Linear-in-means models may be inestimable even when identified

Alex Hayes [1] Keith Levin [1]

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

The linear-in-means model is a canonical tool to estimate peer influence in social networks

$$Y_i = \alpha + \frac{\beta}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} Y_j + \gamma T_i + \frac{\delta}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} T_j + \varepsilon_i. \tag{1}$$

Data:

Network adjacency matrix	A	$\in \mathbb{R}^{n\times n}$
Edge $i \sim j$	A_{ij}	$\in \mathbb{R}$
Treatment	T_i	$\in \{0, 1\}$
Outcome	Y_i	$\in \mathbb{R}$
Confounders	$C_{i\cdot}$	$\in \mathbb{R}^p$
Friend group (latent)	$X_{i\cdot}$	$\in \mathbb{R}^d$

Parameters:

α	intercept	$\in \mathbb{R}$
β	contagion or "endogenous peer effect" or "network autoregression" effect of Y_j	$\in [-1, 1]$
γ	direct effect of T_i	$\in \mathbb{R}$
δ	interference or "exogeneous peer effect" or "contextual peer effect"	$\in \mathbb{R}$

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

TODO

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

Proposition: Fix n. Suppose $\mathbb{E}[\varepsilon \mid T] = 0$ and let

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

Suppose that $|\beta| < 1$ and $\gamma\beta + \delta \neq 0$. If I, G and G^2 are linearly independent in the sense that $aI + bG + cG^2 = 0$ 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.

Proposition: If G has three or more distinct eigenvalues, then I, G and G^2 are linearly independent.

Figure 1. Neighborhood averaging. (Left) A binary covariate T on small network. (Right) The average values of T in each node's neighborhood. For example, node A is connected to nodes B and C, the average value of T in neighborhood centered on A is $1/2$ (the value of T at node A is excluded from this calculation.). Similarly, the average value of T in the neighborhood centered on B is 1.

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

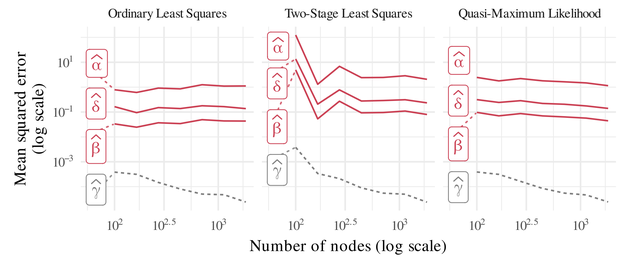

Figure 2. Mean squared error of estimates. Each panel corresponds to a different estimator. Within a panel, the x-axis represents the sample size on a log scale, and the y-axis represents the Monte Carlo estimate of mean squared error, also on a log scale. Each line corresponds to a single coefficient; solid red lines are asymptotically colinear, dashed gray lines are not.

The problem is that peer effects can become asymptotic colinear even when they are linearly independent for every finite sample size

Intuition: TODO COIN FLIP EXAMPLE

Lemma: Suppose that (1) the nodal covariates $T_1, T_2, \ldots, T_n$ are independent with shared mean $\tau \in \mathbb{R}$, and T is independent of A; (2) the centered nodal covariates $\{T_i - \tau : i \in 1, 2, \ldots, n\}$, are independent (ν, b)-subgamma random variables; (3) the regression errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent subgamma random variables with parameters not depending on n, and further are independent of $T_1, ..., T_n$; and (4) the adjacency matrix A contains only non-negative entries and does not contain any self-loops, such that $A_{ii} = 0$ for all $i = 1, 2, \ldots, n$.

If the degrees of the network grow such that

$$\max_{i\in[n]} \frac{1}{d_i^2}\sum_{j=1}^n A_{ij}^2 = o\left(\frac{1}{\nu \log^2 n}\right) \quad \text{and} \quad \max_{i,j\in[n]} \frac{A_{ij}}{d_i} = o\left(\frac{1}{b\log n}\right). \tag{2}$$

then

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

and

$$\max_{i\in[n]} \left\| [GY]_i - \frac{\alpha + (\gamma + \delta)\tau}{1 - \beta} \right\| = o(1) \text{ almost surely.}$$

Simulations and theory show that asymptotic colinearity causes estimators to be inconsistent

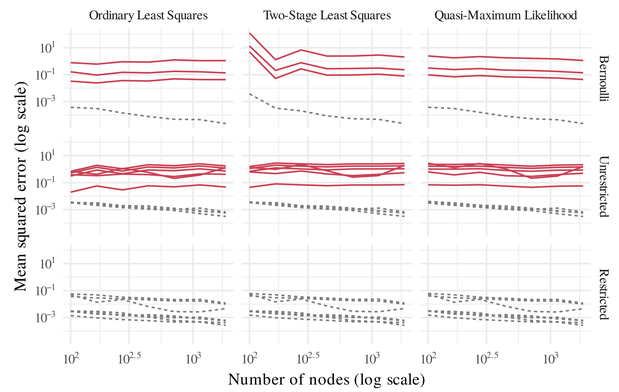

Figure 3. Mean squared error of estimates. Each row of panels denotes a different simulation setting, and each column of panels corresponds to a different estimator. Within a panel, the x-axis represents the sample size on a log scale, and the y-axis represents the Monte Carlo estimate of mean squared error, also on log scale. Each line corresponds to a single coefficient. Solid red lines are asymptotically colinear, dashed gray lines are not.

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

Intuition: avoid might not converge if T and A depend on one another

Random dot product graphs: Let F be a distribution on $\mathbb{R}^d$ such that $0 \leq x^T y$ for all $x, y \in \operatorname{supp} F$ and the convex cone of $\operatorname{supp} F$ is d-dimensional. Draw $X_1, X_2, \ldots, X_n$ independently and identically from F, and collect these in the rows of $X \in \mathbb{R}^{n\times d}$ for ease of notation. Conditional on these n vectors, which we call *latent positions*, generate edges by drawing $\{A_{ij} : 1 \leq i < j \leq n\}$ as independent (ν, b)-subgamma random variables with $\mathbb{E}[A_{ij} \mid X] = \rho X_i^T X_j$, where $\rho \in [0, 1]$. Then we say that A is distributed according to an n-vertex random dot product graph with latent position distribution F, (ν, b)-subgamma edges and sparsity factor ρ. We write $(A, X) \sim \text{RDPG}(F, n)$, with the subgamma and sparsity parameters made clear from the context.

Theorem: Suppose that (A, X) are sampled from a random dot product model where X is rank d with probability 1. Let ε be a vector of mean zero, i.i.d. $(\nu_\varepsilon, b_\varepsilon)$-subgamma random variables, with $(\nu_\varepsilon, b_\varepsilon)$ not depending on n, and let

$$Y = \alpha 1_n + \beta GY + X\gamma + GX\delta + \varepsilon$$

for $\alpha, \beta \in \mathbb{R}$ and $\gamma, \delta \in \mathbb{R}^d$, and the conditions of Proposition **??** hold. 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.

Takeaways

1. Peer effects in linear-in-means models
2. We show there is an asymptotic colinearity problem; this is not an identification failure. Rather, it is a signal-to-noise ratio failure, where the signal-to-noise ratio gets worse and worse despite identification.
3. Only show there is a problem in networks with growing minimum degree
4. It is possible for a parameter to be identified, but for there to be no consistent estimator of that parameter
5. This might explain why the linear-in-means model does poorly in simulations
6. One possible solution is to develop models, like latent space models, that account for dependence between nodal covariates and network structure
7. Another possible solution is to avoid bernoulli designs and using graph cluster randomized designs, ego-cluster designs, or multiple experiments

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`