Design-based causal inference in bipartite experiments

Peng Ding

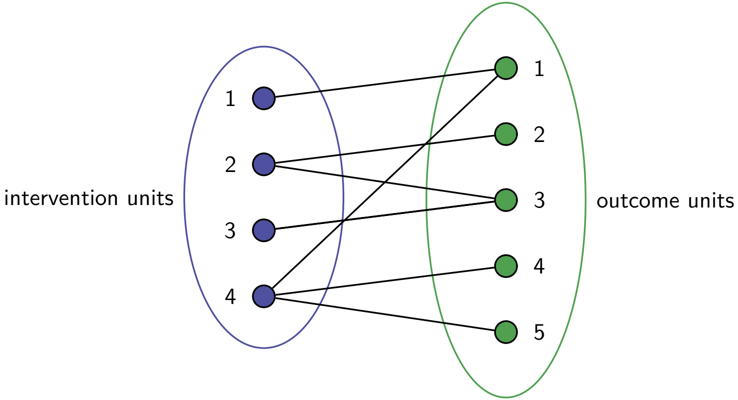
UC Berkeley, Statistics

joint work with Sizhu Lu, Lei Shi, Wenxin Zhang (Berkeley)
and Yue Fang (CUHK Shenzhen)

https://arxiv.org/pdf/2501.09844

Outline

Bipartite graph and bipartite experiment



- Partition the units into “intervention units” and “outcome units”
- They are connected via a bipartite graph
- Randomly assign the treatment over “intervention units”
- Measure the outcomes of “outcome units”

Examples of bipartite experiment: part I

- Trivial example: standard randomized experiment
  - "intervention units" = "outcome units"
- Not so trivial example: cluster-randomized experiment
  - "intervention units" = clusters, e.g., classrooms or villages
  - "outcome units" = individuals, e.g., households or students
  - each cluster can be connected to multiple individuals, whereas each individual can be connected to only one cluster
- Can view bipartite experiment as generalization of cluster experiment
  - each intervention unit can be connected to multiple outcome units
  - each outcome unit can be connected to multiple intervention units

Examples of bipartite experiment: part II

- Install NOx reducing system $\rightarrow$ hospitalization rate
  - intervention units = power plants, installation or not
  - outcome units = neighborhoods, hospitalization rates
  - Zigler and Papadogeorgou (2021)

- Launch a new Facebook Group feature $\rightarrow$ user engagement
  - intervention units = Facebook Groups, new feature or not
  - outcome units = users, engagement
  - Shi et al (2024)

- New pricing mechanism $\rightarrow$ customer satisfaction
  - intervention units = Amazon items, new pricing or not
  - outcome units = customers, satisfaction level
  - Harshaw et al (2024)

Outline

Potential outcomes and total treatment effect

- m intervention units, treatment Z_k, $k = 1, \ldots, m$
- n outcome units, outcome Y_i, $i = 1, \ldots, n$
- Potential outcome $Y_i(z)$, where $z = (z_1, \ldots, z_m) \in \{0, 1\}^m$
- Total treatment effect — a policy-relevant parameter

$$\tau = n^{-1} \sum_{i=1}^{n} \{ Y_i(1_m) - Y_i(0_m) \}$$

  - what if all intervention units receive treatment versus control?
  - reduces to standard ATE $\tau = n^{-1} \sum_{i=1}^{n} \{ Y_i(1) - Y_i(0) \}$ if intervention units = outcome units and SUTVA holds

Simplifying the potential outcomes: bipartite interference

- Each unit has 2^m potential outcomes: too many to make progress
- Need to simplify potential outcomes based on the bipartite graph
- Features of the bipartite graph
  - adjacency matrix $G \in \{0,1\}^{m \times n}$: $G_{ki} = 1$ if outcome unit i is connected to the intervention unit k
  - outcome units connected to intervention unit k: $\mathcal{G}_{k+}$ with $|\mathcal{G}_{k+}| = G_{k+}$
  - intervention units connected to outcome unit i: $\mathcal{G}_{+i}$ with $|\mathcal{G}_{+i}| = G_{+i}$
- Assume "bipartite interference": $Y_i(z) = Y_i(z_{\mathcal{G}_{+i}})$ with subvector $z_{\mathcal{G}_{+i}}$
- Total treatment effect: $\tau = n^{-1} \sum_{i=1}^{n} \{Y_i(1_{G_{+i}}) - Y_i(0_{G_{+i}})\}$

Treatment assignment in bipartite experiment

- We focus on Bernoulli randomization over the intervention units:

$$Z_1, \ldots, Z_m \text{ are IID Bernoulli}(p)$$

- Slightly different from complete randomization: minor difference due to Hajek's coupling argument (Hajek 1960)

- Possible extension to heterogeneous $Z_k \sim \text{Bernoulli}(p_k)$
  - p_k varies across intervention unit k, e.g. stratified randomization
  - more general observational studies

Outline

Hajek estimator

- Recall $\tau = n^{-1} \sum_{i=1}^{n} \{ Y_i(1_{G_{+i}}) - Y_i(0_{G_{+i}}) \}$
- Hajek based on inverse probability weighting: $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$, where

$$
\begin{aligned}
\hat{\mu}_1 &= n^{-1} \sum_{i=1}^{n} \frac{T_i Y_i}{p^{G_{+i}}} \Big/ n^{-1} \sum_{i=1}^{n} \frac{T_i}{p^{G_{+i}}} \\
\hat{\mu}_0 &= n^{-1} \sum_{i=1}^{n} \frac{C_i Y_i}{(1-p)^{G_{+i}}} \Big/ n^{-1} \sum_{i=1}^{n} \frac{C_i}{(1-p)^{G_{+i}}}
\end{aligned}
$$

  - $T_i = \prod_{k \in \mathcal{G}_{+i}} Z_k$, $C_i = \prod_{k \in \mathcal{G}_{+i}} (1 - Z_k)$: "all treatment", "all control"
  - p = probability of Z_k's
  - G_{+i} = number of intervention units connected to outcome unit i

- Horvitz–Thompson estimator: no denominator, poorer performance, simpler asymptotic analysis

Asymptotic properties of the Hajek estimator

- What does asymptotics mean in bipartite experiments?
- m diverges to infinity: a sequence of finite populations
  - design-based inference: randomness driven by Z_k's
  - we need enough randomness from the treatment indicators
  - n grows as m grows: n depends on m
  - bipartite graph grows with dimensions (m, n) grow
- Intuitively, we must have enough units with $T_i = 1$ and $C_i = 1$
  - this depends on the sparsity of the bipartite graph
  - more precise characterization later
- Bounded covariates and potential outcomes: can be relaxed; not the most interesting part of the problem

Consistency of the Hajek estimator

- $\hat{\tau}$ converges to τ if
  - $\max_{1 \leq i \leq n} G_{+i} = O(1)$: Max # intervention units connected to any outcome unit is bounded by a constant (no “super influenced”)
  - $\max_{1 \leq k \leq m} G_{k+}/n = o(1)$: Max # outcome units connected to any intervention unit diverges more slowly than n (no “super influencer”)
- Standard proving strategy based on variance calculation and Markov
- Reasonable assumptions for the power plant example
- Without these assumptions, we might have to move away from τ or impose additional structural assumptions (e.g., Harshaw et al 2024)

Asymptotic normality of the Hajek estimator

$v_n^{-1/2}(\hat{\tau} - \tau) \to \mathcal{N}(0,1)$ if further

- $\sum_{\ell \in [m] \setminus \{k\}} \mathbb{1}\{k, \ell \text{ are connected via an outcome unit}\} \leq B$ for all $k = 1, \ldots, m$, where B is an absolute constant
  - B can diverge slowly but it is a technical issue
  - this is a sparsity condition on the bipartite graph
  - reasonable for the power plant example
  - if not reasonable, we might need alternative estimands and estimators

- $m^{-1/2}(\max_{1 \leq k \leq m} G_{k+}/n)^{-2} v_n \to \infty$ with v_n defined on next page
  - $\max_{1 \leq k \leq m} G_{k+}$ does not diverge to ∞ too fast
  - variance v_n does not converge to 0 too fast
  - more transparent in special cases; see below

- Proof: martingale central limit theorem for polynomials of Z_k's

Asymptotic variance formula of the Hajek estimator

- Centered potential outcomes $\tilde{Y}_i(z) = Y_i(z) - n^{-1}\sum_{i=1}^n Y_i(z)$
- Vectorized potential outcomes $\tilde{Y}(z) = (\tilde{Y}_1(z), \ldots, \tilde{Y}_n(z))^{\mathrm{T}}$
- $n \times n$ matrices related to the bipartite graph:

$$(\Lambda_1)_{i,j} = p^{-|\mathcal{G}_{+ij}|} - 1, \quad (\Lambda_0)_{i,j} = (1-p)^{-|\mathcal{G}_{+ij}|} - 1, \quad (\Lambda_\tau)_{i,j} = \mathbb{1}\{\mathcal{G}_{+ij} \neq \varnothing\}$$

  where $\mathcal{G}_{+ij} = \mathcal{G}_{+i} \cap \mathcal{G}_{+j}$ determines second-order inclusion probabilities

- Asymptotic variance formula

$$v_n = n^{-2}\left\{\tilde{Y}(1_m)^{\mathrm{T}}\Lambda_1\tilde{Y}(1_m) + \tilde{Y}(0_m)^{\mathrm{T}}\Lambda_0\tilde{Y}(0_m) + 2\tilde{Y}(1_m)^{\mathrm{T}}\Lambda_\tau\tilde{Y}(0_m)\right\}$$

Sanity check I: Bernoulli randomization over units

- Intervention units = outcomes units:

$$\mathcal{G}_{+ij} = \mathcal{G}_{+i} \cap \mathcal{G}_{+j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

- Recovers Neyman (1923) and Miratrix et al (2012 Theorem 1):

$$v_n = n^{-2} p(1-p) \sum_{i=1}^{n} \left\{ \frac{\tilde{Y}_i(1)}{p} - \frac{\tilde{Y}_i(0)}{1-p} \right\}^2$$

- Condition $m^{-1/2}(\max_{1 \le k \le m} G_{k+}/n)^{-2} v_n \to \infty$ holds if $n^{3/2} v_n \to \infty$ (easy to hold because $v_n = O(1/n)$ under standard assumptions)

Sanity check II: Bernoulli randomization over clusters

- Outcome units are clustered within intervention units

$$\mathcal{G}_{+ij} = \mathcal{G}_{+i} \cap \mathcal{G}_{+j} = \begin{cases} 1, & \text{if } i, j \text{ belong to the same cluster} \\ 0, & \text{otherwise} \end{cases}$$

- Recovers Su and Ding (2021):

$$v_n = n^{-2} p(1-p) \sum_{k=1}^{m} \left[\sum_{i \in \mathcal{G}_{k+}} \left\{ \frac{\tilde{Y}_i(1)}{p} - \frac{\tilde{Y}_i(0)}{1-p} \right\} \right]^2$$

- Condition $m^{-1/2}(\max_{1 \le k \le m} G_{k+}/n)^{-2} v_n \to \infty$ holds if $m^{3/2}(\text{max cluster size}/\text{ave cluster size})^{-2} v_n \to \infty$ (easy to hold because $v_n = O(1/m)$ under standard assumptions)

Variance estimation: identifiability and upper bound

- Crucial for Wald-type inference based on asymptotic normality
- v_n involves jointly values of the potential outcomes: not identifiable
- Upper bound based on Cauchy–Schwarz:

$$v_{n,\mathrm{UB}} = (v_1^{1/2} + v_0^{1/2})^2$$

  - where $v_1 = n^{-2}\tilde{Y}(1_m)^{\mathrm{T}}\Lambda_1\tilde{Y}(1_m)$ and $v_0 = n^{-2}\tilde{Y}(0_m)^{\mathrm{T}}\Lambda_0\tilde{Y}(0_m)$
  - it follows from $\mathrm{var}(\hat{\mu}_1 - \hat{\mu}_0) \le (\mathrm{var}(\hat{\mu}_1)^{1/2} + \mathrm{var}(\hat{\mu}_0)^{1/2})^2$
  - not same as Neyman (1923) but known in complete randomization
  - only marginals, identifiable

Estimating the upper bound of the variance

- Estimator $\hat{v}_{n,\mathrm{UB}} = (\hat{v}_1^{1/2} + \hat{v}_0^{1/2})^2$ where

$$
\begin{aligned}
\hat{v}_1 &= n^{-2} \sum_{i,j} \frac{T_i T_j (Y_i - \hat{\mu}_1)(Y_j - \hat{\mu}_1)(\Lambda_1)_{i,j}}{p^{|\mathcal{G}_{+i} \cup \mathcal{G}_{+j}|}} \\
\hat{v}_0 &= n^{-2} \sum_{i,j} \frac{C_i C_j (Y_i - \hat{\mu}_0)(Y_j - \hat{\mu}_0)(\Lambda_0)_{i,j}}{(1-p)^{|\mathcal{G}_{+i} \cup \mathcal{G}_{+j}|}}
\end{aligned}
$$

- Variance estimator involves the second-order inclusion probabilities

- Normal-based Wald-type confidence intervals will be conservative

- When is it not conservative?

$$
\tilde{Y}(1_m)^{\mathrm{T}} \Lambda_\tau \tilde{Y}(0_m) = \{\tilde{Y}(1_m)^{\mathrm{T}} \Lambda_1 \tilde{Y}(1_m)\}^{1/2} \{\tilde{Y}(0_m)^{\mathrm{T}} \Lambda_0 \tilde{Y}(0_m)\}^{1/2}
$$

Outline

A class of linearly adjusted estimators

- Covariates X_i for outcome units $i = 1, \ldots, n$
  - can also be features of the bipartite graph, e.g., G_{+i}
  - center it to simplify the presentation: $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i = 0$

- Linearly adjusted estimator $\hat{\tau}(\beta_1, \beta_0) = \hat{\mu}_1(\beta_1, \beta_0) - \hat{\mu}_0(\beta_1, \beta_0)$:

$$\begin{aligned} \hat{\mu}_1(\beta_1, \beta_0) &= n^{-1} \sum_{i=1}^{n} \frac{T_i(Y_i - \beta_1^{\mathrm{T}} X_i)}{p^{G_{+i}}} \Big/ n^{-1} \sum_{i=1}^{n} \frac{T_i}{p^{G_{+i}}} \\ \hat{\mu}_0(\beta_1, \beta_0) &= n^{-1} \sum_{i=1}^{n} \frac{C_i(Y_i - \beta_0^{\mathrm{T}} X_i)}{(1-p)^{G_{+i}}} \Big/ n^{-1} \sum_{i=1}^{n} \frac{C_i}{(1-p)^{G_{+i}}} \end{aligned}$$

  - does not change the probability limit because $\bar{X} = 0$
  - $\hat{\tau}(0, 0)$ reduces to Hajek; can have better choices of β_1, β_0

Oracle estimator

- Easy to derive asymptotic variance formula $v_n(\beta_1, \beta_0)$: $\hat{\tau}(\beta_1, \beta_0)$ is Hajek for potential outcomes $Y_i(1_m) - \beta_1^{\mathrm{T}} X_i$ and $Y_i(0_m) - \beta_0^{\mathrm{T}} X_i$
- Minimize $v_n(\beta_1, \beta_0) =$ oracle estimator: v_n involves joint potential outcomes, not identifiable
- Minimize $n^2\{v_n(\beta_1, \beta_0) - v_n(0, 0)\}$, identifiable and quadratic:

$$\begin{pmatrix}\beta_1\\ \beta_0\end{pmatrix}^{\mathrm{T}} \begin{pmatrix} X^{\mathrm{T}}\Lambda_1 X & X^{\mathrm{T}}\Lambda_\tau X \\ X^{\mathrm{T}}\Lambda_\tau X & X^{\mathrm{T}}\Lambda_0 X\end{pmatrix} \begin{pmatrix}\beta_1\\ \beta_0\end{pmatrix} - 2 \begin{pmatrix} X^{\mathrm{T}}\Lambda_1 \tilde{Y}(1_m) + X^{\mathrm{T}}\Lambda_\tau \tilde{Y}(0_m) \\ X^{\mathrm{T}}\Lambda_0 \tilde{Y}(0_m) + X^{\mathrm{T}}\Lambda_\tau \tilde{Y}(1_m)\end{pmatrix}^{\mathrm{T}} \begin{pmatrix}\beta_1\\ \beta_0\end{pmatrix}$$

- Oracle estimator exists although it may not be unique

A feasible estimator

- An oracle estimator $\hat{\tau}(\tilde{\beta}_1, \tilde{\beta}_0)$ with ("+" denotes the pseudo-inverse):

$$\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_0 \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}\Lambda_1 X & X^{\mathrm{T}}\Lambda_\tau X \\ X^{\mathrm{T}}\Lambda_\tau X & X^{\mathrm{T}}\Lambda_0 X \end{pmatrix}^{+} \begin{pmatrix} X^{\mathrm{T}}\Lambda_1 \tilde{Y}(1_m) + X^{\mathrm{T}}\Lambda_\tau \tilde{Y}(0_m) \\ X^{\mathrm{T}}\Lambda_0 \tilde{Y}(0_m) + X^{\mathrm{T}}\Lambda_\tau \tilde{Y}(1_m) \end{pmatrix}$$

- A feasible estimator $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ with

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}\Lambda_1 X & X^{\mathrm{T}}\Lambda_\tau X \\ X^{\mathrm{T}}\Lambda_\tau X & X^{\mathrm{T}}\Lambda_0 X \end{pmatrix}^{+} \begin{pmatrix} \sum_{i,j} \frac{T_i T_j X_i (Y_j - \hat{\mu}_1)(\Lambda_1)_{i,j}}{p^{|\mathcal{G}_{+i} \cup \mathcal{G}_{+j}|}} + \sum_{i,j} \frac{C_i C_j X_i (Y_j - \hat{\mu}_0)(\Lambda_\tau)_{i,j}}{(1-p)^{|\mathcal{G}_{+i} \cup \mathcal{G}_{+j}|}} \\ \sum_{i,j} \frac{T_i T_j X_i (Y_j - \hat{\mu}_1)(\Lambda_\tau)_{i,j}}{p^{|\mathcal{G}_{+i} \cup \mathcal{G}_{+j}|}} + \sum_{i,j} \frac{C_i C_j X_i (Y_j - \hat{\mu}_0)(\Lambda_0)_{i,j}}{(1-p)^{|\mathcal{G}_{+i} \cup \mathcal{G}_{+j}|}} \end{pmatrix}$$

- $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ and $\hat{\tau}(\tilde{\beta}_1, \tilde{\beta}_0)$ have the same asymptotic distribution as long as $\hat{\beta}_1, \hat{\beta}_0$ converge to $\tilde{\beta}_1, \tilde{\beta}_0$, under some moment conditions

- Analogous estimator for upper bound of variance: $\hat{v}_{n,\mathrm{UB}}(\hat{\beta}_1, \hat{\beta}_0)$

Asymptotic properties of the feasible estimator

- $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ is consistent and asymptotically normal
- $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ has smaller asymptotic variance than $\hat{\tau}$
- Weird issue from design-based inference: true variance and estimate variance differ even in the large-sample limit
  - subtle consequences in complex experiments, e.g. Li and Ding (2020) on "sampling precision" and "estimated precision" in rerandomization
  - $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ reduces the asymptotic variance but the confidence interval may not be narrower because we can only estimate the upper bound of the variance — although this rarely happens; see simulation
  - can minimize $v_n(\hat{\beta}_1, \hat{\beta}_0)$ under the constraint that the estimated variance is also reduced — more complicated to implement

Outline

Power plant example

- Fix the bipartite structure between power plants and nearby areas
- Subsample units to reduce computational burden in simulation
  - $m = 228$ intervention units and $n = 795$ outcome units
  - covariates from the original data of Zigler and Papadogeorgou (2021) and Papadogeorgou et al (2019)
  - true total treatment effect $\tau = -1.266$
- Covariate adjustment improves both true and estimated precision

estimator	point estimator	SE	$\hat{\text{SE}}$	CR	power
$\hat{\tau}$	-1.251	0.136	0.227	98.2%	80.6%
$\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$	-1.202	0.116	0.170	97.7%	86.3%

Outline

Motivation

Causal inference framework for bipartite experiment

Point estimation — Hajek estimator

Covariate adjustment

Simulation based on a real data example

Discussion

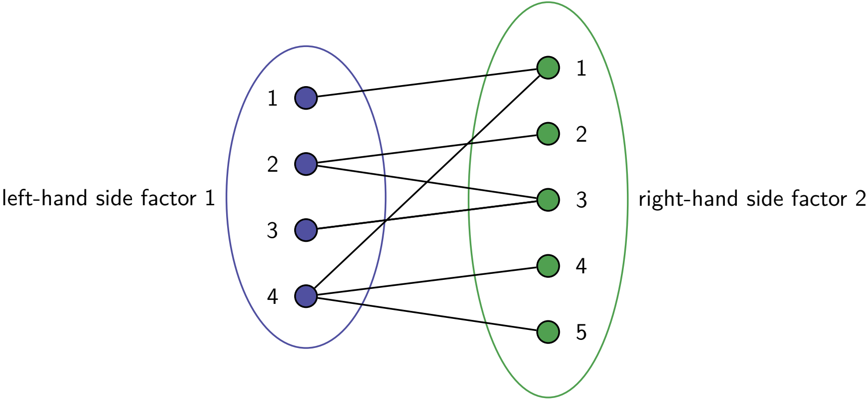
Regression-based versus design-based analyses

- Hajek = WLS of Y_i on (T_i, C_i) with inverse propensity score weights
- We can also include covariates in WLS
- WLS can even provide robust standard errors that work under design-based framework (Gao and Ding 2023 in network experiment)
- Naively using WLS cannot estimate the design-based variance, which involves the second-order inclusion probabilities
- It is an open problem to unify the design-based and regression-based inferences for bipartite experiments

Other estimands and strategies

- Estimating τ requires enough units with $T_i = 1$ and $C_i = 1$
- The weights $p^{-G_{+i}}$ and $(1-p)^{-G_{+i}}$ shrink to 0 exponentially in G_{+i}
- In bipartite graph with large G_{+i}, it is not feasible to estimate τ well without additional assumptions
- May have to make assumptions on "exposure mapping"
- e.g., $Y_i(z) = Y_i(d_i)$ where $d_i = \sum_{k=1}^{m} G_{ki} z_k / \sum_{k=1}^{m} G_{ki}$ is proportion of treated intervention units — requires different estimation strategy

Extension to two-sided randomization



- Motivation from online platforms
  - left-hand side = producers, randomization to factor 1
  - right-hand side = viewers, randomization to factor 2

Causal inference under two-sided randomization

- Extension of the classic factorial design, in particular, split-plot design (Zhao and Ding 2022)
- Richer structure on treatment: factorial
- Richer structure on outcome: outcomes for both sides and for edges
- Related to the "Multiple Randomization Design" proposed by Imbens and Amazon researchers
- Ongoing work

Related papers

- Li, X. and Ding, P. (2020). Rerandomization and regression adjustment. JRSSB
- Su, F. and Ding, P. (2021). Model-assisted analyses of cluster-randomized experiments. JRSSB
- Zhao, A. and Ding, P. (2022). Reconciling design-based and model-based causal inferences for split-plot experiments. AoS
- Gao, M. and Ding. P. (2025+). Causal inference in network experiments: regression-based analysis and design-based properties. JoE
- Lu, S., Shi, L., Fang, Y., Zhang, W. and Ding, P. (2025+) Design-based causal inference in bipartite experiments. ArXiv