Electronic Journal of Statistics
Vol. 12 (2018) 2238–2252
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1448

Confidence intervals for linear unbiased estimators under constrained dependence

Peter M. Aronow

Department of Political Science, Yale University
Department of Biostatistics, Yale School of Public Health
Department of Statistics and Data Science, Yale University
Yale School of Management
87 Trumbull Street, New Haven, Connecticut, U.S.A.
e-mail: peter.aronow@yale.edu; *url:* http://aronow.research.yale.edu/

Forrest W. Crawford

Department of Biostatistics, Yale School of Public Health
Department of Ecology & Evolutionary Biology, Yale University
Yale School of Management
60 College St. New Haven, Connecticut, U.S.A
e-mail: forrest.crawford@yale.edu; *url:* www.crawfordlab.io

José R. Zubizarreta

Department of Health Care Policy, Harvard Medical School
Department of Statistics, Harvard University
180 Longwood Avenue Boston, Massachusetts, U.S.A
e-mail: zubizarreta@hcp.med.harvard.edu; *url:*
https://scholar.harvard.edu/zubizarreta

Abstract: We propose an approach for conducting inference for linear unbiased estimators applied to dependent outcomes given constraints on their independence relations, in the form of a dependency graph. We establish the consistency of an oracle variance estimator when a dependency graph is known, along with an associated central limit theorem. We derive an integer linear program for finding an upper bound for the estimated variance when a dependency graph is unknown, but topological or degree-based constraints are available on one such graph. We develop alternative bounds, including a closed-form bound, under an additional homoskedasticity assumption. We establish a basis for Wald-type confidence intervals that are guaranteed to have asymptotically conservative coverage.

Keywords and phrases: Dependency graph, oracle estimator, variance estimate.

Received December 2017.

Contents

1. Introduction

Researchers often encounter dependent data, where the exact nature of that dependence is unknown, and they wish to make inferences about a feature of the outcome distribution. When the observed data consist of many independent clusters, with possibly dependent outcomes within clusters, standard approaches often remain unbiased and consistent or can be adapted to yield consistent estimators [e.g. 14]. When many independent clusters are not available, current methods typically assume either independence of unit outcomes, or that the dependency structure is known or directly estimable [7, 24, 5, 22, 19]. In many cases, however, researchers may only have limited information about the nature of dependence between units, or perhaps only the number of other units on which a given unit's outcome depends.

Dependency graphs [3] represent a set of non-independence relationships in a set of variables [see also bidirected graphical models [21] and marginal independence models [9]]. Vertices represent individual units and edges represent the possibility of probabilistic dependence. Dependency graphs are useful because they imply marginal non-independence relations in a set of variables, and the class of joint distributions compatible with any dependency graph is flexible. When researchers have partial knowledge of independence relations for a set of variables, it is often easier to incorporate that knowledge into a topological constraint on a dependency graph than to impose restrictions on the space of joint distributions for the variables directly.

In this paper, we study the class of linear estimators that remain unbiased even when outcomes are neither independent nor identically distributed, including the case of ordinary least squares regression coefficients. We develop a framework for constructing confidence intervals for such linear estimators when applied to dependent outcomes, where independence relationships are unknown or partially known, but subject to topological constraints on their dependency graph. We seek an upper bound for the estimated variance of the sum using upper bounds for the degrees of each unit in a dependency graph and a local dependence assumption. We show that this optimization problem can be expressed as an integer linear program for the elements of the adjacency matrix corresponding to a dependency graph. We show that this approach yields asymptotically conservative Wald-type confidence intervals under a normal approximation. We also derive computationally simple bounds, including a closed-form bound, when the random variables are assumed to be homoskedastic.

2. Setting

Consider a simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with no parallel edges or self-loops. Let $|\mathcal{V}| = N$. Associated with each vertex $i \in \mathcal{V}$ is a random variable Y_i, and $\mathcal{G}$ characterizes probabilistic dependencies in the outcomes [e.g., 3].

Definition 1 (Dependency graph)**.** *$\mathcal{G}$ is a dependency graph if for all disjoint sets $\mathcal{V}_1, \mathcal{V}_2 \subset \mathcal{V}$ with no edge in $\mathcal{E}$ connecting a vertex in $\mathcal{V}_1$ to a vertex in $\mathcal{V}_2$, the set $\{Y_i : \ i \in \mathcal{V}_1\}$ is independent from the set $\{Y_j : \ j \in \mathcal{V}_2\}$.*

Suppose $\mathcal{G}$ is a dependency graph and we observe a subset $\mathcal{V}_S \subseteq \mathcal{V}$, where $|\mathcal{V}_S| = n$. Label these observed vertices $1, \dots, n$, and label the unobserved vertices in $\mathcal{V} \setminus \mathcal{V}_S$ arbitrarily by $n+1, \dots, N$. For each $i \in \mathcal{V}_S$, we observe the outcomes $Y_1, \dots, Y_n$, a fixed vector of coefficients $\theta = \theta_1, \dots, \theta_n$, and the degrees $d_i = |\{j : \ \{i,j\} \in \mathcal{E}\}|$ for each $i \in \mathcal{V}_S$.

Definition 2 (Induced subgraph)**.** *For a set of vertices $\mathcal{V}_S \subseteq \mathcal{V}$, the induced subgraph in $\mathcal{G}$ is $\mathcal{G}_S = (\mathcal{V}_S, \mathcal{E}_S)$, where $\mathcal{E}_S = \{\{i,j\} : i \in \mathcal{V}_S, \ j \in \mathcal{V}_S, \ \text{and } \{i,j\} \in \mathcal{E}\}$.*

Let $\mathcal{G}_S = (\mathcal{V}_S, \mathcal{E}_S)$ be the induced subgraph of the observed vertices $\mathcal{V}_S$. It follows that $\mathcal{G}_S$ is also a dependency graph. Let $\mathcal{G}_R = (\mathcal{V}_S, \mathcal{E}_R)$ be a subgraph of $\mathcal{G}_S$, consisting of all the observed vertices in $\mathcal{V}_S$, and a subset of the edges in $\mathcal{E}_S$.

Assumption 1 (Observed data)**.** *We observe the random outcomes $Y_1, \dots, Y_n$, the fixed degrees $d_1, \dots, d_n$, the fixed coefficients $\theta_1, \dots, \theta_n$, and the fixed recruitment graph $\mathcal{G}_R$.*

Let $Y = (Y_1, \dots, Y_n)$, $d = (d_1, \dots, d_n)$, and denote the observed data as $Z = (Y, d, \theta, \mathcal{G}_R)$.

Definition 3 (Linear unbiased estimator (LUE))**.** *An estimator $\hat{\beta}$ is a linear unbiased estimator (LUE) of β if*

$$\hat{\beta} = \frac{1}{n} \sum_{i \in \mathcal{V}_S} \theta_i Y_i$$

where $\mathrm{E}[\hat{\beta}] = \beta$.

Many familiar linear estimators are unbiased in settings with outcomes that are neither independent nor identically distributed. For example, consider the linear regression model $\mathrm{E}[Y_i] = \sum_{j=1}^{p} x_{ij}\beta_j$ for $i = 1, \dots, n$ [e.g., 10, Ch. 1], where x_i and β have dimension $p \times 1$, with the coefficients β estimated by ordinary least squares. Let X be the $n \times p$ matrix of covariates, and let Y be the $n \times 1$ vector of outcomes. The estimated coefficients are $\hat{\beta} = (X'X)^{-1}X'Y$.

Define the $p \times n$ matrix $\Theta = n(X'X)^{-1}X'$. Then for $j = 1, \dots, p$ we can express the vector coefficient estimates as $\hat{\beta} = n^{-1}\Theta Y$, or individually as

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} \Theta_{ji} y_i,$$

as in Definition 3. Additional examples include Horvitz-Thompson-type estimators for a finite population total [11, 20], and the difference-in-means estimator under random assignment of a treatment [18].

In what follows, we wish to conduct inference on β given Z. We proceed by constructing conservative estimators of

$$\text{var}(\hat{\beta}) = \frac{1}{n^2}\sum_{i \in \mathcal{V}_S}^{n}\sum_{j \in \mathcal{V}_S}^{n} \theta_i \theta_j \text{cov}(Y_i, Y_j).$$

We may use the square roots of these estimates as standard error estimators in order to construct Wald-type confidence intervals that are guaranteed to have asymptotic coverage for β at greater than or equal to nominal levels.

3. Variance estimation

The observed subgraph $\mathcal{G}_R$ may not reveal all the edges in $\mathcal{G}_S$ that connect observed vertices. We consider a class of variance estimators that depend on knowledge of $\mathcal{G}_S$, whose structure is represented by an $n \times n$ binary symmetric zero-diagonal adjacency matrix in which rows and columns are ordered by the indices $1, \dots, n$ of the vertices in $\mathcal{V}_S$. We now define some key concepts.

Definition 4 (Compatibility)**.** *The $n \times n$ binary symmetric adjacency matrix A is compatible with the observed data Z if for each $\{i, j\} \in \mathcal{E}_R$, $A_{ij} = A_{ji} = 1$, and for each $i \in \mathcal{V}_S$, $\sum_{j \in \mathcal{V}_S} A_{ij} \leq d_i$.*

The last condition in Definition 4 requires that the degree of i in the subgraph $\mathcal{G}_S$ not be greater than its degree in the full graph $\mathcal{G}$. Let $A^{\mathcal{O}} = \{A^{\mathcal{O}}_{ij}\}$ be the true $n \times n$ adjacency matrix of $\mathcal{G}_S$, where $A^{\mathcal{O}}_{ij} = 1$ if $\{i, j\} \in \mathcal{E}_S$ for $i, j \in \mathcal{V}_S$ and 0 otherwise. Let $\mathcal{A}(Z) = \{A : A \text{ is compatible with } Z\}$ in the sense of Definition 4; it is clear that $A^{\mathcal{O}} \in \mathcal{A}(Z)$.

Definition 5 (Oracle estimator)**.** *For a family of variance estimators $\widehat{V}(A; Z)$ defined for $A \in \mathcal{A}(Z)$, the oracle estimator is $\widehat{V}(A^{\mathcal{O}}; Z)$.*

For a variance estimator $\widehat{V}(A; Z)$, define the set $\mathcal{A}^m = \{A \in \mathcal{A}(Z) : \widehat{V}(A; Z) \text{ is maximized}\}$.

Definition 6 (Maximal compatible estimator)**.** *Let $A^m \in \mathcal{A}^m$. The maximal compatible estimator is $\widehat{V}(A^m; Z)$.*

The maximal compatible estimator provides a sharp upper bound for the oracle estimator because $\widehat{V}(A^{\mathcal{O}}; Z) \leq \widehat{V}(A^m; Z)$, with equality when $A^{\mathcal{O}} \in \mathcal{A}^m$.

We now describe an asymptotic scaling, along with boundedness conditions for outcome values and unit degrees. We will primarily rely on restrictions on a dependency graph to obtain root-n consistency, a central limit theorem, and convergence of the variance estimator. To this end, we describe an asymptotic regime in which the number of units (vertices in $\mathcal{G}$) is increasing, but the maximal dependence between their outcomes is bounded.

Assumption 2 (Asymptotic scaling)**.** *Consider the sequence $(\mathcal{G}, Z)_n$ of nested graphs $\mathcal{G}$ and observed data $Z = (\mathcal{G}_R, Y, d)$, where $\mathcal{G}_R = (\mathcal{V}_S, \mathcal{E}_R)$, $|\mathcal{V}_S| = n$, and $|\mathcal{V}| = N_n \geq n$. Assume there exist finite, positive constants c_1, c_2 such that for every element $(\mathcal{G}, Z)_n$, $\Pr(|\theta_i Y_i - \beta| > c_1) = 0, \forall i \in \mathcal{V}_S$ (bounded outcome values) and $\sum_{j \in \mathcal{V}_S} A^{\mathcal{O}}_{ij} \leq c_2, \forall i \in \mathcal{V}_S$ (bounded degrees in dependency graph). Further assume there exists a finite, positive constant c_3 such that $\lim_{n\to\infty} n\mathrm{var}(\hat{\beta}) = c_3$ (nondegenerate limiting variance). Finally, assume c_4 is a finite positive constant such that $|\theta_i| < c_4$.*

We will proceed by deriving oracle estimators under two sets of nested assumptions. We establish their asymptotic properties, then derive feasible estimators that dominate the oracle estimators.

3.1. General case

We first consider the case where we impose no distributional assumptions on the outcomes beyond the boundedness conditions of Assumption 2. Define the plug-in sample variance $\hat{\sigma}^2_\theta = n^{-1} \sum_{i \in \mathcal{V}_S} (\theta_i Y_i - \hat{\beta})^2$, and the estimator

$$\widehat{V}_1(A; Z) = \frac{1}{n^2} \left[n\hat{\sigma}^2_\theta + \sum_{i \in \mathcal{V}_S} \sum_{j \in \mathcal{V}_S} A_{ij} (\theta_i Y_i - \hat{\beta})(\theta_j Y_j - \hat{\beta}) \right]. \tag{3.1}$$

The corresponding oracle estimator $\widehat{V}_1(A^{\mathcal{O}}; Z)$ is consistent.

Proposition 1. *Under Assumption 2, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr(|n\widehat{V}_1(A^{\mathcal{O}}; Z) - n\mathrm{var}(\hat{\beta})| > \epsilon) = 0.$$

Proof. We follow the general proof strategy of Aronow and Samii [1] to show that when the true dependency structure $A^{\mathcal{O}}$ is known and degrees in the dependency graph are bounded, the sum of covariances of outcomes does not grow too quickly as $n \to \infty$. We will establish mean square convergence of $n\widehat{V}_1(A^{\mathcal{O}}; Z)$ to $n\mathrm{var}(\hat{\beta})$, allowing us to invoke Chebyshev's inequality to prove the proposition. Decompose $\hat{\sigma}^2_\theta = n^{-1} \sum_{i=1}^n \theta_i^2 Y_i^2 - n^{-2} \left(\sum_{i=1}^n \theta_i Y_i\right)^2$. Linearity of expectations implies $\mathrm{E}[\hat{\beta}] = \beta$ and $\mathrm{E}[n^{-1} \sum_{i=1}^n \theta_i^2 Y_i^2] = n^{-1} \sum_{i=1}^n \theta_i^2 \mathrm{E}[Y_i^2]$. Since Assumption 2 guarantees bounded outcomes, and the number of nonzero elements in the covariance matrix of outcome values is $O(n)$, $\mathrm{var}(\hat{\beta}) = O(n^{-1})$ and $\mathrm{var}(n^{-1} \sum_{i=1}^n \theta_i^2 Y_i^2) = O(n^{-1})$, yielding convergence of $\hat{\sigma}^2_\theta$.

Next we address convergence of the second term $n^{-1}\sum_{i\in\mathcal{V}_S}\sum_{j\in\mathcal{V}_S} A^{\mathcal{O}}_{ij}(\theta_i Y_i - \hat\beta)(\theta_j Y_j - \hat\beta)$. Asymptotic unbiasedness follows directly from linearity of expectations and $\text{var}(\hat\beta) = O(n^{-1})$. To establish mean square convergence, we consider the variance

$$\begin{aligned}
&\text{var}\left(\frac{1}{n}\sum_{i\in\mathcal{V}_S}\sum_{j\in\mathcal{V}_S} A^{\mathcal{O}}_{ij}(\theta_i Y_i - \hat\beta)(\theta_j Y_j - \hat\beta)\right)\\
&\quad = \frac{1}{n^2}\sum_{i,j,k,l\in\mathcal{V}_S} \text{cov}\left(A^{\mathcal{O}}_{ij}(\theta_i Y_i - \hat\beta)(\theta_j Y_j - \hat\beta), A^{\mathcal{O}}_{kl}(\theta_k Y_k - \hat\beta)(\theta_l Y_l - \hat\beta)\right)\\
&\quad = \frac{1}{n^2}\sum_{i,j,k,l\in\mathcal{V}_S} A^{\mathcal{O}}_{ij}A^{\mathcal{O}}_{kl}\text{cov}\left((\theta_i Y_i - \hat\beta)(\theta_j Y_j - \hat\beta), (\theta_k Y_k - \hat\beta)(\theta_l Y_l - \hat\beta)\right)
\end{aligned} \tag{3.2}$$

where the last line follows from bilinearity of covariance. Letting

$$\xi_{ijkl} = \text{cov}\big((\theta_i Y_i - \hat\beta)(\theta_j Y_j - \hat\beta), (\theta_k Y_k - \hat\beta)(\theta_l Y_l - \hat\beta)\big),$$

we now examine the conditions under which $\xi_{ijkl} \neq 0$. Expanding the covariance,

$$\begin{aligned}
\xi_{ijkl} &= \text{cov}\big((\theta_i Y_i - \hat\beta)(\theta_j Y_j - \hat\beta), (\theta_k Y_k - \hat\beta)(\theta_l Y_l - \hat\beta)\big)\\
&= \text{E}\big[(\theta_i Y_i - \hat\beta)(Y_j - \hat\beta)(\theta_k Y_k - \hat\beta)(\theta_l Y_l - \hat\beta)\big]\\
&\quad - \text{E}\big[(\theta_i Y_i - \hat\beta)(\theta_j Y_j - \hat\beta)\big]\text{E}\big[(\theta_k Y_k - \hat\beta)(\theta_l Y_l - \hat\beta)\big]\\
&= \theta_i\theta_j\theta_k\theta_l\text{E}[Y_iY_jY_kY_l] - \theta_i\theta_j\theta_k\text{E}[Y_iY_jY_k\hat\beta] - \theta_i\theta_j\theta_l\text{E}[Y_iY_jY_l\hat\beta]\\
&\quad - \theta_i\theta_k\theta_l\text{E}[Y_iY_kY_l\hat\beta] - \theta_j\theta_k\theta_l\text{E}[Y_jY_kY_l\hat\beta] + \theta_i\theta_j\text{E}[Y_iY_j\hat\beta^2]\\
&\quad + \theta_i\theta_k\text{E}[Y_iY_k\hat\beta^2] + \theta_i\theta_l\text{E}[Y_iY_l\hat\beta^2] + \theta_j\theta_k\text{E}[Y_jY_k\hat\beta^2] + \theta_j\theta_l\text{E}[Y_jY_l\hat\beta^2]\\
&\quad + \theta_k\theta_l\text{E}[Y_kY_l\hat\beta^2] - \theta_i\text{E}[Y_i\hat\beta^3] - \theta_j\text{E}[Y_j\hat\beta^3] - \theta_k\text{E}[Y_k\hat\beta^3] - \theta_l\text{E}[Y_l\hat\beta^3]\\
&\quad + \text{E}[\hat\beta^4] - \theta_i\theta_j\theta_k\theta_l\big[\text{E}[Y_iY_j]\text{E}[Y_kY_l] - \theta_i\theta_j\theta_k\text{E}[Y_iY_j]\text{E}[Y_k\hat\beta]\\
&\quad - \theta_i\theta_j\theta_l\text{E}[Y_iY_j]\text{E}[Y_l\hat\beta] + \theta_i\theta_j\text{E}[Y_iY_j]\text{E}[\hat\beta^2] - \theta_k\theta_l\text{E}[Y_i\hat\beta]\text{E}[Y_kY_l]\\
&\quad + \theta_i\theta_l\text{E}[Y_i\hat\beta]\text{E}[Y_l\hat\beta] + \theta_i\theta_k\text{E}[Y_i\hat\beta]\text{E}[Y_k\hat\beta] - \theta_i\text{E}[Y_i\hat\beta]\text{E}[\hat\beta^2]\\
&\quad - \theta_j\theta_k\theta_l\text{E}[Y_j\hat\beta]\text{E}[Y_kY_l] + \theta_j\theta_l\text{E}[Y_j\hat\beta]\text{E}[Y_l\hat\beta] + \theta_j\theta_k\text{E}[Y_j\hat\beta]\text{E}[Y_k\hat\beta]\\
&\quad - \theta_j\text{E}[Y_j\hat\beta]\text{E}[\hat\beta^2] + \theta_k\theta_l\text{E}[\hat\beta^2]\text{E}[Y_kY_l] - \theta_k\text{E}[\hat\beta^2]\text{E}[Y_k\hat\beta] - \theta_l\text{E}[\hat\beta^2]\text{E}[Y_l\hat\beta]\\
&\quad + \text{E}[\hat\beta^2]\text{E}[\hat\beta^2]\big]
\end{aligned} \tag{3.3}$$

Then by root-n consistency of means and Slutsky's Theorem, as $n \to \infty$ expectations involving $\hat\beta$ factorize, yielding, e.g. $\theta_i\text{E}(Y_i\hat\beta) = \theta_i\text{E}(Y_i)\beta + O(n^{-1})$. We therefore combine terms and rewrite (3.3) as

$$\begin{aligned}
\xi_{ijkl} &= \theta_i\theta_j\theta_k\theta_l\text{cov}(Y_iY_j, Y_kY_l)\\
&\quad - \beta\big(\text{cov}(\theta_iY_i\theta_jY_j, \theta_kY_k) + \text{cov}(\theta_iY_i\theta_jY_j, \theta_lY_l)
\end{aligned}$$

$$
\begin{aligned}
&\quad + \operatorname{cov}(\theta_i Y_i, \theta_k Y_k \theta_l Y_l) + \operatorname{cov}(\theta_j Y_j, \theta_k Y_k \theta_l Y_l)\big) \\
&\quad + \beta^2 \big(\operatorname{cov}(\theta_i Y_i, \theta_k Y_k) + \operatorname{cov}(\theta_i Y_i, \theta_l Y_l) \\
&\quad + \operatorname{cov}(\theta_j Y_j, \theta_k Y_k) + \operatorname{cov}(\theta_j Y_j, \theta_l Y_l)\big) + O(n^{-1}) \\
&= \xi'_{ijkl} + O(n^{-1}), \qquad (3.4)
\end{aligned}
$$

where the limiting covariance is denoted ξ'_{ijkl}. This can only be nonzero if at least one of the covariance terms in (3.4) is nonzero. Since $\mathcal{G}_S$ is a dependency graph, this condition is only met when there exists at least one edge between a vertex in the set $\{i, j\}$ and a vertex in the set $\{k, l\}$. Therefore $A^{\mathcal{O}}_{ij} A^{\mathcal{O}}_{kl} \xi'_{ijkl}$ can only be nonzero if

$$\{A^{\mathcal{O}}_{ij} = A^{\mathcal{O}}_{kl} = 1\} \text{ and } \left(\{A^{\mathcal{O}}_{ik} = 1\} \text{ or } \{A^{\mathcal{O}}_{il} = 1\} \text{ or } \{A^{\mathcal{O}}_{jk} = 1\} \text{ or } \{A^{\mathcal{O}}_{jl} = 1\}\right).$$

By Assumption 2, the degree of each vertex in $\mathcal{V}_S$ is bounded by c_2, so the condition is satisfied by at most $4nc_2^3$ terms in the summation in (3.2). In addition, we may compute the remainder term $\sum_{i,j,k,l \in \mathcal{V}_S} A^{\mathcal{O}}_{ij} A^{\mathcal{O}}_{kl} (\xi_{ijkl} - \xi'_{ijkl}) = \sum_{i,j,k,l \in \mathcal{V}_S} A^{\mathcal{O}}_{ij} A^{\mathcal{O}}_{kl} O(n^{-1}) = O(n)$, thus both terms are $O(n)$ before dividing by n^2. Therefore $\operatorname{var}\left(n^{-1} \sum_{i \in \mathcal{V}_S} \sum_{j \in \mathcal{V}_S} A^{\mathcal{O}}_{ij} (\theta_i Y_i - \hat{\beta})(\theta_j Y_j - \hat{\beta})\right) = O(n^{-1})$ and the result follows. □

Proposition 1 is applicable to problems where a dependency graph is known, as it provides a basis for consistent variance estimation, generalizing results for special cases [7, 2]. We now address the case where the true subgraph $\mathcal{G}_S$ is not known, but constraints on the graph are available.

Let $\mathcal{A}^m_1 = \{A \in \mathcal{A}(Z) : \widehat{V}_1(A; Z) \text{ is maximized}\}$ be the set of compatible adjacency matrices that maximize $\widehat{V}_1(A; Z)$. Let $\Theta = \operatorname{diag}(\theta)$ be the $n \times n$ matrix with $\Theta_{ii} = \theta_i$ and $\Theta_{ij} = 0$ for $i \neq j$. We can find an element A^m of $\mathcal{A}^m_1$ by solving the 0-1 integer linear program

$$
\begin{aligned}
&\underset{A}{\text{maximize}} \quad (\Theta Y - \hat{\beta})' A (\Theta Y - \hat{\beta}) \\
&\text{subject to} \quad A1 \preceq d, \quad A \succeq A_R,
\end{aligned}
\qquad (3.5)
$$

where A_R is the adjacency matrix of $\mathcal{G}_R$ and $\preceq$ denotes the element-wise “less-than” relation.

Since A is an adjacency matrix, we can reduce the program and maximize over the decision variables that correspond to the upper or lower triangular elements of A only. Let $\hat{v}_{ij}$ be the ijth element of the sample covariance matrix with $i = 1, ..., n$ and $j = 1, ..., n$. Since the sample covariance matrix is symmetric, we can focus on its upper triangular part and use the decision variable $a_{ij} = 1$ if $\hat{v}_{ij} \neq 0$, and 0 otherwise, for each $i < j$. Based on these decision variables, the integer linear program (3.5) can be written as

$$\begin{aligned} \underset{\boldsymbol{a}}{\text{maximize}} \quad & \sum_{i=1}^{n}\sum_{j=i+1}^{n} \hat{v}_{ij}a_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{i-1} a_{ji} + \sum_{j=i+1}^{n} a_{ij} \leq d_i,\ i = 1, ..., n, \\ & a_{ij} \in \{0,1\},\ i = 1, ..., n,\ j = 1, ..., n,\ i < j, \end{aligned} \tag{3.6}$$

where d_i is the degree, and further simplified with the constraints $A \succeq A_R$ that make some of the decision variables a_{ij} automatically equal to one. The resulting program has at most $n(n-1)/2$ decision variables and in general it is a multidimensional knapsack problem [12]. See chapter 9 of Kellerer et al 13 for an overview of the multidimensional knapsack problem.

This program (3.5) is NP-hard, but it admits a polynomial time approximation scheme (PTAS). Typical PTAS depend heavily on the size of the problem and their running time can be very high (see, e.g., section 9.4.2 of Kellerer et al 12). In spite of this, in standard practice, for example with 1000 observations or less, problem (3.5) can be solved in a few seconds with modern optimization solvers such as Gurobi. To obtain a solution within a provably small optimality gap, these solvers use a variety of techniques, including: linear programming and branch-and-bound procedures to reduce the set of feasible solutions; presolve routines applied prior to the branch-and-bound procedures to reduce the size of the problem; cutting planes methods to remove fractional solutions and tighten the formulation; and a collection of heuristics to find good incumbent solutions in the branch-and-bound [4, 15, 17].

While the true adjacency matrix $A^{\mathcal{O}}$ is not known, an element $A^m \in \mathcal{A}_1^m$ produces a variance estimate $\widehat{V}_1(A^m, Z)$ that is at least as large as the oracle estimator $\widehat{V}_1(A^{\mathcal{O}}; Z)$. As n grows large, the variance estimate $\widehat{V}_1(A^m, Z)$ is conservative: the probability that $n\widehat{V}_1(A^m)$ underestimates $n\text{var}(\hat{\beta})$ by more than $\epsilon > 0$ tends to zero.

Corollary 1. *Given Assumption 2,* $\lim_{n\to\infty} \Pr(n\text{var}(\hat{\beta}) - n\widehat{V}_1(A^m; Z) > \epsilon) = 0$ *for any* $\epsilon > 0$.

Proof is given in the Appendix. Corollary 1 does not imply consistency of $\widehat{V}_1(A^m; Z)$ as an estimator of $\text{var}(\hat{\beta})$, nor does it imply that the estimator converges to any particular limiting value. Rather we have established that, for large n, its distribution will tend to be at least as large as the true variance.

3.2. Alternative bounds under homoskedasticity

When all variances are equal, we can obtain an alternative closed-form bound that is computationally simpler and less sensitive to between-sample variability in the empirical variance-covariance matrix. This estimator essentially only depends on the estimated variance of unit outcomes and the maximum number of edges in a dependency graph.

Assumption 3 (Homoskedasticity)**.** $\text{var}(\theta_i Y_i) = \text{var}(\theta_j Y_j), \forall i, j \in \mathcal{V}$.

Under homoskedasticity, the general estimator $\widehat{V}_1(A^m, Z)$ developed in Section 3.1 provides a conservative variance estimate. A bound that is relatively computationally simple to compute can be derived by noting that when $\text{var}(\theta_i Y_i) = \sigma_\theta^2$, $\text{cov}(\theta_i Y_i, \theta_j Y_j) \leq \sigma_\theta^2 A_{ij}^{\mathcal{O}}$. Define the estimator

$$\widehat{V}_2(A; Z) = \frac{\hat{\sigma}_\theta^2}{n}\left[1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S}\sum_{j\in\mathcal{V}_S} A_{ij}\right]. \tag{3.7}$$

The oracle estimator $\widehat{V}_2(A^{\mathcal{O}}, Z)$ need not be consistent, though it is asymptotically conservative.

Proposition 2. *Given Assumptions 2 and 3,* $\lim_{n\to\infty} \Pr(n\text{var}(\hat{\beta}) - n\widehat{V}_2(A^{\mathcal{O}}; Z) > \epsilon) = 0$ *for any* $\epsilon > 0$.

Proof. We first define an alternative oracle estimator which presumes knowledge of the correlations ρ_i,

$$\widehat{V}_2^*(A^{\mathcal{O}}; Z) = \frac{\hat{\sigma}_\theta^2}{n}\left[1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S}\sum_{j\in\mathcal{V}_S} A_{ij}^{\mathcal{O}}\rho_i\right].$$

Multiplying by n, $n\widehat{V}_2^*(A^{\mathcal{O}}; Z) = \hat{\sigma}_\theta^2\left[1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S}\sum_{j\in\mathcal{V}_S} A_{ij}^{\mathcal{O}}\rho_i\right]$. As in the proof of Proposition 1, $\hat{\sigma}_\theta^2$ converges in mean square. By Assumption 2, $1 \leq 1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S}\sum_{j\in\mathcal{V}_S} A_{ij}^{\mathcal{O}} \leq 1 + c_2$, allowing us to invoke Slutsky's Theorem and Chebyshev's Inequality to show $\lim_{n\to\infty} \Pr(|n\widehat{V}_2^*(A^{\mathcal{O}}; Z) - n\text{var}(\hat{\beta})| < \epsilon) = 0$. The Cauchy-Schwarz Inequality (i.e., all $\rho_i \leq 1$) implies $\widehat{V}_2^*(A^{\mathcal{O}}; Z) \leq \widehat{V}_2(A^{\mathcal{O}}; Z)$ across all sample realizations. The result follows directly. □

As before, we can maximize the estimator $\widehat{V}_2(A; Z)$ over the family of compatible dependency graphs. Define $\mathcal{A}_2^m = \{A \in \mathcal{A}(Z) : \widehat{V}_2(A; Z) \text{ is maximized}\}$, and let $A^m \in \mathcal{A}_2^m$. To find an element of $\mathcal{A}_2^m$, we solve the 0-1 integer linear program

$$\begin{aligned} &\underset{A}{\text{maximize}} && 1'A1 \\ &\text{subject to} && A1 \preceq d, \quad A \succeq A_R, \end{aligned} \tag{3.8}$$

where again A is an arbitrary 0-1 adjacency matrix and A_R is the adjacency matrix of $\mathcal{G}_R$. In order to solve the program (3.8), let $\hat{v}_{ij} = 1$ for every $i = 1, ..., n$ and $j = 1, ..., n$ in (3.6). Note that finding the solution to this problem does not depend on the empirical variance-covariance matrix; the variability of the estimator $\widehat{V}_2(A^m; Z)$ is purely attributable to estimation error in $\hat{\sigma}_\theta^2$.

Since $\widehat{V}_2(A; Z)$ relies only on the number of positive entries in A, we can derive a looser closed-form upper bound by considering the maximum number of edges that can be in E_S. For $i \in \mathcal{V}_S$, let $d_i' = \min\{d_i, n-1\}$ be the degree of

i in $\mathcal{G}$, truncated at $n-1$. Let

$$\widehat{V}_2'(Z) = \frac{\hat{\sigma}_\theta^2}{n}\left[1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S} d_i'\right]. \tag{3.9}$$

The estimator (3.9) does not depend on any particular compatible adjacency matrix.

Lemma 1. *Let $A^m \in \mathcal{A}^m$. Then $\widehat{V}_2(A^{\mathcal{O}}, Z) \le \widehat{V}_2(A^m; Z) \le \widehat{V}_2'(Z)$, with $\widehat{V}_2(A^m; Z) = \widehat{V}_2'(Z)$ when $d_i' = \sum_{j\in\mathcal{V}_S} A_{ij}^m$ for each $i \in \mathcal{V}_S$.*

Proof is given in the Appendix. The upper bound estimators under homoskedasticity are asymptotically conservative.

Corollary 2. *Given Assumptions 2 and 3, then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr(n\text{var}(\hat{\beta}) - n\widehat{V}_2(A^m; Z) > \epsilon) = 0,$$
$$\lim_{n\to\infty} \Pr(n\text{var}(\hat{\beta}) - n\widehat{V}_2'(Z) > \epsilon) = 0.$$

The proof follows from Lemma 1 and the same reasoning employed in the proof of Corollary 1.

4. Wald-type confidence intervals

We now prove that our variance estimates can be used to form valid Wald-type confidence intervals about β. First, we establish a central limit theorem for $\hat{\beta}$ given our asymptotic scaling.

Lemma 2. *Given Assumption 2, $\left(\hat{\beta} - \beta\right) \Big/ \sqrt{\text{var}(\hat{\beta})} \to_d N(0, 1)$.*

Lemma 2, a standard result in applying Stein's method to the setting of local dependence, has been proven by, e.g., Theorem 2.7 of Chen et al [6]. Similarly, we reiterate the well-known basis for Wald-type confidence intervals.

Lemma 3. *Given Assumption 2, if a variance estimator $\widehat{V}(A; Z)$ satisfies*

$$\lim_{n\to\infty} \Pr(|n\widehat{V}(A; Z) - n\text{var}(\hat{\beta})| > \epsilon) = 0,$$

then confidence intervals formed as $\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\widehat{V}(A; Z)}$ will have $100(1-\alpha)\%$ coverage for β in large n.

Lemma 3 follows directly from Lemma 2 and Slutsky's Theorem. We now establish the validity of confidence intervals constructed via Lemma 3.

Proposition 3. *Given Assumption 2, if a variance estimator $\widehat{V}(A; Z)$ satisfies*

$$\lim_{n\to\infty} \Pr(n\text{var}(\hat{\beta}) - n\widehat{V}(A; Z) > \epsilon) = 0,$$

then confidence intervals formed as $\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\widehat{V}(A;Z)}$ *will have at least* $100(1-\alpha)\%$ *coverage for* β *in large* n*.*

Proof. Define a random variable U such that $U = \widehat{V}(A;Z)$ if $\widehat{V}(A;Z) \leq \text{var}(\hat{\beta})$ and $\text{var}(\hat{\beta})$ otherwise. Then $\lim_{n\to\infty} \Pr(|nU - n\text{var}(\hat{\beta})| > \epsilon) = 0$, and by Lemma 3 Wald-type confidence intervals formed with U as a variance estimate will have at least proper coverage. Across every sample realization, $\widehat{V}(A;Z) \geq U$, and thus the coverage of Wald-type confidence intervals using $\widehat{V}(A;Z)$ will be also be at least proper levels. □

It follows that Wald-type confidence intervals constructed using the conservative variance estimators derived in Section 3 yield conservative asymptotic coverage.

Corollary 3. *Given Assumption 2, confidence intervals formed as* $\hat{\beta} \pm z_{1-\alpha/2}$ $\sqrt{\widehat{V}_1(A^m)}$ *have at least* $100(1-\alpha)\%$ *coverage for* β *in large* n*.*

Corollary 4. *Given Assumptions 2 and 3, confidence intervals formed as* $\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\widehat{V}_2(A^m;Z)}$ *or* $\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\widehat{V}_2'(Z)}$ *have at least* $100(1-\alpha)\%$ *coverage for* β *in large* n*.*

Proofs for Corollaries 3 and 4 follow directly from Corollaries 1 and 2 and Proposition 3. Upper bounds for the variance estimates can be obtained by solving a relaxed form of the programs (3.5) and (3.8). By Proposition 3, using such upper bounds as a basis for conservative inference will also yield valid confidence intervals. In practice, the results obtained by modern optimization solvers will be tighter with a provably small optimality gap and thus in problems of moderate size will typically be preferable.

5. Discussion

We have developed conservative estimators for the variance of a linear unbiased estimator under partial observation of a dependency graph and assumptions about the variance of individual outcomes. The variance estimation setting we address here can accommodate a wide variety of dependency and observation assumptions. For example, Assumption 1, which states that we observe $Z = (Y, d, \mathcal{G}_R)$, can be weakened when $\mathcal{G}_R$ is completely unknown. In this case the constraint in the integer linear programs (3.5) and (3.8) becomes $A \succeq 0$ where 0 is the $n \times n$ matrix of all zeros; this constraint is met for all adjacency matrices A, so it becomes superfluous. Alternatively, we may not have full knowledge of the degrees $d = (d_1, \ldots, d_n)$, and instead have only an upper bound d_i^* for each d_i, or a global upper bound $d_i \leq d^*$ for all $i = 1, \ldots, n$. Conservative variance estimation in both of these cases can be achieved (by substituting d_i^* or d^* for d_i) with no change to the programs (3.5) and (3.8) or to the asymptotic results given here. When no information about $\mathcal{G}_R$ or the degrees d is available, setting every $d_i = d^* = n - 1$ delivers a maximally conservative upper bound.

We note here three extensions. First, it is likely possible to extend our results to obtain confidence intervals more generally for asymptotically linear estimators using an empirical analogue of the variance of the influence function as the objective function. Second, a generalization of our results may facilitate conservative inference for causal estimands under interference between units [e.g., 23, 16], given interference that can be characterized by a constrained dependency graph. Finally, the general logic of our approach — maximizing an estimator over a space of graphical structural relations [e.g. 8] — may be profitably be used to obtain conservative interval (or region) estimates for network functionals, including with alternative representations of constraints on the dependency structure (e.g., restrictions compatible with Markov random field-type assumptions). Furthermore, extensions of our approach may not require that constraints involving the adjacency matrix A be linear. More generally, if the constraint is of the form $f(A) \leq 0$, and the system can be solved, it is straightforward to conceptualize a broader class of inferential procedures involving network-topological constraints beyond maximum degree, e.g. triangles, diameter, or clustering.

Acknowledgement

Forrest W. Crawford was supported by NIH grants NICHD DP2 OD022614, NCATS KL2 TR000140, and NIMH P30 MH062294. José R. Zubizarreta acknowledges support from a grant from the Alfred P. Sloan Foundation. We are grateful to Robert Heimer for helpful comments and for providing the SATHCAP data, funded by NIH/NIDA grant U01 DA017387. We also thank Jonathon Baron, Daniel Bienstock, Sarah Hamerling, Edward H. Kaplan, Winston Lin, Luke W. Miratrix, Molly Offer-Westort, Lilla Orr, Cyrus Samii, Thomas Richardson, Jiacheng Wu, and Fan Xia for valuable comments. We express special thanks to Sahand Negahban for important early discussions regarding the formulation of the problem.

Appendix A: Proofs

Proof of 1. Across all sample realizations, $\widehat{V}_1(A^m;Z) \geq \widehat{V}_1(A^{\mathcal{O}};Z)$. Then by Proposition 1,

$$\begin{aligned}
&\lim_{n\to\infty} \Pr(n\mathrm{var}(\hat{\beta}) - n\widehat{V}_1(A^m;Z) > \epsilon) \\
&\leq \lim_{n\to\infty} \Pr(n\mathrm{var}(\hat{\beta}) - n\widehat{V}_1(A^m;Z) + n\widehat{V}_1(A^m;Z) - n\widehat{V}_1(A^{\mathcal{O}};Z) > \epsilon) \\
&= \lim_{n\to\infty} \Pr(n\mathrm{var}(\hat{\beta}) - n\widehat{V}_1(A^{\mathcal{O}};Z) > \epsilon) \\
&= 0,
\end{aligned} \tag{A.1}$$

as claimed. □

Proof of Lemma 1. By definition, $\widehat{V}_2(A;Z) \leq \widehat{V}_2(A^m;Z)$ for every $A \in \mathcal{A}$. Since $A^{\mathcal{O}} \in \mathcal{A}$, it follows that $\widehat{V}_2(A^{\mathcal{O}},Z) \leq \widehat{V}_2(A^m;Z)$. Now let $d_i^m = \sum_{j\in\mathcal{V}_S} A_{ij}^m$ be the degree of i in the adjacency matrix A^m, and note that for every $i \in \mathcal{V}_S$, $d_i^m \leq d_i'$. Then

$$\begin{aligned}
\widehat{V}_2(A^m; Z) &= \frac{\hat{\sigma}_\theta^2}{n}\left[1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S}\sum_{j\in\mathcal{V}_S} A_{ij}^m\right] \\
&= \frac{\hat{\sigma}_\theta^2}{n}\left[1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S} d_i^m\right] \\
&\leq \frac{\hat{\sigma}_\theta^2}{n}\left[1 + \frac{1}{n}\sum_{i\in\mathcal{V}_S} d_i'\right] \\
&= \widehat{V}_2'(Z)
\end{aligned}$$

as claimed. Now consider $A^m \in \mathcal{A}^m$ with the property that $\sum_{j\in\mathcal{V}_S} A_{ij}^m = d_i' = \min\{d_i, n-1\}$ for all $i \in \mathcal{V}_S$. Then $\widehat{V}_2(A^m; Z) = \widehat{V}_2'(Z)$, as claimed. □

Appendix B: Statistical software implementation

We implement this approach in the new statistical software package `depinf` for `R`. `depinf` includes two basic functions: `depgraph`, for finding an adjacency matrix that maximizes the estimated variance of $\hat{\beta}$ given general constraints on the degree of dependence of the observations (these are problems (3.5) and (3.8) above), and `depvar` for calculating the variance estimates (3.1) and (3.7). In both `depgraph` and `depinf`, we give the option to find an exact solution to (3.5) and (3.8) via integer programming, or an approximate solution to the relaxations of (3.5) and (3.8) via linear programming. Naturally, the running time of the approximate solution is lower, but it provides a more conservative variance estimate. In order to solve (3.5) and (3.8), `depgraph` can use three different optimization solvers: CPLEX, GLPK and Gurobi. By default, `depgraph` uses GLPK, which can be downloaded from the `R` repository CRAN. To solve large instances of the problem exactly, we strongly recommend using either CPLEX or Gurobi, which are much faster but require a license and special installation. The `depinf` package can be obtained at `https://github.com/jrzubizarreta/depinf`.

References

[1] Aronow PM, Samii C (2017) Estimating average causal effects under general interference, with application to a social network experiment. Annals of Applied Statistics 11(4):1912–1947 MR3743283

[2] Aronow PM, Samii C, Assenova VA (2015) Cluster-robust variance estimation for dyadic data. Political Analysis 23(4):564–577

[3] Baldi P, Rinott Y (1989) On normal approximations of distributions in terms of dependency graphs. The Annals of Probability pp 1646–1650 MR1048950

[4] Bixby RE, Rothberg E (2007) Progress in computational mixed integer programming—a look back from the other side of the tipping point. Annals of Operations Research 149:37–41 MR2313358

[5] Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. Journal of Human Resources 50(2):317–372

[6] Chen LH, Shao QM, et al (2004) Normal approximation under local dependence. The Annals of Probability 32(3):1985–2028 MR2073183

[7] Conley TG (1999) GMM estimation with cross sectional dependence. Journal of Econometrics 92(1):1–45 MR1707000

[8] Crawford FW, Aronow PM, Zeng L, Li J (2017) Identification of homophily and preferential recruitment in respondent-driven sampling. American Journal of Epidemiology. In Press MR3389977

[9] Drton M, Richardson TS (2008) Binary models for marginal independence. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(2):287–309 MR2424754

[10] Hayashi F (2000) Econometrics. Princeton University Press Princeton, NJ MR1881537

[11] Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260):663–685 MR0053460

[12] Kellerer H, Pferschy U, Pisinger D (2004) Introduction to NP-Completeness of knapsack problems. Springer MR2161720

[13] Kellerer H, Pferschy U, Pisinger D (2004) Knapsack Problems. Springer MR2161720

[14] Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika pp 13–22 MR0836430

[15] Linderoth JT, Lodi A (2010) MILP software. In: Cochran JJ, Cox LA, Keskinocak P, Kharoufeh JP, Smith JC (eds) Wiley Encyclopedia of Operations Research and Management Science, Wiley

[16] Liu L, Hudgens MG (2014) Large sample randomization inference of causal effects in the presence of interference. Journal of the American Statistical Association 109(505):288–301 MR3180564

[17] Nemhauser GL (2013) Integer programming: Global impact. EURO INFORMS July 2013

[18] Neyman J (1923 [1990]) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Reprinted in Statist. Sci. 5:465–472 MR1092986

[19] Ogburn EL, VanderWeele TJ, et al (2017) Vaccines, contagion, and social networks. The Annals of Applied Statistics 11(2):919–948 MR3693552

[20] Overton WS, Stehman SV (1995) The horvitz-thompson theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. The American Statistician 49(3):261–268

[21] Richardson T (2003) Markov properties for acyclic directed mixed graphs. Scandinavian Journal of Statistics 30(1):145–157 MR1963898

[22] Tabord-Meehan M (2015) Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic. arXiv preprint arXiv:1510.07074

[23] Tchetgen EJT, VanderWeele TJ (2012) On causal inference in the presence of interference. Statistical methods in medical research 21(1):55–75 MR2867538

[24] White H (2014) Asymptotic Theory for Econometricians. Academic Press