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

Alex Hayes [1] Keith Levin [1]

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

## Abstract

Linear-in-means models are widely used to investigate peer effects. Identifying peer effects in these models is challenging, but conditions for identification are well-known. However, even when peer effects are identified, they may not be estimable, due to an asymptotic colinearity issue: as sample size increases, peer effects become more and more linearly dependent. We show that asymptotic colinearity occurs whenever nodal covariates are independent of the network and the minimum degree of the network is growing. Asymptotic colinearity can cause estimators to be inconsistent or to converge at slower than expected rates. We also demonstrate that dependence between nodal covariates and network structure can alleviate colinearity issues in random dot product graphs. These results suggest that linear-in-means models are less reliable for studying peer influence than previously believed.

## The linear-in-means model

The linear-in-means model is a canonical approach to estimating social influence in social networks. Suppose there is a network with $n$ nodes, encoded by a symmetric adjacency matrix $A \in \mathbb{R}^{n\times n}$. In binary networks, $A_{ij} = 1$ if nodes $i$ and $j$ form an edge, and $A_{ij} = 0$ otherwise, though we note that our results do not require that $A$ be binary. Each node $i$ is associated with an outcome $Y_i \in \mathbb{R}$ and a covariate $T_i \in \mathbb{R}$. Letting $\mathcal{N}(i) = \{j \in [n] : A_{ij} = 1\}$ denote the neighbors of node $i$ in the network, the treatment and outcome of the neighbors are allowed to influence the outcome of node $i$ as follows:

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

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:

| | | |
|---|---|---|
| $\alpha$ | intercept | $\in \mathbb{R}$ |
| $\beta$ | contagion or "endogenous peer effect" or "network autoregression" effect of $Y_j$ | $\in [-1, 1]$ |
| $\gamma$ | direct effect of $T_i$ | $\in \mathbb{R}$ |
| $\delta$ | interference or "exogeneous peer effect" or "contextual peer effect" | $\in \mathbb{R}$ |

The coefficient $\beta$, typically called the "contagion term", measures how peer outcomes $Y_j$ influence the outcome $Y_i$ at vertex $i$. This is variously referred to elsewhere in the literature as an "exogeneous spatial lag" a "spatial autoregression" or an "endogeneous peer effect" . Similarly, the coefficient $\delta$, typically called the "interference term", measures how peer treatments $T_j$ influence $i$'s outcome $Y_i$. Elsewhere in the literature, $\delta$ is variously referred to as a "contextual peer effect", an "exogeneous peer effect", a "spatial Durbin term," or a "spatially lagged X" term.

## Identifying conditions

### *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 $\alpha, \beta, \gamma$ and $\delta$ 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.*

## Degeneracy in a well-identified model


## Explanation of the degeneracy


Figure 2. 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.

## Semi-parametric network model

### *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 $(\nu, b)$-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*

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

*then*

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

*and*

$$\max_{i\in[n]} \left\| [GY]_i - \eta \right\| = o(1) \;\; \textit{almost surely,}$$

*where*

$$\eta = \frac{\alpha + (\gamma + \delta)\tau}{1 - \beta}. \tag{3}$$

### *Theorem*

*Let $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta})$ be the vector of ordinary least squares estimates of $(\alpha, \beta, \gamma, \delta)$, based on an $n$-by-$n$ network, and suppose that as $n$ grows, the sequence of networks is such that*

$$\|G\|_F^2 = o(n). \tag{4}$$

*Suppose that the adjacency matrix $A$ contains only non-negative entries and does not contain any self-loops, such that $A_{ii} = 0$ for all $i = 1, 2, \ldots, n$; the nodal covariates $T_1, T_2, \ldots, T_n$ are independent with shared mean $\tau \in \mathbb{R}$; the regression errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ and the centered nodal covariates $\{T_i - \tau : i = 1, 2, \ldots, n\}$ are independent sub-Gaussian random variables with parameters not depending on $n$; and the vectors $T$ and $\varepsilon$ are independent given $A$. Then if $\beta = 0$,*

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

*and*

$$|\hat{\delta} - \delta| = \Omega_P\left(\frac{1}{\|G\|_F}\right). \tag{5}$$

*If $\beta \neq 0$,*

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

*Finally, under the stronger growth assumption $\|G\|_F^2 = o(\sqrt{n})$, equation (5) continues to hold for all values of $\beta$.*

TODO

## Dependence between network and covariates can fix the degeneracy (partially)

### Definition (Random Dot Product Graph)

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

## Theory

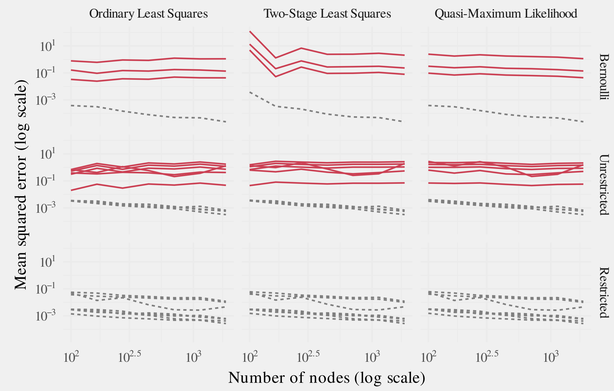


Figure 3. Mean squared error of estimates. Each row of panels denotes a different simulation setting, and each column of panels corresponds to a different estimator. Within a panel, the x-axis represents the sample size on a log scale, and the y-axis represents the Monte Carlo estimate of mean squared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

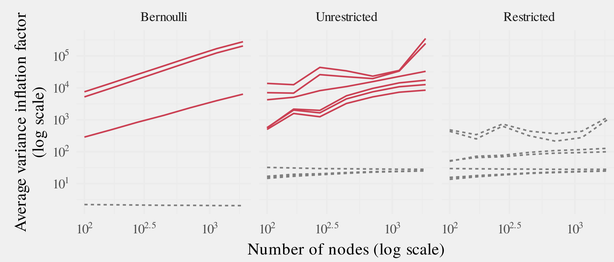


Figure 4. Average variance inflation factors. Each panel denotes a different simulation setting. The x-axis represents the sample size on a log scale, and the y-axis represents the Monte Carlo estimate of mean variance inflation factor, also on log scale. Each line corresponds to a single coefficient. Solid red lines are asymptotically colinear, dashed gray lines are not.

## Takeaways

1. Estimated effects are adjusted for possible confounding by age and church attendance.
2. Estimated effects vary with the chosen dimension $d$ of the latent space
3. Over-specifying $d$ is typically okay, but under-specifying $d$ leads to a failure to capture social structure in $X$
4. Once we capture enough social structure in $X$, we see a significant indirect social effect that leads adolescent girls to smoke more

## References & 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`