are linked through the similarity of individuals' characteristics.

I will also consider the case that $(X_{g,1}, X_{g,2})$ follows bivariate logistic distribution (Gumbel, 1960). The bi-logistic distribution function with parameter $0 < r < 1$ is

$$G(X_{g,1}, X_{g,2}) = \exp\left[-\left(y_{g,1}^{1/r} + y_{g,2}^{1/r}\right)^r\right]$$

where $y_{g,i} = y_{g,i}(X_{g,i}) = \{1 + s_{g,i}(X_{g,i} - a_{g,i})/b_{g,i}\}^{-1/s_{g,i}}$, for $1+s_{g,i}(X_{g,i} - a_{g,i})/b_{g,i} > 0$ and $i = 1, 2$. The marginal distributions are generalized extreme value $G(X_{g,i}) = \exp(-y_{g,i})$ for $i = 1, 2$. The dependence between $X_{g,1}$ and $X_{g,2}$ are measured by r. In the simulation study, I choose $r = 0.5$.

Finally, I will discuss a special case that $X_{g,1}$ and $X_{g,2}$ are uncorrelated but dependent. To be specific, $X_{g,i} \sim U[-1,1]$, and $X_{g,-i} = 1\{X_{g,i} > 0\}X_{g,i} - 1\{X_{g,i} \leq 0\}X_{g,i}$[1]. It can be shown that $X_{g,i}$ and $X_{g,-i}$ are linearly uncorrelated but are dependent in higher order terms.

(iv) The distribution of the disturbance $U_{g,i}$.

Assume

$$\begin{bmatrix} U_{g,1} \\ U_{g,2} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho_u\sigma_u^2 \\ \rho_u\sigma_u^2 & \sigma_u^2 \end{bmatrix}\right)$$

For all the DGPs, $\sigma_u^2 = 0.5, \rho_u = 0.5$ is considered.

(v) The parameter of interest β_0.

β_0 should belong to $(-1, 1)$. For discussing the possible spurious social interaction effect, let the true $\beta_0 = 0.5$.

Given the setup above, Table 2.1 summarizes all the DGPs considered in the Monte Carlo simulation study.

The number of observations is $N = 500$, and the number of simulation repetitions $S = 1000$.

[1] 1(.) denote the indicator function.

Table 2.1: List of DGPs

	dim(X)	Function form $h_0(X)$	Distribution of X	Parameters
DGP 1/3-1-1	1 or 3	$h_0(X) = X\gamma$	Bi-Normal	$\gamma = 1; 0.1, \rho = 0.5$
DGP 1/3-1-2	1 or 3	$h_0(X) = X\gamma$	Bi-Logistic	$\gamma = 1; 0.1, \alpha = 0.5$
DGP 1/3-2-1	1 or 3	$h_0(X) = 1/exp(-X\gamma)$	Bi-Normal	$\gamma = 1, \rho = 0.5$
DGP 1/3-2-2	1 or 3	$h_0(X) = 1/exp(-X\gamma)$	Bi-Logistic	$\gamma = 1, \alpha = 0.5$
DGP 1/3-3-1	1 or 3	$h_0(X) = exp(X\gamma)$	Bi-Normal	$\gamma = 0.5, \rho = 0.5$
DGP 1/3-3-2	1 or 3	$h_0(X) = exp(X\gamma)$	Bi-Logistic	$\gamma = 0.5, \alpha = 0.5$
DGP 1/3-4-1	1 or 3	$h_0(X) = sin(X\gamma) + cos(X\gamma)$	Bi-Normal	$\gamma = 1, \rho = 0.5$
DGP 1/3-4-2	1 or 3	$h_0(X) = sin(X\gamma) + cos(X\gamma)$	Bi-Logistic	$\gamma = 1, \alpha = 0.5$
DGP 3-1/2/3	3	3 function forms	Special case	$\gamma = 1$

2.5.1.2 Estimators

For each simulated dataset, the endogenous social interaction effect, β_0, is estimated by the semiparametric two-step GMM described in the preceding section.

In the first step, the conditional expectations, $\mu_0(x), \nu_0(x), \phi_0(x)$, are estimated using different Machine Learning methods such as LASSO, Post-LASSO (P.L.), Post-Lasso Selection (PL.S), Random Forest (R.F.) and Neural Nets (N.N.). For comparison, I also consider to apply the parametric linear estimator and the nonparametric series estimator with polynomial expansion.

Table 2.2 summaries the methods used in the first step conditional expectations estimation.

Table 2.2: List of Estimation Methods

Methods	Description
Linear	Linear regression with X as explanatory variable
Series	Linear regression with the polynomial expansion as explanatory variable
LASSO	Apply LASSO with the polynomial expansion for each $\mu_0(x), \nu_0(x), \phi_0(x)$
Post-LASSO (P.L.)	Apply Post LASSO for each $\mu_0(x), \nu_0(x), \phi_0(x)$
P.L. Selection (P.L.S)	LASSO for variable selection for $\mu_0(x), \nu_0(x), \phi_0(x)$, then apply Post LASSO
Random Forest (R.F.)	Number of trees=500 and each time randomly select $s = \sqrt{d}$ variables.
Neural Nets (N.N.)	Number of hidden layers=1, with the size for the layer =3

To estimate each of the conditional expectation $\mu_0(x), \nu_0(x), \phi_0(x)$, linear estimators simply apply the linear regression with X as explanatory variables. Series estimators consider using the polynomial expansion of X as explanatory variables which should be more robust under the nonlinear case. However, these traditional nonparametric estimation methods can not deal with the very large number of polynomial terms. In our simulation, we also consider more recent Machine Learning methods in the first step estimation. For example, LASSO can be applied to estimate each of the conditional expectations μ_0, ν_0, and ϕ_0. Post-LASSO refit each of the LASSO selected models to reduce the attenuation bias due to regularization. Post-LASSO Selection(PL.S) uses a slightly different strategy that refits the three regressions using all the variables selected in the first step LASSO estimators. For the Random Forest method, we choose the number of trees to be 500 and for each bootstrapped sample, this simulation study randomly select $s = \sqrt{d}$ control variables from X to fit in each split. The Neural Nets (N.N.) is also applied in the simulated dataset. As far as my knowledge, there is no theoretical results on how to choose the hyperparameters for the Neural Nets. In our Monte Carlo simulation, to avoid the overfitting issue, I choose the number of hidden layers in the network just to be one.

Our Monte Carlo simulations are implemented using R programming. LASSO is implemented using the 'cv.glmnet' package with the penalty parameter chosen by the 10-folded crossvalidation. Random Forest is implemented using the 'randomForest' package. For the hyperparameters values, I choose number of trees (ntree) to be 500 and number of variables randomly sampled at each split (mtry) to be $s = \sqrt{\dim(X)}$. Finally, Neural Nets is implemented using 'neuralnet' package. I choose the number of hidden layers in the network to be one. For that hidden layer, we let the number of hidden units to be 3. Different activation functions are imposed for Neural Nets, such as $\text{logistic}(x) = 1/(1 + exp(-x))$, $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$, and $\text{softplus}(x) = log(1 + exp(x))$.

The second step estimation for the 1-dimensional endogenous social interaction effect, β_0, are constructed using the moment conditions (2.2.4) or orthogonal moment conditions (2.2.6) after plugging the first step nonparametric estimators $\widehat{\vartheta} = (\widehat{\mu}, \widehat{\nu}, \widehat{\phi})$.

For univariate control variable case, i.e. $\dim(X_i) = 1$, as discussed in Section 2.4, β_0

is estimated using exactly one moment function, either the original score (2.2.4) or the orthogonal score (2.2.6). The estimators which are denoted by $\widehat{\beta}$ and $\widehat{\beta}_{db}$ are defined in (2.4.7) and (2.4.16). For the multivariate control variables case, i.e. $\dim(X_i) > 1$ and fixed, we apply a semiparametric GMM estimator for β_0. The original and debiased estimators which are denoted by $\widehat{\beta}^{gmm}$ and $\widehat{\beta}_{db}^{gmm}$ are defined in (2.4.9) and (2.4.18) respectively.

2.5.1.3 Performance measure

Our Monte Carlo simulation study focuses on the performance comparison of semiparametric estimators for β_0 using different first step Machine Learning methods. This paper uses the following measures to evaluate the performance across different methods: Bias, Variance, mean squared error (MSE), mean absolute error (MAE) for estimators of β_0 and empirical size for the corresponding t test. These performance measures are defined as follows:

$$
\begin{aligned}
\text{Bias} &= \frac{1}{S}\sum_{s=1}^{S}\widehat{\beta}_s - \beta_0 \\
\text{VAR} &= \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\beta}_s - \overline{\widehat{\beta}}_s\right)^2 \\
\text{MSE} &= \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\beta}_s - \beta_0\right)^2 = \text{Bias}^2 + \text{VAR} \\
\text{MAE} &= \frac{1}{S}\sum_{s=1}^{S}|\widehat{\beta}_s - \beta_0| \\
\text{Size} &= \sum_{s=1}^{S}\mathbf{1}\left(|t| > z_\alpha\right)/S
\end{aligned}
$$

where S is the number of simulation repetitions. MSE is a practical measure of performance widely used for estimator evaluation, smaller values are preferred. However, MSE has the disadvantage that gives heavily weighting on outliers by squaring each error term. Like variance, MSE does not exist under some cases which are known as the moment problem (Sargan 1978). In the simulation studies, I also report the more robust performance measure

MAE for estimator evaluation under such cases.

To evaluate the performance of t test for β_0,

$$H_0 : \beta = \beta_0 \quad \text{vs.} \quad H_1 : \beta \neq \beta_0,$$

this simulation study reports the empirical size of the t test with nominal size α equal to 5% and 10% (size5% and size10%). The empirical size which measures the proportion of rejection of H_0 in the Monte Carlo simulation should close to the nominal size of α.

2.5.2 Simulation Results

2.5.2.1 Univariate case, $\dim(X) = 1$

We first consider the univariate control variable case, $\dim(X) = 1$. For the linear function $h_0(x) = x\gamma$, we consider four cases, including bivariate normal and bivariate logistic distribution of $X_{g,1}, X_{g,2}$ with $\gamma = 1$ or 0.1, respectively. We also consider the nonlinear function $h_0(x) = \frac{1}{1+e^{-x\gamma}}$, $h_0(x) = e^{x\gamma}$, and $h_0(x) = sin(x\gamma) + cos(x\gamma)$. Table 2.3-Table 2.6 report the performance measures across different estimation methods with linear function form. Table 2.7-2.12 report the results for the nonlinear cases. In the following discussion, we first point out the possible moment problem under the univariate control variable case, and then discuss the performance comparison of the original and debiased estimators ($\widehat{\beta}$ and $\widehat{\beta}_{db}$) given first step estimators. Finally, we focus on the debiased estimator $\widehat{\beta}_{db}$ and compare their performances across different first step estimators applied.

Moment problem and weak identification

From our simulation results in Table 2.5, we find that there might have moment problem under the univariate control variable case. Sargan (1978) shows that the standard two-stage least square (2SLS) estimators possess finite k-th moments if $k \geq d_x d_y - d_z$ where d_x is the number of exogenous regressors, d_y is the number of endogenous variables and d_z is the number of instrumental variables. For our social interaction model with univariate control,

$d_x = \dim(X_{g,i}) = 1$, and $d_y = 2$ for the two endogenous variables $Y_{g,i}$ and $\overline{Y}_{g,-i}$. Assume we use $\overline{X}_{g,-i}$ as IV for $\overline{Y}_{g,-i}$, then $d_z = 1$ which implies that $d_x d_y - d_z = 1$. Thus our MM estimator for β_0 has finite moment up to order 1 and the second moment (Variance or MSE) does not exist.

Consider the simulation results for linear bivariate normal cases with $\gamma = 1$ and $\gamma = 0.1$ reported in Table 2.3 and Table 2.5. Given $\gamma = 0.1$, the VAR and MSE for $\widehat{\beta}$ are all relatively large which is an indicator for the nonexistence of the second moment. In this case, we also report the MAE measure to evaluate model performance. The moment problem is obvious only for the case that $\gamma = 0.1$ (Table 2.5) but not for $\gamma = 1$ (Table 2.3). From the reduced form model, we know that γ is a measure for the strength of the instrumental variable $\overline{X}_{g,-i}$. The smaller γ, the weaker the instrumental variable. We also reported the corresponding first-step F statistics for the semiparametric two step estimator for β_0, which is an empirical measure for the strength of the instruments. From the last line of Table 2.3 and 2.5, we can conclude that under $\gamma = 0.1$, F statistics is small and $\overline{X}_{g,-i}$ tend to be a weak IV for $\overline{Y}_{g,-i}$. Thus, the moment problem is reinforced and becomes obvious under the weak identification case.

The simulation results for linear bivariate logistic cases with $\gamma = 1$ and $\gamma = 0.1$ which are reported in Table 2.4 and Table 2.6 have the similar features as the bivariate normal cases. However the moment problem is not that severe compared with the bivariate normal case. The reason is that the first step F-statistics in Table 2.6 is larger than 10, which indicates the weak identification might not be a problem. Thus, the moment problem is not that severe under the bivariate logistic case.

Besides, if the nuisance parameter $h_0(x)$ is nonlinear, the moment problem might also exist. For example, consider the simulation result for $h_0(x) = 1/(1 + exp(-x))$ that is shown in Table 2.7 and Table 2.8, the conclusion is similar to the linear case which shows that moment problem exists and become obvious under the weak identification case.

Although there is a moment problem for the semiparametric estimator when $\gamma = 0.1$ for linear model and also the nonlinear case with $h_0(x) = 1/(1 + exp(-x))$, the empirical size for the two-sided t-test still has the correct size but is undersized in general.

Plug-in V.S. Debiasing estimator for β_0

As shown in Section 1.5, if we plugging the first step series estimator, the second step debiased estimator $\widehat{\beta}_{db}$ is the same with the plug-in estimator $\widehat{\beta}$. More generally, the variables used in the first step are not restricted to series expansion but we can allow for any variables included as long as the estimation of the first step conditional expectations use the same set of regressors. Thus the first step Post-LASSO selection (P.L.S) estimators should also satisfy this property.

This statement is verified in our Monte Carlo simulation result. From Table 2.3-Table 2.12, it can be learned that for the first step linear, series, Post-LASSO selection (P.L.S) estimators, the second step plug-in estimators and debiased estimators for β_0 have exactly the same results.

For the other first step Machine Learning estimators, such as Post-LASSO (P.L), Random Forest (R.F.) and Neural Nets (N.N), the debiased estimator is different from the plug-in estimator and there are no consistent results that which one performs better. Theoretically, debiased estimators should have smaller bias which also holds in our simulation results in general. However, the additional estimation of $\phi_0(X_{g,i}) = E[X_{g,-i}|X_{g,i}]$ brings some noise into the second step which might lead to an increase in the variance of $\widehat{\beta}$ thus the comparison of MSE becomes unclear. For example, from Table 2.11 and 2.12 we know, if $h_0(x) = sin(x) + cos(x)$, the debiased estimator with the first step Neural Nets performs better than the plug-in estimator in terms of MSE. However, for the results of $h_0(x) = exp(0.5x)$ in Table 2.9 and 2.10, if we apply Neural Nets in the first step, the debiased estimator performs even worse.

Comparison across different first step estimators

From the simulation results, we can learn that in general the parametric linear estimator only has the correct size if the true model is linear but have an over-rejection problem if applied to the nonlinear models. The results indicate that using linear estimators is likely to obtain a significant social interaction effect which is spurious. We can also learn from the simulation result that the distortion is more server if the true function becomes more curvature. However, all the semiparametric debiased estimators do not have such an issue and have the correct size across the nonlinear models considered in the simulation study.

To be specific, first consider the simulation results for weak identification case in linear model with $\gamma = 0.1$ in Table 2.5 and 2.6. We focus on the performance comparison of the debiased estimator for $\widehat{\beta}_{db}$ by plugging in different first step nonparametric estimators. Under this case, MSE can not be a good measure since the second moment for $\widehat{\beta}_{db}$ does not exist. We use MAE and empirical size to evaluate the performance. Among the debiased estimators, $\widehat{\beta}_{db}$ with the first step Random Forest has proper size control under this case but due to the large variance of the estimator, the results might not be reliable. For the other first step estimators, they are all undersized.

The simulation results for the linear model with $\gamma = 1$ in Table 2.3 and 2.4 are relatively more well-behaved for all the estimation methods applied. All the debiased estimators for β_0 under this case have the proper empirical size and moderate MSE. Among them, $\widehat{\beta}_{db}$ with the first step linear estimator performs the best in terms of both MAE and empirical size. The reason is that the true model is linear and apply the linear regression in the first step obtains the oracle estimator. The estimator with series, Post-LASSO selection (P.L.S), Neural Nets have similar performance with the linear estimator under this case. Random Forest is slightly worse since it is more fitted for the nonlinear and many control variables case.

For the nonlinear models considered in the simulation in Table 2.7- Table 2.12, the semiparametric series estimator performs the best in terms of both MSE and empirical size in general. The reason is that for $\dim(X) = 1$, the nonlinear function $h_0(x)$ could be well-approximated by its series expansions without curse of dimensionality. Besides, the Post-LASSO debaised estimator and Neural Nets debiased estimator have the similar performance which are also recommended in the empirical studies.

2.5.2.2 Multivariate case, $\dim(X) = 3$

In this part, we consider the more realistic case that $\dim(X_i) > 1$, here we assume $\dim(X_i) = 3$. In order to have a better understanding of the performance of different estimators under different DGPs, we consider the nuisance function $h_0(x)$ can be linear or nonlinear forms. Table 2.13 - Table 2.16 report the performance measures for $\widehat{\beta}$ and $\widehat{\beta}_{db}$ across different first step estimation methods with linear function form $h_0(x) = x\gamma$. Table 2.17 - Table 2.22 report the results for the nonlinear function, $h_0(x) = \frac{1}{1+e^{-x\gamma}}$, $h_0(x) = e^{x\gamma}$, and $h_0(x) = sin(x\gamma) + cos(x\gamma)$. Similar to the univariate case, in the following discussion, we first discuss the possible weak identification issue, and then compare the performance of the plug-in estimator ($\widehat{\beta}^{gmm}$) and debiased estimator ($\widehat{\beta}_{db}^{gmm}$) given different first step estimators. Finally, we will consider to evaluate the performance of the debiased estimators using different first step nonparametric methods.

Weak identification

We can show that the second moment of estimator for β_0 exists under $\dim(X) = 3$ since $d_x d_y - d_z = 3 * 2 - 3 = 3$. This is verified in our simulation results from Table 2.13 - Table 2.22. The simulation results shows the second moment (VAR or MSE) is bounded across all the cases. However, weak identification might still be an issue and thus t-test has incorrect size under these cases. For example, for linear function with $\gamma = 0.1$ which are reported in Table 2.15 and Table 2.16, the F-stat are relatively low. The variance of these estimators are large compared with the strong identified cases. Besides, for the nonlinear case $h_0(x) = 1/(1 + e(-x))$ reported in Table 2.17, the estimator for β_0 also suffers from the same weak identification problem. One reason for the weak identification issue is because the correlation between the control variables within groups ($\rho = 0.5$ or $r = 0.5$). Thus the residualized instrumental variable is weakly correlated with the endogenous social interactions.

Original V.S. Debiasing estimator for β_0

As shown in Section 1.5, if we plugging the first step series estimator, the second step debiased estimator $\widehat{\beta}_{db}^{gmm}$ is the same with the plug-in estimator $\widehat{\beta}^{gmm}$. This is also true for the Post-LASSO (double) selection estimators which is verified in our Monte Carlo simulation results. For the other first step nonparametric estimators, such as LASSO, Random Forest (R.F.) and Neural Nets (N.N), our simulation results suggest that debiased estimators outperform the plug-in estimators in terms of both MSE and empirical size in general. Thus, estimate β_0 using the orthogonal moment condition instead of the original moment condition in the second step is recommended in the empirical studies.

Comparison across different first step estimators

To compare the performance across different first step estimators, we still focus on the debiased estimator $\widehat{\beta}_{db}$ under the strong identified case. Similar to the univariate strong identification linear cases, in terms of both the MSE and empirical size (5% and 10%), $\widehat{\beta}_{db}^{gmm}$ with first step linear regression estimator performs the best since linear regression in the first step can achieve the oracle estimator. All the other semiparametric estimators given the linear model are all have good performance in terms of both MSE and empirical size.

For the nonlinear cases reported in Table 2.18 - Table 2.22, Post-LASSO debiased estimators and Neural Nets debiased estimators performs the best and they are stable across different DGPs. Random Forest is too risky since it has poor performance under the linear case especially with few regressors.

Simulation results summary

(i) Orthogonal score helps in the performance of two-step estimators for β_0. That is in general $\widehat{\beta}_{db}^{gmm}(\widehat{\beta}_{db})$ outperforms $\widehat{\beta}^{gmm}(\widehat{\beta})$ across different first step estimation methods and different DGPs in the terms of performance measures such as MSE and empirical size.

(ii) The performance of debiased estimators $\widehat{\beta}_{db}^{gmm}(\widehat{\beta}_{db})$ with first step Post-LASSO or Neural Nets estimators are relatively stable across different DGPs and is recommended in the empirical studies.

(iii) In general, the estimators $\widehat{\beta}_{db}$ with first step Random Forest is too risky since its performance is very unstable. But Random Forest methods are more robust to the complexity of the nuisance function $h_0(x)$.

(iv) Neural Nets has several tuning parameters and there are still no theoretical results on how to choose them. And the performance of the second step estimator $\widehat{\beta}$ is sensitive to the choice of some tuning parameters such as the number of hidden layers, size for each layer and also the activation function applied.

2.6 Empirical Example

To illustrate the semiparametric methods developed in the preceding sections, this section considers an empirical example which investigates the endogenous classmates' peer effect on student's performance.

2.6.1 Data and Descriptive Statistics

The data used in this study is from wave 1 (2013-2014) of the China Education Panel Survey (CEPS), which provides large-scale, nationally representative, longitudinal survey datasets. In 2013-2014, CEPS surveys 19,487 students at the 7th and 9th grade from 112 schools in the 2013-2014 academic year, and also survey information from their parents, teachers, and principals.

In this empirical study, I focus on the data for students at the 7th grade only. After dealing with the missing values, the sample includes 7442 students across 167 classes. Table 2.26 summaries the variables used in this empirical study, and Table 2.27 shows the summary statistics of these variables.

The outcome variables I am interest in include the cognitive test score for the student and their level of self confidence. The cognitive test was conducted by the CEPS which included 20 questions. The cognitive test score reported in the dataset was then standardized using the Item Response Theory (IRT) model. Another outcome variable, level of self confidence, was reported by students themselves. The level of self confidence was from $1 \sim 4$, in which 1 denoteed for not confident at all and 4 denoted for very confident. The parameter of interest is the effect of class average cognitive test score or level of self confidence on student's own outcomes.

The dependent variables include the individual only effect X, the individual and contextual effect (W and $\overline{W}$), and also the group effect Υ. For the individual effect, I include the variable whether the students attended kindergarten or not. This variable might determine the student's IQ level, and thus can affect their school performance. In contrast, childhood characteristics of a student's classmates are unlikely to have a direct effect on that student's high school scores. Thus, average class preschool attendance ratio is excluded for the contextual factors and can be used as instrumental variables for the endogenous variable of the average class cognitive score or level of self confidence.

Besides, I also include students' age, gender, one child family or not, minority, hukou status, migrant status, parents' education level, family income as the individual effect W and also their class average $\overline{W}$ as the contextual effect. For the group controls Υ, I consider to use teacher's age, gender, years of schooling, experience, and also tracking school or not.

2.6.2 Data Exploration

For the purpose of data exploration, I plotted the correlation heatmap for the variables discussed in the preceding subsection.

Figure 2.1 plots the correlation map between endogenous variables (cognitive test score and level of self confidence) and the instrumental variables (whether attended preschool or not). From the correlation plot, it can be learned that IV has relatively high correlation with the endogenous variables, which have preliminarily implications that the strength of the IV is relatively strong.

Figure 2.2 plots the correlation map of the control variables $X, W, \overline{W}, \Upsilon$. From the plot, it can be learned that class average father's education level is highly positive correlated with class average family income, class average one child ratio, and father's education. In order to avoid the multicollinearity problem, in the empirical study I remove the average father's education level. And the correlation heatmap becomes Figure 2.3. There is no any strong correlation (absolute value greater than 0.7) any more.

I also plot the correlation map for the controls of only X and W in Figure 2.4. It can be learned from the Figure 2.4 that the correlation among these variables are all relatively low (absolute value smaller than 0.4). The empirical study also considers this case to learn the effect of omitting the contextual variables on the estimation of endogenous social classmates' effect.

2.6.3 Estimation Results

Table 2.28 - Table 2.29 reports the estimation results for the endogenous classmates' effect. I apply the debiased semiparametric estimators with first step linear, series, Post-LASSO, Random Forest, and Neural Nets. The first three columns report the results for only including the individual controls (X, W). The dimension of the original controls $d = 7$ and I use series expansion which include $p = 31$ terms. The second three columns report the results for including the controls of both the individual and contextual effect $(X, W, \overline{W}, \Upsilon)$. The dimension of the controls becomes $d = 14$ and the series terms is $p = 155$.

Table 2.28 shows the estimation results for the endogenous classmates' effect on student's cognitive score. It can be learned that if only the individual controls (X, W) are included, the endogenous classmates' effect on cognitive score is significant. However, after controlling the contextual effect $(\overline{W}, \Upsilon)$, the endogenous classmates' effect becomes insignificant. The magnitude of the debiased estimator for β_0 is different across first step estimation methods applied, but they have the same insignificant conclusion. This empirical result shows that omitting contextual effect might lead to spurious endogenous peer effect.

Table 2.29 reports the estimation results for the endogenous classmates' peer effect on student's level of self confidence. For the empirical results with only the individual controls (X, W), the classmates' level of self confidence also has a significant effect on student's level of self confidence. If the contextual effect $(\overline{W}, \Upsilon)$ are included, the measure of the endogenous classmates' effect becomes different across various first step estimators applied. For first step linear, series or Post-LASSO estimator, the endogenous effect becomes insignificant. Since the linear estimator might suffer from the model misspecification problem and series estimator has the curse of dimensionality problem ($p = 115$), the results based on these first step estimators might be problematic. If the Random Forest or Neural Nets is applied, the endogenous effect is still significant. Also, these two debiased estimators indicate a large $(0.8501, 0.8928)$ significant peer effect on student's level of self confidence.

2.7 Conclusion and Future Research

The existing literature on the social interaction model focuses on the identification issue of the parametric setup. However, the parametric model may be too restrictive and might lead to a spurious or misleading social interaction effect in the empirical studies. This paper studies a semiparametric social interaction model with a parametric linear-in-means endogenous social interaction part and a nonparametric control variables part. To highlight the semiparametric feature of the model, this paper excludes the correlated effect and also imposes restrictions on the contextual effect to avoid the complexity of identification issues.

This paper adopts a semiparametric IV approach to identify the endogenous social interaction effect. Based on the identification condition, I then propose a semiparametric two-step estimator in which the first step nuisance functions could be estimated by any nonparametric (Machine Learning) estimators, such as LASSO, Random Forest and Neural Nets, and then the second step parametric components are estimated by MM/GMM. This paper also considers using the orthogonal moment condition in the second step estimation to reduce the bias induced by the first step Machine Learning estimators.

For the first step of Machine Learning methods, there are no theoretical results to show the performance comparison across different methods applied. This paper uses Monte Carlo simulations to investigate the finite sample performance of our two-step estimators using various first-step nonparametric estimation methods with different DGPs considered. The Monte Carlo simulation results suggest that no estimation method dominates across all the Data Generating Processes (DGPs) considered. However, it is also reflected in the simulation results that the debiased estimators using first step post-LASSO or Neural Nets methods are more reliable and performs relatively well across the settings considered. For this reason, these two debiased estimators are recommended for use in empirical studies.

The discussed group social interaction model can also be extended to a more flexible network social interaction model. The network depicts the connections between individuals and does not need to have the group structural which can be applied to a much richer social structures. Also, the network structural impose certain exclusion restrictions on the model and make the identification to be easier (Bramoullé, Djebbari, and Fortin (2009)). Based on the identification condition, the semiparametric two-step estimation methods can also be applied to the social interaction model with networks. I will consider extending our results for the semiparametric group social interaction model to the semiparametric network social interaction model in future research.

It would also be interesting to apply the cross-fitting methods for the semiparametric estimators to further eliminate the overfitting bias introduced by the first step Machine Learning estimators. It be considered in future research.

Appendix

2.A Simulation Results

DGP1-1-1/2: $\dim(X) = 1$**; Linear** $h(x) = x\gamma$**;** $\gamma = 1$**.**

Table 2.3: $\dim(X) = 1$; Linear $h(x) = x\gamma$; $\gamma = 1$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.0403	-0.0407	-0.0491	-0.0395	-0.0415	0.0006	-0.1068	-0.0369	-0.0423
Vars	0.2943	0.2955	0.2948	0.2958	0.2956	0.4689	0.5166	0.2942	0.2931
MSE	0.2959	0.2972	0.2972	0.2973	0.2973	0.4689	0.5280	0.2955	0.2949
MAE	0.4256	0.4268	0.4273	0.4282	0.4278	0.5438	0.5675	0.4269	0.4266
Size5%	0.0570	0.0580	0.0560	0.0570	0.0550	0.0840	0.0970	0.0530	0.0530
Size10%	0.1240	0.1170	0.1160	0.1130	0.1120	0.1490	0.1630	0.1160	0.1180
F-stats	431.3731	430.9149	430.3033	429.4495	429.4495	403.1003	403.1003	429.5267	429.5267

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with correlation $\rho = 0.5$
*The estimators (Linear, Series, P.L.S, P.L.,R.F., N.N.) are described in Table 2.2 of Section 2.5.1.2
*The performance measure (Bias, Vars, MSE, MAE, Size5%, Size10% are described in Section 2.5.1.3)
*F-stat: F-stats of first stage regression in semiparametric two-step estimator of β_0:
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.4: $\dim(X) = 1$; Linear $h(x) = x\gamma$; $\gamma = 1$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.0130	0.0146	0.0100	0.0324	0.0166	0.3247	-0.0168	0.0187	0.0130
Vars	0.1922	0.1971	0.1961	0.2002	0.1953	0.3897	0.3156	0.1970	0.1962
MSE	0.1924	0.1973	0.1963	0.2013	0.1955	0.4951	0.3159	0.1974	0.1964
MAE	0.3513	0.3548	0.3544	0.3575	0.3535	0.5540	0.4420	0.3555	0.3545
Size5%	0.0470	0.0560	0.0550	0.0600	0.0530	0.1900	0.1040	0.0520	0.0530
Size10%	0.0970	0.1020	0.1020	0.1010	0.0980	0.2630	0.1770	0.1020	0.1020
F-stats	1165.3027	1143.7756	1144.6131	1139.2702	1139.2702	1054.1264	1054.1264	1137.9467	1137.9467

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP1-1-1/2: $\dim(X) = 1$**; Linear** $h(x) = x\gamma$**;** $\gamma = 0.1$**.**

Table 2.5: $\dim(X) = 1$; Linear $h(x) = x\gamma$; $\gamma = 0.1$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	5.6967	-2.7684	-16.4474	-0.8493	-2.8973	13.0369	20.0722	14.0447	-0.6402
Vars	58741	594	258749	2715	670	208866	393283	259791	3312
MSE	58774	602	259019	2716	679	209036	393686	259988	3313
MAE	13.6254	6.7950	24.2230	8.8329	6.9531	26.4334	33.9162	22.2069	8.1225
Size5%	0.0320	0.0330	0.0450	0.0360	0.0340	0.0730	0.0810	0.0320	0.0310
Size10%	0.0650	0.0660	0.0850	0.0690	0.0690	0.1200	0.1150	0.0670	0.0650
F-stats	4.8020	4.8064	4.8782	4.7972	4.7972	4.8514	4.8514	4.8032	4.8032

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.6: $\dim(X) = 1$; Linear $h(x) = x\gamma$; $\gamma = 0.1$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.4451	0.4335	0.3844	0.3925	0.4394	0.6282	0.1649	0.6737	0.4512
Vars	22.8605	23.4415	23.2794	23.4601	23.2211	46.0983	48.4058	22.6058	22.9960
MSE	23.0586	23.6294	23.4272	23.6142	23.4142	46.4930	48.4330	23.0597	23.1995
MAE	3.6963	3.7331	3.7256	3.7422	3.7267	4.6678	4.9146	3.6573	3.6985
Size5%	0.0340	0.0290	0.0290	0.0300	0.0290	0.0570	0.0650	0.0290	0.0320
Size10%	0.0890	0.0910	0.0890	0.0910	0.0910	0.1260	0.1310	0.0800	0.0850
F-stats	11.9591	11.8409	11.7790	11.8056	11.8056	11.9157	11.9157	11.8622	11.8622

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP1-2-1/2: $\dim(X) = 1$**; Nonlinear** $h(x) = 1/(1 + exp(-x))$**.**

Table 2.7: $\dim(X) = 1$; Nonlinear $h(x) = 1/(1 + exp(-x))$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.7942	-0.7996	-0.8469	-0.9116	-0.7824	-1.0005	-1.8059	-0.8400	-0.7990
Vars	22.2020	22.3110	85.7675	86.1319	22.1642	181.2936	325.3122	22.3462	22.2960
MSE	22.8328	22.9504	86.4848	86.9630	22.7764	182.2946	328.5734	23.0518	22.9344
MAE	3.2655	3.2638	3.3753	3.4416	3.2573	5.1020	6.0717	3.2792	3.2599
Size5%	0.0360	0.0340	0.0380	0.0400	0.0350	0.0650	0.0740	0.0320	0.0310
Size10%	0.0780	0.0740	0.0800	0.0780	0.0770	0.1000	0.1120	0.0740	0.0750
F-stats	9.6773	9.7062	9.7208	9.6887	9.6887	9.3972	9.3972	9.6719	9.6719

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.8: $\dim(X) = 1$; Nonlinear $h(x) = 1/(1 + exp(-x))$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.1007	0.4812	0.4222	0.4835	0.4893	0.4835	0.2075	0.4061	0.4813
Vars	28.5679	26.4348	26.6546	27.5114	26.2572	74.7801	65.1258	26.7854	26.3256
MSE	28.5780	26.6663	26.8329	27.7452	26.4966	75.0139	65.1689	26.9504	26.5572
MAE	4.1067	3.9441	3.9599	4.0021	3.9462	5.2375	5.3432	3.9768	3.9416
Size5%	0.0440	0.0280	0.0280	0.0280	0.0250	0.0530	0.0580	0.0290	0.0260
Size10%	0.0920	0.0880	0.0900	0.0920	0.0850	0.1220	0.1280	0.0820	0.0790
F-stats	9.8136	10.7349	10.6365	10.6620	10.6620	10.6785	10.6785	10.6396	10.6396

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP1-3-1/2: $\dim(X) = 1$**; Nonlinear** $h(x) = exp(0.5x)$**.**

Table 2.9: $\dim(X) = 1$; Nonlinear $h(x) = exp(0.5x)$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.0864	-0.0849	-0.0900	-0.0884	-0.0891	0.0103	-0.1441	-0.0839	-0.1020
Vars	1.1419	0.9565	0.9565	0.9570	0.9569	1.5702	1.7260	0.9463	0.9444
MSE	1.1494	0.9637	0.9646	0.9648	0.9649	1.5703	1.7468	0.9533	0.9548
MAE	0.8516	0.7636	0.7648	0.7663	0.7656	0.9915	1.0259	0.7639	0.7625
Size5%	0.0530	0.0550	0.0520	0.0530	0.0540	0.0890	0.0950	0.0500	0.0510
Size10%	0.1160	0.1090	0.1080	0.1060	0.1060	0.1510	0.1630	0.1040	0.1070
F-stats	128.5615	131.4096	131.3602	131.0863	131.0863	124.5216	124.5216	130.7998	130.7998

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.10: $\dim(X) = 1$; Nonlinear $h(x) = exp(0.5x)$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	1.8948	0.0162	-0.0141	-0.0841	-0.0456	1.7581	-0.0819	-0.0943	-0.1330
Vars	3.3856	0.3260	3.5872	2.2782	0.3701	4.1334	1.3812	0.4443	0.4806
MSE	6.9756	0.3263	3.5874	2.2853	0.3722	7.2242	1.3879	0.4532	0.4983
MAE	2.0918	0.4492	0.6119	0.7108	0.4773	1.9047	0.8199	0.5102	0.5159
Size5%	0.3890	0.0510	0.0650	0.0980	0.0440	0.4070	0.1060	0.0640	0.0670
Size10%	0.4840	0.0980	0.1260	0.1660	0.1040	0.4900	0.1840	0.1160	0.1230
F-stats	204.4779	332.5040	330.1172	325.6427	325.6427	305.8845	305.8845	331.6955	331.6955

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP1-4-1/2: $\dim(X) = 1$**; Nonlinear** $h(x) = sin(x) + cos(x)$**.**

Table 2.11: $\dim(X) = 1$; Nonlinear $h(x) = sin(x) + cos(x)$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.0864	-0.0787	-0.0812	-0.0795	-0.0804	-0.2019	-0.1787	-0.1185	-0.0904
Vars	1.6154	0.8040	0.8083	0.8098	0.8058	1.3186	1.4459	0.8369	0.8378
MSE	1.6228	0.8102	0.8149	0.8161	0.8123	1.3594	1.4778	0.8509	0.8460
MAE	1.0014	0.7006	0.7023	0.7035	0.7016	0.9122	0.9392	0.7175	0.7166
Size5%	0.0410	0.0510	0.0520	0.0510	0.0500	0.0720	0.0930	0.0550	0.0560
Size10%	0.0900	0.1080	0.1100	0.1060	0.1040	0.1360	0.1540	0.1080	0.1060
F-stats	114.2840	128.6217	128.4820	127.8354	127.8354	120.5450	120.5450	127.0486	127.0486

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.12: $\dim(X) = 1$; Nonlinear $h(x) = sin(x) + cos(x)$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-5.5566	0.0449	0.0654	0.0735	0.0710	0.3630	-0.0761	0.5614	0.0101
Vars	12.4722	1.2731	1.3359	1.4038	1.3885	2.2872	2.2777	1.8692	1.3956
MSE	43.3478	1.2751	1.3402	1.4092	1.3936	2.4190	2.2835	2.1844	1.3957
MAE	5.7222	0.8813	0.8989	0.9226	0.9156	1.1784	1.1299	1.1518	0.9245
Size5%	0.2890	0.0340	0.0350	0.0360	0.0360	0.0840	0.0820	0.0860	0.0340
Size10%	0.4430	0.0800	0.0830	0.0840	0.0800	0.1750	0.1470	0.1460	0.0830
F-stats	25.6250	90.2842	89.7199	88.8940	88.8940	85.8868	85.8868	88.6245	88.6245

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP1-1-1/2: $\dim(X) = 3$**; Linear** $h(x) = x\gamma$**;** $\gamma = 1$**.**

Table 2.13: $\dim(X) = 3$; Linear $h(x) = x\gamma$; $\gamma = 1$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.0231	0.0199	-0.0197	0.0197	0.0216	1.8247	0.4227	0.0398	0.0336
Vars	0.2714	0.2891	0.2842	0.2934	0.2908	0.3689	0.3794	0.2795	0.2835
MSE	0.2719	0.2895	0.2846	0.2937	0.2913	3.6984	0.5581	0.2810	0.2846
MAE	0.4184	0.4322	0.4285	0.4378	0.4332	1.8259	0.6107	0.4268	0.4294
Size5%	0.0680	0.0670	0.0620	0.0590	0.0610	0.8610	0.1410	0.0650	0.0580
Size10%	0.1140	0.1200	0.1170	0.1200	0.1210	0.9080	0.2180	0.1170	0.1180
F-stats	158.5402	158.6238	157.6166	154.0685	154.0685	152.3965	152.3965	155.2407	155.2407

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.14: $\dim(X) = 3$; Linear $h(x) = x\gamma$; $\gamma = 1$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.0106	0.0114	-0.0116	-0.0020	0.0026	4.0767	0.9923	0.0532	0.0348
Vars	0.1400	0.1475	0.1437	0.1568	0.1529	0.4412	0.3106	0.1468	0.1452
MSE	0.1401	0.1476	0.1438	0.1568	0.1529	17.0609	1.2953	0.1496	0.1464
MAE	0.3003	0.3053	0.3026	0.3138	0.3110	4.0767	1.0021	0.3093	0.3043
Size5%	0.0540	0.0560	0.0560	0.0600	0.0600	1.0000	0.3950	0.0580	0.0600
Size10%	0.0990	0.1010	0.1000	0.1140	0.1040	1.0000	0.5330	0.1000	0.0970
F-stats	531.3653	526.5358	522.7558	500.3013	500.3013	469.4460	469.4460	496.0114	496.0114

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP1-1-1/2: $\dim(X) = 3$**; Linear** $h(x) = x\gamma$**;** $\gamma = 0.1$**.**

Table 2.15: $\dim(X) = 3$; Linear $h(x) = x\gamma$; $\gamma = 0.1$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.2682	0.2758	0.4116	0.2448	0.2996	1.5027	0.5947	0.2644	0.3231
Vars	6.8963	7.2790	6.8239	7.6992	7.2056	7.2106	9.2723	7.1767	6.9271
MSE	6.9682	7.3551	6.9933	7.7591	7.2953	9.4687	9.6260	7.2466	7.0315
MAE	2.0957	2.1500	2.1145	2.2288	2.1390	2.5376	2.4288	2.1376	2.1141
Size5%	0.0680	0.0800	0.0770	0.0790	0.0750	0.1700	0.1060	0.0670	0.0710
Size10%	0.1150	0.1250	0.1240	0.1350	0.1280	0.2230	0.1710	0.1220	0.1210
F-stats	6.8716	6.9039	6.8554	6.8079	6.8079	7.0396	7.0396	6.8850	6.8850

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.16: $\dim(X) = 3$; Linear $h(x) = x\gamma$; $\gamma = 0.1$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.0907	-0.0544	-0.0336	-0.2022	-0.0221	2.8931	0.7712	0.0609	0.0177
Vars	14.5380	15.2292	15.3329	16.2284	14.9188	12.8231	18.7303	14.4723	14.4979
MSE	14.5462	15.2322	15.3340	16.2693	14.9193	21.1930	19.3251	14.4760	14.4982
MAE	3.0325	3.0613	3.0712	3.1745	3.0303	3.6899	3.4444	2.9973	2.9885
Size5%	0.0490	0.0470	0.0480	0.0570	0.0470	0.0920	0.0720	0.0460	0.0450
Size10%	0.0920	0.0920	0.0930	0.1120	0.0940	0.1670	0.1300	0.0870	0.0910
F-stats	5.6354	5.6820	5.5532	5.6053	5.6053	6.1033	6.1033	5.6483	5.6483

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP3-2-1/2: $\dim(X) = 3$**; Nonlinear** $h(x) = 1/(1 + exp(-x))$**.**

Table 2.17: $\dim(X) = 3$; Nonlinear $h(x) = 1/(1 + exp(-x))$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.2635	0.2633	0.4129	0.2391	0.3097	1.3829	0.5435	0.2105	0.3165
Vars	6.5846	6.9436	6.5447	7.3677	6.8657	7.0933	8.9338	6.8431	6.6227
MSE	6.6541	7.0129	6.7151	7.4249	6.9617	9.0056	9.2292	6.8874	6.7228
MAE	2.0519	2.1016	2.0727	2.1800	2.0909	2.4541	2.3784	2.0790	2.0688
Size5%	0.0680	0.0790	0.0740	0.0770	0.0700	0.1550	0.1140	0.0720	0.0690
Size10%	0.1160	0.1260	0.1210	0.1230	0.1310	0.2120	0.1660	0.1180	0.1160
F-stats	7.1158	7.1759	7.1261	7.0777	7.0777	7.2579	7.2579	7.1343	7.1343

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.18: $\dim(X) = 3$; Nonlinear $h(x) = 1/(1 + exp(-x))$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.4052	-0.0217	-0.0576	-0.0124	0.0200	2.5467	0.5212	0.0491	-0.0028
Vars	4.9630	4.6777	4.6396	4.6873	4.6520	3.9928	5.7928	4.7329	4.6676
MSE	5.1272	4.6781	4.6429	4.6875	4.6524	10.4787	6.0644	4.7353	4.6676
MAE	1.8154	1.7108	1.7067	1.7073	1.7049	2.7370	1.9517	1.7268	1.7164
Size5%	0.0620	0.0520	0.0610	0.0580	0.0600	0.1930	0.0760	0.0510	0.0540
Size10%	0.1140	0.1000	0.0990	0.0970	0.0940	0.3100	0.1400	0.1000	0.0980
F-stats	15.6798	17.0872	16.8920	16.7916	16.7916	16.9203	16.9203	16.5507	16.5507

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP3-3-1/2: $\dim(X) = 3$**; Nonlinear** $h(x) = exp(0.5x)$**.**

Table 2.19: $\dim(X) = 3$; Nonlinear $h(x) = exp(0.5x)$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.0302	0.0462	0.0034	0.0114	0.0179	1.7698	0.4239	0.0942	0.0499
Vars	1.0484	0.9021	0.9061	0.9150	0.9116	1.0014	1.1631	0.9099	0.8985
MSE	1.0493	0.9042	0.9061	0.9151	0.9120	4.1336	1.3428	0.9188	0.9010
MAE	0.8184	0.7628	0.7652	0.7690	0.7651	1.8159	0.9313	0.7631	0.7596
Size5%	0.0700	0.0660	0.0620	0.0620	0.0660	0.5010	0.1180	0.0690	0.0640
Size10%	0.1250	0.1260	0.1200	0.1250	0.1230	0.6010	0.1720	0.1240	0.1250
F-stats	47.8174	49.3396	49.1973	48.1536	48.1536	48.9535	48.9535	48.3920	48.3920

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.20: $\dim(X) = 3$; Nonlinear $h(x) = exp(0.5x)$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.9225	-0.0069	-0.0453	-0.1643	-0.0509	4.6727	1.2862	0.0771	-0.0085
Vars	1.6387	1.0877	1.0791	1.7187	1.1071	2.1176	2.0260	1.1876	1.1017
MSE	2.4897	1.0877	1.0812	1.7457	1.1097	23.9520	3.6803	1.1936	1.1017
MAE	1.2505	0.8300	0.8261	0.9382	0.8389	4.6727	1.5223	0.8747	0.8383
Size5%	0.1510	0.0540	0.0580	0.0840	0.0600	0.9610	0.2000	0.0590	0.0570
Size10%	0.2290	0.1040	0.1060	0.1350	0.1120	0.9830	0.2810	0.1130	0.0980
F-stats	61.4364	65.6591	65.9029	64.1136	64.1136	72.3944	72.3944	64.5415	64.5415

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

DGP3-4-1/2: $\dim(X) = 3$**; Nonlinear** $h(x) = sin(x) + cos(x)$**.**

Table 2.21: $\dim(X) = 3$; Nonlinear $h(x) = sin(x) + cos(x)$; Bivariate-Normal

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.1179	0.0397	0.0399	0.0150	0.0131	1.3103	0.2440	0.0295	0.0649
Vars	1.6410	0.9066	0.9087	0.9238	0.9159	1.0468	1.1825	0.8414	0.8487
MSE	1.6549	0.9082	0.9102	0.9240	0.9161	2.7637	1.2421	0.8423	0.8529
MAE	1.0308	0.7611	0.7627	0.7666	0.7642	1.4205	0.8909	0.7388	0.7470
Size5%	0.0490	0.0650	0.0670	0.0620	0.0600	0.2900	0.0690	0.0560	0.0560
Size10%	0.0940	0.1280	0.1260	0.1210	0.1230	0.4030	0.1320	0.1070	0.1070
F-stats	37.3101	42.0433	41.9871	40.9103	40.9103	39.9044	39.9044	41.1875	41.1875

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-normal with parameter $\rho = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.22: $\dim(X) = 3$; Nonlinear $h(x) = sin(x) + cos(x)$; Bivariate-Logistic

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-2.7621	-0.2225	-0.2266	-0.2061	-0.2062	1.1356	0.1000	0.7936	0.0379
Vars	8.8179	2.2922	2.4965	2.6802	2.3755	3.2328	2.2460	1.9738	1.4055
MSE	16.4470	2.3417	2.5479	2.7227	2.4181	4.5224	2.2560	2.6036	1.4069
MAE	3.3108	1.1954	1.2311	1.2562	1.2140	1.6769	1.1723	1.2604	0.9366
Size5%	0.1660	0.0540	0.0520	0.0520	0.0480	0.1510	0.0470	0.1190	0.0470
Size10%	0.2640	0.1100	0.1030	0.0970	0.0930	0.2320	0.0920	0.2150	0.0890
F-stats	10.0470	25.7178	25.2618	24.7216	24.7216	22.8050	22.8050	27.7708	27.7708

*$(X_{g,i}, X_{g,-i})$ follows a bivariate-logistic with parameter $r = 0.5$
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.23: $\dim(X) = 3$; Linear $h(x) = x\gamma$, $\gamma = 1$; Uncorrelated but dependent

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-0.0047	0.0043	-0.0145	0.0670	-0.0044	-1.5700	-0.3952	-0.3381	-0.2736
Vars	0.3587	0.4799	0.4821	0.5490	0.4801	0.5858	0.7437	1.1594	1.1067
MSE	0.3587	0.4799	0.4823	0.5535	0.4802	3.0507	0.8999	1.2737	1.1815
MAE	0.4807	0.5483	0.5506	0.5837	0.5481	1.5755	0.7371	0.8804	0.8526
Size5%	0.0580	0.0580	0.0580	0.0780	0.0590	0.4160	0.0720	0.2540	0.2430
Size10%	0.1010	0.1120	0.1050	0.1330	0.1090	0.5690	0.1280	0.3340	0.3120
F-stats	212.6313	165.6984	164.8807	164.1038	164.1038	116.9175	116.9175	183.1432	183.1432

*$(X_{g,i}, X_{g,-i})$ follows a uncorrelated but dependent distribution
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.24: $\dim(X) = 3$; Nonlinear $h(x) = exp(0.5x)$; Uncorrelated but dependent

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	0.2985	0.0232	-0.0651	0.2820	0.0084	-1.6509	-0.4531	-0.0860	-0.0876
Vars	1.0281	1.6581	1.6801	1.9565	1.6606	1.9296	2.6507	1.8235	1.7394
MSE	1.1172	1.6586	1.6843	2.0360	1.6607	4.6550	2.8560	1.8309	1.7471
MAE	0.8601	1.0110	1.0113	1.1263	1.0110	1.7532	1.2942	1.0610	1.0368
Size5%	0.0750	0.0580	0.0530	0.1020	0.0560	0.0940	0.0520	0.1420	0.1280
Size10%	0.1410	0.1110	0.1050	0.1540	0.1080	0.1830	0.1070	0.2190	0.2150
F-stats	73.1534	49.1132	48.7938	48.5459	48.5459	33.9561	33.9561	74.1576	74.1576

*$(X_{g,i}, X_{g,-i})$ follows a uncorrelated but dependent distribution
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.25: $\dim(X) = 3$; Nonlinear $h(x) = sin(x) + cos(x)$; Uncorrelated but dependent

	Linear	Series	P.L.S	P.L.	P.L.(db)	R.F.	R.F.(db)	N.N.	N.N.(db)
Bias	-1.1763	0.0224	-0.0110	-0.0487	0.0113	-0.6245	0.0400	-0.0252	-0.0518
Vars	1.9181	0.8122	0.8327	0.8485	0.8134	0.9895	1.0288	0.9152	0.9295
MSE	3.3019	0.8127	0.8328	0.8509	0.8136	1.3795	1.0304	0.9159	0.9322
MAE	1.4422	0.7113	0.7176	0.7266	0.7112	0.9222	0.7994	0.7596	0.7669
Size5%	0.0920	0.0590	0.0570	0.0630	0.0580	0.0760	0.0660	0.0490	0.0480
Size10%	0.1810	0.1190	0.1150	0.1060	0.1190	0.1400	0.1100	0.0860	0.0850
F-stats	51.7247	91.2694	89.7274	90.3729	90.3729	76.0486	76.0486	66.3849	66.3849

*$(X_{g,i}, X_{g,-i})$ follows a uncorrelated but dependent distribution
*The number of observations is $n = 500$, and the number of simulation repetitions $S = 1000$.

Table 2.26: Variables

Outcome variables: $Y_{g,i}$:	
	Cognitive test score / Confidence level
Endogenous social outcomes: $\overline{Y}_{g,-i}$	
	Leave out class average of Y_g
Individual only effect: $X_{g,i}$	
	Preschool / preschool family income / Parents' childbearing age
Individual & Contexture controls: $W_{g,i}$	
Student:	age, gender, one child family
	minority, hukou status, migrant status
	avg study hours per day (HW by teacher / parents)
Family:	father's and mother's years of schooling.
	Financial condition
Group controls: Υ_g	
School:	Tracking school
Teacher:	age, gender, years of schooling, experience, title

Table 2.27: Summary Statistics

Variables	Describe	Mean	s.d.	Min	Max
Outcome variables: $Y_{g,i}$					
score	Cognitive test score	0.0444	0.8727	-2.0290	2.3330
confident	Confidence level ($1 \sim 4$)	3.3092	0.6926	1.0000	4.0000
Individual only controls: $X_{g,i}$					
presch	Attended preschool or not	0.8121	0.3906	0.0000	1.0000
Individual controls: $W_{g,i}$					
gender	Gender	0.5227	0.4995	0.0000	1.0000
age	Age	13.1249	0.8864	9.5000	17.8333
hkplace	Hukou	0.7761	0.4169	0.0000	1.0000
only	One child or not	0.4530	0.4978	0.0000	1.0000
inc	Family income level	2.8339	0.5946	1.0000	5.0000
fa_edu	father's education level	10.9878	2.9649	0.0000	18.0000
Group average controls: $W_{g,i}$					
gender.1	Class average gender	0.5227	0.0774	0.2000	0.8696
age.1	Class average age	13.1249	0.3582	12.5123	14.4718
hkplace.1	Class average Hukou	0.7761	0.2269	0.0000	1.0000
only.1	Class average child	0.4530	0.2730	0.0000	1.0000
inc.1	Class average income	2.8339	0.2598	1.9259	3.4062
fa_edu.1	Class average father's educ	10.9878	1.8431	6.6250	15.9375
Group controls: Υ_g					
t_gender	Teacher's gender	0.2816	0.4498	0.0000	1.0000
t_exp	Teacher's experience	15.0414	8.7929	0.0000	45.0000
t_edu	Teacher's education level	15.4714	0.8122	14.0000	18.0000

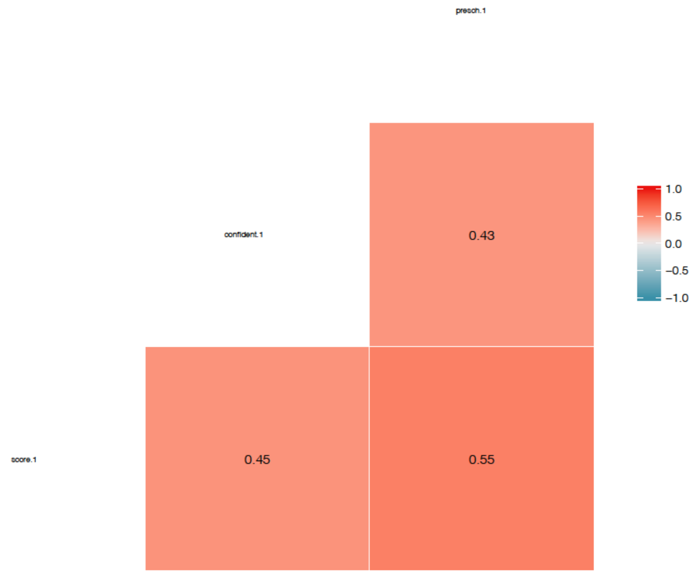

Figure 2.1: Correlation Matrix of Endogenous Variable $\overline{Y}_{g,-i}$ and IV $\overline{X}_{g,-i}$

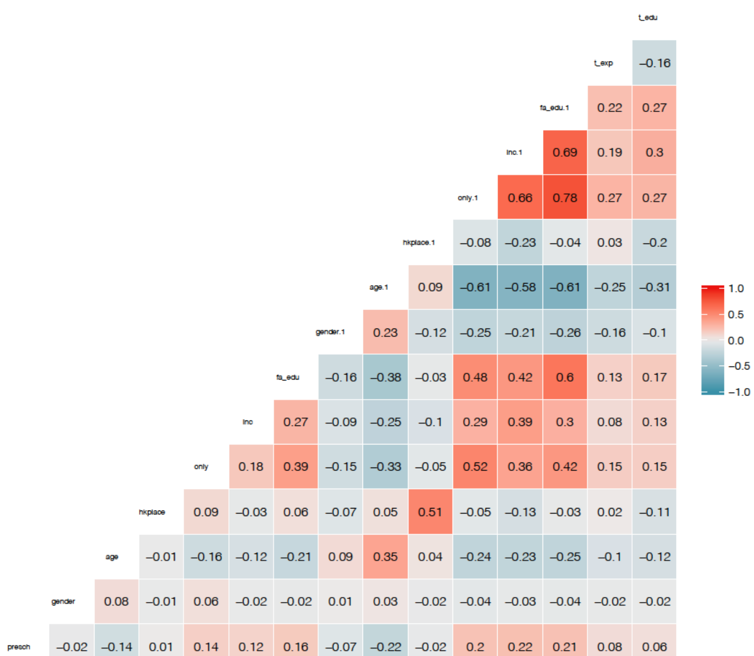

Figure 2.2: Correlation Matrix of Controls $X_{g,i}, W_{g,i}, \overline{W}_{g,i}, \Upsilon_{g,i}$

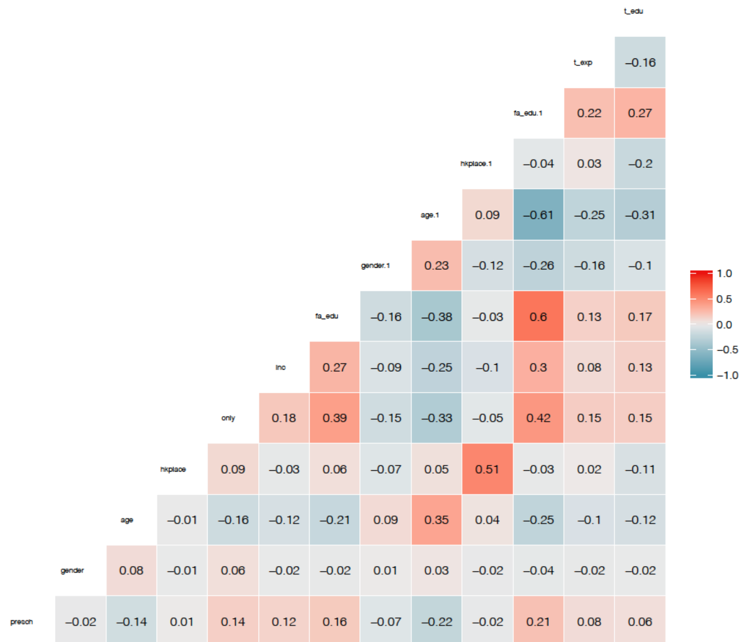

Figure 2.3: Correlation Matrix of Controls $X_{g,i}, W_{g,i}, \widetilde{W}_{g,i}, \Upsilon_{g,i}$
*Note: $\widetilde{W}_{g,i}$ removes the average father's education from $\overline{W}_{g,i}$.

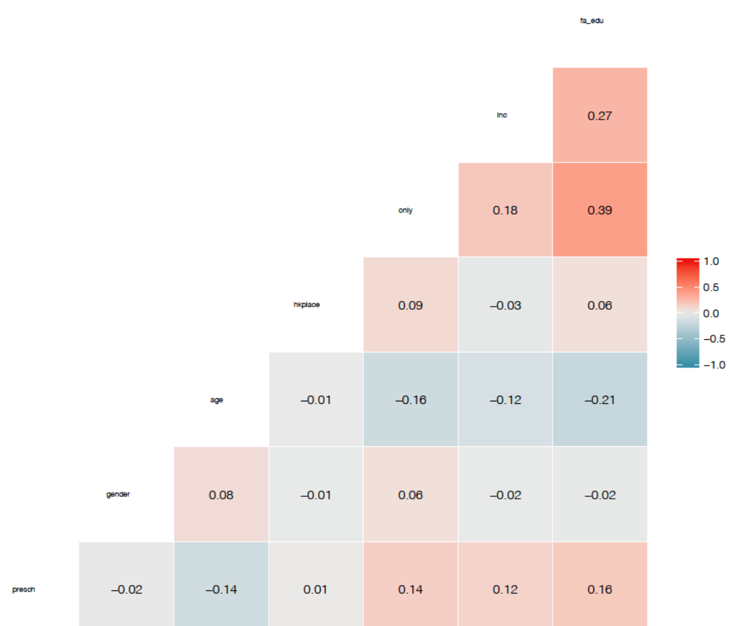

Figure 2.4: Correlation Matrix of Controls $X_{g,i}, W_{g,i}$

Table 2.28: Estimated Results for Endogenous Classmates' Effect on Cognitive Score

Y : Cognitive Score	X, W (d=7)			$X, W, \bar{W}, \Upsilon$ (d=14)		
	$\hat{\beta}$	t	F-stats	$\hat{\beta}$	t	F-stats
Linear	0.7805	15.0310	1411.1197	0.3809	0.8458	25.3490
Series	0.7811	14.8586	1375.3621	0.3973	1.1019	50.8572
Post-LASSO	0.7773	14.7517	1374.3083	0.2558	0.5438	30.3884
Random Forest	0.7215	12.9016	1198.7643	0.2584	0.4018	46.3981
Neural Nets	0.7545	12.2958	944.4589	0.1664	0.2381	14.6908

Table 2.29: Estimated Results for Endogenous Classmates' Effect on Self Confidence

Y :Confidence level	X, W (d=7)			$X, W, \bar{W}, \Upsilon$ (d=14)		
	$\hat{\beta}$	t	F-stats	$\hat{\beta}$	t	F-stats
Linear	0.5611	4.0430	830.9615	0.8037	1.5239	70.2071
Series	0.6123	4.3713	827.9498	0.4519	1.0965	138.2236
Post-LASSO	0.7095	5.6359	991.2625	0.6360	1.1985	78.9778
Random Forest	0.5643	4.0030	770.4734	0.8501	2.7376	234.1050
Neural Nets	0.5864	3.5214	574.8543	0.8928	2.6541	106.4527

Bibliography

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003. 2, 83

Daniel Ackerberg, Xiaohong Chen, and Jinyong Hahn. A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, 94 (2):481–498, 2012. 3, 19, 25, 84, 86

Daniel Ackerberg, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao. Asymptotic efficiency of semiparametric two-step gmm. *Review of Economic Studies*, 81(3):919–943, 2014. 39, 86, 87

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003. 3

Chunrong Ai and Xiaohong Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43, 2007. 3

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. 69

Donald WK Andrews. *Empirical process methods in econometrics*. Citeseer, 1993. 82

Donald WK Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, pages 43–72, 1994. 81, 82

Joshua D Angrist. The perils of peer effects. *Labour Economics*, 30:98–108, 2014. 55, 56

Joshua D Angrist and Kevin Lang. Does school integration generate peer effects? evidence from boston's metco program. *American Economic Review*, 94(5):1613–1634, 2004. 2

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *Stat*, 1050(5):1–26, 2015. 63

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017. 78

Susan Athey, Guido W Imbens, and Stefan Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. Technical report, 2016. 3, 63

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. 78

Eric Auerbach. Identification and estimation of models with endogenous network formation. Technical report, working paper, 2016. 58

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011. 69

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. 70, 71

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012. 63

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014. 3, 63

Marianne Bertrand, Erzo FP Luttmer, and Sendhil Mullainathan. Network effects and welfare cultures. *The Quarterly Journal of Economics*, 115(3):1019–1055, 2000. 1, 61

Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. 70

Lawrence E Blume, William A Brock, Steven N Durlauf, and Yannis M Ioannides. Identification of social interactions. In *Handbook of Social Economics*, volume 1, pages 853–964. Elsevier, 2011. 2

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009. 2, 13, 34, 55, 58, 106

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 76

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 62, 75, 77

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984. 75

William A Brock and Steven N Durlauf. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260, 2001. 2, 3, 57

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007. 3, 19, 82, 83, 86

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003. 3, 82, 83

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1): 649–688, 2015. 63

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018a. 15, 16, 17, 26, 39, 66, 68

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James R Robins. Locally robust semiparametric estimation. 2018b. 25, 66

Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated lasso in high dimensions. Technical report, Working Paper, UCLA, 2019. 71, 72

Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, 101(5):1739–74, 2011. 1, 61

Paul B Ellickson and Sanjog Misra. Supermarket pricing strategies. *Marketing Science*, 27 (5):811–828, 2008. 84

Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981. 63, 79

Theo Gasser and Hans-Georg Müller. Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, pages 23–68. Springer, 1979. 75

Edward L Glaeser, Bruce Sacerdote, and Jose A Scheinkman. Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548, 1996. 1, 61

Arthur S Goldberger. Abnormal selection bias. In *Studies in Econometrics, Time Series, and Multivariate Statistics*, pages 67–84. Elsevier, 1983. 2, 61

Paul Goldsmith-Pinkham and Guido W Imbens. Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264, 2013. 58

Bryan S Graham. Identifying social interactions through conditional variance restrictions. *Econometrica*, 76(3):643–660, 2008. 2, 56, 58

Bryan S Graham and Jinyong Hahn. Identification and estimation of the linear-in-means model of social interactions. *Economics Letters*, 88(1):1–6, 2005. 3, 58

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009. 72, 76, 78, 79, 80

Jerry A Hausman and William E Taylor. Panel data and unobservable individual effects. *Econometrica*, pages 1377–1398, 1981. 58

Chih-Sheng Hsieh and Lung Fei Lee. A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2):301–319, 2016. 58

Hidehiko Ichimura and Sokbae Lee. Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics*, 159(2):252–266, 2010. 82

Ida Johnsson and Hyungsik Roger Moon. Estimation of peer effects in endogenous social networks: control function approach. *Review of Economics and Statistics*, pages 1–51, 2015. 58

Harry H Kelejian and Ingmar R Prucha. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121, 1998. 55

Lung-fei Lee. Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22(4):307–335, 2003. 55

Lung-fei Lee. Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514, 2007. 55, 56

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993. 1, 2, 5, 6, 7, 13, 53, 54, 57, 58

David Marmaros and Bruce Sacerdote. Peer and social networks in job search. *European Economic Review*, 46(4-5):870–879, 2002. 1, 61

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995. 69

KW Newey and D McFadden. Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245, 1994. 3

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382, 1994. 3, 4, 21, 23, 25, 66, 82, 83, 84, 86

Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997. 4, 23

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018. 25, 26, 27, 29

Whitney K Newey, Fushing Hsieh, and James Robins. Undersmoothing and bias corrected functional estimation. 1998. 22

Ariel Pakes and Steven Olley. A limit theorem for a smooth class of semiparametric estimators. *Journal of Econometrics*, 65(1):295, 1995. 82, 83

Xi Qu and Lung-fei Lee. Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics*, 184(2):209–232, 2015. 58

John Rice and Murray Rosenblatt. Integrated mean squared error of a smoothing spline. *Journal of Approximation Theory*, 33(4):353–369, 1981. 75

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, pages 931–954, 1988. 6

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. 80

Bruce Sacerdote. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704, 2001. 1, 61

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. 69

Mark R Segal. Machine learning benchmarks and random forest regression. Technical report, working paper, 2004. 78

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 62, 69

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. 63