Communication-efficient distributed estimation of causal effects with high-dimensional data

Xiaohan Wang, Jiayi Tong, Sida Peng, Yong Chen, Yang Ning

Abstract

We propose a communication-efficient algorithm to estimate the average treatment effect (ATE), when the data are distributed across multiple sites and the number of covariates is possibly much larger than the sample size in each site.

Settings

- Data distributed in m sites, each site has n samples. Samples are homogenous. For each sample, we observe a 3– tuple (T_{ij}, Y_{ij}, X_{ij}).
- $Y_{ij}(1)$ and $Y_{ij}(0)$ denote the potential outcomes.

$$Y_{ij} = T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0)$$

Algorithm

Algorithm Distributed high-dimensional propensity score estimation on J_1

On the leading site $j = 1$, compute the following initial propensity score estimator

$$\bar{\theta}_{J_1} = \arg\min_{\theta \in \mathbb{R}^p} Q_1(\theta) + \lambda_{\text{ps,ini}} \|\theta\|_1 ,$$

where $\lambda_{\text{ps,ini}}$ is a regularization parameter and $Q_1(\theta)$ is defined as:

$$Q_j(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\{(1 - T_{ij})X_{ij}^T\theta + T_{ij}/\exp(X_{ij}^T\theta)\right\}.$$

Broadcast $\bar{\theta}_{J_1}$ to all sites in J_1.
for each site $j \in J_1$ **do**
Compute the gradient $\nabla Q_j(\bar{\theta}_{J_1})$ and broadcast to the leading site.
end for
Construct the surrogate loss

$$\tilde{Q}(\theta, \bar{\theta}) = Q_1(\theta) - \left\langle \theta, \left(\nabla Q_1(\bar{\theta}) - \nabla Q_N(\bar{\theta})\right)\right\rangle,$$

where

$$\nabla Q_N(\bar{\theta}_{J_1}) = \frac{1}{|J_1|}\sum_{j \in J_1} \nabla Q_j(\bar{\theta}_{J_1}).$$

Then, compute the penalized surrogate propensity score estimator

$$\tilde{\theta}_{J_1} = \arg\min_{\theta \in \mathbb{R}^p} \tilde{Q}(\theta, \bar{\theta}_{J_1}) + \lambda_{\text{ps}} \|\theta\|_1 , \qquad (1)$$

where λ_{ps} is a regularization parameter.

Algorithm Distributed high-dimensional outcome estimation on J_1

On the leading site $j \in J_1$, say $j = 1$, compute the initial outcome model estimator

$$\bar{\beta}_{J_1} = \arg\min_{\theta \in \mathbb{R}^p} L_1(\beta, \tilde{\theta}_{J_2}) + \lambda_{\text{om,ini}} \|\beta\|_1 ,$$

where $\lambda_{\text{om,ini}}$ is a regularization parameter and

$$L_j(\beta, \tilde{\theta}_{J_2}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_{ij}}{\exp(X_{ij}^T\tilde{\theta}_{J_2})}\left(Y_{ij} - X_{ij}^T\beta\right)^2\right\}.$$

Broadcast $\bar{\beta}_{J_1}$ to all sites in J_1.
for each site $j \in J_1$ **do**
Compute the gradient $\nabla L_j(\bar{\beta}_{J_1}, \tilde{\theta}_{J_2})$, where the derivative is taken with respect to $\bar{\beta}_{J_1}$ and broadcast to the leading site.
end for
Define the surrogate loss for the outcome model as:

$$\tilde{L}(\beta, \bar{\beta}_{J_1}, \tilde{\theta}_{J_2}) = L_1(\beta, \tilde{\theta}_{J_2}) - \left\langle \theta, \left(\nabla L_1(\bar{\beta}_{J_1}, \tilde{\theta}_{J_2}) - \nabla L_N(\bar{\beta}_{J_1}, \tilde{\theta}_{J_2})\right)\right\rangle$$

where $\tilde{L}_N(\bar{\beta}_{J_1}, \tilde{\theta}_{J_2}) = \frac{1}{|J_1|}\sum_{j \in J_1} L_j(\bar{\beta}_{J_1}, \tilde{\theta}_{J_2})$.
Compute the penalized surrogate outcome estimator

$$\tilde{\beta}_{J_1} = \arg\min_{\beta \in \mathbb{R}^p} \tilde{L}(\beta, \bar{\beta}_{J_1}, \tilde{\theta}_{J_2}) + \lambda_{\text{om}} \|\beta\|_1 , \qquad (2)$$

where λ_{om} is a regularization parameter.

Algorithm Distributed high-dimensional ATE estimation with covariate balancing (disthdCBPS)

Compute $\tilde{\theta}_{J_1}$ by Algorithm 1 on J_1 and broadcast $\tilde{\theta}_{J_1}$ to all sites in J_2 and J_3.
Compute $\tilde{\beta}_{J_2}$ by Algorithm 2 on J_2 with $\tilde{\theta}_{J_1}$ and broadcast $\tilde{\beta}_{J_2}$ to all sites in J_3.
for each site j in J_3 **do**
Calculate the AIPW estimator of $\tau_1^* = \mathbb{E}[Y_{ij}(1)]$

$$\tilde{\tau}_{1,j} = \frac{1}{n}\sum_{i=1}^{n}\left\{X_{ij}^T\tilde{\beta}_{J_2} + \frac{T_{ij}}{\pi(X_{ij}^T\tilde{\theta}_{J_1})}\left(Y_{ij} - X_{ij}^T\tilde{\beta}_{J_2}\right)\right\}.$$

end for
Aggregate the local AIPW estimators in J_3

$$\tilde{\tau}_{1,J_3} = \frac{1}{|J_3|}\sum_{j \in J_3} \tilde{\tau}_{1,j}.$$

Similarly, compute $\tilde{\tau}_{1,J_1}$ and $\tilde{\tau}_{1,J_2}$. The final estimator of τ_1^* is $\tilde{\tau}_1 = (\tilde{\tau}_{1,J_1} + \tilde{\tau}_{1,J_2} + \tilde{\tau}_{1,J_3})/3$.

Assumptions

In this section, we present and discuss the assumptions under which our theoretical results are proved.

1. Unconfoundedness: The treatment assignment is unconfounded, i.e., $\{Y_{ij}(0), Y_{ij}(1)\} \perp T_{ij} \mid X_{ij}$.
2. Strict Overlap: There exists a constant $c_0 > 0$ such that $c_0 \leqslant \mathbb{P}(T_{ij} = 1|X_{ij}) \leqslant 1 - c_0$.
3. Design: The minimal and maximal eigenvalues of $\mathbb{E}[X_{ij}X_{ij}^T]$ are contained in a bounded interval that does not contain zero.
4. Model: X_{ij} has mean 0 and a bounded sub-Gaussian norm. Moreover, $\varepsilon_{ij}^* = Y_{ij}(1) - X_{ij}^T\beta^*$ also has a bounded sub-Gaussian norm.
5. Sparsity: Let $s_1 = \|\theta^*\|_0$, and $s_2 = \|\beta^*\|_0$. Assume that
$$\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee mn)}{\sqrt{mn}} + \frac{\sqrt{mn}\sqrt{s_1^3 s_2(s_1 \vee s_2)}\log^3(p \vee n)}{n^2} + \frac{\sqrt{s_2}(s_1 \vee s_2)^{3/2}\sqrt{\log(p \vee mn)\log^4(p \vee n)}}{n} = o(1)$$
as $s_1, s_2, p, m, n \to \infty$.
6. Variance: We assume that there exists some constant $c_1 > 0$ such that $\mathbb{E}(\varepsilon_{ij}^{*2}|X_{ij}) \geqslant c_1$, $\mathbb{E}\left(X_{ij}^T\beta^*\right)^4 = O(s_2^2)$.

Asymptotic distribution under correct model specifications

Theorem Under Assumptions 1-6, we have

$$\sup_{x \in \mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{mn}(\tilde{\tau}_1 - \tau_1^*)}{\sqrt{\hat{V}}} \leqslant x\right) - \Phi(x)\right| \leqslant E_1 + E_2 + E_3, \qquad (3)$$

where

$$E_1 = \frac{M}{n^s}, \quad E_2 = C_L\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee mn)}{\sqrt{mn}},$$

$$E_3 = C_L\Big(\frac{\sqrt{mn}\sqrt{s_1^3 s_2(s_1 \vee s_2)}\log^3(p \vee n)}{n^2} + \frac{\sqrt{s_2}(s_1 \vee s_2)^{3/2}\sqrt{\log(p \vee mn)\log^4(d \vee n)}}{n}\Big).$$

Here C_L is a sufficiently large constant, M is a constant depending on C_L and $\Phi(x)$ is the c.d.f of $N(0, 1)$.

Asymptotic distribution under misspecified model

In this subsection, we study the robustness of the proposed estimator when not both model are correctly specified.

We define the estimand obtained via Algorithm 1 as:

$$\theta^o = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[Q_j(\theta)].$$

Lemma Under Assumptions 1-6, with θ^* replaced by θ^o, the proposed estimator satisfies

$$|\tilde{\tau}_1 - \hat{\tau}_{1,\text{ps}}^o| \leqslant C_L\left(\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee mn)}{mn} + \frac{\sqrt{s_1^3 s_2(s_1 \vee s_2)}\log^3(p \vee n)}{n^2} + \frac{\sqrt{s_2}(s_1 \vee s_2)^{3/2}\sqrt{\log(p \vee mn)\log^4(p \vee n)}}{n\sqrt{mn}}\right)$$

with probability at least $1 - \frac{M}{n^s}$, where C_L is a sufficiently large constant and M is another constant depending on C_L.
When the outcome model is misspecified, we define the estimand obtained via Algorithm 2 as:

$$\beta^o = \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}[L_j(\beta, \theta^*)],$$

Lemma Under Assumptions 1-6, with β^* replaced by β^o, the proposed estimator satisfies

$$|\tilde{\tau}_1 - \hat{\tau}_{1,\text{om}}^o| \leqslant C_L\left(\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee mn)}{mn} + \frac{\sqrt{s_1^3 s_2(s_1 \vee s_2)}\log^3(p \vee n)}{n^2} + \frac{\sqrt{s_2}(s_1 \vee s_2)^{3/2}\sqrt{\log(p \vee mn)\log^4(p \vee n)}}{n\sqrt{mn}}\right)$$

with probability at least $1 - \frac{M}{n^s}$, where C_L is a sufficiently large constant and M is another constant depending on C_L.

Simulation

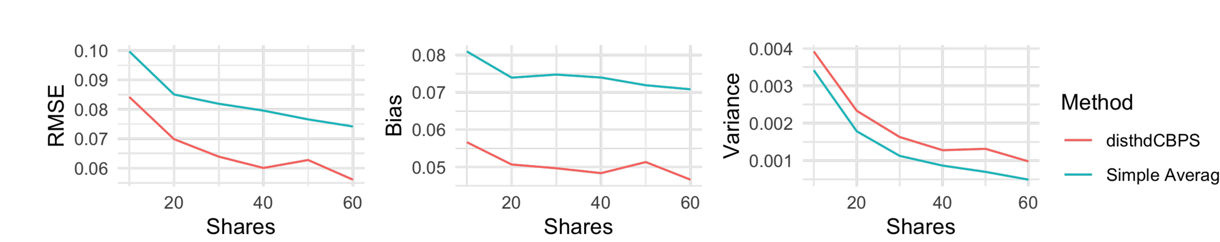

Figure: (A) Correctly specified models

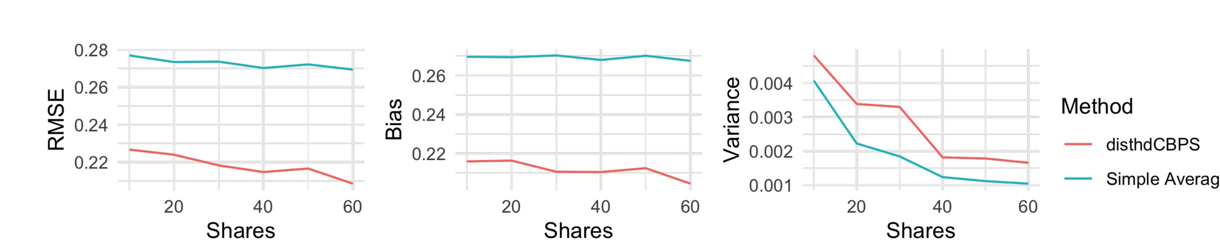

Figure: (B) Misspecified potential outcome model

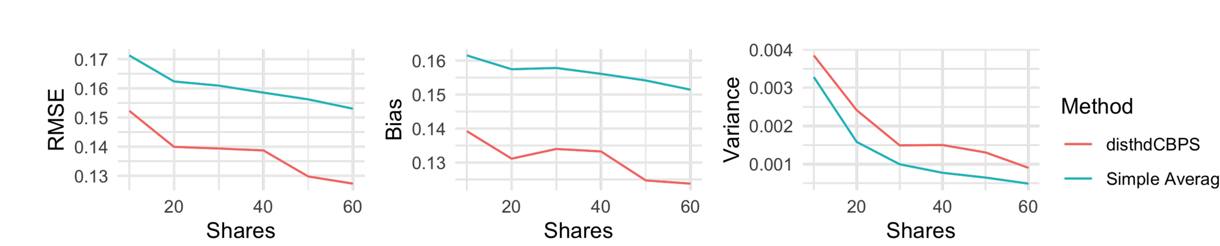

Figure: (C) Misspecified propensity score model