

A Framework for Mediation Analysis with Massive Data

Haixiang Zhang* and Xin Li

Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

Abstract

During the past few years, mediation analysis has gained increasing popularity across various research fields. The primary objective of mediation analysis is to examine the direct impact of exposure on outcome, as well as the indirect effects that occur along the pathways from exposure to outcome. There has been a great number of articles that applied mediation analysis to data from hundreds or thousands of individuals. With the rapid development of technology, the volume of available data increases exponentially, which brings new challenges to researchers. Directly conducting statistical analysis for large datasets is often computationally infeasible. Nonetheless, there is a paucity of findings regarding mediation analysis in the context of big data. In this paper, we propose utilizing subsampled double bootstrap and divide-and-conquer algorithms to conduct statistical mediation analysis on large-scale datasets. The proposed algorithms offer a significant enhancement in computational efficiency over traditional bootstrap confidence interval and Sobel test, while simultaneously ensuring desirable confidence interval coverage and power. We conducted extensive numerical simulations to evaluate the performance of our method. The practical applicability of our approach is demonstrated through two real-world data examples.

Keywords: Big data; Divide-and-Conquer; Mediation effects; Structural equation modeling; Subsampled double bootstrap

1 Introduction

Mediation analysis plays a crucial role in understanding the causal mechanism by which an independent variable X influences a dependent variable Y through an intermediate variable

*Corresponding author. Email: haixiang.zhang@tju.edu.cn (Haixiang Zhang)

(mediator) M . The topic of mediation analysis is widely researched in various fields, including but not limited to psychology, economics, epidemiology, medicine, sociology and behavioral science. Baron and Kenny (1986) laid the foundation for the development of mediation analysis in academic literature. Afterwards, numerous papers have been published to advance the development of mediation analysis. e.g., mediation effects with incomplete data (Wang and Zhang, 2011; Zhang and Wang, 2013), propensity score based methods (Jo *et al.*, 2011; Coffman, 2011), longitudinal mediation models (Soest and Hagtvat, 2011; Gunzler *et al.*, 2014), quantile mediation effects (Shen *et al.*, 2014), network mediation analysis (Liu *et al.*, 2021; Che *et al.*, 2021), high-dimensional mediation effects (Zhang *et al.*, 2016; Zhang *et al.*, 2021a; Zhang *et al.*, 2021b), Bayesian approaches for estimating mediation effects (Sun *et al.*, 2021; Miočević and Golchi, 2021). For more related results on mediation analysis, we refer to three reviewing papers by MacKinnon *et al.* (2007), Preacher (2015) and Zhang *et al.* (2022).

With the advancement of technology, there has been an exponential increase in the volume of available data in recent year. When dealing with massive datasets, the computational burden becomes onerous if we directly apply traditional statistical methods. There have been recent advancements in statistical methodologies for analyzing large-scale datasets. For example, Chen and Xie (2014) proposed a divide-and-conquer algorithm for generalized linear models with large-scale datasets; Sengupta *et al.* (2016) introduced a novel subsampled double bootstrap method for massive data. However, there is a lack of research on mediation analysis using large datasets. One recent example of studies that performed mediation analysis on large dataset is attributed to Hou *et al.* (2022). The authors proposed a method (PSE-MR) to identify and estimate path-specific effects of body mass index (BMI) on cardiovascular disease (CVD) through multiple causally ordered and non-ordered mediators (e.g., lipids) using summarized genetic data. This large dataset includes 694,649 participants from the Genetic Investigation of ANthropometric Traits (GIANT). There are two primary challenges that researchers may encounter when conducting mediation analysis on large datasets: The construction of bootstrap-based confidence intervals and the performance of Sobel test require high computing resources, while computation speed is relatively low. To address this issue, we present and assess the efficiency of two techniques that have the poten-

tial to reduce the computational burden associated with conducting mediation analyses on large datasets: subsampled double bootstrap (Sengupta *et al.*, 2016) for confidence intervals and the divide-and-conquer algorithm (Chen and Xie, 2014) for Sobel test. Specifically, the subsampled double bootstrap method is proposed to reduce the computational costs of traditional bootstrap-based mediation confidence intervals. Meanwhile, the divide-and-conquer approach is employed to alleviate the computational burden of the conventional Sobel test. Our aim is to efficiently conduct mediation analysis on large datasets, with a primary focus on speed, utilizing subsampled double bootstrap and divide-and-conquer methods.

The subsequent sections of this paper are structured as follows: In Section 2, we provide a comprehensive overview of the definitions and notations for linear and logistic mediation models within the potential outcomes framework. In Section 3, a refined approach of subsampled double bootstrap is employed to construct confidence intervals for mediation effects. In Section 4, we introduce a Sobel mediation testing procedure based on the divide-and-conquer approach. In Sections 5 and 6, we provide simulations and real-world data examples to validate the effectiveness of our methodologies. Section 7 provides concluding remarks.

2 Models and Notations

2.1 Linear Mediation Model

In this section, we will review the counterfactual (or potential outcome) framework for mediation models that involve multiple continuous mediators and a continuous outcome variable. Based on the notations of counterfactuals derived from literature on causal mediation analysis (Vanderweele and Vansteelandt, 2009), Let $\mathbf{M}(x)$ denote the potential value of a d -dimensional vector of mediators, represented as $(M_1(x), \dots, M_d(x))'$, under exposure level x . Denote $Y(x, \mathbf{m})$ as the potential outcome when x is the exposure and $\mathbf{m} = (m_1, \dots, m_d)'$ are the mediators. The linear mediation models, represented by counterfactual notations, are defined as follows:

$$Y(x, \mathbf{m}) = c + \gamma x + \beta_1 m_1 + \dots + \beta_d m_d + \boldsymbol{\theta}'\mathbf{Z} + \epsilon, \quad (2.1)$$

$$M_k(x) = c_k + \alpha_k x + \boldsymbol{\eta}'_k \mathbf{Z} + e_k, \quad k = 1, \dots, d, \quad (2.2)$$

where $\mathbf{Z} = (Z_1, \dots, Z_q)'$ represents a vector of confounding variables or covariates; γ represents the impact of exposure on the outcome while controlling for mediators and covariate; $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)'$ represents the relationship between mediators and outcome; $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)'$ represents the association between exposure and mediators while adjusting for $\mathbf{Z}$; $\boldsymbol{\theta}$ and $\boldsymbol{\eta}_k$ denote the regression coefficients of $\mathbf{Z}$; c and c_k are intercepts; ϵ is mean zero normal error, $\mathbf{e} = (e_1, \dots, e_d)'$ is a mean zero normal vector with covariance matrix $\boldsymbol{\Sigma}_e$.

To decompose the effect of an exposure on outcome into direct and indirect effects, we will first review some fundamental concepts within the framework of causal mediation analysis.

- Natural direct effect (NDE):

$$\begin{aligned} \text{NDE} &= E[Y(x, \mathbf{M}(x^*))] - E[Y(x^*, \mathbf{M}(x^*))] \\ &= \int_{\mathbf{m}} \{E(Y|x, \mathbf{m}) - E(Y|x^*, \mathbf{m})\} P(\mathbf{M} = \mathbf{m}|x^*) d\mathbf{m}. \end{aligned}$$

- Natural indirect effect (NIE):

$$\begin{aligned} \text{NIE} &= E[Y(x, \mathbf{M}(x))] - E[Y(x, \mathbf{M}(x^*))] \\ &= \int_{\mathbf{m}} E(Y|x, \mathbf{m}) \{P(\mathbf{M} = \mathbf{m}|x) - P(\mathbf{M} = \mathbf{m}|x^*)\} d\mathbf{m}. \end{aligned}$$

- Total effect (TE):

$$\begin{aligned} \text{TE} &= E[Y(x, \mathbf{M}(x))] - E[Y(x^*, \mathbf{M}(x^*))] \\ &= \text{NIE} + \text{NDE}. \end{aligned}$$

For the purpose of identifying causal effects, the causal mediation analysis literature requires adherence to four fundamental assumptions (Vanderweele and Vansteelandt, 2009).

(C.1) Stable Unit Treatment Value Assumption (SUTVA). There is no multiple versions of exposures and there is no interference between subjects, which implies that the observed variables are identical to the potential variables corresponding to the actually observed exposure level. i.e., $\mathbf{M} = \sum_x \mathbf{M}(x) I(X = x)$, and $Y = \sum_x \sum_{\mathbf{m}} Y(x, \mathbf{m}) I(X = x, \mathbf{M} = \mathbf{m})$, where $I(\cdot)$ is the indicator function.

(C.2) There are no measurement errors in the mediators and the outcome.

(C.3) Sequential ignorability: (i) $Y(x, \mathbf{m}) \perp X | \mathbf{Z}$, i.e., no unmeasured confounding between exposure and the potential outcome. (ii) $Y(x, \mathbf{m}) \perp \mathbf{M} | \{X, \mathbf{Z}\}$, i.e., no unmeasured confounding for the mediator-outcome relationship after adjusting for the exposure. (iii) $\mathbf{M}(x) \perp X | \mathbf{Z}$, i.e., no unmeasured confounding for the exposure effect on all the mediators. (iv) $Y(x, \mathbf{m}) \perp \mathbf{M}(x^*) | \mathbf{Z}$, i.e., no exposure-induced confounding between mediators and the potential outcome.

(C.4) The mediators are assumed to be causally independent. i.e., it is not allowed that one mediator is the cause of another.

Theorem 1 *Under the assumptions (C.1)-(C.4), if the mediation models (2.1) and (2.2) are correctly specified, then we have*

$$\begin{aligned} \text{NDE} &= \gamma(x - x^*), \\ \text{NIE} &= \sum_{k=1}^d \alpha_k \beta_k (x - x^*), \\ \text{TE} &= \left(\gamma + \sum_{k=1}^d \alpha_k \beta_k \right) (x - x^*). \end{aligned}$$

The results of Theorem 1 can be similarly derived as equations (2.6), (2.7) and (2.8) in Song (2020), thus the proof details are omitted here.

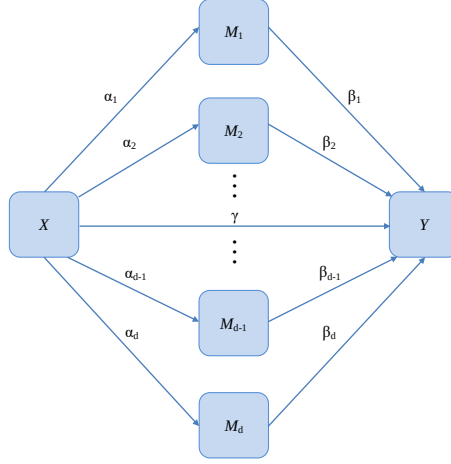


Figure 1: A scenario involving multiple mediators (excluding confounding variables).

Based on Theorem 1 and MacKinnon *et al.* (2020), we know that the total effect of X on Y is $\gamma + \sum_{k=1}^d \alpha_k \beta_k$, where γ is the *direct effect*, $\sum_{k=1}^d \alpha_k \beta_k$ is the *indirect effect*, and $\alpha_k \beta_k$ can be interpreted as the causal indirect effect transmitted by the k th mediator M_k along the pathway $X \rightarrow M_k \rightarrow Y$ (see Figure 1) for $k = 1, \dots, d$. i.e., the product-coefficient method of SEM-based mediation framework has the same causal interpretation as counterfactual approach when there is no X-M interaction term.

2.2 Logistic Mediation Model

For the case of a binary outcome, we consider the counterfactual mediation models that involve continuous mediators and a binary outcome:

$$P\{Y(x, \mathbf{m}) = 1\} = \frac{\exp\{c + \gamma x + \beta_1 m_1 + \dots + \beta_d m_d + \boldsymbol{\theta}' \mathbf{Z}\}}{1 + \exp\{c + \gamma x + \beta_1 m_1 + \dots + \beta_d m_d + \boldsymbol{\theta}' \mathbf{Z}\}}, \quad (2.3)$$

$$M_k(x) = c_k + \alpha_k x + \boldsymbol{\eta}'_k \mathbf{Z} + e_k, \quad k = 1, \dots, d, \quad (2.4)$$

where $Y(x, \mathbf{m}) \in \{0, 1\}$ is the potential outcome, $\mathbf{e} = (e_1, \dots, e_d)'$ is a vector of normal errors with mean zero and covariance matrix $\boldsymbol{\Sigma}_e$, other variables and parameters are similarly

defined as the linear mediation models in Section 2.1. Following VanderWeele and Vansteelandt (2010), we adopt the odds ratio scale for definitions of direct and indirect effects when the outcome is rare. Based on similar counterfactual notations as Section 2, the natural direct effect on the odds ratio scale (NDE^{OR}) has the form

$$\text{NDE}^{\text{OR}} = \frac{\text{P}\{Y(x, \mathbf{M}(x^*)) = 1\} / [1 - \text{P}\{Y(x, \mathbf{M}(x^*)) = 1\}]}{\text{P}\{Y(x^*, \mathbf{M}(x^*)) = 1\} / [1 - \text{P}\{Y(x^*, \mathbf{M}(x^*)) = 1\}]}; \quad (2.5)$$

the natural indirect effect on the odds ratio scale (NIE^{OR}) is given by

$$\text{NIE}^{\text{OR}} = \frac{\text{P}\{Y(x, \mathbf{M}(x)) = 1\} / [1 - \text{P}\{Y(x, \mathbf{M}(x)) = 1\}]}{\text{P}\{Y(x, \mathbf{M}(x^*)) = 1\} / [1 - \text{P}\{Y(x, \mathbf{M}(x^*)) = 1\}]}; \quad (2.6)$$

the total effect on the odds ratio scale (TE^{OR}) is

$$\text{TE}^{\text{OR}} = \frac{\text{P}\{Y(x, \mathbf{M}(x)) = 1\} / [1 - \text{P}\{Y(x, \mathbf{M}(x)) = 1\}]}{\text{P}\{Y(x^*, \mathbf{M}(x^*)) = 1\} / [1 - \text{P}\{Y(x^*, \mathbf{M}(x^*)) = 1\}]} \quad (2.7)$$

From the above three definitions, we have this expression $\text{TE}^{\text{OR}} = \text{NIE}^{\text{OR}} \cdot \text{NDE}^{\text{OR}}$. That is to say, $\log(\text{TE}^{\text{OR}}) = \log(\text{NIE}^{\text{OR}}) + \log(\text{NDE}^{\text{OR}})$ with the log scale.

Theorem 2 *Assuming that conditions (C.1)-(C.4) are satisfied and the event is rare, then according to models (2.3) and (2.4), we can conclude that*

$$\begin{aligned} \text{NDE}^{\text{OR}} &= \exp\{\gamma(x - x^*)\}, \\ \text{NIE}^{\text{OR}} &= \exp\left\{\sum_{k=1}^d \alpha_k \beta_k (x - x^*)\right\}, \\ \text{TE}^{\text{OR}} &= \exp\left\{\left(\gamma + \sum_{k=1}^d \alpha_k \beta_k\right)(x - x^*)\right\}. \end{aligned}$$

Given the rarity of the outcome and assuming that (C.1)-(C.4) are satisfied, Theorem 2 suggests that we may employ the product-of-coefficients $\alpha_k \beta_k$ to describe the mediation effect directly along the causal pathway $X \rightarrow M_k \rightarrow Y$ (see Figure 1). This stems from the conventional regression-based approach to mediation proposed by Baron and Kenny (1986). We recommend referring to VanderWeele and Vansteelandt (2010) for a more comprehensive understanding of logistic mediation models with a single mediator.

3 Subsampled Double Bootstrap Confidence Intervals

The Bootstrap (Efron, 1979) is a widely used and powerful statistical tool, but its computational demands can be extremely high for large-scale datasets. Sengupta *et al.* (2016) proposed a subsampled double bootstrap (SDB) method to deal with this problem. Its key idea is to randomly draw a small subset of the full data and use it to construct a full-size resample by generating repetition frequencies for each data point in the subset. Although the resample has the same sample size as the original full data, it only contains a small number of unique data points. The resamples are utilized in the same manner as the conventional Bootstrap. Notice that the computational cost of SDB mainly arises from the small subset rather than the full data set. Therefore, utilizing the SDB method significantly reduces computational costs when performing Bootstrap for massive data sets.

Suppose we have a set of n independent and identically distributed samples, denoted as $\mathcal{D}_n = \{(X_i, \mathbf{M}_i, Y_i, \mathbf{Z}_i)\}_{i=1}^n$. The confidence interval for the mediation effect is crucial in understanding the statistical impact of a mediator at a specified level of reliability, such as Biesanz *et al.* (2010), Preacher and Selig (2012) and Kisbu-Sakarya *et al.* (2014). Our objective is to utilize the SDB method in the context of massive data to construct confidence intervals for the causal mediation effects $\{\alpha_k \beta_k\}_{k=1}^d$. Following Sengupta *et al.* (2016), we define the “root function” of SDB method as

$$T_k = \sqrt{n}(\hat{\alpha}_k \hat{\beta}_k - \alpha_k \beta_k), \quad (3.1)$$

where $\alpha_k \beta_k$ is the true mediation effect of the k th mediator, $\hat{\alpha}_k$ and $\hat{\beta}_k$ are the estimators of α_k and β_k , respectively. Let $q_v^{(k)}$ represent the v -quantile of T_k , which can be estimated using the SDB procedure as follows: We first randomly select a subset of b points without replacement, $\mathcal{D}_b^* = \{(X_1^*, \mathbf{M}_1^*, Y_1^*, \mathbf{Z}_1^*), \dots, (X_b^*, \mathbf{M}_b^*, Y_b^*, \mathbf{Z}_b^*)\}$, from the full data set $\mathcal{D}_n$, where $b = n^r$ for some $0 < r < 1$. It is noteworthy that the elements in $\mathcal{D}_b^*$ are distinct data points. By repeating this procedure S times, we obtain S subsets denoted as $\mathcal{D}_b^{*(s)} = \{(X_1^{*(s)}, \mathbf{M}_1^{*(s)}, Y_1^{*(s)}, \mathbf{Z}_1^{*(s)}), \dots, (X_b^{*(s)}, \mathbf{M}_b^{*(s)}, Y_b^{*(s)}, \mathbf{Z}_b^{*(s)})\}$ for $s = 1, \dots, S$. Second, we generate a resample of size n denoted by $(\mathcal{D}_b^{*(s)}, \mathbf{W}_b^{*(s)})$, where $\mathbf{W}_b^{*(s)} = (w_1^{(s)}, \dots, w_b^{(s)})'$ is a vector representing the frequencies of $\{(X_i^{*(s)}, \mathbf{M}_i^{*(s)}, Y_i^{*(s)}, \mathbf{Z}_i^{*(s)})\}_{i=1}^b$ in the resample for $s = 1, \dots, S$. The weight vector $\mathbf{W}_b^{*(s)}$ satisfies $w_1^{(s)} + \dots + w_b^{(s)} = n$. That is to say, the

vector $(w_1^{(s)}, \dots, w_b^{(s)})'$ follows from $PN(n, \frac{1}{b}, \dots, \frac{1}{b})$, where $PN(n, \frac{1}{b}, \dots, \frac{1}{b})$ represents an n -trial uniform multinomial distribution over b objects. Third, let $\hat{\alpha}_k^{*(s)}$ and $\hat{\beta}_k^{*(s)}$ be the estimators based on $(\mathcal{D}_b^{*(s)}, \mathbf{W}_b^{*(s)})$. Calculate the values of root functions

$$T_k^{*(s)} = \sqrt{n} \{ \hat{\alpha}_k^{*(s)} \hat{\beta}_k^{*(s)} - \hat{\alpha}_k \hat{\beta}_k \}, \quad s = 1, \dots, S, \quad (3.2)$$

where $\hat{\alpha}_k$ and $\hat{\beta}_k$ are the estimators derived from the full data $\mathcal{D}_n$.

Denote $q_v^{*(k)}$ as the v -quantile of $\{T_k^{*(s)}\}_{s=1}^S$, which estimates the theoretical quantile $q_v^{(k)}$ of T_k . Subsequently, a single $(1 - \delta)$ -confidence interval for $\alpha_k \beta_k$ is estimated as

$$\text{CI}_{\text{single},k} = [\hat{\alpha}_k \hat{\beta}_k - n^{-1/2} q_{1-\delta/2}^{*(k)}, \hat{\alpha}_k \hat{\beta}_k + n^{-1/2} q_{\delta/2}^{*(k)}], \quad (3.3)$$

where δ represents the level of significance. Utilizing the Bonferroni technique (Alt and Spruill, 1977), we can construct Bonferroni adjusted confidence intervals for mediators M_k 's as follows:

$$\text{CI}_{\text{adjusted},k} = [\hat{\alpha}_k \hat{\beta}_k - n^{-1/2} q_{1-\delta/2d}^{*(k)}, \hat{\alpha}_k \hat{\beta}_k + n^{-1/2} q_{\delta/2d}^{*(k)}], \quad (3.4)$$

where $k = 1, \dots, d$. These intervals satisfy $P\{\alpha_1 \beta_1 \in \text{CI}_{\text{adjusted},1}, \dots, \alpha_d \beta_d \in \text{CI}_{\text{adjusted},d}\} = 1 - \delta$. To clarify, the confidence intervals presented in (3.4) have undergone Bonferroni correction to account for multiple testing.

We suggest employing the subsequent weighted least squares estimation equations for a linear mediation model:

$$Q_1^{(s)}(c, \gamma, \boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^b w_i^{(s)} \{Y_i^{*(s)} - c - \gamma X_i^{*(s)} - \beta_1 M_{i1}^{*(s)} - \dots - \beta_d M_{id}^{*(s)} - \theta_1 Z_{i1}^{*(s)} - \dots - \theta_q Z_{iq}^{*(s)}\}^2, \quad (3.5)$$

and

$$Q_{2k}^{(s)}(c_k, \alpha_k, \boldsymbol{\eta}_k) = \sum_{i=1}^b w_i^{(s)} \{M_{ik}^{*(s)} - c_k - \alpha_k X_i^{*(s)} - \eta_{k1} Z_{i1}^{*(s)} - \dots - \eta_{kq} Z_{iq}^{*(s)}\}^2, \quad k = 1, \dots, d, \quad (3.6)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)'$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$ and $\boldsymbol{\eta}_k = (\eta_{k1}, \dots, \eta_{kq})'$. The weighted least squares estimators $\hat{\beta}_k^{*(s)}$ and $\hat{\alpha}_k^{*(s)}$ can be easily obtained by minimizing (3.5) and (3.6), respectively, for $k = 1, \dots, d$.

For the logistic mediation model, we propose the following weighted negative log-likelihood function:

$$Q_3^{(s)}(c, \gamma, \boldsymbol{\beta}, \boldsymbol{\theta}) = - \sum_{i=1}^b w_i^{(s)} [Y_i^{*(s)} \log p_i^{*(s)}(c, \gamma, \boldsymbol{\beta}, \boldsymbol{\theta}) + (1 - Y_i^{*(s)}) \log \{1 - p_i^{*(s)}(c, \gamma, \boldsymbol{\beta}, \boldsymbol{\theta})\}]^2, \quad (3.7)$$

where

$$p_i^{*(s)}(c, \gamma, \boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\exp\{c + \gamma X_i^{*(s)} + \beta_1 M_{i1}^{*(s)} + \cdots + \beta_d M_{id}^{*(s)} + \theta_1 Z_{i1}^{*(s)} + \cdots + \theta_q Z_{iq}^{*(s)}\}}{1 + \exp\{c + \gamma X_i^{*(s)} + \beta_1 M_{i1}^{*(s)} + \cdots + \beta_d M_{id}^{*(s)} + \theta_1 Z_{i1}^{*(s)} + \cdots + \theta_q Z_{iq}^{*(s)}\}}.$$

The corresponding estimators $\{\hat{\alpha}_k^{*(s)}\}_{k=1}^d$ and $\{\hat{\beta}_k^{*(s)}\}_{k=1}^d$ can be obtained through the minimization of (3.6) and (3.7), respectively, corresponding to their respective parameters.

Algorithm 1 summarizes the SDB-based method proposed in this study.

Algorithm 1 SDB-based Confidence Intervals for Mediation Effects

- *Step 1:* We randomly select a subset $\mathcal{D}_b^{*(s)}$ of size b ($b \ll n$) without replacement from the full data $\mathcal{D}_n$.
 - *Step 2:* Generate a resample $(\mathcal{D}_b^{*(s)}, \mathbf{W}_b^{*(s)})$, where $\mathbf{W}_b^{*(s)} = (w_1^{(s)}, \dots, w_b^{(s)})'$ is a vector representing the frequencies of $\{(X_i^{*(s)}, \mathbf{M}_i^{*(s)}, Y_i^{*(s)}, \mathbf{Z}_i^{*(s)})\}_{i=1}^b$ in the resample, Specifically, $(w_1^{(s)}, \dots, w_b^{(s)}) \sim PN(n, \frac{1}{b}, \dots, \frac{1}{b})$.
 - *Step 3:* Calculate the estimators $\hat{\alpha}_k^{*(s)}$ and $\hat{\beta}_k^{*(s)}$, together with the statistic $T_k^{*(s)}$ in (3.2).
 - *Step 4:* Repeat *Steps 1-3* for S times (e.g. $S = 500$), and calculate the critical values $q_{1-\delta/2}^{*(k)}$, $q_{\delta/2}^{*(k)}$, $q_{1-\delta/2d}^{*(k)}$ and $q_{\delta/2d}^{*(k)}$. Output the confidence intervals as given in (3.3) and (3.4).
-

4 Divide-and-Conquer Sobel Test

In this section, our focus is on estimating the mediation effects $\alpha_k \beta_k$'s and conducting multiple tests for

$$H_{0k} : \alpha_k \beta_k = 0 \leftrightarrow H_{Ak} : \alpha_k \beta_k \neq 0, \quad k = 1, \dots, d, \quad (4.1)$$

within the context of big data. Assume that we have a set of independent and identically distributed (i.i.d.) samples, denoted as $\mathcal{D}_n = \{(X_i, \mathbf{M}_i, Y_i, \mathbf{Z}_i)\}_{i=1}^n$. One of the most commonly

used methods for (4.1) is Sobel's test (Sobel, 1982), which employs

$$T_k = \frac{\hat{\alpha}_k \hat{\beta}_k}{\hat{\sigma}_{\alpha_k \beta_k}}. \quad (4.2)$$

Here, the estimator $\hat{\alpha}_k$ is obtained through ordinary least squares regression with a variance of $\hat{\sigma}_{\alpha_k}^2$. The estimator $\hat{\beta}_k$ is obtained through ordinary least squares regression for continuous outcomes and maximum likelihood estimation for binary outcomes, with a variance of $\hat{\sigma}_{\beta_k}^2$. According to Sobel (1982), the standard variance of $\hat{\alpha}_k \hat{\beta}_k$ is given by

$$\hat{\sigma}_{\alpha_k \beta_k} = \{\hat{\alpha}_k^2 \hat{\sigma}_{\beta_k}^2 + \hat{\beta}_k^2 \hat{\sigma}_{\alpha_k}^2\}^{1/2}, \quad k = 1, \dots, d. \quad (4.3)$$

Let $\Omega = \{k : \alpha_k \beta_k \neq 0, k = 1, \dots, d\}$ denote the index set of significant mediators. The Bonferroni adjusted p -value for testing $H_{0k} : \alpha_k \beta_k = 0$ is calculated as

$$P_k = 2d\{1 - \Phi(|T_k|)\}, \quad (4.4)$$

where T_k is defined in (4.2) and $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0, 1)$. The estimated index set of significant mediators is denoted as $\hat{\Omega} = \{k : P_k < 0.05, k = 1, \dots, d\}$ under the significance level of 0.05.

When the sample size n is extremely large, the computational burden of T_k given in (4.2) becomes onerous. To address this issue, we employ a divide-and-conquer approach (Battey *et al.*, 2018; Shi *et al.*, 2018; Volgushev *et al.*, 2019) when computing the statistics T_k 's. We partition the entire dataset $\mathcal{D}_n$ into J disjoint subsets, denoted as $\mathcal{D}_n^{(1)}, \dots, \mathcal{D}_n^{(J)}$, and estimate parameters based on each subset $\mathcal{D}_n^{(j)}$ randomly selected from the full data. For $j = 1, \dots, J$, let $\hat{\alpha}_k^{(j)}$ denote the least squares estimator obtained from $\mathcal{D}_n^{(j)}$, and let $\hat{\beta}_k^{(j)}$ denote the maximum likelihood estimator obtained from $\mathcal{D}_n^{(j)}$. Furthermore, the estimated standard errors of $\hat{\alpha}_k^{(j)}$ and $\hat{\beta}_k^{(j)}$ are denoted by $\hat{\sigma}_{\alpha_k}^{(j)}$ and $\hat{\sigma}_{\beta_k}^{(j)}$, respectively. Similar to (4.3), we obtain $\hat{\sigma}_{\hat{\alpha}_k \hat{\beta}_k}^{(j)} = [\{\hat{\alpha}_k^{(j)} \hat{\sigma}_{\beta_k}^{(j)}\}^2 + \{\hat{\beta}_k^{(j)} \hat{\sigma}_{\alpha_k}^{(j)}\}^2]^{1/2}$ for $k = 1, \dots, d$ and $j = 1, \dots, J$. The final estimator for M_k 's mediation effect is

$$\hat{\alpha}_k \hat{\beta}_k = \frac{1}{J} \sum_{j=1}^J \hat{\alpha}_k^{(j)} \hat{\beta}_k^{(j)}, \quad (4.5)$$

and its estimated standard error is given by

$$\hat{\sigma}_{\alpha_k \beta_k} = \frac{1}{J} \left\{ \sum_{j=1}^J (\hat{\sigma}_{\hat{\alpha}_k \hat{\beta}_k}^{(j)})^2 \right\}^{1/2}. \quad (4.6)$$

Algorithm 2 presents a summary of the proposed divide-and-conquer Sobel test for mediation effects.

Algorithm 2 Divide-and-Conquer Sobel Test

- *Step 1:* Divide the entire dataset $\mathcal{D}_n$ into J disjoint subsamples of equal size n/J , denoted as $\mathcal{D}_n^{(1)}, \mathcal{D}_n^{(2)}, \dots$ and $\mathcal{D}_n^{(J)}$, respectively.
 - *Step 2:* For each $\mathcal{D}_n^{(j)}$, we calculate the estimators $\hat{\alpha}_k^{(j)}, \hat{\beta}_k^{(j)}$ and $\hat{\sigma}_{\alpha_k\beta_k}^{(j)}$. The final estimators of $\hat{\alpha}_k\hat{\beta}_k$ and $\hat{\sigma}_{\alpha_k\beta_k}$ are given in (4.5) and (4.6), respectively.
 - *Step 3:* Conduct multiple tests for the null hypothesis $H_{0k} : \alpha_k\beta_k = 0$, where $k = 1, \dots, d$. If P_k is less than 0.05, reject the hypothesis H_{0k} based on (4.4). i.e., $\hat{\Omega} = \{k : P_k < 0.05, k = 1, \dots, d\}$.
-

5 Simulation Study

5.1 SDB Confidence Intervals

In this section, we perform simulations to assess the performance of SDB-based confidence intervals as discussed in Section 2. The mediators M_k 's are generated from

$$M_k = c_k + \alpha_k X + \boldsymbol{\eta}'_k \mathbf{Z} + e_k, \quad k = 1, \dots, d. \quad (5.1)$$

We consider two types of outcomes (continuous and binary) as follows:

- Linear model:

$$Y = c + \gamma X + \beta_1 M_1 + \dots + \beta_d M_d + \boldsymbol{\theta}' \mathbf{Z} + \epsilon; \quad (5.2)$$

- Logistic model:

$$P\{Y = 1\} = \frac{\exp\{\gamma X + \beta_1 M_1 + \dots + \beta_d M_d + \boldsymbol{\theta}' \mathbf{Z}\}}{1 + \exp\{\gamma X + \beta_1 M_1 + \dots + \beta_d M_d + \boldsymbol{\theta}' \mathbf{Z}\}}. \quad (5.3)$$

With the help of R software, we generate random samples from models (5.1), (5.2) and (5.3), where $\boldsymbol{\alpha} = (0, 0.2, 0, 0.1, 0.15)'$, $\boldsymbol{\beta} = (0, 0, 0.2, 0.1, 0.15)'$, $c = c_k = \gamma = 0.5$ and $\boldsymbol{\theta} = \boldsymbol{\eta}_k = (1, 1)'$. We generate $\mathbf{e} = (e_1, e_2, e_3, e_4, e_5)'$ from $N_5(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (0.5^{|i-j|})$. ϵ follows from $N(0, 4)$, $\mathbf{Z} = (Z_1, Z_2)'$ with Z_1 and Z_2 being generated from $N(0, 2)$. We consider three different scenarios for generating exposure X :

Case 1: The variable X is normally distributed with mean 0 and standard deviation 1.

Case 2: The variable X is derived from the t-distribution with 5 degrees of freedom.

Case 3: The variable X is generated from an exponential distribution with rate 1.

As suggested by the reviewer, it would be informative to compare the performance of SDB confidence intervals with that of conventional percentile Bootstrap-based confidence intervals using full data (denoted as “Bootstrap”). The subset size for the SDB method is chosen as $b = n^{0.7}$, where the full data sample size is $n = 10^5$. The results presented in Tables 1-4 are based on 500 repetitions.

In Table 1, we present the coverage probabilities (CPs) of individual mediators’ 95% confidence intervals obtained through SDB and Bootstrap methods, respectively. Based on the findings presented in Table 1, all CPs are acceptable compared to 0.95, except for scenarios in which both the exposure-mediator and mediator-outcome effects are zero. Ideally, the coverage probability is 0.95, where a coverage probability between 0.925 and 0.975 is deemed acceptable based on Bradley’s liberal robustness criterion (Bradley, 1978; Valente and MacKinnon, 2017; Miočević *et al.*, 2021). Based on this criterion, the SDB and Bootstrap perform similar in terms of CPs.

In Table 2, we present the mean lengths of estimated confidence intervals for the SDB and Bootstrap methods. It can be observed from Table 2 that the SDB method yields slightly longer confidence intervals than the Bootstrap method. However, these differences are very small and negligible in practice. To investigate the performance of Bonferroni-adjusted confidence intervals for both SDB and Bootstrap methods, we present the CPs and average lengths of estimated confidence intervals in Tables 3 and 4, which yield comparable conclusions to those drawn from Tables 1 and 2. In terms of precision, the proposed SDB method is comparable to the Bootstrap method in constructing confidence intervals for mediation effects.

We perform a second simulation to compare the computational efficiency of the proposed SDB method with that of the traditional percentile Bootstrap method. The data is generated as the first simulation with Case I, except that $\alpha = (0.5, \dots, 0.5)'$ and $\beta = (0.5, \dots, 0.5)'$ for $d = 5, 10$ and 20 , respectively. The subset size for SDB method is determined as $b = n^{0.7}$, where the full data size is $n = 10^5$. The computation is implemented by a computer

Table 1: The coverage probabilities of estimated confidence intervals with level $0.95^\dagger$.

Models	(α_k, β_k)	SDB			Bootstrap		
		Case I	Case II	Case III	Case I	Case II	Case III
Linear	$(\alpha_1, \beta_1) = (0, 0)$	1	1	1	0.996	1	0.992
	$(\alpha_2, \beta_2) = (0.2, 0)$	0.956	0.944	0.972	0.952	0.942	0.962
	$(\alpha_3, \beta_3) = (0, 0.2)$	0.956	0.948	0.942	0.948	0.934	0.932
	$(\alpha_4, \beta_4) = (0.1, 0.1)$	0.924	0.948	0.964	0.922	0.938	0.958
	$(\alpha_5, \beta_5) = (0.15, 0.15)$	0.958	0.958	0.950	0.950	0.956	0.944
Logistic	$(\alpha_1, \beta_1) = (0, 0)$	1	1	1	1	1	1
	$(\alpha_2, \beta_2) = (0.2, 0)$	0.956	0.946	0.964	0.964	0.938	0.960
	$(\alpha_3, \beta_3) = (0, 0.2)$	0.972	0.960	0.972	0.952	0.944	0.938
	$(\alpha_4, \beta_4) = (0.1, 0.1)$	0.956	0.946	0.956	0.944	0.936	0.954
	$(\alpha_5, \beta_5) = (0.15, 0.15)$	0.962	0.968	0.948	0.948	0.964	0.944

† “SDB” denotes our method with subsampled double bootstrap algorithm; “Bootstrap” denotes the conventional percentile Bootstrap method with full data.

Table 2: The average lengths ($\times 10^4$) of estimated confidence intervals with level 0.95[†].

Models	(α_k, β_k)	SDB			Bootstrap		
		Case I	Case II	Case III	Case I	Case II	Case III
Linear	$(\alpha_1, \beta_1) = (0, 0)$	6.698	5.196	6.695	1.362	1.075	1.399
	$(\alpha_2, \beta_2) = (0.2, 0)$	49.71	49.43	49.75	49.22	49.10	49.26
	$(\alpha_3, \beta_3) = (0, 0.2)$	25.58	19.89	25.60	24.59	19.07	24.55
	$(\alpha_4, \beta_4) = (0.1, 0.1)$	28.47	26.99	28.58	27.55	26.36	27.60
	$(\alpha_5, \beta_5) = (0.15, 0.15)$	41.77	39.81	41.78	41.41	39.53	41.19
Logistic	$(\alpha_1, \beta_1) = (0, 0)$	9.972	7.764	10.07	2.041	1.547	2.054
	$(\alpha_2, \beta_2) = (0.2, 0)$	73.57	73.45	74.34	72.22	72.62	72.95
	$(\alpha_3, \beta_3) = (0, 0.2)$	27.05	20.93	26.93	24.73	19.09	24.49
	$(\alpha_4, \beta_4) = (0.1, 0.1)$	39.95	38.63	40.41	38.29	37.64	38.73
	$(\alpha_5, \beta_5) = (0.15, 0.15)$	58.95	57.42	59.40	57.63	56.69	58.02

[†] The meanings of “SDB” and “Bootstrap” are given in Table 1. The numbers are increased by 10^4 times.

Table 3: The coverage probabilities of Bonferroni adjusted confidence intervals[†].

Models	SDB			Bootstrap		
	Case I	Case II	Case III	Case I	Case II	Case III
Linear	0.950	0.948	0.966	0.946	0.938	0.940
Logistic	0.970	0.964	0.958	0.956	0.934	0.952

[†] The meanings of “SDB” and “Bootstrap” are given in Table 1; The confidence level is 0.95.

Table 4: The average lengths ($\times 10^4$) of Bonferroni adjusted confidence intervals with level 0.95[†].

Models	(α_k, β_k)	SDB			Bootstrap		
		Case I	Case II	Case III	Case I	Case II	Case III
Linear	$(\alpha_1, \beta_1) = (0, 0)$	10.11	7.829	10.05	2.047	1.606	2.088
	$(\alpha_2, \beta_2) = (0.2, 0)$	64.93	64.27	64.73	63.83	63.36	63.76
	$(\alpha_3, \beta_3) = (0, 0.2)$	34.13	26.57	34.11	31.73	24.67	31.68
	$(\alpha_4, \beta_4) = (0.1, 0.1)$	37.86	35.38	37.97	35.55	33.92	35.59
	$(\alpha_5, \beta_5) = (0.15, 0.15)$	54.83	51.99	54.53	53.56	50.99	53.01
Logistic	$(\alpha_1, \beta_1) = (0, 0)$	14.99	11.66	15.11	3.058	2.333	3.071
	$(\alpha_2, \beta_2) = (0.2, 0)$	95.77	95.48	97.31	93.00	93.91	94.47
	$(\alpha_3, \beta_3) = (0, 0.2)$	36.93	28.58	36.89	31.93	24.71	31.64
	$(\alpha_4, \beta_4) = (0.1, 0.1)$	53.25	51.14	53.90	49.56	48.57	50.05
	$(\alpha_5, \beta_5) = (0.15, 0.15)$	77.33	75.09	77.82	74.23	73.34	74.99

[†] The meanings of “SDB” and “Bootstrap” are given in Table 1. The numbers are increased by 10^4 times.

with 64GB memory (running R code). In Table 5, we present the average computation times (in seconds) for SDB and Bootstrap methods based on 10 repetitions, without taking into account data generation. The results demonstrate that the proposed SDB method exhibits significantly higher computational efficiency compared to the Bootstrap procedure. By Sengupta *et al.* (2016), increasing the size of the subset b results in greater benefits in terms of statistical accuracy, but at a higher computational cost. For a given computational time budget, it is still unclear how to choose an optimal subset size that balances statistical accuracy and running time for SDB procedure. From a practical perspective, the practitioner may choose the largest possible subset size within a given computational time budget.

Table 5: The comparison of computation time (in seconds) with Case I[†].

Models	SDB			Bootstrap		
	$d = 5$	$d = 10$	$d = 20$	$d = 5$	$d = 10$	$d = 20$
Linear	16.292	26.615	48.951	180.72	326.38	633.98
Logistic	20.211	32.108	60.498	299.17	472.72	917.27

[†] The meanings of “SDB” and “Bootstrap” are given in Table 1; d is the number of mediators.

5.2 DC-based Sobel Test

In this section, we perform simulations to assess the efficacy of the Sobel test method based on divide-and-conquer (DC) approach. We consider two types of outcomes and generate random samples from models (5.1), (5.2) and (5.3). The settings are similar to those in Section 5.1, except that a linear model is used with $\boldsymbol{\alpha} = (0, 0.1, 0, 0.025, 0.035)'$ and $\boldsymbol{\beta} = (0, 0, 0.15, 0.025, 0.035)'$, as well as a logistic model with $\boldsymbol{\alpha} = (0, 0.15, 0, 0.025, 0.035)'$ and $\boldsymbol{\beta} = (0, 0, 0.15, 0.035, 0.05)'$. The sample size for full data is chosen as $n = 10^5$. The number of blocks are $J = 1, 5, 50$ and 100 , respectively. We would like to emphasize that the scenario with $J = 1$ is based on the entire sample, without any subsampling, and thus corresponds to a standard analysis. The results in Tables 6-9 are based on 500 repetitions.

Tables 6-8 present the estimated bias (Bias), which is calculated as the average of mediation effect estimates minus the true value, and mean square error (MSE) of mediation effect

Table 6: Bias ($\times 10^6$) and MSE ($\times 10^8$; in parenthesis) for mediation effects with Case 1[†].

Models	(α_k, β_k)	$J = 1$	$J = 5$	$J = 50$	$J = 100$
Linear	(0, 0)	-0.2101	1.4898	1.7893	-2.8824
		(0.0534)	(0.2237)	(1.7845)	(3.6766)
	(0.1, 0)	-14.388	-13.879	-21.925	-17.121
		(36.628)	(36.547)	(39.541)	(41.749)
	(0, 0.15)	-43.309	-45.686	-46.746	-39.853
		(23.569)	(23.747)	(24.905)	(27.193)
	(0.025, 0.025)	13.504	14.310	7.8653	7.6704
		(2.9042)	(3.0187)	(4.8931)	(6.1670)
	(0.035, 0.035)	6.1792	2.5440	-8.8570	-9.4346
		(6.7853)	(6.8542)	(8.9651)	(10.821)
Logistic	(0, 0)	0.3281	-1.6809	-13.570	-16.801
		(0.0798)	(0.4096)	(3.7180)	(8.1516)
	(0.15, 0)	-86.649	-84.675	-87.789	-105.19
		(172.21)	(172.38)	(183.22)	(190.29)
	(0, 0.15)	-46.762	-44.841	-38.833	-28.515
		(22.282)	(22.873)	(27.155)	(30.603)
	(0.025, 0.035)	14.052	14.863	19.785	28.341
		(5.9699)	(6.4280)	(11.044)	(14.347)
	(0.035, 0.05)	2.0318	1.7772	16.018	38.367
		(12.071)	(12.633)	(16.705)	(21.584)

[†] $J = 1$ is the same as a standard analysis without subsamples. The Bias and MSE are increased by 10^6 and 10^8 times, respectively.

estimates given by

$$\text{MSE}(\alpha_k\beta_k) = \frac{1}{500} \sum_{j=1}^{500} \left\{ \hat{\alpha}_k^{(j)} \hat{\beta}_k^{(j)} - \alpha_k\beta_k \right\}^2,$$

where $\hat{\alpha}_k^{(j)}$ and $\hat{\beta}_k^{(j)}$ are the corresponding estimates from the j th repetition, $j = 1, \dots, 500$. It can be seen from Tables 6-8 that all estimators are unbiased towards the true values of mediation effects. As the number of blocks increases, the differences in Bias (or MSE) are very small and can be considered negligible in practical applications.

Table 9 reports the results of family-wise error rate (FWER) and Power when conducting multiple testing with the Algorithm 2, where

$$H_{0k} : \alpha_k\beta_k \neq 0 \leftrightarrow H_{1k} : \alpha_k\beta_k \neq 0, \quad k = 1, \dots, 5,$$

at a significance level of 0.05. Specifically, the Power and FWER are defined as

$$\text{Power} = \frac{1}{|\Omega|} \sum_{k \in \Omega} \mathbb{P}(P_k < 0.05), \quad (5.4)$$

and

$$\text{FWER} = \mathbb{P}(\exists k \in \Omega^c : P_k < 0.05), \quad (5.5)$$

where $\Omega = \{4, 5\}$, $\Omega^c = \{1, 2, 3\}$ and P_k 's are obtained by the divide-and-conquer algorithm. The results in Table 9 indicate that Power decreases as the number of blocks increases. The FWERs are much smaller than the significance level of 0.05. Essentially, our method involves a tradeoff between statistical efficiency and computational efficiency, which is a common issue for divide-and-conquer based procedures.

We conduct the second simulation to investigate the computational efficiency of Algorithm 2. The generation of data is the same as the first simulation with Case 1, except that $\boldsymbol{\alpha} = (0.5, \dots, 0.5)'$ and $\boldsymbol{\beta} = (0.5, \dots, 0.5)'$, where $d = 5, 50$ and 100 , respectively. The computation is carried out by a computer with 64GB of memory. To simulate scenarios involving multiple machines simultaneously, we vary the sample sizes as $n, n/5, n/50$ and $n/100$ corresponding to J values of $1, 5, 50$ and 100 , respectively. Here $J = 1$ is actually the conventional analysis of all the data (i.e., no subsamples). In Table 10, we report the

Table 7: Bias ($\times 10^6$) and MSE ($\times 10^8$; in parenthesis) for mediation effects with Case 2[†].

Models	(α_k, β_k)	$J = 1$	$J = 5$	$J = 50$	$J = 100$
Linear	(0, 0)	0.6877	1.2694	0.1059	5.2722
		(0.0241)	(0.1211)	(1.3666)	(2.5342)
	(0.1, 0)	-1.1405	-1.988	-0.3675	-2.9085
		(41.227)	(41.337)	(42.253)	(44.342)
	(0, 0.15)	3.8327	1.6769	3.6794	2.1754
		(13.338)	(13.310)	(15.247)	(16.092)
	(0.025, 0.025)	3.0014	5.3530	5.4643	7.1933
		(3.0211)	(3.0332)	(4.0692)	(5.4192)
	(0.035, 0.035)	-8.7985	-8.7657	-5.1591	-4.1719
		(6.1348)	(6.2202)	(7.6394)	(8.8628)
Logistic	(0, 0)	-1.0418	-3.3285	-5.7510	-7.6715
		(0.0417)	(0.2051)	(2.4418)	(4.6024)
	(0.15, 0)	19.296	19.469	17.297	10.674
		(168.19)	(168.59)	(174.62)	(181.89)
	(0, 0.15)	8.8453	10.717	11.768	4.7968
		(12.992)	(13.215)	(16.153)	(17.771)
	(0.025, 0.035)	2.8607	2.8227	15.713	23.402
		(5.4237)	(5.6156)	(7.7186)	(10.395)
	(0.035, 0.05)	-2.6734	-1.8057	9.0643	11.053
		(10.983)	(11.058)	(14.024)	(16.714)

[†] $J = 1$ is the same as a standard analysis without subsamples. The Bias and MSE are increased by 10^6 and 10^8 times, respectively.

Table 8: Bias ($\times 10^6$) and MSE ($\times 10^8$; in parenthesis) for mediation effects with Case 3[†].

Models	(α_k, β_k)	$J = 1$	$J = 5$	$J = 50$	$J = 100$
Linear	(0, 0)	-0.5258	-0.7788	-1.3620	-4.4085
		(0.0359)	(0.2017)	(1.8676)	(3.7627)
	(0.1, 0)	2.4277	1.4731	-9.7915	-12.549
		(41.698)	(42.180)	(44.403)	(45.680)
	(0, 0.15)	5.5580	6.9381	9.8424	1.7548
		(19.516)	(19.412)	(21.151)	(24.043)
	(0.025, 0.025)	-7.7499	-7.3152	-16.176	-12.278
		(3.2775)	(3.5264)	(5.5125)	(8.0929)
	(0.035, 0.035)	-2.8638	-6.5763	-5.0265	-0.2281
		(5.4613)	(5.5391)	(7.4047)	(10.017)
Logistic	(0, 0)	1.7192	2.8207	0.4644	-10.200
		(0.0691)	(0.4125)	(4.1914)	(8.6075)
	(0.15, 0)	-34.695	-34.809	-27.133	-11.657
		(173.65)	(172.98)	(181.44)	(191.11)
	(0, 0.15)	-41.219	-42.612	-53.293	-67.098
		(23.108)	(23.644)	(28.319)	(32.728)
	(0.025, 0.035)	-12.907	-12.109	-7.6015	-16.501
		(5.9377)	(6.1271)	(9.8584)	(14.336)
	(0.035, 0.05)	15.605	21.180	36.543	51.505
		(11.948)	(12.029)	(15.972)	(21.353)

[†] $J = 1$ is the same as a standard analysis without subsamples. The Bias and MSE are increased by 10^6 and 10^8 times, respectively.

Table 9: The family-wise error rate (FWER) and Power for multiple testing[†].

Models	Settings		$J = 1$	$J = 5$	$J = 50$	$J = 100$
Linear	Case 1	FWER	0.024	0.020	0.018	0.010
		Power	0.954	0.935	0.649	0.476
	Case 2	FWER	0.016	0.016	0.014	0.010
		Power	0.945	0.936	0.770	0.635
	Case 3	FWER	0.018	0.018	0.018	0.018
		Power	0.937	0.914	0.658	0.480
Logistic	Case 1	FWER	0.028	0.028	0.024	0.022
		Power	0.952	0.924	0.667	0.511
	Case 2	FWER	0.020	0.020	0.010	0.008
		Power	0.955	0.946	0.787	0.636
	Case 3	FWER	0.014	0.018	0.020	0.016
		Power	0.934	0.915	0.655	0.500

[†] $J = 1$ is the same as a standard analysis without subsamples.

computation times of Algorithm 2 without accounting for data generation. These times are the average of 10 repetitions and are measured in seconds. According to Table 10, the computational efficiency of the divide-and-conquer algorithm increases with the utilization of additional computational resources. In consideration of the tradeoff between statistical efficiency and computational speed, it is recommended to minimize the use of computational resources within the specified time budget.

Table 10: The comparison of computation time (in seconds) with Case 1[†].

Models	dimension	$J = 1$	$J = 5$	$J = 50$	$J = 100$
Linear	$d = 5$	0.352	0.071	0.016	0.014
	$d = 50$	3.292	0.575	0.092	0.070
	$d = 100$	6.543	1.116	0.208	0.149
Logistic	$d = 5$	0.785	0.155	0.025	0.015
	$d = 50$	7.044	1.389	0.238	0.133
	$d = 100$	21.279	3.735	0.607	0.335

[†] $J = 1$ is the same as a standard analysis without subsamples; d is the number of mediators.

6 Application

6.1 Continuous Outcomes

In this section, we apply the SDB-based confidence intervals and DC-based Sobel Test to a large dataset about the P2P lending platform. The P2P is the abbreviation of peer to peer, which is a new type of network lending platform. The dataset is about the loan transaction data of Lending Club Company from 2007 to 2015, which is public available at <https://www.kaggle.com/wendykan/lending-club-data>. Herzenstein *et al.* (2008) demonstrated a significant correlation between the borrower’s housing property and interest rates. It can be reasonably anticipated that the total loan amount and repayment period are correlated with borrower’s housing property, while the interest rate is correlated with both

loan amount and repayment period. Therefore, it is interesting to investigate whether loan amount and repayment period are along the causal pathways from the borrower’s housing property to the interest rate. After eliminating missing values from the dataset, the resulting sample size is $n = 842,322$.

For the purpose of analysis, the exposure variable X represents the borrower’s housing property, which includes rental properties, owned properties and mortgaged properties. Specifically, we define $X = -1$ for rented or mortgaged properties and $X = 1$ for owned properties. The variable Y represents the interest rate on the loan. It is expected that the status “rent” or “mortgage” should have an opposite effect on Y compared to the status “own”. To characterize this attribute, we represent the binary exposure X as -1 and 1 rather than encoding it as 0 and 1 . The mediator M_1 represents the total amount of the loan (in units of ten thousand dollars). The mediator M_2 represents the repayment period of the loan, where $M_2 = 1$ if the repayment period is 36 months and $M_2 = 2$ if the repayment period is 60 months. Moreover, we designate Z_1 as the borrower’s annual income (in units of ten thousand dollars), while Z_2 denotes the borrower’s years of work experience. Our purpose is to explore whether the two mediators, M_1 and M_2 , have significant mediated effects on the pathways linking housing status (X) to loan interest rates (Y).

We first employ the least squares method to fit the mediation linear models (as shown in Figure 2) with full data:

$$\begin{aligned} Y &= 29.667 - 0.025X + 0.019M_1 + 0.066M_2 - 0.009Z_1 - 3.169Z_2 + \epsilon, \\ M_1 &= 1.745 - 0.009X + 0.043Z_1 - 0.114Z_2 + e_1, \\ M_2 &= 2.082 - 0.009X + 0.006Z_1 - 0.159Z_2 + e_2, \end{aligned}$$

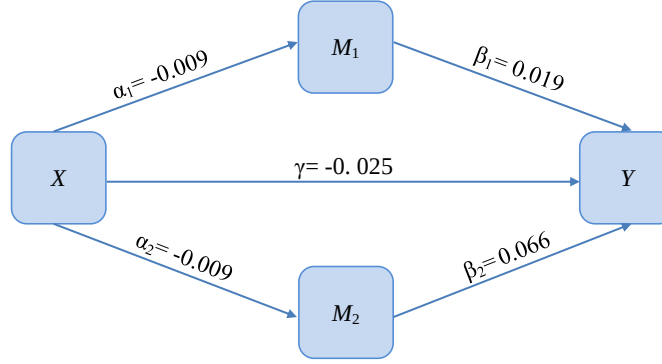


Figure 2: The multiple mediators model in Section 6.1 (omitted confounding variables).

Table 11: The estimated confidence intervals of mediation effects in Section 6.1.

Estimates	Methods	CI _{single}	CI _{adjusted}
$\hat{\alpha}_1\hat{\beta}_1 = -0.00017$	SDB	$[-0.00024, -0.00008]$	$[-0.00026, -0.00006]$
	Bootstrap	$[-0.00024, -0.00011]$	$[-0.00071, -0.00046]$
$\hat{\alpha}_2\hat{\beta}_2 = -0.00059$	SDB	$[-0.00071, -0.00046]$	$[-0.00073, -0.00043]$
	Bootstrap	$[-0.00071, -0.00046]$	$[-0.00073, -0.00045]$

† The meanings of “SDB” and “Bootstrap ” are given in Table 1.

According to Figure 2, after adjusting for other variables, exposure X has a negative direct effect ($\gamma = -0.025$) on the outcome Y . That is to say, when applying for a loan, homeowners would be eligible for a lower interest rate (Herzenstein *et al.*, 2008). Borrowers who possess a residential property exhibit a tendency towards requiring a smaller loan amount ($\alpha_1 = -0.009$) and opting for a shorter repayment period ($\alpha_2 = -0.009$). The interest rate has a positive correlation with borrower’s loan amount ($\beta_1 = 0.019$) and repayment period ($\beta_2 = 0.066$), but it has a negative relationship with both annual incomes and working years.

We apply the proposed SDB approach to estimate confidence intervals for mediation effects, with a subset size of $b = n^{0.8}$ and a confidence level of 0.95. For comparison, we also

Table 12: Estimates, standard errors and p -values in Section 6.1[†].

	$J = 1$	$J = 3$	$J = 5$	$J = 10$
$\hat{\alpha}_1\hat{\beta}_1$	-0.00017	-0.00106	-0.00129	-0.00047
$\hat{\sigma}_{\alpha_1\beta_1}$	0.00003	0.00013	0.00018	0.00018
P_1	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0.01551
$\hat{\alpha}_2\hat{\beta}_2$	-0.00059	-0.00120	-0.00131	-0.00162
$\hat{\sigma}_{\alpha_2\beta_2}$	0.00006	0.00010	0.00012	0.00017
P_2	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

[†] $J = 1$ is the same as a standard analysis without subsamples.

provide the confidence intervals obtained by traditional percentile Bootstrap with full data. Table 11 presents the estimated confidence intervals of mediation effects, indicating that both estimators for mediation effects significantly deviate from zero at a confidence level of 0.95. i.e., both the loan amount (M_1) and repayment period (M_2) can be considered as two significant mediators along the pathways from housing status (X) to loan interest rates (Y).

Furthermore, we employ the proposed divide and conquer method (the Algorithm 2) to estimate indirect effects and conduct multiple tests:

$$H_{0i} : \alpha_i\beta_i = 0 \leftrightarrow H_{1i} : \alpha_i\beta_i \neq 0, \quad i = 1 \text{ and } 2.$$

The significance level is set at 0.05, while the number of blocks is selected as $J = 1, 3, 5$ and 10 respectively. Note that the scenario with $J = 1$ is the same as a standard analysis without subsamples. Table 12 presents the estimates, standard errors, and p -values of indirect effects across various J values. As J increases, the standard errors are observed to grow larger overall. This phenomenon provides evidence for the trade-off between statistical efficiency and computational speed inherent in divide-and-conquer algorithms.

6.2 Binary Outcomes

In this section, we will employ our proposed two methods for analyzing a large dataset about the loan data of Lending Club from 2007 to 2015, which is publicly available at

<https://www.kaggle.com/wendykan/lending-club-data>. According to Kumar (2007), the borrower's annual income is correlated with the total loan amount, while the loan status of the borrower is also associated with the total loan amount. In particular, borrowers with a higher loan amount are at an increased risk of defaulting. Additionally, it is anticipated that the borrower's annual income will have a correlation with the loan's interest rate, and those with higher rates may struggle to meet repayment deadlines (Serrano-Cinca *et al.*, 2015; Emekter *et al.*, 2015). A relevant question is whether loan amount and interest rate mediate the impact of annual income on default status. To achieve this objective, we define exposure X as the borrower's annual income (in ten thousand dollars), and the binary outcome Y represents the loan status of the borrower, where $Y = 1$ indicates default and $Y = 0$ indicates timely repayment. The mediator variables M_1 and M_2 represent the total loan amount (in ten thousand dollars) and the interest rate of the loan, respectively. We set the repayment period as a covariate Z_1 , where $Z_1 = 1$ for a period of 36 months and $Z_1 = 2$ for a period of 60 months. For the purpose of analysis, we have standardized the three continuous variables X , M_1 and M_2 with mean zero and unit variance, respectively. After removing missed data, the full data sample size is $n = 825,994$.

First we estimate the parameters in models (2.3) and (2.4) with full data, which are presented as follows (see Figure 3):

$$\begin{aligned} P(Y_i = 1) &= \frac{\exp\{-4.260 - 0.367X + 0.044M_1 + 2.225M_2 - 0.323Z_1\}}{1 + \exp\{-4.260 - 0.367X + 0.044M_1 + 2.225M_2 - 0.323Z_1\}}, \\ M_1 &= 0.146 + 0.234X + 0.418Z_1 + e_1, \\ M_2 &= 0.594 - 0.047X + 0.299Z_1 + e_2. \end{aligned}$$

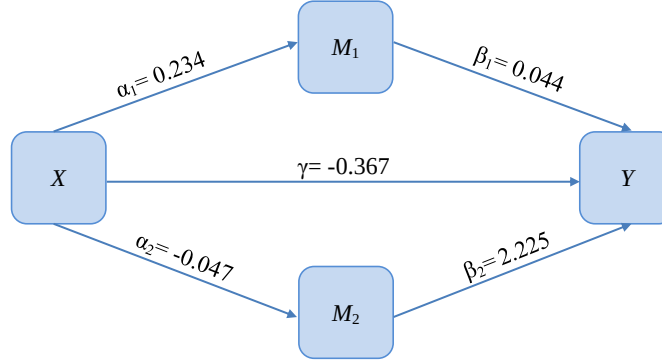


Figure 3: The multiple mediators model in Section 6.2 (omitted confounding variables).

In view of Figure 3, the exposure X has a negative direct effect ($\gamma = -0.367$) on the outcome Y . Although the interest rate may be lower for a borrower with a high annual income ($\alpha_2 = -0.047$), an applicant with higher income possesses the confidence and capacity to borrow larger amounts ($\alpha_1 = 0.234$). Essentially, the risk of default has a significant positive correlation with the loan amount ($\beta_1 = 0.044$) and the interest rate ($\beta_2 = 2.225$). That is to say, if the loan amount and interest rate are higher, it may lead to borrowers defaulting on the loan.

Table 13: The estimated confidence intervals of mediation effects in Section 6.2.

Estimates	Methods	CI _{single}	CI _{adjusted}
$\hat{\alpha}_1\hat{\beta}_1 = 0.01035$	SDB	[0.00363, 0.01659]	[0.00204, 0.01753]
	Bootstrap	[0.00392, 0.01753]	[0.00309, 0.01853]
$\hat{\alpha}_2\hat{\beta}_2 = -0.10453$	SDB	[-0.11567, -0.09191]	[-0.11701, -0.08884]
	Bootstrap	[-0.12216, -0.09225]	[-0.12434, -0.09114]

† The meanings of “SDB” and “Bootstrap ” are given in Table 1.

In Table 13, we implement the proposed SDB method to estimate confidence intervals for mediation effects. The subset size is set at $b = n^{0.8}$ and the confidence level is 0.95. In

Table 14, we employ the proposed divide and conquer approach (Algorithm 2) to estimate indirect effects and conduct multiple tests:

$$H_{0i} : \alpha_i \beta_i = 0 \leftrightarrow H_{1i} : \alpha_i \beta_i \neq 0, \quad i = 1 \text{ and } 2.$$

The significance level is set at 0.05, and the number of blocks is selected as $J = 1, 3, 5$ and 10, respectively. Similar conclusions to those presented in Tables 11 and 12 can be drawn from our findings here.

Table 14: Estimates, standard errors and p -values in Section 6.2[†].

	$J = 1$	$J = 3$	$J = 5$	$J = 10$
$\hat{\alpha}_1 \hat{\beta}_1$	0.01035	0.01026	0.01012	0.01138
$\hat{\sigma}_{\alpha_1 \beta_1}$	0.00264	0.00275	0.00273	0.00288
P_1	0.00018	0.00038	0.00043	0.00016
$\hat{\alpha}_2 \hat{\beta}_2$	-0.10453	-0.10716	-0.10682	-0.11106
$\hat{\sigma}_{\alpha_2 \beta_2}$	0.00128	0.00131	0.00131	0.00135
P_2	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$

[†] $J = 1$ is the same as a standard analysis without subsamples.

7 Concluding Remarks

In this paper, we have investigated the statistical inference of mediation effects using large-scale datasets. First we proposed a SDB-based algorithm to construct confidence intervals of mediation effects. Although we used $b = n^{0.7}$ for the SDB algorithm in the simulation, there are some other possible choices for the subset sizes (e.g., $b = n^{0.6}$ or $b = n^{0.8}$) according to the available computing resources at hand. Then we presented a divide-and-conquer algorithm to estimate and test the mediation effects. Simulations and two real-world examples were provided to demonstrate the usefulness of our proposed methods in practical applications. As pointed out by the reviewer that we have considered the scenario of causally independent mediators for both SDB and divide-and-conquer algorithms. When the mediators are

uncausally correlated, the product of individual parameters fails to fully capture individual causal indirect effects (Wang *et al.*, 2013; Jerolon *et al.*, 2021).

There exist several topics to be studied in the future. First, the effect size measures is an important topic for mediation analysis (Preacher and Kelley, 2011), which is out of the scope for this manuscript. It is interesting to consider the statistical inference for effect size measures in the context of massive data. Second, the mediation analysis with exposure-mediator interaction is an important topic, such as Valeri and VanderWeele (2013) and Rijnhart *et al.* (2021). This manuscript is mainly focused on fast calculation for mediation effects without interactions between exposure and mediators. It is desirable to extend the proposed method to the situation that there exists exposure-mediator interactions with massive data. Third, it is commonly required strong unconfoundedness assumptions for the identification of direct and indirect effects in mediation analysis. i.e., conventional mediation methods assume all confounders can be measured, which is often unverifiable in the case of large datasets. It is interesting to investigate how to perform mediation analysis for massive dataset with hidden confounders (Cheng *et al.*, 2022). Fourth, when the mediators are uncausally related, the joint indirect effect could be of main interest instead of the individual indirect effect (Wang *et al.*, 2013; Jerolon *et al.*, 2021). In this case, Sobel’s multivariate delta method needs to take into account the covariance among the mediators (Sobel, 1982). Under the framework of massive data, it is interesting to consider how to perform valid inference for the joint indirect effect with uncausally related mediators.

Acknowledgement

The authors would like to thank the Editor, the Associate Editor and the reviewers for their constructive and insightful comments that greatly improved the manuscript.

Appendix

In this Appendix, we give the proof details of Theorem 2. Note that VanderWeele and Vansteelandt (2010) has provided the expressions of NDE^{OR} , NIE^{OR} and TE^{OR} for logistic

mediation model with one mediator. Taking the log scale on (2.5), we have

$$\begin{aligned}\log(\text{NDE}^{\text{OR}}) &= \log \left\{ \frac{P\{Y(x, \mathbf{M}(x^*)) = 1\} / [1 - P\{Y(x, \mathbf{M}(x^*)) = 1\}]}{P\{Y(x^*, \mathbf{M}(x^*)) = 1\} / [1 - P\{Y(x^*, \mathbf{M}(x^*)) = 1\}]} \right\} \\ &= \text{logit}(P\{Y(x, \mathbf{M}(x^*)) = 1\}) - \text{logit}(P\{Y(x^*, \mathbf{M}(x^*)) = 1\}),\end{aligned}\quad (\text{A.1})$$

where $\text{logit}(p) = \log(\frac{p}{1-p})$ for $p \in (0, 1)$. Under the assumptions (C.1)-(C.4) and the outcome is rare, we get that

$$\begin{aligned}\text{logit}(P\{Y(x, \mathbf{M}(x^*)) = 1\}) &\approx \log(P\{Y(x, \mathbf{M}(x^*)) = 1\}) \\ &= \log \left(\int_{\mathbf{m}} P\{Y(x, \mathbf{m}) = 1 | \mathbf{Z}, \mathbf{M}(x^*) = \mathbf{m}\} P\{\mathbf{M}(x^*) = \mathbf{m}\} d\mathbf{m} \right) \\ &= \log \left(\int_{\mathbf{m}} P\{Y(x, \mathbf{m}) = 1 | \mathbf{Z}\} P\{\mathbf{M}(x^*) = \mathbf{m} | \mathbf{Z}\} d\mathbf{m} \right) \\ &= \log \left(\int_{\mathbf{m}} P\{Y = 1 | x, \mathbf{m}, \mathbf{Z}\} P\{\mathbf{M} = \mathbf{m} | x^*, \mathbf{Z}\} d\mathbf{m} \right) \\ &\approx \log \left(\int_{\mathbf{m}} \exp\{c + \gamma x + \boldsymbol{\beta}'\mathbf{m} + \boldsymbol{\theta}'\mathbf{Z}\} P\{\mathbf{M} = \mathbf{m} | x^*, \mathbf{Z}\} d\mathbf{m} \right) \\ &= \log \left(\exp\{c + \gamma x + \boldsymbol{\theta}'\mathbf{Z}\} \int_{\mathbf{m}} \exp\{\boldsymbol{\beta}'\mathbf{m}\} P\{\mathbf{M} = \mathbf{m} | x^*, \mathbf{Z}\} d\mathbf{m} \right) \\ &= \log \left(\exp\{c + \gamma x + \boldsymbol{\theta}'\mathbf{Z}\} E[e^{\boldsymbol{\beta}'\mathbf{M}} | x^*, \mathbf{Z}] \right) \\ &= \log \left(\exp\{c + \gamma x + \boldsymbol{\theta}'\mathbf{Z}\} \exp \left\{ \boldsymbol{\beta}'\tilde{\mathbf{c}} + \boldsymbol{\alpha}'\boldsymbol{\beta}x^* + \sum_{k=1}^d \beta_k \boldsymbol{\eta}'_k \mathbf{Z} + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_e \boldsymbol{\beta} \right\} \right) \\ &= c + \gamma x + \boldsymbol{\theta}'\mathbf{Z} + \boldsymbol{\beta}'\tilde{\mathbf{c}} + \boldsymbol{\alpha}'\boldsymbol{\beta}x^* + \sum_{k=1}^d \beta_k \boldsymbol{\eta}'_k \mathbf{Z} + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_e \boldsymbol{\beta},\end{aligned}\quad (\text{A.2})$$

where $\tilde{\mathbf{c}} = (c_1, \dots, c_d)'$, and $\boldsymbol{\Sigma}_e$ is the covariance matrix of mean-zero normal vector $\mathbf{e} = (e_1, \dots, e_d)'$ in (2.4). Similarly, we can derive that

$$\begin{aligned}\text{logit}(P\{Y(x^*, \mathbf{M}(x^*)) = 1 | \mathbf{Z}\}) \\ = c + \gamma x^* + \boldsymbol{\theta}'\mathbf{Z} + \boldsymbol{\beta}'\tilde{\mathbf{c}} + \boldsymbol{\alpha}'\boldsymbol{\beta}x^* + \sum_{k=1}^d \beta_k \boldsymbol{\eta}'_k \mathbf{Z} + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_e \boldsymbol{\beta},\end{aligned}\quad (\text{A.3})$$

and

$$\begin{aligned}\text{logit}(P\{Y(x, \mathbf{M}(x)) = 1 | \mathbf{Z}\}) \\ = c + \gamma x + \boldsymbol{\theta}'\mathbf{Z} + \boldsymbol{\beta}'\tilde{\mathbf{c}} + \boldsymbol{\alpha}'\boldsymbol{\beta}x + \sum_{k=1}^d \beta_k \boldsymbol{\eta}'_k \mathbf{Z} + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_e \boldsymbol{\beta}.\end{aligned}\quad (\text{A.4})$$

Based on (A.1), (A.2) and (A.3), the calculation of the following expression is straightforward,

$$\log(\text{NDE}^{\text{OR}}) = \gamma(x - x^*).$$

Therefore, by taking exponential arithmetic we have

$$\text{NDE}^{\text{OR}} = \exp\{\gamma(x - x^*)\}.$$

In addition, (A.2), (A.3) and (A.4) lead to that

$$\begin{aligned} \log(\text{NIE}^{\text{OR}}) &= \log \left\{ \frac{P\{Y(x, \mathbf{M}(x)) = 1 | \mathbf{Z}\} / [1 - P\{Y(x, \mathbf{M}(x)) = 1 | \mathbf{Z}\}]}{P\{Y(x, \mathbf{M}(x^*)) = 1 | \mathbf{Z}\} / [1 - P\{Y(x, \mathbf{M}(x^*)) = 1 | \mathbf{Z}\}]} \right\} \quad (\text{A.5}) \\ &= \text{logit}(P\{Y(x, \mathbf{M}(x)) = 1 | \mathbf{Z}\}) - \text{logit}(P\{Y(x, \mathbf{M}(x^*)) = 1 | \mathbf{Z}\}) \\ &= \boldsymbol{\alpha}'\boldsymbol{\beta}(x - x^*), \end{aligned}$$

and

$$\begin{aligned} \log(\text{TE}^{\text{OR}}) &= \log \left\{ \frac{P\{Y(x, \mathbf{M}(x)) = 1 | \mathbf{Z}\} / [1 - P\{Y(x, \mathbf{M}(x)) = 1 | \mathbf{Z}\}]}{P\{Y(x^*, \mathbf{M}(x^*)) = 1 | \mathbf{Z}\} / [1 - P\{Y(x^*, \mathbf{M}(x^*)) = 1 | \mathbf{Z}\}]} \right\} \quad (\text{A.6}) \\ &= \text{logit}(P\{Y(x, \mathbf{M}(x)) = 1 | \mathbf{Z}\}) - \text{logit}(P\{Y(x^*, \mathbf{M}(x^*)) = 1 | \mathbf{Z}\}) \\ &= (\gamma + \boldsymbol{\alpha}'\boldsymbol{\beta})(x - x^*). \end{aligned}$$

Taking exponential arithmetic for both (A.5) and (A.6), we can get $\text{NIE}^{\text{OR}} = \exp\{\boldsymbol{\alpha}'\boldsymbol{\beta}(x - x^*)\}$ and $\text{TE}^{\text{OR}} = \exp\{(\gamma + \boldsymbol{\alpha}'\boldsymbol{\beta})(x - x^*)\}$. This ends the proof.

References

- Alt, F. and Spruill, C. (1977). A comparison of confidence intervals generated by the scheffé and bonferroni methods. *Communications in Statistics-Theory and Methods* **6**, 1503–1510.
- Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**, 6, 1173.
- Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics* **46**, 3, 1352–1382.

- Biesanz, J., Falk, C., and Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research* **45**, 661–701.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology* **31**, 144–152.
- Che, C., Jin, I., and Zhang, Z. (2021). Network mediation analysis using model-based eigenvalue decomposition. *Structural Equation Modeling* **28**, 148–161.
- Chen, X. and Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **24**, 4, 1655–1684.
- Cheng, L., Guo, R., and Liu, H. (2022). Causal mediation analysis with hidden confounders. *WSDM’22: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* 113–122.
- Coffman, D. L. (2011). Estimating causal effects in mediation analysis using propensity scores. *Structural Equation Modeling* **18**, 357–369.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26.
- Emekter, R., Tu, Y., Jirasakuldech, B., and Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics* **47**, 1, 54–70.
- Gunzler, D., Tang, W., Lu, N., Wu, P., and Tu, X. (2014). A class of distribution-free models for longitudinal mediation analysis. *Psychometrika* **79**, 543–568.
- Herzenstein, M., Andrews, R. L., Dholakia, U. M., Lyandres, E., *et al.* (2008). The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities. *Boston University School of Management Research Paper* **14**, 6, 1–36.
- Hou, L., Yu, Y., Sun, X., Liu, X., Yu, Y., Li, H., and Xue, F. (2022). Causal mediation analysis with multiple causally non-ordered and ordered mediators based on summarized genetic data. *Statistical Methods in Medical Research* **31**, 1263–1279.

- Jerolon, A., Baglietto, L., Birmele, E., Alarcon, F., and Perduca, V. (2021). Causal mediation analysis in presence of multiple mediators uncausally related. *The International Journal of Biostatistics* **17**, 191–221.
- Jo, B., Stuart, E. A., MacKinnon, D. P., and Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research* **46**, 425–452.
- Kisbu-Sakarya, Y., MacKinnon, D. P., and Miočević, M. (2014). The distribution of the product explains normal theory mediation confidence interval estimation. *Multivariate Behavioral Research* **49**, 3, 261–268.
- Kumar, S. (2007). Bank of one: Empirical analysis of peer-to-peer financial marketplaces. *AMCIS 2007 Proceedings* **305**.
- Liu, H., Jin, I., Zhang, Z., and Yuan, Y. (2021). Social network mediation analysis: A latent space approach. *Psychometrika* **86**, 272–298.
- MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology* **58**, 593–614.
- MacKinnon, D. P., Valente, M. J., and Gonzalez, O. (2020). The correspondence between causal and traditional mediation analysis: The link is the mediator by treatment interaction. *Prevention Science* **21**, 147–157.
- Miočević, M. and Golchi, S. (2021). Bayesian mediation analysis with power prior distributions. *Multivariate Behavioral Research* DOI:10.1080/00273171.2021.1935202.
- Miočević, M., Levy, R., and MacKinnon, D. (2021). Different roles of prior distributions in the single mediator model with latent variables. *Multivariate Behavioral Research* **56**, 20–40.
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology* **66**, 825–852.
- Preacher, K. J. and Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods* **16**, 93–115.

- Preacher, K. J. and Selig, J. P. (2012). Advantages of monte carlo confidence intervals for indirect effects. *Communication Methods and Measures* **6**, 2, 77–98.
- Rijnhart, J. J., Valente, M. J., MacKinnon, D. P., Twisk, J. W., and Heymans, M. W. (2021). The use of traditional and causal estimators for mediation models with a binary outcome and exposure-mediator interaction. *Structural Equation Modeling* **28**, 345–355.
- Sengupta, S., Volgushev, S., and Shao, X. (2016). A subsampled double bootstrap for massive data. *Journal of the American Statistical Association* **111**, 515, 1222–1232.
- Serrano-Cinca, C., Gutiérrez-Nieto, B., and López-Palacios, L. (2015). Determinants of default in P2P lending. *PloS one* **10**, 10, e0139427.
- Shen, E., Chou, C.-P., Pentz, M. A., and Berhane, K. (2014). Quantile mediation models: A comparison of methods for assessing mediation across the outcome distribution. *Multivariate Behavioral Research* **49**, 471–485.
- Shi, C., Lu, W., and Song, R. (2018). A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association* **113**, 524, 1698–1709.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology* **13**, 290–312.
- Soest, T. and Hagtvet, K. (2011). Mediation analysis in a latent growth curve modeling framework. *Structural Equation Modeling* **18**, 289–314.
- Song, Y. (2020). Bayesian methods in high-dimensional sparse mediation analysis. *PhD dissertation, University of Michigan* 1–162.
- Sun, R., Zhou, X., and Song, X. (2021). Bayesian causal mediation analysis with latent mediators and survival outcome. *Structural Equation Modeling* **28**, 778–790.
- Valente, M. and MacKinnon, D. (2017). Comparing models of change to estimate the mediated effect in the pretest-posttest control group design. *Structural Equation Modeling* **24**, 428–450.

- Valeri, L. and VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods* **18**, 137–150.
- Vanderweele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface* **2**, 457–468.
- VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology* **172**, 1339–1348.
- Volgushev, S., Chao, S.-K., and Cheng, G. (2019). Distributed inference for quantile regression processes. *The Annals of Statistics* **47**, 3, 1634–1662.
- Wang, L. and Zhang, Z. (2011). Estimating and testing mediation effects with censored data. *Structural Equation Modeling* **18**, 18–34.
- Wang, W., Nelson, S., and Albert, J. M. (2013). Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Statistics in Medicine* **32**, 4211–4228.
- Zhang, H., Chen, J., Feng, Y., Wang, C., Li, H., and Liu, L. (2021a). Mediation effect selection in high-dimensional and compositional microbiome data. *Statistics in Medicine* **40**, 885–896.
- Zhang, H., Hou, L., and Liu, L. (2022). A review of high-dimensional mediation analyses in DNA methylation studies. In Guan, Weihua (Ed.), *Epigenome-Wide Association Studies: Methods and Protocols* DOI: 10.1007/978-1-0716-1994-0.
- Zhang, H., Zheng, Y., Hou, L., Zheng, C., and Liu, L. (2021b). Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics* **37**, 3815–3821.
- Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L., and Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32**, 3150–3154.

Zhang, Z. and Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika* **78**, 154–184.