Optimized Covariance Design for AB Test on Social Network under Interference

Qianyi Chen, Bo Li, Tsinghua SEM
Lu Deng, Yong Wang, Tencent Inc.

NEURAL INFORMATION PROCESSING SYSTEMS

Better experimental design on social network with scalable optimization

Existing network experimental design research was mostly based on the unbiased Horvitz-Thompson (HT) estimator with substantial data **trimming** to ensure unbiasedness at the price of **high resultant estimation variance**.
The analyses of bias and variance in most existing works are presented in rate form, which is not only **weakly connected to network topology** but also hardly instruct us how to design experiment.
Many of them rely on a mathematical programming or a sequential optimization, which may be **inefficient** and can't scale well for large network.

- We derive a **well-crafted upper bound** for the MSE of the HT estimator that **decouples** the estimation of causal mechanism and experimental design. This enables us to optimize the experimental design by minimizing this bound, in which the covariance matrix of treatment vector acts as decision variables.
- We propose the formulation of **covariance optimization** problem with reparameterization of covariance matrix and constraints that guarantees legitimate sampling subject to optimized covariance matrix, and a **projected gradient descent** algorithm is proposed for solving the optimization problem.
- We conduct **systematic simulation** on two social networks and compare our method with several methods proposed recently under a range of settings, providing credible reference for the effectiveness of our method.

Basic setting of our work

Our basic scheme is cluster-level randomization with $E[z_i] = 1/2$

The interested estimand is global average treatment effect (GATE)

$$\tau := \frac{1}{n}\sum_{i\in[n]}\left(Y_i(\mathbf{1}) - Y_i(\mathbf{0})\right)$$

We consider **standard** HT estimator (without exposure indicator)

$$\hat{\tau} = \frac{1}{n}\sum_{i\in[n]}\left(\left(\frac{z_i}{\mathbb{E}\left[z_i\right]} - \frac{(1-z_i)}{\mathbb{E}\left[1-z_i\right]}\right)Y_i(\boldsymbol{z})\right)$$

The considered linear potential outcome model

$$Y_i(\boldsymbol{z}) = \alpha_i + \beta_i z_i + \gamma\sum_{j\in N_i} z_j$$

The C matrix characterize within/between cluster connections

$$\boldsymbol{C}_{ij} = |\{(u,v) : (u,v)\in\mathcal{E}, u\in S_i, v\in S_j\}|$$

Bias and variance analysis under linear potential outcome model

Bias of HT estimator

$$\mathbb{E}[\hat{\tau}] - \tau = \frac{\gamma}{n}\left(4\,\mathrm{trace}(\boldsymbol{C}\,\mathrm{Cov}[\boldsymbol{t}]) - \sum_{i,j\in[K]}\boldsymbol{C}_{ij}\right)$$

Well-crafted variance bound

$$\mathrm{Var}[\hat{\tau}] \le \frac{8\gamma^2\left(\omega^2+4\right)}{n^2}\,\mathrm{trace}\left(\boldsymbol{d}\boldsymbol{d}^T\left(\mathrm{Cov}[\boldsymbol{t}] + \frac{1}{4}\mathbf{1}\mathbf{1}^T\right)\right)$$

This variance bound is concerned with a constant ω that characterizes the comparability between interference and direct treatment effect. (ω is fixed as 1 in all of our simulations, and can be adjusted according to belief.)

Optimize the covariance matrix through minimizing MSE upper bound

We guarantee the covariance to be valid in the optimization process with parameterization enlightened by following **Grothendieck's identity**

Let x, y be n-dimensional real unit vectors and let $g = (g_1, \ldots, g_n) \sim N(0, I_n)$ be an n-dimensional standard Gaussian vector. Then,

$$\mathbb{E}[\mathrm{sign}(\langle x, g\rangle)\,\mathrm{sign}(\langle y, g\rangle)] = \frac{2}{\pi}\arcsin(\langle x, y\rangle)$$

The final optimization problem is

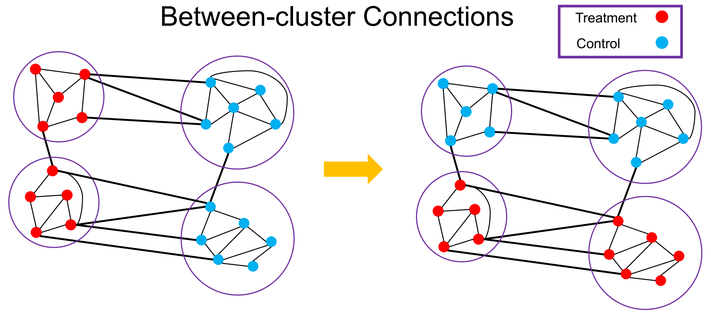
$$\begin{aligned}
\min_{R}\quad & M(R) = B(X(R))^2 + \bar{V}_\omega(X(R)) \\
\text{s.t.}\quad & X(R) = \frac{\arcsin(RR^T)}{2\pi} \\
& (RR^T)_{i,j} \in [-1,1] \quad \forall i\neq j,\ i,j\in[K] \\
& (RR^T)_{i,i} = 1 \qquad\quad \forall i\in[K]
\end{aligned}$$

A simple **projected gradient descent** can be applied, where **row-normalization** is applied on R in every epoch.

After the optimization problem is solved, we can sample from a **multivariate Bernoulli** distribution directly through a reparameterization of the multivariate Gaussian distribution.

$$\boldsymbol{t} = \frac{\mathbf{1} + \mathrm{sgn}\left(R^*\mathcal{N}(\mathbf{0}, I_K)\right)}{2}$$

Treatment Covariance Adaptive to Between-cluster Connections



The information of clusters is summarized as cluster sizes in former work (e.g. independent block randomization), and we extend it greatly to the whole C matrix.

$$\underbrace{\begin{pmatrix}1252\\622\\342\\328\end{pmatrix}}_{\text{Cluster sizes}} \xrightarrow{\text{Extend}} \underbrace{\begin{pmatrix}36610 & 102 & 308 & 18\\ 102 & 5834 & 261 & 334\\ 308 & 261 & 25868 & 473\\ 18 & 334 & 473 & 8312\end{pmatrix}}_{\text{C matrix}} \xrightarrow{\text{Optimize}} \underbrace{\begin{pmatrix}0.2498 & -0.1579 & -0.1038 & -0.2490\\ -0.1579 & 0.2498 & 0.1906 & 0.1584\\ -0.1038 & 0.1906 & 0.2498 & 0.1043\\ -0.2490 & 0.1584 & 0.1043 & 0.2498\end{pmatrix}}_{\text{Optimized covariance matrix}}$$

Some simulation results

Table 1: The average bias, standard deviation and MSE of HT estimator under linear model

gamma		0.5			1.0			2.0	
metric	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
method									
Ber	-0.293	0.521	0.358	-0.588	0.584	0.688	-1.178	0.709	1.893
CR	-0.292	0.409	0.253	-0.586	0.459	0.554	-1.177	0.562	1.702
ReAR	-0.393	0.227	0.206	-0.700	0.251	0.554	-1.317	0.303	1.829
PSR	-0.295	0.235	0.143	-0.587	0.264	0.415	-1.179	0.323	1.496
IBR	-0.298	0.273	0.164	-0.593	0.308	0.447	-1.181	0.380	1.541
IBR-p	-0.294	0.232	**0.141**	-0.596	0.261	0.423	-1.185	0.318	1.507
OCD	-0.198	0.411	0.209	-0.388	0.469	**0.371**	-0.764	0.585	**0.926**

Table 2: The average bias, standard deviation and MSE of HT estimator under multiplicative model

gamma		0.5			1.0			2.0	
metric	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
method									
Ber	-0.365	0.348	0.255	-0.736	0.394	0.698	-1.475	0.493	2.421
CR	-0.368	0.235	0.191	-0.744	0.274	0.629	-1.477	0.336	2.297
ReAR	-0.402	0.178	0.194	-0.809	0.174	0.685	-1.548	0.226	2.450
PSR	-0.366	0.134	0.152	-0.738	0.153	0.569	-1.479	0.192	2.227
IBR	-0.369	0.155	0.161	-0.737	0.178	0.576	-1.484	0.221	2.252
IBR-p	-0.368	0.163	0.163	-0.739	0.185	0.581	-1.482	0.232	2.252
OCD	-0.258	0.040	**0.069**	-0.517	0.050	**0.271**	-1.034	0.054	**1.073**