Distributed mediation analysis with communication-efficiency

Shaomin Li
School of Mathematics and Statistics
Beijing Jiaotong University
Beijing, 100044
smli1@bjtu.edu.cn

Abstract

We study the mediation analysis under the distributed framework, where data are stored and processed across different worker machines due to storage limitations or privacy concerns. Building upon the classic Sobel's test and MaxP test, we introduce the distributed Sobel's test and distributed MaxP test, respectively. These tests are both communication-efficient and easy to implement. Theoretical analysis and numerical experiments show that, compared to the global test obtained by pooling all data together, the proposed tests achieve nearly identical power, independent of the number of machines. Furthermore, based on these two distributed test statistics, many enhanced mediation tests derived from the Sobel's or MaxP tests can be easily adapted to the distributed system. We apply our method to an educational study, testing whether the effect of high school mathematics on college-level Probability and Mathematical Statistics courses is mediated by Calculus. Our method successfully detects the mediation effect, which would not be identifiable using data from only the first or second class, highlighting the advantage of our approach.

1 Introduction

Mediation analysis is regarded as a prevalent tool to dissect a mediation relationship between exposures and outcomes, and it has been widely applied in different fields, such as epidemiology [29], economics [1], psychology [18], education [11], and many others. Baron and Kenny[3] provided the basis for the advance of mediation analysis. They proposed a conventional regression-based approach, commonly referred to as the causal steps method, to examine the logical relationships among exposure, mediator, and outcome variables linking in a causal chain. While the causal steps method established necessary conditions for causal inference, it did not provide a joint test of the indirect effect of exposure on the outcome through a mediator. To test the mediation effect, the classic tests are Sobel's test [31] and the MaxP test [24]. However, these tests suffer from low statistical power, as it overlooks the impact of the composite null structure in mediation analysis [25]. To address the aforementioned issues, Nuijten et al.[26] proposed an approach based on Bayesian models. Zhang[37] developed two data-adaptive tests that outperform both Sobel's test and MaxP test. These studies were limited to a small number of mediators. For high-dimensional mediators, one usually first estimates the proportions of sub-null hypotheses and generates weighted p-values [7, 8, 10, 17, 23].

In many applications, data is collected independently by numerous agents and organizations. Due to limitations in resources and concerns about privacy, only summary statistics can be shared between sites, while raw data remains inaccessible. To address this, distributed learning algorithms have gained attention in multi-source studies, enabling collaborative analyses without compromising the

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

privacy of individual-level data. There has been substantial research on distributed estimation, which can be categorized into two main approaches: the one-shot average method [6, 13, 14, 32, 38] and the multi-round iteration method [12, 19, 20, 35, 36]. However, research on distributed statistical testing is still in its early stages. For example, Battey et al.[4] explored distributed versions of Wald and Rao's score tests for sparse high-dimensional schemes. Zhao et al.[39] proposed a distributed specification test for massive datasets and analyzed the impact of divide-and-conquer strategies on nonparametric testing. Du et al.[9] addressed one-sample mean testing within distributed frameworks. Li et al.[22] introduced a distributed conditional independence test aimed at discovering causal relationships. Cai et al.[5] studied distributed nonparametric goodness-of-fit testing under differential privacy constraints in a white-noise-with-drift model. In the context of mediation analysis research, meta-analysis is often employed to combine results from multiple studies. This is typically done by either computing a weighted average of p-values or test statistics [30], or by averaging the estimated mediation effects using a weighted approach [21]. However, these methods often lack a detailed and rigorous statistical framework to support their use.

The distributed tests discussed above all rely on a simple weighted average approach. One limitation of this method is that, when the total sample size N is fixed, the power of the test declines sharply as the number of data blocks K increases, as shown in the simulation. In this paper, we develop a distributed framework for testing mediation effects, which overcomes this limitation. Specifically, our main contributions can be summarized as follows:

- We propose communication-efficient distributed versions of the classic Sobel's test statistic T_{Sobel} and the MaxP test statistic T_{MaxP}, denoted as T^{Dis}_{Sobel} and T^{Dis}_{MaxP}, respectively. After presenting the theoretical properties of these statistics, we provide an easy-to-implement algorithm. Both theoretical analysis and numerical experiments demonstrate that the proposed tests achieve nearly identical power to the global test, which is obtained by pooling all data togethera process difficult to implement in practice due to data-sharing limitations.
- Our method lays the foundation for the further development of distributed mediation analysis. Although the classic Sobel's and MaxP tests are conservative and underpowered, primarily because they do not account for the composite null nature of the mediation effect, many enhanced mediation tests are build upon Sobel's or MaxP tests. Consequently, the distributed versions of these enhanced tests can be derived from our distributed Sobel's or MaxP tests.
- We apply our method to an educational study, demonstrating that high school mathematics' effect on college-level Probability and Mathematical Statistics is mediated by Calculus. The data are sourced from three classes, with only summary information of local data available from each class due to student privacy concerns. Our distributed test successfully detects the mediation effect, which would be undetectable using local tests from just the first or second class.

2 Preliminaries: Mediation Analysis

Let A represent the exposure variable, Y the continuous outcome, M the continuous mediator, and $X \in \mathbb{R}^p$ the vector of additional covariates used to adjust for potential confounding. In their seminal work, Baron and Kenny[3] proposed the following linear structural equation models for the outcome and mediator:

$$Y = \beta_0 + \beta_A A + \beta M + \beta_X^T \mathbf{X} + \epsilon_Y, \tag{1}$$

$$M = \gamma_0 + \gamma A + \gamma_X^T \mathbf{X} + \epsilon_M, \tag{2}$$

where ϵ_Y and ϵ_M represent error terms with zero means and constant variances, assumed to be uncorrelated with one another.

The mediation effect is the Natural Indirect Effect (NIE), a concept introduced by [28] and [27], which quantifies the effect of the exposure on the outcome that is mediated through the mediator. For continuous mediator and outcome variables, the mediation effect is defined as the product $\beta\gamma$, where β and γ are the coefficients from the respective equations. Graphically, the mediation effect corresponds to the NIE conveyed through the chain $A \to M \to Y$, as illustrated in the directed acyclic graph (DAG) in Figure 1. The modern causal inference framework relies on the following standard identification assumptions to estimate the mediation effect [34, 33].

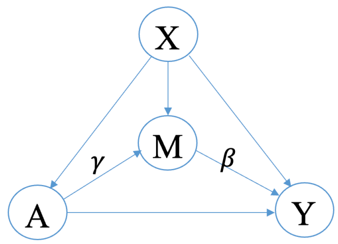

Figure 1: The causal graph with treatment A, outcome Y, mediator M, and confounder X. γ is the effect of A on M, and β is the effect of M on Y.

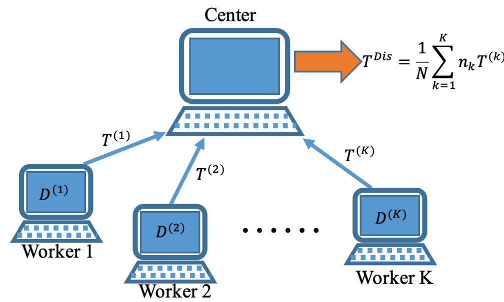

Figure 2: A distributed framework with K worker machines connected to a central processor. The k-th machine computes a local statistic $T^{(k)}$ using its local data $D^{(k)}$, and then transmits it to the central processor. The central processor subsequently computes the distributed statistic T^{Dis} based on the K local statistics.

Assumptions

A1. There are no unmeasured exposure-outcome confounders given X;

A2. There are no unmeasured mediator-outcome confounders given $(X; A)$;

A3. There are no unmeasured exposure-mediator confounders given X;

A4. There is no effect of exposure that confounds the mediator-outcome relationship;

A5. There is no exposure and mediator interaction on the outcome.

To test the existence of a mediation effect, we can formalize the problem as a hypothesis testing problem:

$$H_0 : \beta\gamma = 0 \quad \text{versus} \quad H_1 : \beta\gamma \neq 0. \tag{3}$$

The null hypothesis is composite and can be decomposed into three disjoint cases:

$$\text{Case 1. } \beta \neq 0, \gamma = 0; \text{Case 2. } \beta = 0, \gamma \neq 0; \text{Case 3. } \beta = 0, \gamma = 0.$$

The classic tests, Sobel's test and MaxP test, are constructed from the t-test statistics, $T_\beta = \hat{\beta}/\hat{\sigma}_\beta$, and $T_\gamma = \hat{\gamma}/\hat{\sigma}_\gamma$, where $\hat{\beta}$ and $\hat{\gamma}$ are least squared estimators, and $\hat{\sigma}_\beta$ and $\hat{\sigma}_\gamma$ are the estimated standard errors for $\hat{\beta}$ and $\hat{\gamma}$, respectively. Under mild regularity conditions, we have $T_\beta \xrightarrow{d} N(0,1)$ if $\beta = 0$, and $T_\gamma \xrightarrow{d} N(0,1)$ if $\gamma = 0$. Building on these two t-statistics, a Wald-type Sobel's test statistic [31] is constructed as

$$T_{Sobel} = \frac{\hat{\beta}\hat{\gamma}}{\sqrt{\hat{\gamma}^2\hat{\sigma}_\beta^2 + \hat{\beta}^2\hat{\sigma}_\gamma^2}} = \frac{T_\beta T_\gamma}{\sqrt{T_\beta^2 + T_\gamma^2}}.$$

Alternatively, the MaxP test statistic [24] can be defined as:

$$T_{MaxP} = \max\{2 - 2\Phi(|T_\beta|), 2 - 2\Phi(|T_\gamma|)\},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

According to [23], in cases 1 and 2, T_{Sobel} asymptotically follows $N(0,1)$, while in case 3, T_{Sobel} asymptotically follows $N(0,1/4)$. In cases 1 and 2, T_{MaxP} asymptotically follows a uniform distribution on $[0,1]$, and in case 3, T_{MaxP} asymptotically follows a $Beta(2,1)$ distribution. In practice, these two tests do not account for case 3, making them overly conservative and underpowered. Numerous enhanced tests have been proposed based on these methods, such as those in [17, 7, 23, 8, 10], though this lies beyond the scope of our discussion.

3 Distributed Mediation Analysis

In the distributed framework, the total N data are splitted and stored in K worker machines (locations) connected to a central processor as shown in Figure 2, where $D^{(k)} = (Y^{(k)}, A^{(k)}, M^{(k)}, X^{(k)}) = \{(Y_{ki}, A_{ki}, M_{ki}, X_{ki})\}_{i=1}^{n_k}$ denotes the data in the k-th machine, $k = 1, \cdots, K$, and $N = \sum_{k=1}^{K} n_k$. Assume that there exist positive constants c_1 and c_2 such that $c_1 \leq \inf_{k_1,k_2} n_{k_1}/n_{k_2} \leq \inf_{k_1,k_2} n_{k_1}/n_{k_2} \leq c_2$, and K can be either finite or diverging to infinity as long as $K/\min_k n_k \to 0$ as $N \to \infty$. Each worker machine extracts statistics based on the local data and transmits it to the central machine. The final statistic is obtained by suitably aggregating the local statistics.

Note that both the Sobel's and MaxP tests are based on T_β and T_γ. To derive the distributed versions of these tests, we first compute the distributed versions of T_β and T_γ. Specifically, we calculate the local values $T_\beta^{(k)}$ and $T_\gamma^{(k)}$ on each machine, transmit them to the central machine, and then obtain the distributed statistics by weighted averaging:

$$\overline{T}_\beta = \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}} T_\beta^{(k)}, \quad \overline{T}_\gamma = \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}} T_\gamma^{(k)}. \tag{4}$$

The properties of the distributed $\overline{T}_\beta$ and $\overline{T}_\gamma$ are outlined below, based on Lemma 25 of [16].

Proposition 3.1. *Under assumptions A1-A5 and with $K = o(\sqrt{N})$, as $N \to \infty$,*

$$\overline{T}_\beta - \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}} \frac{\beta}{\hat{\sigma}_\beta^{(k)}} \xrightarrow{d} N(0,1), \quad \overline{T}_\gamma - \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}} \frac{\gamma}{\hat{\sigma}_\gamma^{(k)}} \xrightarrow{d} N(0,1).$$

Based on $\overline{T}_\beta$ and $\overline{T}_\gamma$, we can construct Sobel's test and MaxP test in the distributed framework, as shown below:

$$T_{Sobel}^{Dis} = \frac{\overline{T}_\beta \overline{T}_\gamma}{\sqrt{(\overline{T}_\beta)^2 + (\overline{T}_\gamma)^2}}, \tag{5}$$

and

$$T_{MaxP}^{Dis} = \max\{2 - 2\Phi(|\overline{T}_\beta|), 2 - 2\Phi(|\overline{T}_\gamma|)\}. \tag{6}$$

The asymptotic properties of these distributed statistics are analogous to those of the global statistics, as demonstrated in the following theorem.

Theorem 3.1. *Under Assumptions A1-A5, the null hypothesis, and $K = o(\sqrt{N})$, as $N \to \infty$, we have the following results.*

In Cases 1 and 2,

$$T_{Sobel}^{Dis} \xrightarrow{d} N(0,1), \quad \textit{and} \quad T_{MaxP}^{Dis} \xrightarrow{d} Uniform(0,1).$$

In Case 3,

$$T_{Sobel}^{Dis} \xrightarrow{d} N(0,1/4), \quad \textit{and} \quad T_{MaxP}^{Dis} \xrightarrow{d} Beta(2,1).$$

The proof of Theorem 3.1 follows a similar approach to the proofs of Results 1(b) and 2(b) in [23], and is therefore omitted here. According the theorem, given the significance level α, we reject H_0 if $|T_{Sobel}^{Dis}| > Z_{1-\alpha/2}$ or $T_{MaxP}^{Dis} < \alpha$, where $Z_{1-\alpha/2}$ denotes the $1 - \alpha/2$ percentile of the standard normal distribution. The distributed test procedure is outlined in Algorithm 1. From the algorithm, we can observe that each machine only needs to transmit two scalars to the central machine. This not only minimizes transmission costs but also significantly protects data privacy. The following theorem demonstrates that the asymptotic power of the distributed test, based on the local summary statistics, is equivalent to the asymptotic power of the global test that aggregates all the data.

To investigate the power of the two distributed statistics, we first introduce the necessary notations. Let $\mu_\beta \equiv E[\overline{T}_\beta] = \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}} E[T_\beta^{(k)}]$ and $\mu_\gamma \equiv E[\overline{T}_\gamma] = \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}} E[T_\gamma^{(k)}]$. According to [23], $E[T_\beta^{(k)}] \approx \sqrt{n_k}\beta \frac{\sigma_M}{\sigma_Y} \sqrt{1 - R_{M|A,X}^2}$, $E[T_\gamma^{(k)}] \approx \sqrt{n_k}\gamma \frac{\sigma_A}{\sigma_M} \sqrt{1 - R_{A|X}^2}$, where

Algorithm 1: The Distributed Test of Mediation Effects.

Step 1. For $k = 1, ..., K$, compute the local statistics $T_\beta^{(k)}$ and $T_\gamma^{(k)}$, then transmit them to the central machine.

Step 2. In the central machine, compute the distributed statistics $\bar{T}_\beta$ and $\bar{T}_\gamma$ using (4), then compute the distributed Sobel statistic T_{Sobel}^{Dis} and MaxP statistic T_{MaxP}^{Dis} using (5) and (6), respectively.

Step 3. Given the significance level α, if $|T_{Sobel}^{Dis}| > Z_{1-\alpha/2}$ or $T_{MaxP}^{Dis} < \alpha$, reject H_0.

$\sigma_A = \sqrt{Var(A)}$, $R^2_{A|X}$ and $R^2_{M|A,X}$ are the coefficients of determination by regressing A on X, and regressing M on $(A, X^T)^T$, respectively. Thus, $\mu_\beta \approx \sqrt{N}\beta\frac{\sigma_M}{\sigma_Y}\sqrt{1-R^2_{M|A,X}}$ $\mu_\gamma \approx \sqrt{N}\gamma\frac{\sigma_A}{\sigma_M}\sqrt{1-R^2_{A|X}}$.

Theorem 3.2. *Under the same assumptions as in Theorem 3.1, the powers of T_{Sobel}^{Dis} and T_{MaxP}^{Dis} can be calculated analytically as*

$$\int\int_{\{\frac{1}{x^2}+\frac{1}{y^2}\leq\frac{1}{Z_{1-\alpha/2}}\}} \frac{1}{2\pi}e^{-\frac{(x-\mu_\gamma)^2}{2}-\frac{(y-\mu_\beta)^2}{2}}dxdy$$

and

$$\{\Phi(\mu_\beta - Z_{1-\alpha/2}) + \Phi(-\mu_\beta - Z_{1-\alpha/2})\}\{\Phi(\mu_\gamma - Z_{1-\alpha/2}) + \Phi(-\mu_\gamma - Z_{1-\alpha/2})\},$$

respectively.

The proof of Theorem 3.2 is analogous to the proofs of Results 1(c) and 2(c) in [23], and is therefore not repeated here. Furthermore, the analytical powers of the proposed distributed tests are identical to the powers of the global tests in [23], provided that $K = o(\sqrt{N})$. This property is also corroborated by subsequent simulations. Notably, for a given total sample size N, the power of our tests remains unaffected by the number of machines K, as long as $K = o(\sqrt{N})$. This behavior can be attributed to the fact that $T_\beta^{(k)}$ and $T_\gamma^{(k)}$ (for $k = 1, \dots, K$) exhibit an approximately linear form. In contrast, many test statistics for other hypothesis testing problems (e.g., goodness-of-fit tests) are degenerate U-statistics, which leads to a reduction in power for their distributed versions as K increases, as discussed in [39, 2, 6, 15, 5].

Remark 3.1. *The condition $K = o(\sqrt{N})$ is commonly used in distributed statistical inference and federated learning. This is equivalent to $K/n_k \to 0$ as $n_k \to \infty$, meaning the number of sites K is much smaller than the sample size within each site. This setup is frequently encountered in practice, particularly in meta-analysis, where researchers aggregate conclusions from dozens of studies, each with hundreds or even thousands of data points. When $K = o(\sqrt{N})$ is not satisfied, test power may decrease, and Type I error could become uncontrollable. In practice, determining this boundary is quite challenging. Our simulations reveal that when $N = 2^{11}$ and $K = 32$ (less than $\sqrt{N} \approx 45$), our method yields results similar to the global method. However, when $K = 64$, the difference between our method and the global method increases. This suggests that our method is somewhat tolerant of larger values of K. Therefore, we recommend ensuring that $K < \sqrt{N}/2$ in practice, which should be sufficiently effective.*

4 Simulation Studies

In this section, we conduct extensive simulation studies to evaluate the performance of the proposed distributed Sobel test and MaxP test. We generated the p-dimensional exposure variable $A \sim Bernouli(0.5)$, the covariate $X \sim N(0, \Sigma_X)$ with $p = 3$ and $\Sigma_X = (|0.5|^{j-l})_{p\times p}$. The mediator M and the outcome Y were simulated as follows

$$Y = A + \beta M + \beta_X^T X + \epsilon_Y, \quad \epsilon_Y \sim N(0, 2), \tag{7}$$

$$M = \gamma A + \gamma_X^T X + \epsilon_M, \quad \epsilon_M \sim N(0, 1), \tag{8}$$

where $\beta_X = (1, -0.5, 1)^T$, $\gamma_X = (0.5, 1, -1)^T$. Under the null hypothesis, we consider three scenarios: (1) $(\beta, \gamma) = (0.2, 0)$; (2) $(\beta, \gamma) = (0, 0.2)$; (3) $(\beta, \gamma) = (0, 0)$. Under the alternatives, we set $(\beta, \gamma) = (0.2, 0.05)$, $(0.1, 0.1)$, and $(0.05, 0.2)$. We compared the proposed distributed tests with the global tests (denoted as T_{Sobel}^G and T_{MaxP}^G), which use all the data pooled together, the local tests from the first machine (denoted as $T_{Sobel}^{(1)}$ and $T_{MaxP}^{(1)}$), and the simple average tests that average the K local statistics (denoted as $T_{Sobel}^{ave} = \sum_{k=1}^K \sqrt{\frac{n_k}{N}} T_{Sobel}^{(k)}$ and $T_{MaxP}^{ave} = \sum_{k=1}^K \frac{n_k}{N} T_{MaxP}^{(k)}$). The significance level $\alpha = 0.05$. We conducted the tests in R Studio on a MacBook Pro with an M2 CPU, and each experiment was repeated 2,000 times to calculate the empirical sizes and powers of the tests.

4.1 Balanced Data

In this setting, we set the sample size in each machine are the same, that is, $n = N/K$. We fix the total sample size $N = 2^{11}$, and set the number of machines $K = 1, 2, 4, 8, 16, 32, 64$. The empirical sizes are presented in Figure 3. We can see that all the tests can control the type I error, while the empirical sizes of the six tests are close to 0 when $\beta = \gamma = 0$. This is because in this case, the six tests are too conservative. As the number of machine K increases, the empirical sizes of $T_{Sobel}^{(1)}$ and $T_{MaxP}^{(1)}$ tends to decrease, while our proposed distributed test are not influenced by K. This is because for fixed N, as K increases, the sample size in the first machine n will decreases, thus, the local tests may distance away from their asymptotic null distribution. As contrast, the asymptotic null distributions of our distributed tests is not influenced by K as long as $K = o(n)$.

Figure 4 shows the empirical powers. As expected, the powers of the local tests ($T_{Sobel}^{(1)}$ and $T_{MaxP}^{(1)}$) and average tests (T_{Sobel}^{ave} and T_{MaxP}^{ave}) decrease as K increases, while the proposed distributed test is less affected by K and performs similarly to the global tests. It is also evident that the performance of T_{MaxP}^{ave} is even worse than that of $T_{MaxP}^{(1)}$. In the leftmost four plots of Figure 4, we can observe that when K is large (e.g., $K = 64$), the power of our method is slightly lower than that of the global tests. This may be due to the fact that when N is fixed and K is large, the condition $K = o(\sqrt{N})$ might not be satisfied, causing the distribution of our test statistic to deviate somewhat from the asymptotic distribution.

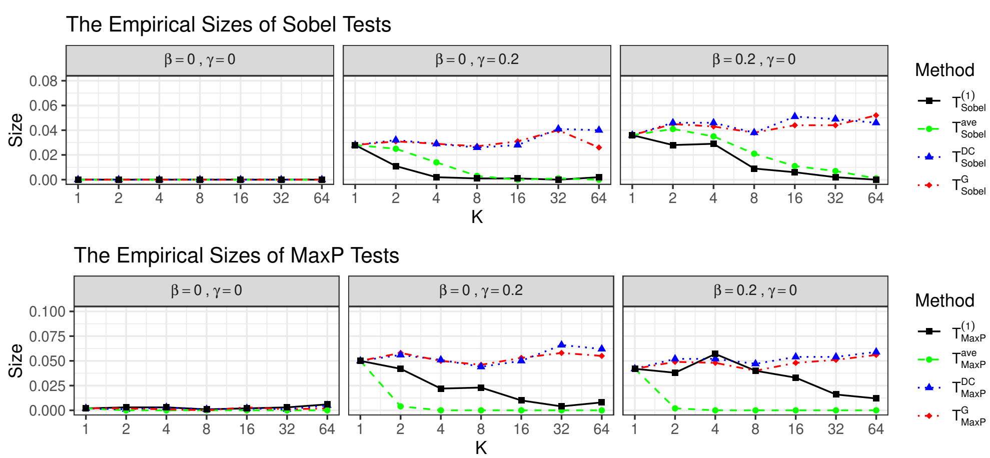

Figure 3: The empirical sizes of the eight tests under three null hypothesis and seven choices of K, with the total sample size N fixed at 2^{11}.

4.2 Unbalanced Data

In this setting, we first generate the local sample sizes n_k randomly from 50 to 150 for $k = 1, 2, \ldots, 32$. Then, we generate a total of $\sum_{k=1}^K n_k$ data points. Next, we calculate the local tests ($T_{Sobel}^{(1)}$ and $T_{MaxP}^{(1)}$) based on n_1 samples, the global tests ($T_{Sobel}^{(G)}$ and $T_{MaxP}^{(G)}$) based on $\sum_{k=1}^K n_k$ samples, the average tests (T_{Sobel}^{ave} and T_{MaxP}^{ave}) based on the first K local test statistics, and the dis-

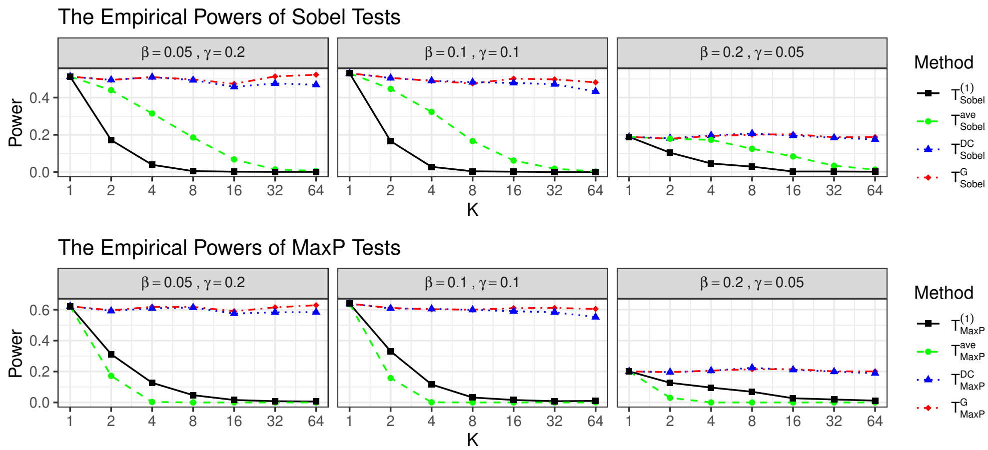

Figure 4: The empirical powers of the eight tests under three alternative hypothesis and seven choices of K, with the total sample size N fixed at 2^{11}.

tributed tests (T_{Sobel}^{Dis} and T_{MaxP}^{Dis}) based on the first K summary local statistics, for $K = 1, 2, \ldots, 32$. The empirical sizes and powers are presented in Figures 5 and 6, respectively.

We observe that the sizes of the local tests are close to zero, as they only use n_1 samples. As K increases, the total sample size N also increases, and the empirical sizes of the global tests and distributed tests approach the nominal significance levels in scenarios 1 and 2. However, in scenario 3 ($\beta = 0, \gamma = 0$), increasing the sample size does not affect the empirical sizes of the Sobel and MaxP tests. We can also see that the size and power of T_{MaxP}^{ave} are close to zero, indicating that averaging the local p-values does not improve the power of the test.

As expected, the powers of the local tests in the first machine are close to zero, and the powers of the global and distributed tests increase as K increases. Notably, the power of the distributed tests is nearly identical to that of the global tests, despite using only the local summary statistics. This highlights the advantage of the proposed distributed tests: they do not require pooling all the data together, yet they achieve the same performance as the global tests, which pool all the data.

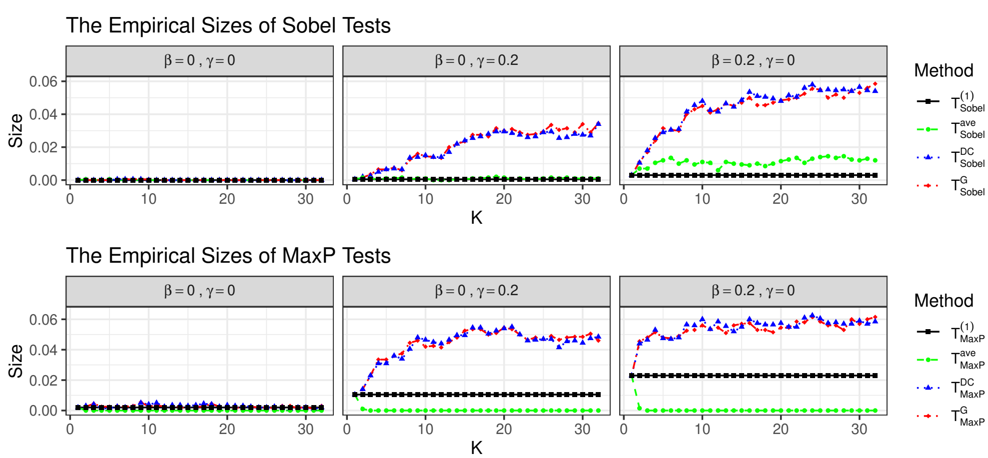

Figure 5: The empirical sizes of the six tests under three null hypothesis. The number of machines K increases from 1 to 32 , and the total sample size N increases with K.

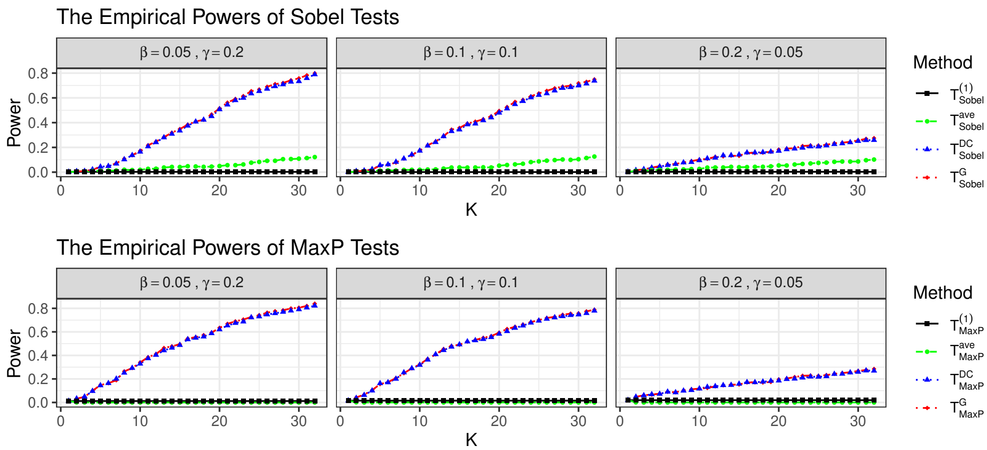

Figure 6: The empirical powers of the six tests under three alternative hypothesis. The number of machines K increases from 1 to 32, and the total sample size N increases with K.

4.3 Heterogeneous case

While the proposed test primarily focuses on the homogeneous case, we also investigate its behavior in the heterogeneous setting. In this scenario, we consider a situation where the distribution of X varies across different centers, and the parameters γ and β differ for each machine, while maintaining the consistency of their signs across all centers. Specifically, we define $X^{(k)} \sim N(\mu_k 1_p, \Sigma_X)$, where $\mu_1, ..., \mu_K \sim_{i.i.d} U(-3, 3)$, and $1_p = (1, ..., 1)^T$ is a p-dimensional vector. The true parameter values for machine k are given by:

$$\beta_k = 0.05 + e^{(k)}_{\beta,1}, 0.1 + e^{(k)}_{\beta,2}, 0.2 + e^{(k)}_{\beta,3};$$

$$\gamma_k = 0.05 + e^{(k)}_{\gamma,1}, 0.1 + e^{(k)}_{\gamma,2}, 0.2 + e^{(k)}_{\gamma,3},$$

where $e^{(k)}_{\beta,1}, e^{(k)}_{\gamma,1} \sim_{i.i.d} U(-0.02, 0.02)$, $e^{(k)}_{\beta,2}, e^{(k)}_{\gamma,2} \sim_{i.i.d} U(-0.04, 0.04)$, and $e^{(k)}_{\beta,3}, e^{(k)}_{\gamma,3} \sim_{i.i.d} U(-0.08, 0.08)$.

The results from this simulation are shown in Figure 7, and are consistent with those obtained in the homogeneous case (Figure 4). Specifically, our method performs similarly to the global test, with noticeable differences only when K is particularly large ($K = 64$). This indicates that our method is well-suited to handle heterogeneity across centers.

5 Real Data Analysis

We examined the impact of Gaokao mathematics (A) performance on college-level "Probability and Mathematical Statistics" course grades (Y), with college "Calculus" course grades serving as a mediator (M). The *Gaokao* is the National Higher Education Entrance DExamination in China, and the results of the Gaokao are often seen as a key determinant of a student's future academic and career opportunities in China. In this study, we sought to explore whether there is a mediating effect in which Gaokao mathematics performance influences calculus grades, which in turn affect the grades in the probability and statistics course. This research holds important implications for both the learning and teaching of Probability and Mathematical Statistics. The confounders (X) include learning interest and classroom efficiency.

We distributed surveys to three classes at a university. To protect student privacy, the surveys were organized and the statistical calculations were conducted by the instructors of each class. The results are as follows.

Class 1: $n_1 = 34, T^{(1)}_{\beta} = 4.654908, T^{(1)}_{\gamma} = 1.723737;$

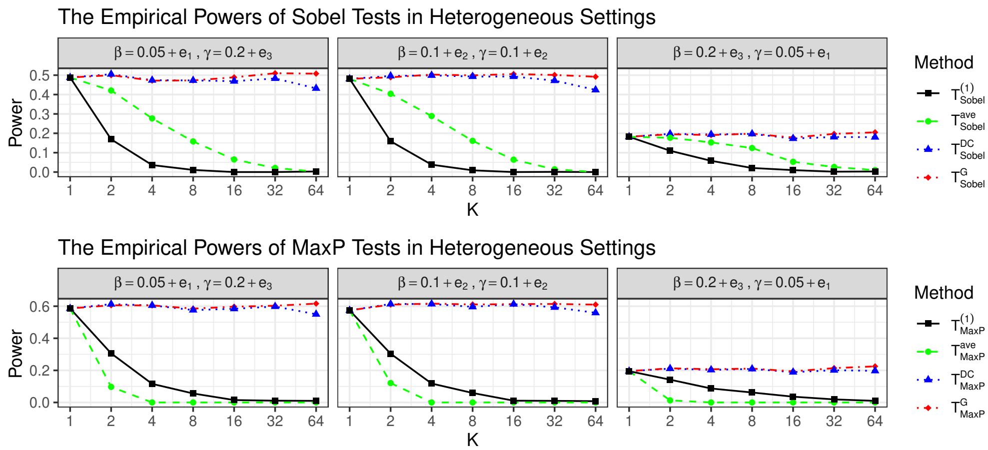

Figure 7: The empirical powers of the eight tests under three alternative hypothesis in heterogeneous settings, with the total sample size N fixed at 2^{11}.

Class 2: $n_2 = 14$, $T_\beta^{(2)} = 0.530207$, $T_\gamma^{(2)} = 0.981202$;

Class 3: $n_3 = 17$, $T_\beta^{(3)} = 4.829768$, $T_\gamma^{(3)} = 3.274552$.

The sample sizes of Class 2 and Class 3 are relatively small. Fortunately, the total sample size $N = n_1 + n_2 + n_3 = 65$ is sufficient. Although we do not have the original data for each class, the test results obtained using our proposed method are very close to the results from pooling the original data from all three classes. Let Dis_{ij} represent the distributed test using the local test statistics from classes i and j, for $(i, j) = (1, 2)$, $(1, 3)$, $(2, 3)$, and Dis_{123} represent the distributed test statistic using the local test statistics from all three classes. The corresponding p-values for these tests are shown in Table 1.

Table 1: The p-values of the tests.

Method	Class1	Class2	Class3	Dis_{12}	Dis_{13}	Dis_{23}	Dis_{123}
Sobel	0.106	0.6409	0.0067	0.0732	0.0032	0.0152	0.0032
MaxP	0.0848	0.596	0.0011	0.0476	0.001	0.002	0.0007

From this table, we can see that for a given significance level of $\alpha = 0.05$, when each class is tested individually, the results are as follows: the tests for Class 1 and Class 2 both fail to detect the mediation effect with Calculus as the mediator, while the test for Class 3 detects the mediation effect. However, by combining the test statistics from Class 1 and Class 2 using our proposed method, the resulting distributed MaxP test yields a p-value < 0.05, indicating the presence of a mediation effect. This means that although the individual tests for Class 1 and Class 2 failed to detect the mediation effect, our method, which combines their local test statistics, successfully identified the mediation effect through the distributed test. Furthermore, the Sobel and MaxP tests for Dis_{13}, Dis_{23}, and Dis_{123} all detected the presence of the mediation effect, and the p-values of the Dis_{123} are the smallest.

In conclusion, this study successfully identified the mediation effect between Gaokao mathematics performance, Calculus grades, and Probability and Mathematical Statistics course grades by pooling test statistics from multiple classes. Despite mixed results from individual class tests, our method overcame the limitations of small sample sizes and revealed that Gaokao mathematics performance influences statistics course grades through Calculus. This finding has important implications for teaching and learning in higher education mathematics courses.

The above results highlight a key application of distributed testing, especially in healthcare settings. Many hospitals face limitations in data availability due to factors like privacy concerns and regulatory restrictions, preventing them from sharing raw data. Our distributed method addresses this by only requiring hospitals to share test statistics rather than raw data. This allows the detection of mediation effects that might be missed in individual analyses, making it a valuable tool for healthcare institutions with limited data and privacy constraints.

6 Conclusion

In this article, we construct a distributed framework for mediation analysis. Building on Sobel's and MaxP tests, we introduce their distributed version, which are communication-efficient and easy to implement. Theoretical analysis and experiments show that the distributed tests achieve nearly identical power to global tests as long as $K = o(\sqrt{N})$. We apply our method to an educational study, demonstrating its ability to detect a mediation effect that would not be identifiable using data from only the first or second class, highlighting the advantage of our approach.

There are still several limitations. First, the Sobel's and MaxP tests are conservative and underpowered, and focus on one-dimensional mediator, so the proposed distributed test inherits these disadvantages. Fortunately, there are several enhanced tests based on Sobel's or MaxP tests, and therefore, their distributed versions can be easily derived from our distributed Sobel's or MaxP tests. For instance, suppose there are J candidate mediator variables $M_1, \ldots, M_j$, leading to J hypothesis tests $H_{0j} : \beta_j\gamma_j = 0$. Let $T^G_{MaxP,j}$ represents the global test statistics for the j-th test. Dai et al. [7] propose the HDMT method, which constructs the test statistic:

$$T_{HDMT,j} = (\hat{\pi}_{01} + \hat{\pi}_{10})T^G_{MaxP,j} + \hat{\pi}_{00}(T^G_{MaxP,j})^2,$$

where $\hat{\pi}_{01}$, $\hat{\pi}_{10}$, and $\hat{\pi}_{00}$ are the estimated proportions of $(\beta_j = 0, \gamma_j \neq 0)$, $(\beta_j \neq 0, \gamma_j = 0)$, and $(\beta_j = 0, \gamma_j = 0)$ across all J tests, respectively. In the distributed framework, $T^G_{MaxP,j}$ can be replaced by the proposed distributed test statistic $T^{Dis}_{MaxP,j}$, and $\hat{\pi}_{01}$ can be approximated by the simple average $\frac{1}{K}\sum_{k=1}^{K} \hat{\pi}^{(k)}_{01}$.

Second, our study assumes that the outcome Y is continuous. If Y is binary, we could fit a logistic model

$$\text{logit}(Pr(Y = 1|A, M, X)) = \beta_0 + \beta_A A + \beta_M M + \beta_X^T X$$

instead of the linear model in equation (1), and compute the local maximum likelihood estimate (MLE) of β, from which we could derive the distributed $\overline{T}_\beta$ based on the MLE. However, as shown in [19], the properties of the distributed $\overline{T}_\beta$ may be influenced by K. Therefore, improvements to our method are needed. These directions will be explored in future work.

7 Acknowledgment

This research is supported by the National Natural Science Foundation of China (12101056),the National Statistical Science Research Project (2022LY040), and the Talent Fund of Beijing Jiaotong University (2023XKRC008).

References

[1] Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

[2] Ernest Atta-Asiamah and Mingao Yuan. Distributed inference for degenerate u-statistics. *Stat*, 8(1):e234, 2019.

[3] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.

[4] Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics*, 46(3):1352–1383, 2018.

[5] T Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Federated nonparametric hypothesis testing with differential privacy constraints: Optimal rates and adaptive tests. *arXiv preprint arXiv:2406.06749*, 2024.

[6] Song Xi Chen and Liuhua Peng. Distributed statistical inference for massive data. *The Annals of Statistics*, 49(5):2851–2869, 2021.

[7] James Y Dai, Janet L Stanford, and Michael LeBlanc. A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 117(537):198–213, 2022.

[8] Jiarong Ding and Xuehu Zhu. Amdp: an adaptive detection procedure for false discovery rate control in high-dimensional mediation analysis. *Advances in Neural Information Processing Systems*, 36:65906–65935, 2023.

[9] Bin Du, Junlong Zhao, and Xin Zhang. Hypothesis testing of one sample mean vector in distributed frameworks. *Communications in Statistics-Simulation and Computation*, pages 1–18, 2024.

[10] Jiacong Du, Xiang Zhou, Dylan Clark-Boucher, Wei Hao, Yongmei Liu, Jennifer A Smith, and Bhramar Mukherjee. Methods for large-scale single mediator hypothesis testing: Possible choices and comparisons. *Genetic Epidemiology*, 47(2):167–184, 2023.

[11] Altay Eren. Uncovering the links between prospective teachers personal responsibility, academic optimism, hope, and emotions about teaching: A mediation analysis. *Social Psychology of Education*, 17:73–104, 2014.

[12] Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, 118(542):1000–1010, 2023.

[13] Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *Annals of Statistics*, 47(6):3009–3031, 2019.

[14] Tianyu Guo, Sai Praneeth Karimireddy, and Michael I Jordan. Collaborative heterogeneous causal inference beyond meta-analysis. *arXiv preprint arXiv:2404.15746*, 2024.

[15] Bingyao Huang, Yanyan Liu, and Liuhua Peng. Distributed inference for two-sample u-statistics in massive data analysis. *Scandinavian Journal of Statistics*, 50(3):1090–1115, 2023.

[16] Cheng Huang and Xiaoming Huo. A distributed one-step estimator. *Mathematical Programming*, 174:41–76, 2019.

[17] Yen-Tsung Huang. Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics*, 13(1):60–84, 2019.

[18] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010.

[19] Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

[20] Rémi Khellaf, Aurélien Bellet, and Julie Josse. Federated causal inference: Multi-study ate estimation beyond meta-analysis. *arXiv preprint arXiv:2410.16870*, 2024.

[21] Hopin Lee, Markus Hübscher, G Lorimer Moseley, Steven J Kamper, Adrian C Traeger, Gemma Mansell, and James H McAuley. How does pain lead to disability? a systematic review and meta-analysis of mediation studies in people with back and neck pain. *Pain*, 156(6):988–997, 2015.

[22] Loka Li, Ignavier Ng, Gongxu Luo, Biwei Huang, Guangyi Chen, Tongliang Liu, Bin Gu, and Kun Zhang. Federated causal discovery from heterogeneous data. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Zhonghua Liu, Jincheng Shen, Richard Barfield, Joel Schwartz, Andrea A Baccarelli, and Xihong Lin. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, 117(537):67–81, 2022.

[24] David P MacKinnon, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West, and Virgil Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83–104, 2002.

[25] David P MacKinnon, Chondra M Lockwood, and Jason Williams. Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39(1):99–128, 2004.

[26] Michèle B Nuijten, Ruud Wetzels, Dora Matzke, Conor V Dolan, and Eric-Jan Wagenmakers. A default bayesian hypothesis test for mediation. *Behavior Research Methods*, 47:85–97, 2015.

[27] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

[28] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

[29] Kenneth J Rothman and Sander Greenland. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95(S1):S144–S150, 2005.

[30] Katherine RK Saunders, Sabine Landau, Louise M Howard, Helen L Fisher, Louise Arseneault, Geraldine FH McLeod, and Sian Oram. Past-year intimate partner violence perpetration among people with and without depression: an individual participant data (ipd) meta-mediation analysis. *Social psychiatry and psychiatric epidemiology*, 58(12):1735–1747, 2023.

[31] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290–312, 1982.

[32] Zhenheng Tang, Yonggang Zhang, Peijie Dong, Yiu-ming Cheung, Amelie Zhou, Bo Han, and Xiaowen Chu. Fusefl: One-shot federated learning through the lens of causality with progressive model fusion. *Advances in Neural Information Processing Systems*, 37:28393–28429, 2024.

[33] Linda Valeri and Tyler J VanderWeele. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with sas and spss macros. *Psychological Methods*, 18(2):137–150, 2013.

[34] Tyler VanderWeele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2:457–468, 2009.

[35] Shuyuan Wu, Danyang Huang, and Hansheng Wang. Quasi-newton updating for large-scale distributed learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1326–1354, 2023.

[36] Qiaoling Ye, Arash A Amini, and Qing Zhou. Federated learning of generalized linear causal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6623–6636, 2024.

[37] Haixiang Zhang. Efficient adjusted joint significance test and sobel-type confidence interval for mediation effect. *Structural Equation Modeling: A Multidisciplinary Journal*, 32(1):93–104, 2025.

[38] Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013.

[39] Yanyan Zhao, Changliang Zou, and Zhaojun Wang. A scalable nonparametric specification testing for massive data. *Journal of Statistical Planning and Inference*, 200:161–175, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**The claims are supported with theoretical and empirical results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **[TODO]**We have discussed the limitations in the last paragraph in section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: **[TODO]**The assumptions are listed, but the complete proofs are not provided, as they are analogous to the proofs of Results 1 and 2 in [23] and are therefore omitted here.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: **[TODO]**The code provided in the supplemental materials can be used to reproduce results in both simulation study and real data anaalysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: **[TODO]**The code is in the supplementary matarials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**All steps for data generation, model fitting, evaluation, etc., are discussed in Section 4. Full details provided with the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: **[TODO]**The outcome in our simulation is the empirical power of a test, and error bars are typically not used in this context.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: **[TODO]**We have provided the information in the first paragraph in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **[TODO]**We conduct in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: **[TODO]**We have discussed the societal impacts of the work in Introduction. We believe that our distributed framework for testing mediation effects does not have any negative societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: **[TODO]**The paper does not utilize existing assets. Instead, the real application relies on the summarized information from the survey.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: **[TODO]** The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: **[TODO]**The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: **[TODO]** The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: **[TODO]**The LLM is used only for writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.