Estimation of Endogenous Treatment Effects with Social Interactions

Xueyuan Liu*

February 23, 2022

Abstract

This paper estimates the treatment effect when we relax the standard stable-unit treatment value assumption and consider that individuals and their peers endogenously determine treatment choices. In the presence of peer effects, a group of people decides whether to participate in a program simultaneously. The outcome variable is then realized after their participation decisions, depending on their own treatment status and the peers' choices. Since peer effects both exist in first-stage treatment decisions and second-stage outcome equations, we construct joint propensity score functions of an individual and her peers' treatment statuses. With the joint propensity scores, we identify and estimate the treatment effect on outcome variables using a system of equations of expected outcomes. The possible set of joint propensity scores is generally large as multiple peers could influence one individual, thereby generating debiasing challenges to derive the influence function, which is orthogonal to first-stage joint propensity scores. We propose to approximate debiasing the estimator using numerical derivatives. It is also shown that this estimation method could be extended to heterogeneous treatment effects and endogenous networks.

1 Introduction

The standard stable-unit treatment value assumption (SUTVA), which requires that the potential outcome on one unit should be unaffected by the particular assignment of treatments to the peers (Rubin (1978)), is common in treatment literature. However, it does not necessarily hold in the presence of social interactions since spillover effects from

*Department of Economics, University of California at Los Angeles, Los Angeles, CA 90095. E-mail: liuxueyuan@g.ucla.edu.

peers' treatments also affect one's own outcome. Moreover, if we consider the case that individuals endogenously determine treatment status, peer effect also plays an important role in treatment choices. For instance, a vast literature shows that health behaviors, such as smoking and drinking, are correlated among spouses. Fletcher and Marksteiner (2017) show that smoking behavior is imitated by the spouse, thus increasing the spouse's tobacco use using two randomized experiments. Manski (2000) provides several mechanisms to explain why people within groups tend to behave similarly.

This paper aims to identify and estimate the endogenous treatment effect on outcomes with social interactions. We allow for both peer effects in treatment decisions and spillover effects in potential outcomes under a group structure. Some studies focus on the latter one and circumvent the former one by assuming exogenous treatment assignments. Manski (2013) and Lazzati (2015) indicate that the potential outcome should be a function of the vector of treatment status, covering all states of people in the network. Leung (2019) proposes an estimator to allow for the weak dependence data structure with a single large network and develop network robust standard errors. Vazquez-Bare (2017) identifies the spillover effects within groups given that treatment is randomly assigned. To consider peer effects in treatment decisions, Balat and Han (2018) and Hoshino and Yanagi (2018) model treatment choices as a binary game with a set of threshold-crossing rules. Balat and Han (2018) derive bounds on the average treatment effects (ATEs) under nonparametric shape restrictions, showing that instrument variation compensates strategic substitution. Hoshino and Yanagi (2018) focus on two-player games and impose assumptions that allow interference to be directly identified. Jackson, Yu, and Lin (2020) studies propensity score matching with peer influence. They model social interactions as a linear function and assume no spillover effects in potential outcomes. They propose a nested pseudo-likelihood algorithm to estimate adjusted propensity scores and the corresponding ATEs after matching. Our setting is different from the aforementioned related papers since we do not impose restrictions on treatment participation games within groups. Our goal is to estimate the parameters of interest which are in post-game outcome equations.

Complications arise due to the presence of first-stage peer effects in treatments and second-stage spillover effects. We propose to use joint propensity score functions, which are conditional choice probabilities (CCPs) of one's own treatment and peers' treatments. The functional form of joint propensity scores is left unspecified and thus, flexible. For example, they could be a result of a game-theoretical interaction model. Although we do not impose restrictions on peer effects in the first stage, we need some specific exposure mappings of spillover effects on outcome in the second stage, which are sufficient statistics to summarize the parameters of interest. These mappings could be quite flexible, such as the percentage

of treated peers, the maximum intensity of treatment among peers, etc..Relying on joint propensity scores and exposure mappings, we construct a system of expected outcome equations by varying the realizations of instrumental variables (IVs). This equation system can be proven to have a unique solution under some rank conditions on the joint propensity score matrix. Therefore we can recover the treatment effects by solving the equation system.

Another advantage of joint propensity score functions is to avoid violations of exclusion condition and relevance condition for IVs. The standard way to tackle endogeneity is to find out an IV for the endogenous variable, for instance, the treatment choice. However, it is not easy to implement in the context of peer effects since the exclusion condition is often violated. The instrument we have for one's own treatment status is likely to affect the outcome through other covariates, including the peers' treatment status and the link choice. Moreover, the relevance condition for one's own IV does not necessarily hold as an individual's choices could be affected by IVs for the peers' choices. Without considering the vector of IVs of all individuals within the group, the estimators could be severely biased. Using joint propensity score functions, we allow for flexible treatment response functions (CCPs) to IVs of individuals and peers. The intuition is to employ a method similar to standard two-stage-least-square except that we project joint propensity score functions to the space of the vector of IVs.

One technical issue is to debias our estimators since the joint propensity scores are functions of group-level characteristics, which are high dimensional even in small groups and settings with relatively few relevant demographic characteristics, all these things that ultimately determine treatment participation would be conditioned upon. However, the possible set of joint propensity scores is generally large as multiple peers could influence one individual, thereby generating debiasing challenges to derive the influence function, which is orthogonal to first-stage joint propensity scores. We provide theoretical results that numerical derivatives could approximate the adjustment terms of influence function. This means that our estimators are easy-to-implement, even with a large joint propensity score matrix.

We extend a quasi-panel data approach (Graham and Hahn (2005)) to this paper and consider a group structure. It is shown that heterogeneous treatment effects are also identified using within-group variations and under assumptions that there exists a peer who has symmetric response functions (Angrist and Imbens (1994), Vytlacil (2002), Heckman, Urzua, and Vytlacil (2006), Florens et al. (2008), Heckman and Pinto (2018), Lee and Salanié (2018)). We also provide identification results for a special case when individuals form the network endogenously according to the treatment programs (Kline and Tamer (2018)). The remainder of the paper is organized as follows. In Section 2, we introduce a model of spillover

and present the main idea using a simple two-player case. In Section 3, we formally illustrate the general model. treatment effects are discussed in Section 4. In Section 5, we consider endogenous networks and present theoretical results of numerical derivatives in Section 6. We conduct Monte Carlo simulations in section 7. In Section 8, we replicate Cutler and Glaeser (2010) and compare our estimator's finite sample performance against an estimator that ignores social interactions.

2 A model of spillover

2.1 Identification of a simple case

In the simplest possible case, we illustrate the identification strategy with a group of two. Take individual 1 in the group as an example, she is affected by individual 2's treatment. For instance, when we study the effect of smoking on health status, individual 1's smoking behavior may affect individual 2's health regardless of individual 2's smoking behavior.

Let Y_{gi} be the outcome variable for individual i in group g, $i = 1$ indicates individual 1, and $i = 2$ represents individual 2. T_{g1} and T_{g2} are binary variables that indicate the decisions to participate in a program, or treatment take-up. $T_{gi} = 1$ if unit i in group g receives the active treatment, and $T_{gi} = 0$ if unit i receives the control treatment. Using the potential outcome notation popularized by Rubin (1974), let $Y_{g1}(0, t_{g2})$ denote the outcome for individual 1 under control given the individual 2's treatment t_{g2}, and $Y_{g1}(1, t_{g2})$ is the outcome under treatment. To simplify notation, we suppress the index g henceforth, $\mathbf{X} \equiv (X_1, X_2)$ is the vector of covariates of the group. The parameters of interest are:

$$\beta_{t_2} \equiv Y_1(1, t_2) - Y_1(0, t_2)$$
$$\gamma_{t_1} \equiv Y_1(t_1, 1) - Y_1(t_1, 0)$$

The β_1 is treatment effect given partner receives treatment ($t_2 = 1$), the β_0 measures the treatment effect given partner is not treated ($t_2 = 0$). The γ_1 captures the spillover effect given one's own choice is to be treated ($t_1 = 1$), and the γ_0 is the spillover effect for individuals in control group ($t_1 = 0$).These parameters of interest β_1, β_0, γ_1 and γ_0 are not necessarily equal. In the context of social interaction, the treatment effect also varies with neighbors' behaviors. Take smoking for example, consider two scenarios: case 1. people in the community smoke everywhere, the health status of an individual who lives in the neighborhood could be very poor even if he does not smoke, implying that the treatment effect of smoking on his own health status could be very small; case 2. people in

the community do not smoke at all, the individual's health status could be very good when he quit smoking, suggesting that the treatment effect of smoking on his own health status could be very large. These differences also exist among spillover effects. To recover these parameters, we specify a model of spillover for individual 1 as follows:

$$Y_1(0, t_2) = \mu_1(0, t_2, \mathbf{X}) + \varepsilon_1$$
$$Y_1(1, t_2) = \mu_1(1, t_2, \mathbf{X}) + \varepsilon_0$$

The outcomes variable is a function of the treatments of all individuals in the group, This is relatively more similar to a standard response function in models without social interactions, but with vector outcomes and treatments. The model of the social interaction presumably places additional structure on this response function. This type of treatment response refers to the fact that, from the perspective of the entire group, the group of individuals is "treated" by the treatments of all individuals in the group. Suppose that the object of interest is the "production function" of health outcomes as a function of the vector of treatments in a group. The treatments could be smoking behavior, provision or subsidization of immunization, some sort of public health awareness campaign, or anything else that an individual can manipulate that affects health outcomes. The outcome response function μ_i is indexed by individual ID and treatment status, for example, $\mu_1(1, t_2, X_1)$ is the response function for individual 1 being treated. This means that we allow for heterogeneous response functions within groups. For simplicity, we identify the average treatment effect only for individual 1.We assume that the error term ε is associated with treatment status, for which we normalize $E[\varepsilon_1|\mathbf{X}] = E[\varepsilon_0|\mathbf{X}] = 0$. Notice that $\varepsilon_0 \neq \varepsilon_1$, and we denote $\eta \equiv \varepsilon_1 - \varepsilon_0$.

The outcomes are realized in the second stage after treatment participation in the first stage, which is a equilibrium generated by a structural model, for example, a collective model for household decisions or a entry game with strategic substitution.We allow for flexible form of interactions in treatment decisions and introduce $\mathbf{Z} \equiv (Z_1, Z_2)$as the vector of binary instrumental variable to identify the treatment effect on outcome : $Z_i \in \{0, 1\}$ for $i = 1, 2$. Specifically, the treatments are determined by $\mathbf{X}, \mathbf{Z}$ and a set of unobservables including one's own unobserved heterogeneity v_1, the peer's unobserved heterogeneity v_2, and the

group heterogeneity α_g.

$$T_1 = \mu_{t1}(\underbrace{X_1, X_2, Z_1, Z_2}_{observables}, \underbrace{v_1, v_2, \alpha_g}_{unobservables})$$
$$T_2 = \mu_{t2}(\underbrace{X_2, X_1, Z_2, Z_1}_{observables}, \underbrace{v_2, v_1, \alpha_g}_{unobservables})$$

μ_{ti} is indexed by individual ID because we also allow for heterogeneous responses in treatment choices._

Assumption 1

- 1.1 $(\varepsilon_0, \varepsilon_1) \perp \mathbf{Z} \mid \mathbf{X}$
- 1.2 $\varepsilon_0 = \varepsilon_1$

To identify the parameters, we take expectations of outcome variables at both sides with respect to $\mathbf{X}, \mathbf{Z}$, under assumption 1, we have the following system of equations of expected outcome:

$$\mathbb{E}[Y_1|\mathbf{Z}, \mathbf{X}] = \sum_{(t_1,t_2)\in T_1\times T_2} \mu(t_1, t_2, \mathbf{X}) \times \mathbb{P}(T_1 = t_1, T_2 = t_2|\mathbf{Z}, \mathbf{X}) \tag{1}$$

We denote the propensity score matrix as:

$$\mathbb{P}(\mathbf{X}) = \begin{bmatrix} P_{11}(1,1 \mid \mathbf{X}) & P_{10}(1,1 \mid \mathbf{X}) & P_{01}(1,1 \mid \mathbf{X}) & P_{00}(1,1 \mid \mathbf{X}) \\ P_{11}(1,0 \mid \mathbf{X}) & P_{10}(1,0 \mid \mathbf{X}) & P_{01}(1,0 \mid \mathbf{X}) & P_{00}(1,0 \mid \mathbf{X}) \\ P_{11}(0,1 \mid \mathbf{X}) & P_{10}(0,1 \mid \mathbf{X}) & P_{01}(0,1 \mid \mathbf{X}) & P_{00}(0,1 \mid \mathbf{X}) \\ P_{11}(0,0 \mid \mathbf{X}) & P_{10}(0,0 \mid \mathbf{X}) & P_{01}(0,0 \mid \mathbf{X}) & P_{00}(0,0 \mid \mathbf{X}) \end{bmatrix}$$

Elements of the matrix are propensity score functions, which can be identified from the data:

$$P_{t_1t_2}(z_1, z_2 \mid \mathbf{X}) = \mathbb{E}[T_1 = t_1, T_2 = t_2 \mid Z_1 = z_1, Z_2 = z_2, \mathbf{X}]$$

The conditional expectation matrix is denoted by

$$\mathbb{Y}(\mathbf{X}) = [\mathbb{E}[Y \mid 1,1,\mathbf{X}], \mathbb{E}[Y \mid 1,0,\mathbf{X}], \mathbb{E}[Y \mid 0,1,\mathbf{X}], \mathbb{E}[Y \mid 0,0,\mathbf{X}]]'$$

We require that (z_1, z_2) has at least 4 support points, which holds for binary instrumental variables, to identify the fundamental parameters

$$\Theta(\mathbf{X}) = [\mu(1,1,\mathbf{X}), \mu(1,0,\mathbf{X}), \mu(0,1,\mathbf{X}), \mu(0,0,\mathbf{X})]'$$

and $\beta_1(x), \beta_0(x), \gamma_1(x), \gamma_0(x)$ can be recovered from $\Theta(\mathbf{X})$.

PROPOSITION 1 Under Assumption 1, $\beta_{t_2}(x)$ and $\gamma_{t_1}(x)$ could be identified from

$$\mathbb{P}(\mathbf{X})\Theta(\mathbf{X}) = \mathbb{Y}(\mathbf{X})$$

using binary instrumental variables, where $\Theta(\mathbf{X}) = [\mu(1,1,\mathbf{X}), \mu(1,0,\mathbf{X}), \mu(0,1,\mathbf{X}), \mu(0,0,\mathbf{X})]'$, $\beta_{t_2}(x) = \mu(1,t_2,\mathbf{X}) - \mu(0,t_2,\mathbf{X})$, $\gamma_{t_1}(x) = \mu(t_1,1,\mathbf{X}) - \mu(t_1,0,\mathbf{X})$.

Proof. *See the appendix.*

We use an example to illustrate how treatment choices could be rationalized, although the mechanism of treatment assignment is not our focus. The identification strategy and estimation method we propose are not restricted to the following example.

EXAMPLE 1 *Following Manski (1993), we consider treatment decision as a a binary choice model with social interactions.*

$$T_1 = 1\{X_1\delta_{x1} + X_2\delta_{x2} + Z_1\delta_{z1} + Z_2\delta_{z2} + \rho\mathbb{E}(T_2|X_1, X_2, Z_1, Z_2) + \alpha_g + v_1 \geq 0\}$$

where $1\{.\}$ *is a logical operator that returns one when the argument is true and zero otherwise,* $\delta_2 = (\delta_{x2}, \delta_{z2})$ *is contextual effect,* ρ *is endogenous effect,* α_g *is unobserved group effects,* v_1 *is correlated effect.* T_1 *and* T_2 *are equilibrium outcomes of the static game, which are determined by* $Z_1, Z_2, X_1, X_2, v_1, v_2, \alpha_g$ *We assume either there is a unique equilibrium or multiple equilibria with a well-defined equilibrium selection process.* T_1 *and* T_2 *are functions of group level exogenous variables* $\mathbf{X}$*, instrumental variables* $\mathbf{Z}$ *and error term* ϖ*.*

2.2 Estimation

To estimate $\Theta \equiv \mathbb{E}[\Theta(\mathbf{X})]$, we estimate the propensity score functions and conditional outcome functions in the conditional expectation matrix nonparametrically, treating them as nuisance parameters. Conditional on $\mathbf{X}_g$,we project $\mathbf{T}_g$ and $\mathbf{Y}_g$ onto the space of $\mathbf{Z_g}$, which is a vector of instrumental variables. We estimate each each element in propensity

score matrix $\mathbf{P}(\mathbf{X}_g)$ and conditional outcome matrix $\mathbf{Y}(\mathbf{X}_g)$ as follows:

$$\begin{aligned}
\mathbf{h^p}\left(\mathbf{z_g}, \mathbf{X_g}\right) &\triangleq \mathbb{E}\left[Y_{gi} \mid \mathbf{Z_g} = \mathbf{z}_g, \mathbf{X}_g\right] \\
&= \frac{\mathbb{E}\left[Y_{gi} * \mathbb{I}_{gi}\left(\mathbf{Z_g} = \mathbf{z_g}\right) \mid \mathbf{X}_g\right]}{\mathbb{E}\left[\mathbb{I}_{gi}\left(\mathbf{Z_g} = \mathbf{z_g}\right) \mid \mathbf{X}_g\right]} = \frac{u_{\mathbf{z_g}}\left(\mathbf{X}_g\right)}{d_{\mathbf{z_g}}\left(\mathbf{X}_g\right)} \\
\mathbf{h^y}\left(\mathbf{z_g}, \mathbf{X_g}\right) &\triangleq \mathbb{E}\left[\mathbb{I}_{gi}(T_{gi} = t_{gi}, S_{gi} = s_{gi}) \mid \mathbf{Z_g} = \mathbf{z}_g, \mathbf{X}_g\right] \\
&= \frac{\mathbb{E}\left[\mathbb{I}_{gi}(T_{gi} = t_{gi}, S_{gi} = s_{gi}) * \mathbb{I}_{gi}\left(\mathbf{Z_g} = \mathbf{z_g}\right) \mid \mathbf{X}_g\right]}{\mathbb{E}\left[\mathbb{I}_{gi}\left(\mathbf{Z_g} = \mathbf{z_g}\right) \mid \mathbf{X}_g\right]} = \frac{n_{\mathbf{z_g}}\left(\mathbf{X}_g\right)}{d_{\mathbf{z_g}}\left(\mathbf{X}_g\right)}
\end{aligned}$$

For every $\mathbf{z}_g \in \mathcal{Z}$, we estimate $\mathbf{h^m}\left(\mathbf{z}_g, \mathbf{X}_g\right)$ for m=p, y respectively. The total number of nuisance parameters depends on how we define the measure of spillover effect $\mathbf{S_g}$. We will therefore estimate those parameters using the following nonparametric estimators by plugging in different components.

ASSUMPTION 2 For all $(z, x) \in \mathcal{Z} \times \mathcal{X}$, the $\mathbb{E}\left[\mathbb{I}_{gi}\left(\mathbf{Z_g} = \mathbf{z_g}\right) \mid \mathbf{X}_g\right]$ is bounded away from zero.

Assumption 2 guarantees that we have well-behaved estimators since the propensity score matrix is positive definite. The intuition is that we need an instrument allocation scheme generating sufficient variation within groups, full rank and differentiable propensity score matrix.

The fundamental parameters are defined as can be estimated using the following formula:

$$\widehat{\Theta}(\mathbf{X_g}) = \left[\widehat{\mathbf{P}}(\mathbf{Z_g}, \mathbf{X_g})'\widehat{\mathbf{P}}(\mathbf{Z_g}, \mathbf{X_g})\right]^{-1}\left[\widehat{\mathbf{P}}(\mathbf{Z_g}, \mathbf{X_g})'\widehat{\mathbf{Y}}(\mathbf{Z_g}, \mathbf{X_g})\right].$$

To simplify the notation, we define $\widehat{\Theta}(\mathbf{X_g}) = f(\widehat{\mathbf{h}}^{\mathbf{m}}(\mathbf{z}_g, \mathbf{X}_g))$. And $\mathbf{h_0^m}(\mathbf{z}_g, \mathbf{X}_g)$ as the expectation functions evaluated at true value. Denote $\mathbb{I}_{\mathbf{z_g}} = \mathbb{I}(\mathbf{Z_g} = \mathbf{z_g})$, $\mathbf{W}_g \triangleq (\mathbf{Y_g}, \mathbb{I}_{\mathbf{g}}(\mathbf{T_g} = \mathbf{t_g}, \mathbf{S_g} = \mathbf{s_g}))', n = n_{\mathbf{z_g}}\left(\mathbf{X}_g\right), d = d_{\mathbf{z_g}}\left(\mathbf{X}_g\right), u = u_{\mathbf{z_g}}\left(\mathbf{X}_g\right)$.The adjustment term for nuisance estimators is

$$\begin{aligned}
\alpha_0(\mathbf{X_g}) = \delta(\mathbf{z_g}, \mathbf{X_g})[\mathbf{W_g} - \mathbf{h_0}(\mathbf{z_g}, \mathbf{X_g})] = \sum \mathbb{I}_{\mathbf{z_g}}[&\frac{\partial f}{\partial h^p} \times \left\{\frac{1}{d}\left(T * \mathbb{I}_{\mathbf{z_g}} - n\right) - \frac{n}{d^2}\left(\mathbb{I}_{\mathbf{z_g}} - d\right)\right\} \\
&+ \frac{\partial f}{\partial h^m} \times \left\{\frac{1}{v}\left(Y * \mathbb{I}_{\mathbf{z_g}} - u\right) - \frac{u}{d^2}\left(\mathbb{I}_{\mathbf{z_g}} - d\right)\right\}
\end{aligned}$$

THEOREM 1. Under assumptions 2, if conditions for proposition 1 and regularity conditions in appendix hold, we have:

$$\sqrt{G}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \Omega\right).$$

where

$$\widehat{\theta}_n = \frac{1}{G}\sum_{g=1}^{G}\widehat{\Theta}(\mathbf{X}_g) = \frac{1}{G}\sum_{g=1}^{G}[f(\mathbf{h_0}(\mathbf{Z}_g, \mathbf{X}_g)) + \alpha_0(\mathbf{X_g})] + o_p(1)$$

$$\Omega = Var\{f(\mathbf{h_0}(\mathbf{Z}, \mathbf{X})) - E[f(\mathbf{h_0}(\mathbf{Z}, \mathbf{X}))]\} + \alpha_0(\mathbf{X})$$

Proof. *See the appendix.*

3 General Model

Suppose we consider a general case when one individual is affected by more than one friend. In that case, we can naturally extend the previous analysis exploiting an index to summarize spillover effects, for instance, the ratio of treated friends. Consider n_g units in group g, $d_{ij} = 1$ if i is connected to j, otherwise $d_{ij} = 0$. $a_{ij} = \frac{d_{ij}}{\sum_{j\neq i} d_{ij}}$ if $i \neq j$. Otherwise $a_{ij} = 0$. k is $i's$ degree in group g, or number of neighbors. $\sum_{j\neq i} a_{ij}T_j$ is the fraction of $i's$ treated neighbors. Then the outcome equations become

$$Y_i(1, \sum_{j\neq i} a_{ij}T_j) = \mu_i(1, \sum_{j\neq i} a_{ij}T_j, \mathbf{X}) + \varepsilon_i$$
$$Y_i(0, \sum_{j\neq i} a_{ij}T_j) = \mu_i(0, \sum_{j\neq i} a_{ij}T_j, \mathbf{X}) + \varepsilon_i$$

If the network structure is within group undirected connection, that is, group members are connected to each other, there are $2k$ equations when i has $k-1$ connected neighbors, and there are $2k$ fundamental parameters to be solved. Following the same logic in previous sections, the rank of coefficient matrix can be full, depending on the values of the propensity score. The sufficient condition for the coefficient matrix has full column rank is $\det \boldsymbol{P} \neq 0$. This inequality can be tested using observed propensity score functions. When this is the case, the linear system of equations either has no solution, or a unique solution. If there is a unique solution, then the fundamental parameters are point identified. If the network structure is directed connection, that is, every individual has heterogeneous degree (number of neighbors), then propensity score function is different across different individuals, then there are 2^k equations when i has $k-1$ connected neighbors, and there are $2k$ fundamental parameters to be solved. Take $k = 3$ for example, if the group size is 3 and individuals are connected, we have 8 equations. The sufficient condition is that the coefficient matrix has full column rank.

$\sum_{j\neq i} a_{ij}T_j$ is the exposure mapping of the vector of $i\prime s$ neighbors' treatment status,

which is widely used in network literature, however, the functional restriction can be relaxed in this paper. Propensity score matching is based on a two-dimension index: one's own treatment and the measure of neighbors' treatments, the former variable is binary while the latter one is a discrete variable. The identification replies on full rank condition of propensity score matrix, not the functional form of exposure mapping.

We denote $R_{gi} = (Y_{gi}, Z_{gi}, T_{gi}, X_{gi}, n_{gi})$ as the data on unit i in group g available to the econometrician and assume that we observe a large sample of R_{gi}. Additionally, we assume the the group network structure of each sampled unit i is observed, that is, the set of units i connected to j.

Assumption 3 (Sampling and moments)

1. For $g = 1, \ldots, G; i = 1, \ldots, n_g, \mathbf{R}_g = \left(\mathbf{Y}'_g, \mathbf{Z}'_g, \mathbf{X}'_g, n_g\right)$ is a random sample.
2. Across different groups, $\mathbf{R}_g$ are independent and identically distributed for $g = 1, \ldots, G$, all of the observations have finite second moments.
3. Group size n_g is fixed across g.

For each unit i in group , for $i = 1, \ldots, n_g$. Let T_{gi} indicate whether the treatment of interest was received, with $T_{gi} = 1$ if unit i in group g receives the active treatment, and $T_{gi} = 0$ if unit i receives the control treatment. Using the potential outcome notation popularized by Rubin (1974), let $Y_{gi}(0, s)$ denote the outcome for each unit i under control, where $s = \phi(\mathbf{T}_{g_{-i}})$ is the exposure mapping of the vector of $i\prime s$ neighbors' treatment status, and $Y_{gi}(1, s)$ is the outcome under treatment. To precisely define the parameters of interest, we maintain the exchangeability assumption.

Assumption 4 (Exchangeability)

Let $\mathbf{T}_{-gi}, \widetilde{\mathbf{T}}_{-gi} \in \mathcal{D}_g$ such that $\mathbf{1}'_{-gi}\mathbf{T}_{-gi} = \mathbf{1}'_{-gi}\widetilde{\mathbf{T}}_{-gi}$. Then for each

$$\mathbb{E}\left[Y_{gi}\left(T_{gi}, \mathbf{T}_{-gi}\right)\right] = \mathbb{E}\left[Y_{gi}\left(T_{gi}, \widetilde{\mathbf{T}}_{-gi}\right)\right]$$

The exchangeability assumption is very commonly invoked in the literature, it means that potential outcomes depend on how many peers, but not which ones, are treated, peers are said to be exchangeable. Then we have $s = \sum_{j \neq i} a_{gij} T_{gj}$ is the fraction of treated neighbors for individual i.

Assumption 5

1. (Exogeneity) $(\varepsilon_{0gi}, \varepsilon_{1gi}) \perp \mathbf{Z}_g \mid \mathbf{X}_g$
2. $E\left[\eta_{gi} \mid T_{gi} = 1, \mathbf{X}_g\right] = 0$

We have the general formula for identification:

$$\widetilde{\mathbb{E}}\left[Y_{gi} \mid \mathbf{Z}_g, \mathbf{X}_g\right] = \sum_{(t,s)\in(t\times s)} \mathbb{E}\left[\Theta\left(t_{gi}, s_{gi}\right) \mid \mathbf{X}_g\right] \mathbb{P}\left[T_{gi} = t_{gi}, S_{gi} = s_{gi} \mid \mathbf{Z}_g, \mathbf{X}_g\right]$$

We denote the matrix form of propensity score as $\mathbf{P}(\mathbf{X}_g)$, and matrix form of conditional expectation of the left hand side as $\mathbf{Y}(\mathbf{X_g})$, each element of the matrix can be treated as known since they are observable:

$$\mathbb{E}\left[\mathbf{1}_{gi}(T_{gi} = t_{gi}, S_{gi} = s_{gi}) \mid \mathbf{Z}_g, \mathbf{X}_g\right] = \mathbb{P}\left[T_{gi} = t_{gi}, S_{gi} = s_{gi} \mid \mathbf{Z}_g, \mathbf{X}_g\right]$$

Assumption 6 (Rank Condition 2)

The propensity score matrix $\mathbf{P}(\mathbf{X}_g)$ has full column rank for all $.x. \in \mathcal{X}$.

Proposition 2 (Identification) Under assumption 3, 4, 5 and 6, the endogenous treatment effects and the spillover effects are identified. Specifically,

$$\gamma(t, \mathbf{X}_g) = \mu(t, 1, \mathbf{X}_g) - \mu(t, 0, \mathbf{X}_g)$$
$$ATE(s, \mathbf{X}_g) = \mu(1, s, \mathbf{X}_g) - \mu(0, s, \mathbf{X}_g)$$

Remark 1 *We do not require the exposure mapping $s = \phi(\mathbf{T}_{g_{-i}})$ to be mean number of treated neighbors, it could have flexible functional form to capture heterogeneous interactions, for example, the maximum of peers' behaviors.*

4 Heterogeneous treatment effects

Now we relax assumption 1.2 to consider a heterogeneous case when the group size is 2:

$$\mathbb{E}\left[Y_1|\mathbf{Z}, \mathbf{X}\right] = \underbrace{\sum_{(t_1,t_2)\in T_1\times T_2} \mu(t_1, t_2, \mathbf{X}) \times \mathbb{P}(T_1 = t_1, T_2 = t_2|\mathbf{Z}, \mathbf{X})}_{(a)} + \underbrace{\mathbb{E}\left(\eta|T_1 = 1, \mathbf{X}\right) \mathbb{P}(T_1 = 1|\mathbf{Z}, \mathbf{X})}_{(b)}$$

The (a) contains parameters of interest: $\beta_{t_2}(x)$ and $\gamma_{t_1}(x)$, the (b) depends on unobservables within groups including individual heterogeneities v_1, v_2 and group fixed effect α_g. To disentangle the (a) from the (b) we need some identification conditions utilizing information from peers, in this section we simplify the model by imposing the following assumption, which allows us to ignore the (b) and focus on the (a).

ASSUMPTION 7 1. (Independence) $(\varepsilon_0, \varepsilon_1) \perp \mathbf{Z} \mid \mathbf{X}$

2. (Symmetric response function) Individual 1 and 2 have symmetric treatment response function μ_t and outcome response function μ.

Assumption 7 can be generalized to groups with $n_g(\geq 2)$ individuals, identification requires the existence of at least one group member whose response functions are symmetric to the agent.

$$
\begin{aligned}
(2) \qquad \mathbb{E}\left[Y_1|Z_1, Z_2, \mathbf{X}\right] = &\left[\mu(1,1,\mathbf{X}) - \mu(0,1,\mathbf{X})\right]\mathbb{P}(T_1 = 1, T_2 = 1|Z_1, Z_2, \mathbf{X}) \\
&+ \left[\mu(1,0,\mathbf{X}) - \mu(0,0,\mathbf{X})\right]\mathbb{P}(T_1 = 1, T_2 = 0|Z_1, Z_2, \mathbf{X}) \\
&+ \left[\mu(1,0,\mathbf{X}) - \mu(0,0,\mathbf{X})\right]\mathbb{P}(T_1 = 1, T_2 = 0|Z_1, Z_2, \mathbf{X}) \\
&+ \left[\mu(0,1,\mathbf{X}) - \mu(0,0,\mathbf{X})\right]\mathbb{P}(T_2 = 1|Z_1, Z_2, \mathbf{X}) \\
&+ \mu(0,0,\mathbf{X}) + \mathbb{E}\left(\varepsilon_0\right) + \mathbb{E}\left(\eta|T_1 = 1|Z_1, Z_2, \mathbf{X}\right)\mathbb{P}(T_1 = 1|z, z, \mathbf{X})
\end{aligned}
$$

And for the symmetric individual:

$$
\begin{aligned}
(3) \qquad \mathbb{E}\left[Y_2|Z_1, Z_2, \mathbf{X}\right] = &\left[\mu(1,1,\mathbf{X}) - \mu(0,1,\mathbf{X})\right]\mathbb{P}(T_2 = 1, T_1 = 1|Z_1, Z_2, \mathbf{X}) \\
&+ \left[\mu(1,0,\mathbf{X}) - \mu(0,0,\mathbf{X})\right]\mathbb{P}(T_2 = 1, T_1 = 0|Z_1, Z_2, \mathbf{X}) \\
&+ \left[\mu(0,1,\mathbf{X}) - \mu(0,0,\mathbf{X})\right]\mathbb{P}(T_1 = 1|Z_1, Z_2, \mathbf{X}) \\
&+ \mu(0,0,\mathbf{X}) + \mathbb{E}\left(\varepsilon_0\right) + \mathbb{E}\left(\eta|T_2 = 1|Z_1, Z_2, \mathbf{X}\right)\mathbb{P}(T_1 = 1|z, z, \mathbf{X})
\end{aligned}
$$

We set $z_1 = z_2 = z, x_1 = x_2 = x$.

$$
\begin{aligned}
\mathbb{E}\left[Y_1 - Y_2|z, z, \mathbf{X}\right] = &\left[\mu(1,0,\mathbf{X}) - \mu(0,0,\mathbf{X})\right]\left[\mathbb{P}(T_1 = 1, T_2 = 0|z, z, \mathbf{X}) - \mathbb{P}(T_2 = 1, T_1 = 0|z, z, \mathbf{X})\right] \\
&+ \left[\mu(0,1,\mathbf{X}) - \mu(0,0,\mathbf{X})\right]\left[\mathbb{P}(T_2 = 1|z, z, \mathbf{X}) - \mathbb{P}(T_1 = 1|z, z, \mathbf{X})\right] \\
&+ \mathbb{E}\left(\eta|T_1 = 1, \mathbf{X}\right)\mathbb{P}(T_1 = 1|z, z, \mathbf{X}) - \mathbb{E}\left(\eta|T_2 = 1, \mathbf{X}\right)\mathbb{P}(T_2 = 1|z, z, \mathbf{X})
\end{aligned}
$$

Notice that $\mathbb{P}(T_1 = 1, T_2 = 0|z, z, \mathbf{X}) \neq \mathbb{P}(T_2 = 1, T_1 = 0|z, z, \mathbf{X})$, even if individual 1 and individual 2 have symmetric propensity score function and outcome function. We have $\mathbb{P}(T_1 = 1|z, z, \mathbf{X}) = \mathbb{P}(T_2 = 1|z, z, \mathbf{X})$ but

$$
\mathbb{P}(T_1 = 1, T_2 = 0|z, z, \mathbf{X}) \neq \mathbb{P}(T_1 = 1|z, z, \mathbf{X}) \times \mathbb{P}(T_2 = 0|z, z, \mathbf{X})
$$

due to interdependent treatment choices within groups. Then:

$$
\mu(1,0,\mathbf{X}) - \mu(0,0,\mathbf{X}) = \frac{\mathbb{E}\left[Y_1 - Y_2|z, z, \mathbf{X}\right]}{\left[\mathbb{P}(T_1 = 1, T_2 = 0|z, z, \mathbf{X}) - \mathbb{P}(T_2 = 1, T_1 = 0|z, z, \mathbf{X})\right]}
$$

We have four fundamental parameters: $\mu(1,1,\mathbf{X}), \mu(0,1,\mathbf{X}), \mu(0,0,\mathbf{X}), \mathbb{E}\left(\eta|T_1=1,\mathbf{X}\right)$, which can be identified if if the following propensity score matrix $\mathbb{P}(\mathbf{Z},\mathbf{X})$ has full rank:

$$\mathbb{P}(\mathbf{Z},\mathbf{X}) = \begin{bmatrix} P\left(T_1=1,T_2=1 \mid 1,0,\mathbf{X}\right) & P\left(T_1=1 \mid 0,1,\mathbf{X}\right) & 1 & P\left(T_1=1 \mid 1,0,\mathbf{X}\right) \\ P\left(T_1=1,T_2=1 \mid 0,0,\mathbf{X}\right) & P\left(T_1=1 \mid 0,0,\mathbf{X}\right) & 1 & P\left(T_1=1 \mid 0,0,\mathbf{X}\right) \\ P\left(T_1=1,T_2=1 \mid 1,1,\mathbf{X}\right) & P\left(T_1=1 \mid 1,1,\mathbf{X}\right) & 1 & P\left(T_1=1 \mid 1,1,\mathbf{X}\right) \\ P\left(T_1=1,T_2=1 \mid 0,1,\mathbf{X}\right) & P\left(T_1=1 \mid 1,0,\mathbf{X}\right) & 1 & P\left(T_1=1 \mid 0,1,\mathbf{X}\right) \end{bmatrix}$$

Thus, we conclude that identification condition in this part is similar to previous case, albeit different manipulation of matrices.

$$\underbrace{\begin{bmatrix} P\left(T_1=1,T_2=1|1,0,\mathbf{X}\right) & P\left(T_1=1|0,1,\mathbf{X}\right) & 1 & P\left(T_1=1|1,0,\mathbf{X}\right) \\ P\left(T_1=1,T_2=1|0,0,\mathbf{X}\right) & P\left(T_1=1|0,0,\mathbf{X}\right) & 1 & P\left(T_1=1|0,0,\mathbf{X}\right) \\ P\left(T_1=1,T_2=1|1,1,\mathbf{X}\right) & P\left(T_1=1|1,1,\mathbf{X}\right) & 1 & P\left(T_1=1|1,1,\mathbf{X}\right) \\ P\left(T_1=1,T_2=1|0,1,\mathbf{X}\right) & P\left(T_1=1|1,0,\mathbf{X}\right) & 1 & P\left(T_1=1|0,1,\mathbf{X}\right) \end{bmatrix}}_{\mathbb{P}(\mathbf{Z},\mathbf{X})} \times \underbrace{\begin{bmatrix} \mu(1,1,\mathbf{X})-\mu(0,1,\mathbf{X}) \\ \mu(0,1,\mathbf{X})-\mu(0,0,\mathbf{X}) \\ \mu(0,0,\mathbf{X}) \\ \mathbb{E}\left(\eta|T_1=1,\mathbf{X}\right) \end{bmatrix}}_{\Theta} = \underbrace{\begin{bmatrix} \widehat{\mathbb{E}}\left[Y_1|1,0,\mathbf{X}\right] \\ \widehat{\mathbb{E}}\left[Y_1|0,0,\mathbf{X}\right] \\ \widehat{\mathbb{E}}\left[Y_1|1,1,\mathbf{X}\right] \\ \widehat{\mathbb{E}}\left[Y_1|0,1,\mathbf{X}\right] \end{bmatrix}}_{Y(\mathbf{Z},\mathbf{X})}$$

PROPOSITION 3 Under Assumption 7, $\beta_{t_2}(x)$ and $\gamma_{t_1}(x)$ could be identified using binary instrumental variables.

5 Endogenous Networks

If we relax the assumption that the network is exogenously given, that is, individuals choose the treatment status and corresponding networks. Therefore, the instruments $\mathbf{Z}_g$ for treatment status could also affect network formation. But they affect outcomes only through t and s , t is the endogenous treatment effect, s measures the endogenous spillover effect. To simplify the analysis, we consider simplest possible parametric models.

This assumption applies to Manski-type model of social interaction:

$$Y_{gi} = \beta T_{gi} + \gamma \sum_{j \neq i} a_{gij} T_{gj} + h\left(X_{gi}, W_{gi}, \bar{W}_{g,-i}, \Upsilon_g\right) + \varepsilon_{gi} \tag{4}$$

where Y_{gi} is the outcome variable, X_{gi} and W_{gi} are vectors of exogenous individual characteristics, Υ_g is the group characteristics. $\bar{W}_{g,-i} = \frac{1}{n_g-1}\sum_{j\in I_g, j\neq i} W_{gj}$ denotes the leave-i-out average outcome within the gth group. The instrumental variables affect the outcome equation through treatment variables and link choice. The former captures the effect of treatment on the outcome, and the latter is the effect of treatment on network formation induced by

assignments. If we restrict the endogenous link choice within each group, we can treat both treatment and spillover effects as endogenous. Under these circumstances, our identification condition still holds. The propensity function is identified as a function of group-level instruments and characteristics. Let d_{gij} denotes the indicator function; if utility is transferable, individuals i and j form a link in group g according to the rule:

$$d_{gij} = \mathbf{I}\left(\alpha + \beta Z_{gi} + \beta Z_{gj} + \gamma F_{gij} + B_{gij} - \varsigma_{gij} > 0\right) \tag{5}$$

The term inside the indicator function is the net social surplus associated with a link between i and j. Agents form a link if the net utility from doing so is positive. According to the literature on network formation, let η_i and ι_i be individual-specific latent variables. $F_{gij} = 1$ if i and j have any friends in common and zero otherwise.The latent social distance between individuals i and j is measured by the distance function $h(\eta_i, \iota_i)$, let the pair-specific unobserved heterogeneity term $B_{gij} = v_{gi} + v_{gj} - h(\eta_i, \iota_i) = B_{gji}$. ς_{gij} is utility shock which is independently and identically distributed across pairs.

$$T_{gi} = \mathbb{I}\{Z_{gi} + \sum_{j \neq i} a_{gij} Z_{gj} + \mathbf{X_{gi}} + \sum_{j \neq i} a_{gij} \mathbf{X_{gj}} > v_{gi}\} \tag{6}$$

The pair-specific unobserved heterogeneity B_{gij}, heterogeneity in choice equation v_{gi} and error term in outcome equation ε_{gi} may be correlated. But the critical identification condition still holds since instrumental variables affect outcome variables only through $t = t_{gi}$ and $s = \sum_{j \neq i} a_{gij} t_{gj}$. And both t and s are nontrivial functions of instrumental variables. Therefore we can easily extend the previous estimators to this case. However, if instrumental variables affect outcome equations through other covariates. Take linear-in-means model as an example:

$$Y_{gi} = \beta T_{gi} + \gamma \sum_{j \neq i} a_{gij} T_{gj} + \kappa X_{gi} + \vartheta \sum_{j \neq i} a_{gij} X_{gj} + \varepsilon_{gi} \tag{7}$$

$\mathbf{Z}_g$ affects the outcome variable Y_{gi} through T_{gi} and $\sum_{j \neq i} a_{gij} T_{gj}$, but also $\sum_{j \neq i} a_{gij} X_{gj}$. We can not naturally extend previous analysis to this situation. We leave this case for future research.

6 Approximate debiasing with numerical derivatives

The technical issue we need to address is to debias our estimators since the nuisance parameters are functions of group-level characteristics, which are high dimensional even in small groups and settings with relatively few relevant demographic characteristics, all these things that ultimately determine treatment participation would be conditioned upon. It involves deriving propensity scores from very large dimensional information and relies on machine learning techniques, for instance, random forests, neural net and dropout training etc.. There are two to debias the first stage nuisance parameters estimated using machine learning method for high-dimensional models: cross-fitting and Neyman-orthogonality (Neyman (1959)). According to (Chernozhukov et al. (2018)), sample splitting is not required if the nuisance parameters are estimated by l_1-regularized methods. For the other machine learning techniques, cross-fitting is useful for asymtotic results. A two-stage estimation procedure is orthogonal if the influence function is orthogonal to the nuisance parameters (Neyman (1959)). We use numerical derivatives to approximate the orthogonal influence function.

In our model, suppose that we can solve for $\alpha_0(\mathbf{X_g})$, which is a adjustment term, then we have:

$$\begin{aligned}
\sqrt{G}\left(\widetilde{\theta}_n - \theta_0\right) &= \frac{1}{\sqrt{G}}\sum_{g=1}^{G}[f(\widehat{\mathbf{h}}(\mathbf{Z_g}, \mathbf{X_g})) + \widehat{\alpha}(\mathbf{X_g}) - \mathbb{E}\{f(\mathbf{h_0}(\mathbf{Z}_g, \mathbf{X}_g))\}] \\
\text{(i)} \qquad &= \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\{f(\mathbf{h_0}(\mathbf{Z}_g, \mathbf{X}_g)) - \mathbb{E}\{f(\mathbf{h_0}(\mathbf{Z}_g, \mathbf{X}_g))\}\} \\
\text{(ii)} \qquad &+ \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\{\mathbf{f}_{\mathbf{h}_0(\mathbf{Z_g}, \mathbf{X_g})}(\mathbf{W_g} - \mathbf{h_0}(\mathbf{Z_g}, \mathbf{X_g}))\} \\
&+ o_p(1)
\end{aligned}$$

The (i) is the leading term, the (ii) depends on estimation errors and bias induced by overfitting. We denote the first derivatives as $\mathbf{f}_{\mathbf{h}_0(\mathbf{z_g}, \mathbf{X_g})}$. Chernozhukov et al. (2018) propose a sample splitting procedure to swap the roles of main and auxiliary samples to obtain multiple estimates and then average the results cross-fitting to remove the bias caused by overfitting. Estimation errors can vanish under a broad range of machine learning methods. In some cases, we can solve for the debiased estimators conceptually, like in the first example, however, it is challenging to compute even in the simplest possible case with group size equals 2. Because we have 36 nuisance parameters to estimate and debias, the closed form solutions are too tedious to implement in practice. Moreover, it is possible that we can not

obtain the formula even conceptually for the general model due to possible nonlinear social interactions in both treatment participation and outcome equations. We suggest that numerical derivatives can be used to replace the closed-form orthogonal score. Following Hong et al. (2016),we examine the uniform consistency for first-step machine-learning estimators.

Since:

$$(ii) = \frac{1}{\sqrt{G}} \sum_{g=1}^{G} \{\mathbf{f}_{\mathbf{h}_0(\mathbf{Z_g},\mathbf{X_g})}[\mathbf{W_g} - \mathbf{h_0}(\mathbf{Z_g},\mathbf{X_g})]\} + o_p(1)$$

$$= \frac{1}{\sqrt{G}} \sum_{g=1}^{G} \{\mathbf{f}_{\mathbf{h}_0(\mathbf{Z_g},\mathbf{X_g})}\mathbf{u_g}\} + o_p(1)$$

We denote the directional derivative of f with respect to $\mathbf{h}$ in the direction $\mathbf{w} \in H_n - \mathbf{h_0}$. $L_{k,p}^{\epsilon_n,\mathbf{w}}$ is the kth derivative of f for a d-dimensional $\mathbf{h}$,that makes use of a pth order two-sided formula:

$$L_{k,p}^{\epsilon_n,\mathbf{w}} f(\mathbf{h}) = \frac{1}{\epsilon_n} \sum_{l_1=-p}^{p} \dots \sum_{l_D=-p}^{p} c_{l_1 \dots l_D} f(\mathbf{h} + \sum_{j=1}^{k} l_j \mathbf{w} \epsilon_n).$$

LEMMA 2. Suppose that (i) A mean value expansion of order $2p+1$ applies to the limiting function $f(\mathbf{h})$ uniformly in a neighborhood of $\mathbf{h_0}$. For all sufficiently small $|\epsilon|$ and $r = 2p+1$,

$$\underset{d(\mathbf{h},\mathbf{h_0})=o(1),\mathbf{h},\mathbf{w}\in H_n}{sup} \left| f\left(\mathbf{h} + \sum_{j=1}^{d} l_j \mathbf{w} \epsilon_n\right) - \sum_{l=0}^{r} \frac{\epsilon^l}{l!} f^{(l)}(\mathbf{h}) \right| = O\left(|\epsilon|^{r+1}\right)$$

(ii) The estimated functions $\left\|\widehat{\mathbf{h}} - \mathbf{h}_0\right\|_{p,2} \leqslant \tau_n$, $\tau_n^2 \sqrt{n} \leqslant \delta_n$ for each $n \geqslant n_0$ and all $P \in P_n$, where $\Delta_n \searrow 0, \delta_n \searrow 0$, and $\tau_n \searrow 0$ be sequences of constants approaching zero from above at a speed at most polynomial in n.

For each n, the class of functions $\mathcal{F}_n = \{f(\mathbf{h}(\mathbf{Z}_g, \mathbf{X}_g) + \epsilon_n \mathbf{w}) - f(\mathbf{h}(\mathbf{Z}_g, \mathbf{X}_g) - \epsilon_n \mathbf{w}), \mathbf{h}, \mathbf{w} \in H_n\}$ is suitably measurable and its uniform covering entropy obeys

$$\log N\left(\delta, \mathcal{F}_n, \mathbf{L}_1\right) \leq s_n \log\left(\frac{a_n}{\delta}\right)$$

We can set $\tau_n = O(n^{-1/4}), s_n = 1$ and $a_n = e$ if we employ data-splitting according to Chernozhukov et al. (2017).

(iii) Suppose that $0 \in H_n$ and for $\epsilon_n \to 0$,and some $C > 0$,

$$\sup_{|\mathbf{h}|,|\mathbf{w}|<C,\mathbf{h},\mathbf{w}\in H_n} Var\{f(\mathbf{h}(\mathbf{Z}_g, \mathbf{X}_g) + \epsilon_n \mathbf{w}) - f(\mathbf{h}(\mathbf{Z}_g, \mathbf{X}_g) - \epsilon_n \mathbf{w})\} = O(\epsilon_n)$$

Then we have:

$$\sup_{d(\mathbf{h},\mathbf{h_0})=o(1),\mathbf{h},\mathbf{w}\in H_n} |\frac{1}{\sqrt{G}} \sum_{g=1}^{G} \{L_{1,p}^{\epsilon_n,\mathbf{w}} f(\mathbf{h}(\mathbf{Z}_g, \mathbf{X}_g))\mathbf{u_g}\} - \mathbb{E}\{L_{1,p}^{\epsilon_n,\mathbf{w}} f(\mathbf{h}(\mathbf{Z}_g, \mathbf{X}_g))\mathbf{u_g}\}| = o_p(1)$$

in $\mathbf{Z}, \mathbf{X}$ and $\mathbf{w}$, d(.) be the metric generated by the L^1 norm, provided that $\epsilon_n \to 0$ and $\frac{n\epsilon_n}{N^2 \log n} \to \infty$.

Proof. Followed by Pollard (1984).

7 Simulation

In the Monte Carlo Simulation studies, we consider the semiparametric model with group size $n_g = 2$ across all the groups.

Outcome equation

$$\begin{bmatrix} Y_{g1} \\ Y_{g2} \end{bmatrix} = \beta_0 \begin{bmatrix} T_{g1} \\ T_{g2} \end{bmatrix} + \begin{bmatrix} f(T_{g2}, X_{g1}) \\ f(T_{g1}, X_{g2}) \end{bmatrix} + \begin{bmatrix} \varepsilon_{g1} \\ \varepsilon_{g2} \end{bmatrix}, \quad g = 1, \cdots, G$$

Choice equation

$$T_{g1} = \mathbb{I}(p(X_{g1}, X_{g2}, Z_{g1}, Z_{g2}) - v_{g1} \geq 0)$$

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

(i) The dimensionality of control variables X. In our simulation study, we consider the DGPs with univariate control variable case, $dim(X) = 4$.

(ii) The function forms of nonparametric nuisance function $p()$, $f()$.

We consider both the linear and a variety of nonlinear function forms. For example, for the special case, we ignore interaction effects in the outcome equation and consider linear function $f(X) = X\gamma$ with $\gamma = 0.2$ or 1. Also for the nonlinear function, we consider $f(X) = \frac{1}{1+e^{-X\gamma}}$ or $f(X) = \exp(X\gamma)$. If interaction effects are included, $f(T, X) = T\gamma_1 + X\gamma_2$, $f(T, X) = \frac{1}{1+e^{-T\gamma_1 - X\gamma_2}}$, $f(T, X) = \exp(T\gamma_1 + X\gamma_2)$ respectively.

(iii) The distribution of the control variables X

Assume we generate the group pair regressors $(X_{g1}; X_{g2})$ independently from Multivariate Gaussian distribution.

$$\begin{bmatrix} X_{g1} \\ X_{g2} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_x \\ \rho_x & 1 \end{bmatrix} \right)$$

where ρ measures the within group correlations of X_{g1} and X_{g2}. we consider the case with $\rho_x = 0.5$, implying that groups are linked through the similarity of individuals' characteristics.

(iv) The distribution of the control variables Z.

Assume we generate the group pair regressors $(Z_{g1}; Z_{g2})$ independently from 0 and 1. To avoid the degeneracy problem, we set $P(Z_{g1} = 1) = 0.7$, $P(Z_{g2} = 1) = 0.4$.

(v) The distribution of the control variables v_{gi} and ε_{gi}.

Assume we generate the group pair regressors $(v_{gi}; \varepsilon_{gi})$ independently from Multivariate Gaussian distribution.

$$\begin{bmatrix} v_{gi} \\ \varepsilon_{gi} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{v\varepsilon} \\ \rho_{v\varepsilon} & 1 \end{bmatrix} \right)$$

The disturbance $\varepsilon_{gi} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ where σ_ε^2 measures the noise/signal level. We set $\sigma_\varepsilon^2 = 1$ for all the DGPs.

(vi) The parameter of interest $\beta_0 = 0.5$.

We denote $\mathbf{W} = (T_-, X)'$, $\mathbf{Q_x} = (X, X_-)'$, $\mathbf{Q_z} = (Z, Z_-)', \tau = (\tau_1, \tau_2) = (0.1, 1)$, $\pi_\mathbf{x} = (\pi_1, \pi_2) = (0.1, 1)$,$\pi_\mathbf{z} = (\pi_3, \pi_4) = (0.1, 1)$, $\pi = (\pi_\mathbf{x}, \pi_\mathbf{z})$, $\mathbf{Q} = (\mathbf{Q_x'}, \mathbf{Q_z'})'$.

The number of observations we use is G = 200. The number of simulation repetitions S = 1000. For each simulated dataset, we estimate the endogenous treatment effect by applying the two-step semiparametric estimation methods described in the preceding section. Data Generating Process is as follows: DGP 1: $f(\mathbf{W}) = \tau\mathbf{W}, p(\mathbf{Q}) = \pi\mathbf{Q}$.

DGP 1: dim (X)=1, $\rho_x = \rho_{v\varepsilon} = 0.5$

	Linear	Series	LASSO	Post-LASSO	Neural Net
Estimators considering interactions					
Bias	0.090	0.074	0.044	0.059	0.019
Vars	0.332	0.345	0.513	0.283	0.241
MSE	0.341	0.350	0.515	0.286	0.241
MAE	0.431	0.438	0.497	0.423	0.388
Estimators ignoring interactions					
Bias	1.015	-0.247	0.682	0.530	-0.414
Vars	1.642	0.462	137.408	0.113	0.052
MSE	2.672	0.523	137.873	0.393	0.223
MAE	1.289	0.527	4.569	0.549	0.422

8 Empirical Study

Large literatures in public health and health economics provide evidence that health behaviors, such as smoking, are highly correlated among peers. We attempt to separate the treatment effect of subject smoking and spillover effect on health outcomes using data from the Current Population Survey (CPS). CPS contains tobacco supplement data for information on smoking rates and workplace smoking bans, which can be used as an instrument. In this part, we replicate Cutler and Glaeser (2010), compared with the estimator we proposed, OLS and 2SLS tend to overestimate/underestimate the treatment effect of subject smoking on health outcomes.

Cutler and Glaeser (2010)

Linear	Series	LASSO	Post-LASSO	Neural Net
Estimators considering interactions				
-0.041	-0.132	0.005	0.181	0.013
Estimators ignoring interactions				
-0.046	-0.073	-0.491	-0.328	-0.097

9 Conclusion

This paper shows that the endogenous treatment effect could be separated from spillover effects using propensity score functions, allowing for flexible social interactions in both choice equations and outcome equations. We show that the estimators can be easily applied to heterogeneous treatment effects and endogenous networks. Our estimation method does

not rely on the restrictive exclusion condition, so it is easy to implement in practice. The estimator we proposed is semiparametric because we do not restrict the functional form of the social interactions, both in choice equation and outcome equation. An important part we have not dealt with is to consider the multi-valued and continuous treatment variables. If we think that the strong monotonicity still holds under the ordered treatment case, we can directly generalize the binary case to multi-nominal ones. We have not analyzed complex network structure, for example, the asymmetric neighbors' influence. Throughout this paper, we assume that the interaction effects are symmetric. It is also a potential research direction in the future.

10 Appendix

10.1 Assumptions

ASSUMPTION 8 Suppose that $\mathcal{X}$, and $\mathcal{Y}$ are Cartesian products of closed intervals.

ASSUMPTION 9 Suppose that $P_k(z,x) = P_{z,k_z}(z) \otimes P_{1,k_1}(x_1) \otimes \cdots \otimes P_{d_x,k_{d_x}}(x_{d_x})$. This implies that $k = k_z * \prod_{l=1}^{d_x} k_l$.

ASSUMPTION 10

If k denotes the number of series basis functions used to approximate an unknown function of x, then let k_l , k_z denote the numbers of series basis functions used to approximate the x_l component and z component in the Cartesian space, respectively. For simplicity, we assume that k_l and k_z grow at the same rate.

ASSUMPTION 11

Suppose that and are continuously differentiable of order $d \geq 2$ on the support .

ASSUMPTION 12 Suppose that for a positive integer $\sigma_m \geq 1$, there exist a constant $\alpha_m > 2$ and pseudo-true series coefficients $\xi_{0,k_m} \in \mathbb{R}^{k_m}$ such that

$$\left|h_0^m - P'_{k_m}\xi_{0,k_m}\right|_{\sigma_m} \leq Ck_m^{-\alpha_m}$$

for all positive integers k_m and $m = g, y$.

From Newey (1994), for power series, Assumption 6 is satisfied with $\alpha_m = d/(d_x + d_z)$ and $\sigma_m = 0$.

ASSUMPTION 13 There exist a finite constant $C > 0$ for $m = g, y$, such that the eigenvalues of $E\left[P_{k_m}(z,x)P_{k_m}(z,x)'\right]$ are bounded from above by C and bounded from below by C^{-1} uniformly in k_m. And

$$\left(\frac{k_m}{G} + k_m^{-d/(d_x+d_z)}\right)^{1/2} \zeta_{0,k_m} \to 0, G \to \infty.$$

where $\sup_{x\in\mathcal{X}} \|P_{k_m}(z,x)\| \leq \zeta_{0,k_m}$.

ASSUMPTION 14 $\mathbb{E}\left[\left\|u_g^m\right\|^2 \mid \mathbf{Z}_g, \mathbf{X}_g\right]$ are bounded for $m = p, y$, where the residuals are defined as

$$u_g^y = \mathbf{Y_{gi}} - \mathbf{E}\left[\mathbf{1_{gi}}(\mathbf{T_{gi}} = \mathbf{t_{gi}}, \mathbf{S} = \mathbf{s}) \mid \mathbf{Z_g}, \mathbf{X_g}\right]$$

10.2 Propositions

Proof of Proposition 1. By Theorem 3.11 in , if the joint propensity matrix $\mathbb{P}(\mathbf{X})$ has a full column rank, then there is a unique solution to the system of equations: $\mathbb{P}(\mathbf{X})\Theta = \mathbb{Y}(\mathbf{X})$.

LEMMA 3.

$\left\|\widehat{Q}_{k_m} - Q_{k_m}\right\| = O_p\left(\zeta_{0,k_m} {k_m}^{1/2} G^{-1/2}\right) = o_p(1)$, where $\widehat{Q}_{k_m} = \frac{1}{G}\sum_{g=1}^{G} P_{k_m}\left(\mathbf{Z}_g, \mathbf{X}_g\right) P_{k_m}\left(\mathbf{Z}_g, \mathbf{X}_g\right)'$ and $Q_{k_m} = \mathbb{E}\left[P_{k_m}(\mathbf{Z}, \mathbf{X})P_{k_m}(\mathbf{Z}, \mathbf{X})'\right]$.

PROOF. By i.i.d assumption,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{Q}_{k_m} - Q_{k_m}\right\|^2\right] \\
=&\mathbb{E}\left[\left\|G^{-1}\sum_{g=1}^{G} P_{k_m}\left(\mathbf{Z}_g, \mathbf{X}_g\right) P_{k_m}\left(\mathbf{Z}_g, \mathbf{X}_g\right)' - \mathbb{E}\left[P_{k_m}(\mathbf{Z}, \mathbf{X})P_{k_m}(\mathbf{Z}, \mathbf{X})'\right]\right\|^2\right] \\
\leq& G^{-1}\sum_{j_1, j_2} \mathbb{E}\left[p_{j_1}^2(\mathbf{Z}, \mathbf{X})p_{j_2}^2(\mathbf{Z}, \mathbf{X})\right] = G^{-1}\sum_{j=1}^{k} \mathbb{E}\left[p_j^2(\mathbf{Z}, \mathbf{X})\left\|P_{k_m}(\mathbf{Z}, \mathbf{X})\right\|^2\right] \\
\leq& \zeta_{0,k_m}^2 \mathbb{E}\left[\left\|P_{k_m}(\mathbf{Z}, \mathbf{X})\right\|^2\right] G^{-1} = O\left(\zeta_{0,k_m}^2 k_m G^{-1}\right) = o_p(1).
\end{aligned}$$

■

LEMMA 4. Under assumptions 8-14, we have

$$\left\|\hat{\xi}_{k_m} - \xi_{0,k_m}\right\| = O_p\left({k_m}^{1/2} G^{-1/2} + {k_m}^{-\alpha_m}\right)$$

PROOF. Let $\mathbf{H}_{m,g} = \left(h_0^m\left(X_1\right), \ldots, h_0^m\left(X_G\right)\right)'$, $\mathbf{u}_{m,g} = \left(u_1^m, \ldots, u_g^m\right)'$.

$$\begin{aligned}\hat{\xi}_{k_m} - \xi_{0,k_m} &= \left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1}\mathbf{P}'_{m,g}\mathbf{Y}_{m,g} - \xi_{0,k_m} \\ &= \left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1}\mathbf{P}'_{m,g}\left(\mathbf{H}_{m,g} - \mathbf{H}_{0,k_m}\right) + \left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1}\mathbf{P}'_{m,g}\mathbf{u}_{m,g}\end{aligned}$$

$$\begin{aligned}&\left\|\left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1}\mathbf{P}'_{m,g}\mathbf{u}_{m,g}\right\|^2 \\ &= \frac{\mathbf{u}'_{m,g}\mathbf{P}_{m,g}\left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1/2}\left(\widehat{Q}_{k_m}\right)^{-1}\left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1/2}\mathbf{P}'_{m,g}\mathbf{u}_{m,g}}{N} \\ &\leq \left(G\lambda\min\left(\widehat{Q}_{k_m}\right)\right)^{-1}\left\|\left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1/2}\mathbf{P}'_{m,g}\mathbf{u}_{m,g}\right\|^2 = O_p\left(k_m G^{-1}\right)\end{aligned}$$

$$\begin{aligned}&\left\|\left(\mathbf{P}'_{m,g}\mathbf{P}_{m,g}\right)^{-1}\mathbf{P}'_{m,g}\left(\mathbf{H}_{m,g} - \mathbf{H}_{0,k_m}\right)\right\|^2 \\ &\leq \left(G\lambda_{\min}\left(\widehat{Q}_{k_m}\right)\right)^{-1}\left(\mathbf{H}_{m,g} - \mathbf{H}_{0,k_m}\right)'\left(\mathbf{H}_{m,g} - \mathbf{H}_{0,k_m}\right) \\ &= O_p\left({k_m}^{-2\alpha_m}\right)\end{aligned}$$

■

Under assumptions 8-14, we have

$$\left\|\widehat{h}^m - h_0^m\right\| = O_p\left(\zeta_{0,k_m}\left({k_m}^{1/2}G^{-1/2} + {k_m}^{-\alpha_m}\right)\right)$$

PROOF. By the triangle inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}\left\|\widehat{h}^m - h_0^m\right\| &= \left\|\widehat{h}^m - h_{0,k_m}^m\right\| + \left\|h_{0,k_m}^m - h_0^m\right\| \\ &= \left\|P_{k_m}(z,x)\right\|\left\|\hat{\xi}_{k_m} - \xi_{0,k_m}\right\| + \left\|h_{0,k_m}^m - h_0^m\right\| \\ &= O_p\left(\zeta_{0,k_m}\left({k_m}^{1/2}G^{-1/2} + {k_m}^{-\alpha_m}\right)\right)\end{aligned}$$

■

LEMMA 5. (Linearity) (i) For all $\mathbf{h^m}$ with $|\mathbf{h^m} - \mathbf{h_0^m}|_0$ sufficiently small,

$$\left|\mathbf{f}(\mathbf{h^m}(\mathbf{Z},\mathbf{X})) - \mathbf{f}(\mathbf{h_0^m}(\mathbf{Z},\mathbf{X})) - \mathbf{D}(\mathbf{Z},\mathbf{X},\mathbf{h^m} - \mathbf{h_0^m})\right|_0 \leq b^m(\mathbf{Z},\mathbf{X})\left|\mathbf{h^m} - \mathbf{h_0^m}\right|_0^2$$

(ii) $\mathbb{E}[b^m(\mathbf{Z},\mathbf{X})]\sqrt{G}\,|\mathbf{h^m}-\mathbf{h_0^m}|_0^2 \to 0$ PROOF. Let $D(\mathbf{Z},\mathbf{X},\mathbf{h^m}-\mathbf{h_0^m}) = f_{h_0^m}(\mathbf{h^m}-\mathbf{h_0^m})$, $b^m(\mathbf{Z},\mathbf{X}) = \sup_{(z,x)\in\mathcal{Z}\times\mathcal{X}} |D^2 f(h_0^m(\mathbf{Z},\mathbf{X}))|$, we have

$$|\mathbf{f}(\mathbf{h^m}(\mathbf{Z},\mathbf{X})) - \mathbf{f}(\mathbf{h_0^m}(\mathbf{Z},\mathbf{X})) - \mathbf{D}(\mathbf{Z},\mathbf{X},\mathbf{h^m}-\mathbf{h_0^m})|_0 \leq b^m(\mathbf{Z},\mathbf{X})\,|\mathbf{h^m}-\mathbf{h_0^m}|_0^2$$

$$\mathbb{E}[b^m(\mathbf{Z},\mathbf{X})]\sqrt{G}\,|\mathbf{h^m}-\mathbf{h_0^m}|_0^2 = \mathbb{E}[b^m(\mathbf{Z},\mathbf{X})]\sqrt{G}\zeta_{0,k_m}\left({k_m}^{1/2}G^{-1/2} + {k_m}^{-\alpha_m}\right) \to 0$$

■

(Stochastic equicontinuity) (i)

(ii) PROOF. (i)

$$\left\|\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\Delta\left(\mathbf{Z_g},\mathbf{X_g},\mathbf{h^m_{0,k_m}}-\mathbf{h_0^m}\right)\right\|^2 = \mathbb{E}[\Delta\left(\mathbf{Z},\mathbf{X},\mathbf{h^m_{0,k_m}}-\mathbf{h_0^m}\right)]$$
$$= O_p\left(K^{-2\alpha}\right) = o_p(1)$$

(ii)

$$\left\|\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\Delta\left(\mathbf{Z_g},\mathbf{X_g},\widehat{\mathbf{h}}^{\mathbf{m}}-\mathbf{h_0^m}\right)\right\| = \left\|\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\Delta_g^K\right)'\left(\hat{\xi}_{k_m}-\xi_{0,k_m}\right)\right\|$$
$$= O_p\left(\left(\sum_{k_m=1}^{K_m}|p_{k_mK}|_d^2\right)^{1/2}\left[(K_m/G)^{1/2}+K_m^{-\alpha_m}\right]\right)$$
$$= o_p(1)$$

■

(Mean-square continuity) (i)

(ii) (i)

$$
\begin{aligned}
&\mathbf{f_{h_0}}(\mathbf{Z_g}, \mathbf{X_g})(\widehat{\mathbf{h}}^{\mathbf{m}} - \mathbf{h}_{\mathbf{0}}^{\mathbf{m}}) \\
&= \mathbf{f_{h_0}}(\mathbf{Z_g}, \mathbf{X_g})(\widehat{\mathbf{h}}^{\mathbf{m}} - \mathbf{h}_{\mathbf{0,k_m}}^{\mathbf{m}}) + o_p(1) \\
&= \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})(\widehat{\mathbf{h}}^{\mathbf{m}} - \mathbf{h}_{\mathbf{0,k_m}}^{\mathbf{m}}) + o_p(1) \\
&= \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})\mathbf{P_{k_m}}(\mathbf{Z}, \mathbf{X})'\hat{\mathbf{Q}}_{\mathbf{k_m}}^{-\mathbf{1}} \frac{\mathbf{1}}{\mathbf{G}} \sum_{\mathbf{g=1}}^{\mathbf{G}} \mathbf{P_{k_m}}(\mathbf{Z_g}, \mathbf{X_g})(\mathbf{W_g} - \mathbf{h}_{\mathbf{0,k_m}}^{\mathbf{m}}) + o_p(1) \\
&= \frac{\mathbf{1}}{\mathbf{G}} \sum_{\mathbf{g=1}}^{\mathbf{G}} \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})\mathbf{P_{k_m}}(\mathbf{Z}, \mathbf{X})'\hat{\mathbf{Q}}_{\mathbf{k_m}}^{-\mathbf{1}}\mathbf{P_{k_m}}(\mathbf{Z_g}, \mathbf{X_g})\varepsilon_{\mathbf{g}} + o_p(1)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left\| \frac{1}{G} \sum_{g=1}^{G} \alpha_K(\mathbf{Z_g}, \mathbf{X_g}) - \mathbb{E}[D\left(\mathbf{Z_g}, \mathbf{X_g}, \widehat{\mathbf{h}}^{\mathbf{m}} - \mathbf{h}_{\mathbf{0}}^{\mathbf{m}}\right) \right\| \\
&= \mathbb{E}\left[\left\| \frac{1}{G} \sum_{g=1}^{G} \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})\varepsilon_{\mathbf{g}} - \mathbb{E}[\mathbf{f_{h_0}}(\mathbf{Z_g}, \mathbf{X_g})(\widehat{\mathbf{h}}^{\mathbf{m}} - \mathbf{h}_{\mathbf{0}}^{\mathbf{m}})] \right\| \right] \\
&= \mathbb{E}\left[\left\| \frac{1}{G} \sum_{g=1}^{G} \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})\varepsilon_{\mathbf{g}} - \mathbb{E}[\frac{\mathbf{1}}{\mathbf{G}} \sum_{\mathbf{g=1}}^{\mathbf{G}} \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})\mathbf{P_{k_m}}(\mathbf{Z}, \mathbf{X})'\hat{\mathbf{Q}}_{\mathbf{k_m}}^{-\mathbf{1}}\mathbf{P_{k_m}}(\mathbf{Z_g}, \mathbf{X_g})\varepsilon_{\mathbf{g}}] \right\| \right] \\
&= \mathbb{E}\left[\left\| \frac{1}{G} \sum_{g=1}^{G} \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})\varepsilon_{\mathbf{g}} - \mathbb{E}[\mathbf{f_{h_{0,k_m}}}(\mathbf{Z}, \mathbf{X})\mathbf{P_{k_m}}(\mathbf{Z}, \mathbf{X})'\hat{\mathbf{Q}}_{\mathbf{k_m}}^{-\mathbf{1}}\mathbf{P_{k_m}}(\mathbf{Z}, \mathbf{X})\varepsilon] \right\| \right] \\
&= \mathbb{E}\left[\left\| \frac{1}{G} \sum_{g=1}^{G} \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})\varepsilon_{\mathbf{g}} - \mathbb{E}[\mathbf{f_{h_{0,k_m}}}(\mathbf{Z}, \mathbf{X})\varepsilon] \right\| \right] = o_p(1)
\end{aligned}
$$

(ii)

$$
\begin{aligned}
&\left\| \frac{1}{G} \sum_{g=1}^{G} \alpha_K(\mathbf{Z_g}, \mathbf{X_g}) - \frac{1}{G} \sum_{g=1}^{G} \alpha_0(\mathbf{Z_g}, \mathbf{X_g}) \right\| \\
&= \left\| \frac{1}{G} \sum_{g=1}^{G} \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})(\varepsilon_{\mathbf{g}} + \mathbf{h}_{\mathbf{0}}^{\mathbf{m}} - \mathbf{h}_{\mathbf{0,k_m}}^{\mathbf{m}}) - \frac{1}{G} \sum_{g=1}^{G} \mathbf{f_{h_0}}(\mathbf{Z_g}, \mathbf{X_g})(\varepsilon_{\mathbf{g}}) \right\| \\
&\leq \mathbb{E}\left\| [\mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g}) - \mathbf{f_{h_0}}(\mathbf{Z_g}, \mathbf{X_g})]\varepsilon_{\mathbf{g}} \right\| + \mathbb{E}\left\| \mathbf{f_{h_{0,k_m}}}(\mathbf{Z_g}, \mathbf{X_g})(\mathbf{h}_{\mathbf{0}}^{\mathbf{m}} - \mathbf{h}_{\mathbf{0,k_m}}^{\mathbf{m}}) \right\| \\
&= o_p(1)
\end{aligned}
$$

Proof of Theorem 1. Follows from the above lemmas.