Estimating peer influence: limitations of linear-in-means models

Alex Hayes [1] Keith Levin [1]

[1]Department of Statistics, University of Wisconsin-Madison

Understanding social influence is a fundamental problem in a highly connected society

Contagion: if my friends get sick, I am more likely to get sick

Direct effect: if I get vaccinated, I am less likely to get sick

Interference: if my friends get vaccinated, I am less likely to get sick

The linear-in-means model is a canonical tool to estimate peer influence in network data

$$\underbrace{Y_i}_{\text{sick?}} = \alpha + \beta \underbrace{\frac{1}{d_i}\sum_{j\in\mathcal{N}(i)} Y_j}_{\substack{\text{fraction}\\ \text{sick}\\ \text{friends}}} + \gamma T_i + \delta \underbrace{\frac{1}{d_i}\sum_{j\in\mathcal{N}(i)} T_j}_{\substack{\text{fraction}\\ \text{vaccinated}\\ \text{friends}}} + \varepsilon_i$$

Data:

Outcome	(sick?)	$Y_i \in \{0,1\}$
Treatment	(vaccinated?)	$T_i \in \{0,1\}$
Edge $i \sim j$	(friends?)	$A_{ij} \in \{0,1\}$
Degree	(num friends)	$d_i \in \{0,1,2,\ldots\}$

Parameters:

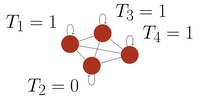
Base rate	$\alpha \in \mathbb{R}$
Contagion	$\beta \in (-1,1)$
Direct effect	$\gamma \in \mathbb{R}$
Interference	$\delta \in \mathbb{R}$

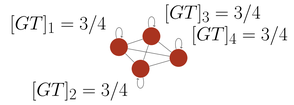
Letting $G = D^{-1}A$ be the row-normalized adjacency matrix, can express in matrix-vector form:

$$Y = \alpha 1_n + \beta GY + T\gamma + GT\delta + \varepsilon$$

Linear-in-means models are famously suspectible to the "reflection problem," an identification failure due to colinearity

In highly structured networks, the peer effect terms can be perfectly colinear, such that peer effects cannot be estimated from the data. For example, in a fully connected network:





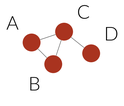
$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & GY_1 & 1 & 3/4 \\ 1 & GY_2 & 0 & 3/4 \\ 1 & GY_3 & 1 & 3/4 \\ 1 & GY_4 & 1 & 3/4 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Can't distinguish base rate α from interference δ due to colinearity

It's widely believed that colinearity problems can be avoided when certain identifying conditions, such as intransitivity, hold

Proposition: Suppose $\mathbb{E}[\varepsilon \mid T] = 0$ and that $|\beta| < 1$ and $\gamma\beta + \delta \neq 0$. For any fixed n, if I, G and G^2 are linearly independent (i.e., $aI + bG + cG^2 = 0$ only if $a = b = c = 0$), then α, β, γ and δ are identified. If I, G and G^2 are linearly dependent and no node is isolated, then $(\alpha, \beta, \gamma, \delta)$ are not identified.

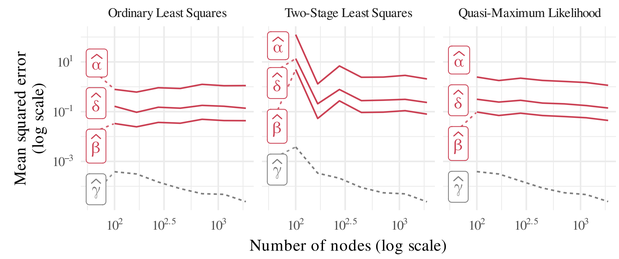
Intuition: Peer influence is identified when there are **open triangles** ("intransitivity") in the network



Closed: $A \leftrightarrow B \leftrightarrow C \leftrightarrow A$
Open: $B \leftrightarrow C \leftrightarrow D \nleftrightarrow B$

We show that peer effects can be inestimable even when identifying conditions hold

Three popular estimators all fail to estimate parameters at expected $\sqrt{n}$ rates in simulations, even though half of all possible triangles in the network are open



The problem is that peer effect terms are getting more and more colinear as the number of nodes in the network increases

Intuition: Imagine vaccination is a coin flip for every node (i.e., a Bernoulli design). As the network grows, the fraction of vaccinated friends ≈ 0.5 for every single node.

$$\lim_{d_i\to\infty} \underbrace{[GT]_i}_{\substack{\text{fraction}\\ \text{vaccinated}\\ \text{friends}}} = \lim_{d_i\to\infty} \underbrace{\frac{1}{d_i}\sum_{j\in\mathcal{N}(i)} T_j}_{\substack{\text{average of } d_i\\ \text{i.i.d. coin flips}}} = \frac{1}{2} \tag{1}$$

This also causes the contagion term to be near-colinear with the intercept, since GY is a repeated diffusion of T and GT over the network

$$\begin{aligned} Y &= \alpha 1_n + \beta GY + \gamma T + \delta GT + \varepsilon \\ &= (I - \beta G)^{-1}(\alpha 1_n + \gamma T + \delta GT + \varepsilon) \\ &= \sum_{k=0}^{\infty} \beta^k G^k (\alpha 1_n + \gamma T + \delta GT + \varepsilon) \end{aligned}$$

Lemma: Suppose that (1) the nodal covariates $T_1, T_2, \ldots$ are independent with shared mean $\tau \in \mathbb{R}$, and T is independent of A; (2) the nodal covariates are sub-gamma random variables; (3) the regression errors $\varepsilon_1, \varepsilon_2, \ldots$ are independent subgamma random variables independent of T.

If the minimum degree of the network grows at a $\omega(\log n)$ rate, then there exists $\eta \in \mathbb{R}$ such that

$$\max_{i\in[n]} \left|[GT]_i - \tau\right| = o(1) \text{ and } \max_{i\in[n]} \left|[GY]_i - \eta\right| = o(1) \text{ almost surely.}$$

Asymptotic colinearity leads to inconsistency because the signal-to-noise ratio gets worse and worse with growing sample size

Theorem: Under the same conditions as the Lemma, let $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta})$ be the vector of ordinary least squares estimates of $(\alpha, \beta, \gamma, \delta)$. Suppose that the degrees of the network are such that $\|G\|_F^2 = o(n)$. Then if $\beta = 0$,

$$\min\{|\hat{\alpha} - \alpha|, |\hat{\beta} - \beta|\} = \Omega_P(1)$$

and

$$|\hat{\delta} - \delta| = \Omega_P\left(\frac{1}{\|G\|_F}\right). \tag{2}$$

If $\beta \neq 0$,

$$\min\{|\hat{\alpha} - \alpha|, |\hat{\beta} - \beta|\} = \Omega_P\left(\frac{1}{\|G\|_F}\right).$$

Under the stronger growth assumption $\|G\|_F^2 = o(\sqrt{n})$, eq. (2) holds for all values of β.

Intuition: In networks with growing minimum degree, ordinary least squares estimates of α, β and δ are either inconsistent, or at best consistent at $\sqrt{n/d_{\min}}$ rates, where $d_{\min} = \min_{i\in[n]} d_i$.

When nodal covariates are strongly associated with network structure, it is sometimes possible to avoid asymptotic colinearity

Intuition: the fraction of vaccinated peers in eq. (1) might not converge to a column of constants if treatment T depends highly on position in the network A

Random dot product graphs: Suppose $X_1, \ldots, X_n \in \mathbb{R}^d$ are i.i.d. from a distribution F such that $0 \leq x^T y < 1$ for all $x, y \in \operatorname{supp} F$. Then $\mathbb{P}(A_{ij} \mid X_i, X_j) = X_i^T X_j$.

Theorem: Suppose that (A, X) are sampled from a random dot product model where $X \in \mathbb{R}^{n\times d}$ is full-rank with high probability. Let

$$Y = \alpha 1_n + \beta GY + X\gamma + GX\delta + \varepsilon \tag{3}$$

for $\alpha, \beta \in \mathbb{R}$ and $\gamma, \delta \in \mathbb{R}^d$. Suppose that X has $k \geq 2d$ distinct rows. Then, under suitable technical conditions, the columns of design matrix corresponding to $(\alpha, \beta, \delta_1, \delta_2, \ldots, \delta_d)$ are asymptotically colinear. If any two elements of $(\alpha, \beta, \delta_1, \delta_2, \ldots, \delta_d)$ are equal to zero, there is no asymptotic colinearity.

Simulation: In a network generated according to eq. (3) with no coefficients set to zero (*Unrestricted* model), there are still colinearity and estimation issues. However, when two coefficients from $(\alpha, \beta, \delta_1, \delta_2, \ldots, \delta_d)$ set to zero (*Restricted* model), popular estimators recover regression coefficients at expected $\sqrt{n}$ rates.

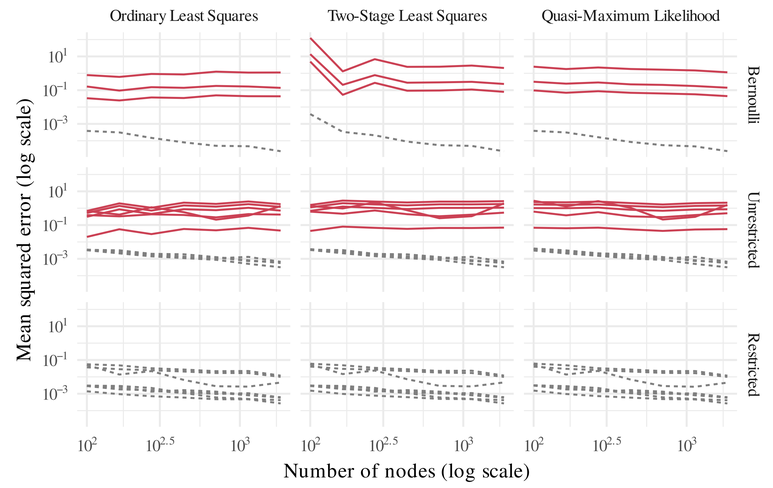


Figure 1. Red lines represent estimation error for asymptotically aliased regression coefficients, and gray lines represent estimation error for asymptotic un-aliased coefficients.

Takeaways

1. It can be impossible to estimate peer effects using linear-in-means models, even when all parameters in the model are identified. This is primarily an issue for Bernoulli designs in dense networks, or under models with growing minimum degree.
2. In observational data from random dot product models (like stochastic blockmodels), it may be possible to avoid colinearity issues by including latent network structure in the regression
3. In controlled experiments, use network-specific designs, like graph-cluster or ego-cluster randomization, to avoid the colinearity issues of Bernoulli designs

Want to learn more? Have a comment? Pre-print & contact info

Hayes, Alex and Keith Levin (Oct. 2024). *Peer Effects in the Linear-in-Means Model May Be Inestimable Even When Identified*. arXiv: `2410.10772 [stat]`. URL: `http://arxiv.org/abs/2410.10772` (visited on 10/15/2024).

`alex.hayes@wisc.edu`
`https://www.alexpghayes.com`