Identification and Estimation of Social Interactions in Endogenous Peer Groups

Shuyang Sheng (UCLA)   Xiaoting Sun (SFU)

ASSA/ES North American Winter Meeting
January 4, 2021

Introduction

- We consider a linear-in-means social interaction model with non-overlapping groups (Manski, 1993).
  - The peer groups can be endogenous.
- Examples that lead to endogenous peer groups:
  - College admission
  - Student assignment
  - Neighborhood choice
  - Human resources
- The objective of the paper:
  - Correct for the selection bias due to endogenous group formation
  - Identify and estimate the causal peer effects

In This Paper

- We characterize group formation by a two-sided many-to-one matching model.

- We derive a nonparametric form of the selection bias, which depends on the preferences and qualification indices in group formation.

- The peer effects can be identified using typical methods once we control for the selection bias as in a sample selection model.

- We propose a two-stage distribution-free estimation method:
  - 1st stage: estimate the group formation parameters semiparametrically
  - 2nd stage: estimate the social interactions by semiparametric two-step GMM

Related Literature

- Social interactions with exogenous groups or networks:
  - Manski (1993), Moffitt (2001), Brock and Durlauf (2001), Lee (2007), Graham (2008), Bramoulle, Djebbari, and Fortin (2009), De Giorgi, Pellizzari, and Redaelli (2010), Lin (2010), Liu and Lee (2010), Laschever (2011), Lee, Li, and Lin (2014), Blume et al. (2018), Cohen-Cole, Liu, and Zenou (2018), among others.
- Social interactions with endogenous networks:
  - Goldsmith-Pinkham and Imbens (2013), Arduini, Patacchini, and Rainone (2015), Qu and Lee (2015), Hsieh and Lee (2016), Hsieh and van Kippersluis (2018), Auerbach (2019), Hsieh, Lee, and Boucher (2019), Johnsson and Moon (2019).
- Social interactions with endogenous unilateral groups:
  - Brock and Durlauf (2003, 2007).
- Our paper focuses on endogenous bilateral groups:
  - Many-to-one matching: Azevedo and Leshno (2016), He, Sinha, and Sun (2020).
- Sample selection models:
  - Heckman (1979), Das, Newey, and Vella (2010), Newey (2019).

Model

Notation

- Finite # of non-overlapping groups: $\{1, 2, \ldots, G\}$
- Each group has (exogenous) capacity n_g
- Large number of individuals: $\{1, 2, \ldots, n\}$

Model

Social Interactions

- Consider a linear-in-means social interaction model (Manski, 1993)

$$
\begin{aligned}
y_i &= \overline{y}_{g_i}\gamma_1 + (\overline{x}_{g_i})^{'}\gamma_2 + x_i^{'}\gamma_3 + \epsilon_i \\
&= \Big(\sum_{j \neq i} w_{ij} y_j\Big)\gamma_1 + \Big(\sum_{j \neq i} w_{ij} x_j\Big)^{'}\gamma_2 + x_i^{'}\gamma_3 + \epsilon_i, \qquad (1)
\end{aligned}
$$

where

$$g_i : \text{ the group that individual } i \text{ joins,}$$

$$
w_{ij} = \begin{cases} \frac{1}{n_{g_i}-1} & if\ g_i = g_j \\ 0 & if\ g_i \neq g_j. \end{cases}
$$

Model

Group Formation/Matching

- Most literature assumes unilateral group formation, where individuals unilaterally choose groups (e.g., Brock and Durlauf, 2003).
- We consider bilateral group formation, where individuals choose groups but also have to be qualified for groups (He, Sinha, and Sun, 2020).
    - E.g., college admission, neighborhood choice
- Many-to-one matching between individuals and “groups”

Model

Group Formation/Matching

- Utility of individual i when joins in group g:

$$u_{ig} = z_i^{'} \delta_g^u + \xi_{ig}, \tag{2}$$

- Qualification for individual i of joining in group g:

$$v_{gi} = z_i^{'} \delta_g^v + \eta_{gi}. \tag{3}$$

- $\xi_i = (\xi_{i1}, \ldots, \xi_{iG}), \eta_i = (\eta_{1i}, \ldots, \eta_{Gi})$
- $(\varepsilon_i, \xi_i, \eta_i) \overset{i.i.d.}{\sim} F$

Endogeneity in Groups

Simulation example

- Market Parameters:
  - No. of students: 2000
  - No. of schools: 3
  - School capacities: 495, 462, 495
- Student Preferences:

$$u_{ig} = \alpha_g + \delta^u_{1,g} d_{ig} + \delta^u_{2,g} z_i + \xi_{ig}$$

  - $\alpha_1 = 3.5$, $\alpha_2 = 1.5$, $\alpha_3 = 0$
  - $\xi_{ig} \sim i.i.d.$ type I extreme value, outside option of value $\xi_{i0} \sim i.i.d.$ type I extreme value.
- Student Qualifications:

$$v_{gi} = \delta^v_{1,g} w_{ig} + \delta^v_{2,g} z_i + \eta_{gi}$$

  - exogenous group formation: $\eta_{gi} = \tilde{\eta}_{gi}$ v.s. endogenous group formation: $\eta_{gi} = \epsilon_i + \tilde{\eta}_{gi}$
  - $\epsilon_i \sim i.i.d.\, N(0,1)$, $\tilde{\eta}_{gi} \sim i.i.d.\, N(0,1)$, $\epsilon_i \perp \tilde{\eta}_{gi}$
- Linear-in-means model:

$$y_i = \left(\overline{x}_{g_i}\right)' \beta + \epsilon_i$$

  - $(z_i, x_i) \sim i.i.d.\, N\left(\begin{pmatrix} 6 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}\right)$
- Group formed by the deferred acceptance algorithm

Endogeneity in Groups

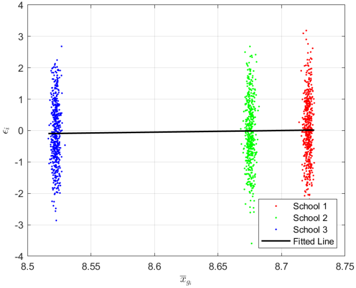

Figure: Exogenous group formation

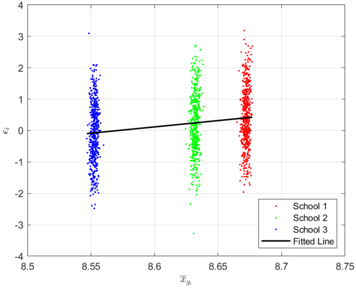

Figure: Endogenous group formation

avg. of 200 MC samples	exogenous groups	endogenous groups
Social effects: $\hat{\beta}$ (true value =1)	0.998	2.399
Reg of ϵ_i on $\overline{x}_{g_i}$: t-stat	-0.011	5.899

Selection Bias

- Notation:
  - $\boldsymbol{z}_n = (z_1, \dots, z_n)'$: $n \times d_z$ matrix
  - $\boldsymbol{g}_n = (g_1, \dots, g_n)'$: $n \times 1$ vector
  - $\boldsymbol{z}_{-i} = (z_1, \dots z_{i-1}, z_{i+1}, \dots, z_n)'$: $(n-1) \times d_z$ matrix
  - $\boldsymbol{g}_{-i} = (g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_n)'$: $(n-1) \times 1$ vector

- The selection bias is

$$E\left(\epsilon_i | \boldsymbol{x}_n, \boldsymbol{g}_n\left(\boldsymbol{z}_n, \boldsymbol{\xi}_n, \boldsymbol{\eta}_n\right)\right)$$

- Sources of endogeneity: ϵ_i and (ξ_i, η_i) are correlated
  - leading to correlation between ϵ_i and $(g_i, \boldsymbol{g_{-i}})$

Equilibrium in Group Formation

Stable Matching

- A stable matching is characterized by “cut-offs” (Azevedo & Leshno, 2016)
  - Cut-off for group g:

$$p_g = \inf_{\{i:g_i=g\}} v_{gi}, \; g = 1, \ldots, G$$

- An individual can only choose among the groups that she qualifies for

$$C_i(p) = \{g : v_{gi} \geq p_g\}$$

Equilibrium in Group Formation

Stable Matching

- Stability means that each individual is matched with her most-preferred feasible group:

  $$g_i = arg \max_{k \in C_i(p)} u_{ik},$$

  which is equivalent to

  $$1\left(g_i = g\right) = \underbrace{1\left(v_{gi} \geq p_g\right)}_{g \text{ is feasible to } i} \cdot \prod_{k \neq g} \underbrace{1\left(u_{ik} < u_{ig} \, or \, v_{ki} < p_k\right)}_{k \text{ is not preferred over } g \text{ or } k \text{ is not feasible to } i} .$$

- The equilibrium cut-offs clear the market.

Selection Bias

Cut-offs as Sufficient Statistics

- **Assumption 1.** *IID.* $(x_i, z_i, \epsilon_i, \xi_i, \eta_i)$ are i.i.d. for all $i = 1, \ldots, n$.
- **Lemma 1:** Under Assumption 1, for any deterministic cut-off $\boldsymbol{p}^0$**,** we have

$$E\left(\epsilon_i | \boldsymbol{x}_n, \boldsymbol{g}_n\left(\boldsymbol{z}_n, \boldsymbol{\xi}_n, \boldsymbol{\eta}_n; \boldsymbol{p}^0\right)\right) = E\left(\epsilon_i | x_i, g_i\left(z_i, \xi_i, \eta_i; \boldsymbol{p}^0\right)\right).$$

- **Intuition:**

$$\begin{aligned} & E\left(\epsilon_i | \boldsymbol{x}_n, \boldsymbol{g}_n\left(\boldsymbol{z}_n, \boldsymbol{\xi}_n, \boldsymbol{\eta}_n; \boldsymbol{p}^0\right)\right) \\ = \; & E\left(\epsilon_i | x_i, \boldsymbol{x}_{-i}, g_i(z_i, \xi_i, \eta_i; \boldsymbol{p}^0), \boldsymbol{g}_{-i}(\boldsymbol{z}_{-i}, \boldsymbol{\xi}_{-i}, \boldsymbol{\eta}_{-i}; \boldsymbol{p}^0)\right) \\ = \; & E\left(\epsilon_i | x_i, g_i(z_i, \xi_i, \eta_i; \boldsymbol{p}^0)\right). \end{aligned}$$

Selection Bias

Limiting Approximation

- **Assumption 2.** *SMOOTHNESS*. The joint cdf of $(\epsilon_i, \xi_i, \eta_i)$ is continuously differentiable.

- **Lemma 2:** Under Assumptions 1-2, we have

$$E\left(\epsilon_i | \boldsymbol{x}_n, \boldsymbol{g}_n\left(\boldsymbol{z}_n, \boldsymbol{\xi}_n, \boldsymbol{\eta}_n; \boldsymbol{p}_n\right)\right) \xrightarrow{p} E\left(\epsilon_i | x_i, g_i\left(z_i, \xi_i, \eta_i; \boldsymbol{p}\right)\right).$$

  - $\boldsymbol{p}_n = (p_{n,1}, \ldots, p_{n,G})$: $G \times 1$ vector of equilibrium cut-offs in a finite market with n individuals
  - $\boldsymbol{p} = (p_1, \ldots, p_G)$: $G \times 1$ vector of equilibrium cut-offs in a limiting market where $n \to \infty$
  - **Intuition:**

$$\boldsymbol{p}_n \xrightarrow{p} \boldsymbol{p}, \text{ and } \boldsymbol{p} \text{ is unique (Azevedo and Leshno, 2016)}$$

Selection Bias

Nonparametric Form

- Recall that the group assignment is determined by

$$1\left(g_i = g\right) = \underbrace{1\left(v_{gi} \geq p_g\right)}_{i \text{ qualifies for } g} \times \prod_{k \neq g} \underbrace{1\left(u_{ik} < u_{ig} \; or \; v_{ki} < p_k\right)}_{i \text{ prefers } g \text{ to } k \text{ or } i \text{ does not qualify for } k}$$

$$= 1\left(v_{gi} \geq p_g\right) \times \prod_{k \neq g} \left(1 - 1\left(u_{ik} \geq u_{ig}\right) 1\left(v_{ki} \geq p_k\right)\right)$$

- This can be seen as a multi-variate selection rule (Das, Newey, and Vella, 2003).

- Define the propensity scores of the selection rules:
  - $\pi_k^v = P\left(v_{gi} \geq p_g | z_i\right)$
  - $\pi_{gk}^u = P\left(u_{ik} \geq u_{ig} | z_i\right)$
  - $\pi_g = (\pi_1^v, \ldots, \pi_G^v, \pi_{g1}^u, \ldots, \pi_{g(g-1)}^u, \pi_{g(g+1)}^u, \ldots, \pi_G^u)$: a collection of $2G - 1$ propensity scores associated with group g.

Selection Bias

Nonparametric Form

- **Assumption 3.** *EXOGENEITY.* $(\epsilon_i, \xi_i, \eta_i)$ are independent of (x_i, z_i).
- **Assumption 4.** *MONOTONICITY.* The cdf of (ξ_i, η_i) is strictly increasing.

Selection Bias

Nonparametric Form

- **Proposition 1:** Under Assumptions 1-4, for each group g, there exists a function λ_g such that

$$E\left(\epsilon_i | x_i, z_i, g_i = g; p\right) = \lambda_g(\pi_g).$$

  - **Example:** For $G = 2$,

$$\begin{aligned} & E(\epsilon_i | z_i, g_i = 1; p) \\ =& E(\epsilon_i | \eta_{1i} \geq p_1 - z_i^{'}\delta_1, \xi_{i2} - \xi_{i1} \leq z_i^{'}\beta_1 - z_i^{'}\beta_2 \text{ or } \eta_{2i} < p_2 - z_i^{'}\delta_2) \\ =& E(\epsilon_i | 1 - F_{\eta_1}(\eta_{1i}) \leq \pi_1, F_{\xi_2 - \xi_1}(\xi_{i2} - \xi_{i1}) < \pi_2 \text{ or } F_{\eta_2}(\eta_{2i}) < \pi_3) \end{aligned}$$

- **Corollary 1:** Under Assumptions 1-4, for each group g, there exists a function h_g such that

$$E\left(\epsilon_i | x_i, z_i, g_i = g; p\right) = h_g\left(z_i^{'}\delta^v, z_i^{'}\Delta\delta_g^u\right),$$

  where $\delta^v = \left(\delta_1^v, \ldots, \delta_G^v\right)$, and $\Delta\delta_g^u = \left(\delta_1^u - \delta_g^u, \ldots, \delta_G^u - \delta_g^u\right)$.

Identification

Parameters of interest

- Group formation

$$\begin{aligned} u_{ig} &= z_i^{'}\delta_g^u + \xi_{ig} \\ v_{gi} &= z_i^{'}\delta_g^v + \eta_{gi} \end{aligned}$$

- Social interactions

$$E\left(\boldsymbol{y}_n|\boldsymbol{x}_n,\boldsymbol{z}_n,\boldsymbol{w}_n\right) = \left(I - \gamma_1\boldsymbol{w}_n\right)^{-1}\left(\boldsymbol{w}_n\boldsymbol{x}_n\gamma_2 + \boldsymbol{x}_n\gamma_3 + \boldsymbol{h}(\boldsymbol{\tau}_n)\right),$$

where $\boldsymbol{\tau}_n = (\tau_1,\ldots,\tau_n)'$, $\tau_i = \left(z_i^{'}\delta^v, z_i^{'}\Delta\delta_{g_i}^u\right)'$.

Identification of δ

Exclusion restrictions

- Partition z_i such that $z_i = (x_i^{'}, z_{2,i}^{'}, z_{3,i}^{'})^{'}$, where $z_{2,i} = (z_{2,i1}, \cdots, z_{2,iG})'$ and $z_{3,i} = (z_{3,i1}, \cdots, z_{3,iG})'$
- $z_{2,ig}$ acts as a demand shifter

$$u_{ig} = z_{2,ig} + x_i^{'}\delta_g^u + \xi_{ig}$$

- $z_{3,ig}$ acts as a qualification shifter

$$v_{gi} = z_{3,ig} + x_i^{'}\delta_g^v + \eta_{gi}$$

Identification of δ

He, Sinha, and Sun (2020)

- **Assumption 5.** *Exclusion Restrictions, and at least one continuous covariate in each index.*
  - For each group g, $z_{2,ig}$ and $z_{3,ig}$ are continuous random variables.
  - $Supp(z_i)$ is not contained in a proper linear subspace of $\mathbb{R}^{d_z}$.
  - $z_{2,i1}, \cdots, z_{2,iG}, z_{3,i1}, \cdots, z_{3,iG}$ are linearly independent.

- **Theorem.** Under Assumptions 1-5, $\delta = (\delta^u, \delta^v)$ is identified.

Identification of γ

- The key is to identify the endogenous peer effect γ_1.

- For i and j in the same group g, the different effects of z_i on y_i and y_j can be used to cancel out h_g. For example,

$$\frac{\frac{\partial E(y_i|\boldsymbol{x}_n,\boldsymbol{z}_n,\boldsymbol{w}_n)}{\partial z_{2,ik}}}{\frac{\partial E(y_j|\boldsymbol{x}_n,\boldsymbol{z}_n,\boldsymbol{w}_n)}{\partial z_{2,ik}}} = \frac{-\left(I-\gamma_1\boldsymbol{w}_n\right)^{-1}_{ii}\frac{\partial h_g\left(z_i^{'}\delta^v,z_i^{'}\Delta\delta_g^u\right)}{\partial z_{2,ik}}}{-\left(I-\gamma_1\boldsymbol{w}_n\right)^{-1}_{ji}\frac{\partial h_g\left(z_i^{'}\delta^v,z_i^{'}\Delta\delta_g^u\right)}{\partial z_{2,ik}}} = \frac{\left(I-\gamma_1\boldsymbol{w}_n\right)^{-1}_{ii}}{\left(I-\gamma_1\boldsymbol{w}_n\right)^{-1}_{ji}}$$

- Therefore, γ_1 can be identified from the ratio equation.

- **Intuition**: Use friends' z_j in group formation as an IV for friends' y_j. The relevance of IV comes from the presence of the selection bias.

Identification of γ

- Once γ_1 is identified, the derivatives of h_g are identified using equations like

$$\frac{\partial E\left(y_i|\boldsymbol{x}_n,\boldsymbol{z}_n,\boldsymbol{w}_n\right)}{\partial z_{2,ik}} = -\left(I-\gamma_1\boldsymbol{w}_n\right)^{-1}_{ii}\frac{\partial h_g\left(z_i^{'}\delta^v, z_i^{'}\Delta\delta_g^u\right)}{\partial z_{2,ik}}.$$

  Assuming that h_g functions have full support, the values of h_g are identified up to a constant.

- γ_2 and γ_3 are identified as long as $\boldsymbol{w}_n\boldsymbol{x}_n$ and $\boldsymbol{x}_n$ are linearly independent.

- **Theorem.** Under Assumptions 1-5, γ is identified.

Estimation

First stage: Semiparametric estimation

- Based on the constructive identification results in He, Sinha, and Sun (2020)

- Construct system of equations in terms of Average Derivative Estimators (Powell, Stock, and Stoker, 1989)

$$\begin{aligned}\mathbb{E}_z\left[\frac{\partial \mathbb{P}(g_i = g|z_i)}{\partial z_i}\right] \quad &= \quad \delta_g^u \times \mathbb{E}_z\left[\frac{\partial \mathbb{P}(g_i = g|z_i)}{\partial z_{2i,g}}\right] \\ &\quad\quad + \delta_g^v \times \mathbb{E}_z\left[\frac{\partial \mathbb{P}(g_i = g|z_i)}{\partial z_{3i,g}}\right]\end{aligned}$$

- Estimator: $\hat{\delta} = (\hat{\delta^u}, \hat{\delta^v})$

Estimation

Second stage: Semiparametric estimation

- The social interaction model can be written as

$$y_i = X_i \gamma + h_{g_i}\left(\tau_{i,g_i}\right) + \nu_i,$$

  - X_i is the ith row of $\boldsymbol{X}_n = [\boldsymbol{w}_n \boldsymbol{y}_n, \boldsymbol{w}_n \boldsymbol{x}_n, \boldsymbol{x}_n]$
  - $\nu_i = \epsilon_i - \mathbb{E}\left[\epsilon_i | \boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{w}_n\right]$.

- Partial out the selection bias

$$y_i - \mathbb{E}\left[y_i | \tau_{i,g_i}\right] = \left(X_i - \mathbb{E}\left[X_i | \tau_{i,g_i}\right]\right) \gamma + \nu_i.$$

- We derive the conditional moment restrictions

$$\mathbb{E}\left[y_i - \mathbb{E}\left[y_i | \tau_{i,g_i}\right] - \left(X_i - \mathbb{E}\left[X_i | \tau_{i,g_i}\right]\right) \gamma | Z_i, \tau_{i,g_i}\right] = 0,$$

  where Z_i is the ith row of $\boldsymbol{Z}_n = [\boldsymbol{w}_n \boldsymbol{z}_n, \boldsymbol{w}_n \boldsymbol{x}_n, \boldsymbol{x}_n]$, i.e., the instruments.

Estimation

Second stage: Semiparametric estimation

- Stack the moment conditions

$$
\begin{aligned}
\mathbb{E}\left[y_i - s_{g_i}^{y}\left(\tau_{i,g_i}\right) - \left(X_i - s_{g_i}^{X}\left(\tau_{i,g_i}\right)\right)\gamma | Z_i, \tau_{i,g_i}\right] &= 0 \\
\mathbb{E}\left[y_i - s_{g_i}^{y}\left(\tau_{i,g_i}\right)|\tau_{i,g_i}\right] &= 0 \\
\mathbb{E}\left[X_i - s_{g_i}^{X}\left(\tau_{i,g_i}\right)|\tau_{i,g_i}\right] &= 0
\end{aligned}
$$

- We apply the efficient semiparametric two-step GMM in Ackerberg, Chen, Hahn and Liao (2014) to estimate γ.

Monte Carlo Simulations

- Market Parameters:
  - No. of students: 2000
  - No. of schools: 2
  - School capacities: 725, 750
- Student Preferences:

$$u_{ig} = \alpha_g + d_{ig} + \delta_s s_{ig} + \xi_{ig}$$

  - $\alpha_1 = 2$, $\alpha_2 = 4$. $\xi_{ig} \sim i.i.d.$ type I extreme value, outside option of value $\xi_{i0} \sim i.i.d.$ type I extreme value.
- Student Qualifications:

$$v_{gi} = w_{ig} + \delta_m m_i + \eta_{gi}$$

  - $\eta_{gi} = 2\epsilon_i + \tilde{\eta}_{gi}$, $\epsilon_i \sim i.i.d.\ N(0,1)$, $\tilde{\eta}_{gi} \sim i.i.d.\ N(0,1)$.
- Linear-in-means model:

$$y_i = (\overline{x}_{g_i})^{'} \beta + \epsilon_i$$

- $(m_i, x_i) \sim i.i.d.\ N\left(\begin{pmatrix} 6 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 8 \\ 8 & 81 \end{pmatrix}\right)$, $\begin{pmatrix} s_{i1} \\ s_{i2} \end{pmatrix} = \begin{pmatrix} \tilde{s}_{i1} \\ 0.25 * log(x_i) + \tilde{s}_{i2} \end{pmatrix}$ and $\tilde{s}_{ig} \sim i.i.d. N(5,6)$.
- Group formed by the deferred-acceptance algorithm.

Monte Carlo Simulations

Results

	Mean	Std
Bias of social effect β		
OLS	0.365	0.813
Sieve (order 2)	0.066	3.233
First-stage parameters (true value: 1)		
$\delta_{s,1}$	1.151	0.343
$\delta_{s,2}$	1.143	0.360
$\delta_{m,1}$	0.979	0.289
$\delta_{m,2}$	0.989	0.266

Conclusions

- This papers studies a linear-in-means social interaction model with endogenous peer groups. The endogeneity is due to the correlation between the unobservables in group formation and the unobservable in social interactions.
- Under a two-sided many-to-one matching framework, we explicitly characterize the group formation using a multivariate selection rule.
- We derive a nonparametric form of the selection bias, which depends on the preferences and qualification indices in group formation.
- We propose a two-stage distribution-free estimation strategy, where the first stage is semiparametric estimation of the matching model, and the second stage is semiparametric two-step GMM estimation of the social interaction model.
- The asymptotic distribution of our estimator is work-in-progress.