Joint modeling, variable selection and multiply robust estimation in mediation analysis with multiple mediators

by

Lijia Wang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Geneviève Lefebvre Professeure titulaire, Département de mathématiques, Université du Québec à Montréal (UQAM)
Supervisor(s):	Dr. Yeying Zhu Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo Dr. Richard J. Cook Professor, Department of Statistics and Actuarial Science, University of Waterloo
Internal Member:	Dr. Cecilia Cotton Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo Dr. Liqun Diao Assistant Professor, Department of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Dr. Ashok Chaurasia Associate Professor, School of Public Health Sciences, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis explores topics in causal mediation analysis with multiple possibly related mediators. The goal of this thesis is to propose innovative methodologies for joint modeling of multiple uncausally related mediators, selecting mediators from high-dimensional candidates while simplifying their dependency structures and performing multiply robust estimations to uncover causal effects of interest.

Causal mediation analysis aims to enhance understanding of the effects of an exposure on an outcome by examining direct and indirect effects. In settings where multiple mediators are involved, the relations among these mediators play an important role. Traditional studies focus on the scenario that the multiple mediators are either related under specified causal structures or independent given baseline covariates. Our studies focus on multiple uncausally related mediators, where the mediators are associated with each other conditioning on pre-treatment covariates and treatment but there is no causal ordering among them.

In Chapter 2, we begin by reviewing and expanding upon the concept of mediators that are uncausally related, followed by the introduction of causal effects defined under such settings and the associated identification assumptions. We propose to jointly model the uncausally related mediators using copula functions. An important advantage of employing copula functions in joint modeling is the significant flexibility it offers, as this method allows for multiple mediators to have different distributions and be correlated in various ways. Subsequently, we propose methods estimating causal effects within this framework.

In Chapter 3, we center our attention on the sparse mediation phenomenon, where only a handful of true mediators, from a pool of possibly high-dimensional candidates, exhibit nonzero indirect effects. We propose a LASSO-based penalization technique that selects the true mediators by considering their indirect effects. Acknowledging that the selected mediators often still exhibit complex dependency structures even after selection, our method also simplifies these structures by selecting non-zero correlation entries within the correlation matrix using a similar penalized estimation technique. To facilitate the correlation structure selection, we transform the correlation matrix selection problem into a standard variable selection problem within the framework of a linear model. Moreover, our proposed method allows the mediator selection and the dependency structure selection processes, to be conducted either via either a parallel or a sequential approach. The grouped and individual causal effects are defined under such settings with estimation approaches discussed.

In Chapter 4, we discuss the issue of model misspecification within the context of causal mediation analysis. Following the discussion, we propose two ways of constructing multiply

robust estimators. In causal mediation analysis, typically three working models must be specified: the treatment model, the mediator model, and the response model. Both of our multiply robust estimation methods yield consistent estimation of the causal quantities of interest, provided that any two out of the three models are correctly specified.

For each proposed method introduced in Chapters 2, 3 and 4, we provide theoretical results with proofs of the consistency and other properties. We also derive large sample properties and investigate finite sample properties via simulations. Each chapter includes an application of the proposed method to a genetic study in psychiatry to investigate DNA methylation loci as mediators on the causal path between childhood trauma and stress reactivity. In Chapter 2, the proposed method estimates the mediation effects of three DNA loci on the Kit ligand gene. Chapter 3 extends this analysis and applies the proposed mediator selection method to the entire DNA methylation dataset, revealing 12 mediating loci, with 10 showing a strong association. We estimate the grouped indirect effect from them and the individual effects of the remaining two loci. In Chapter 4, we employ our multiply robust estimation methods to re-evaluate the mediation effects of these 12 loci, demonstrating enhanced robustness to previous findings.

Acknowledgements

I would like to extend my heartfelt gratitude to everyone who helped me make this thesis possible. First and foremost, I want to express my deepest appreciation to my esteemed supervisors, Dr. Yeying Zhu and Dr. Richard J. Cook. Undertaking my Ph.D. studies under your supervision has been the most enriching and fulfilling four years of my life. Whenever I encountered challenges, you were always by my side, ever ready to offer your utmost selfless help. I shall always remember the moment when Professor Zhu provided me with the most crucial instructions I needed, even when she had just recovered from COVID and her voice was still weakened. I will never forget the time when Professor Cook persistently met with me and helped me solve one urgent problem after another, all while dealing with his injured shoulder. I am deeply grateful for your unwavering commitment to my academic growth, and I am profoundly thankful for the countless hours you've devoted to nurturing my personal development. Your guidance has been instrumental in shaping the person I have become.

I would also like to extend my gratitude to the committee members, Dr. Cecilia Cotton, Dr. Liqun Diao, Dr. Ashok Chaurasia, and Dr. Geneviève Lefebvre, for your valuable suggestions that have contributed to polishing my thesis to its best shape. I am also thankful to all the professors in our department who imparted the knowledge necessary for the successful completion of this thesis and my doctoral studies. I am also thankful to my colleagues, classmates, and friends, including Yuliang Shi, Yiran Wang, Yunfeng Yang, Jie Jian, Xianwei Li, Edward Yu, and others, who aided me in overcoming challenges throughout my research and the completion of my thesis.

Finally, I would like to express my appreciation to my parents Weihong Li and Qiusheng Sun, and my grandparents You Li and Cuiling Zhang. Without their tremendous understanding and encouragement over the past few years, it would be impossible for me to complete my studies. I wish to particularly express my gratitude to my grandparents, who gave me all their love. Despite their limited education, their invaluable support has been the cornerstone of my success. In honor of their love and respect for them, I would like to convey the following sentences in Chinese: "亲爱的爸爸妈妈，尊敬的姥姥姥爷，礼佳谨以此毕业论文表达我最诚挚的感激之情。感谢您们一直以来对我的无微不至的关怀和支持。正是来自于您们无尽的爱，鼓舞着我完成了博士学业，也激励了我完成这份博士毕业论文的撰写。" (in English: Dear Mom and Dad, respected Grandparents, I extend my deepest gratitude to you through this graduation thesis. I am sincerely thankful for your unwavering care and support throughout my journey. It is your boundless love that has inspired me to accomplish my doctoral studies and has motivated me to complete this doctoral thesis.)

Dedication

This thesis is dedicated to my parents and my grandparents, who offered me your boundless love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General introduction of causal inference and the potential outcomes framework

Causal analysis aims to identify and evaluate causal effects resulting from a treatment (in an experiment) or exposure (in an observational study) on the outcome of interest. The concept of causality is believed by many philosophers to be metaphysical [127, 82, 18]. However, in contemporary statistical studies, causal effects are defined in terms of counterfactual relations. To begin with, Robin et al. [106] state that "the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1". For example, an interesting research topic in the field of social sciences is examining the impact of college education on individuals' income. A related research question in the domain of causal inference is to evaluate the causal effect of attending college on a person's income at the age of 40. According to the definition, such a causal effect can be calculated by comparing the income that an individual might earn at age 40 if they had attended college versus if they had not attended. Throughout this thesis, we explore causal inference and causal mediation analysis, applicable to both experimental and observational studies. Therefore, unless otherwise specified, the terms "treatment" and "exposure" are used interchangeably to refer to the assignment of study groups (either treatment or control) for each subject.

As per the definition, the different possible outcomes under varying treatment assignments, such as the two potential income levels at age 40, play a pivotal role in defining

the causal effect. These alternative outcomes are defined by Rubin et al. [106] as the potential outcomes. Moreover, such definitions of the causal effect are established within the potential outcomes framework, which is the cornerstone of modern causal inference studies [106, 46]. The preliminary concept of the potential outcomes framework was first introduced by Jerzy Neyman in his Master's thesis in 1923 [87], though such introduction was albeit in the context of completely randomized experiments [112]. Rubin et al. [106] later expanded this into a comprehensive framework for understanding causation in both observational and experimental studies [115]. Consequently, the potential outcomes framework, along with its associated models, is often referred to as the Rubin Causal Model (RCM). However, in real life, it is impossible to observe both potential outcomes at once: one of the potential outcomes is always unobservable. This unobserved outcome is sometimes referred to as the counterfactual outcome, in contrast to the observed (factual) one. This is why the potential outcomes framework is also known as the counterfactual framework, particularly when emphasizing the contrast between observed and unobserved outcomes. This dilemma that "one of the potential outcomes is always unobservable" is often denoted as the "fundamental problem of causal inference" [46].

Another essential challenge in causal analysis is the confounding issue. A confounding issue refers to a situation where the observed relationship between the treatment and the outcome may not be a direct cause-and-effect relationship but is instead influenced by a third variable [139]. The third variable that affects both the exposure and the outcome is called confounder. The existence of the confounder can lead to a misleading or incorrect conclusion about the true causal relationship between the treatment and the outcome. In other words, the confounding variable can create the appearance of causality when there may not be a true causal connection. Moreover, for most studies, there usually exists a set of confounding variables or even a high-dimensional set of such variables, which imposes extra difficulties in eliminating the confounding issue.

A straightforward approach to address the confounding issue is the regression approach, which is to include confounders as predictors in the regression model expressing the effect of treatment on the outcome. However, this approach has a significant drawback: it often proves challenging to accurately specify the underlying regression models. The relationship between the confounders and the outcome is hard to determine, particularly if there exist multiple, potentially high-dimensional, confounders. An alternative strategy to tackle this challenge is to utilize methods based on the propensity score, which transforms the problem of modeling the relationship between the covariates and the outcome to modeling the treatment assignment conditioning on covariates. Such methods are commonly employed in observational research, especially within the scope of causal inference. The propensity score is a single numerical value assigned to each individual in a study, representing their

likelihood or probability of receiving the treatment. Furthermore, Rosenbaum et al. [102] show that the propensity score is a balancing score such that the conditional distribution of covariates given the propensity score is the same for the treated and the control. Therefore, by conditioning on the propensity score, it is assumed that observations from different arms of the study, whether the treatment or control group, have an equal likelihood of receiving the treatment assignment. Consequently, comparing subjects with similar propensity score values across different arms is akin to conducting comparisons between subjects that had been randomly assigned to each arm. In essence, this mimics the process of analyzing data from randomized trials, where each subject's assignment to either the treatment or control arm is random, which effectively eliminates bias induced by confounding.

Some typical ways of incorporating the propensity score in causal analysis include inverse probability weighting (IPW) [57, 44], stratification [102], and matching [104, 105]. Particularly, the IPW method is used widely due to its favorable mathematical properties. Specifically, it employs a smoothing weighting approach to maintain the continuous nature of propensity values. Propensity values are typically assessed on continuous scales, and IPW assigns unique continuous-scale weights to each subject, preserving this continuity. In contrast, methods like matching or stratification, either match or group subjects based on similar propensity values. Since finding two subjects with exactly the same propensity scores is often impractical, applying those methods leads to the continuous propensity values being transformed into categorical scales during the process. Additionally, the IPW method is relatively easy to implement. Such an approach works by assigning each observed data point a weight that is proportional to the inverse of the probability of receiving its observed treatment assignment, so that after the process, we are treating each data point as if they were assigned to each arm (treatment or control) randomly. Therefore, calculating the average causal effect with inverse propensity weighting can eliminate the bias due to confounding between the treatment assignment and the outcome.

In mathematical notation, we denote the binary exposure or treatment indicator as $T = t \in \{0, 1\}$ and the outcome as Y. The treatment T may affect Y directly and/or through the mediator M. We denote $Y(t)$ as the potential value of the variable Y observed under the treatment t. $E\{Y(t)\}$ is used to represent the expected potential outcome, where the expectation is taken with respect to the distribution of the potential value in the population [42]. Then according to the definition, the average causal effect is $E\{Y(1)\} - E\{Y(0)\}$. Additionally, $\boldsymbol{X}$ is used to denote the covariates.

1.2 General introduction of causal mediation analysis

On top of the aforementioned causal effects, researchers in most fields of studies are often not only interested in the impact of a treatment or exposure on an outcome, but also the underlying mechanism involved within the process [28]. Furthermore, in many cases, there exist intermediate variables positioned along the causal pathway between the treatment and the outcome. When specific conditions are met, these intermediate variables are referred as mediators [54]. An understanding of the relationship among treatment, mediator(s), and outcome provides insights into how the treatment precisely influences the outcome. Causal mediation analysis offers an essential tool for disentangling the effects of a treatment on an outcome via a variety of paths through either the mediator(s) or the treatment itself [13, 136].

Though the rigorous definition of causal mediation analysis is a relatively recent concept in statistics or biostatistics, the initial analysis of intermediate variables, which is akin to modern mediation analysis, has a long history predating the modern concepts of causal inference and the potential outcomes framework. Such analysis has been employed in various fields, primarily in early studies on psychology and other social sciences [151, 152, 153]. Baron and Kenny [4], along with some other scholars[64, 59], contributed some initial work on mediation analysis. Baron and Kenny [4] were the first to clarify the criteria for considering a variable a mediator. Then, they introduced the mediation effects in terms of a series of regression coefficients and provided methods for estimating and testing the effects. To illustrate their models, we consider the simplest mediation scenario as presented by Baron and Kenny. In this scenario, there is a binary exposure or treatment T, a continuous mediator M, and a continuous outcome Y. The exposure may affect the outcome directly and/or through the mediator. This mediation scenario can be represented graphically using a causal diagram (see Figure 1.1). Such a figure is also called the directed acyclic graph (DAG) [90].

Based on the mediation problem, they propose several models, which are,

$$
\begin{aligned}
Y &= \beta_1 + \tau T + \varepsilon_1 \\
Y &= \beta_2 + \tau' T + \beta M + \varepsilon_2 \\
M &= \beta_3 + \alpha T + \varepsilon_3,
\end{aligned}
$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$, for $i = 1, 2, 3$. Following their proposed models, they identify the total effect as τ, the direct effect as τ', and the indirect effect (mediation effect) as $\alpha\beta$. Notice that when Baron and Kenny propose their model, they do not utilize the potential outcomes framework; instead, they purely define the mediation effects as functions of

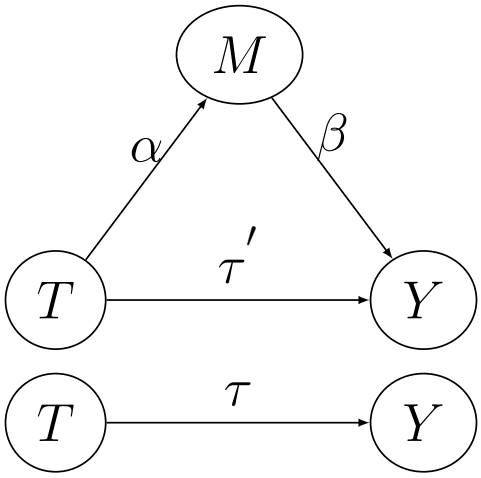


Figure 1.1: The causal diagram of Baron and Kenny's model

regression coefficients. In addition, they do not discuss causal effects, where confounding issues may exist in real applications.

James et al. [59], Judd et al. [64] and MacKinnon et al. [81] also make contributions on either methodologies or applications of mediation analysis using regression-based approaches. The traditional regression-based mediation effect models are summarized as the structural equation modeling (SEM) framework, according to Imai et al. [54], Gunzler et al. [37] and other scholars [149, 19].

The conceptualization of the potential outcomes framework greatly influences studies on mediation analysis and modern mediation analysis is based on that. Several studies define the causal mediation effects in more rigorous ways. We introduce some additional notations. The notation $M(t)$ is used to denote the potential value of mediator M under treatment assignment t and $Y(t, m)$ is used to denote the potential value of the outcome Y when the associated treatment is assigned to the level of t and the mediator has taken value m. Following Pearl[91], the controlled indirect effect $CIE(t, m, m') = E\{Y(t, m)\} - E\{Y(t, m')\}$, which depicts the changes of the potential outcome if controlling the treatment to the level t and changing the mediator value from m to m'. Similarly, the controlled direct effect $CDE(t, t', m) = E\{Y(t, m)\} - E\{Y(t', m)\}$, which depicts the changes of the potential outcome if setting the mediator to the level of m and changing the treatment level from t' to t. However, compared with the controlled effect, another definition called the natural effect is more widely used. Under the natural effect setting, the nested potential outcome $Y(t, M(t'))$ is used to denote the potential value of the outcome Y if the associated treatment is assigned t and the mediator takes the value that it should

have taken had the treatment affecting the mediator been assigned t'. Then the natural indirect effect $NIE(t)$ is defined as $E\{Y(t, M(t))\} - E\{Y(t, M(t'))\}$, which represents the changes of the expected outcome if the treatment assignment is assigned t and the value of the mediator changes from the value it should have taken under treatment t to that under t'. The natural direct effect $NDE(t)$ is defined as $E\{Y(t, M(t))\} - E\{Y(t', M(t))\}$, which represents the changes of the expected outcome if the treatment assignment is changed from t to t' and the value of the mediator fixed at the value it should have taken under treatment t. Because the natural effects are more commonly used, people sometimes simplify the natural indirect effect (NIE) as indirect effect (IE) and the natural direct effect (NDE) as direct effect (DE).

In terms of the estimation, Imai et al. [54], in their groundbreaking work on causal mediation analysis, argue that the SEM framework is problematic. Instead, in a series of studies [55, 54, 56], based on the new counterfactual thinking, Imai et al. [54, 55, 56] propose a unified framework for the definition, identification, estimation, application, as well as sensitivity analysis of causal mediation effects. Meanwhile, Pearl et al. [92, 91, 93] propose the mediation formula that non-parametrically estimates causal mediation effects. Other scholars also contribute to the identification and estimation of causal mediation effects [137, 1, 163], application of causal mediation analysis [43] as well as performing sensitivity analysis under the mediation settings[138]. In the causal mediation analysis framework, we call the model that associates the mediators to the treatment and baseline covariates the "mediator model", and the model that associates the outcome with the treatment, the mediators and the baseline covariates the "response model". Sometimes, the "response model" is also called the "outcome model" and the two terms are used interchangeably in this thesis.

1.3 Joint modeling of multiple uncausally related mediators

In early studies on mediation analysis, researchers mainly focus on the case of a single mediator[136]. Moving forward to the past two decades, quite a number of papers[136, 16, 124] extend the single-mediator framework to account for multiple mediators. In the presence of multiple mediators, the relationship among the multiple mediators plays an essential role when conducting analysis. However, traditional studies on causal mediation analysis of multiple mediators mostly concentrate on the scenarios in which the mediators are either sequentially causally related [124, 136] (the causal pathways among the mediators can be clearly identified), or independent with each other conditioning on covariates[124].

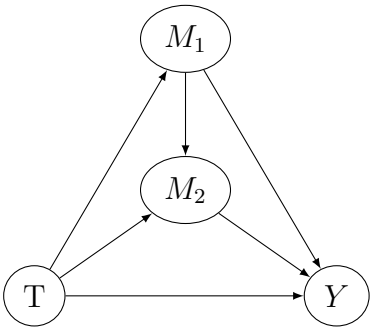


Figure 1.2: A DAG of two sequentially causally related mediators

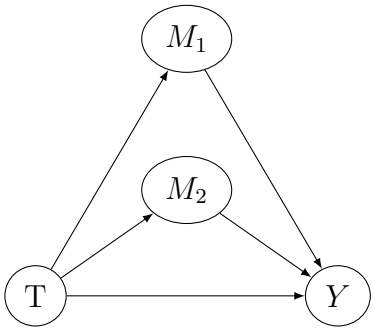


Figure 1.3: A DAG of two causally unrelated mediators

Among them, one of the most well-known estimating frameworks is the study of Tingley et al. [132] that assumes the multiple mediators are sequentially causally related. See Figure 3.3 for a DAG illustrating an example of two mediators sequentially causally related. In the example, M_1 is causally affected by T, while M_2 is causally affected by both M_1 and T.

Recently, a new concept that mediators are uncausally related has been proposed and formalized by Jerolon et al. [60]. Even before the formal definition of "uncausally" related is proposed, there are already some studies on modeling uncausally related mediators, even though these studies do not explicitly state that the multiple mediators are uncausally related. For example, Wang et al. [147] propose modeling bivariate mediators when one is continuous and the other one is discrete; Huang et al. [52] study mediation analysis on survival outcome when multiple mediators are involved and Yu et al. [156, 158] proposes using the multivariate normal framework to jointly model multiple mediators. Also, some applied studies can be found [26, 47, 168, 155] utilizing the idea of uncausally correlated mediators in applications without explicitly formalizing the concept. In short, uncausally correlated mediators depict the scenario that the mediators are conditionally dependent given the treatment and measured covariates, but regarding their dependencies, it is hard to establish causalities [60]. For example, there are no deterministic conclusions regarding whether M_1 causally affects M_2 or vice versa. Such a new concept offers higher practical value, as in many real-life scenarios, drawing the causal relationships among the mediators is often challenging.

In this thesis, we propose a framework for modeling uncausally related mediators based on the work of Jerolon et al. [60]. When modeling multiple mediators, one of the most important issues is modeling the correlation structure, because failing to do so may lead to biased estimations of the causal effects or invalid estimation of standard errors [60, 147].

In our proposed method, we utilize copula functions to model the dependencies among mediators.

There have been pioneer works utilizing copula functions for mediation analysis. The literature on copula-based methodologies in causal inference and mediation analysis encompasses a rich array of significant contributions. Zheng et al. [166] dig into the intricate domain of multi-treatment causal inference, employing copula functions to disentangle complex causal relationships involving multiple interventions. Their work revolutionizes the understanding of multifaceted treatment effects by considering dependencies and interactions among various interventions, enhancing the precision of causal inference in complex settings. Falkenström et al. [23] significantly advance causal inference by focusing on dependent variable modeling. Through the wide use of copula functions, they illuminate the associations between the dependent variable and other factors within a causal framework, providing nuanced insights that surpass traditional linear approaches. Meanwhile, Huang et al. [50] pioneer the application of copula-based methodologies in semi-competing risk modeling within mediation analysis. Their work addresses the challenges posed by multiple risk factors influencing competing events, showcasing copula functions' utility in modeling complex risk relationships and refining the precision of causal inference. In another study, Vanderweele et al. [136] introduce the innovative use of the Plackett copula for handling dichotomous mediators in mediation analysis, offering a robust tool to comprehend and quantify indirect effects in diverse causal mediation models. Together, these studies demonstrate the versatility and efficacy of copula functions across various domains.

A copula function can be used to connect the associated marginal distributions to form the joint distribution. Following [86], we define $C : [0,1]^d \to [0,1]$ a d-dimensional copula if C is a joint cumulative distribution function of a d-dimensional random vector on the unit cube $[0,1]^d$ with uniform marginals. Sklar's theorem [119, 21] further states that every multivariate cumulative distribution function $F_{Z_1,\ldots,Z_d}(z_1,\ldots,z_d) = \Pr[Z_1 \leq z_1,\ldots,Z_d \leq z_d]$ of a random vector $\{Z_1,\ldots,Z_d\}$ can be expressed in terms of its marginal cumulative distribution function $F_{Z_i}(z_i) = \Pr[Z_i \leq z_i]$ and a copula C. Indeed:

$$F_{Z_1,\ldots,Z_d}(z_1,\ldots,z_d) = C(F_{Z_1}(z_1),\ldots,F_{Z_d}(z_d)\,;\boldsymbol{\varrho}) = C(u_1,\ldots,u_d\,;\boldsymbol{\varrho}).$$

In the equation, all the $u_1,\ldots,u_d$ are realizations of probability integral transformations $U_1,\ldots,U_d$ such that

$$U_i = F_{Z_i}(Z_i) \sim \text{Uniform}(0,1),$$

where $\boldsymbol{\varrho}$ is the correlation structure of random variables $Z_1,\ldots,Z_d$. In some literature, $\boldsymbol{\varrho}$ is considered a part of the copula and therefore is omitted in the equation, but here to

emphasize the modeling of the correlation parameters among the variables, we put $\boldsymbol{\varrho}$ into the equation. In case that the multivariate distribution has a density $f_{Z_1,\ldots,Z_d}$, and if this is available, it holds further that

$$f_{Z_1,\ldots,Z_d}(z_1,\ldots,z_d) = c(F_{Z_1}(z_1),\ldots,F_{Z_d}(z_d)\,;\boldsymbol{\varrho})\prod_{i=1}^{d} f_{Z_i}(z_i) = c(u_1,\ldots,u_d\,;\boldsymbol{\varrho})\prod_{i=1}^{d} f_{Z_i}(z_i),$$

where c denotes the density corresponding to the copula C. Some recent studies on copula include Kolve et al. [71], Song et al. [120] and Zimmer et al. [169]. The copula technique has also been applied widely in many areas of research to model the dependencies among multivariate variables, especially in finance, biomedical and social science studies[89, 65, 22].

1.4 Mediator selection and dependency structure simplification for high-dimensional causal mediation analysis

In real-life problems, a common issue we encounter involves high-dimensional mediators. For instance, consider a human neuroimaging study [2] aimed at investigating the relationship between brain regions and pain perception. In this context, brain neural connectors serve as natural mediators, facilitating the transition from a stimulus to pain perception. These neural connectors exhibit high dimensionality, which poses a challenge for analysis. Another example can be found in a study exploring the mediation effect of DNA methylation on the relationship between childhood trauma and the development of long-term psychiatric disorders [69, 49]. DNA methylation is postulated to mediate the process of connecting childhood trauma to psychiatric outcomes. However, DNA methylation occurs at multiple DNA loci, often in large numbers, resulting in high-dimensional mediators. This example also serves as the real data application problem that we are going to address throughout the thesis. A detailed introduction to the study is going to be presented in Section 2.8. Conducting causal mediation analysis on original datasets that involve high-dimensional mediators can lead to estimation problems. The high dimensionality typically results in large standard errors for estimated parameters, leading to inaccurate estimations. Moreover, in some cases, the number of mediators may even exceed the sample size, a situation commonly referred to as the "$p > n$" problem. Under such circumstances, traditional analysis methods become invalid, and issues such as non-invertible design matrices can hinder the fitting of response models.

On the other hand, the high dimensionality of mediators gives rise to the challenge of proper interpretation. In practical scenarios, not all variables classified as 'mediators' truly fulfill that role. In our earlier example of pain perception, the human brain is a complex structure with numerous circuits, yet only a specific brain region mediates the process from stimulus to pain perception. In the case of DNA methylation, it is unlikely that all DNA methylation loci mediate the process, but usually, only a small number of them play the role of mediators. The phenomenon that among the large number of variables that should be regarded as candidate mediators, it is usually the case that only a small number of them contribute non-zero indirect effects from the treatment to the outcome, is called the sparse mediation problem by Zhao et al. [165]. In this thesis, we only refer to the mediators with nonzero indirect effects as true mediators, whereas, we regard the original large set of variables where true mediators are picked from as candidate mediators. For the sake of precise estimation and clearer interpretation, the essential step of mediator selection is crucial before conducting causal analysis. This selection process also results in dimension reduction that benefits estimations and statistical inferences in subsequent steps. The exploration of potential relationships among the exposure, candidate mediators, and the outcome, along with the selection of potential true mediators before causal analysis, constitutes the process of 'exploratory mediation analysis' (EMA), following Van Kesteren et al.[134].

Several methods have been proposed to address the issue. A natural approach involves fitting a series of univariate regression models, where each model associates one mediator with either the treatment or the outcome [7, 78]. Such techniques are commonly categorized as 'filter methods' following Guyon et al. [38]. However, these methods have limitations in that they only consider the marginal effect of each mediator individually, overlooking potential relationships among multiple mediators. In certain scenarios, a group of mediators may jointly exhibit significant mediating effects, even if their individual marginal effects are not significant when considered separately. A more comprehensive approach is to extend the Structural Equation Model (SEM) framework introduced in Section 1.2 and incorporate multivariate regression models to account for multiple mediators (Preacher et al. [96]). In the multivariate regression model proposed by Preacher et al. [96], the multiple mediators are treated as a multivariate vector and the multivariate regression model is therefore fitted to assess the mediation effects. Mediator selection can be addressed by examining the mediators associated with non-zero regression coefficients. Figure (1.4) illustrates the process. In the figure, $M_1 \ldots, M_J$ are candidate mediators and each one is associated with coefficient α linking it with the treatment and coefficient β linking it with the outcome. Selection can be made by investigating if $\alpha_j \beta_j = 0$ for all $j = 1, \ldots, J$. However, such a method also comes with drawbacks. For example, if the dimension of

mediators is higher than the sample size, regression models may not be valid to fit due to the design matrices being non-invertible.

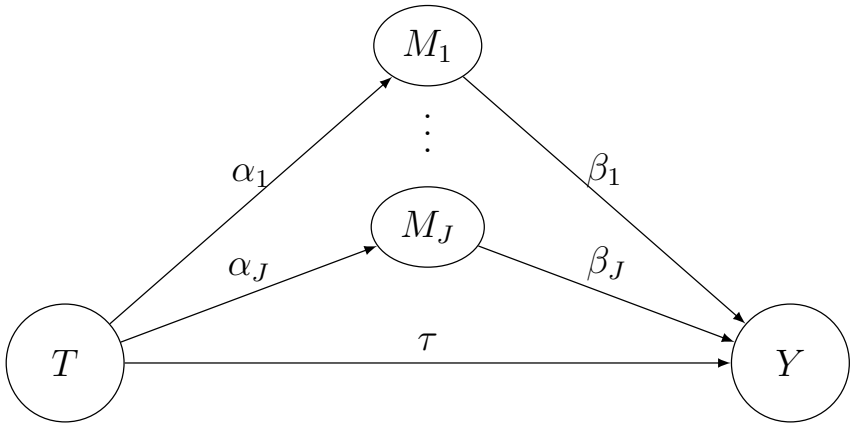


Figure 1.4: An extended SEM framework for mediator selection

Other methods approach the issue from different perspectives. Instead of conducting mediator selection, these methods focus on projecting the original mediators into linear combinations. By transforming mediators from their original high-dimensional space into potential lower-dimensional linear combinations, these approaches effectively reduce the dimension of the mediator space. Huang et al. [51] employ the principal component analysis (PCA) to reduce the dimensionality of the multivariate mediators first and then perform mediation analysis on the reduced linear combinations. Zhao et al. [164] extend the method introduced by Huang et al. [51] by replacing the regular PCA with the sparse PCA technique to enhance estimation efficiency. These PCA-based methods work by projecting the original mediator values to a reduced space, which leads to dimension reduction and efficiency gains. However, when considering the problem under the "sparse mediation" settings, these methods may not be ideal, as they do not inherently facilitate the proper selection of true mediators from the candidate pool.

Some studies propose mediator selection and dimension reduction simultaneously using penalized estimation techniques. Shojaie et al. [117] employs the LASSO technique [130] to penalize effects through each possible mediator candidate connecting the treatment to the outcome in directed acyclic graphs. Their method fits a set of multiple models relating candidate mediators to either the treatment or outcome. Then they perform variable selection for each model via the standard LASSO penalization method. Zhao et al. [165] extend the standard LASSO framework by modifying the penalization term and propose a pathway LASSO method to select mediators with strong connections with

both the treatment and the outcome. Their new penalization terms are constructed by adding the absolute values of the intersections of coefficients ($|\alpha\beta|$ in their paper) and second-ordered terms of α and β to the original first-order ones, where α and β are the regression coefficients associated with the mediator variables in the mediator models and the response model respectively. Our method shares a similar spirit with their work in terms of mediator selection, but we modify the penalization terms for better computation efficiency. We remove the second-order term to make the penalization target the indirect effects only and thus, the penalized objective function is much simpler to solve. We further add a second part of selection to the dependency structures, which is introduced in the next paragraph. Moreover, the proposed method not only picks up true mediators from the set of candidate variables, but also picks up the proper covariates from the remaining variables to be included in the response model.

Other penalization-based variable selection methods include the de-biased LASSO approach proposed by Gao et al.[30], the minimax concave penalty approach proposed by Zhang et al. [161], and the adaptive LASSO approach proposed by Zhang [162]. The de-biased LASSO approach [30] first performs sure independence screening (SIS) to choose a relatively small number of mediators that are most associated with the outcome or the exposure, based on p-values from linear regression, then fit the outcome model for the remaining mediators using de-biased LASSO approach. Such a step is followed by fitting the mediator models using linear regression among those mediators that have both survived the screening (in step 1) and been identified by the LASSO (in step 2). Finally, the global indirect effect is estimated by summing the mediation contributions and the direct effect is estimated by subtracting the global indirect effect from an estimate of the total effect. The difference between our proposed method and the de-biased LASSO is that, we penalized a joint objective function including the contribution from both the mediator model and the response model simultaneously. The minimax concave penalty approach works [160] by first reducing the pool of potential mediators from a very large to a moderate number that is less than the sample size, followed by conducting the variable selection with the minimax concave penalty and finally carrying out joint significance testing for the mediation effects. We do not impose the concave restriction for the objective function and our objective function targets the indirect effects of interest only. The adaptive LASSO approach [161] works by first obtaining an initial model fit using either LASSO or elastic net depending on the user. Then, estimates from this fit are used to compute the adaptive weights used in the adaptive LASSO. Once the final adaptive LASSO estimates are obtained for the outcome model, estimates for the selected mediator models are obtained by linear regression. The mediation contributions (indirect effects) are computed as products of coefficients from the two models, the joint indirect effect is estimated by summing the mediation contribu-

tions, and the direct effect is estimated by subtracting the global indirect effect from an estimate of the total effect. Similarly, compared to the adaptive LASSO technique, the objective function of our method consists of both the mediator model and the response model. Moreover, our proposed method penalizes the indirect effects of interest in a more straightforward way.

Once the true mediators have been selected, estimations of causal effects of interest (throughout the true mediator(s)) become a focal point for researchers. It's worth noting that their interest often extends beyond the overall joint mediation effects resulting from all mediators collectively when more than one true mediator is selected, but also on the individual mediation effects and the effects mediated by specific subsets of mediators [147]. Expanding on this notion, typically a large number of mediation effects can be defined. In fact, the count of conceivable mediation effects grows exponentially as the number of mediators increases. However, not every effect carries significance or requires estimation. In practice, only a select few of these effects hold significance and warrant precise estimation. To illustrate, consider a scenario involving three mediators. Among the estimable mediation effects, there are three individual effects, each stemming from a single mediator. Additionally, there are three grouped effects, arising from pairs of mediators. Finally, the joint effect encompassing all mediators is also estimable. While in total seven mediation effects can be derived in this simple example, only a subset of meaningful ones merit careful estimation. In this thesis, unless provided with additional information (e.g. in a specific context for a real problem), we believe that the grouped mediation effects from multiple mediators are only meaningful and worth precise estimating if the component mediators contributed to those effects exhibit associations with each other. On the other hand, we also believe that only for mediators that are not associated with others, their corresponding individual effects are meaningful and need to be estimated. Moreover, we assume that among the possible high-dimensional mediators, only a small subset of them are associated with each other. Figure 1.5 illustrates an example involving 3 mediators. In this example, M_1 and M_2 are uncausally related while M_3 is independent of the others. Though there are in total 7 causal effects that can be defined, among the 7 estimable effects, only the grouped indirect effect from M_1 and M_2, the individual indirect effect from M_3 and the direct effect are meaningful enough and worth estimation.

Following such notions, it remains imperative to understand the dependency structures among the mediators, even after selecting them from a large set of candidates, before delving into the estimation of causal effects. In a similar manner, challenges arising from high dimensionality continue to exist, owing to the quadratic surge in the number of correlations in accordance with the number of mediators, even if only pairwise correlations are considered. As a consequence, the reduction in dimensionality achieved by selecting me-

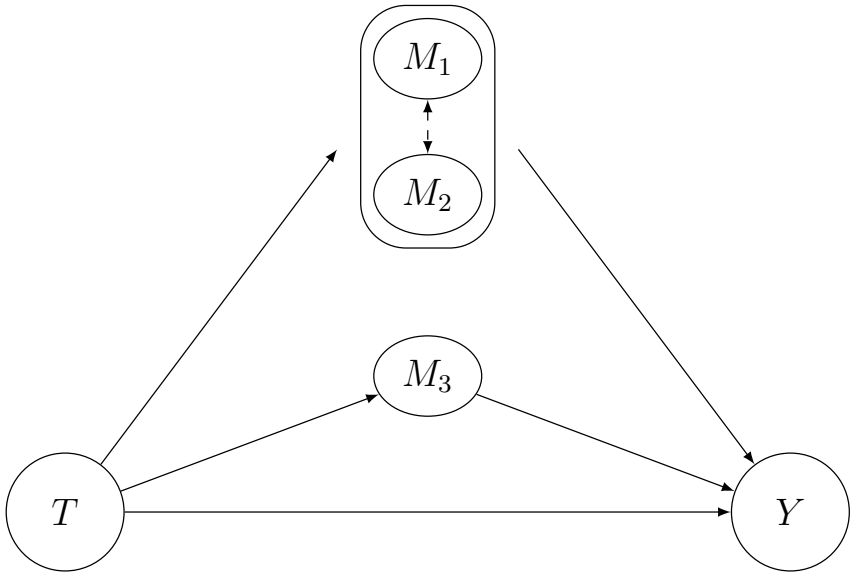


Figure 1.5: A 3 mediator illustrating example

diators from the initial high-dimensional candidates does not entirely alleviate the issue. The exponential growth rate of correlations implies that the dimension of the correlation parameter remains sufficiently substantial to pose challenges. To illustrate, in the DNA methylation example, there is an initial mediator candidate pool of 385882 variables. Even if only 100 mediators are selected from this set, there still exist 4950 pairs of correlations demanding estimation, and 2^{100} different estimable causal effects can be defined. To address this challenge, we propose an innovative approach that encompasses the simultaneous selection of mediators from high-dimensional candidates while implementing selection and dimension reduction on their dependency structures. This dual-process strategy seeks to effectively pick up the correct mediators and simplify their correlation patterns within high-dimensional mediator contexts, where causal effects can be estimated based on that. Following Jerolon et al. [61], we estimate the grouped indirect effects for mediators showing strong correlations and individual effects for mediators that are believed independent of the others.

1.5 Multiply robust methods of causal mediation analysis

As introduced in Section 1.1, addressing the confounding issues is an essential part of causal inference studies. The two most common approaches addressing the issue are the regression approach and the inverse-propensity weighting approach. Depending on the different approaches, two models need to be specified: 1. a response model, which is a regression model linking the treatment and baseline covariates with the outcome; and 2. a treatment model (also called the propensity model), which models exposure/treatment assignment conditioning on baseline covariates. When used individually to estimate causal effects of interest, for whichever approach, consistent estimation of the causal effects requires the corresponding statistical model to be correctly specified. However, when the underlying model that expresses the effect of the confounder(s) on the outcome or the treatment is misspecified, none of the approaches is able to eliminate the bias caused by confounding. In addition, model misspecification frequently arises due to the high dimensionality of confounders and limited knowledge of the causal mechanism [140, 145].

The concept of doubly robust estimation, introduced by Robins et al. [100], is designed to mitigate selection bias arising from uncontrolled nonresponse and attrition, nonrandom treatment assignment in observational studies and noncompliance in randomized experiments [67]. The doubly robust (DR) estimation approaches work by combining the aforementioned two models together, such that consistent estimates of the causal effects can be drawn if either one of the two models is correctly specified [29, 67]. Therefore, combining the two models together provides double protection from confounding. Robins et al. [100] and Rotnitzky et al. [103] propose augmented inverse probability-weighted estimators in models with missing data, where Scharfstein et al. [114] show the double robustness property of such estimators. Kang et al. [67] and Schafer et al. [113] extend the estimating framework to causal inference settings and further propose different forms of estimators that enjoy double robust properties. According to their studies, two of the most popular ways of constructing DR estimators are the inverse propensity weighted regression (WR) approach and the augmented inverse propensity weighting (AIPW) approach. Suppose for each subject, we observe a treatment assignment (T), a vector of covariates $\boldsymbol{X}$ (that has a leading $\mathbf{1}$) and an outcome Y. The response model refers to the model associating the conditional mean of outcome Y with treatment T and covariates X. If we denote $\tau = E\{Y(1) - Y(0)\}$, as the average causal effect, then the two methods estimate τ in a doubly robust way as follows:

- The WR approach assumes a response model under treatment ($T = 1$): $E\{Y(1)\} =$

$\boldsymbol{X}'\boldsymbol{\beta}^1$; and a model under control ($T=0$): $E\{Y(0)\} = \boldsymbol{X}'\boldsymbol{\beta}^0$. Here, $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^1$ are estimated by solving the following two estimating equations:

$$P_n\left\{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}(Y-\boldsymbol{X}'\boldsymbol{\beta}^1)\boldsymbol{X}\right\}=0, \tag{1.1}$$

and

$$P_n\left\{\frac{I(T=0)}{P(T=0|\boldsymbol{X})}(Y-\boldsymbol{X}'\boldsymbol{\beta}^0)\boldsymbol{X}\right\}=0, \tag{1.2}$$

where the probabilities on the denominators are the probability of being assigned treatment (control) arms and P_n is a short-form notation for $1/n\sum_{i=1}^{n}$, where the sample of n independent realization of $\boldsymbol{X}, T, Y$ is denoted $\{\boldsymbol{X}_i, T_i, Y_i, i=1,\ldots,n\}$. Solutions to the above two estimating equations are denoted as $\hat{\beta}^0_{WR}$ and $\hat{\beta}^1_{WR}$ respectively, then,

$$\hat{\tau}_{WR} = P_n\{\boldsymbol{X}'\hat{\boldsymbol{\beta}}^1_{WR} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}^0_{WR}\} \tag{1.3}$$

is a DR estimator for τ;

- The AIPW approach assumes the response model as $E\{Y|\boldsymbol{X}\} = \mu^1_Y(\boldsymbol{X})$ for $T=1$ (under treatment) and $E\{Y|\boldsymbol{X}\} = \mu^0_Y(\boldsymbol{X})$ for $T=0$ (under control). We let $\hat{\tau}_{1,AIPW}$ be an AIPW estimator for the expected potential outcome $E\{Y(1)\}$ and $\hat{\tau}_{0,AIPW}$ be an AIPW estimator for the expected potential outcome $E\{Y(0)\}$, where

$$\hat{\tau}_{1,AIPW} = P_n\left\{\frac{I(T=1)}{Pr(T=1|\boldsymbol{X})}[Y-\mu^1_Y(\boldsymbol{X})]+\mu^1_Y(\boldsymbol{X})\right\}, \tag{1.4}$$

and

$$\hat{\tau}_{0,AIPW} = P_n\left\{\frac{I(T=0)}{Pr(T=0|\boldsymbol{X})}[Y-\mu^0_Y(\boldsymbol{X})]+\mu^0_Y(\boldsymbol{X})\right\}, \tag{1.5}$$

then

$$\hat{\tau}_{AIPW} = \hat{\tau}_{1,AIPW} - \hat{\tau}_{0,AIPW} \tag{1.6}$$

is a DR estimator for the average causal effect τ.

The intuition of both approaches is as follows: the WR approach weighs each observation to the aforementioned inverse of their propensity to create a pseudo sample mimicking random sampling. Therefore, fitting regression models on the weighted dataset is similar to working with random sampling data. So the weights provide protection on the consistent estimation of the regression coefficients $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^1$; the AIPW approach using the aforementioned inverse propensity weights to correct for bias remains in the residuals ($Y - \mu_Y^0(\boldsymbol{X})$ and $Y - \mu_Y^1(\boldsymbol{X})$) and thus correct for the bias of the estimation of τ. Some other approaches for constructing doubly robust estimators include combining the regression model with the propensity score matching and stratification approach. The propensity score matching algorithm matches each subject from the treated group to one from the control group with a "similar" propensity score. The stratification approach splits the data into multiple strata, where within each strata, data points have "similar" values of propensity scores. Then causal effects are calculated by making comparisons between data from the treatment and control groups within each stratum. Both approaches adjust the observational dataset to a dataset that aims to mimic a randomized experiment, in a way that treatment status is randomly assigned to each subject. There are other ways of constructing doubly robust estimators. See Schafer et al.[67] for more details on doubly robust approaches in causal analysis.

When it comes to mediation analysis, in addition to the previously mentioned treatment model (propensity model) and the response model, we require an additional set of mediator models. As a consequence, since our framework comprises more than two models, the concept of 'double robustness' is not valid, but instead, researchers propose the concept of 'multiple robustness (MR)'. However, there exist varying perspectives among researchers on the definition of multiply robust methods. Han et al. [39] define multiply robust methods as: for an estimation method that consists of multiple (usually more than two) working models, consistent estimations can be achieved under the correct specification of one or multiple working models. On the other hand, Wang et al. [146] use the term 'multiple robustness' to denote the situation where consistent estimations are achieved as long as one of the multiple working models is correctly specified. For our proposed method, it is required that at least two out of the three models that are going to be introduced are correctly specified, in order to achieve consistent estimations of the causal effects and/or potential outcomes.

In this thesis, we extend the aforementioned two ways of constructing DR estimators (the WR approach and the AIPW approach) to mediation analysis settings and propose two ways of constructing MR estimators for causal mediation analysis with multiple possibly correlated mediators.

1.6 Structure of the thesis

This thesis is organized as follows: In Chapter 2, we first review the concept of multiple uncausally related mediators, define causal effects of interest under such settings and provide assumptions on identifiability. We then propose a copula-based regression method that jointly models the multiple uncausally related mediators. Finally, we illustrate estimations of causal effects of interest under the proposed joint modeling framework. In Chapter 3, we introduce the proposed penalization-based method that is able to both select mediators from the high-dimensional candidates and simplify their dependency structures. In Chapter 4, we propose two ways of constructing multiply robust estimators for causal mediation analysis when multiple possibly correlated mediators exist.

Chapter 2

Causal mediation analysis of multiple uncausally related mediators

2.1 Introduction

We consider the causal model presented in Figure 2.1 that includes a treatment or exposure variable T, J uncausally related mediators $M_1, \ldots, M_J$, an outcome Y and pre-treatment covariates $\boldsymbol{X}$. The bold format notations in this thesis are used to denote vectors (e.g. $\boldsymbol{X}$ denotes the vector of covariates). The causal relationships among the variables are represented by solid arrows while non-causal associations are represented using bi-directed dashed arrows. In this setting, the exposure may have a direct effect on the outcome and/or an indirect effect through one or more of the mediator(s). Lowercase letters are used to denote the corresponding realizations of the random variables. From Figure 2.1, the treatment has a causal effect on both the mediators and the outcome and the outcome is causally affected by the mediators. The multiple mediators are uncausally related. In our model, as presented in Figure 2.1, we are considering all the mediators to be uncausally related.

We focus on a binary treatment (or exposure) with $T \in \{0, 1\}$, and let $M_j(t_j)$ denote the potential value taken by the jth mediator if the corresponding treatment were assigned to level $t_j \in \{0, 1\}$, $j = 1, \ldots, J$. Extensions to deal with continuous exposure variables are possible. We use the notation $Y(t, \boldsymbol{m}) = Y(t, m_1, \ldots, m_J)$ to denote the potential value of the outcome when the corresponding treatment is assigned to level t and each of the mediators M_j takes value m_j, where $\boldsymbol{m}$ denotes the vector $\{m_1, \ldots, m_J\}$. The nested potential outcome notation $Y(t_0, M_1(t_1), \ldots, M_J(t_J))$ is used to define natural effects [90]; a more

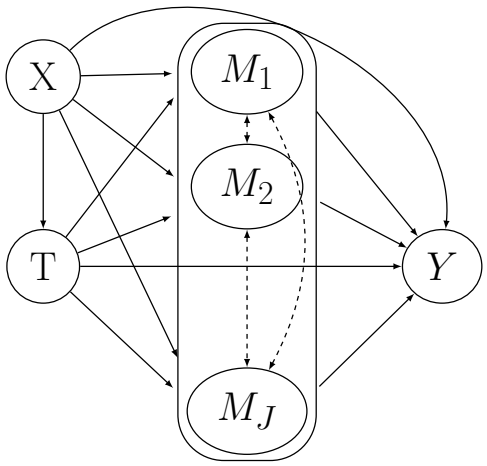


Figure 2.1: The causal diagram of multiple uncausally related mediators

compact notation for $Y(t_0, M_1(t_1), \ldots, M_J(t_J))$ is $Y(t_0, \boldsymbol{M}(\boldsymbol{t}))$, where $\boldsymbol{t} = \{t_1, \ldots, t_J\}$; this notation denotes the potential value of Y under the scenario that the treatment was set to t_0 and the mediator M_j takes a value that it would be obtained if the treatment associated with it were assigned to the level t_j respectively for $j = 1, \ldots, J$. While it is impossible to observe $Y(t_0, \boldsymbol{M}(\boldsymbol{t}))$ if $t_i \neq t_j$ for some $i \neq j$, defining the potential outcome in this way makes it possible to define causal effects using flexible expressions involving the potential values, including the individual indirect effect, which is defined in terms of counter-factual values. See Wang et al.[147] and Lange et al.[73] for details on the definition of causal effects using nested potential outcomes with multiple mediators. If $t_j = t'$, where $t' \in \{0, 1\}$ for all $j = 1, \ldots, J$, then we denote $Y(t, M_1(t'), \ldots, M_J(t'))$ as $Y(t, \boldsymbol{M}(t'))$. Note that in this notation, t' is not in a bold format since it does not denote a vector involving multiple values, but rather a single value.

The total effect of T on Y is

$$TE = E\left\{Y(1, \boldsymbol{M}(1))\right\} - E\left\{Y(0, \boldsymbol{M}(0))\right\}, \tag{2.1}$$

which represents the expected change in outcome Y when T is set to 1 versus 0. Here the expectation is taken with respect to the distribution of the potential value in the population [42].

Following Pearl[91], the natural direct effect $DE(\boldsymbol{t}) = DE(t_1, \ldots, t_J)$ denotes the ex-

pected change in outcome Y from changing T from control (0) to treatment (1) while setting all mediators at whatever value they would have obtained if their corresponding treatment levels were set to levels $t_1, \ldots, t_J$. It is defined as,

$$DE(\boldsymbol{t}) = E\left\{Y(1, \boldsymbol{M}(\boldsymbol{t}))\right\} - E\left\{Y(0, \boldsymbol{M}(\boldsymbol{t}))\right\}.$$

The natural direct effect reflects the expected change in the outcome from changing the treatment only. When there are J mediators, there are 2^{J+1} potential outcomes, so there are 2^J natural direct effects. However, one can only observe the potential outcomes when treatment assignments for the mediators are all equal to the received treatment, e.g. only potential outcomes with the form $Y(1, \boldsymbol{M}(1))$ (under the treatment arm) or $Y(0, \boldsymbol{M}(0))$ (under the control arm) are observable. Researchers aim to evaluate the effect when mediators within the same individual are simultaneously observed under the same treatment level t (see [147]). Therefore a simplified definition $DE(t)$ (as a simplified notation for $DE(t, \ldots, t)$) is often used:

$$DE(t) = E\left\{Y(1, \boldsymbol{M}(t))\right\} - E\left\{Y(0, \boldsymbol{M}(t))\right\}. \tag{2.2}$$

Such a definition represents the direct effect when controlling all the mediators to be the values that they would have taken under treatment status t. In practice, when evaluating the direct effect, it is often reasonable to keep the mediators at the control level, since it is more common for researchers to regard control as the reference level [147]. Therefore $DE(0)$ is often more of interest. For ease of notation, we denote $DE(0)$ as DE unless stated otherwise. At treatment t, the joint indirect effect is

$$IE(t) = E\left\{Y(t, \boldsymbol{M}(1))\right\} - E\left\{Y(t, \boldsymbol{M}(0))\right\}, \tag{2.3}$$

which reflects the causal effect of the treatment on the outcome through the mediator. Finally, we note that $TE = DE(0) + IE(1) = DE(1) + IE(0)$, so the joint natural indirect effect and the natural direct effect sum to the total causal effect. If there are no treatment-mediator interactions, we have that $DE = DE(1) = DE(0)$ and $IE = IE(1) = IE(0)$.

In models with multiple mediators, interest may lie in the causal indirect effect(s) through an individual or a set of mediator(s) [147]. The individual indirect effect captures the causal indirect effect only through a particular mediator while the indirect effect through a subset of mediators captures the causal indirect effects from a subset of mediators. We use $IE_j(t_0, t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_J)$ to denote the mediation effect through the jth mediator (M_j), with exposure set to be t_0 and values of the other mediators except for M_j to be fixed at values that they would have attained under exposure levels

$t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_J$, respectively. Then,

$$\begin{aligned} & IE_j(t_0, t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_J) \\ =& E\left\{Y(t_0, M_1(t_1), \ldots, M_{j-1}(t_{j-1}), M_j(1), M_{j+1}(t_{j+1}), \ldots, M_J(t_J))\right\} \\ & - E\left\{Y(t_0, M_1(t_1), \ldots, M_{j-1}(t_{j-1}), M_j(0), M_{j+1}(t_{j+1}), \ldots, M_J(t_J))\right\}. \end{aligned} \tag{2.4}$$

Similarly, we can define 2^J indirect effects corresponding to 2 treatment levels and 2^{J-1} settings for the $J-1$ mediators. For example, in the setting of two mediators, if we fix t to 1, we can define the individual indirect effect through mediator M_1 as either

$$E\{Y(1, M_1(1), M_2(0))\} - E\{Y(1, M_1(0), M_2(0))\}$$

or

$$E\{Y(1, M_1(1), M_2(1))\} - E\{Y(1, M_1(0), M_2(1))\},$$

which are also called the 'exit' and 'entrance' indirect effects by Fan et al. [154]. To retain the additive property as in Wang et al. [147], we define individual indirect effects as,

$$\begin{aligned} IE_j =& IE_j(1, 1, \ldots, 1, 0, \ldots, 0) \\ =& E\left\{Y(1, M_1(1), \ldots, M_{j-1}(1), M_j(1), M_{j+1}(0), \ldots, M_J(0))\right\} \\ & - E\left\{Y(1, M_1(1), \ldots, M_{j-1}(1), M_j(0), M_{j+1}(0), \ldots, M_J(0))\right\}. \end{aligned} \tag{2.5}$$

With this definition, $IE = IE_1 + IE_2 + \cdots + IE_J$, and the indirect effect from a set of mediators can be defined as the sum of the individual indirect effects through the components. Moreover, if $\mathcal{A}$ is the set of mediators of interests, the joint indirect effects through this set are calculated as $IE_{\mathcal{A}} = \sum_{i \in \mathcal{A}} IE$.

When defining the causal mediation effects through individual or group of mediators, the ordering of the mediators matters: different orderings lead to different definitions based on 2.5. We acknowledge that this is a major limitation of the definition 2.5 for the indirect effects. Other possible definitions with similar forms (i.e. expected differences between two or more potential outcomes) can be handled similarly with the proposed method. For example, one may also define

$$\begin{aligned} IE_j^* =& IE_j(0, \ldots, 0, 1, 0, \ldots, 0) \\ =& E\left\{Y(1, M_1(0), \ldots, M_{j-1}(0), M_j(1), M_{j+1}(0), \ldots, M_J(0))\right\} \\ & - E\left\{Y(1, M_1(0), \ldots, M_{j-1}(0), M_j(0), M_{j+1}(0), \ldots, M_J(0))\right\}, \end{aligned}$$

such that IE_j^* captures the change applied to M_j only while all the other mediators remain at the baseline levels.

In mediation analysis with multiple causally related mediators, following Imai et al. [55, 58], the hierarchical potential outcomes are usually adopted to investigate the effect contributed by each mediator in orders. For example, when dealing with problems as depicted in 3.3, the hierarchical potential outcome has the form $Y(t_0, M_2(t_2, M_1(t_1)))$ reflecting the fact that M_1 causally affects M_2. Our proposed framework can be extended to incorporate causally related mediators as well. For example, in a problem involving 3 mediators, where M_1 and M_2 are uncausally related but they both causally affect M_3 (as illustrated in Figure 2.2), the potential outcome can be adjusted to the form $Y(t_0, M3(t_3, M1(t_1), M_2(t_2)))$ and similar analysis can be proceeded with the adjusted forms of potential outcomes. However, this flexibility is not explored here as the focus of this thesis is on dealing primarily with uncausally related mediators.

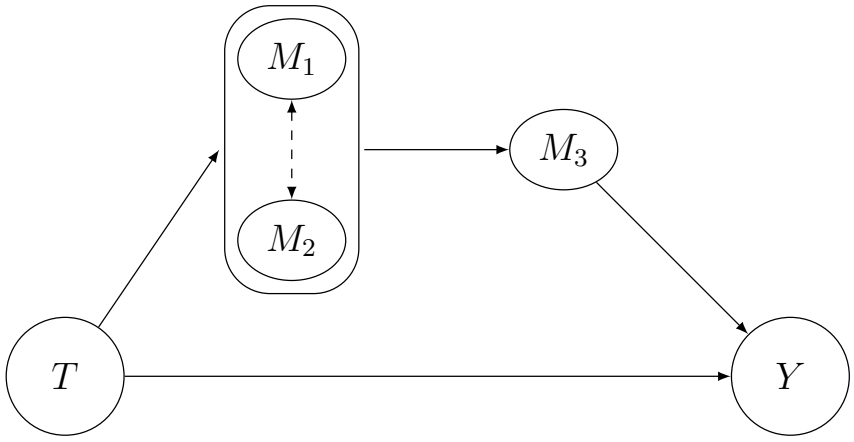
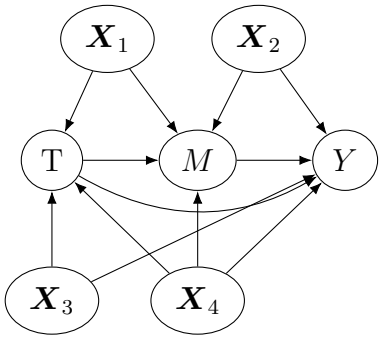


Figure 2.2: A scenario where M_1 and M_2 are uncausally related and they causally affect M_3

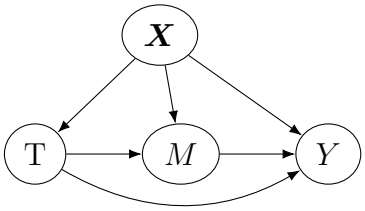
2.2 Assumption and Identifiability

2.2.1 Brief review of simple mediation

We first begin with a brief review of simple mediation analysis. Following Imai et al. [54], when there is only one mediator M, causal effects based on the potential outcomes framework can be identified and estimated via approaches built on the mediation formula



(a) The DAG with different confounding structures when a single mediator exists



(b) The simplified DAG when a single mediator exists

Figure 2.3: The DAG with different confounding structures when a single mediator exists

(Formula (2.6), to be introduced) under a set of assumptions. Imai et al. call the set of assumptions *Sequential Ignorability Assumptions (SIA)*. These assumptions state that:

(a) conditioning on the observed pretreatment covariates, the observed treatment is independent of all potential values of the outcome and mediating variables, i.e. the treatment assignment is assumed to be ignorable; and

(b) the observed mediator is independent of all potential outcomes given the observed treatment and pretreatment covariates, i.e. the mediator is ignorable.

To be specific, following Imai et al. [55], the SIA state as follows:

Assumption 1 (The Sequential Ignorability Assumptions (SIA)). [55]

The following two statements of conditional independence hold:

1. $\{Y(t', m), M(t)\} \perp T | \boldsymbol{X} = \boldsymbol{x}$
2. $Y(t', m) \perp M | \boldsymbol{X} = \boldsymbol{x}, T = t$

where $0 < Pr(T = t|\boldsymbol{X} = \boldsymbol{x}) < 1$ and $0 < P(M(t) = m|\boldsymbol{X} = \boldsymbol{x}, T = t) < 1$, for $t, t' \in \mathcal{T}$ and $m \in \boldsymbol{\mathcal{M}}$.

Figure 2.3a presents the mediation structure when only a single mediator is involved. Under such a scenario, there could be several sets of pre-treatment covariates (e.g. $\boldsymbol{X}_i (i = 1, \ldots, 4)$), which should be addressed respectively in the assumptions. For simplicity, we consider a simplified case such that $\boldsymbol{X}$ is the union of the covariates included in the model, as presented in Figure 2.3b. We also require $\boldsymbol{X}$ to fulfill the conditions that:

1. $\boldsymbol{X}$ should be pre-treatment covariates; and

2. $\boldsymbol{X}$ confound one or some of the pathways among T to Y, T to M and M to Y.

Further, Imai et al. [54] show that under *SIA*, the distribution of the potential outcome conditioning on the respective covariates, can be estimated via the formula:

$$f(Y(t, M(t'))|\boldsymbol{X} = \boldsymbol{x}) = \int_{\boldsymbol{\mathcal{M}}} f(Y|\boldsymbol{X} = \boldsymbol{x}, T = t, M = m) \, \mathrm{d} \, F(m|\boldsymbol{X} = \boldsymbol{x}, T = t'). \quad (2.6)$$

In the above equation, $f(Y|M = m, T = t, X = x)$ denotes the distribution of outcome Y given mediator, treatment and baseline covariates and $F(m|\boldsymbol{X} = \boldsymbol{x}, T = t')$ denotes the conditional distribution of the mediator M given treatment and baseline covariates. When facing discrete random variables, one should replace the integration by summation. Therefore, the average direct and indirect effects are calculated as follows,

$$\begin{aligned} & IE(t) \\ = & \int_{\mathcal{X}} \Big\{ \int_{\boldsymbol{\mathcal{M}}} E(Y|\boldsymbol{X} = \boldsymbol{x}, T = t, M = m) \, \mathrm{d} \, F(m|\boldsymbol{X} = \boldsymbol{x}, T = 1) \\ & - \int_{\boldsymbol{\mathcal{M}}} E(Y|\boldsymbol{X} = \boldsymbol{x}, T = t, M = m) \, \mathrm{d} \, F(m|\boldsymbol{X} = \boldsymbol{x}, T = 0) \Big\} \, \mathrm{d} \, F_{\boldsymbol{X}}(\boldsymbol{x}), \end{aligned}$$

$$\begin{aligned} & DE(t) \\ = & \int_{\mathcal{X}} \Big\{ \int_{\boldsymbol{\mathcal{M}}} E(Y|\boldsymbol{X} = \boldsymbol{x}, T = 1, M = m) \, \mathrm{d} \, F(m|\boldsymbol{X} = \boldsymbol{x}, T = t) \\ & - \int_{\boldsymbol{\mathcal{M}}} E(Y|\boldsymbol{X} = \boldsymbol{x}, T = 0, M = m) \, \mathrm{d} \, F(m|\boldsymbol{X} = \boldsymbol{x}, T = t) \Big\} \, \mathrm{d} \, F_{\boldsymbol{X}}(\boldsymbol{x}). \end{aligned}$$

Similarly, when the random variables are discrete, the integration should be replaced by summation. Such estimation formulas are called the "mediation formula" by Pearl [92].

2.2.2 Identifiability of multiple uncausally related mediators

Jerolon et al. [60] formalize the concept of uncausally related mediators and extend *SIA* to account for the situation when multiple uncausally related mediators exist. When extending *SIA*, the biggest challenge is that the mediators themselves are the confounders of each other. In addition, one of the most prevailing reasons for mediators uncausally

related is the existence of unmeasured covariates U causally affecting all or some of the mediators. Under such situations, U are unobserved confounders between the mediators and the outcome, so SIA does not hold [60]. The extended version of *SIA* to account for multiple uncausally related mediators, is called the *Sequential Ignorability for Multiple Mediators Assumption (SIMMA)* by Jerolon et al. [60]. In the original version of SIMMA by Jerolon et al., mediators are categorized into the ones of primary interest and the others. In this thesis, we do not distinguish them. Therefore, by generalizing SIMMA from Jerolon et al. [60], we propose our version of SIMMA, which is stated as follows:

Assumption 2 (The Sequential Ignorability for Multiple Mediators Assumption (SIMMA)). The following two statements hold:

1. $\{Y(t_0, m_1, m_2, \ldots, m_J), M_1(t_1), \ldots, M_J(t_J)\} \perp T | \boldsymbol{X} = \boldsymbol{x}$
2. $Y(t_0, m_1, m_2, \ldots, m_J) \perp \{M_1, \ldots, M_J\} | \boldsymbol{X} = \boldsymbol{x}, T = t$

where $0 < P(T = t|\boldsymbol{X} = \boldsymbol{x}) < 1$, and for any $j \in \{0, \ldots, J\}$, $0 < P(M_j(t) = m_j|\boldsymbol{X} = \boldsymbol{x}, T = t) < 1$, $t_j \in \mathcal{T}$ and $m_j \in \boldsymbol{\mathcal{M}}$.

The *SIMMA* indicates that:

(a) conditioning on the observed pre-treatment covariates, the observed treatment is independent of all potential values of the outcome and all mediating variables; and

(b) the observed mediators are jointly independent of all potential outcomes given the observed treatment and pre-treatment covariates.

An intuitive understanding of SIMMA is that, this assumption treats the multiple mediators, no matter uncausally related or not, as a whole and considers the joint distribution of all mediators. Our modeling method is also based on such an idea and considers modeling the set of mediators jointly. One needs to be cautious that, both SIA and SIMMA can be violated even with randomized treatment assignment, this is because the randomization can only solve the confounding issue for the first part (confounding between treatment and mediator, or treatment and the outcome), it cannot solve the confounding issue between mediator and outcome. For detailed discussion related to randomization and the ignorability assumptions, please refer to Imai et al. [55, 54]

Under *SIMMA*,

$$
\begin{aligned}
&E\{Y(t_0, M_1(t_1), \ldots, M_J(t_J))\} \\
&= \int_{\boldsymbol{\mathcal{X}}} \left\{ \int \cdots \int_{m_1,\ldots,m_J} \int E\{Y|\boldsymbol{X} = \boldsymbol{x}, T = t_0, M_1 = m_1, \ldots, M_J = m_J\} \right. \\
&\left. d^J F_{M_1 \ldots M_J}(m_1, \ldots, m_J|\boldsymbol{X} = \boldsymbol{x}, T_1 = t_1, \ldots, T_J = t_J) \right\} \mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}),
\end{aligned}
$$

which can be simplified as

$$
\int_{\boldsymbol{\mathcal{X}}} \left\{ \int \cdots \int_{\boldsymbol{\mathcal{M}}} \int E\{Y|\boldsymbol{x}, t_0, \boldsymbol{m}\} \, \mathrm{d}F(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{t}) \right\} \, \mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}). \tag{2.7}
$$

Similarly, one may replace the integration by summation to account for discrete random variables. With the estimated potential outcomes, causal effects of interest, as functions of potential outcomes, are calculated respectively via (2.1), (2.3), (2.2), (2.5).

Having the *SIMMA* is not adequate for the individual indirect effects (and its extension e.g. the indirect effects through a subset of mediators) to be identifiable. We therefore propose the following assumptions.

Assumption 3 (Invariant correlations among mediators). The conditional correlations among the multiple mediators do not vary with any arbitrary combinations of treatment assignments $t_1, \ldots, t_J$.

$$
Cor(M_1(t_1), \ldots, M_J(t_J)|\boldsymbol{X}) = \boldsymbol{\varrho}
$$

for any $t_1, \ldots, t_J \in \{0, 1\}^J$, where $Cor(\cdot)$ denotes the correlation matrix among the following elements and $\boldsymbol{\varrho}$ denotes the conditional correlation matrix among mediators.

Assumption 3 guarantees that $\boldsymbol{\varrho}$ is fixed and not affected by $\boldsymbol{t}$, which makes estimation of correlation structure among mediators feasible. Moreover, Assumption 3 also guarantees that such an estimated correlation structure could be used when imputing the potential outcomes under counterfactual scenarios, particularly, under counterfactual combinations of treatment schemes (i.e. $t_i \neq t_j$ for some $i \neq j$). Even though we may relax Assumption 3 to some extent, we still need to assume two fixed correlations under either the treatment or control arm and it is impossible to drop Assumption 3 completely. Assumption 3R states the relaxed version of Assumption 3.

Assumption 3R (Invariant correlations among mediators (relaxed version)). The conditional correlations among the multiple mediators under either the treatment or control setting are assumed to be fixed.

$$
Cor(M_1(0), \ldots, M_J(0)|\boldsymbol{X}) = \boldsymbol{\varrho}_0 \qquad \text{and} \qquad Cor(M_1(1), \ldots, M_J(1)|\boldsymbol{X}) = \boldsymbol{\varrho}_1
$$

The reason why we cannot drop Assumption 3 completely is that the mediator values under arbitrary counterfactual combinations of treatment assignments ($t_i \neq t_j$ for some $i \neq j$) are never observed. Therefore, without Assumption 3, the correlation structure among mediators cannot be estimated and the proposed analysis cannot proceed. Hence, Assumption 3 provides fundamentals for the proposed analysis framework, particularly, if our interest lies within the estimation of individual indirect effects, though such an assumption may sometimes be violated in practice and is untestable. Nevertheless, if our interests lie only within the direct and joint indirect effects, and we are not interested in individual indirect effects, then Assumption 3 can be dropped. We acknowledge that this is a main limitation of the proposed method. In addition, throughout this chapter, we assume Assumption 3 is satisfied and also $\boldsymbol{\varrho}$ is a fixed but unstructured matrix. If Assumption 3R, rather than Assumption 3 is adopted, then in Section 2.3.2, when estimating the correlation structure, we model two different correlation matrices under either the treatment or control arm separately. Following the same logic, a partial test can be proposed to test if variance among mediators does not vary with changing exposure(s), i.e. testing Assumption 3 versus Assumption 3R, using observed data. Assumption 3 cannot be fully tested due to unobservable counterfactual values.

Moreover, the consistency and the positivity assumptions are also assumed analogous to most studies for causal analysis. The consistency assumption implies that an individual's potential outcome, given their observed exposure history, aligns precisely with their observed outcome [99]. On the other hand, the positivity assumption dictates that there exists a non-zero (i.e., positive) probability of receiving each level of exposure for every possible combination of exposure and confounding variables present among individuals in the population [41]. These two assumptions are widely adopted in studies involving causal inferences. While integral to our analysis, we refrain from delving further into extensive discussions regarding these assumptions.

2.3 Method

Formula (2.7) involves modeling both the distribution of mediators conditioning on treatment and covariates ($F(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T})$) and the outcome conditioning treatment, covariates and mediators ($E\{Y|\boldsymbol{X},T,\boldsymbol{M}\}$). We call the model for $F(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T})$ the mediator model and it is denoted as $\mathcal{M}_M$. Throughout this thesis, we assume $\boldsymbol{M}$ has a density and we denote the density as $f(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T})$. We call the model for $E\{Y|\boldsymbol{X},T,\boldsymbol{M}\}$ the response model, which is denoted as $\mathcal{M}_Y$. With the estimated $\hat{F}(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T})$ and $\hat{\mu}_Y(\boldsymbol{X},T,\boldsymbol{M})$, potential outcomes are imputed via (2.7) and causal effects of interest are estimated according to

their defining forms.

2.3.1 Model

We denote the marginal conditional CDF and the respective PDF of M_j as $F_j(m_j|t, \boldsymbol{x}; \boldsymbol{\theta}_j)$ and $f_j(m_j|t, \boldsymbol{x}; \boldsymbol{\theta}_j)$ respectively, where $\boldsymbol{\theta}_j$ denotes the marginal parameters. This thesis proposes a general framework for joint modeling using copula functions and we denote their joint model as $\mathcal{M}_M$. If we let $\mathcal{C}$ be the associated copula function (note that here C denotes the conditional copula function given the treatment and baseline covariates) and c be the corresponding density function, then we have

$$F(\boldsymbol{m}|\boldsymbol{X}, T; \boldsymbol{\theta}_M) = C(F_1(m_1|\boldsymbol{X}, T; \boldsymbol{\theta}_1), \ldots, F_J(m_J|\boldsymbol{X}, T; \boldsymbol{\theta}_J), \boldsymbol{\varrho}), \tag{2.8}$$

and

$$\begin{aligned} f(\boldsymbol{m}|\boldsymbol{X}, T; \boldsymbol{\theta}_M) = c(F_1(m_1|\boldsymbol{X}, T; \boldsymbol{\theta}_1), \ldots, F_J(m_J|\boldsymbol{X}, T; \boldsymbol{\theta}_J), \boldsymbol{\varrho}) \\ f_1(m_1|\boldsymbol{X}, T; \boldsymbol{\theta}_1) \cdots f_J(m_J|\boldsymbol{X}, T; \boldsymbol{\theta}_J), \end{aligned} \tag{2.9}$$

where $\boldsymbol{\varrho}$ denotes parameters for the conditional correlations depending on the copula function. One of the advantages of using a copula function for joint modeling is that the marginal distributions and the dependence structures can be modeled separately. When choosing copula functions, one must ensure that Assumption 2 is satisfied. We consider the Gaussian copula function with an unstructured dependence, under which, Assumption 2 is satisfied. Under such case, we also denote the conditional covariance as $\boldsymbol{\Sigma}$.

For the joint response model $\mathcal{M}_Y$, we assume

$$E\{Y|\boldsymbol{X}, T, \boldsymbol{M}\} = \mu_Y(\boldsymbol{X}, T, \boldsymbol{M}; \boldsymbol{\theta}_Y), \tag{2.10}$$

where $\boldsymbol{\theta}_Y$ denotes the parameters in $\mathcal{M}_Y$.

In the following illustration, there are occasions that both i and j appear in the subscript of the mediator or treatment variables, and the first subscript is used to denote the ith subject and the second one to denote the jth mediator, where $j = 1, \ldots, J$. We assume the covariates have a dimension of p, so x_{ir} denotes the rth covariate of subject $i, r = 1, \ldots, p, i = 1, \ldots, n$. In addition, we denote the column vector $\{M_{i1}, \ldots, M_{iJ}\}'$ as $\boldsymbol{M}_{i\cdot}$, and the same rule applies to other variables with two subscripts. Similarly, we denote vector $\{M_{1j}, \ldots, M_{nJ}\}$ as $\boldsymbol{M}_{\cdot j}$ and the same rule also applies to other variables with two subscripts. For example, the notation $\boldsymbol{M}_{\cdot j}$ denotes the values of the jth mediator across all subjects and $\boldsymbol{M}_{i\cdot}$ denotes the values of all mediators within subject i. Finally, a bold form

variable without subscript is used to denote the entire data matrix. Furthermore, in our model, we can allow for different covariates to be used for each marginal or the response model, but for simplicity, we denote all covariates including a leading **1** associated with the intercept as $\boldsymbol{X}$.

2.3.2 Estimation

We begin with the estimation of $\mathcal{M}_M$. A straightforward way to optimize the joint likelihood function. If denoting the n observations of independent samples as $(\boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x})$, the full data log-likelihood for $\mathcal{M}_M$ is

$$
\begin{aligned}
& l_M(\boldsymbol{\theta}_M; \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}) \\
=& \log \left\{ \prod_{i=1}^{n} f(\boldsymbol{m}_{i.} | t_i, \boldsymbol{x}_i ; \boldsymbol{\theta}_M) \right\} \\
=& \log \left\{ \prod_{i=1}^{n} \left[c\left(F_1(m_{i1}|t_i, \boldsymbol{x}_i; \boldsymbol{\theta}_1), \ldots, F_J(m_{iJ}|t_i, \boldsymbol{x}_i; \boldsymbol{\theta}_J); \boldsymbol{\varrho}\right) \prod_{j=1}^{J} f_j(m_{ij}|t_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j) \right] \right\} \\
=& \sum_{i=1}^{n} \left\{ \log \left[c\left(F_1(m_{i1}|t_i, \boldsymbol{x}_i; \boldsymbol{\theta}_1), \ldots, F_J(m_{iJ}|t_i, \boldsymbol{x}_i; \boldsymbol{\theta}_J); \boldsymbol{\varrho}\right) \right] + \sum_{j=1}^{J} \log \left[f_j(m_{ij}|t_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j) \right] \right\}.
\end{aligned}
\tag{2.11}
$$

Optimizing (2.11) yields the MLE,

$$
\hat{\boldsymbol{\theta}}_M = \underset{\boldsymbol{\theta}_M}{\operatorname{argmax}}\, l_M(\boldsymbol{\theta}_M;\ \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}).
$$

When certain requirements as stated by Gijbels et al. [32] are satisfied, alternative methods can be implemented to simplify the estimation process. Here we propose using the two-stage approach [62, 32]. Such a procedure is also called the method of inference functions for margins (IFM) [62]. For this approach, we first estimate the parameters of marginal models. Let $l_j(\boldsymbol{\theta}_j; \boldsymbol{t}, \boldsymbol{m}_{.j}, \boldsymbol{x})$ be the log-likelihood of M_j where

$$
l_j(\boldsymbol{\theta}_j;\ \boldsymbol{t}, \boldsymbol{m}_{.j}, \boldsymbol{x}) = \sum_{i=1}^{n} \log \left\{ f_j(m_{ij}|t_i, \boldsymbol{x}_i; \boldsymbol{\theta}_j) \right\}.
$$

For each $j = 1, \ldots, J$, we estimate $\boldsymbol{\theta}_j$ as $\tilde{\boldsymbol{\theta}}_j = \underset{\boldsymbol{\theta}_j}{\operatorname{argmax}}\, l_j(\boldsymbol{\theta}_j; \boldsymbol{t}, \boldsymbol{m}_{.j}, \boldsymbol{x})$. Then plugging these estimators into the joint likelihood in (2.11), we estimate the dependency parameters as

$$
\tilde{\boldsymbol{\varrho}} = \underset{\boldsymbol{\varrho}}{\operatorname{argmax}}\, l_M(\boldsymbol{\varrho};\ \tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_J, \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}).
$$

When the two-stage method is used, under conditions stated by Joe et al. [63, 62], we obtain consistent estimators. However, the variances of the estimators will mostly be different from the ones obtained by MLE and usually, we have slightly larger variances, which can be treated as the cost we pay for easier computation. As a special case, when the copula is Gaussian, the estimators under the two-stage approach are identical to the ones obtained by MLE. For more details on the two-stage method, one may also refer to Joe et al. [63, 62].

For either the direct or the two-stage estimation method, we denote the estimated parameter as $\hat{\boldsymbol{\theta}}_M$ for simplicity. Once we obtain the estimated parameters, we plug them in (2.8) and (2.9) to get estimates of the joint conditional distributions of mediators, such that,

$$\hat{F}(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T}) = F(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T};\hat{\boldsymbol{\theta}}_M), \quad \text{and} \quad \hat{f}(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T}) = f(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{T};\hat{\boldsymbol{\theta}}_M). \tag{2.12}$$

Estimation of $\mathcal{M}_Y$ is straightforward. If $l_Y(\boldsymbol{\theta}_Y;\boldsymbol{y}|\boldsymbol{t},\boldsymbol{m},\boldsymbol{x})$ is denoted as the log-likelihood, then

$$l_Y(\boldsymbol{\theta}_Y;\boldsymbol{y}|\boldsymbol{t},\boldsymbol{m},\boldsymbol{x}) = \sum_{i=1}^{n} \log\left\{f_{Y_i}(y_i|t_i,\boldsymbol{m}_{i\cdot},\boldsymbol{x}_i;\boldsymbol{\theta}_Y)\right\}.$$

Therefore,

$$\hat{\boldsymbol{\theta}}_Y = \underset{\boldsymbol{\theta}_Y}{\operatorname{argmax}}\, l_Y(\boldsymbol{\theta}_Y;\boldsymbol{y}|\boldsymbol{t},\boldsymbol{m},\boldsymbol{x}).$$

When deciding whether to use the direct optimization of the joint likelihood or the two-stage approach, the trade-off is efficiency versus computation burden. When the two-stage approach is used, the variances of the estimators are usually larger than the ones obtained from optimizing the joint likelihood directly, and such an efficiency loss is usually the cost we pay for less computation.

2.3.3 Estimation of causal effects

Potential outcomes are then imputed following (2.7), where true values are replaced by estimated ones, and hence, causal effects of interest are calculated via their definitions as shown by (2.1)-(2.5). Following (2.7), the expected potential outcome $E\{Y(t_0,\boldsymbol{M}(\boldsymbol{t}))\}$ is imputed as

$$\int_{\boldsymbol{\mathcal{X}}} \left\{ \int_{\boldsymbol{m}} \cdots \int \hat{\mu}_Y(\boldsymbol{x},t_0,\boldsymbol{m};\hat{\boldsymbol{\theta}}_Y)\, \mathrm{d}\, \hat{F}(\boldsymbol{m}|\boldsymbol{x},\boldsymbol{t};\hat{\boldsymbol{\theta}}_M) \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}). \tag{2.13}$$

Integration over the distribution of baseline covariates $\boldsymbol{X}$ can be replaced by averaging over the distribution of $\boldsymbol{X}$ in the data. Therefore, an estimated version of (2.13) based on data is

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\boldsymbol{m}}\cdots\int \hat{\mu}_Y(\boldsymbol{x}_i,t_0,\boldsymbol{m};\hat{\boldsymbol{\theta}}_Y)\,\mathrm{d}\,\hat{F}(\boldsymbol{m}|\boldsymbol{x}_i,\boldsymbol{t};\hat{\boldsymbol{\theta}}_M)\right\}. \tag{2.14}$$

The causal effects of interest are estimated from (2.1)-(2.5).

The inner integration with respect to $\boldsymbol{m}$ can be carried out analytically, numerically or via the Monte-Carlo [54] approach. The Monte-Carlo approach works as follows. For each subject i, Monte-Carlo samples of size N are drawn from the fitted distribution $\hat{F}(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{t}_i)$. The vector of mediators in the kth Monte-Carlo sample is denoted as $\boldsymbol{m}_i^{(k)}$, $k=1,\ldots,N$. Then, kth Monte-Carlo outcome is taken to be the value $\hat{\mu}_{Y,i}^{(k)} = \hat{y}(t_0,\boldsymbol{m}_i^{(k)},\boldsymbol{X}_i)$. Next, by averaging over the N Monte-Carlo samples ($N^{-1}\sum_{k=1}^{N}\hat{\mu}_i^{(k)}$), we obtain an estimated outcome for individual i. Finally, by averaging over the n subjects, we obtain an estimate of $E\{Y(t_0,\boldsymbol{M}(\boldsymbol{t}))\}$ as desired.

In addition, we also show that, under some particular model specifications, closed forms of the results can be obtained, which is introduced in Section 2.5.

A summary of the algorithm (when a two-stage approach is used) is provided as follows:

Algorithm 1. Algorithm of the copula-based estimator using the two-stage approach:

Step 1: Solve for mediator model marginal parameters $\boldsymbol{\theta}_j$ by optimizing mediator marginal data likelihood, then obtain a fitted marginal distribution for each mediator.

Step 2: For a given copula function, solve for correlation parameters $\boldsymbol{\varrho}$ and obtain a fitted joint distribution of the mediators by combining the fitted margins with the estimated correlation parameters.

Step 3: Solve for response model parameters $\boldsymbol{\theta}_Y$ and obtain the fitted conditional mean model of response.

Step 4: Impute the expected potential outcome under each combination of treatment assignments.

Step 5: Estimate the causal effects of interest.

2.4 Theoretical Properties

2.4.1 Consistency

In this section, we provide the consistency property of the proposed estimators. The following two lemmas (Lemma 1 and Lemma 2) state the consistency of the estimators of the parameters.

Lemma 1 (Consistent estimation of the parameters of the mediator model)**.** Provided that the joint distribution model of mediators is correctly specified, under regularity conditions listed in Appendix A,

$$\hat{\boldsymbol{\theta}}_M \xrightarrow{p} \boldsymbol{\theta}_M \quad \text{as} \quad n \to \infty.$$

Lemma 2 (Consistent estimation of the parameters of the response model)**.** Provided that the conditional mean model for the response is correctly specified, under regularity conditions listed in Appendix A,

$$\hat{\boldsymbol{\theta}}_Y \xrightarrow{p} \boldsymbol{\theta}_Y \quad \text{as} \quad n \to \infty.$$

The proofs of the two lemmas utilize the consistency properties of MLE. Because both models are estimated via a likelihood-based approach, under regularity conditions (see Appendix A), we naturally have the consistency of the estimated parameters [11].

Theorem 1 (Consistency of the estimated causal effects)**.** Provided that the models for both the joint distribution of mediators and the conditional mean of the response are correctly specified, when the regularity conditions listed in Appendix A are satisfied, we have that, as $n \to \infty$,

$$\widehat{TE} \xrightarrow{p} TE, \quad \widehat{DE} \xrightarrow{p} DE, \quad \widehat{IE} \xrightarrow{p} IE, \quad \text{and} \quad \widehat{IE}_j \xrightarrow{p} IE_j \quad \text{for} \quad j \in \{1, \ldots, J\}.$$

The proof of Theorem 1 is shown in Appendix B, where the general idea is to combine the continuous mapping theorem as well as the law of large number [11] with Lemma 1 and 2 to show the overall consistency.

2.4.2 Asymptotic normality

The following theorem helps us obtain the asymptotic variance-covariance matrix of estimated parameters from $\mathcal{M}_M$ and $\mathcal{M}_Y$.

Theorem 2 (Diagonal block pattern of the variance-covariance matrix of estimated parameters)**.** If we let $Var(\hat{\boldsymbol{\theta}}_M|\boldsymbol{X},\boldsymbol{T}) = \boldsymbol{V}_M$ (the variance-covariance matrix of $\hat{\boldsymbol{\theta}}_M$), $Var(\hat{\boldsymbol{\theta}}_Y|\boldsymbol{X},\boldsymbol{T}) = \boldsymbol{V}_Y$ (the variance-covariance matrix of $\hat{\boldsymbol{\theta}}_Y$) and $\boldsymbol{\theta}_{MY} = \{\boldsymbol{\theta}'_M\boldsymbol{\theta}'_Y\}'$, then,

$$\boldsymbol{V} = Var(\hat{\boldsymbol{\theta}}_{MY}|\boldsymbol{X},\boldsymbol{T}) = \begin{pmatrix} \boldsymbol{V}_M & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_Y \end{pmatrix}.$$

When a standard MLE is used, an asymptotic version of $\boldsymbol{V}_M$ can be obtained by taking the inverse of the observed Fisher-information of $\mathcal{M}_M$ (denoted as $\boldsymbol{I}_M$). If a two-stage approach is used, the asymptotic $\boldsymbol{V}_M$ is obtained differently, see Shih et al. [116] and Joe et al. [62, 63] for details. An asymptotic version of $\boldsymbol{V}_Y$ can be obtained by taking the inverse of the observed Fisher-information of $\mathcal{M}_Y$ in a similar way (denoted as $\boldsymbol{I}_Y$). The importance of Theorem 2 is that it reduces the complexity of the correlation structure of the estimated parameters. In addition, a corollary of Theorem 2 is that $Cor(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_Y|\boldsymbol{X},\boldsymbol{T}) = \boldsymbol{0}$. The proof is in Appendix B.

On top of Theorem 2, the multivariate delta method could be utilized to obtain the asymptotic distribution of the estimated causal effects. Using $\lambda(\hat{\boldsymbol{\theta}}_{MY})$ to represent any estimated causal effects, as a function of estimated parameters $\hat{\boldsymbol{\theta}}_{MY}$, we have,

$$\sqrt{n}(\lambda(\hat{\boldsymbol{\theta}}_{MY}) - \lambda(\boldsymbol{\theta}_{MY})) \xrightarrow{d} \mathcal{N}(0, \nabla'\lambda(\boldsymbol{\theta}_{MY})\boldsymbol{V}\,\nabla\,\lambda(\boldsymbol{\theta}_{MY})). \tag{2.15}$$

Closed-form expressions of point estimators and variances under particular model settings are introduced in the next session.

2.5 Closed form results under linear and Gaussian model settings

In this section, we show that under some particular model settings, explicit mathematical results can be derived with respect to the proposed method.

2.5.1 The mediator model

We begin with the introduction of marginal models in $\mathcal{M}_M$. We assume:

$$\mu_{ij}(\boldsymbol{X}_i, T_i; \boldsymbol{\Psi}_j) = \boldsymbol{X}'_i\boldsymbol{\eta}_j + \alpha_j T_i,\ i = 1,\ldots,n,\ j = 1,\ldots,J, \tag{2.16}$$

where α_j is the unknown coefficient of treatment. When introducing the specified models, we assume $\boldsymbol{X}$ is a $p+1$-dimensional vector of covariates including a leading $\mathbf{1}$, so $\boldsymbol{\eta}_j = \{\eta_{j0}, \ldots, \eta_{jp}\}'$ is the $p+1$-dimensional vector of regression coefficients with the first element represents the intercept. We denote the mean parameters of the j-th marginal model as $\boldsymbol{\Psi}_j$ and in this case $\boldsymbol{\Psi}_j = \{\boldsymbol{\eta}_j', \alpha_j\}'$. Additionally, $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_1', \ldots, \boldsymbol{\Psi}_J'\}'$ denotes mean parameters from all marginal models. While we could allow treatment-covariates, between covariates interactions and higher-order terms of the covariates to be included in the model, for simplicity we only consider the settings as shown in (2.16). Therefore, if we let $\boldsymbol{W}_i$ denotes the covariates used for each marginal model of μ_{ij}, then under our case, $\boldsymbol{W}_i = \{\boldsymbol{X}_i, T_i\}'$.

We further assume a Gaussian copula is used to model the joint distribution of mediators, which is equivalent to assume

$$\boldsymbol{M}_i|\boldsymbol{X}_i, T_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),\ i = 1, \ldots, n, \tag{2.17}$$

where "MVN" stands for multivariate normal distribution, $\boldsymbol{\mu}_i = \{\mu_{i1}, \ldots, \mu_{iJ}\}'$ and $\boldsymbol{\Sigma} = Cov(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i)$. We additionally let $Cor(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i) = \boldsymbol{\varrho}$, so that $\boldsymbol{\Sigma}$ and $\boldsymbol{\varrho}$ denotes the conditional variance-covariance matrix and correlation matrix of mediators respectively. Furthermore, we denote $Var(M_{ik}|\boldsymbol{X}_i, T_i) = \sigma_k^2$ and $Cor(M_{ik}, M_{il}|\boldsymbol{X}_i, T_i) = \rho_{kl}$ for $k = 1, \ldots, J$, $l = 1, \ldots, J$ and $k \neq l$, so that the k, lth element of $\boldsymbol{\varrho}$ is ρ_{kl} for $k \neq l$ and 1 for $k = l$, and the k, lth element of $\boldsymbol{\Sigma}$ is $\rho_{kl}\sigma_k\sigma_l$ for $k \neq l$ and $\sigma_k^2(\sigma_l^2)$ for $k = l$. Under these settings, $\boldsymbol{\theta}_j = \{\boldsymbol{\Psi}_j', \sigma_j\}'$ and $\boldsymbol{\theta}_M = \{\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_J', \rho_{12}, \ldots, \rho_{J-1,J}\}'$.

The following theorem helps us obtain the closed-form solution of $\boldsymbol{\Psi}_M$ and its variances.

Theorem 3 (Multivariate regression model)**.** We assume, for each subject i ($i = 1, \ldots, n$), the p-dimensional outcome $\boldsymbol{Y}_i|\boldsymbol{X}_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are the marginal mean and variance-covariance parameters. For the jth margin ($j = 1, \ldots, p$), we assume $\mu_{ij} = \sum_{k=0}^{q_j} \beta_{jk} X_{ijk}$, where for each $k = 0, \ldots, q_j$, X_{ijk} denotes the kth covariate with $X_{ij0} = 1$, β_{jk} is the corresponding coefficient with β_{j0} representing the intercept and q_j denotes the total number of covariates for the jth marginal model.

If we let the design matrix of subject i be

$$\overline{\boldsymbol{X}}_i = \begin{pmatrix} \underbrace{\begin{matrix} X_{i10}, & \ldots, & X_{i1q_1}, \\ 0, & \ldots, & 0, \\ & \ddots, & \\ 0, & \ldots, & 0, \end{matrix}}_{\text{marginal 1}} & \underbrace{\begin{matrix} 0, & \ldots, & 0, \\ X_{i20}, & \ldots, & X_{i2q_2}, \\ & \ddots, & \\ 0, & \ldots, & 0, \end{matrix}}_{\text{marginal 2}} & \ldots, & \underbrace{\begin{matrix} 0, & \ldots, & 0 \\ 0, & \ldots, & 0 \\ & \ddots, & \\ X_{ip0}, & \ldots, & X_{ipq_p} \end{matrix}}_{\text{marginal p}} \end{pmatrix} \tag{2.18}$$

such that the j-th row denotes covariates associated with the jth marginal model and the corresponding coefficient vector $\boldsymbol{\beta}$ be

$$\boldsymbol{\beta} = \{ \underbrace{\beta_{10}, \ldots, \beta_{1q_1}}_{\text{marginal 1}}, \underbrace{\beta_{20}, \ldots, \beta_{2q_2}}_{\text{marginal 2}}, \ldots, \underbrace{\beta_{p0}, \ldots, \beta_{pq_p}}_{\text{marginal p}} \}' \tag{2.19}$$

such that the first 1 to q_1+1 elements are coefficients for the 1st marginal model, the q_1+2 to q_1+q_2+1 elements are coefficients for the 2nd model and so on, until the last q_p+1 elements are coefficients associated with the the pth model, then we have

$$\boldsymbol{Y}_i \sim MVN(\overline{\boldsymbol{X}}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

The estimated values

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \overline{\boldsymbol{X}}_i'\overline{\boldsymbol{X}}_i\right)^{-1} \left(\sum_{i=1}^{n} \overline{\boldsymbol{X}}_i'\boldsymbol{Y}_i\right) \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \left\{(\boldsymbol{Y}_i - \overline{\boldsymbol{X}}_i\boldsymbol{\beta})(\boldsymbol{Y}_i - \overline{\boldsymbol{X}}_i\boldsymbol{\beta})'\right\}.$$

The score vectors are

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{\overline{\boldsymbol{X}}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \overline{\boldsymbol{X}}_i\boldsymbol{\beta})\right\};$$

and

$$S(\boldsymbol{\Sigma}) = -\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{i=1}^{n} \left\{(\boldsymbol{Y}_i - \overline{\boldsymbol{X}}_i\boldsymbol{\beta})(\boldsymbol{Y}_i - \overline{\boldsymbol{X}}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-2}\right\}.$$

The Information matrix and Fisher information matrix for $\boldsymbol{\beta}$ is given as

$$\mathcal{I}(\boldsymbol{\beta}) = I(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{\overline{\boldsymbol{X}}_i'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{X}}_i\right\}.$$

Proof of Theorem 1 is in Appendix B. In our case, if we let

$$\overline{\boldsymbol{W}}_i = \begin{pmatrix} X_{i10}, & \ldots, & X_{i1q_1}, & T_i, & 0, & \ldots, & 0 \\ & & \vdots & & \ddots & \vdots & \\ 0, & & \ldots, & & X_{iJ0}, & \ldots, & X_{iJq_J}, & T_i \end{pmatrix},$$

denotes the re-organized design matrix for the mediator model illustrated in Theorem 3, then we have

$$\hat{\boldsymbol{\Psi}}_M = \left(\sum_{i=1}^{n} \overline{\boldsymbol{W}}_i' \boldsymbol{\Sigma}^{-1} \overline{\boldsymbol{W}}_i \right)^{-1} \left(\sum_{i=1}^{n} \overline{\boldsymbol{W}}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{M}_i \right),$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ (\boldsymbol{M}_i - \overline{\boldsymbol{W}}_i \boldsymbol{\Psi}_M)(\boldsymbol{M}_i - \overline{\boldsymbol{W}}_i \boldsymbol{\Psi}_M)' \boldsymbol{\Sigma}^{-2} \right\},$$

And we have the score vectors:

$$S_M(\boldsymbol{\Psi}_M) = \sum_{i=1}^{n} S_{M,i}(\boldsymbol{\Psi}_M) = \sum_{i=1}^{n} \left\{ \overline{\boldsymbol{W}}_i' \boldsymbol{\Sigma}^{-1} (\boldsymbol{M}_i - \overline{\boldsymbol{W}}_i \boldsymbol{\Psi}_M) \right\};$$

as well as the Fisher information matrix

$$\mathcal{I}_M(\boldsymbol{\Psi}_M) = \sum_{i=1}^{n} \mathcal{I}_{M,i}(\boldsymbol{\Psi}_M) = \sum_{i=1}^{n} \left\{ \overline{\boldsymbol{W}}_i' \boldsymbol{\Sigma}^{-1} \overline{\boldsymbol{W}}_i \right\}.$$

2.5.2 Linear additive response model

Next, we let

$$\mu_{Y,i}(\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\Psi}_Y) = \boldsymbol{X}_i' \boldsymbol{\gamma} + \tau T_i + \boldsymbol{M}_{i\cdot}' \boldsymbol{\beta},\ i = 1, \ldots, n, \tag{2.20}$$

where τ is the unknown coefficient of the treatment assignment T_i, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_J\}'$ is the J-dimensional vector of regression coefficients on the vector of mediators and $\boldsymbol{\gamma} = \{\gamma_0, \ldots, \gamma_p\}'$ is the $p+1$-dimensional vector of regression coefficients on the vector of covariates including the intercept. Similarly, we denote the mean paramters in $\mathcal{M}_Y$ as $\boldsymbol{\Psi}_Y$ and in this case, $\boldsymbol{\Psi}_Y = \{\boldsymbol{\gamma}', \tau, \boldsymbol{\beta}'\}'$. We also assume $Var(Y_i | \boldsymbol{X}_i, \boldsymbol{M}_i, T_i) = \sigma_Y^2$. For this model, $\boldsymbol{\theta}_Y = \{\boldsymbol{\Psi}_Y', \sigma_Y\}'$.

Under this model setting, if we let $\boldsymbol{Z}_i = \{\boldsymbol{X}_i, T_i, \boldsymbol{M}_i\}'$ to denote the covariates used for $\mathcal{M}_Y$, then,

$$\hat{\boldsymbol{\Psi}}_Y = \left(\sum_{i=1}^{n} \boldsymbol{Z}_i' \boldsymbol{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i' Y_i \right) \quad \text{and} \quad \mathcal{I}_Y = \sum_{i=1}^{n} \{ \boldsymbol{Z}_i' \boldsymbol{Z}_i \}.$$

From (2.7),

$$
\begin{aligned}
& E\{Y(t_0, M_1(t_1), \ldots, M_J(t_J))\} \\
&= \int_{\boldsymbol{\mathcal{X}}} \left\{ \int \cdots \int_{\boldsymbol{m}} \mu_Y(\boldsymbol{x}, t_0, \boldsymbol{m}; \boldsymbol{\Psi}_Y) \, \mathrm{d}\, F(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{\theta}_M) \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_{\boldsymbol{\mathcal{X}}} \left\{ \int \cdots \int_{m_1,\ldots,m_J} (\boldsymbol{\gamma}'\boldsymbol{x} + \tau t_0 + \boldsymbol{\beta}'\boldsymbol{m}) \, \mathrm{d}\, F(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{\theta}_M) \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_{\boldsymbol{\mathcal{X}}} \left\{ \mathrm{E}_{\boldsymbol{M}}\big(\boldsymbol{\gamma}'\boldsymbol{x} + \tau t_0 + \boldsymbol{\beta}'\boldsymbol{M} \mid \boldsymbol{x}, \boldsymbol{t}\big) \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_{\boldsymbol{\mathcal{X}}} \left\{ \gamma_0 + \tau t_0 + \sum_{j=1}^{J} \beta_j \mathrm{E}\big(M_j|t_j, \boldsymbol{x}\big) + \boldsymbol{\gamma}'\boldsymbol{x} \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}) \\
&= \int_{\boldsymbol{\mathcal{X}}} \left\{ \gamma_0 + \tau t_0 + \sum_{j=1}^{J} \beta_j \big(\boldsymbol{\eta}_j'\boldsymbol{x} + \alpha_j t_j\big) + \boldsymbol{\gamma}'\boldsymbol{x} \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}).
\end{aligned}
$$

The third step is because of the linear additive property of expectation and in the last step, we plug in the mean value of the mediator from its marginal mean model.

It follows that, the individual indirect effect IE_j (defined in (2.5)), the joint indirect effect (defined in (2.3)) and the direct effect DE (defined in (2.2)), are shown to be:

$$
IE_j = \alpha_j\beta_j, \quad IE = \sum_{j=1}^{J} \alpha_j\beta_j, \quad DE = \tau.
$$

From the above formula, we see that, if we assume the multiple uncausally related mediators as causally related, then our method yields the same results on the individual indirect effects as Imai et al. [54] and VanderWeele et al. [139]. This alignment in outcomes is coincidental since the formulas for computing individual indirect effects are identical to formulas in Imai et al. [54] and VanderWeele et al. [139] (the product approach), despite our focus on uncausally related mediators while their work centers on causally related mediators. However, differences arise in defining the joint indirect effect and direct effect due to the distinct focal points of our method compared to the aforementioned works. Variances of estimated causal effects are calculated following (2.15). For example, $\widehat{IE}_j = \lambda_{IE_j}(\hat{\boldsymbol{\theta}}_{MY}) = \hat{\alpha}_j\hat{\beta}_j$, then

$$
\widehat{IE}_j = \hat{\alpha}_j\hat{\beta}_j \xrightarrow{d} N(\alpha_j\beta_j, \nabla'\lambda_{IE_j}(\hat{\boldsymbol{\theta}}_{MY})\boldsymbol{V}_{IE_j}\nabla\lambda_{IE_j}(\hat{\boldsymbol{\theta}}_{MY})),
$$

where $\nabla\lambda_{IE_j}(\hat{\boldsymbol{\theta}}_{MY}) = \{\hat{\beta}_j, \hat{\alpha}_j\}'$ and $\boldsymbol{V}_{IE_j}$ is calculated by extracting the corresponding elements from $\boldsymbol{V} = \mathrm{diag}(\mathcal{I}_M^{-1}, \mathcal{I}_Y^{-1})$.

2.5.3 Response model with mediator-treatment interactions

Suppose mediator-treatment interactions are included in the response model and (2.20) changes to

$$\mu_{Y,i} = \boldsymbol{X}_i'\boldsymbol{\gamma} + \tau_0 T_i + \boldsymbol{M}_i'\boldsymbol{\beta} + \sum_{j=1}^{J} \tau_j T_i M_{ij},$$

where the parameter τ_j reflects the mediator-treatment interactions. Under this scenario, we have,

$$IE_j = \alpha_j\beta_j + \alpha_j\tau_j, \quad IE = \sum_{j=1}^{J} \left(\alpha_j\beta_j + \alpha_j\tau_j\right),$$

and,

$$DE|\boldsymbol{X}_i = \tau_0 + \sum_{j=1}^{J} \tau_j \left(\alpha_j + \boldsymbol{X}_i'\boldsymbol{\eta}_j\right).$$

A detailed derivation can be found in Appendix C. We can see that when the response model includes mediator-treatment interactions, the direct effect varies with baseline covariates. Under such settings, the population-level effects can be treated as the expectation of conditional (individual-level) effects. It can also be understood as a double-expectation procedure, such that,

$$DE = E_{\boldsymbol{X}}\left\{\tau_0 + \sum_{j=1}^{J} \tau_j \left(\alpha_j + \boldsymbol{X}'\boldsymbol{\eta}_j\right)\right\}.$$

In practice, the population-level effects can be calculated using the sample mean, i.e.

$$\widehat{DE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\tau}_0 + \sum_{j=1}^{J} \hat{\tau}_j \left(\hat{\alpha}_j + \boldsymbol{X}_i'\hat{\boldsymbol{\eta}}_j\right)\right\}.$$

2.5.4 Response model with mediator-mediator interactions

Suppose mediator-mediator interactions are included in the response model to accommodate the uncausal correlations among mediators, where (2.20) changes to,

$$\mu_{Y,i} = \boldsymbol{X}_i'\boldsymbol{\gamma} + \tau T_i + \sum_{j=1}^{J} \beta_j M_{ij} + \sum_{\substack{k=1,\ldots,J \\ l=1,\ldots,J \\ k\leq l}} \beta_{k,l} M_{ik} M_{il},$$

where the parameters $\beta_{k,l}$ reflect the interactions among mediators.

Under this scenario, we have,

$$IE_j|T_i, \boldsymbol{X}_i = \alpha_j\beta_j + \alpha_j\left\{ \sum_{\substack{k=1,\ldots,J \\ k\neq j}} \beta_{j,k}\left(\alpha_k T_i + \boldsymbol{X}_i'\boldsymbol{\eta}_k\right) \right\},$$

and

$$\widehat{IE}_j = \sum_{i=1}^{n} \left\{ \hat{\alpha}_j\hat{\beta}_j + \hat{\alpha}_j\left\{ \sum_{\substack{k=1,\ldots,J \\ k\neq j}} \hat{\beta}_{j,k}\left(\hat{\alpha}_k T_i + \boldsymbol{X}_i'\hat{\boldsymbol{\eta}}_k\right) \right\} \right\}.$$

We can see from the results that the individual indirect effect for the jth mediator depends not only on the marginal distribution of the jth mediator itself but also on the margins of other mediators. A detailed derivation can be found in Appendix C. The joint indirect effect can be derived similarly. In addition, the direct effect can be shown to be τ as well. We do not further consider more complicated models.

2.6 Results under some non-normal settings

In this section, we consider some particular cases where either (both) the mediator or (and) the response are non-normally distributed. To accommodate the non-normality, the linear model assumptions for both the mediator model and the response model are replaced by the generalized linear model assumptions. To be specific, (2.16) is changed to

$$\mu_{ij}(\boldsymbol{X}_i, T_i; \boldsymbol{\Psi}_j) = g(\boldsymbol{X}_i'\boldsymbol{\eta}_j + \alpha_j T_i),\ i = 1, \ldots, n,\ j = 1, \ldots, J,$$

where g is a known link function. And (2.20) is replaced by

$$\mu_{Y,i}(\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\Psi}_Y) = h(\boldsymbol{X}_i'\boldsymbol{\gamma} + \tau T_i + \boldsymbol{M}_{i\cdot}'\boldsymbol{\beta}),\ i = 1, \ldots, n,$$

where h is a known link function. In the following, we introduce details under some specific model settings.

2.6.1 One mediator is log-normally distributed

In our simulation studies, we include a scenario where one of the mediators is assumed to be log-normally distributed and the other one is normally distributed. The log-normal distribution is a convenient and useful model used widely in engineering, sciences, medicine, as well as economics. For example, in biology, the measures of size of living tissue (length, skin area, weight) can be well approximated by log-normal distribution [53]. Here we provide analytical results.

We assume the distributions of the mediators are: $M_{i1}|T_i, \boldsymbol{X}_i \sim \text{Lognormal}(\mu_{i1}, \sigma_1^2)$ and $M_{i2}|T_i, \boldsymbol{X}_i \sim \text{Normal}(\mu_{i2}, \sigma_2^2)$. For M_{i2}, $E(M_{i2}|T_i, \boldsymbol{X}_i) = \mu_{i2}$ and $Var(M_{i2}|T_i, \boldsymbol{X}_i) = \sigma_2^2$. However, for M_{i1}, $E(M_{i1}|T_i, \boldsymbol{X}_i) \neq \mu_{i1}$ and $Var(M_{i1}|T_i, \boldsymbol{X}_i) \neq \sigma_1^2$. Instead,

$$E(M_{i1}|T_i, \boldsymbol{X}_i) = \exp(\mu_{i1} + \frac{\sigma_1^2}{2}) \quad \text{and} \quad Var(M_{i2}|T_i, \boldsymbol{X}_i) = \{\exp(\sigma_1^2) - 1\}\exp(2\mu_{i1} + \sigma_1^2).$$

Furthermore, for M_1, a log link is used and for M_2, an identity link is used. Therefore,

$$E(M_{i1}|T_i, \boldsymbol{X}_i) = \exp(\boldsymbol{X}_i'\boldsymbol{\eta}_1 + \alpha_1 T_i + \sigma_1^2/2) \quad \text{and} \quad \mu_{i2} = E(M_{i2}|T_i, \boldsymbol{X}_i) = \boldsymbol{X}_i'\boldsymbol{\eta}_2 + \alpha_2 T_i.$$

Further, we assume $Cor(M_{i1}, M_{i2}|T_i, \boldsymbol{X}_i) = \rho$.

For simplicity, the response model is assumed to be linear additive, such that,

$$\mu_{Y,i} = \boldsymbol{X}_i'\boldsymbol{\gamma} + \tau T_i + \beta_1 M_{i1} + \beta_2 M_{i2}.$$

Therefore, for the individual potential outcome (the effect conditioning on covariates $\boldsymbol{X}_i$), we have

$$\begin{aligned}
&Y_i(t_0, M_{i1}(t_1), M_{i2}(t_2)) \\
=&\iint_{m_1,m_2} (\tau t_0 + \beta_1 m_1 + \beta_2 m_2 + \boldsymbol{X}_i'\boldsymbol{\gamma})\ \mathrm{d}^2\ F_{M_{i1},M_{i2}}(m_1, m_2|t_1, t_2, \boldsymbol{X}_i) \\
=&\tau t_0 + \beta_1 \mathrm{E}\left(M_{i1}|t_1, \boldsymbol{X}_i\right) + \beta_2 \mathrm{E}\left(M_{i2}|t_2, \boldsymbol{X}_i\right) + \boldsymbol{X}_i'\boldsymbol{\gamma} \\
=&\tau t_0 + \beta_1 \exp(\alpha_1 t_1 + \boldsymbol{X}_i'\boldsymbol{\eta}_1 + \sigma_1^2/2) + \beta_2(\alpha_2 t_2 + \boldsymbol{X}_i'\boldsymbol{\eta}_2) + \boldsymbol{X}_i'\boldsymbol{\gamma}.
\end{aligned}$$

It follows that,

$$IE_{1i}|\boldsymbol{X}_i = \beta_1\{\exp(\boldsymbol{X}_i'\boldsymbol{\eta}_1 + \sigma_1^2/2)(\exp(\alpha_1) - 1)\}, \tag{2.21}$$

$$IE_{2i}|\boldsymbol{X}_i = \beta_2\alpha_2,$$

where $IE_{ji}|\boldsymbol{X}_i, j = 1, \ldots, J$ denotes the individual indirect effect conditioning on baseline covariates $\boldsymbol{X}_i$. The conditional joint indirect effect is the combination of the two, i.e.

$$IE_i|\boldsymbol{X}_i = \beta_1\{\exp(\boldsymbol{X}_i'\boldsymbol{\eta}_1 + \sigma_1^2/2)(\exp(\alpha_1) - 1)\} + \beta_2\alpha_2,$$

where $IE_i|\boldsymbol{X}_i$ denotes the joint indirect effect conditioning on baseline covariates $\boldsymbol{X}_i$. And the conditional direct effect,

$$DE_i|\boldsymbol{X}_i = \tau,$$

where similarly, $DE_i|\boldsymbol{X}_i$ denotes the direct effect conditioning on baseline covariates $\boldsymbol{X}_i$. In this case, the conditional direct effect equals the unconditional one.

Similarly, the population-level effects are the expectation of the individual-level effects, with expectation taken with respect to the distribution of the covariates (i.e. $\boldsymbol{X}$). One can also regard such a procedure as a double expectation procedure since the individual-level effects are the effects conditioning on covariates. The estimated causal effects of interest are obtained by replacing the parameters with their estimated values and calculating the sample mean, for example:

$$\widehat{IE}_1 = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\beta}_1\{\exp(\boldsymbol{X}_i'\hat{\boldsymbol{\eta}}_1 + \hat{\sigma}_1^2/2)(\exp(\hat{\alpha}_1) - 1)\}\right\},$$

$\widehat{IE}$ and $\widehat{DE}$ are calculated similarly.

When it comes to estimating the variances of the causal effects of interest, one needs to be cautious that, under these non-linear model settings, the variation of the covariates ($\boldsymbol{X}_i$) needs to be incorporated, because the effects usually involve covariates. The procedure of obtaining the variances is as follows: We first obtain the asymptotic variance-covariance matrix of parameters $\boldsymbol{V}$ following Theorem 2. Under the model settings, if we denote $\boldsymbol{\theta} = \{\boldsymbol{\theta}_M', \boldsymbol{\theta}_Y'\}'$, where $\boldsymbol{\theta}_M = \{\alpha_1, \boldsymbol{\eta}_1', \sigma_1, \alpha_2, \boldsymbol{\eta}_2', \sigma_2, \rho\}'$ and $\boldsymbol{\theta}_Y = \{\tau, \beta_1, \beta_2, \boldsymbol{\gamma}'\}'$, then,

$$\boldsymbol{V} = \text{diag}(\mathcal{I}_M^{-1}, \mathcal{I}_Y^{-1}).$$

Then, following (2.15), the conditional variances are approximated using the delta-method. For example, if we denote $\widehat{IE}_1|\boldsymbol{X} = \lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{X})$ to emphasize it is a function of both estimated parameters $\hat{\boldsymbol{\theta}}$ and baseline covariates $\boldsymbol{X}$, then,

$$\widehat{IE}_1|\boldsymbol{X} = \lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{X}) \xrightarrow{d} N\big(\lambda(\boldsymbol{\theta}, \boldsymbol{X}),\ \nabla'\lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{X})\,\boldsymbol{V}\,\nabla\lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{X})\big). \tag{2.22}$$

where, if only considering elements appear in the estimated effect,

$$\nabla\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X}) = \begin{pmatrix} \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\alpha_1} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\boldsymbol{\eta}_1} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\sigma_1} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\alpha_2} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\boldsymbol{\eta}_2} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\sigma_2} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\tau} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\beta_1} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\beta_2} \\ \frac{\partial\lambda(\hat{\boldsymbol{\theta}},\boldsymbol{X})}{\partial\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \beta_1\exp(\boldsymbol{\eta}_1'\boldsymbol{X}+\sigma_1^2/2)\exp(\alpha_1) \\ \beta_1\exp(\boldsymbol{\eta}_1'\boldsymbol{X}+\sigma_1^2/2)\{\exp(\alpha_1)-1\}\boldsymbol{X} \\ \beta_1\exp(\boldsymbol{\eta}_1'\boldsymbol{X}+\sigma_1^2/2)\{\exp(\alpha_1)-1\}\sigma_1 \\ 0 \\ \mathbf{0} \\ 0 \\ 0 \\ \exp(\boldsymbol{\eta}_1'\boldsymbol{X}+\sigma_1^2/2)\{\exp(\alpha_1)-1\} \\ 0 \\ \mathbf{0} \end{pmatrix}$$

In applications, we replace the variance-covariance matrix of the parameters $\boldsymbol{V}$ in the middle of the formula with the estimated $\widehat{\boldsymbol{V}}$.

For the unconditional variance, the final step is to utilize the law of total variance, such that,

$$Var(\widehat{IE}_1) = E\{Var(\widehat{IE}_1|\boldsymbol{X})\} + Var\{E(\widehat{IE}_1|\boldsymbol{X})\}.$$

In practice, without assuming the distribution of $\boldsymbol{X}$,

1. $E\{Var(\widehat{IE}_1|\boldsymbol{X})\}$ is obtained from caluating sample mean of $Var(\widehat{IE}_1|\boldsymbol{X}_i)$, with each of $Var(\widehat{IE}_1|\boldsymbol{X}_i)$ approximated by (2.22).

2. $Var\{E(\widehat{IE}_1|\boldsymbol{X})\}$ is obtained from taking sample variance of $E(\widehat{IE}_1|\boldsymbol{X}_i)$, with each of $E(\widehat{IE}_1|\boldsymbol{X}_i)$ approximated by the point estimate following (2.21).

A detailed derivation on the variance of the population-level causal effects of interest (with a focus on IE_1) is provided in Appendix C.2.

An alternative approach for variance estimation is to apply the multivariate Taylor's expansion to approximate the exponential function in order to get rid of the non-linearity. The results are provided in Appendix C.2. In scenario 3 of the simulation study (Section 2.6.3), a comparison between the results from direct calculation and approximation using the Taylor's expansion is provided.

2.6.2 The outcome is binary

We assume a logistic model for the outcome (for simplicity, we assume no mediator-treatment or mediator-mediator interactions), such that:

$$\log\left\{\frac{P\{Y_i=1|T_i,\boldsymbol{M}_i,\boldsymbol{X}_i\}}{1-P\{Y_i=1|T_i,\boldsymbol{M}_i,\boldsymbol{X}_i\}}\right\}=\boldsymbol{X}_i'\boldsymbol{\gamma}+\tau T_i+\boldsymbol{M}_i'\boldsymbol{\beta}.$$

Then the response probability from the logistic model can be formulated as:

$$P(Y_i=1|T_i,\boldsymbol{M}_i,\boldsymbol{X}_i)=\text{expit}\left\{\tau T_i+\boldsymbol{\beta}'\boldsymbol{M}_i+\boldsymbol{\gamma}'\boldsymbol{X}_i\right\}=\frac{\exp(\boldsymbol{X}_i'\boldsymbol{\gamma}+\tau T_i+\boldsymbol{M}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i'\boldsymbol{\gamma}+\tau T_i+\boldsymbol{M}_i'\boldsymbol{\beta})}.$$

The potential outcome can be shown as,

$$\begin{aligned}&E\{Y(t_0,M_1(t_1),\ldots,M_J(t_J))\}\\&=\int_{\mathcal{X}}\left\{\int\cdots\int_{\boldsymbol{m}}\left(\frac{\exp(\boldsymbol{X}'\boldsymbol{\gamma}+\tau t_0+\boldsymbol{m}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}'\boldsymbol{\gamma}+\tau t_0+\boldsymbol{m}'\boldsymbol{\beta})}\right)\,\mathrm{d}\,F_{\boldsymbol{M}}(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{t};\boldsymbol{\theta}_Y)\right\}\,\mathrm{d}\,F_{\boldsymbol{X}}(\boldsymbol{x})\end{aligned}\tag{2.23}$$

It is hard to derive an explicit formula for the integration. Therefore, under this scenario, we turn to the Monte Carlo method to numerically calculate the result, which is briefly summarized in Section 2.3.5. In addition, readers may refer to Imai et al. [55] for more details.

2.6.3 Some of the mediators are discrete

In this section, we describe a special setting where some mediators are discrete. We introduce a latent variable approach. Without loss of generality, we may assume that the mediator M_j is discrete and the others are continuous. Suppose M_j can take value 0 or 1, the latent variable approach assumes that there exists a latent continuous variable M_j^* that is associated with M_j and

$$M_j=\begin{cases}0, & \text{if } M_j^*\in(-\infty,s_j^*]\\1, & \text{if } M_j^*\in(s_j^*,\infty)\end{cases}$$

where s_j^* is the unknown threshold. When M_j takes more than two values, the proposed method can still be applied by re-coding M_j into a number of binary variables. In a more

complicated case where M_j takes value from $K+1$ ordered values, $s_{j0}, \ldots, s_{jK}$ cut points are required. The latent variable approach can still be applied, except that a total of K ordered thresholds are needed to model M_j, which is represented as,

$$M_j = \begin{cases} s_{j,0}, & \text{if } M_j^* \in (-\infty, s_{j,1}^*] \\ \vdots & \\ s_{j,k}, & \text{if } M_j^* \in (s_{j,k}^*, s_{j,k+1}^*] \\ \vdots & \\ s_{j,K}, & \text{if } M_j^* \in (s_{j,K}^*, \infty) \end{cases}$$

where $s_{j,1}^* < \cdots < s_{j,K}^*$ are unknown thresholds.

For simplicity, we consider the case of two mediators, M_1 and M_2, where M_1 is discrete taking values $s_1, \ldots, s_K$ and M_2 is continuous. The following can be extended to accommodate circumstances including more than two mediators. This is also the scenario we include in the simulation study (Section 2.7). M_1^* is assumed to be the latent continuous variable associated with M_1 with threshold values $s_1^*, \ldots, s_K^*$. Then, the joint distribution of M_1 and M_2 is given by:

$$\begin{aligned} &P(M_{i1} = m_1, M_{i2} \leq m_2) \\ &= \begin{cases} F_{M_{i1}^*, M_{i2}}(s_1^*, m_2) & \text{if } m_1 = s_1 \\ \vdots & \vdots \\ F_{M_{i1}^*, M_{i2}}(s_{k+1}^*, m_2) - F_{M_{i1}^*, M_{i2}}(s_k^*, m_2) & \text{if } m_1 = s_k \\ \vdots & \vdots \\ F_{M_{i2}}(m_2) - F_{M_{i1}^*, M_{i2}}(s_K^*, m_2) & \text{if } m_1 = s_K \end{cases} \end{aligned} \tag{2.24}$$

and the joint density is given by taking derivatives such that,

$$f_{M_{i1}, M_{i2}}(m_1, m_2) = \frac{\partial\, P(M_{i1} = m_1, M_{i2} \leq m_2)}{\partial\, m_2}. \tag{2.25}$$

We further assume that the joint distribution of M_1^* and M_2 is determined by a copula C, for individual i, the model for $F_{M_{i1}^*, M_{i2}}(m_1^*, m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_M)$ is denoted as

$$F_{M_{i1}^*, M_{i2}}(m_1^*, m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_M) = C(F_{M_{i1}^*}(m_1^* | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}),$$

Following (2.24),

$$
\begin{aligned}
&P(M_{i1} = m_1, M_{i2} \le m_2 | T_i, \boldsymbol{X}_i) \\
&= \begin{cases}
F_{M^*_{i1}, M_{i2}}(s^*_1, m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_M) & \text{if } m_1 = s_1 \\
\vdots & \vdots \\
F_{M^*_{i1}, M_{i2}}(s^*_{k+1}, m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_M) & \\
\quad - F_{M^*_{i1}, M_{i2}}(s^*_k, m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_M) & \text{if } m_1 = s_k \\
\vdots & \vdots \\
F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2) - F_{M^*_{i1}, M_{i2}}(s^*_K, m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_M) & \text{if } m_1 = s_K
\end{cases} \\
&= \begin{cases}
C(F_{M^*_{i1}}(s^*_1 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) & \text{if } m_1 = s_1 \\
\vdots & \vdots \\
C(F_{M^*_{i1}}(s^*_{k+1} | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) & \\
\quad - C(F_{M^*_{i1}}(s^*_k | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) & \text{if } m_1 = s_k \\
\vdots & \vdots \\
F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2) & \\
\quad - C(F_{M^*_{i1}}(s^*_K | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) & \text{if } m_1 = s_K
\end{cases}
\end{aligned}
$$

From (2.25),

$$
\begin{aligned}
&f_{M_{i1}, M_{i2}}(m_1, m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_M) \\
&= \begin{cases}
c(F_{M^*_{i1}}(s^*_1 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) & \\
\quad f_{M^*_{i1}}(s^*_1 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1) f_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2) & \text{if } m_1 = s_1 \\
\vdots & \vdots \\
\big\{ c(F_{M^*_{i1}}(s^*_{k+1} | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) f_{M^*_{i1}}(s^*_{k+1} | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1) & \\
\quad - c(F_{M^*_{i1}}(s^*_k | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) f_{M^*_{i1}}(s^*_k | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1) \big\} & \\
\quad f_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2) & \text{if } m_1 = s_k \\
\vdots & \vdots \\
\big\{ 1 - c(F_{M^*_{i1}}(s^*_K | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1), F_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2); \boldsymbol{\varrho}) f_{M^*_{i1}}(s^*_K | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_1) \big\} & \\
\quad f_{M_{i2}}(m_2 | T_i, \boldsymbol{X}_i; \boldsymbol{\theta}_2) & \text{if } m_1 = s_K
\end{cases}
\end{aligned}
$$

Conventionally, the marginal distribution for modeling the latent variable can be chosen as either the Normal distribution based on a probit copula model [12, 36], or logistic

distribution based on a logit copula model [88]. It is also possible to choose a (generalized) t-distribution [72] to model $F_{M_1^*}$ and that would induce a robit regression model for the binary outcome and therefore a robit copula model [77]. Moreover, the dependency parameter $\boldsymbol{\varrho}$ represents the dependencies between the latent variable M_1^* and the continuous variable M_2[17].

Maximum likelihood estimation approach can be used to estimate the parameters, such that,

$$\hat{\boldsymbol{\theta}}_M = \underset{\boldsymbol{\theta}_M}{\operatorname{argmax}} \sum_{i=1}^{n} \log f_{M_{i1},M_{i2}}(m_1, m_2|\boldsymbol{t}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_M).$$

The two-stage approach could also be used to simplify the estimation, but with a modified procedure. One may refer to Joe et al. [62, 63] for details. However, it should be noted that, using the two-stage approach will yield a different variance estimation. For details on estimating the copula joint distribution with the latent variable approach, one may refer to [17].

2.6.4 Under more general settings

If we do not specify any particular forms of the mediator model or the response model, we still need to turn to the Monte Carlo method for the estimation. Bootstrap can always be used to obtain the variance estimations.

2.7 Simulation study

We perform simulation studies to test the performance of the proposed method and make comparisons between the proposed method and the existing methods. We have three scenarios with 1000 replications for each ($N = 1000$).

We acknowledge that there are existing methods performing similar analyses. However, these methods impose different assumptions and perform the analysis under different mechanisms. Therefore, we cannot compare the results obtained from each one directly. Nevertheless, we report the results obtained from the other three methods (listed by the names of their R packages): package 'mediation'[132], 'mma'[157] and 'medflex'[125] for references. The package 'mediation' [132] is one of the most widely used packages for performing causal mediation analysis. This package conducts the analysis by fitting a set

of linear or nonlinear regressions to both the outcome and the mediators and calculates the causal effects from the estimated regression coefficients. For example, in a situation with two mediators, before applying the package, the user first needs to order them. We denote the ordered ones as M_1 and M_2. Then the package fits the following regressions sequentially: first

$$E\{M_1|\boldsymbol{X}, T\} = \boldsymbol{X}'\boldsymbol{\eta}_1 + \alpha_1 T,$$

followed by

$$E\{M_2|M_1, \boldsymbol{X}, T\} = \boldsymbol{X}'\boldsymbol{\eta}_2 + \alpha_1 T + \lambda M_1$$

and finally

$$E\{Y|\boldsymbol{X}, M_1, M_2, T\} = \boldsymbol{X}'\boldsymbol{\gamma} + \tau T + \beta_1 M_1 + \beta_2 M_2.$$

Results are obtained by investigating functions of the regression coefficients α, λ, τ and β. Confidence intervals are obtained by bootstrap for this package. This package also provides flexible ways to conduct sensitivity analysis. The major difference between the proposed method and package 'mediation' is that, package 'mediation' fits regression for each mediator one at a time while treating the other mediators as post-treatment confounders, while the proposed method models the joint distribution of the multiple mediators simultaneously.

The package 'mma' was designed specifically for multiple mediation analysis with details of the method described in Yu et al. [156]. This method improves traditional methods to enable the consideration of multiple mediators by fitting several linear or nonlinear mediator models simultaneously and uses linear or nonlinear predictive models for estimating mediation effects. Moreover, package 'mma' incorporates nonparametric algorithms like the spline functions or the 'LOESS' regression to fit each model. The major difference between the proposed method and package 'mma' is that, package 'mma' assumes a multivariate normal distribution on the mediators only. For example, Formula (2.26), (2.27) and (2.28) from Yu et al. [156] describe the three models to be fit by the package if there are two mediators. We adjust the three formulas with our notations for convenience:

$$M_{1i} = g_1(T_i, X_i) + \epsilon_{1i}; \tag{2.26}$$

$$M_{2i} = g_2(T_i, X_i) + \epsilon_{2i}; \tag{2.27}$$

and

$$Y_i = h(M_{1i}, M_{2i}, T_i, X_i) + \epsilon_3. \tag{2.28}$$

In the above three formulas from Yu et al. [156], $g_1(\cdot)$ and $g_2(\cdot)$ are non-parametric mediator models associating the mediators with the exposure and the covariates, while $h(\cdot)$ is a non-parametric response model that associates the response with the mediators, the exposure and the covariates. The default choice of nonparametric functions $g_1(\cdot), g_2(\cdot)$ and $h(\cdot)$ are spline functions. Additionally, in Yu et al. [156], $(\epsilon_{1i}, \epsilon_{2i})$ are assumed following a multivariate normal distribution such that

$$\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2, & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2, & \sigma_2^2 \end{pmatrix} \right),$$

and are independent with ϵ_{3i}, which are iid $N(0, \sigma^2)$, for $i = 1, \ldots, n$. Another difference between our proposed model and package 'mma' is that, package 'mma' does not fit the correlation model and ρ is treated as a nuisance parameter. Similar to package 'mediation', confidence intervals are obtained by bootstrap.

The package 'medflex' [125] is no longer available in CRAN currently. 'medflex' implements mediation analysis embedded within the natural effect models, which are a novel class of causal models that directly parameterize the path-specific effects of interest. The analysis framework of package 'medflex' is quite different from our proposed one and we provide a brief introduction to it for reference. Following Steen et al. [125], the second equation of Section 3 elaborates the model that package 'medflex' fits when there is one mediator, which is

$$E\{Y(x, M(x^*))|C\} = \beta_0 + \beta_1 x + \beta_2 x^* + \beta_3 C. \tag{2.29}$$

In (2.29), x and x^* are used to denote exposure (treatment) under either factual or counterfactual status, and C is used to denote baseline covariates. Though there is no explicit formula in Steen et al. [125] elaborating the models for multiple mediators under linear settings, from Equation (6) (the model for binary outcome), we can infer the model for multiple mediators to be as follows,

$$E\{Y(x, L(x^*), M(x^*; L(x^*)))|C\} = \beta_0 + \beta_1 x + \beta_2 x^* + \beta_3 C, \tag{2.30}$$

where L is used to denote the first mediator under exposure (treatment) x^* and M is used to denote the second mediator under exposure (treatment) x^* and the first mediator L evaluated at $L(x^*)$. Therefore, package 'medflex' assumes hierarchical orders of the

mediators (modeling M as a function of L and then Y as a function of both M and L), which is different from our proposed method. On top of that, since the potential outcomes $Y(x, M(x^*))$ and $Y(x, L(x^*), M(x^*; L(x^*)))$ are unobservable when $x \neq x^*$, (2.29) and (2.30) cannot be fitted directly with the observed data. Therefore, package 'medflex' uses either the imputation method that imputes such outcomes under counterfactual scenarios or the weighting approach that assigns each observation a weight and fit (2.29) and (2.30) on the weighted data. However, when there is only a single mediator, package 'medflex' offers both the weighting approach and the imputation approach to estimate the natural causal effects, but when multiple mediators exist, the package only allows the imputation approach. For details on the imputation or weighting approach, one may refer to Steen et al. [125]

We evaluate the performance of each approach in terms of the biases on both the point estimation and the variance estimation. The raw bias was calculated by taking the difference between the average estimated and the true value of a statistic. Suppose the true value of the statistic of interest is θ_0 and the estimated value for each replication i $(i = 1, \ldots, N)$ is $\hat{\theta}^{(i)}$, then the average estimated value of the statistic is $\hat{\theta} = 1/n \sum_{i=1}^{n} \hat{\theta}^{(i)}$, so Bias $= \hat{\theta} - \theta_0$. One may also calculate the relative bias that is defined as: Relative Bias $=$ $|\text{Bias}|/\theta_0$. These performance metrics tell us information on how accurate our proposed method as well as the existing method are. In terms of the variance estimation, we calculate the average estimated standard error from each replication. The SE is calculated from two approaches, via (2.15) and via bootstrap. We also calculated the empirical standard error of the estimated values, which is: Empirical SE $= \sqrt{1/(n-1) \sum_{i=1}^{n} (\hat{\theta}^{(i)} - \hat{\theta})^2}$. When assessing the accuracy of the variance estimation, one may compare the estimated standard error with the empirical standard error.

2.7.1 Scenario 1: 2 mediators distributed as Normal with different correlations

In this scenario, we consider 2 mediators that are both normally distributed but with different correlations. Our goal is to test the model performance under the situation that there are relatively strong, moderate, weak, and even no or negative correlations between the uncausally related mediators.

We consider a sample size of 200 pseudo-observations. For each simulated observation, we first generate three covariates $X_{i,1}, X_{i,2}$ and $X_{i,3} \overset{i.i.d}{\sim} N(0, 1)$. The treatment assignment for subject i is generated from a logistic model where the probability of receiving treatment $P(T_i = 1) = \text{expit}(0.3X_{i,1} + 0.3X_{i,2} + 0.3X_{i,3})$. After generating the

Scenario	estimated statistics	Point estimation		Variance			Coverage rate	
		average value	bias	empirical SE	SE (theoretical)	SE (bootstrap)	coverage (theoretical)	coverage (bootstrap)
Strong correlation, $\rho = 0.9$	IE of M1	1.005	0.005	0.222	0.221	0.228	0.923	0.926
	IE of M2	1.013	0.013	0.218	0.222	0.228	0.929	0.933
	Joint IE	2.018	0.018	0.286	0.293	0.293	0.937	0.941
	Total DE	0.993	-0.007	0.176	0.163	0.167	0.915	0.919
Moderate correlation, $\rho = 0.5$	IE of M1	0.996	-0.004	0.174	0.168	0.172	0.934	0.937
	IE of M2	0.995	-0.005	0.176	0.169	0.172	0.933	0.938
	Joint IE	1.991	-0.009	0.284	0.266	0.270	0.933	0.933
	Total DE	1.005	0.005	0.17	0.167	0.171	0.952	0.948
Weak correlation, $\rho = 0.1$	IE of M1	1	0	0.162	0.163	0.166	0.942	0.947
	IE of M2	0.992	-0.008	0.159	0.163	0.166	0.952	0.952
	Joint IE	1.992	-0.008	0.236	0.237	0.241	0.946	0.948
	Total DE	0.997	-0.003	0.178	0.174	0.179	0.949	0.946
No correlation, $\rho = 0$	IE of M1	1.001	0.001	0.167	0.164	0.167	0.939	0.944
	IE of M2	1.005	0.005	0.167	0.165	0.168	0.936	0.948
	Joint IE	2.006	0.006	0.233	0.232	0.237	0.949	0.945
	Total DE	0.992	-0.008	0.186	0.178	0.183	0.937	0.943
Negative correlation, $\rho = -0.5$	IE of M1	0.994	-0.006	0.168	0.169	0.172	0.940	0.943
	IE of M2	0.993	-0.007	0.164	0.169	0.172	0.950	0.951
	Joint IE	1.987	-0.013	0.203	0.205	0.211	0.949	0.950
	Total DE	1.005	0.005	0.204	0.204	0.211	0.949	0.957

Table 2.1: Performance of proposed method under scenario 1

covariates and treatment assignments, we proceed to the generation of multiple mediators. The mean and the variance across each mediator are set to be the same, such that $E(M_{i,1}) = E(M_{i,2}) = 0.5 + 1T_i + 0.3X_{i,1} + 0.3X_{i,2} + 0.3X_{i,3}$ and $Var(M_1) = Var(M_2) = 1^2$. With respect to the correlation between mediators ($\rho = Cor(M_{i,1}, M_{i,2})$), we set $\rho = 0.9, 0.5, 0.2, 0, -0.5$ to represent strong, moderate, weak, no and negative correlations. The two uncausally related mediators are generated using a bi-variate normal distribution. With the mediators generated, we proceed to the final step where the outcome is calculated as $Y_i = 0.5 + 1M_{i,1} + 1M_{i,2} + 1T_i + 0.3X_{i,1} + 0.3X_{i,2} + 0.3X_{i,3} + \varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$. The true values of the three types of effects are generated according to the definition and with true potential outcomes.

The simulation results for the proposed method are shown in Table 2.1, whereas the performances of existing methods can be accessed in Table 2.2 and 2.3.

package	scenarios	estimated statistics	Point estimation		Variance		Coverage rate
			average value	bias	empirical SE	SE	
package 'mediation'	Strong correlation, $\rho = 0.9$	IE of M1	1.010	0.010	0.224	0.228	0.945
		IE of M2	1.006	0.006	0.223	0.227	0.948
		Joint IE	2.016	0.016	0.290		
		Total DE	0.996	-0.004	0.173		
	Moderate correlation, $\rho = 0.5$	IE of M1	0.997	-0.003	0.173	0.17	0.936
		IE of M2	0.994	-0.006	0.172	0.171	0.943
		Joint IE	1.992	-0.008	0.280		
		Total DE	1.005	0.005	0.168		
	Weak correlation, $\rho = 0.1$	IE of M1	1	0	0.160	0.165	0.948
		IE of M2	0.992	-0.008	0.159	0.165	0.948
		Joint IE	1.993	-0.007	0.235		
		Total DE	0.998	-0.002	0.178		
	No correlation, $\rho = 0$	IE of M1	1	0	0.167	0.166	0.945
		IE of M2	1.006	0.006	0.167	0.167	0.942
		Joint IE	2.006	0.006	0.233		
		Total DE	0.991	-0.009	0.184		
	Negative correlation, $\rho =$ -0.5	IE of M1	0.994	-0.006	0.168	0.170	0.945
		IE of M2	0.992	-0.008	0.164	0.171	0.961
		Joint IE	1.986	-0.014	0.205		
		Total DE	1.007	0.007	0.202		

Table 2.2: Performance of package 'mediation' under scenario 1

package	scenario	estimated statistics	Point estimation		Variance		Coverage rate
			average value	bias	empirical SE	SE	
package 'mma'	Strong correlation, $\rho = 0.9$	IE of M1	1.232	0.232	0.255	0.269	0.897
		IE of M2	1.231	0.231	0.264	0.266	0.888
		Joint IE	2.457	0.457	0.326	0.337	0.737
		DE	0.995	-0.005	0.196	0.189	0.939
	Moderate correlation, $\rho = 0.5$	IE of M1	1.236	0.236	0.196	0.197	0.793
		IE of M2	1.232	0.232	0.198	0.198	0.800
		Joint IE	2.469	0.469	0.321	0.316	0.687
		DE	1.007	0.007	0.186	0.191	0.953
	Weak correlation, $\rho = 0.1$	IE of M1	1.211	0.211	0.182	0.188	0.829
		IE of M2	1.198	0.198	0.177	0.187	0.845
		Joint IE	2.411	0.411	0.279	0.287	0.716
		DE	0.995	-0.005	0.198	0.195	0.938
	No correlation, $\rho = 0$	IE of M1	1.286	0.286	0.191	0.193	0.701
		IE of M2	1.293	0.293	0.195	0.195	0.690
		Joint IE	2.578	0.578	0.290	0.299	0.510
		DE	0.986	-0.014	0.203	0.200	0.935
	Negative correlation, $\rho = -0.5$	IE of M1	1.233	0.233	0.193	0.198	0.804
		IE of M2	1.235	0.235	0.187	0.197	0.793
		Joint IE	2.465	0.465	0.267	0.279	0.618
		DE	1.007	0.007	0.210	0.220	0.950
package 'medflex'	Strong $\rho = 0.9$	Joint IE	2.016	0.016	0.290	0.162	0.720
		DE	0.995	-0.005	0.171	0.292	0.999
	Moderate $\rho = 0.5$	Joint IE	1.992	-0.008	0.28	0.166	0.742
		DE	1.005	0.005	0.168	0.264	0.998
	Weak $\rho = 0.1$	Joint IE	1.993	-0.007	0.235	0.174	0.849
		DE	0.998	-0.002	0.177	0.236	0.992
	No $\rho = 0$	Joint IE	2.007	0.007	0.233	0.177	0.845
		DE	0.990	-0.010	0.181	0.230	0.987
	Negative $\rho = -0.5$	Joint IE	1.987	-0.013	0.202	0.202	0.945
		DE	1.007	0.007	0.200	0.203	0.949

Table 2.3: Performance of package 'mma' and 'medflex' under scenario 1

2.7.2 Scenario 2: 2 mediators with 1 distributed as Normal and the other one as Lognormal

The second scenario is to test the performance of the proposed model when the mediators are not normally distributed. We design the study such that the two mediators follow different distributions to see if the proposed method can capture the changes precisely. For the proposed method, we specify the models according to their true distributions respectively. However, when performing the analysis using the other packages, we can only proceed by treating all mediators as normal since it is the only way supported. In other words, we are performing the analysis under incorrect model assumptions when investigating the other methods. We expect that, under this scenario, the proposed method achieves much lower bias and higher CI coverages since the proposed method takes advantage of capturing the true distributions in terms of model assumptions.

We fix the correlation to be $\rho = 0.5$ to avoid distractions. The data-generating procedures and the parameters are the same, except that, this time we assume the first mediator $(M_{i,1})$ to be distributed as a log-normal distribution.

The simulation results for the proposed method are shown in Table 2.4, whereas the results from other packages can be accessed in Table 2.5. Under this scenario, the package 'mediation' cannot provide variance estimations for the joint IE or DE and the package 'medflex' cannot provide estimations on individual IE. Therefore, the corresponding cells in the table are left blank.

We also conduct simulations where M_1 is treated as normally distributed and apply the proposed method. These simulations reflect common practice where, either unintentionally or due to lack of precise information, the distribution of the mediator(s) is assumed to be normal though in reality it is not. The results obtained under this scenario are presented in Table 2.6.

2.7.3 Scenario 3: 2 mediators distributed as Normal with binary outcome

This scenario is designed to test the performance of the proposed model when the response model is no longer linear. We assume a logistic model for the outcome. The data generating procedure is similar to scenario 1 except that, (1) we fix the correlation to be 0.5 to avoid distraction and (2) we generate the outcome values as $Y_i \sim Bernoulli(\pi_i)$ where $\pi_i = \text{expit}(0.5 + 1M_{i,1} + 1M_{i,2} + 1T_i + 0.3X_{i,1} + 0.3X_{i,2} + 0.3X_{i,3})$. In this scenario, the

scenario	estimated statistics	Point estimation		Variance			Coverage Rate	
		average value	bias	empirical SE	SE (theoretical)	SE (bootstrap)	coverage (theoretical)	coverage (bootstrap)
Monte Carlo	IE of M1	5.298	-0.002	0.937		0.971		0.957
	IE of M2	0.994	-0.006	0.198		0.191		0.939
	Joint IE	6.392	0.092	1.012		1.041		0.953
	Total DE	0.999	-0.001	0.166		0.167		0.952
theoretical (without approx)	IE of M1	5.326	0.026	0.435	0.423	0.434	0.942	0.951
	IE of M2	1	0	0.074	0.074	0.074	0.953	0.950
	Joint IE	6.326	0.026	0.469	0.455	0.465	0.939	0.951
	Total DE	1.001	0.001	0.070	0.073	0.074	0.955	0.949
theoretical (1st order approx)	IE of M1	4.644	-0.656	0.390	0.378	0.388	0.583	0.565
	IE of M2	1	0	0.074	0.074	0.074	0.953	0.970
	Joint IE	5.644	-0.656	0.426	0.412	0.420	0.620	0.885
	Total DE	1.001	0.001	0.070	0.073	0.074	0.955	0.945
theoretical (2nd order approx)	IE of M1	5.287	-0.013	0.432	0.420	0.432	0.941	0.950
	IE of M2	1	0	0.074	0.074	0.074	0.953	0.950
	Joint IE	6.287	-0.013	0.466	0.452	0.464	0.938	0.950
	Total DE	1.001	0.001	0.070	0.073	0.074	0.955	0.925

Table 2.4: Performance of the proposed method under scenario 2

package	estimated statistics	Point estimation		Variance		Coverage rate
		average value	bias	empirical SE	SE	
package 'mediation'	IE of M1	5.205	-0.095	1.212	1.250	0.903
	IE of M2	0.996	-0.004	0.174	0.169	0.943
	Joint IE	6.201	-0.100	1.270		
	Total DE	1.004	0.004	0.196		
package 'mma'	IE of M1	6.524	1.224	1.567	1.568	0.884
	IE of M2	1.238	0.238	0.199	0.193	0.785
	Joint IE	7.762	1.462	1.629	1.640	0.842
	Total DE	0.983	-0.017	0.407	0.415	0.951
package 'medflex'	Joint IE	6.104	-0.196	1.267	0.162	0.195
	Total DE	1	0	0.163	1.278	0.999

Table 2.5: Performance of other methods under scenario 2

Estimated statistics	Point estimation		Variance		Coverage rate
	average value	bias	empirical SE	estimated SE	
IE of M_1	5.109	-0.191	1.209	1.223	0.926
IE of M_2	0.995	-0.005	0.173	0.164	0.961
Joint IE	6.104	-0.196	1.267	1.278	0.885
Total DE	1.000	0	0.163	0.162	0.974

Table 2.6: Performance of the proposed method under scenario 2 when M_1 is treated as normally distributed

proposed method can provide estimations in either the probability scale or the odds-ratio scale. When considering the odds ratio scale, the effects are calculated by comparing $\log\left\{\frac{P[Y(t_0,M_1(t_1),\ldots,M_J(t_J))=1]}{1-P[Y(t_0,M_1(t_1),\ldots,M_J(t_J))=1]}\right\}$ under different combinations of $t_0,\ldots,t_J$. The calculation results under the odds-ratio scale are equivalent to the coefficients obtained from the logistic regression outcome model. Two reasons contribute to this phenomenon: 1. under the odds-ratio scale, the effects are obtained by contradicting linear predictors of the expected potential outcomes; 2. the mediator models are assumed to be linear and in each mediator model, the α coefficient is assumed to be 1. When considering calculations under the probability scale, the effects are calculated by taking the difference between the potential outcomes (defined in (2.23)) under different combinations of $t_0,\ldots,t_J$ directly.

With respect to other packages, the package 'mediation' allows for binary outcomes when the model only includes one mediator, and results are reported as risk differences. However, when considering multiple mediators, the package only supports those mediators to be causally related and models them with hierarchical linear models. Therefore, the package 'mediation' does not address our scenario. For the others, we proceed with the analysis following their respective model assumptions and it turns out that both packages 'mma' and 'medflex' provide results on the odds-ratio scale. The simulation results of the proposed method are shown in Table 2.7 while the results from other methods are provided in Table 2.8.

2.7.4 Discussion on simulation results

In the first scenario, we include two mediators that are normally distributed with different correlations. Under this scenario, both our proposed model and the other existing ones perform well with low bias and variation, as well as close to 95% CI coverage rates. This

Calculation method	estimated statistics	Point estimation		Variance		coverage rate
		average value	bias	empirical SE	SE (bootstrap)	
probability scale	IE of M1	0.094	-0.001	0.03	0.031	0.949
	IE of M2	0.070	0	0.025	0.026	0.940
	Joint IE	0.164	-0.001	0.033	0.035	0.942
	DE	0.083	-0.026	0.053	0.052	0.942
odds-ratio scale	IE of M1	1.099	0.099	0.397	0.528	0.952
	IE of M2	1.097	0.097	0.410	0.531	0.959
	Joint IE	2.197	0.197	0.568	0.785	0.955
	DE	0.995	-0.005	1.593	1.539	0.945

Table 2.7: Performance of the proposed method under scenario 3

package	estimated statistics	Point estimation		Variance		coverage rate
		average value	bias	empirical SE	SE (bootstrap)	
package 'mma'	IE of M1	1.470	0.470	0.523	0.634	0.963
	IE of M2	1.463	0.463	0.499	0.645	0.982
	Joint IE	2.884	0.884	0.748	0.936	0.917
	DE	1.029	0.029	1.535	1.858	0.941
package 'medflex'	Joint IE	1.485	-0.515	0.322	0.519	0.862
	DE	0.861	-0.139	1.075	0.312	0.653

Table 2.8: Performance of other methods under scenario 3 (under odds-ratio scale)

is as expected, because under such model settings, the effects of interest for all methods are framed in similar ways and they estimate them with similar approaches. However, in terms of variance estimation, our proposed method provides not only the ones obtained from bootstrap but also the ones calculated by formulas based on large sample properties, which are introduced in Sections 2.4 and 2.5, while the other packages only provide results obtained by bootstrap. Additionally, the results obtained from both methods are similar and are also similar to the empirical SE, which indicates that the calculations from the large sample properties illustrated in Section 2.4.2 are robust. if we take into consideration that calculations using the bootstrap method are much more time-consuming compared with calculations using large sample formulas, it is always recommended to calculate the SE via large sample formulas provided in Section 2.4.2. Such large sample variance calculation formulas are also major contributions of our proposed method as it does not require conducting the time-consuming bootstrap anymore, which is the conventional way of obtaining variance calculations of most other packages. Moreover, the proposed method can

provide variance estimations on the joint indirect effect, their individual ones, and on the direct effect, while the latter two cannot be obtained from package 'mediation'. Similarly package 'medflex' also only provides variance estimation on the joint indirect effect and the direct effect and is slightly less precise than the proposed method. Finally, when comparing with package 'mma', since 'mma' incorporates nonparametric algorithms like the spline functions for the model fitting, in simple linear cases, the bias is even larger. Additionally, the proposed method outperforms them in terms of precision when correlations among mediators exist.

In scenario 2, under the circumstance that one of the mediators is non-normally distributed, the proposed method provides estimations with small biases and high CI coverage rates. This is also expected since the proposed method is designed to utilize the flexibility of copula functions modeling non-normal distributions. For the other methods, because their estimations are based on incorrect model assumptions, the biases are much larger and the CI coverage rates are also different from expected. Futhermore, those results are obtained with knowledge of the mediator distributions. However, from Table 2.6, we see that without such knowledge of the mediator distributions, if we mistakenly treat M_1 as normally distributed, then the proposed method performs as poor as the other existing ones.

In scenario 3, the proposed method also provides desirable estimation performances, with low bias and close to 95% CI coverage rates in terms of both the probability and odds-ratio scale. Package 'mma' also provides comparable results, though it only provides estimations under odds-ratio scales. The performance of package 'medflex' is relatively poor. Though it provides relatively comparable point estimations with package 'mma', the CI coverage rates are far from 95%. We suspect it is because of the relatively more complicated estimation framework that induces more uncertainty.

2.8 Real data application

In psychological studies, experiencing childhood trauma is believed to play an important role in leading to psychiatric disorder [69]. However, the question of the mechanism of how persistent adverse effects of brain functioning are imposed by childhood trauma is still not well studied [49]. Recently, some researches suggest that DNA methylation is likely acting as a key factor, associated with both childhood trauma and long-term psychological disorder, especially adult stress reactivity and behavior [14, 148]. On one hand, long-term changes on brain stress reaction is believed to be partially caused by epigenetic alternations [101, 68, 85]. On the other hand, there are several studies revealing long-lasting effects

of environmental risk factors on DNA methylation changes [143]. However, most of them concentrate on a single gene [143] and the persistent effect of childhood trauma is unlikely the consequence of epigenetic changes of a single gene [83]. A preliminary study by Houtepen et al. [49] suggests that three loci in the Kit ligand gene (KITLG; cg27512205, cg05608730, cg26179948) have the strongest association between both exposure of childhood trauma and stress reactivity. However, Houtepen et al. [49] do not conduct joint mediation analysis on the three DNA methylation loci, but only investigate the effect of the single gene loci (cg27512205), which is suggested as having the strongest association.

In this section, following Houtepen et al. [49], we apply the proposed causal mediation analysis method on the same dataset, to identify the joint mediation effect of the three DNA methylation loci, as well as the individual indirect effect of each of them. The dataset consists of 85 healthy individuals recruited from the general population at the University Medical Center, Utrecht, the Netherlands [49]. In the dataset, only participants not taking any prescription medication and having not been enrolled in stress-related research before participation are included. For details on the dataset, one may refer to Houtepen et al. [49]. The stress procedure is performed by using a version of the Trier Social Stress Test (TSST) as a stress induction task. This task includes the public speaking test (PST) and arithmetic task [70]. The cortisol levels were measured with an in-house radioimmunoassay. In total eight saliva samples (Salivette) are collected within a 90-minute time period [144]. Measurements is calculated by the area under the curve (AUC) in terms of the increase (AUCi) of cortisol. Childhood trauma exposure was assessed using the short version of the Childhood Trauma Questionnaire (CTQ) [6, 129]. Genome-wide DNA methylation levels were measured by using Illumina Infinium HumanMethylation450K BeadChip (Illumina) arrays [123]. For details, please refer to Houtepen et al. [49]. In our analysis, age, sex, as well as the ethnicity of the subjects are also included as covariates, as suggested by researches [40, 48, 84]. Some descriptive statistics of the variables in our analysis can be found in Table 3.22.

In our analysis, a normal copula is chosen to jointly model the three gene loci as mediators. The residual histograms as well as normal Q-Q plots of the three mediators are shown in Figure 2.4. We see that normality assumptions are satisfied and therefore a normal copula is suitable. The correlation structure between the three mediators are shown in Table 2.10, which indicates weak to moderate pairwise correlations among all three gene loci.

The results of mediation analysis are shown in Table 2.11 (using bootstrap CI) and 2.12 (using theoretical CI). The three loci jointly has significant mediation effect on the causal pathway of childhood trauma to stress reactivity. The results with variances obtained via either bootstrap or theoretical derivations are similar and therefore, we use the results

characteristics	mean	range
sex (% of female)	50.589	
age (in years)	33.800	18 to 69
stress (cortisol stress reactivity, AUCi[1])	243.460	-1029.850 to 1876.280
trauma (total CTQ[2] score)	31.906	24 to 63
cg05608730 (methylation in percentage)	0.380	0.293 to 0.469
cg26179948 (methylation in percentage)	0.093	0.093 to 0.159
cg27512205 (methylation in percentage)	0.125	0.125 to 0.188

Table 2.9: Descriptive statistics of variables

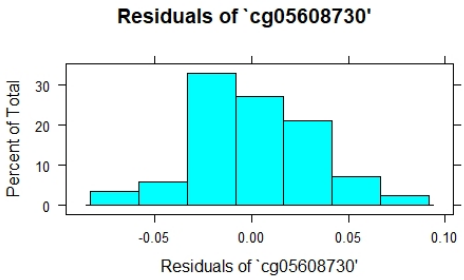
	cg05608730	cg26179948	cg27512205
cg05608730	1.000	0.219	0.322
cg26179948	0.219	1.000	0.425
cg27512205	0.322	0.425	1.000

Table 2.10: Pearson correlations between the three methylation loci

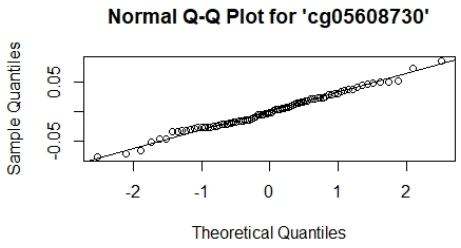
with theoretical variances for later discussions. The joint indirect effect from the three mediators is -9.167, with a 95% confidence interval of $(-15.925, -2.408)$. However, when considering the individual indirect effect through each of the mediators, none of them is identified as having significant mediation effect at 95% significance level. However, under 90% significance level, two loci ('cg05608730' and 'cg26179948') are shown to have strong mediation effect. In the end, the direct effect is estimated as -5.49, with a 95% confidence interval of $(-14.514, 3.535)$, which indicates non-significance at $\alpha = 0.05$.

	point estimation	SE theoretical	95%CI.LB. theoretical	95%CI.UB. theoretical	90%CI.LB. theoretical	90%CI.UB. theoretical
IE of: cg05608730	-3.512	1.968	-7.369	0.344	-6.739	-0.285
IE of: cg26179948	-3.190	1.906	-6.926	0.547	-6.316	-0.064
IE of: cg27512205	-2.464	1.718	-5.832	0.903	-5.282	0.354
joint IE	-9.167	3.448	-15.925	-2.408	-14.822	-3.512
DE	-5.490	4.604	-14.514	3.535	-13.04	2.061

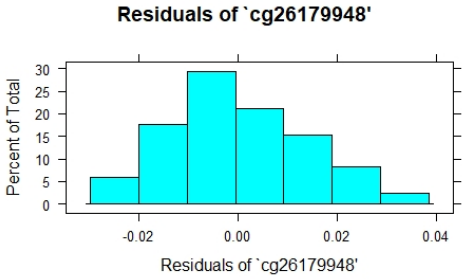
Table 2.11: Mediation effects of three methylation loci (using theoretical CI)



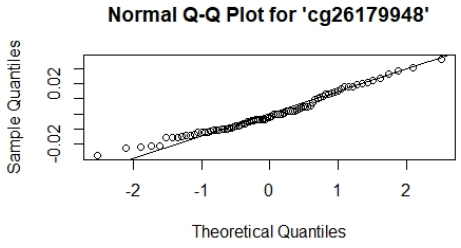
Histogram of ‘cg05608730’



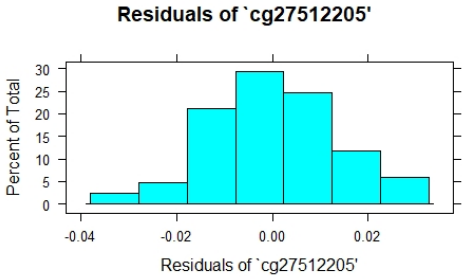
Q-Q plot of ‘cg05608730’



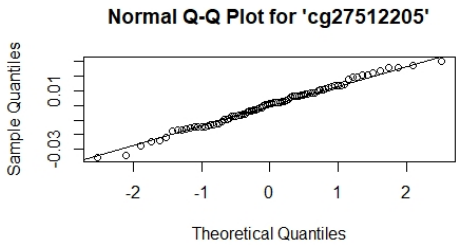
Histogram of ‘cg26179948’



Q-Q plot of ‘cg26179948’



Histogram of ‘cg27512205’



Q-Q plot of ‘cg27512205’

Figure 2.4: Residual histogram and normal Q-Q plot of the three methylation loci

	point estimation	SE bootstrap	95%CI.LB. bootstrap	95%CI.UB. bootstrap	90%CI.LB. bootstrap	90%CI.UB. bootstrap
IE of: cg05608730	-3.512	1.756	-6.954	-0.070	-6.392	-0.632
IE of: cg26179948	-3.190	1.829	-6.775	0.395	-6.190	-0.190
IE of: cg27512205	-2.464	1.783	-5.959	1.031	-5.389	0.460
joint IE	-9.167	3.865	-16.741	-1.592	-15.505	-2.829
DE	-5.490	5.207	-15.695	4.716	-14.029	3.050

Table 2.12: Mediation effects of three methylation loci (using bootstrap CI)

2.9 Discussion

In this chapter, we first clarify conceptual definitions of uncausally related mediators based on the idea of Jerolon et al. [60] We then introduce the causal effects and lay out the assumptions for estimating the causal effects. The main contribution of the work is that we propose a joint modeling method performing causal mediation analysis under the scenario that multiple uncausally correlated mediators are presented. The method utilizes copula framework to model the joint distribution of the multiple uncausally correlated mediators. The main advantages of our model are that: (1) our proposed model allows for large flexibility in terms of the distributions of either the mediators or the response; (2) our proposed method is able to estimate the individual indirect effects explicitly and (3) theoretical expressions include the point estimation and standard error can be given under particular settings.

Our proposed method also comes with certain limitations that warrant acknowledgment. The primary limitation pertains to the stringent assumption that the correlation between mediators remains constant across different treatment assignments. Real-world scenarios often involve more complex and dynamic relationships among variables, and this assumption may not always hold.

Furthermore, the definition of individual indirect effects in the proposed method is not entirely general. It relies heavily on the order in which mediators are considered, meaning that altering the order can yield different results. To enhance the robustness and applicability of our approach, future research could explore more flexible and generalized definitions of causal effects, as well as develop estimation methods that can handle a broader range of cases. This would be particularly valuable when dealing with multiple mediators that are interrelated, and their causal correlation structures are both uncertain and intricate.

Chapter 3

Mediator and dependency structure selection for high-dimensional causal mediation analysis

3.1 Introduction

In Chapter 2, we discussed causal mediation analysis when multiple mediators exist and introduced the concept of multiple uncausally related mediators. We also discussed estimations of causal effects of interest in such contexts. However, as introduced in Section 1.4, real-life problems are often more complicated and it is usually infeasible to conduct causal mediation analysis with the observed data directly. Proper mediator selection and dependency structure simplification are usually needed prior to conducting such analysis. In this chapter, we propose a penalization-based technique that selects both mediators from the high-dimensional candidates and their dependency structures. Estimations of causal effects and further analysis are then conducted on the selected data based on the learned data structure.

Figure 3.1 presents a DAG that illustrates the problem we are considering in this Chapter. Multiple candidate mediators are possibly high-dimensional. We use M and N to denote all the variables that are candidate mediators. Among them, only M are true mediators. Further, among the true mediators, some of them are uncausally related to each other while others are independent. We further suppose the uncausally related mediators form L groups, where within each group, mediators are uncausally related but across groups, they are independent. We denote $M_1^{(l)}, \ldots, M_{J_l}^{(l)}$ as uncausally related mediators

in group l, where $J_1, \ldots, J_L$ denote the group sizes. For the independent mediators, we denote them as $M_1, \ldots, M_J$, where J denotes the total number of independent mediators. Finally, $N_1^{(1)}, \ldots, N_{R_1}^{(1)}$ and $N_1^{(2)}, \ldots, N_{R_2}^{(2)}$ are not mediators because $N_1^{(1)}, \ldots, N_{R_1}^{(1)}$ do not affect Y though they are affected by T and on the contrary $N_1^{(2)}, \ldots, N_{R_2}^{(2)}$ are not affected by T though they affect Y.

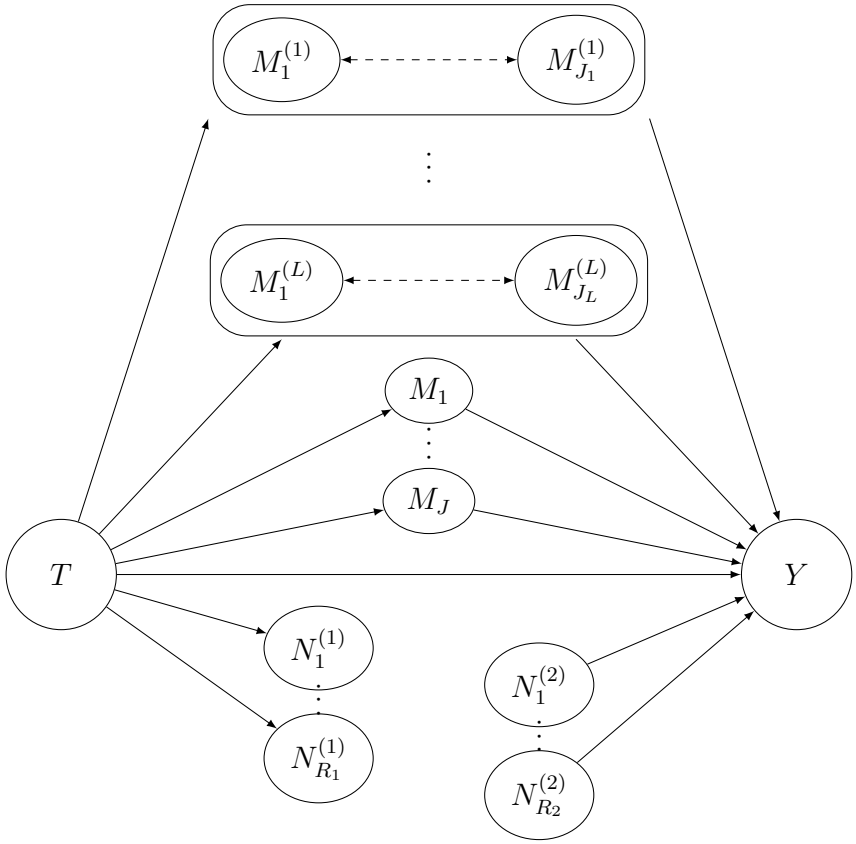


Figure 3.1: Mixture of uncausally related and independent mediators

Similar to Chapter 2, in the following illustration, a subscript i is often added to denote the subject i, where $i = 1, \ldots, n$ and n is the sample size. For example, $Y_i(t_0, M_1(t_1), \ldots, M_J(t_J))$ is used to denote the respective potential outcome for subject i. There are also occasions that both i and j appear in the subscript of the mediator or treatment variables, and the first subscript is used to denote the ith subject and the second one to denote the jth mediator. We assume the covariates have a dimension of p, so the notation x_{ir}, $r = 1, \ldots, p$ denotes the rth covariate of subject i. In addition, we denote the vector $\{M_{i1}, \ldots, M_{iJ}\}'$

as $\boldsymbol{M}_{i.}$, and the same rule applies to other variables with two subscripts. Similarly, we denote vector $\{M_{1j}, \ldots, M_{nj}\}'$ as $\boldsymbol{M}_{.j}$ and the same rule also applies to other variables with two subscripts. Readers should not be confused with the notations. The former denotes the values of all mediators within subject i and the latter denotes the values of the jth mediator across all subjects. Finally, a bold form variable without subscript is used to denote the entire data matrix.

3.2 Method

We propose a penalized variable selection technique that both selects the mediators from candidates and their dependency structures. After these two procedures, we proceed to the estimation of causal effects based on the learned causal structure of the data. Here we assume the mediators and the response are all continuous and our proposed method can be extended to deal with other cases with proper adjustments.

3.2.1 Model

Suppose there are J candidate mediators and we denote them as $\boldsymbol{M} = \{M_1, \ldots, M_J\}'$. We first introduce a marginal model for each potential mediator. We let $\mu_{ij} = E(M_{ij}|T_i, \boldsymbol{X_i})$, then the marginal model is written as follows:

$$\mu_{ij} = \boldsymbol{X}_i'\boldsymbol{\eta}_j + \alpha_j T_i,\ i = 1, \ldots, n;\ j = 1, \ldots, J, \tag{3.1}$$

where $\boldsymbol{X}_i = \{1, X_{i1}, \ldots, X_{iP}\}'$, α_j is the unknown coefficient of treatment, $\boldsymbol{\eta_j} = \{\eta_{j0}, \ldots, \eta_{jp}\}'$ is the $p+1$-dimensional vector of regression coefficients including the intercept. We can also allow treatment-covariates and between covariates interactions and higher order terms of the covariates in the model. We also specify the variance $Var(M_{ij}|T_i, \boldsymbol{X}_i) = \sigma_j^2$.

We then assume the candidate variables follow a multivariate normal distribution. To be specific, we assume $\boldsymbol{M}_{i.}|T_i, \boldsymbol{X}_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu_i} = \{\mu_{i1}, \ldots, \mu_{iJ}\}'$. The (k,l)th entry of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma}_{kl} = \rho_{kl}\sigma_k\sigma_l$ for $k \neq l$ and $\boldsymbol{\Sigma}_{kk} = \sigma_k^2$. The corresponding conditional correlation matrix is assumed to be $\boldsymbol{\varrho}$ and we denote $Cor(M_{ik}, M_{il}|T_i, \boldsymbol{X}_i) = \rho_{kl}$.

We then let $\mu_{Y,i} = E\{Y_i|T_i, \boldsymbol{M}_{i.}, \boldsymbol{X}_i\}$ and assume

$$\mu_{Y,i} = \boldsymbol{X}_i'\boldsymbol{\gamma} + \tau T_i + \boldsymbol{M}_{i.}'\boldsymbol{\beta}, \quad i = 1, \ldots, n, \tag{3.2}$$

where τ is the unknown coefficient of the treatment assignment T_i, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_J\}'$ is the J-dimensional vector of regression coefficients on the vector of candidate mediators and

$\boldsymbol{\gamma} = \{\gamma_0, \ldots, \gamma_p\}'$ is the $p+1$-dimensional vector of regression coefficients on the vector of covariates including the intercept. We also assume $Var(Y_i|\boldsymbol{X}_i, \boldsymbol{M}_{i.}, T_i) = \sigma_Y^2$. We further assume $Y_i|T_i, \boldsymbol{M}_{i.}, \boldsymbol{X}_i \sim N(\mu_{Y,i}, \sigma_Y^2)$

3.2.2 Mediator Selection

Though the mediators are possibly correlated, we first assume an independent working correlation among them. We impose a square loss and a penalization term to form an objective function. Denoting the loss function as $\mathcal{L}$, we have,

$$\mathcal{L}_M = \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{J}\left\{M_{ij} - \alpha_j T_i - \boldsymbol{X}_i'\boldsymbol{\eta}_j\right\}^2}_{(1)} + \underbrace{\sum_{i=1}^{n}\left\{Y_i - \tau T_i - \boldsymbol{\beta}'\boldsymbol{M}_{i.} - \boldsymbol{X}_i'\boldsymbol{\gamma}\right\}^2}_{(2)}. \tag{3.3}$$

In (3.3), term (1) calculates the squared error loss from the mediator model and term (2) calculates the squared error loss from the response model. We then add an L_1 penalized term to the loss function to form the objective function, which is to solve that,

$$\underset{\alpha_j,\beta_j,j=1,\ldots,J}{\operatorname{argmax}} \mathcal{L}_M/n - \lambda_1 \sum_{j=1}^{J}|\alpha_j\beta_j| - \lambda_2 \sum_{j=1}^{J}(|\alpha_j| + |\beta_j|). \tag{3.4}$$

In Equation (3.4), λ_1 and λ_2 are the penalization parameters that need to be tuned. The first penalization term penalizes the absolute value of the indirect effect, while the second term penalizes the absolute values of coefficients associated with each candidate mediator in both the mediator and the response models. The two penalization parameters, λ_1 and λ_2 control the relative magnitude of each penalization. The first penalization term is relatively more important, as it directly shrinks the indirect effect ($\alpha\beta$) for each candidate mediator, aligning with our primary goal of mediator selection. However, in certain scenarios, there are candidate mediators that exhibit a large value in one of α and β, while simultaneously having a very small value in the other. Such candidates are generally not considered as mediators, but employing a model with only the first penalization term may mistakenly select them as mediators. Therefore, in these cases, relying solely on the first penalization term is insufficient for precise mediator selection. The second penalization term serves to aid the selection process by shrinking one of the coefficients to zero. Subsequently, some values of α and β will be shrunk to 0. We then select variables associated with non-zero values of $\alpha\beta$ and regard them as mediators. Figure 3.2 illustrates the process of mediator selection, where red dashed lines represent zero effect ($|\alpha_2| = |\beta_3| = |\alpha_J| = |\beta_J| = 0$). In the figure, only M_1 will be selected as mediators following the proposed selection rules.

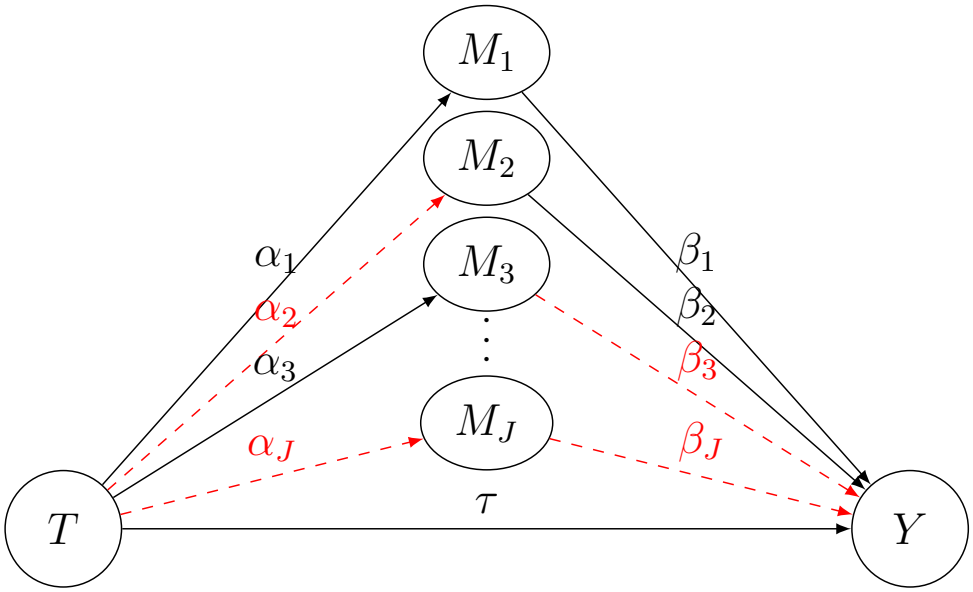


Figure 3.2: An illustration figure of the mediator selection process

3.2.3 Dependency structure selection

In this section, we present the method for selecting the dependency structure among the candidate mediators. We assess their dependency structures by analyzing their correlation matrices and choose to retain only the nonzero elements found in the off-diagonal entries of these matrices. There are two methods for handling the selection problem. The first approach is that we select off-diagonal entries from the correlation matrix of the candidate mediators and then extract a sub-matrix that only contains the elements related to the selected mediators obtained in Section 3.2.2. Since for this method, we select mediators and correlations from the same candidates and such two processes can be done independently without interfering with each other, we call such a method the parallel approach or the independent approach in some rare cases. Another approach is called the two-stage approach as it involves selecting mediators and dependency structures in two steps. The first stage is to select the mediators from the candidates. Then in the second stage, we select the correlations of the mediators obtained from stage one. In this section, we are going to introduce the first approach (the parallel approach), while the second approach (the two-stage approach) is going to be introduced in Section 3.2.5.

For the parallel approach, we first standardize the candidate mediators before making any selections. For each $j = 1, \ldots, J$, we estimate $\hat{\mu}_{ij}$ using model (3.1) and also $\hat{\sigma}_j$ using

the sample variance. Thereafter, we calculate the standardized values of M_{ij} (denoting as M^*_{ij}) by

$$M^*_{ij} = (M_{ij} - \hat{\mu}_{ij})/\hat{\sigma}_{ij}.$$

We denote $\boldsymbol{M}^* = \{M^*_1, \ldots, M^*_J\}$ with realization $\boldsymbol{m}^* = \{m^*_1, \ldots, m^*_J\}$. For subject i, we denote $\boldsymbol{m}^*_i = \{m^*_{i1}, \ldots, m^*_{iJ_t}\}$.

We calculate the log-likelihood of the standardized candidate variables as

$$l_{\boldsymbol{M}^*}(\boldsymbol{\varrho}; \boldsymbol{m}^*) = \sum_{i=1}^{n} \log\{f_{\boldsymbol{M}^*}(\boldsymbol{m}^*_i; \boldsymbol{\varrho})\} = \sum_{i=1}^{n} \log\{\phi(\boldsymbol{m}^*_i; \boldsymbol{\varrho})\}, \tag{3.5}$$

where $\phi(\cdot; \boldsymbol{\varrho})$ denotes the density of a multivariate normal distribution with marginal mean 0, marginal variance 1 and correlation structure $\boldsymbol{\varrho}$. Because of standardization, (3.8) is a function of $\boldsymbol{\varrho}$. We then add another L_1 penalized term and the objective function becomes,

$$\underset{\boldsymbol{\varrho}}{\text{argmax}}\ l_{\boldsymbol{M}^*}(\boldsymbol{\varrho}; \boldsymbol{m}^*)/n - \eta \sum_{k=1}^{J-1} \sum_{l=k+1}^{J} |\rho_{kl}|, \tag{3.6}$$

where similarly, η is the penalization parameter for this objective function and needs to be tuned. The penalization term will penalize some of ρ_{kl}'s to zero. We denote the selected correlation matrix as $\boldsymbol{\varrho}^s$, then the selected correlation matrix associated with the mediators is obtained by extracting the corresponding entries from $\boldsymbol{\varrho}^s$ according to the mediator selection results from Section 2.2. We take the non-zero correlation coefficients as an indication of correlated mediators and the zero ones as uncorrelated. Under the multivariate Gaussian settings, uncorrelated mediators can be regarded as independent ones. In addition, we group correlated mediators into several groups as shown by Figure 3.1. This is done by investigating the block-wise structure of the selected correlation matrix. After all the steps, we complete the selection of the dependency structure of the mediators.

3.2.4 Dependency structure selection on correlation matrices using standard LASSO technique

Solving (3.6) can be hard because the function is non-convex. To account for this issue, we propose a method that re-codes the data into a particular data frame where the regular

LASSO technique (variable selection under linear model settings) can be applied. The method works as follow: for each observation i and each pair of $\{k, l\} \in 1, \ldots, J$ where $k < l$, we generate $W_{ikl} = M^*_{ik} M^*_{il}$. We then let $\boldsymbol{Z} = \{Z_{12}, \ldots, Z_{J-1,J}\}'$ be a $J(J-1)/2 \times 1$ vector associated with each generated W_{ikl}, such that the entry $Z_{kl} = 1$ and the values for all the other entries are zero. We provide an example illustrating the process: we suppose M_1, M_2 and M_3 are 3 mediator candidates and we have 2 observations. The original data frame is shown in Table 3.1. The re-coded data frame after the aforementioned process is

	M_1	M_2	M_3
$i = 1$	m_{11}	m_{12}	m_{13}
$i = 2$	m_{21}	m_{22}	m_{23}

Table 3.1: Original data frame in illustrative example

shown in table 3.2. We then perform the standard LASSO technique on the re-coded data

	Z_{12}	Z_{13}	Z_{23}
W_{112}	1	0	0
W_{113}	0	1	0
W_{123}	0	0	1
W_{212}	1	0	0
W_{213}	0	1	0
W_{223}	0	0	1

Table 3.2: Re-coded data frame in illustrative example

frame. We assume

$$E(W_{ikl}) = \sum_{k=1}^{J-1} \sum_{l=k+1}^{J} \rho_{kl} Z_{kl}.$$

The objective function is to solve,

$$\underset{\boldsymbol{\varrho}}{\text{argmax}} \sum_{i=1}^{n} (W_{ikl} - \sum_{k=1}^{J-1} \sum_{l=k+1}^{J} \rho_{kl} Z_{kl})^2 / n - \phi \sum_{k=1}^{J-1} \sum_{l=k+1}^{J} |\rho_{kl}|, \tag{3.7}$$

where $\boldsymbol{\varrho} = \{\rho_{kl}; k, l \in \{1, \ldots, J\}, k < l\}$ and ϕ is the tuning parameter for this case.

Ideally, solving (3.7) would yield the same selection results as solving (3.6). This is because the correlation coefficient between two random variables can be transformed to

the expectation of their product, provided that the two variables have mean 0 and variance 1. For instance, considering two random variables X and Y with $E(X) = E(Y) = 0$ and $Var(X) = Var(Y) = 1$, by definition, we have

$$Cor(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(X)}} = E(XY).$$

Therefore, following our procedures, we obtain a random vector $\boldsymbol{M}^*$ that consists of random variables that all have mean 0 and variance 1, selecting the correlations $\rho_{k,l}$ among them is equivalent to picking up the non-zero pairwise products among them. With such a technique, we transfer the correlation selection problem to the variable selection issue on the mean model, which can be handled in more convenient ways using the standard LASSO technique and existing software packages.

Moreover, since the correlation coefficient always takes on values within $[-1.1]$, in practice, we may obtain estimated results outside the range when fitting the penalized model as shown in (3.7) under extreme cases. An ad hoc method tackling this issue is performing a transformation on the variable with range constraints. Any function mapping $[-1.1] \to \mathcal{R}$ satisfies our requirement. Some examples include $g(x) = 2\text{arctan}(x)/\pi$ or $g(x) = 2\Phi(x) - 1$, where $\Phi(\cdot)$ is the CDF function of a standard normal distribution.

3.2.5 Dependency structure selection in a two-stage approach

In this section, we introduce the dependency selection method in a two-stage approach. The two-stage approach reduces computational burden because we only need to focus on a relatively low-dimensional sub-matrix that reflects the correlations of the selected mediators, rather than the original correlation matrix generated by all candidates. Furthermore, modeling the correlation structure of only the selected variables would decrease the dimensionality of the parameters quadratically. However, the less computational burden of using a two-stage approach comes with a cost of less robustness. For the two-stage approach, if we do not select the mediators correctly in the first stage, then the dependency structure selection in the second stage can be adversely impacted.

The procedure is similar to the one introduced in Section 2.3 except that we working on the selected mediators from Section 2.2 only rather than all possible candidates. After mediator selection, we denote the selected mediators among all candidates as $\boldsymbol{M}^s = \{M_1^s, \ldots, M_{J_s}^s\}$, where J_s denotes the total number of selected mediators. Similarly, we standardize the selected mediators first by subtracting the estimated conditional mean and dividing by the estimated conditional variance. We denote the standardized mediators as

$\boldsymbol{M}^{s*} = \{M_1^{s*}, \ldots, M_{J_s}^{s*}\}$ with realization $\boldsymbol{m}^{s*} = \{m_1^{s*}, \ldots, m_{J_s}^{s*}\}$. Similarly, for subject i, we denote $\boldsymbol{m}_i^{s*} = \{m_{i1}^{s*}, \ldots, m_{iJ_s}^{s*}\}$. We calculate the log-likelihood of the standardized selected mediators as

$$l_{\boldsymbol{M}^{s*}}(\boldsymbol{\varrho}^s; \boldsymbol{m}^{s*}) = \sum_{i=1}^{n} \log\{f_{\boldsymbol{M}^{s*}}(\boldsymbol{m}_i^{s*}; \boldsymbol{\varrho}^s)\}. \tag{3.8}$$

We then select the correlation coefficients by solving

$$\underset{\boldsymbol{\varrho}^s}{\operatorname{argmax}}\ l_{\boldsymbol{M}^{s*}}(\boldsymbol{\varrho}^s; \boldsymbol{m}^{s*})/n - \eta \sum_{k,l \in \{1,\ldots,J_s\}; k<l} \sum |\rho_{kl}|. \tag{3.9}$$

Similarly, the penalization term will penalize some of ρ_{kl}'s to zero.

When choosing between the parallel and the two-stage approach, the trade-off is robustness versus computation burden. The advantage of the parallel approach is that, we have more robustness in terms of correlation selection because the correlation selection step does not depend on the results from the mediator selection step. However, the disadvantage is that it imposes more computational burden because of modeling the correlation of the high dimensional candidates. On the contrary, the two-stage approach requires less computation, but the results are less robust. The rule of thumb is that, if computational burden is not an issue, one should always go with the parallel approach.

3.2.6 Parameter tuning

A remaining issue of the proposed method is tuning the parameters that appeared in each model. We implement the cross-validation method for parameter tuning similar to the algorithm used by R package 'glmnet' [27].

The first step is to determine candidate values of tuning parameters. For the mediator selection process, there are two tuning parameters as shown in (3.4). We first re-code the two penalization parameters λ_1 and λ_2 as λ and ω, such that,

$$\lambda_1 = \lambda\omega \qquad \text{and} \qquad \lambda_2 = \lambda(1-\omega).$$

In the updated parametrization, the parameter λ indicates the overall strength of penalization while ω represents the relative importance of λ_1 and λ_2. For a larger ω, we impose stronger penalization on the indirect effect compared with the coefficient absolute values, while a smaller one indicates vice versa. We then obtain the candidate values for the parameters that form a tuning grid. For λ, we first find λ_{max}, which is the smallest possible

value of λ such that all the coefficients are shrunk to zero. λ_{max} acts as an upper bound for candidate values of λ. Then, with a user-specified value r_λ, which represents the ratio of the smallest candidate value of λ to λ_{max}, we specify the range of λ to be $(\lambda_{max} r_\lambda, \lambda_{max})$. A typical example of r_λ is 1/10000. Finally, with another user-specified number n_λ, which specifies the number of candidate values for λ, we create a length-n_λ sequence of candidate values for λ. In this sequence, the distances between the logarithm of each element are placed equally within the range $(\log(\lambda_{max} r_\lambda), \log(\lambda_{max}))$. For example, if $\lambda_{max} = 1$, $r_\lambda = 1/1000$ and $n_\lambda = 4$, then the sequence of candidates for λ is $\{0.001, 0.01, 0.1, 1\}$. The candidate values of the parameter ω are an equally distanced sequence between 0 and 1. For example, $\omega = \{0, 0.2, 0.4, \ldots, 1\}$. Each pair of candidate values λ and ω forms a tuning candidate that uniquely determines λ_1 and λ_2. Therefore the two candidate sets expand into a two-dimensional tuning grid for the mediator selection model. For tuning the parameter of the dependency structure selection, the procedures are similar. In both (3.6) and (3.9), we pick up the candidate values of η similar to the ways we pick up candidate values of λ.

Then, we begin the cross-validation process. Suppose a K-fold cross-validation method is used, we split the dataset into K fractions randomly. For each $k = 1, \ldots, K$, we regard the kth fraction of data as the validation data and the remaining $K - 1$ fractions as the training data. For each candidate value of the parameters, we fit the model on the training dataset (i.e. fitting objective function (3.4)), and validate the model performance on the validating data. In the validating step, we predict the results on the validation set with the fitted parameters and calculate the mean square error (MSE), as the performance measurement. In terms of determining the optimal value for the tuning parameters, we adopt the one standard error (1SE) criteria following the suggestion by Fang et al. [25]. We pick the values of the parameters such that under the particular values, the average validation MSE is within 1 standard error from the smallest one and the model is in the most parsimony form. With the determined tuning parameter, we fit the entire model with the whole dataset.

3.2.7 Estimation of causal effects

We notice that the complete data involves 4 components: the treatment (exposure), the mediator(s), the response and the covariate(s). If we assume the exposure and covariates are given and denote the complete data likelihood as $L(\boldsymbol{\theta}_{MY}; \boldsymbol{y}, \boldsymbol{m} | \boldsymbol{t}, \boldsymbol{x})$, we can factor the likelihood as

$$L(\boldsymbol{\theta}_{MY}; \boldsymbol{y}, \boldsymbol{m} | \boldsymbol{t}, \boldsymbol{x}) = L_Y(\boldsymbol{\theta}_Y; \boldsymbol{y} | \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}) \cdot L_{\boldsymbol{M}}(\boldsymbol{\theta}_M; \boldsymbol{m} | \boldsymbol{t}, \boldsymbol{x}), \tag{3.10}$$

where $\boldsymbol{\theta}_Y$ denotes the parameters for the response model, $\boldsymbol{\theta}_M$ denotes the parameters for the mediator model and $\boldsymbol{\theta}_{MY} = \{\boldsymbol{\theta}_M, \boldsymbol{\theta}_Y\}$ denotes the parameters used for the entire model. Under the multivariate normal setting, $\boldsymbol{\theta}_M = \{\boldsymbol{\alpha}', \boldsymbol{\eta}', \boldsymbol{\varrho}\}'$ and $\boldsymbol{\theta}_Y = \{\tau, \boldsymbol{\beta}', \boldsymbol{\gamma}'\}'$. We call the model for $L_Y(\boldsymbol{\theta}_Y; \boldsymbol{y}|\boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x})$ the response model and $L_{\boldsymbol{M}}(\boldsymbol{\theta}_M; \boldsymbol{m}|\boldsymbol{t}, \boldsymbol{x})$ the mediator model.

For mediators that fall within groups, we model their group-wise joint distributions. Suppose $M_1^{(l)}, \ldots, M_{J_l}^{(l)}$ are the mediators in group l, first the marginal model for mediator $M_j^{(l)}$ is fitted similarly to (3.1), such that,

$$\mu_{ij}^{(l)} = \boldsymbol{X}_i' \boldsymbol{\eta}_j^{(l)} + \alpha_j^{(l)} T_i, \ i = 1, \ldots, n; \ j = 1, \ldots, J_l, \tag{3.11}$$

The marginal standard deviation is then calculated from the residuals and they are denoted as $\sigma_j^{(l)}$. After that, we calculate the joint conditional density of mediators within group l following our multivariate normal assumption. If we denote the joint conditional density of mediators within group l as $f_{\boldsymbol{M}}^{(l)}(\boldsymbol{m}^{(l)}|\boldsymbol{X}, T)$, the corresponding conditional likelihood function is $L_{\boldsymbol{M}}^{(l)}(\boldsymbol{\theta}^{(l)}; \boldsymbol{m}^{(l)}|\boldsymbol{t}, \boldsymbol{x}) = \prod_{i=1}^n f_{\boldsymbol{M}}^{(l)}(\boldsymbol{m}_i^{(l)}|\boldsymbol{x}_i, t_i)$. $\boldsymbol{\theta}^{(l)}$ refers to parameters used in grouped mediator model l and $\boldsymbol{\theta}^{(l)\prime} = \{\alpha_j^{(l)}, \boldsymbol{\eta}_j^{(l)\prime}, j = 1, \ldots, J_l\}$.

For mediators that are determined not associated with any others, we fit their individual mediator models respectively. Suppose $M_j, j = 1, \ldots, J$ is a mediator that is not associated with any others, the mean model based on (3.1) is assumed to be

$$\mu_{ij} = \boldsymbol{X}_i' \boldsymbol{\eta}_j + \alpha_j T_i, \ i = 1, \ldots, n; \ j = 1, \ldots, J. \tag{3.12}$$

The standard deviation is then calculated from the residuals and they are denoted as σ_j. Following our assumption, the conditional density of M_j is denoted as $f_{\boldsymbol{M}_j}(\boldsymbol{m}_j|\boldsymbol{X}, \boldsymbol{T})$ and the corresponding conditional likelihood function is $L_{\boldsymbol{M}_j}(\boldsymbol{\theta}_j; \boldsymbol{m}_j|\boldsymbol{x}, \boldsymbol{t}) = \prod_{i=1}^n f_{\boldsymbol{M}_j}(\boldsymbol{m}_{ij}|\boldsymbol{x}_i, t_i)$, where $\boldsymbol{\theta}_j$ refers to parameters used in model for individual mediator j and $\boldsymbol{\theta}_j' = \{\alpha_j, \boldsymbol{\eta}_j'\}$. Combining the likelihood of mediators that fall within groups and the individual ones, we then have the mediator model likelihood shown as follows:

$$L_{\boldsymbol{M}}(\boldsymbol{\theta}_M; \boldsymbol{m}|\boldsymbol{t}, \boldsymbol{x}) = \prod_{l=1}^{L} L_{\boldsymbol{M}}^{(l)}(\boldsymbol{\theta}^{(l)}; \boldsymbol{m}^{(l)}|\boldsymbol{x}, \boldsymbol{t}) \prod_{j=1}^{J} L_{\boldsymbol{M}_j}(\boldsymbol{\theta}_j; \boldsymbol{m}_j|\boldsymbol{x}, \boldsymbol{t}),$$

where $\boldsymbol{\theta}_M' = \{\boldsymbol{\theta}^{(l)\prime}(l = 1 \ldots, L), \boldsymbol{\theta}_j'(j = 1, \ldots, J)\}$ denotes the parameters used in mediator model.

The response model is also updated to include only the selected mediators. The updated response model based on (3.13) is

$$\mu_{Y,i} = \boldsymbol{X}_i' \boldsymbol{\gamma} + \tau T_i + \boldsymbol{M}_{i\cdot}^{s'} \boldsymbol{\beta}^s, \quad i = 1, \ldots, n, \tag{3.13}$$

where $\boldsymbol{\beta}^{s\prime} = \{\boldsymbol{\beta}^{(l)\prime}(l = 1, \ldots, L), \boldsymbol{\beta}_j^{\prime}(j = 1, \ldots, J)\}$. $\boldsymbol{\beta}^{(l)}$ denotes the coefficients in the response model that correspond to the mediators in group l and $\boldsymbol{\beta}_j$ denotes the coefficient in the response model that corresponds to the individual mediator j. The conditional likelihood function for the response is $L_Y(\boldsymbol{\theta}_Y; \boldsymbol{y}|\boldsymbol{t}, \boldsymbol{m}, \boldsymbol{x}) = \prod_{i=1}^{n} f_Y(y_i|t_i, \boldsymbol{m}_i, \boldsymbol{x}_i)$, where $\boldsymbol{\theta}_Y^{\prime} = \{\tau, \boldsymbol{\beta}^{s\prime}, \boldsymbol{\gamma}^{\prime}\}$ denotes parameters appear in the response model. Following (3.10), we have the entire data conditional likelihood and parameters can be estimated through maximizing the logarithm of $L(\boldsymbol{\theta}_{MY}; \boldsymbol{y}, \boldsymbol{m}|\boldsymbol{t}, \boldsymbol{x})$.

In our settings, following [61], for mediators that are associated with each other, we group them into multiple groups and estimate group-wise joint indirect effects. For mediators that are not associated with any others, we estimate their individual indirect effects respectively. Furthermore, under the normality assumptions are that assumed throughout this paper, closed-form expressions can be attracted. Denote $IE^{(l)}$ as the grouped indirect effect of group l, then

$$IE^{(l)} = \sum_{j=1}^{J_l} \alpha_j^{(l)} \boldsymbol{\beta}_j^{(l)}.$$

And denote IE_j as the individual indirect effect of mediator j, then

$$IE_j = \alpha_j \boldsymbol{\beta}_j.$$

The direct effect (DE) is τ and the total effect (TE) is the summation of all grouped and individual indirect effects as well as the direct effect. Variances of the causal effects of interest can be estimated by treating them as functions of estimated parameters of the model and applying the delta-method on the asymptotic model parameter variance-covariance matrix. The asymptotic variance-covariance matrix of model parameters is derived by taking Fisher information of the joint model as shown in (3.10). It should be noticed that though the point estimations are the same for our proposed method and models assuming working independent or unspecified correlation structures among the mediators, the variance estimations are different. Our model yields a more precise estimation when the correlation structure among mediators is correctly specified.

Such an estimation procedure is similar to the one introduced in Chapter 2 except that we are now considering some of the mediators as grouped while others individually. For the grouped ones, we consider them uncausally related. In addition, in the next chapter, we are going to introduce multiple robust estimation methods. Researchers may also choose to use these multiple robust estimators for additional robustness. One needs to be cautious that, the variances and confidence intervals constructed by using the large

sample properties introduced in Chapter 2 with selected data only reflect the post-selection variations, and variations that occur due to the selection procedures are not incorporated in those confidence intervals.

3.3 Simulation study

To evaluate the performance of our proposed method in terms of mediator selection, correlation selection and mediators grouping, as well as estimation and inference in finite samples, we perform simulation studies of several scenarios.

3.3.1 Simulation setup

Throughout each scenario, we assume there are n subjects in the dataset and for each subject, we observe J candidate mediators. Furthermore, we denote M_{ij} as the jth candidate mediator for subject i and assume that $M_{ij}|t_i \sim N(\mu_{ij}, \sigma_j^2)$. We let $\mu_{ij} = E\{M_{ij}|t_i\}$ and assume that

$$\mu_{ij} = \alpha_j t_i,\ i = 1, \ldots, n;\ j = 1, \ldots, J. \tag{3.14}$$

σ_j is always assumed to be 1. In addition, across each scenario, for the response, we assume $Y_i|t_i, \boldsymbol{m}_{i\cdot} \sim N(\mu_{Y,i}, \sigma_Y^2)$ and let $\mu_{Y,i} = E\{Y_i|t_i, \boldsymbol{m}_{i\cdot}\}$. In addition, the response model across each scenario is assumed to be

$$\mu_{Y,i} = \tau t_i + \boldsymbol{m}_{i\cdot}'\boldsymbol{\beta}, \quad i = 1, \ldots, n, \tag{3.15}$$

and we always let $\tau = 1$ and $\sigma_Y = 1$. With respect to the correlation structure in the simulated dataset, we follow Zhao et al. [165] and generate a sparse correlation matrix to reflect the sparsity of the dependency structures among the mediators. If we have n mediators, there should be $n(n-1)/2$ pairs of correlations among them. However, to reflect the sparsity, we only let 20% of the off-diagonal entries of the correlation matrix be randomly chosen as nonzero. The values of the entries are chosen from $\{0, \rho\}$.

We implement a $2 \times 3 \times 3$ study design. The first "2" represents two different dimensions of the candidate mediators (J), which are specified to be 10 and 200 respectively. Since the sample size n is fixed at 100, the two scenarios represent either the "$n > p$" or the "$n < p$" case respectively. Under either scenario, only M_6, M_7 and M_8 are true mediators. To be specific, the models are specified as follows: under either scenario, for $j = 6, \ldots, 10$,

we let $\alpha_j = \alpha \neq 0$, and we let $\alpha_j = 0$ for all other values of j. With respect to β, we let $\beta_j = \beta \neq 0$ only for $j = 4, \ldots, 8$ and we let $\beta_j = 0$ for all the other values of j. In addition, under either scenario, only the first 10 candidate mediators are assumed to be correlated. The second "3" represents three different signal levels of the mean models, which are strong, moderate and weak respectively. Under the strong signal scenario, we specify that $\alpha = \beta = 1$, whereas for the moderate and weak signal scenario, we let $\alpha = \beta = 0.7$ and $\alpha = \beta = 0.4$ respectively. It should be noticed that under our model settings, the signal of indirect effects is the product of α and β. Table 3.3 illustrates the values for α_j and β_j under the 10-candidate mediators scenario, where the elements in blue are corresponding to real mediators. The third "3" represents three different signal levels of the correlations. We specify $\rho = 0.8, 0.5$ and 0.2 to reflect the strong, moderate and weak signals. Table 3.4 provides an example of a generated correlation matrix among the candidate mediators for the 10 candidate mediators scenario, where the sub-matrix in blue is the correlation matrix among the three real mediators. In addition, Table 3.5 presents the causal effects and Figure 3.3 illustrates the structure of the three mediators across all scenarios in the simulation study.

j	α_j	β_j
1	0	0
2	0	0
3	0	0
4	0	β
5	0	β
6	α	β
7	α	β
8	α	β
9	α	0
10	α	0

Table 3.3: An illustration of α and β under the $J = 10$ setting

For each simulation scenario, $N = 1000$ replications are performed. Under the "$n > p$" scenarios, both the two-stage and parallel approaches are evaluated. We report the performance of mediator selection, correlation selection and estimation of causal effects. For the mediator selection, we report the true positive (TP) and false positive (FP) number of mediators identified for each approach. For fairness of comparison, the corresponding rates (the TP and FP numbers divided by the true numbers) are also calculated. Particularly, when evaluating the performance of correlation selection using a two-stage approach, the

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	1	0	0	ρ	0	0	ρ	0	ρ	0
M2	0	1	0	0	0	ρ	0	0	0	0
M3	0	0	1	0	ρ	0	ρ	0	0	0
M4	ρ	0	0	1	0	0	0	0	0	ρ
M5	0	0	ρ	0	1	ρ	0	ρ	0	0
M6	0	ρ	0	0	ρ	1	ρ	0	0	0
M7	ρ	0	ρ	0	0	ρ	1	0	0	0
M8	0	0	0	0	ρ	0	0	1	0	0
M9	ρ	0	0	0	0	0	0	0	1	0
M10	0	0	0	ρ	0	0	0	0	0	1

Table 3.4: An illustration of the correlation matrix of the candidate mediators under the $J = 10$ scenario

causal effect	value
Grouped IE (of M_6 and M_7)	$2\alpha\beta$
Individual IE (of M_8)	$\alpha\beta$
IE of all other candidate mediators	0
DE	1
TE	$1 + 3\alpha\beta$

Table 3.5: Causal effects in simulation study across scenarios

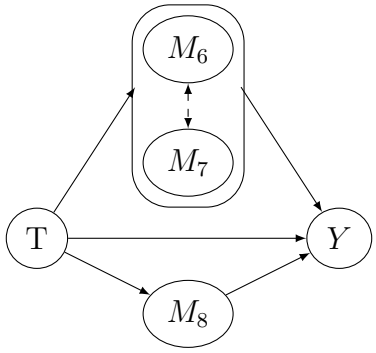


Figure 3.3: A DAG illustrating the causal relationship in the simulation study

denominator for TP and FP rates are different each time. This is because the correlation structure of the selected mediators varies with results obtained in the first stage of

mediator selection. Therefore, TP and FP numbers of the correlation selection under a two-stage approach are not reported because these values are meaningless. For the estimation results, we evaluate the bias and the empirical standard error. Additionally, for statistical inference purposes, we also calculate the standard error of the estimated causal effects using the method described in Section 2.6. The coverage rate of a 95% confidence interval is also evaluated using the asymptotic standard error.

We also compare the simulation results from the proposed method with several existing ones. The existing packages only perform mediator selection and do not perform the dependency structure selection procedure, so we only compare results on mediator selections. The package 'hdmed' was used when performing simulations with existing methods. This package contains a collection of various methods performing mediator selection from high-dimensional candidates and user may choose different methods that best fit their needs by specifying different commands from the package. The existing methods that are included in our simulation study include the de-biased LASSO approach proposed by Gao et al.[30], the minimax concave penalty approach proposed by Zhang et al. [161], and the adaptive LASSO approach proposed by Zhang [162]. The simulation parameters settings are the same when compared with existing methods. Similarly, for each simulation scenario, $N = 1000$ replications are performed. The tuning procedures are conducted automatically using functions embedded in each corresponding package.

3.3.2 Results

When $p < n$, across each scenario, in terms of both mediator selection and correlation selection, the proposed method shows high TP rates and low FP rates. Table 3.6 shows the performances of mediator selection across scenarios under the 10 candidate mediators setting. Generally, the proposed method performs better when the signal gets larger (values of α and β represent the signal for mediator selection). For the same values of α and β, the proposed method performs better in terms of mediator selection with a smaller value of ρ. In comparison, Table 3.7 presents the results of mediator selection for the three existing methods: the de-biased LASSO method, the minimax concave penalty method and the adaptive LASSO method under the $p = 10$ scenario. We can see that in terms of the true positive, all of the methods perform well and such phenomena are observed across different settings (different α, β and ρ). The performance of the proposed method is comparable to existing ones in terms of the true positive rates. In terms of false positive, when $p = 10$, all the methods do not perform well compared with the proposed one, with false positive rates greater than 0.3.

The mediator selection procedure is shared for both parallel and two-stage approaches. Table 3.8 shows the performances of correlation selection using the parallel approach under the 10 candidate mediators setting while Table 3.9 shows the performances of correlation selection using a two-stage approach. In Table 3.9, both the unconditional and the conditional results are presented, where the unconditional parts show the results of correlation selection in all 1000 replications regardless of mediator selection results, whereas the conditional parts present the results of correlation selection when only replications with correct mediator selection are considered. The "No." column (7th column) of Table 3.9 indicates the number of replications with correct mediator selection among the total 1000 simulation replications. Similar to the mediator selection results, the proposed method performs better when the signal gets larger, in terms of the correlation selection (values of ρ reflect the signal for correlation selection). When comparing conditional results with unconditional results, we see that the model performs better on correlation selection if the model correctly identifies the mediators.

In terms of causal effects estimations, Table 3.10 presents the performance of causal effects estimations using the parallel approach across each simulation replication regardless of mediator selection and correlation selection results (unconditional results), while Table 3.11 presents the results of causal effects estimations using the same parallel approach and conditioning on the replications with correct mediator selection results and Table 3.12 presents the results of causal effects estimations using the same parallel approach and considering the replications with both correct mediator selection and correct correlation selection results. Similarly, Table 3.13, 3.14 and 3.15 present the results of causal effects estimation using the two-stage approach under the unconditional, conditioning on correct mediator selection and conditioning on both correct mediator selection and correct correlation selection situations. For each table, "ESE" refers to the empirical standard error (i.e. square root of the empirical variance of the estimated causal quantities across each replication) while "ASE" refers to the average estimated standard error (i.e. sample average of the estimated standard error for each simulated sample). We see that the biases become smaller and the coverage rates become closer to 95% each time when we narrow our focus to the conditional results. For example, in the scenario that $\alpha = \beta = 1$ and $\rho = 0.5$, the unconditional performances are summarized from all 1000 replications. However, among the 1000 replications, for 932 of them, the proposed method correctly identified the three mediators M_6, M_7 and M_8. If we only consider these 932 replications that correctly select the three mediators, performances improve. Furthermore, among the 932 replications, 928 of them both identify mediators and select correlations correctly. If we further narrow our scope to these 928 replications only, then the performances improve further. Such a phenomenon is observed across both the parallel and the two-stage approach.

mean model	correlation	TP	TP rate	FP	FP rate
$\alpha = \beta = 1$	$\rho = 0.8$	3.000	1.000	0.055	0.008
	$\rho = 0.5$	3.000	1.000	0.066	0.009
	$\rho = 0.2$	3.000	1.000	0.126	0.018
$\alpha = \beta = 0.7$	$\rho = 0.8$	2.992	0.997	0.133	0.019
	$\rho = 0.5$	2.990	0.997	0.141	0.020
	$\rho = 0.2$	2.995	0.998	0.198	0.028
$\alpha = \beta = 0.4$	$\rho = 0.8$	2.718	0.906	0.414	0.059
	$\rho = 0.5$	2.712	0.904	0.462	0.066
	$\rho = 0.2$	2.710	0.903	0.455	0.065

Table 3.6: Results of mediator selection under the $J = 10$ scenario

When considering the $p > n$ settings, Table 3.18 shows the performances of mediator selection across scenarios under the 200 candidate mediators setting. The results obtained are similar to the ones observed under the $J = 10$ scenarios, with high true positive rates and low false positive rates. The proposed method performs better when the signal gets larger and for the same values of α and β, the proposed method performs better in terms of mediator selection with a smaller value of ρ. In contrast, Table 3.17 presents the results of mediator selection for the three existing methods under the $p = 200$ scenarios. Similar trends are observed for the $p = 200$ settings as well. We can see that in terms of the true positive, all of the methods perform well and such phenomena are observed across different settings (different α, β, ρ and p). The performance of the proposed method is comparable to existing ones in terms of the true positive rates. In terms of false positive, however, under the $p > n(p = 200)$ scenario, the false positive rates drop to below 0.05. We believe it is the tunning logic embedded in each method contributes to such phenomena. When it comes to the trade-off between false positives and false negatives, all the existing methods adopt the logic that it is worse to omit a true variable than mistakenly select a redundant one. Therefore, when p is relatively low, the existing methods tend to select redundant variables. Apart from that, the selection performances get better when the signals get larger, and for the same values of α and β, each method performs better in terms of mediator selection with a smaller value of ρ. Table 3.18 shows the performances of correlation selection under the 200 candidate mediators setting. Under the settings with 200 candidate mediators, only the two-stage approach is used for correlation selection. This is because the parallel approach involves performing selection on the 200×200 correlation matrix of the candidates, which turns out to be infeasible due to the large computational burden. Similarly, results presented in Table 3.18 consist of two parts: the unconditional

method	scenario		TP	TP rate	FP	FP rate
The de-biased LASSO	$\alpha = \beta = 1$	$\rho = 0.8$	3.000	1.000	2.150	0.307
		$\rho = 0.5$	3.000	1.000	2.160	0.309
		$\rho = 0.2$	3.000	1.000	2.240	0.320
	$\alpha = \beta = 0.7$	$\rho = 0.8$	2.970	0.990	2.200	0.314
		$\rho = 0.5$	3.000	1.000	2.230	0.319
		$\rho = 0.2$	3.000	1.000	2.270	0.324
	$\alpha = \beta = 0.4$	$\rho = 0.8$	2.120	0.707	1.700	0.243
		$\rho = 0.5$	2.890	0.963	1.960	0.280
		$\rho = 0.2$	2.950	0.983	2.130	0.304
The minimax concave penalty	$\alpha = \beta = 1$	$\rho = 0.8$	3.000	1.000	2.090	0.299
		$\rho = 0.5$	3.000	1.000	2.170	0.310
		$\rho = 0.2$	3.000	1.000	2.150	0.307
	$\alpha = \beta = 0.7$	$\rho = 0.8$	2.930	0.977	2.290	0.327
		$\rho = 0.5$	3.000	1.000	2.230	0.319
		$\rho = 0.2$	3.000	1.000	2.260	0.323
	$\alpha = \beta = 0.4$	$\rho = 0.8$	2.100	0.700	1.690	0.241
		$\rho = 0.5$	2.880	0.960	2.300	0.329
		$\rho = 0.2$	2.940	0.980	2.430	0.347
The adaptive LASSO	$\alpha = \beta = 1$	$\rho = 0.8$	3.000	1.000	2.500	0.357
		$\rho = 0.5$	3.000	1.000	2.500	0.357
		$\rho = 0.2$	3.000	1.000	2.430	0.347
	$\alpha = \beta = 0.7$	$\rho = 0.8$	2.980	0.993	2.740	0.391
		$\rho = 0.5$	3.000	1.000	2.740	0.391
		$\rho = 0.2$	3.000	1.000	2.700	0.386
	$\alpha = \beta = 0.4$	$\rho = 0.8$	2.720	0.907	2.740	0.391
		$\rho = 0.5$	2.960	0.987	2.750	0.393
		$\rho = 0.2$	2.960	0.987	2.810	0.401

Table 3.7: Performance of existing methods on mediator selections under $p = 10$ scenario

ones (results on correlation selection regardless of mediator selection) and the conditional ones (results on correlation selection if only focusing on the ones with correct mediator selection). Furthermore, the model performs better on correlation selection if provided that the model correctly identifies the mediators.

In terms of estimation of causal effects and statistical inferences, Table 3.19 presents the results of estimations and statistical inferences of causal effects of interest. Table 3.20

mean model	correlation	TP	TP rate	FP	FP rate
$\alpha = \beta = 1$	$\rho = 0.8$	9.000	1.000	0.000	0.000
	$\rho = 0.5$	9.000	1.000	0.065	0.002
	$\rho = 0.2$	7.541	0.838	3.416	0.098
$\alpha = \beta = 0.7$	$\rho = 0.8$	9.000	1.000	0.000	0.000
	$\rho = 0.5$	9.000	1.000	0.051	0.001
	$\rho = 0.2$	7.342	0.816	3.142	0.090
$\alpha = \beta = 0.4$	$\rho = 0.8$	9.000	1.000	0.002	0.000
	$\rho = 0.5$	9.000	1.000	0.057	0.002
	$\rho = 0.2$	7.495	0.833	3.114	0.089

Table 3.8: Results of dependency structure selection using the parallel approach under the $J = 10$ scenario

Settings		Unconditional		Conditional		
mean model	correlation	TP rate	FP rate	TP rate	FP rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	1.000	0.000	1.000	0.000	949
	$\rho = 0.5$	1.000	0.017	1.000	0.017	936
	$\rho = 0.2$	0.858	0.193	0.873	0.201	882
$\alpha = \beta = 0.7$	$\rho = 0.8$	1.000	0.003	1.000	0.001	886
	$\rho = 0.5$	1.000	0.014	1.000	0.014	867
	$\rho = 0.2$	0.817	0.191	0.839	0.198	819
$\alpha = \beta = 0.4$	$\rho = 0.8$	1.000	0.029	1.000	0.002	567
	$\rho = 0.5$	1.000	0.062	1.000	0.006	486
	$\rho = 0.2$	0.816	0.190	0.841	0.189	466

Table 3.9: Results of dependency structure selection using the two-stage approach under $J = 10$ scenario

presents the performances of causal effects estimations calculated from replications with correct mediator selection results while Table 3.21 presents the performances of causal effects estimations calculated based on replications with both correct mediator selection and correct correlation selection. Similarly, the last column records the number of replications with correct mediator selection only and both correct mediator selection and correct correlation selection. We see that the proposed method performs poorly when considering the general (unconditional) results of causal effects estimations. That is due to the inclusion of replications with incorrect mediator selection and (or) correlation selection. If we only consider replications with correct mediator selection and both correct mediator selection

and correct correlation selection, the performances improve under each case. Other results are similar to the ones obtained under $J = 10$ scenarios.

3.4 Apply the proposed method to a psychological study

In psychological studies, experiencing childhood trauma is believed to be a key factor that leads to long-term psychiatric disorder [69]. However, the underlying mechanism of persistent adverse effects of brain functioning induced by childhood trauma is still under investigation [49]. Some researchers suggest that DNA methylation is likely mediating the process between childhood trauma and long-term psychological disorder, especially adult stress reactivity and behavior [14].

Epigenetic changes may happen at multiple gene loci [83] and one of the challenges faced by researchers is the high dimensionality of genetic data. Some preliminary studies identify a few gene loci that may have strong associations between both exposure of childhood trauma and long-term stress reactivity [49], but to the best of our knowledge, few studies perform variable selection and correlation reduction on gene loci data under causal mediation analysis framework.

In this section, we apply the proposed method on the dataset used by Houtepen et al. [49], to identify causal mediation effects of DNA methylation loci. The dataset consists of 85 healthy individuals recruited at the University Medical Center, Utrecht, the Netherlands. In the dataset, only participants not taking any prescription medication and having not been enrolled in stress-related research before participation are included. The stress procedure is performed by using a version of the Trier Social Stress Test (TSST) as a stress induction task, which includes the public speaking test (PST) and arithmetic task [70]. The cortisol levels were measured with an in-house radioimmunoassay. In total eight saliva samples (Salivette) are collected within 90 minute time period [144]. Measurements are calculated by the area under the curve (AUC) in terms of the increase (AUCi) of cortisol. Childhood trauma exposure was assessed using the short version of the Childhood Trauma Questionnaire (CTQ) [6]. Genome-wide DNA methylation levels were measured by using Illumina Infinium HumanMethylation450K BeadChip (Illumina) arrays [123]. For details on the dataset, one may refer to Houtepen et al. [49]. In addition, age and sex of the subjects are also included as covariates as suggested by Heim et al. [40]. Some descriptive statistics of the variables in our analysis can be found in Table 3.22.

The original dataset consists of 385882 methylation variables in total, which makes us

infeasible to proceed with the original dataset directly due to the ultra-high dimension. In order to scale down the computational burden, we first apply sure independence screening (Fan et al. [24]) to reduce the dimension to a reasonable scale, before applying the proposed method. Following Perara et al. [94] and Cai et al. [10], we select a total number of m candidate mediators with the largest values of $\alpha_j\beta_j$ from the following model:

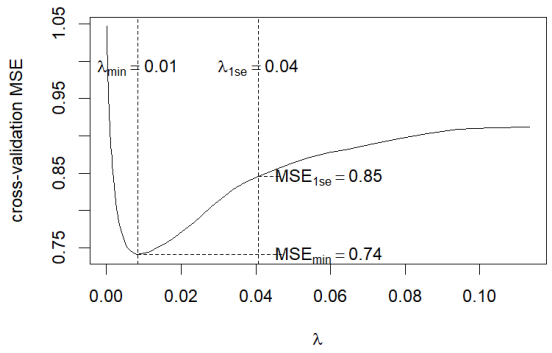
$$M_j = \boldsymbol{X}'\boldsymbol{\gamma}_j + \alpha_j T + \varepsilon_j, \text{ for } j = 1, \ldots, 385882$$

$$Y = \boldsymbol{X}'\boldsymbol{\eta} + \tau T + \boldsymbol{M}'\boldsymbol{\beta} + \epsilon;$$

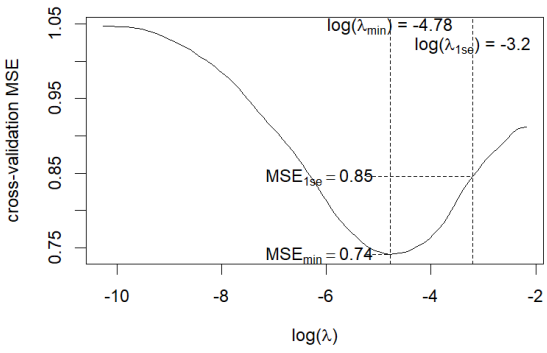
When it comes to determining m, there is a trade-off between estimation precision and computational burden. The more variables we include, the lower the chances we omit an important variable, but on the contrary, the higher the computational burden we have to face. Different scholars propose different strategies for determining m. We specify two different values of m and apply the proposed method under each setting respectively. For the first value, as suggested by Cai et al. [10], we consider $m_1 = 200$ candidate mediators. However, noticing that $m_1 = 200$ exceeds the sample size ($n = 85$), as a reference, we also follow the suggestion by Perara et al. [94], and select $m_2 = 2n/\log(n) \approx 38$ candidate mediators. In the following, we report the results based on $m1 = 200$. Results obtained under $m_2 = 38$ can be found in Appendix D Additional results from real data application.

Figure 3.4a and 3.4b shows the (10-fold) cross-validation mean square error (MSE) versus the original and natural logarithm of the penalization parameter λ. When $\lambda = 0.01$, we obtain the smallest value of cross-validation MSE. We adopt the "1SE" criteria for choosing penalization parameters. Following this criteria, instead of choosing λ to be 0.01, we choose $\lambda = 0.04$, because under such a value of λ, we obtain the most parsimony model while the corresponding cross-validation MSE is within one standard error of the smallest one. Figure 3.5a and 3.5b shows values of $\alpha\beta$ versus λ and $\log(\lambda)$, and it is clear that when λ increases, $\alpha\beta$ decreases and eventually shrinks to zero with sufficiently large value of λ.

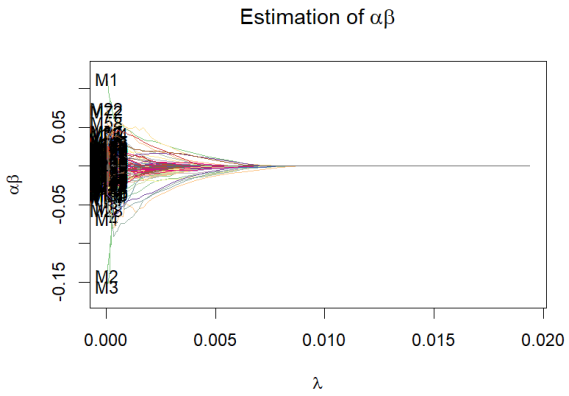
When considering 200 candidate variables, the proposed method identifies 15 DNA methylation loci that have mediation effects on the process of childhood trauma to persistent psychological disorders on stress reaction. Of the 15 DNA methylation loci, 13 of them (cg18634806, cg25626453, cg00096307, cg06992213, cg25448067, cg26657045, cg03643137, cg01696984, cg05292310, cg06001786, cg09211256, cg05051734, cg16180796) form a group due to associations among them and similar mediation effects. Within each group, the pairwise association between any two DNA methylation loci is greater than 0.1. These 13 loci have a grouped indirect effect of -10.802 (95$CI: (-20.401, -1.203)). The result can be interpreted as: when changing the treatment status for the 20 methylation loci from control to treated, while holding all the other factors including the overall treatment at the



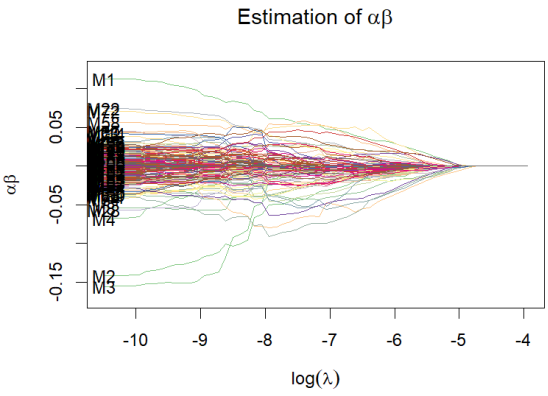
(a) Cross-validation MSE versus λ



(b) Cross-validation MSE versus log(λ)



(a) $\alpha\beta$ versus λ



(b) $\alpha\beta$ versus log(λ)

same level, the response will exhibit a change of -10.802, with a 95% CI of (-20.401, -1.203). Such a result indicates a very strong grouped mediation effect of the 20 DNA methylation loci. In addition, two single DNA methylation loci cg05608730 and cg00578039 show individual indirect effects of -3.305 and 2.010 with 95% CI (-6.976, 0.365) and (-1.061, 5.082) respectively. The individual effects from the two single loci do not show significant indirect effects on the pathway from childhood trauma to long-term psychiatric disorder. The direct effect of childhood trauma exposure is -2.555, with a 95% CI of (-12.977, 7.887), which is also insignificant. Figure 3.6 shows the mediation structure identified by the proposed method. Table 3.23 illustrates details on the grouped and individual indirect effects of DNA methylation loci with 95% confidence intervals. In this chapter, we only

provide results on the estimation of causal effects using conventional methods. In Chapter 4, we further provide results on the estimation of causal effects using the multiple robust methods that are proposed in that Chapter.

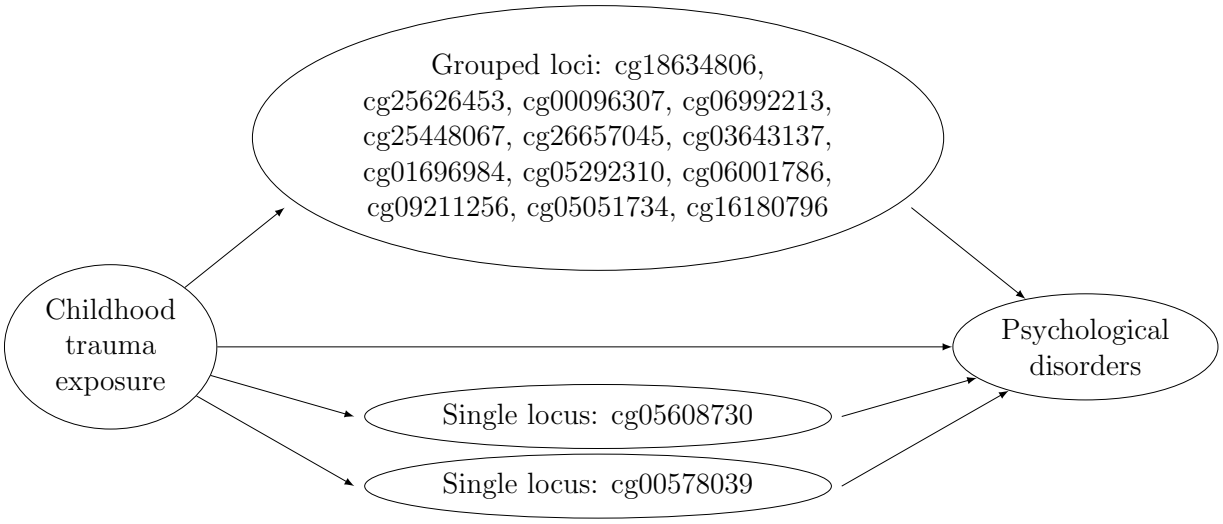


Figure 3.6: Mediation structure of DNA methylation loci on the process of childhood trauma exposure to long-term psychological disorders

Moreover, in Chapter 4, we will revisit this problem and analyze the causal effects of interest using the multiply robust methods proposed in that chapter. However, the current version of the MR estimators proposed in Chapter 4 requires the exposure variable to be binary, but in the data, the exposure (childhood trauma) is recorded as the total score of the Childhood Trauma Questionnaire (CTQ) and is measured on a continuous scale. To tackle the issue, we use the Bernstein and Fink [5] cut-off points to categorize continuous exposures into binary ones. In this study, we regard each individual with a total higher than 41 as being exposed and the opposite as being unexposed. For detailed discussions on the cut-off values, please refer to Section 4.5 of this thesis. Table 3.24 presents the results if the exposure is re-categorized into binary measures. It can be seen that both the direction and the level of significance for each estimated causal effect remain the same, with only the magnitude being changed due to the transformation of exposure type.

3.5 Discussion

In this chapter, we address the challenge of conducting causal mediation analysis in the presence of high-dimensional correlated candidate mediators. We introduce a penalization-based technique that serves a dual purpose: it helps select the most relevant mediators from the pool of high-dimensional candidates and the dependency structures among them as well. This selection approach enhances the precision and efficiency of the following statistical analysis and inference procedures.

On one hand, by penalizing the potential indirect effects throughout each candidate mediator directly, our proposed method effectively identifies the mediators exhibiting non-zero indirect effects in a straightforward way. On the other hand, we transform the problem of selecting dependency structures among the mediators into a linear penalization-based variable selection problem. This transformation eliminates the need for performing optimization that involves matrix calculations while ensuring the precise and fast selection of non-zero elements within the correlation matrix. By working with the selected dataset, researchers have a better understanding of the causal mechanism underlining the problem. This not only enhances their understanding of the problem at hand but also facilitates better explanations. Moreover, the selection process offers advantages for subsequent statistical inferences by reducing computation costs, particularly in scenarios involving high-dimensional data.

There are also limitations to the proposed approach. In terms of the mediator selection, the objective function is constructed by augmenting the loss function with a penalization term corresponding to the potential indirect effect through each candidate mediator. As a consequence, the objective function may not always be convex and therefore finding the global optima is not guaranteed for most cases. Additionally, in practice, the optimization procedure could be time-consuming and sensitive to starting values. Due to the nature of the optimization problem, it is possible that with poor starting values, the algorithm may face convergence issues of boundary solutions. When implementing the proposed method, we recommend running the proposed algorithm with randomized starting values to alleviate the problem. One may also refer to Friedman et al. [27], Simon et al. [118] and Tibshirani et al.[131] regarding optimization issues in similar penalization-based problems. Additionally, concerning the selection of dependence structures, our method currently focuses solely on pairwise correlations. In practice, it is possible to obtain a special phenomenon where the dependency relationships among the mediators are hard to identify. For example, suppose there are 3 mediators in total and we denote them as M_1, M_2 and M_3, it is possible that M_1 and M_2 are correlated, M_1 and M_3 are correlated but M_2 and M_3 are not. At present, there is no better way to deal with such a special phenomenon. Under such a special scenario,

the ad hoc method we are currently using is to treat all of the three mediators as a group and estimate their grouped indirect effect. We acknowledge this is another limitation of the proposed method. Moreover, we require a common true list of mediators to exist for all subjects in the study. In the DNA methylation study that is discussed in Section 3.4, such an assumption is satisfied naturally. However, this assumption may not always hold, particularly, if the list of mediators is not of the same type. Future research can explore these areas to address these limitations and further refine our approach. In addition, since the main contribution of this study is the novel approach of grouping mediators based on their correlations and assessing the grouped effects for those highly correlated ones, such a strategy can be extended to other problems. For example, in a study involving multiple, potentially high-dimensional correlated exposures (treatments), a similar strategy can be implemented to group exposures first before assessing their grouped/individual effect(s). Future research may concentrate on extending the current framework.

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.005	0.285	0.263	0.925
		Individual IE (of M8)	-0.003	0.185	0.186	0.935
		Joint IE	-0.004	0.348	0.323	0.924
		DE	0.004	0.269	0.265	0.946
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.006	0.266	0.232	0.913
		Individual IE (of M8)	-0.010	0.165	0.164	0.938
		Joint IE	-0.003	0.316	0.285	0.932
		DE	-0.002	0.214	0.219	0.940
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.005	0.246	0.226	0.923
		Individual IE (of M8)	-0.001	0.165	0.159	0.936
		Joint IE	0.004	0.306	0.278	0.927
		DE	-0.002	0.207	0.213	0.960
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.003	0.200	0.186	0.928
		Individual IE (of M8)	0.005	0.133	0.133	0.942
		Joint IE	0.004	0.256	0.231	0.914
		DE	-0.017	0.213	0.212	0.953
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.003	0.189	0.163	0.917
		Individual IE (of M8)	0.002	0.121	0.116	0.942
		Joint IE	0.004	0.282	0.203	0.915
		DE	0.003	0.182	0.185	0.955
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.005	0.167	0.159	0.937
		Individual IE (of M8)	0.000	0.109	0.112	0.953
		Joint IE	0.007	0.214	0.197	0.934
		DE	0.003	0.178	0.182	0.958
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.075	0.310	0.119	0.884
		Individual IE (of M8)	0.012	0.148	0.079	0.871
		Joint IE	0.327	1.533	0.169	0.862
		DE	0.000	0.168	0.170	0.944
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.031	0.221	0.098	0.841
		Individual IE (of M8)	0.000	0.097	0.066	0.890
		Joint IE	0.149	0.953	0.136	0.853
		DE	0.003	0.159	0.160	0.951
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.038	0.244	0.096	0.838
		Individual IE (of M8)	0.005	0.106	0.065	0.901
		Joint IE	0.202	1.130	0.137	0.859
		DE	0.010	0.160	0.158	0.942

Table 3.10: Overall performance of causal effects estimations under $J = 10$ scenario using the parallel approach

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.002	0.285	0.279	0.950	952
		Individual IE (of M8)	-0.003	0.183	0.186	0.936	
		Joint IE	-0.005	0.340	0.335	0.945	
		DE	0.006	0.266	0.265	0.947	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.004	0.265	0.259	0.944	932
		Individual IE (of M8)	-0.009	0.164	0.164	0.939	
		Joint IE	-0.005	0.309	0.306	0.948	
		DE	0.001	0.211	0.219	0.943	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.002	0.243	0.236	0.942	875
		Individual IE (of M8)	0.002	0.162	0.160	0.937	
		Joint IE	0.004	0.297	0.287	0.945	
		DE	-0.001	0.205	0.213	0.963	
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.002	0.198	0.196	0.942	886
		Individual IE (of M8)	0.008	0.130	0.133	0.949	
		Joint IE	0.010	0.236	0.237	0.941	
		DE	-0.014	0.215	0.212	0.949	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.001	0.182	0.182	0.945	866
		Individual IE (of M8)	0.004	0.119	0.116	0.942	
		Joint IE	0.004	0.216	0.216	0.961	
		DE	0.004	0.178	0.186	0.958	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.009	0.163	0.165	0.952	817
		Individual IE (of M8)	0.003	0.108	0.113	0.955	
		Joint IE	0.013	0.198	0.201	0.962	
		DE	0.002	0.173	0.182	0.966	
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.019	0.101	0.115	0.975	561
		Individual IE (of M8)	0.008	0.075	0.081	0.970	
		Joint IE	0.028	0.129	0.141	0.970	
		DE	-0.007	0.166	0.171	0.948	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.015	0.085	0.107	0.977	485
		Individual IE (of M8)	0.004	0.061	0.068	0.977	
		Joint IE	0.019	0.107	0.126	0.981	
		DE	0.000	0.155	0.161	0.971	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.011	0.088	0.097	0.973	452
		Individual IE (of M8)	0.006	0.057	0.067	0.987	
		Joint IE	0.018	0.104	0.118	0.971	
		DE	0.003	0.159	0.159	0.942	

Table 3.11: Performance of causal effects estimations under $J = 10$ scenario using the parallel approach conditioning on correct mediator selection

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.002	0.285	0.279	0.950	952
		Individual IE (of M8)	-0.003	0.183	0.186	0.936	
		Joint IE	-0.005	0.340	0.335	0.945	
		DE	0.006	0.266	0.265	0.947	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.004	0.265	0.259	0.944	928
		Individual IE (of M8)	-0.009	0.164	0.164	0.939	
		Joint IE	-0.005	0.309	0.306	0.948	
		DE	0.001	0.212	0.219	0.943	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	-0.003	0.241	0.240	0.946	606
		Individual IE (of M8)	0.001	0.158	0.159	0.946	
		Joint IE	-0.002	0.287	0.288	0.954	
		DE	0.002	0.202	0.213	0.967	
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.002	0.198	0.196	0.942	886
		Individual IE (of M8)	0.008	0.130	0.133	0.949	
		Joint IE	0.010	0.236	0.237	0.941	
		DE	-0.014	0.215	0.212	0.949	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.000	0.182	0.182	0.945	865
		Individual IE (of M8)	0.005	0.119	0.116	0.942	
		Joint IE	0.004	0.216	0.216	0.962	
		DE	0.004	0.178	0.186	0.958	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.006	0.161	0.169	0.955	538
		Individual IE (of M8)	-0.002	0.102	0.112	0.957	
		Joint IE	0.003	0.191	0.203	0.967	
		DE	0.005	0.175	0.182	0.968	
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.019	0.101	0.115	0.975	561
		Individual IE (of M8)	0.008	0.075	0.081	0.970	
		Joint IE	0.028	0.129	0.141	0.970	
		DE	-0.007	0.166	0.171	0.948	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.015	0.085	0.107	0.977	485
		Individual IE (of M8)	0.004	0.061	0.068	0.977	
		Joint IE	0.019	0.107	0.126	0.981	
		DE	0.000	0.155	0.161	0.971	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.009	0.087	0.099	0.974	313
		Individual IE (of M8)	0.005	0.058	0.067	0.987	
		Joint IE	0.013	0.104	0.120	0.971	
		DE	0.003	0.153	0.159	0.949	

Table 3.12: Performance of causal effects estimations under $J = 10$ scenario using the parallel approach conditioning on both correct mediator selection and correct correlation selection

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.005	0.285	0.263	0.925
		Individual IE (of M8)	-0.003	0.185	0.186	0.935
		Joint IE	-0.004	0.348	0.323	0.926
		DE	0.004	0.269	0.265	0.946
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.006	0.266	0.232	0.913
		Individual IE (of M8)	-0.010	0.165	0.164	0.938
		Joint IE	-0.003	0.314	0.285	0.929
		DE	-0.002	0.214	0.219	0.940
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.005	0.246	0.226	0.923
		Individual IE (of M8)	-0.001	0.165	0.159	0.936
		Joint IE	0.003	0.306	0.278	0.928
		DE	-0.002	0.207	0.213	0.960
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.004	0.201	0.186	0.928
		Individual IE (of M8)	0.005	0.133	0.133	0.942
		Joint IE	0.002	0.256	0.231	0.912
		DE	-0.017	0.213	0.212	0.953
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.005	0.190	0.163	0.917
		Individual IE (of M8)	0.002	0.121	0.116	0.942
		Joint IE	-0.002	0.240	0.203	0.910
		DE	0.003	0.182	0.185	0.955
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.005	0.167	0.159	0.937
		Individual IE (of M8)	0.000	0.109	0.112	0.953
		Joint IE	0.008	0.213	0.197	0.933
		DE	0.003	0.178	0.182	0.958
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.006	0.132	0.109	0.893
		Individual IE (of M8)	-0.007	0.087	0.076	0.864
		Joint IE	-0.012	0.182	0.140	0.866
		DE	0.000	0.168	0.170	0.944
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.013	0.122	0.092	0.838
		Individual IE (of M8)	-0.007	0.077	0.065	0.873
		Joint IE	-0.021	0.162	0.120	0.850
		DE	0.003	0.159	0.160	0.951
	$\rho = 0.2$	Grouped IE (of M6 and M7)	-0.016	0.116	0.089	0.847
		Individual IE (of M8)	-0.006	0.073	0.063	0.890
		Joint IE	-0.022	0.151	0.115	0.867
		DE	0.010	0.160	0.158	0.942

Table 3.13: Overall performance of causal effects estimations under $J = 10$ scenario using the two-stage approach

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.002	0.286	0.279	0.949	949
		Individual IE (of M8)	-0.004	0.183	0.186	0.936	
		Joint IE	-0.005	0.341	0.335	0.946	
		DE	0.006	0.267	0.266	0.946	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.007	0.266	0.258	0.943	936
		Individual IE (of M8)	-0.010	0.164	0.164	0.938	
		Joint IE	-0.003	0.310	0.306	0.947	
		DE	-0.001	0.212	0.219	0.941	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.003	0.243	0.234	0.939	882
		Individual IE (of M8)	0.001	0.163	0.160	0.937	
		Joint IE	0.005	0.293	0.287	0.947	
		DE	0.000	0.206	0.213	0.964	
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.004	0.196	0.196	0.945	886
		Individual IE (of M8)	0.006	0.130	0.133	0.948	
		Joint IE	0.010	0.236	0.237	0.941	
		DE	-0.016	0.214	0.212	0.950	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.001	0.181	0.182	0.943	867
		Individual IE (of M8)	0.004	0.120	0.116	0.940	
		Joint IE	0.003	0.216	0.216	0.958	
		DE	0.002	0.178	0.186	0.960	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.007	0.165	0.164	0.947	819
		Individual IE (of M8)	0.003	0.108	0.113	0.954	
		Joint IE	0.009	0.199	0.201	0.961	
		DE	0.004	0.174	0.182	0.962	
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.018	0.102	0.115	0.975	567
		Individual IE (of M8)	0.008	0.074	0.081	0.972	
		Joint IE	0.027	0.130	0.141	0.968	
		DE	-0.006	0.165	0.171	0.951	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.015	0.085	0.107	0.977	486
		Individual IE (of M8)	0.005	0.061	0.068	0.975	
		Joint IE	0.020	0.107	0.127	0.979	
		DE	0.000	0.154	0.161	0.973	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.012	0.089	0.097	0.968	466
		Individual IE (of M8)	0.007	0.058	0.067	0.989	
		Joint IE	0.019	0.106	0.119	0.972	
		DE	0.004	0.160	0.159	0.944	

Table 3.14: Performance of causal effects estimations under $J = 10$ scenario using the two-stage approach conditioning on correct mediator selection

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.002	0.286	0.279	0.949	949
		Individual IE (of M8)	-0.004	0.183	0.186	0.936	
		Joint IE	-0.005	0.341	0.335	0.946	
		DE	0.006	0.267	0.266	0.946	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.004	0.265	0.259	0.943	910
		Individual IE (of M8)	-0.010	0.165	0.164	0.938	
		Joint IE	-0.006	0.309	0.306	0.946	
		DE	-0.001	0.212	0.219	0.942	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.002	0.238	0.241	0.948	502
		Individual IE (of M8)	0.000	0.158	0.159	0.956	
		Joint IE	0.002	0.277	0.289	0.964	
		DE	0.002	0.203	0.213	0.964	
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.004	0.196	0.196	0.945	885
		Individual IE (of M8)	0.006	0.130	0.133	0.948	
		Joint IE	0.010	0.236	0.237	0.941	
		DE	-0.016	0.214	0.212	0.950	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.000	0.181	0.182	0.943	847
		Individual IE (of M8)	0.004	0.119	0.116	0.940	
		Joint IE	0.004	0.215	0.216	0.961	
		DE	0.002	0.179	0.186	0.959	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.006	0.171	0.170	0.944	444
		Individual IE (of M8)	0.000	0.101	0.112	0.962	
		Joint IE	0.006	0.201	0.203	0.962	
		DE	0.005	0.171	0.182	0.966	
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.018	0.101	0.115	0.975	566
		Individual IE (of M8)	0.008	0.075	0.081	0.972	
		Joint IE	0.027	0.130	0.141	0.968	
		DE	-0.006	0.165	0.171	0.951	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.015	0.085	0.107	0.977	480
		Individual IE (of M8)	0.004	0.061	0.068	0.975	
		Joint IE	0.020	0.107	0.127	0.979	
		DE	0.001	0.154	0.161	0.973	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.007	0.089	0.100	0.969	259
		Individual IE (of M8)	0.004	0.060	0.067	0.988	
		Joint IE	0.011	0.109	0.120	0.961	
		DE	0.003	0.155	0.159	0.946	

Table 3.15: Performance of causal effects estimations under $J = 10$ scenario using the two-stage approach conditioning on both correct mediator selection and correct correlation selection

mean model	correlation	TP	TP rate	FP	FP rate
$\alpha = \beta = 1$	$\rho = 0.8$	3.000	1.000	0.326	0.002
	$\rho = 0.5$	3.000	1.000	0.337	0.002
	$\rho = 0.2$	3.000	1.000	0.430	0.002
$\alpha = \beta = 0.7$	$\rho = 0.8$	2.996	0.999	0.296	0.002
	$\rho = 0.5$	2.996	0.999	0.348	0.002
	$\rho = 0.2$	2.997	0.999	0.432	0.002
$\alpha = \beta = 0.4$	$\rho = 0.8$	2.118	0.706	0.191	0.001
	$\rho = 0.5$	2.200	0.733	0.223	0.001
	$\rho = 0.2$	2.202	0.734	0.335	0.002

Table 3.16: Results of mediator selection under $J = 200$ scenario

method	scenario		TP	TP rate	FP	FP rate
The de-biased LASSO	$\alpha = \beta = 1$	$\rho = 0.8$	2.990	0.997	5.970	0.030
		$\rho = 0.5$	2.960	0.987	6.310	0.032
		$\rho = 0.2$	3.000	1.000	6.940	0.035
	$\alpha = \beta = 0.7$	$\rho = 0.8$	2.990	0.997	7.080	0.036
		$\rho = 0.5$	2.980	0.993	7.140	0.036
		$\rho = 0.2$	3.000	1.000	7.090	0.036
	$\alpha = \beta = 0.4$	$\rho = 0.8$	2.580	0.860	7.760	0.039
		$\rho = 0.5$	2.830	0.943	8.280	0.042
		$\rho = 0.2$	2.840	0.947	8.830	0.045
The minimax concave penalty	$\alpha = \beta = 1$	$\rho = 0.8$	2.940	0.980	4.730	0.024
		$\rho = 0.5$	2.960	0.987	3.820	0.019
		$\rho = 0.2$	3.000	1.000	3.320	0.017
	$\alpha = \beta = 0.7$	$\rho = 0.8$	2.560	0.853	7.110	0.036
		$\rho = 0.5$	2.980	0.993	4.950	0.025
		$\rho = 0.2$	3.000	1.000	4.880	0.025
	$\alpha = \beta = 0.4$	$\rho = 0.8$	1.760	0.587	7.790	0.040
		$\rho = 0.5$	2.530	0.843	10.690	0.054
		$\rho = 0.2$	2.720	0.907	11.930	0.061
The adaptive LASSO	$\alpha = \beta = 1$	$\rho = 0.8$	2.940	0.980	2.040	0.010
		$\rho = 0.5$	2.970	0.990	2.090	0.011
		$\rho = 0.2$	3.000	1.000	2.150	0.011
	$\alpha = \beta = 0.7$	$\rho = 0.8$	2.940	0.980	2.290	0.012
		$\rho = 0.5$	2.920	0.973	2.400	0.012
		$\rho = 0.2$	2.950	0.983	2.530	0.013
	$\alpha = \beta = 0.4$	$\rho = 0.8$	2.400	0.800	3.160	0.016
		$\rho = 0.5$	2.700	0.900	4.150	0.021
		$\rho = 0.2$	2.690	0.897	5.070	0.026

Table 3.17: Performance of existing methods on mediator selections under $p = 200$ scenario

Settings		Unconditional		Conditional		
mean model	correlation	TP rate	FP rate	TP rate	FP rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	1.000	0.000	1.000	0.000	765
	$\rho = 0.5$	1.000	0.015	1.000	0.014	740
	$\rho = 0.2$	0.822	0.173	0.876	0.202	670
$\alpha = \beta = 0.7$	$\rho = 0.8$	1.000	0.002	1.000	0.001	791
	$\rho = 0.5$	1.000	0.015	1.000	0.013	733
	$\rho = 0.2$	0.798	0.181	0.838	0.199	678
$\alpha = \beta = 0.4$	$\rho = 0.8$	1.000	0.061	1.000	0.000	425
	$\rho = 0.5$	1.000	0.117	1.000	0.010	395
	$\rho = 0.2$	0.833	0.244	0.843	0.201	345

Table 3.18: Results of correlation selection under $J = 200$ scenario using the two-stage approach

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.009	0.284	0.260	0.922
		Individual IE (of M8)	-0.008	0.164	0.158	0.940
		Joint IE	-0.015	0.365	0.311	0.914
		DE	-0.003	0.186	0.177	0.944
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.002	0.266	0.230	0.912
		Individual IE (of M8)	-0.017	0.161	0.157	0.928
		Joint IE	-0.019	0.338	0.284	0.902
		DE	-0.007	0.189	0.178	0.937
	$\rho = 0.2$	Grouped IE (of M6 and M7)	-0.003	0.245	0.224	0.920
		Individual IE (of M8)	-0.008	0.165	0.157	0.930
		Joint IE	-0.012	0.326	0.281	0.910
		DE	-0.006	0.187	0.183	0.941
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.007	0.198	0.184	0.927
		Individual IE (of M8)	-0.006	0.113	0.111	0.938
		Joint IE	-0.013	0.251	0.219	0.905
		DE	-0.008	0.170	0.159	0.923
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.009	0.184	0.162	0.915
		Individual IE (of M8)	-0.004	0.117	0.110	0.937
		Joint IE	-0.012	0.239	0.200	0.898
		DE	-0.003	0.168	0.160	0.944
	$\rho = 0.2$	Grouped IE (of M6 and M7)	-0.004	0.166	0.157	0.931
		Individual IE (of M8)	-0.005	0.108	0.110	0.947
		Joint IE	-0.001	0.218	0.197	0.918
		DE	-0.005	0.169	0.162	0.944
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.052	0.176	0.098	0.704
		Individual IE (of M8)	-0.037	0.096	0.055	0.663
		Joint IE	-0.089	0.225	0.115	0.681
		DE	-0.002	0.155	0.147	0.931
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.054	0.156	0.084	0.685
		Individual IE (of M8)	-0.031	0.094	0.056	0.700
		Joint IE	-0.087	0.203	0.103	0.687
		DE	-0.005	0.152	0.148	0.938
	$\rho = 0.2$	Grouped IE (of M6 and M7)	-0.058	0.149	0.079	0.689
		Individual IE (of M8)	-0.031	0.092	0.056	0.716
		Joint IE	-0.091	0.194	0.100	0.686
		DE	-0.003	0.158	0.147	0.922

Table 3.19: Performance of causal effects estimations under $J = 200$ scenario using the two-stage approach

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.004	0.284	0.278	0.950	765
		Individual IE (of M8)	-0.004	0.159	0.158	0.941	
		Joint IE	-0.008	0.331	0.320	0.950	
		DE	-0.005	0.185	0.177	0.941	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.001	0.265	0.257	0.946	740
		Individual IE (of M8)	-0.018	0.161	0.157	0.922	
		Joint IE	-0.018	0.304	0.302	0.953	
		DE	-0.007	0.184	0.178	0.951	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	-0.002	0.242	0.233	0.937	670
		Individual IE (of M8)	-0.006	0.170	0.157	0.925	
		Joint IE	-0.008	0.297	0.284	0.936	
		DE	-0.005	0.185	0.183	0.945	
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.002	0.194	0.196	0.956	791
		Individual IE (of M8)	0.000	0.110	0.112	0.942	
		Joint IE	-0.002	0.224	0.225	0.946	
		DE	-0.007	0.166	0.159	0.930	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.012	0.181	0.180	0.950	733
		Individual IE (of M8)	-0.002	0.113	0.111	0.944	
		Joint IE	-0.014	0.215	0.212	0.947	
		DE	-0.003	0.164	0.160	0.951	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.001	0.165	0.163	0.944	678
		Individual IE (of M8)	-0.005	0.108	0.110	0.945	
		Joint IE	-0.003	0.194	0.198	0.953	
		DE	-0.002	0.172	0.163	0.941	
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.024	0.096	0.116	0.960	425
		Individual IE (of M8)	0.008	0.059	0.067	0.976	
		Joint IE	0.032	0.113	0.134	0.955	
		DE	-0.008	0.154	0.149	0.934	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.018	0.081	0.106	0.972	395
		Individual IE (of M8)	0.006	0.052	0.065	0.987	
		Joint IE	0.024	0.092	0.125	0.980	
		DE	-0.002	0.149	0.149	0.957	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.013	0.080	0.096	0.974	345
		Individual IE (of M8)	0.006	0.054	0.066	0.988	
		Joint IE	0.019	0.092	0.117	0.968	
		DE	-0.001	0.154	0.149	0.948	

Table 3.20: Performance of causal effects estimations under $J = 200$ scenario using the two-stage approach conditioning on correct mediator selection

mean model	correlation	causal effect	bias	ESE	ASE	coverage rate	No.
$\alpha = \beta = 1$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.004	0.284	0.278	0.950	765
		Individual IE (of M8)	-0.004	0.159	0.158	0.941	
		Joint IE	-0.008	0.331	0.320	0.950	
		DE	-0.005	0.185	0.177	0.941	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.003	0.265	0.258	0.945	723
		Individual IE (of M8)	-0.018	0.161	0.157	0.920	
		Joint IE	-0.020	0.305	0.302	0.953	
		DE	-0.007	0.184	0.178	0.952	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.002	0.238	0.239	0.948	385
		Individual IE (of M8)	-0.003	0.170	0.157	0.938	
		Joint IE	-0.001	0.287	0.286	0.945	
		DE	-0.002	0.179	0.183	0.958	
$\alpha = \beta = 0.7$	$\rho = 0.8$	Grouped IE (of M6 and M7)	-0.002	0.194	0.196	0.956	790
		Individual IE (of M8)	0.000	0.110	0.112	0.942	
		Joint IE	-0.002	0.224	0.225	0.946	
		DE	-0.007	0.166	0.159	0.930	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	-0.011	0.182	0.181	0.950	716
		Individual IE (of M8)	-0.003	0.113	0.111	0.943	
		Joint IE	-0.014	0.214	0.212	0.948	
		DE	-0.003	0.165	0.160	0.950	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.002	0.165	0.168	0.953	364
		Individual IE (of M8)	-0.003	0.102	0.110	0.953	
		Joint IE	0.000	0.194	0.201	0.967	
		DE	-0.002	0.171	0.163	0.945	
$\alpha = \beta = 0.4$	$\rho = 0.8$	Grouped IE (of M6 and M7)	0.024	0.096	0.116	0.960	425
		Individual IE (of M8)	0.008	0.059	0.067	0.976	
		Joint IE	0.032	0.113	0.134	0.955	
		DE	-0.008	0.154	0.149	0.934	
	$\rho = 0.5$	Grouped IE (of M6 and M7)	0.018	0.081	0.106	0.972	389
		Individual IE (of M8)	0.006	0.052	0.065	0.987	
		Joint IE	0.024	0.092	0.125	0.979	
		DE	-0.001	0.148	0.149	0.959	
	$\rho = 0.2$	Grouped IE (of M6 and M7)	0.007	0.077	0.098	0.979	187
		Individual IE (of M8)	0.004	0.054	0.066	0.995	
		Joint IE	0.011	0.092	0.118	0.973	
		DE	0.001	0.145	0.149	0.952	

Table 3.21: Performance of causal effects estimations under $J = 200$ scenario using the two-stage approach conditioning on both correct mediator selection and correct correlation selection

Table 3.22: Descriptive statistics for selected variables

Variable	mean	SD	min, max
sex (% of female)	0.506	0.503	0, 1
age (in years)	33.800	15.900	18, 69
stress (cortisol stress reactivity, AUCi tn:1)	243.460	420.613	-1029.85, 1876.28
trauma (total CTQtn:2 score)	31.906	8.228	24, 63

[1] area under the curve (AUC) with respect to the increase
[2] Childhood Trauma Questionnaire

	value	95% CI
Grouped IE of cg18634806, cg25626453, cg00096307, cg06992213, cg25448067, cg26657045, cg03643137, cg01696984, cg05292310, cg06001786, cg09211256, cg05051734, cg16180796	-10.802	(-20.401, -1.203)
Individual IE of cg05608730	-3.305	(-6.976, 0.365)
Individual IE of cg00578039	2.01	(-1.061, 5.082)
DE	-2.555	(-12.997, 7.887)

Table 3.23: grouped and individual indirect effects of DNA methylation loci

causal effects	value	95% CI
Grouped IE of cg18634806, cg25626453, cg00096307, cg06992213, cg25448067, cg26657045, cg03643137, cg01696984, cg05292310, cg06001786, cg09211256, cg05051734, cg16180796	-287.022	(-544.167,-29.877)
Single IE of cg05608730	-86.928	(-161.754,-12.103)
Single IE of cg00578039	32.465	(-27.678,92.608)
DE	-48.609	(-225.783,128.565)

Table 3.24: Estimated grouped and individual indirect effects of DNA methylation loci when treatment (childhood trauma) is treated as binary

Chapter 4

Multiply Robust Estimation for Mediation Analysis with Multiple Mediators

4.1 Introduction

In Chapter 3, we introduce methods for mediator selection and dependency structure simplification, which are essential steps prior to estimations of causal effects. In this chapter, we focus on the estimation of causal effects and conducting related statistical inferences on the selected dataset. We propose two ways of constructing multiple robust estimators in causal mediation analysis based on weighted regression (WR) and augmented inverse propensity weighting (AIPW).

Following the introduction in Section 1.5, we extend the aforementioned two DR estimation models and propose two ways of constructing MR estimators for causal mediation analysis with multiple possibly correlated mediators. First, three working models must be specified:

1. The treatment model: Modeling $P(T = 1|\boldsymbol{X})$. We denote it as $\mathcal{M}_P$ and specify $P(T = 1|\boldsymbol{X}) = \pi(\boldsymbol{X}; \boldsymbol{\theta}_P)$, where $\boldsymbol{\theta}_P$ denotes the parameters used in $\mathcal{M}_P$. The treatment model is also called the propensity model by some scholars and in this thesis, they are used interchangeably.

2. The mediator joint conditional density model: $f_{\boldsymbol{M}|\boldsymbol{X},T}(\boldsymbol{m}|\boldsymbol{X}, T)$. We denote it as $\mathcal{M}_M$ and simplify it as $f(\boldsymbol{m}|\boldsymbol{X}, T; \boldsymbol{\theta}_M)$, where $\boldsymbol{\theta}_M$ denotes the parameters used in

$\mathcal{M}_M$. We also need to model the joint CDF $F_{\boldsymbol{M}|\boldsymbol{X},T}(\boldsymbol{m}|\boldsymbol{X},T)$ and it is simplified as $F(\boldsymbol{m}|\boldsymbol{X},T;\boldsymbol{\theta}_M)$ similarly.

3. The response model: $E(Y|\boldsymbol{X},T,\boldsymbol{M})$. We denote it as $\mathcal{M}_Y$ and specify $E(Y|\boldsymbol{X},T,\boldsymbol{M}) = \mu_Y(\boldsymbol{X},T,\boldsymbol{M};\boldsymbol{\theta}_Y)$, where $\boldsymbol{\theta}_Y$ denotes the parameters used in $\mathcal{M}_Y$.

In the following, the notation $\mathcal{M}_A\&\mathcal{M}_B$ are used to denote models A and B are correctly specified.

For the MR estimators discussed in this chapter, we still require the SIMMA that is introduced in Chapter 2. Additionally, under SIMMA, the following theorem is essential for ensuring the MR properties of the two proposed estimators:

Theorem 4 (The Sequential Ignorability for Multiple Mediators Assumption with Propensity Score and Conditional Distribution of Mediators)**.** For any outcome Y, multiple mediators $\boldsymbol{M}$ and treatment indicator T,

1. $\{Y(t_0,\boldsymbol{m}), M_1(t_1),\ldots,M_J(t_J)\} \perp T|PS(\boldsymbol{X})$
2. $Y(t_0,\boldsymbol{m}') \perp \{M_1,\ldots,M_J\}\,|F(\boldsymbol{m}|\boldsymbol{X},T)$

where $0 < PS(\boldsymbol{X}) < 1$, $0 < F(\boldsymbol{m}|\boldsymbol{X},T) < 1$.

In the second part of the theorem, $F(\boldsymbol{m}|\boldsymbol{X},T)$ denotes the conditional joint distribution of the mediators given treatment indicator and covariates. We assume $\boldsymbol{M}$ has density $f(\boldsymbol{m}|\boldsymbol{X},T)$. The proof of the first half of the theorem is an extension of Rubin et al. [104, 106]. Our proof concentrates on the second half and it is shown in Appendix E.

4.2 Method 1: weighted regression (WR) approach

The first MR estimator we propose is based on weighted regression. This estimator is constructed by estimating two sets of weighted regression functions using different weights respectively. Two sets of regression models are considered: the set of mediator models and the response model. The mediator regression models are estimated with weights equal to the inverse of the propensity score and the outcome regression model is fitted with weights given by the product of the inverse propensity score and inverse of the conditional density of the mediators given treatment and covariates. Causal effect estimates can be then calculated from the functions of the estimated regression parameters from different models.

4.2.1 Models

For the weighted regression MR estimator, we require the mediators to be linearly related to the treatment and the outcome to be linearly related to both the treatment and the mediators, conditioning on the covariates.

Starting with $\mathcal{M}_M$, we first introduce the marginal models. If we let $\mu_{ij} = E(M_{ij}|T_i, \boldsymbol{X_i})$, then under the linearity assumption, we let

$$\mu_{ij}(\boldsymbol{X}_i, T_i; \boldsymbol{\Psi}_j) = g(\boldsymbol{X}_i'\boldsymbol{\eta}_j) + \alpha_j T_i,\ i = 1, \ldots, n;\ j = 1, \ldots, J,$$

where $\boldsymbol{\eta_j} = \{\eta_{j0}, \ldots, \eta_{jp}\}'$ is the vector of coefficients associating with the covariates, α_j is the coefficient of treatment, $g(\cdot)$ is some known functions associating μ_{ij} and $\boldsymbol{X}_i'\boldsymbol{\eta}_j$ and $\boldsymbol{\Psi}_j = \{\boldsymbol{\eta}_j', \alpha_j\}'$ denote the mean parameters for the marginal model. The linearity assumption means that, the conditional mean μ_{ij} must be linearly related to the treatment T_i, though we do not require such a strictly linear relationship between μ_{ij} and $\boldsymbol{X}_i$. However, for the remainder of this chapter, without loss of generality, we may assume the following form of the marginal model for simplicity:

$$\mu_{ij}(\boldsymbol{X}_i, T_i; \boldsymbol{\Psi}_j) = \boldsymbol{X}_i'\boldsymbol{\eta}_j + \alpha_j T_i,\ i = 1, \ldots, n;\ j = 1, \ldots, J, \tag{4.1}$$

We also assume

$$Var(M_{ij}|\boldsymbol{X}_i, T_i) = \sigma_j^2, \tag{4.2}$$

and further

$$M_{ij}|\boldsymbol{X}_i, T_i \sim N(\mu_{ij}, \sigma_j^2), \tag{4.3}$$

Therefore, for each subject i, if we denote the marginal conditional distribution of M_{ij} as $F_j(M_{ij}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_j)$, then $F_j(M_{ij}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_j) = \Phi\{[M_{ij} - \mu_{ij}(\boldsymbol{X}_i, T_i)]/\sigma_j\}$, where $\Phi(\cdot)$ denotes the CDF of a standard normal distribution and $\boldsymbol{\theta}_j$ denotes parameters for marginal distribution model of the jth mediator and in this case $\boldsymbol{\theta}_j = \{\boldsymbol{\Psi}_j', \sigma_j\}'$. We also denote the associated density as $f_j(M_{ij}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_j)$, so $f_j(M_{ij}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_j) = \phi\{[M_{ij} - \mu_{ij}(\boldsymbol{X}_i, T_i)]/\sigma_j\}/\sigma_j$, where $\phi(\cdot)$ denotes the density of a standard normal distribution.

We further assume that, for the joint conditional distribution of $\boldsymbol{M}_i$,

$$\boldsymbol{M}_i|\boldsymbol{X}_i, T_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \tag{4.4}$$

where $MVN(\cdot)$ denotes multivariate normal distribution, where the conditional mean vector is $\boldsymbol{\mu}_i = \{\mu_{ij}, j = 1, \ldots, J\}'$. If we denote the (k, l)th entry of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}_{kl}$, where

$k, l = 1, \ldots, J$, then $\boldsymbol{\Sigma}_{kl} = \sigma_k^2$ for $k = l$ and $\boldsymbol{\Sigma}_{kl} = \rho_{kl}\sigma_k\sigma_l$ for $k \neq l$. It follows that, $Cor(M_{ik}, M_{il}|\boldsymbol{X}_i, T_i) = \rho_{kl}$ and we also use $\boldsymbol{\varrho}$ to denote the conditional correlation matrix. Additionally, we use $\boldsymbol{\theta}_M$ to denote the parameters used for the joint mediator distribution model and in this case $\boldsymbol{\theta}_M = \{\boldsymbol{\theta}_j', \rho_{jk}, j = 1, \ldots, J, k = j+1, \ldots, J\}'$. We also use $\boldsymbol{\Psi}_M$ to denote the mean parameters from all marginal models (i.e. $\boldsymbol{\Psi}_M = \{\boldsymbol{\Psi}_j', j = 1, \ldots, J\}'$). Furthermore, we denote the joint conditional CDF of $\boldsymbol{M}_i$ as $F(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_M)$, then

$$F(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_M) = \boldsymbol{\Phi}\{\boldsymbol{\Sigma}^{-\frac{1}{2}}[\boldsymbol{M}_i - \boldsymbol{\mu}_i(\boldsymbol{X}_i, T_i)]\}, \tag{4.5}$$

where $\boldsymbol{\Phi}(\cdot)$ denotes the CDF of a standard multivariate normal distribution. As a consequence, the conditional density has the form

$$f(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_M) = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}\{\boldsymbol{\Sigma}^{-\frac{1}{2}}[\boldsymbol{M}_i - \boldsymbol{\mu}_i(\boldsymbol{X}_i, T_i)]\}, \tag{4.6}$$

where $\boldsymbol{\phi}(\cdot)$ denotes the density of a standard multivariate normal distribution.

For $\mathcal{M}_Y$, we let $\mu_{Y,i} = E\{Y_i|\boldsymbol{X}_i, T_i, \boldsymbol{M}_{i\cdot}\}$. Similarly, under the linear model assumption, we assume

$$\mu_{Y,i}(\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\theta}_Y) = h(\boldsymbol{X}_i'\boldsymbol{\gamma}) + \tau T_i + \boldsymbol{M}_{i\cdot}'\boldsymbol{\beta} \quad i = 1, \ldots, n,$$

where τ is the unknown coefficient of the treatment indicator T_i, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_J\}'$ is the J-dimensional vector of regression coefficients on the vector of mediators, $\boldsymbol{\gamma} = \{\gamma_0, \ldots, \gamma_p\}'$ is the $p+1$-dimensional vector of regression coefficients on the vector of covariates including the intercept, $h(\cdot)$ is some known function associating $\mu_{Y,i}$ with $\boldsymbol{X}_i'\boldsymbol{\gamma}$ and $\boldsymbol{\Psi}_Y = \{\boldsymbol{\gamma}', \tau, \boldsymbol{\beta}'\}'$ denote the mean parameters in the response model. In a similar manner, the linearity assumption means that, $\mu_{Y,i}$ must be linearly related to the treatment T_i and $\boldsymbol{M}_i$, though we do not impose such a strict restriction between $\mu_{Y,i}$ and $\boldsymbol{X}_i$. However, for the remainder of this chapter, without loss of generality, we may assume the following form of the response model for simplicity:

$$\mu_{Y,i}(\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\theta}_Y) = \boldsymbol{X}_i'\boldsymbol{\gamma} + \tau T_i + \boldsymbol{M}_{i\cdot}'\boldsymbol{\beta} \quad i = 1, \ldots, n, \tag{4.7}$$

A natural extension of the proposed method is that, the normality assumptions for the mediator distributions can be dropped. The proposed method is still valid if (4.3) and (4.4) are dropped, but (4.1) and (4.2) are kept and (4.5) is changed to

$$F(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_M) = \mathcal{C}[F_1(M_{i1}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_1), \ldots, F_J(M_{iJ}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_j); \boldsymbol{\varrho}], \tag{4.8}$$

where $\mathcal{C}(\cdot)$ denotes any copula function that factorizes the joint distribution of $\boldsymbol{M}$ into marginal distributions and correlation structures. It follows that, (4.6) is changed accordingly to

$$f(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_M) = c[F_1(M_{i1}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_1), \ldots, F_J(M_{iJ}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_j); \boldsymbol{\varrho}] \prod_{j=1}^{J} f_j(M_{ij}|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_j), \tag{4.9}$$

where $c(\cdot)$ denotes the density of copula $\mathcal{C}(\cdot)$.

We allow the covariates $\boldsymbol{X}$ in (4.1), (4.8), (4.9) and (4.7) to be different and we also allow interactions and higher order terms of covariates to be included in $\boldsymbol{X}$ with proper adjustment of coefficients. For simplicity of notations, we denote all covariates including a leading $\mathbf{1}$ as $\boldsymbol{X}$. In addition, in (4.10) and (4.11), $\boldsymbol{W}_i = \{\boldsymbol{X}_i', T_i\}'$ and $\boldsymbol{Z}_i = \{\boldsymbol{X}_i', T_i, \boldsymbol{M}_i'\}'$, which denote the covariates appears in (4.1) and (4.7) respectively. Throughout the thesis, $\boldsymbol{X}$, $\boldsymbol{W}$ and $\boldsymbol{Z}$ are used to denote covariates for $\mathcal{M}_P$, $\mathcal{M}_M$ and $\mathcal{M}_Y$ respectively, with a subscript i denotes individual value.

4.2.2 Estimations

The proposed method estimates $\boldsymbol{\Psi}_j$ by solving the following weighted regression estimating equations

$$P_n\left\{\omega_{M,i}[M_{ij} - \mu_{ij}(\boldsymbol{X}_i, T_i; \boldsymbol{\Psi}_j)]\boldsymbol{W}_i\right\} = \mathbf{0}, \quad \text{where} \quad \omega_{M,i} = \frac{1}{P(T_i|\boldsymbol{X}_i)} \tag{4.10}$$

and estimates parameters $\boldsymbol{\theta}_Y$ by solving another sets of weighted regression estimating equations

$$P_n\left\{\omega_{Y,i}[Y_i - \mu_{Y,i}(\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\theta}_Y)]\boldsymbol{Z}_i\right\} = \mathbf{0}, \quad \text{where} \quad \omega_{Y,i} = \frac{1}{P(T_i|\boldsymbol{X}_i)f(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i)}. \tag{4.11}$$

ω_M and ω_Y are called the weights for the mediator model and the response (outcome) model respectively, with a subscript i to denote the ith individual value. The causal effects of interest can be calculated as functions of the estimated parameters, as introduced in Chapter 1 of this thesis. Additionally, one may also use the stabilized weights for the estimation, by changing

$$\omega^*_{M,i} = \frac{P(T_i)}{P(T_i|\boldsymbol{X}_i)} \qquad \text{and} \qquad \omega^*_{Y,i} = \frac{f(\boldsymbol{M}_i|T_i)P(T_i)}{P(T_i|\boldsymbol{X}_i)f(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i)}. \tag{4.12}$$

In the above equation, the numerator values denote the marginal distribution of the treatment and the mediators.

In practice, we estimate the parameters and the causal effects of interest using iterative ways that are introduced as follows.

Algorithm 2. Algorithm of the MR estimator using weighted regression approach

Step 1: We first estimate the parameters for the treatment model as

$$\hat{\boldsymbol{\theta}}_P = \underset{\boldsymbol{\theta}_P}{\operatorname{argmax}} \sum_{i=1}^{n} \{T_i \log[\pi(\boldsymbol{X}_i; \boldsymbol{\theta}_P)] + (1 - T_i) \log[1 - \pi(\boldsymbol{X}_i; \boldsymbol{\theta}_P)]\},$$

and with $\hat{\boldsymbol{\theta}}_P$, we fit $\hat{\pi}_i(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}_P) = \pi(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}_P)$.

Step 2: We then solve (4.10) by solving:

$$P_n \{\hat{\omega}_{M,i}[M_{ij} - \mu_{ij}(\boldsymbol{X}_i, T_i; \boldsymbol{\Psi}_j)]\boldsymbol{W}_i\} = \mathbf{0}, \tag{4.13}$$

for all $j = 1, \ldots, J$, where

$$\hat{\omega}_{M,i} = \frac{I(T_i = 1)}{\hat{\pi}_i(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}_P)} + \frac{I(T_i = 0)}{1 - \hat{\pi}_i(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}_P)}$$

is an estimated version of $\omega_{M,i}$ in (4.10). We denote the solution of $\boldsymbol{\Psi}_j$ as $\hat{\boldsymbol{\Psi}}_j$ and particularly, the solution of α_j as $\hat{\alpha}_{j,WR}$.

Step 3: With $\hat{\boldsymbol{\Psi}}_j$ from Step 2, using a two-stage approach, we estimate $\boldsymbol{\Sigma}$ by

$$\hat{\boldsymbol{\Sigma}} = \underset{\boldsymbol{\Sigma}}{\operatorname{argmax}} \sum_{i=1}^{n} l_{W,M,i}(\boldsymbol{M}_i | \boldsymbol{X}_i, T_i; \hat{\boldsymbol{\Psi}}_M, \boldsymbol{\Sigma}), \tag{4.14}$$

where

$$l_{W,M,i}(\boldsymbol{M}_i | \boldsymbol{X}_i, T_i) = \hat{\omega}_{M,i}(\boldsymbol{X}_i, T_i; \hat{\boldsymbol{\theta}}_P) l_{M,i}(\boldsymbol{M}_i | \boldsymbol{X}_i, T_i; \hat{\boldsymbol{\Psi}}_M, \boldsymbol{\Sigma})$$

is the weighted log-likelihood function and

$$l_{M,i}(\boldsymbol{M}_i | \boldsymbol{X}_i, T_i; \hat{\boldsymbol{\Psi}}_M, \boldsymbol{\Sigma}) = \log [f(\boldsymbol{M}_i | \boldsymbol{X}_i, T_i; \hat{\boldsymbol{\Psi}}_M, \boldsymbol{\Sigma})]$$

is the log-likelihood function of $\boldsymbol{M}_i$ conditioning on $\boldsymbol{X}_i$ and T_i.

Step 4: we estimate $\boldsymbol{\theta}_Y$ in (4.7) by solving:

$$P_n\big\{\hat{\omega}_{Y,i}[Y_i - \mu_Y(\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\theta}_Y)]\boldsymbol{Z}_i\big\} = \mathbf{0}, \tag{4.15}$$

where

$$\hat{\omega}_{Y,i} = \frac{\hat{\omega}_{M,i}(\boldsymbol{X}_i, T_i; \hat{\boldsymbol{\theta}}_P)}{\hat{f}_i(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \hat{\boldsymbol{\theta}}_M)},$$

is an estimated version of $\omega_{Y,i}$ in (4.11). The solutions of $\boldsymbol{\theta}_Y$ to estimating equations (4.15) are denoted as $\hat{\boldsymbol{\theta}}_Y$, and particularly, the solutions of τ and $\boldsymbol{\beta}$ are denoted as $\hat{\tau}_{WR}$ and $\hat{\boldsymbol{\beta}}_{WR}$ respectively, where $\hat{\boldsymbol{\beta}}_{WR} = \{\hat{\beta}_{1,MR}, \ldots, \hat{\beta}_{J,MR}\}'$.

Step 5: Causal effects of interest are estimated under the model assumptions introduced in Section 3.2.1 and according to definitions introduced in Chapter 1 of this thesis. For example,

$$\widehat{IE}_{j,MR1} = \hat{\alpha}_{j,WR}\hat{\beta}_{j,WR}, \text{ for each } j = 1, \ldots, J;$$

$$\widehat{IE}_{MR1} = \sum_{j=1}^{J} \hat{\alpha}_{j,WR}\hat{\beta}_{j,WR}, \quad \text{and} \quad \widehat{DE}_{MR1} = \hat{\tau}_{WR}.$$

The grouped indirect effects of mediators in group $\mathcal{G}$ are estimated as follows:

$$\widehat{IE}_{\mathcal{G},MR1} = \sum_{j\in\mathcal{G}} \hat{\alpha}_{j,WR}\hat{\beta}_{j,WR}.$$

4.2.3 Using Matching or Stratification Approaches

One may also construct a multiply robust estimation via a matching approach. When matching is used, one can generate a matched dataset with subjects matched in pairs. Depending on different methods of the matching algorithm used, the matched dataset can include observations such that each (or multiple) observation(s) from the treatment group is matched with an (or multiple) observation(s) from the control group with similar estimated propensity scores as well as values of inverse joint conditional density of mediators (within a pre-specified caliber). For details on how the matching process is conducted, one can refer to Stuart, E. A. (2010)[126]. When the matching approach is used to construct a doubly robust estimator, the parameters estimated via weighting approaches should be replaced by conducting the same estimation on the respective matched dataset.

4.2.4 Theoretical Properties

In this section, we derive the consistency and asymptotic properties of the MR estimators constructed by the weighted regression approach.

Consistency of the weighted regression MR estimator

The following theorem states the DR property of weighted regression estimators in causal analysis settings.

Theorem 5 (DR property of the weighted regression estimator for average causal effect)**.** If for each subject, we observe $\boldsymbol{X}, T$ and Y, and we assume (1.1) and (1.2), then $\hat{\tau}_{WR}$ calculated as (1.3) is a DR estimator for τ.

We then proceed with the DR property in mediation analysis settings. We start with the following lemmas.

Lemma 3 (DR property of the mediator marginal conditional mean model)**.** If the marginal conditional mean model for each jth mediator is assumed as (4.1), then $\alpha_j = E\{M_j(1) - M_j(0)\}$, where $M(t)$ denotes the potential value of the mediator M_j under treatment (exposure) status $t = 0, 1$, and the expectation is taken with respect to covariates in the population. Additionally, if $\boldsymbol{\theta}_j$ are estimated by solving (4.10), then the estimator $\hat{\alpha}_{j,WR}$ is a DR estimator for the causal effect $E\{M_j(1) - M_j(0)\}$.

The first part of Lemma 3 states that, if the marginal conditional mean model for each mediator is assumed as (4.1), then the coefficient α_j always equals the causal effect defined as $E\{M_j(1) - M_j(0)\}$, which can be interpreted as the change of the expected potential value of the mediator if the treatment (exposure) is changed from 0 to 1. This can be shown since under our model assumptions, $E\{M_j|T = 0, \boldsymbol{X}\} = \boldsymbol{X}'\boldsymbol{\eta}_j$ and $E\{M_j|T = 1, \boldsymbol{X}\} = T + \boldsymbol{X}'\boldsymbol{\eta}_j$. The second part of Lemma 3 states that if the model parameters $\boldsymbol{\theta}_j$ in (4.1) are estimated by solving the weighted regression estimating equations (4.10), then $\hat{\alpha}_{j,WR}$, as an estimator of coefficient α_j, is also a DR estimator for the causal effect $E\{M_j(1)\} - E\{M_j(0)\}$. Proof of the DR property is an application of Theorem 5 with details provided in Appendix E.

The following lemma states the DR property of the estimated parameter in the response model.

Lemma 4 (DR property of the response model)**.** If the response model is assumed as in (4.7), then $\tau = E\{Y(1,m) - Y(0,m)\}$ and $\beta = [E\{Y(t,m) - Y(t,m')\}]/(m-m')$ for any t, m and m', where $Y(t,m)$ denotes the potential outcome under treatment (exposure) t and mediator value m, and the expectation is taken with respect to covariates in the population. Additionally, if $\boldsymbol{\theta}_Y$ are estimated by solving (4.11), then the estimators $\hat{\tau}_{WR}$ and $\hat{\beta}_{WR}$ are DR estimators of causal effects $E\{Y(1,m)-Y(0,m)\}$ and $[E\{Y(t,m)-Y(t,m')\}]/(m-m')$ respectively.

Similar to the proof of Lemma 3, the first part of Lemma 4 states that, if the response model is assumed as in (4.7), then under the model assumptions, the coefficients τ and β always equal the causal effect defined as $E\{Y(1,m) - Y(0,m)\}$ and $[E\{Y(t,m) - Y(t,m')\}]/(m-m')$ respectively, where the former one can be regarded as the controlled direct effect (the change of the expected outcome if the treatment (exposure) is changed from 0 to 1 and the mediator value is kept at m) and the later one can be regarded as the change of the expected outcome for one unit change of the mediator value and the treatment (exposure) remains unchanged at value t. These are due to our linear additive model assumption for the response as shown in (4.7). The second part of Lemma 4 states that if the model parameters $\boldsymbol{\theta}_Y$ in (4.7) are estimated by solving the weighted regression estimating equations (4.11), then $\hat{\tau}_{WR}$ and $\hat{\beta}_{WR}$, as estimators of the coefficient τ and β, are also DR estimators for their respective causal effects, which are also proved by utilizing Theorem 5 with details provided in Appendix E.

We then introduce the following lemmas stating the consistency of MLE and its associated plug-in estimators.

Lemma 5 (Consistency of the propensity model)**.** Under regularity conditions and provided that the model is correctly specified, we have, for each subject i

$$\hat{\boldsymbol{\theta}}_P \xrightarrow{p} \boldsymbol{\theta}_P \quad \text{and} \quad \widehat{\pi}_i(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}_P) \xrightarrow{p} \pi(\boldsymbol{X}_i), \quad \text{as} \quad n \to \infty.$$

Lemma 6 (Consistency of the mediator model)**.** Under regularity conditions and provided that the model is correctly specified, we have, for each subject i

$$\hat{\boldsymbol{\theta}}_M \xrightarrow{p} \boldsymbol{\theta}_M \quad \text{and} \quad \widehat{f}(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \hat{\boldsymbol{\theta}}_M) \xrightarrow{p} f(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\theta}_M), \quad \text{as} \quad n \to \infty.$$

Lemma 7 (Consistency of the response model)**.** Under regularity conditions and provided that the model is correctly specified, we have, for each subject i

$$\hat{\boldsymbol{\theta}}_Y \xrightarrow{p} \boldsymbol{\theta}_Y \quad \text{and} \quad \widehat{\mu}_Y(Y_i|\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_Y) \xrightarrow{p} \mu_Y(Y_i|\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\theta}_Y), \quad \text{as} \quad n \to \infty.$$

These lemmas are straightforward to prove and hold due to the consistency property of MLE and the continuous mapping theorem.

In the end, by summarizing Lemma 3 - 7, we have the MR property of the estimated causal effects of interest, which is stated by the following theorem.

Theorem 6 (MR property of the weighted regression estimator)**.** For the method introduced in Section 4.2, from the following 3 models:

1. The propensity model (the treatment (exposure) model): $P(T = t|\boldsymbol{X}; \boldsymbol{\theta}_P)$;
2. The mediator joint conditional density model: $f(\boldsymbol{m}|\boldsymbol{X}, T; \boldsymbol{\theta}_M)$;
3. The response model: $E(Y|\boldsymbol{X}, T, \boldsymbol{M}; \boldsymbol{\theta}_Y)$.

provided that two out of three of the aforementioned models are correctly specified, the causal effects of interest are consistently estimated, i.e., as $n \to \infty$,

$$\hat{TE} \xrightarrow{p} TE, \quad \widehat{DE} \xrightarrow{p} DE, \quad \widehat{IE} \xrightarrow{p} IE, \quad \text{and} \quad \widehat{IE}_j \xrightarrow{p} IE_j \quad \text{for} \quad j \in \{1, \ldots, J\}.$$

Consistency of the estimated grouped indirect effect is a corollary of Theorem 6.

Asymptotic properties of the weighted regression MR estimator

In the proposed method, the parameters consist of three parts: $\boldsymbol{\theta}_P$ for the propensity model ($\mathcal{M}_P$); $\boldsymbol{\theta}_M$ for the mediator model ($\mathcal{M}_M$) and $\boldsymbol{\theta}_Y$ for the response model ($\mathcal{M}_Y$). We denote $\boldsymbol{\theta} = \{\boldsymbol{\theta}'_P, \boldsymbol{\theta}'_M, \boldsymbol{\theta}'_Y\}'$ as the vector consisting all parameters used in the model. We further denote $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ as the asymptotic variance-covariance matrix of the estimated parameters and $\boldsymbol{V}(\boldsymbol{\theta})$ are obtained using the robust covariance estimation (the sandwich estimation formula). For the three models and associated three sets of parameters, there are three sets of estimating functions correspondingly. Throughout the illustration, we denote all estimating equations using the notation $\boldsymbol{U}(\cdot)$.

We begin with $\boldsymbol{\theta}_P$. The estimating equation for $\boldsymbol{\theta}_P$ is denoted as $\boldsymbol{U}_P(\boldsymbol{\theta}_P)$, so $\hat{\boldsymbol{\theta}}_P$ is the solution to equation

$$\boldsymbol{U}_P(\boldsymbol{\theta}_P) = \sum_{i=1}^{n} \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P) = \boldsymbol{0}, \tag{4.16}$$

where $\boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)$ is the treatment model estimating equation for each individual. In the proposed method, we do not specify any particular form for $\mathcal{M}_P$ (e.g. $\mathcal{M}_P$ can be either

logistic regression, tree-based model or more sophisticated non-parametric models), so (4.16) is the most general expression for $\boldsymbol{U}_P$. Detailed expressions of $\boldsymbol{U}_P$ under logistic model settings are presented in Section 3.2.5.

With respect to $\boldsymbol{\theta}_M$, we start with the mean parameters. $\hat{\boldsymbol{\Psi}}_M = \{\hat{\boldsymbol{\Psi}}_1', \ldots, \hat{\boldsymbol{\Psi}}_J'\}'$, where each of $\hat{\boldsymbol{\Psi}}_j$ is the solutions to the equation

$$\boldsymbol{U}_{\boldsymbol{\Psi},j}(\boldsymbol{\Psi}_j, \boldsymbol{\theta}_P) = \sum_{i=1}^{n} \boldsymbol{U}_{W,\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j, \boldsymbol{\theta}_P; M_{ij}, \boldsymbol{X}_i, T_i) = \boldsymbol{0},$$

where $\boldsymbol{U}_{W,\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j, \boldsymbol{\theta}_P; M_{ij}, \boldsymbol{X}_i, T_i)$ is the weight-adjusted estimating equation for mean parameters of the jth mediator for each individual such that

$$\boldsymbol{U}_{W,\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j, \boldsymbol{\theta}_P; M_{ij}, \boldsymbol{X}_i, T_i) = \omega_{M,i}(\boldsymbol{\theta}_P; \boldsymbol{X}_i)\boldsymbol{U}_{\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j; M_{ij}, \boldsymbol{X}_i, T_i),$$

and

$$\boldsymbol{U}_{\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j; M_{ij}, \boldsymbol{X}_i, T_i) = [M_{ij} - \mu_{ij}(\boldsymbol{X}_i, T_i; \boldsymbol{\Psi}_j)]\boldsymbol{W}_i.$$

Detailed expressions of $\boldsymbol{U}_{\boldsymbol{\Psi},j,i}$ under the linear model settings are presented in Section 3.2.5.

For the covariance parameters,

$$\boldsymbol{U}_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P) = \sum_{i=1}^{n} \boldsymbol{U}_{W,\boldsymbol{\Sigma},i}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P) = \boldsymbol{0},$$

where $\boldsymbol{U}_{W,\boldsymbol{\Sigma},i}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P)$ is the weight-adjusted estimating equation for the covariance parameters of the mediator model for each individual, such that,

$$\boldsymbol{U}_{W,\boldsymbol{\Sigma},i}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P) = \text{vec}\left\{\frac{\partial}{\partial \boldsymbol{\Sigma}} l_{W,\boldsymbol{M},i}(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P)\right\},$$

where vec$\{\boldsymbol{A}\}$ denotes vectorization of a matrix $\boldsymbol{A}$ and $l_{W,\boldsymbol{M},i}(\boldsymbol{M}_i|\boldsymbol{X}_i, T_i; \boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P)$ is the weighted log-likelihood function of $\boldsymbol{M}_i$ given $\boldsymbol{X}_i$ and T_i from (4.14). Notice that the estimating equation for covariance (correlation) parameters can be different under different model settings.

Therefore, by stacking all estimating equations, we have

$$\boldsymbol{U}'_M(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P) = [\boldsymbol{U}'_{\boldsymbol{\Psi},1}(\boldsymbol{\Psi}_1, \boldsymbol{\theta}_P), \ldots, \boldsymbol{U}'_{\boldsymbol{\Psi},J}(\boldsymbol{\Psi}_J, \boldsymbol{\theta}_P), \boldsymbol{U}'_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P)]',$$

and for each individual

$$\boldsymbol{U}'_{W,M,i}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P) = [\boldsymbol{U}'_{W,\boldsymbol{\Psi},1,i}(\boldsymbol{\Psi}_1, \boldsymbol{\theta}_P), \ldots, \boldsymbol{U}'_{W,\boldsymbol{\Psi},J,i}(\boldsymbol{\Psi}_J, \boldsymbol{\theta}_P), \boldsymbol{U}'_{W,\boldsymbol{\Sigma},i}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P)]'.$$

For $\boldsymbol{\theta}_Y$, $\hat{\boldsymbol{\theta}}_Y$ are the solutions to the equation

$$\boldsymbol{U}_Y(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) = \sum_{i=1}^{n} \boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) = \mathbf{0}.$$

where $\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P)$ is the weight-adjusted estimating equation for response model parameter $\boldsymbol{\theta}_Y$ for each individual such that

$$\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) = \omega_{Y,i}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P)\boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y)$$

and

$$\boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y) = [Y_i - \mu_{Y,i}(\boldsymbol{X}_i, T_i, \boldsymbol{M}_i; \boldsymbol{\theta}_Y)]\boldsymbol{Z}_i.$$

Detailed expressions of $\boldsymbol{U}_{Y,i}$ under the linear model settings are presented in Section 3.2.5.

Combining all estimating equations, we have,

$$\begin{aligned}
&\boldsymbol{U}(\boldsymbol{\theta}) \\
=&\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}) \\
=&\sum_{i=1}^{n} \left[\boldsymbol{U}'_{P,i}(\boldsymbol{\theta}_P), \boldsymbol{U}'_{W,M,i}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P), \boldsymbol{U}'_{W,Y,i}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P)\right]' \\
=&\left[\sum_{i=1}^{n} \boldsymbol{U}'_{P,i}(\boldsymbol{\theta}_P), \sum_{i=1}^{n} \boldsymbol{U}'_{W,M,i}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P), \sum_{i=1}^{n} \boldsymbol{U}'_{W,Y,i}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P)\right]'.
\end{aligned}$$

Due to the unbiased property of estimating equations, under regularity conditions and provided that the models are all correctly specified, $\boldsymbol{\theta}$ is consistent and asymptotically normal, such that,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{N}(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta})) \quad \text{as} \quad n \to \infty.$$

The asymptotic variance-covariance matrix has the form

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathcal{I}^{-1}(\boldsymbol{\theta})\boldsymbol{C}(\boldsymbol{\theta})\mathcal{I}(\boldsymbol{\theta}),$$

where

$$\mathcal{I}(\boldsymbol{\theta}) = E\left\{-\partial \boldsymbol{U}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'\right\}, \qquad \text{and} \qquad \boldsymbol{C}(\boldsymbol{\theta}) = E\left\{\boldsymbol{U}_i(\boldsymbol{\theta})\boldsymbol{U}_i'(\boldsymbol{\theta})\right\}.$$

In the equation, $\mathcal{I}(\boldsymbol{\theta})$ can be partitioned that,

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\theta}) &= -E\begin{pmatrix} \frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'} & \frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y'} \\ \frac{\partial \boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'} & \frac{\partial \boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y'} \\ \frac{\partial \boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'} & \frac{\partial \boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y'} \end{pmatrix} \\
&= -E\begin{pmatrix} \frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'} & \mathbf{0} & \mathbf{0} \\ \frac{\partial \boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'} & \mathbf{0} \\ \frac{\partial \boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'} & \frac{\partial \boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y'} \end{pmatrix} \\
&= \begin{pmatrix} -E\left(\frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'}\right) & \mathbf{0} & \mathbf{0} \\ -E\left(\frac{\boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'}\right) & -E\left(\frac{\boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'}\right) & \mathbf{0} \\ -E\left(\frac{\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'}\right) & -E\left(\frac{\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'}\right) & -E\left(\frac{\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y'}\right) \end{pmatrix}.
\end{aligned}$$

If we further denote

$$\mathcal{I}_P(\boldsymbol{\theta}_P) = -E\left(\frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'}\right),$$

$$\mathcal{I}_{M,P}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P) = -E\left(\frac{\boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'}\right),$$

$$\mathcal{I}_{M,M}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P) = -E\left(\frac{\boldsymbol{U}_{W,M,i}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'}\right);$$

and

$$\mathcal{I}_{Y,P}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P) = -E\left(\frac{\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'}\right),$$

$$\mathcal{I}_{Y,M}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P) = -E\left(\frac{\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'}\right),$$

$$\mathcal{I}_{Y,Y}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P) = -E\left(\frac{\boldsymbol{U}_{W,Y,i}(\boldsymbol{\theta}_Y,\boldsymbol{\theta}_M,\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y'}\right),$$

then $\mathcal{I}(\boldsymbol{\theta})$ can be partitioned to

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{I}_P(\boldsymbol{\theta}_P) & \mathbf{0} & \mathbf{0} \\ \mathcal{I}_{M,P}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P) & \mathcal{I}_{M,M}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P) & \mathbf{0} \\ \mathcal{I}_{Y,P}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) & \mathcal{I}_{Y,M}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) & \mathcal{I}_{Y,Y}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) \end{pmatrix}$$

With $\boldsymbol{V}(\boldsymbol{\theta})$, the delta-method could be applied to get the asymptotic distributions of the estimated causal effects of interest, such that

$$\sqrt{n}[\kappa(\hat{\boldsymbol{\theta}}) - \kappa(\boldsymbol{\theta})] \xrightarrow{d} N(0, \nabla'\kappa(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})\nabla\kappa(\boldsymbol{\theta})), \tag{4.17}$$

where $\kappa(\boldsymbol{\theta})$ denotes the causal effects of interest, that can be expressed as functions of parameters $\boldsymbol{\theta}$ and $\nabla\kappa(\boldsymbol{\theta})$ denotes the gradient of $\kappa(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$.

4.2.5 Closed form results under particular model settings

The propensity model

A logistic regression model is used to model the treatment. We assume:

$$g[\pi(\boldsymbol{X}; \boldsymbol{\theta}_P)] = \text{logit}[P(T = 1|\boldsymbol{X}; \boldsymbol{\theta}_P)] = \boldsymbol{X}'_P\boldsymbol{\xi},\ i = 1, \ldots, n, \tag{4.18}$$

where $\text{logit}(p) = \log[p/(1-p)]$ is the logit link function, $\boldsymbol{\theta}_P = \boldsymbol{\xi}$ denotes the parameters used for the propensity model and $\boldsymbol{X}$ denotes the $p+1$-dimensional covariates that include an intercept.

Under such a setting, MLE can be used to estimate $\boldsymbol{\theta}_P$, with the score vector

$$S_P(\boldsymbol{\theta}_P) = \sum_{i=1}^{n} S_{P,i}(\boldsymbol{\theta}_P) = \sum_{i=1}^{n} [T_i - \text{expit}(\boldsymbol{X}'_i\boldsymbol{\theta}_P)]\boldsymbol{X}_i;$$

so the estimating equations equal to the score equations such that $\boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P) = S_{P,i}(\boldsymbol{\theta}_P)$. Further,

$$\mathcal{I}_P(\boldsymbol{\theta}_P) = \frac{\partial\, S_P(\boldsymbol{\theta}_P)}{\partial\, \boldsymbol{\theta}_P} = \sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}'_i\text{expit}(\boldsymbol{X}'_i\boldsymbol{\theta}_P)[1 - \text{expit}(\boldsymbol{X}'_i\boldsymbol{\theta}_P)],$$

where $\text{expit}(\cdot)$ is the inverse of the logit link function such that $\text{expit}(x) = g^{-1}(x) = \exp(x)/[1 + \exp(x)]$.

The mediator model

The multivariate normal distribution is assumed for the joint distribution of mediators. The marginal models are assumed as in (4.1). Following (4.13),

$$\hat{\boldsymbol{\Psi}}_j = \left(\sum_{i=1}^{n} \omega_{M,i} \boldsymbol{W}_i \boldsymbol{W}_i' \right)^{-1} \left(\sum_{i=1}^{n} \omega_{M,i} \boldsymbol{W}_i M_{ij} \right),$$

for $j = 1, \ldots, J$. Since $\hat{\mu}_{ij} = \boldsymbol{W}_i' \hat{\boldsymbol{\theta}}_j$ and $\hat{\boldsymbol{\mu}}_i = \{\hat{\mu}_{i1}, \ldots, \hat{\mu}_{iJ}\}'$, then,

$$\hat{\boldsymbol{\Sigma}} = \left[\sum_{i=1}^{n} \omega_{M,i} (\boldsymbol{M}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{M}_i - \hat{\boldsymbol{\mu}}_i)' \right] / \left(\sum_{i=1}^{n} \omega_{M,i} \right)$$

Additionally, we have, the estimating equations

$$\boldsymbol{U}_{W,\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j, \boldsymbol{\theta}_P) = \omega_{M,i}(\boldsymbol{\theta}_P) \boldsymbol{W}_i (M_{ij} - \boldsymbol{W}_i' \boldsymbol{\Psi}_j).$$

and

$$\boldsymbol{U}_{W,\boldsymbol{\Sigma},i}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P) = \operatorname{vec} \left\{ \omega_{M,i}(\boldsymbol{\theta}_P) [\boldsymbol{\Sigma} - (\boldsymbol{M}_i - \boldsymbol{\mu}_i)(\boldsymbol{M}_i - \boldsymbol{\mu}_i)'] \right\}.$$

Furthermore, we have,

$$\frac{\partial \boldsymbol{U}_{W,\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j, \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} = \boldsymbol{W}_i (M_{ij} - \mu_{ij}) \left\{ -\frac{T_i}{\pi_i^2(\boldsymbol{X}_i; \boldsymbol{\theta}_P)} + \frac{1 - T_i}{[1 - \pi_i(\boldsymbol{X}_i; \boldsymbol{\theta}_P)]^2} \right\} \frac{\partial \pi_i(\boldsymbol{X}_i; \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P},$$

where

$$\frac{\partial \pi_i(\boldsymbol{X}_i; \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} = \frac{\partial \operatorname{expit}(\boldsymbol{X}_i' \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} = \boldsymbol{X}_i \frac{\exp(\boldsymbol{X}_i' \boldsymbol{\theta}_P)}{[1 + \exp(\boldsymbol{X}_i' \boldsymbol{\theta}_P)]^2},$$

and

$$\frac{\partial \boldsymbol{U}_{W,\boldsymbol{\Psi},j,i}(\boldsymbol{\Psi}_j, \boldsymbol{\theta}_P)}{\partial \boldsymbol{\Psi}_j} = \omega_{M,i}(\boldsymbol{\theta}_P) \boldsymbol{W}_i \boldsymbol{W}_i'.$$

The higher-order derivative of $\boldsymbol{U}_{W,\boldsymbol{\Sigma},i}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P)$ (Particularly $\partial \boldsymbol{U}_{W,\boldsymbol{\Sigma},i}(\boldsymbol{\Sigma}, \boldsymbol{\Psi}_M, \boldsymbol{\theta}_P)/\partial \boldsymbol{\Sigma}$) involves complex matrix calculation and is therefore omitted here.

The response model

The response model is assumed as in (4.7), then

$$\hat{\boldsymbol{\theta}}_Y = \left(\sum_{i=1}^{n} \omega_{Y,i} \boldsymbol{Z}_i \boldsymbol{Z}_i'\right)^{-1} \left(\sum_{i=1}^{n} \omega_{Y,i} \boldsymbol{Z}_i \boldsymbol{Y}_i'\right)$$

Additionally, we have the score vector

$$\boldsymbol{U}_{Y,W,i}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) = \omega_{Y,i}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_P)(Y_i - \mu_{Yi})\boldsymbol{Z}_i,$$

and

$$\begin{aligned}
&\mathcal{I}_{Y,P}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) \\
=&\frac{\partial \boldsymbol{U}_Y(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} \\
=&\sum_{i=1}^{n} (Y_i - \mu_{Yi}) \boldsymbol{Z}_i \frac{\partial \omega_{Y,i}}{\partial \boldsymbol{\theta}_P} \\
=&\sum_{i=1}^{n} (Y_i - \mu_{Yi}) \frac{1}{f_{\boldsymbol{M},i}(\boldsymbol{\theta}_M)} \left\{ -\frac{T_i}{\pi_i^2(\boldsymbol{\theta}_P)} + \frac{1 - T_i}{[1 - \pi_i(\boldsymbol{\theta}_P)]^2} \right\} \boldsymbol{Z}_i \frac{\partial \pi_i(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P},
\end{aligned}$$

where

$$\frac{\partial \pi_i(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} = \frac{\partial \operatorname{expit}(\boldsymbol{X}_i' \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} = \boldsymbol{X}_i \frac{\exp(\boldsymbol{X}_i' \boldsymbol{\theta}_P)}{[1 + \exp(\boldsymbol{X}_i' \boldsymbol{\theta}_P)]^2},$$

$$\begin{aligned}
&\mathcal{I}_{Y,M}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) \\
=&\frac{\partial \boldsymbol{U}_Y(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M} \\
=&\sum_{i=1}^{n} (Y_i - \mu_{Yi}) \boldsymbol{Z}_i \frac{\partial \omega_{Y,i}}{\partial \boldsymbol{\theta}_M} \\
=&\sum_{i=1}^{n} (Y_i - \mu_{Yi}) \frac{1}{f_{\boldsymbol{M},i}^2(\boldsymbol{\theta}_M)} \left[\frac{T_i}{\pi_i(\boldsymbol{\theta}_P)} + \frac{1 - T_i}{1 - \pi_i(\boldsymbol{\theta}_P)} \right] \boldsymbol{Z}_i \frac{\partial f_{\boldsymbol{M},i}(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M},
\end{aligned}$$

and

$$\mathcal{I}_{Y,Y}(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P) = \frac{\partial \boldsymbol{U}_Y(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M, \boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y} = \sum_{i=1}^{n} \omega_{Y,i}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_M) \boldsymbol{Z}_i \boldsymbol{Z}_i'.$$

4.3 Method 2: augmented inverse propensity weighting (AIPW) approach

Another method of constructing MR estimator we consider is utilizing the augmented inverse propensity weighting (AIPW) approach. For this approach, we estimate the three models, the propensity model ($\mathcal{M}_P$), the mediator model ($\mathcal{M}_M$) and the response model ($\mathcal{M}_Y$) separately, but the potential outcome is imputed in a sophisticated way combining the three models such that MR properties are achieved. We extend the idea of Tchetgen et al. [128] to accommodate multiple mediators.

4.3.1 Model

For the AIPW approach, we do not impose any particular forms for $\mathcal{M}_P$, $\mathcal{M}_M$ and $\mathcal{M}_Y$ and we write the models in the general forms. The estimation procedures are different from the WR approach: the three models are estimated separately first, then the potential outcomes are imputed as follows using the three fitted models, and hereby the causal effects of interest are then estimated with multiple robustness.

For potential outcomes with the form $Y(t, \boldsymbol{M}(t))$, if we denote it as $\kappa(t)$, where $t = 0, 1$, then it is estimated as follow:

$$\kappa(t) = P_n\left\{\frac{I(T=t)}{P(T=t|\boldsymbol{X})}[Y - r_1(t,\boldsymbol{X})] + r_1(t,\boldsymbol{X})\right\}, \tag{4.19}$$

where $r_1(t, \boldsymbol{X})$ is an imputation model for potential outcome $Y(t, \boldsymbol{M}(t))$ conditioning on $\boldsymbol{X}$ and it is obtained using the mediation formula introduced in Chapter 1, which is stated below (throughout this chapter, we assume $\boldsymbol{M}$ has a density),

$$\begin{aligned} r_1(t,\boldsymbol{X}) &= \int_{\boldsymbol{m}} E(Y|T=t,\boldsymbol{X},\boldsymbol{M})\ \mathrm{d}F(\boldsymbol{m}|T=t,\boldsymbol{X}) \\ &= \int_{\boldsymbol{m}} E(Y|T=t,\boldsymbol{X},\boldsymbol{M})\ f(\boldsymbol{m}|T=t,\boldsymbol{X})\ \mathrm{d}\boldsymbol{m}. \end{aligned} \tag{4.20}$$

For potential outcome with the form $Y(t, \boldsymbol{M}(t'))$, where $t \neq t'$, if we denote it as $\kappa(t, t')$, then it is estimated as:

$$\begin{aligned} \kappa(t,t') =& P_n\left\{\frac{I(T=t)f(\boldsymbol{M}|T=t',\boldsymbol{X})}{P(T=t|\boldsymbol{X})f(\boldsymbol{M}|T=t,\boldsymbol{X})}[Y - E(Y|T=t,\boldsymbol{X},\boldsymbol{M})]\right. \\ &\left. + \frac{I(T=t')}{P(T=t'|\boldsymbol{X})}[E(Y|T=t,\boldsymbol{X},\boldsymbol{M}) - r_2(t,t',\boldsymbol{X})] + r_2(t,t',\boldsymbol{X})\right\}, \end{aligned} \tag{4.21}$$

where similar to (4.20), $r_2(t, t', \boldsymbol{X})$ is an imputation model for potential outcome $Y(t, \boldsymbol{M}(t'))$ conditioning on $\boldsymbol{X}$ and it has the form

$$\begin{aligned} r_2(t, t', \boldsymbol{X}) &= \int_{\boldsymbol{m}} E(Y|T = t, \boldsymbol{X}, \boldsymbol{M}) \, \mathrm{d}F(\boldsymbol{m}|T = t', \boldsymbol{X}) \\ &= \int_{\boldsymbol{m}} E(Y|T = t, \boldsymbol{X}, \boldsymbol{M}) \, f(\boldsymbol{m}|T = t', \boldsymbol{X}) \, \mathrm{d}\boldsymbol{m}. \end{aligned} \tag{4.22}$$

Note that $r_1(t, \boldsymbol{X}) = r_2(t, t, \boldsymbol{X})$ and $r_1(t, \boldsymbol{X})$ is a simplified notation. Additionally, both imputation models $r_1(t, \boldsymbol{X})$ and $r_2(t, t', \boldsymbol{X})$ are functions of $\mathcal{M}_M$ and $\mathcal{M}_Y$.

Estimation of causal effects is then conducted by investigating functions of estimated potential outcomes, according to their definitions. For example, in a linear framework, according to our definition in Chapter 1,

$$IE(1) = \kappa(1) - \kappa(1, 0),$$

and

$$DE(0) = \kappa(1, 0) - \kappa(0).$$

The following algorithm summarizes the estimation procedure in practice.

Algorithm 3. Algorithm of the MR estimator using augmented inverse propensity weighting approach

Step 1: We estimate the three model parameters separately, which are denoted as $\tilde{\boldsymbol{\theta}}_P$, $\tilde{\boldsymbol{\theta}}_M$ and $\tilde{\boldsymbol{\theta}}_Y$ respectively, to distinguish them from parameters estimated by weighted regression approach introduced in Section 4.2.

Step 2: We fit the three models separately, which are denoted as:

1. The fitted treatment model: $\tilde{\pi}_i(\boldsymbol{X}_i; \tilde{\boldsymbol{\theta}}_P) = P(T = 1|\boldsymbol{X}_i; \tilde{\boldsymbol{\theta}}_P)$;
2. The fitted mediator model under treatment t:

$$\tilde{f}_i(\boldsymbol{M}_i|\boldsymbol{X}_i, t; \tilde{\boldsymbol{\theta}}_M) = f_i(\boldsymbol{M}_i|\boldsymbol{X}_i, t; \tilde{\boldsymbol{\theta}}_M)$$

3. The fitted response model under treatment t:

$$\tilde{\mu}_Y(\boldsymbol{X}_i, t, \boldsymbol{M}_i; \tilde{\boldsymbol{\theta}}_Y) = E(Y_i|\boldsymbol{X}_i, t, \boldsymbol{M}_i; \tilde{\boldsymbol{\theta}}_Y)$$

Similarly, the fitted models are denoted with a tilde mark to distinguish them from the ones introduced in Section 4.2.

Step 3: Following (4.19) - (4.22), we impute the potential outcomes with the fitted models from Step 2. For example: for the potential outcome $Y(t, \boldsymbol{M}(t))$, where $t = 0, 1$, if we denote the imputed value as $\tilde{\kappa}(t)$, then

$$\tilde{\kappa}(t) = P_n\Big\{\frac{I(T_i = t)}{T_i\,\tilde{\pi}_i(\boldsymbol{X}_i;\tilde{\boldsymbol{\theta}}_P) + (1-T_i)[1-\tilde{\pi}_i(\boldsymbol{X}_i;\tilde{\boldsymbol{\theta}}_P)]}[Y_i - \tilde{r}_1(t, \boldsymbol{X}_i;\tilde{\boldsymbol{\theta}}_M, \tilde{\boldsymbol{\theta}}_Y)] + \tilde{r}_1(t, \boldsymbol{X}_i;\tilde{\boldsymbol{\theta}}_M, \tilde{\boldsymbol{\theta}}_Y)\Big\},$$

where

$$\tilde{r}_1(t, \boldsymbol{X}_i;\tilde{\boldsymbol{\theta}}_M, \tilde{\boldsymbol{\theta}}_Y) = \int_{\boldsymbol{m}} \tilde{\mu}_{Y,i}(\boldsymbol{X}_i, t, \boldsymbol{m};\tilde{\boldsymbol{\theta}}_Y)\ \tilde{f}(\boldsymbol{m}|\boldsymbol{X}_i, t;\tilde{\boldsymbol{\theta}}_M)\ \mathrm{d}\boldsymbol{m}$$

Step 4: Calculate the causal effects of interest using functions of imputed potential outcomes according to their definitions.

4.3.2 Statistical inference

We denote $\boldsymbol{\theta} = \{\boldsymbol{\theta}'_P, \boldsymbol{\theta}'_M, \boldsymbol{\theta}'_Y\}'$ as the vector consisting all parameters used in the model. Note that $\boldsymbol{\theta}, \boldsymbol{\theta}_P, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_Y$ are different from those in Section 4.2. Similar to Section 4.2, we denote $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ as the asymptotic variance-covariance matrix of $\boldsymbol{\theta}$ that can be obtained as follows. We denote the overall estimating equation as $\boldsymbol{U}(\boldsymbol{\theta})$, with estimating equations for $\mathcal{M}_P$ as $\boldsymbol{U}_P(\boldsymbol{\theta}_P)$, for $\mathcal{M}_M$ as $\boldsymbol{U}_M(\boldsymbol{\theta}_M)$ and for $\mathcal{M}_Y$ as $\boldsymbol{U}_Y(\boldsymbol{\theta}_Y)$. It follows that (for simplicity of notations, we omit the data elements inside each function),

$$\begin{aligned}
&\boldsymbol{U}(\boldsymbol{\theta}) \\
=&\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}) \\
=&\sum_{i=1}^{n} \left[\boldsymbol{U}'_{P,i}(\boldsymbol{\theta}_P), \boldsymbol{U}'_{M,i}(\boldsymbol{\theta}_M), \boldsymbol{U}'_{Y,i}(\boldsymbol{\theta}_Y)\right]' \\
=&\left[\sum_{i=1}^{n} \boldsymbol{U}'_{P,i}(\boldsymbol{\theta}_P), \sum_{i=1}^{n} \boldsymbol{U}'_{M,i}(\boldsymbol{\theta}_M), \sum_{i=1}^{n} \boldsymbol{U}'_{Y,i}(\boldsymbol{\theta}_Y)\right]'
\end{aligned} \tag{4.23}$$

Since we do not specify particular forms for $\mathcal{M}_P$, $\mathcal{M}_M$ and $\mathcal{M}_Y$, we leave (4.23) as the general form of estimating equations for the models. Detailed expressions of $\boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)$, $\boldsymbol{U}_{M,i}(\boldsymbol{\theta}_M)$ and $\boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y)$ under specific model settings are presented in Section 4.3.3. Due to the unbiased property of estimating equations, under regularity conditions and provided that the models are all correctly specified, $\boldsymbol{\theta}$ is consistent and asymptotically normal, such that,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\theta})) \quad \text{as} \quad n \to \infty.$$

The asymptotic variance-covariance matrix has the form

$$\boldsymbol{V}(\boldsymbol{\theta}) = \mathcal{I}^{-1}(\boldsymbol{\theta})\boldsymbol{C}(\boldsymbol{\theta})\mathcal{I}(\boldsymbol{\theta}), \tag{4.24}$$

where

$$\mathcal{I}(\boldsymbol{\theta}) = E\left[-\partial \boldsymbol{U}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'\right], \qquad \text{and} \qquad \boldsymbol{C}(\boldsymbol{\theta}) = E\left[\boldsymbol{U}_i(\boldsymbol{\theta})\boldsymbol{U}_i'(\boldsymbol{\theta})\right].$$

In the equation, $\mathcal{I}(\boldsymbol{\theta})$ can be partitioned that,

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\theta}) &= -E\begin{pmatrix}
\frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_M'} & \frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_Y'} \\
\frac{\partial \boldsymbol{U}_{M,i}(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{M,i}(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M'} & \frac{\partial \boldsymbol{U}_{M,i}(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_Y'} \\
\frac{\partial \boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y)}{\partial \boldsymbol{\theta}_P'} & \frac{\partial \boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y)}{\partial \boldsymbol{\theta}_M'} & \frac{\partial \boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y)}{\partial \boldsymbol{\theta}_Y'}
\end{pmatrix} \\
&= \begin{pmatrix}
-E\left[\frac{\partial \boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P'}\right] & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{0} & -E\left[\frac{\partial \boldsymbol{U}_{M,i}(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M'})\right] & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{0} & -E\left[\frac{\partial \boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y)}{\partial \boldsymbol{\theta}_Y'}\right]
\end{pmatrix} \\
&= \begin{pmatrix}
\mathcal{I}_P(\boldsymbol{\theta}_P) & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{0} & \mathcal{I}_M(\boldsymbol{\theta}_M) & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{0} & \mathcal{I}_Y(\boldsymbol{\theta}_Y)
\end{pmatrix},
\end{aligned}$$

If MLE is used to fit $\mathcal{M}_P$, $\mathcal{M}_M$ and $\mathcal{M}_Y$, then $\boldsymbol{U}_{P,i}(\boldsymbol{\theta}_P)$, $\boldsymbol{U}_{M,i}(\boldsymbol{\theta}_M)$ and $\boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y)$ can equal to their respective Score equations (denoted as $S'_{P,i}(\boldsymbol{\theta}_P)$, $S'_{M,i}(\boldsymbol{\theta}_M)$ and $S'_{Y,i}(\boldsymbol{\theta}_Y)$), and $\mathcal{I}_p$, $\mathcal{I}_M$ and $\mathcal{I}_Y$ equals to the Fisher information matrix for each model respectively.

It follows that, from (4.19), by applying the delta method, we obtain:

$$\sqrt{n}[\hat{\kappa}_{MR}(t) - \kappa_{MR}(t)] \xrightarrow{d} N(0, \nabla'\kappa(t;\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})\nabla\kappa(t;\boldsymbol{\theta})), \tag{4.25}$$

where $\nabla\kappa(t;\boldsymbol{\theta})$ denotes the gradient of $\kappa(t;\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ that is to be derived in the following. Additionally, following (4.21) and applying the delta method, we have:

$$\sqrt{n}[\hat{\kappa}_{MR}(t,t') - \kappa_{MR}(t,t')] \xrightarrow{d} N(0, \nabla'\kappa(t,t';\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})\nabla\kappa(t,t';\boldsymbol{\theta})), \tag{4.26}$$

where $\nabla\kappa(t,t';\boldsymbol{\theta})$ denotes the gradient of $\kappa(t,t';\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ that is to be derived in the following. The asymptotic variances of the estimated causal effects are calculated in similar ways except for replacing κ by functions of it. For example, in a linear framework, according to our definition in Chapter 2,

$$\sqrt{n}[\widehat{IE}(1) - IE(1)] \xrightarrow{d} N(0, \nabla'\{\kappa(1) - \kappa(1,0)\}\boldsymbol{V}(\boldsymbol{\theta})\nabla\{\kappa(1) - \kappa(1,0)\}), \tag{4.27}$$

and

$$\sqrt{n}[\widehat{DE}(0) - DE(0)] \xrightarrow{d} N(0, \nabla'\{\kappa(1,0) - \kappa(0)\}\boldsymbol{V}(\boldsymbol{\theta})\nabla\{\kappa(1,0) - \kappa(0)\}). \tag{4.28}$$

The derivations of $\nabla\kappa(t;\boldsymbol{\theta})$, $\nabla\kappa(t,t';\boldsymbol{\theta})$ and $\boldsymbol{V}$ are as follows. We begin with $\nabla\kappa(t;\boldsymbol{\theta})$, since $\nabla\kappa(t;\boldsymbol{\theta}) = \sum_{i=1}^{n}\nabla\kappa(t,\boldsymbol{X}_i;\boldsymbol{\theta})$, for any $\kappa(t,\boldsymbol{X};\boldsymbol{\theta})$,

$$\begin{aligned}
&\nabla\kappa(t,\boldsymbol{X};\boldsymbol{\theta}) \\
=&\frac{\partial\kappa(t,\boldsymbol{X};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\
=&\frac{\partial\left\{\frac{I(T=t)}{\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}[Y - r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)] + r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)\right\}}{\partial\,\boldsymbol{\theta}} \\
=&I(T=t)Y[-\pi(\boldsymbol{X};\boldsymbol{\theta}_P)^{-2}]\frac{\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial\,\boldsymbol{\theta}} - I(T=t)r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)[-\pi(\boldsymbol{X};\boldsymbol{\theta}_P)^{-2}]\frac{\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial\,\boldsymbol{\theta}} \\
&-\frac{I(T=t)}{\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}\frac{\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}} + \frac{\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}}
\end{aligned}$$

If we write in a vector format,

$$\nabla\kappa(t,\boldsymbol{X};\boldsymbol{\theta}) = \begin{pmatrix} \partial\,\kappa(t,\boldsymbol{X};\boldsymbol{\theta})/\partial\,\boldsymbol{\theta}_P \\ \partial\,\kappa(t,\boldsymbol{X};\boldsymbol{\theta})/\partial\,\boldsymbol{\theta}_M \\ \partial\,\kappa(t,\boldsymbol{X};\boldsymbol{\theta})/\partial\,\boldsymbol{\theta}_Y \end{pmatrix},$$

where

$$\begin{aligned}
&\frac{\partial\,\kappa(t,\boldsymbol{X};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}_P} \\
=&I(T=t)Y[-\pi^{-2}(\boldsymbol{X};\boldsymbol{\theta}_P)]\frac{\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial\,\boldsymbol{\theta}_P} \\
&-I(T=t)r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)[-\pi^{-2}(\boldsymbol{X};\boldsymbol{\theta}_P)]\frac{\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial\,\boldsymbol{\theta}_P}
\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial\,\kappa(t,\boldsymbol{X};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}_M}\\
=&-I(T=t)\pi^{-1}(\boldsymbol{X};\boldsymbol{\theta}_P)\frac{\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_M}+\frac{\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_M}
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial\,\kappa(t,\boldsymbol{X};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}_Y}\\
=&-I(T=t)\pi^{-1}(\boldsymbol{X};\boldsymbol{\theta}_P)\frac{\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_Y}+\frac{\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_Y}
\end{aligned}
$$

respectively. The gradient functions involved include:

- The gradient of propensity model: $\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)/\partial\,\boldsymbol{\theta}_P$;
- The gradient of the potential outcome imputation model: $\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)/\partial\,\boldsymbol{\theta}_M$ and $\partial\,r_1(t,\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)/\partial\,\boldsymbol{\theta}_Y$.

Similarly, $\nabla\kappa(t,t';\boldsymbol{\theta})=\sum_{i=1}^{n}\nabla\kappa(t,t',\boldsymbol{X}_i;\boldsymbol{\theta})$, then for any $\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})$, we have:

$$
\nabla\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})=\begin{pmatrix}\partial\,\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})/\partial\,\boldsymbol{\theta}_P\\ \partial\,\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})/\partial\,\boldsymbol{\theta}_M\\ \partial\,\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})/\partial\,\boldsymbol{\theta}_Y\end{pmatrix},
$$

where

$$
\begin{aligned}
&\frac{\partial\,\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}_P}\\
=&\frac{I(T=t)f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)}{f(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)}[Y-\mu_Y(\boldsymbol{X},t,\boldsymbol{M};\boldsymbol{\theta}_Y)]\frac{\partial\,\pi^{-1}(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial\,\boldsymbol{\theta}_P}\\
&+I(T=t')[\mu_Y(\boldsymbol{X},t,\boldsymbol{M};\boldsymbol{\theta}_Y)-r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)]\frac{\partial\,[1-\pi(\boldsymbol{X};\boldsymbol{\theta}_P)]^{-1}}{\partial\,\boldsymbol{\theta}_P}\\
=&\frac{I(T=t)f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)}{f(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)}[Y-\mu_Y(\boldsymbol{X},t,\boldsymbol{M};\boldsymbol{\theta}_Y)][-\pi^{-2}(\boldsymbol{X};\boldsymbol{\theta}_P)]\,\frac{\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial\,\boldsymbol{\theta}_P}\\
&+I(T=t')[\mu_Y(\boldsymbol{X},t,\boldsymbol{M};\boldsymbol{\theta}_Y)-r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)][1-\pi(\boldsymbol{X};\boldsymbol{\theta}_P)]^{-2}\,\frac{\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial\,\boldsymbol{\theta}_P};
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial\,\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}_M}\\
=&I(T=t)\pi^{-1}(\boldsymbol{X};\boldsymbol{\theta}_P)[Y-\mu_Y(\boldsymbol{X},t,\boldsymbol{M})]f^{-1}(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)\,\frac{\partial\,f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)}{\partial\,\boldsymbol{\theta}_M}\\
&+I(T=t)\pi^{-1}(\boldsymbol{X};\boldsymbol{\theta}_P)[Y-\mu_Y(\boldsymbol{X},t,\boldsymbol{M})]f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)\,\frac{\partial\,f^{-1}(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)}{\partial\,\boldsymbol{\theta}_M}\\
&-I(T=t')[1-\pi(\boldsymbol{X};\boldsymbol{\theta}_P)]^{-1}\,\frac{\partial\,r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_M}+\frac{\partial\,r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_M}\\
=&I(T=t)\pi^{-1}(\boldsymbol{X};\boldsymbol{\theta}_P)[Y-\mu_Y(\boldsymbol{X},t,\boldsymbol{M})]f^{-1}(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)\,\frac{\partial\,f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)}{\partial\,\boldsymbol{\theta}_M}\\
&-I(T=t)\pi^{-1}(\boldsymbol{X};\boldsymbol{\theta}_P)[Y-\mu_Y(\boldsymbol{X},t,\boldsymbol{M})]f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)\\
&\quad f^{-2}(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)\,\frac{\partial\,f(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)}{\partial\,\boldsymbol{\theta}_M}\\
&-I(T=t')[1-\pi(\boldsymbol{X};\boldsymbol{\theta}_P)]^{-1}\,\frac{\partial\,r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_M}+\frac{\partial\,r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_M}
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial\,\kappa(t,t',\boldsymbol{X};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}_Y}\\
=&-\frac{I(T=t)f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)}{\pi(\boldsymbol{X};\boldsymbol{\theta}_P)f(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)}\,\frac{\partial\,\mu_Y(\boldsymbol{X},t,\boldsymbol{M};\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_Y}\\
&+\frac{I(T=t')}{1-\pi(\boldsymbol{X};\boldsymbol{\theta}_P)}\left[\frac{\partial\,\mu_Y(\boldsymbol{X},t,\boldsymbol{M};\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_Y}-\frac{\partial\,r(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_Y}\right]\\
&+\frac{\partial\,r(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)}{\partial\,\boldsymbol{\theta}_Y}
\end{aligned}
$$

respectively. The gradient functions involved include:

- The gradient of the propensity model: $\partial\,\pi(\boldsymbol{X};\boldsymbol{\theta}_P)/\partial\,\boldsymbol{\theta}_P$;
- The gradient of the mediator model: $\partial\,f(\boldsymbol{M}|\boldsymbol{X},t;\boldsymbol{\theta}_M)/\partial\,\boldsymbol{\theta}_M$ and $\partial\,f(\boldsymbol{M}|\boldsymbol{X},t';\boldsymbol{\theta}_M)/\partial\,\boldsymbol{\theta}_M$;
- The gradient of the response model: $\partial\,\mu_Y(\boldsymbol{X},t,\boldsymbol{M};\boldsymbol{\theta}_Y)/\partial\,\boldsymbol{\theta}_Y$;
- The gradient of the potential outcome imputation model: $\partial\,r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)/\partial\,\boldsymbol{\theta}_M$ and $\partial\,r_2(t,t',\boldsymbol{X};\boldsymbol{\theta}_M,\boldsymbol{\theta}_Y)/\partial\,\boldsymbol{\theta}_Y$.

Closed-form expressions of the gradients in the aforementioned equations, together with the estimated causal effects and variance under specific model settings are given in the next section.

4.3.3 Closed form results under particular model settings

We provide a standard specification of the three models used for MR estimations and derive closed-form expressions in this section. We also derive the closed-form asymptotic variances of the MR estimators under standard settings.

The propensity model

We assume the same model settings as with the weighted regression approach, which is introduced in Section 4.2.5. Therefore, we have the same derivations and conclusions as in Section 4.2.5.

The mediator model

We have similar model settings as with the weighted regression approach, which is introduced in Section 4.2.1. We assume the same marginal model as (4.1), (4.2) and (4.3), as well as the same joint distribution as (4.4).

However, the estimation results are different since there are no weights involved for the AIPW approach. Following Theorem 3 in Chapter 1, if we let

$$\overline{\boldsymbol{W}}_i = \begin{pmatrix} X_{i10}, & \ldots, & X_{i1q_1}, & T_i, & 0, & \ldots, & 0 \\ & & \vdots & & \ddots & \vdots & \\ 0, & & \ldots, & & X_{iJ0}, & \ldots, & X_{iJq_J}, & T_i \end{pmatrix},$$

which denotes the re-organized design matrix for the mediator model illustrated in the theorem, we have

$$\tilde{\boldsymbol{\Psi}}_M = \left(\sum_{i=1}^{n} \overline{\boldsymbol{W}}_i' \overline{\boldsymbol{W}}_i \right)^{-1} \left(\sum_{i=1}^{n} \overline{\boldsymbol{W}}_i' \boldsymbol{M}_i \right),$$

where $\boldsymbol{\Psi}_M$ denotes the mean parameters in a similar way and

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \left[(\boldsymbol{M}_i - \boldsymbol{\mu}_i)(\boldsymbol{M}_i - \boldsymbol{\mu}_i)' \right],$$

Additionally, we have

$$\boldsymbol{U}_{\boldsymbol{M},i}(\boldsymbol{\theta}_M) = [\boldsymbol{U}'_{\boldsymbol{\Psi},i}(\boldsymbol{\theta}_M), \boldsymbol{U}'_{\boldsymbol{\Sigma},i}(\boldsymbol{\theta}_M)]',$$

where

$$\boldsymbol{U}_{\boldsymbol{\Psi},i}(\boldsymbol{\theta}_M) = \overline{\boldsymbol{W}}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{M}_i - \boldsymbol{\mu}_i);$$

and

$$\boldsymbol{U}_{\boldsymbol{\Sigma},i}(\boldsymbol{\theta}_M) = \text{vec}\left[\boldsymbol{\Sigma} - (\boldsymbol{M}_i - \boldsymbol{\mu}_i)(\boldsymbol{M}_i - \boldsymbol{\mu}_i)'\right].$$

Furthermore, we have

$$\frac{\partial \boldsymbol{U}_{\boldsymbol{\Psi},i}(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\Psi}_M} = \overline{\boldsymbol{W}}_i'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{W}}_i.$$

Similarly, the higher-order derivative of $\boldsymbol{U}_{\boldsymbol{\Sigma},i}(\boldsymbol{\theta}_M)$ involves complex matrix calculation and is therefore omitted here.

The response model

The response model is assumed to be the same as introduced in Section 4.2.1. However, the estimation is different as there are no weights involved.

For the AIPW approach, we have

$$\tilde{\boldsymbol{\theta}}_Y = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.$$

Additionally, we have

$$\boldsymbol{U}_{Y,i}(\boldsymbol{\theta}_Y) = [Y_i - \mu_{Yi}(\boldsymbol{\theta}_Y)]\boldsymbol{Z}_i, \quad \text{and} \quad \mathcal{I}_{Y,i}(\boldsymbol{\theta}_Y) = \boldsymbol{Z}_i\boldsymbol{Z}_i'.$$

The imputation model

$r_1(t, \boldsymbol{X})$ and $r_2(t, t', \boldsymbol{X})$ are calculated in similar ways under our model settings and $r_1(t, \boldsymbol{X}) = r_2(t, t, \boldsymbol{X})$. Without loss of generality, we only show the results from $r_2(t, t', \boldsymbol{X})$. Under (4.22),

$$\begin{aligned} r_2(t, t', \boldsymbol{X}) &= \int_{\boldsymbol{m}} \mu_Y(\boldsymbol{X}, t, \boldsymbol{m}; \boldsymbol{\theta}_Y) f(\boldsymbol{m}|\boldsymbol{X}, t'; \boldsymbol{\theta}_M)\, \mathrm{d}\boldsymbol{m} \\ &= \tau t + \sum_{j=1}^{J} [\alpha_j t' + \boldsymbol{X}'\boldsymbol{\eta}_j]\beta_j + \boldsymbol{X}'\boldsymbol{\gamma}. \end{aligned}$$

Gradients and Asymptotic variance

Following Section 1.3.2, we derive the gradient functions for each model. Under the model settings,

The gradient of the propensity model:

$$\frac{\partial \pi(\boldsymbol{X};\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} = \frac{\partial \operatorname{expit}(\boldsymbol{X}'\boldsymbol{\theta}_P)}{\partial \boldsymbol{\theta}_P} = \frac{\exp(\boldsymbol{X}'\boldsymbol{\theta}_P)}{1+\exp(\boldsymbol{X}'\boldsymbol{\theta}_P)}\boldsymbol{X};$$

The gradient of the mediator model:

$$\frac{\partial f(\boldsymbol{M}|\boldsymbol{X},T;\boldsymbol{\theta}_M)}{\partial \boldsymbol{\Psi}_M} = \frac{\partial \boldsymbol{\Sigma}^{-\frac{1}{2}}\phi[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{M}-\boldsymbol{W}'\boldsymbol{\Psi}_M)]}{\partial \boldsymbol{\Psi}_M} = \phi'[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{M}-\boldsymbol{W}'\boldsymbol{\Psi}_M)]\boldsymbol{\Sigma}^{-1}\boldsymbol{W}';$$

and

$$\begin{aligned}
&\frac{\partial f(\boldsymbol{M}|\boldsymbol{X},T;\boldsymbol{\theta}_M)}{\partial \boldsymbol{\Sigma}} \\
=&\frac{\partial \boldsymbol{\Sigma}^{-\frac{1}{2}}\phi[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{M}-\boldsymbol{W}'\boldsymbol{\Psi}_M)]}{\partial \boldsymbol{\Sigma}} \\
=&-\frac{1}{2}\boldsymbol{\Sigma}^{-\frac{3}{2}}\phi[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{M}-\boldsymbol{W}'\boldsymbol{\Psi}_M)] - \frac{1}{2}\boldsymbol{\Sigma}^{-2}\phi'[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{M}-\boldsymbol{W}'\boldsymbol{\Psi}_M)](\boldsymbol{M}-\boldsymbol{W}'\boldsymbol{\Psi}_M)
\end{aligned}$$

The gradient of the response model:

$$\frac{\partial \mu_Y(\boldsymbol{X},T,\boldsymbol{M};\boldsymbol{\theta}_Y)}{\partial \boldsymbol{\theta}_Y} = \frac{\partial \boldsymbol{Z}'\boldsymbol{\theta}_Y}{\partial \boldsymbol{\theta}_Y} = \boldsymbol{Z},$$

The gradient of the potential response model:

$$\frac{\partial r_2(t,t',\boldsymbol{X})}{\partial \boldsymbol{\theta}_M} = \beta_j[t',\boldsymbol{\eta}_1',\ldots,\boldsymbol{\eta}_J']',$$

and

$$\frac{\partial r_2(t,t',\boldsymbol{X})}{\partial \boldsymbol{\theta}_Y} = [t,(\alpha_1 t'+\boldsymbol{X}'\boldsymbol{\eta}_1),\ldots,(\alpha_J t'+\boldsymbol{X}'\boldsymbol{\eta}_J),\boldsymbol{X}']'.$$

It therefore follows that, by (4.24), (4.25) and (4.26), asymptotic distributions of the estimated potential outcomes are obtained. And following (4.27) and (4.28), asymptotic distributions of the estimated causal effects are obtained in similar ways.

4.4 Simulation study

4.4.1 Simulation study set up

We conduct simulation studies to investigate the properties of the proposed MR methods. We introduce the general setups across each scenario first.

- For each subject i, we assume there are three covariates. The correct covariates are denoted as $X_{i,1}, X_{i,2}$ and $X_{i,3}$ and they are all generated from standard normal distributions independently.
- The exposure is generated as, $P(T_i = 1) = \text{expit}(0.3X_{i,1} + 0.3X_{i,2} + 0.3X_{i,3})$.
- For each subject, three mediators are generated following $E(M_{i,1}) = E(M_{i,2}) = E(M_{i,3}) = 0.5 + 1T_i + 0.3X_{i,1} + 0.3X_{i,2} + 0.3X_{i,3}$, with $Var(M_1) = Var(M_2) = Var(M_3) = 1^2$.
- With respect to their correlations, we specify $\rho_{1,2} = 0.5$ and $\rho_{1,3} = \rho_{2,3} = 0$, so that only M_1 and M_2 is correlated and M_3 is independent with either of them.
- The outcome $Y_i = 0.5 + 1M_{i,1} + 1M_{i,2} + 1M_{i,3} + 1T_i + 0.3X_{i,1} + 0.3X_{i,2} + 0.3X_{i,3} + \varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$.

We specify the correlation structure to mimic the scenario that appears in Chapter 2. Under these settings, we investigate the grouped indirect effect of M_1 and M_2, the individual indirect effect of M_3 and the direct effect.

To test the multiple robustness, we need to introduce the misspecified models, which require specifying the incorrect covariates. The incorrect covariates are denoted as Z_1, Z_2, Z_3. Following Kang et al.[67], Z_1, Z_2, Z_3 are generated as:

$$Z_1 = \exp(X_1/2); \quad Z_2 = X_2/\exp(1 + X_1) + 10; \quad \text{and} \quad Z_3 = (X_1 \cdot X_3/25 + 0.6)^3.$$

The incorrect covariates are generated such that the distributions of the incorrect covariates are similar to the true ones. In the following, "correct specification" means using the correct covariates $\boldsymbol{X}$ for the model while "incorrect specification" means using the incorrect counterparts $\boldsymbol{Z}$.

4.4.2 Scenarios

Because there are in total three models involved ($\mathcal{M}_P$, $\mathcal{M}_M$ and $\mathcal{M}_Y$) in each of the proposed MR estimators, we design $2^3 = 8$ scenarios to test the performance of the proposed method under each combination of model specifications. The notation "&" denotes correct model specification among different models. For example, "$\mathcal{M}_P\&\mathcal{M}_M$" denotes the scenario that the propensity and the mediator model are correctly specified while the response model is not. The term "None" denotes none of the three models is correctly specified.

For each scenario, both of the two proposed MR estimation methods are examined. In order to better examine the consistency property of either method, we design the study to investigate the behaviour of each method when the sample size increases. We specify two sample sizes: $n = 100$ reflects a moderate sample size while $n = 1000$ reflects a large one.

Performance measurements include bias and the 95% confidence interval coverage rates. Additionally, for each simulation study, we calculate the empirical standard error ("ESE", the square root of the empirical variance of the estimated causal quantities across each replication) and the average estimated standard error ("ASE", the sample average of the estimated standard error for each simulated sample). The standard deviations are calculated using methods presented in sections 3.1.5 and 3.2.4. For each study, $m = 1000$ Monte-Carlo replications are conducted.

4.4.3 Simulation study results and discussions

Table (4.1) and (4.2) present the simulation results of the WR approach under $n = 100$ and $n = 1000$ settings respectively while Table (4.3) and (4.4) present the simulation results of the AIPW approach under $n = 100$ and $n = 1000$ settings respectively.

In terms of the point estimates, we see that, for both methods, when all of the three models are correctly specified or at least two among the three models are correctly specified, we can get small biases for the estimations of the causal effects of interest. The biases reduce as the sample size increases. However, if only one of the three models is correctly specified or none of the three models is correctly specified, the biases tend to be much larger and they do not reduce as the sample size grows larger.

In terms of the variance estimates and the 95% CI coverage rates, for both methods and across different sample sizes, when all of the three models are correctly specified, the average estimated SDs are close to the ESEs and the 95% CI coverage rates are close to 0.95. When two among the three models are correctly specified, across different sample sizes, the SD estimations are slightly worse than when all of the models are correctly specified.

Similarly, the 95% CI coverage rates are slightly further from 0.95 compared with the results obtained under the scenarios that all models are correctly specified. On the other hand, if only one or none of the three models is correctly specified, the average estimated SDs are much different from their empirical counterparts and the 95% CI coverage rates deviate much from the true value of 0.95, with the worst results obtained when none of the three models is correctly specified. In addition, such differences are not mitigated as the sample size grows, on the contrary, they are enlarged when the sample size gets larger.

Both the WR and the AIPW methods achieve multiple robustness such that when two or more models among the three are correctly specified, the biases tend to become smaller with the increasing sample sizes. Additionally, the 95% CI coverage rates are close to 0.95 if two or more models are correctly specified compared with other cases. If we take a closer look at the two methods, we do not find significant differences between them. In terms of the bias, the estimated SD and the 95% CI coverage rate, both methods provide results in very comparable scales.

model	causal effect	estimate	bias	ESE	ASE	95% CI coverage
$\mathcal{M}_P\&\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	1.995	-0.005	0.410	0.395	0.941
	Individual IE	0.991	-0.009	0.257	0.233	0.905
	DE	1.008	0.008	0.311	0.373	0.972
$\mathcal{M}_P\&\mathcal{M}_M$	Grouped IE	2.069	0.069	0.420	0.415	0.952
	Individual IE	1.032	0.032	0.264	0.249	0.927
	DE	0.853	-0.147	0.318	0.423	0.982
$\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	1.993	-0.007	0.410	0.412	0.952
	Individual IE	0.991	-0.009	0.259	0.242	0.912
	DE	1.009	0.009	0.310	0.383	0.975
$\mathcal{M}_P\&\mathcal{M}_Y$	Grouped IE	2.033	0.033	0.423	0.502	0.972
	Individual IE	1.011	0.011	0.261	0.291	0.953
	DE	1.006	0.006	0.310	0.375	0.975
$\mathcal{M}_P$	Grouped IE	2.172	0.172	0.437	0.527	0.973
	Individual IE	1.117	0.117	0.276	0.312	0.951
	DE	0.782	-0.218	0.315	0.400	0.951
$\mathcal{M}_M$	Grouped IE	2.066	0.066	0.419	0.430	0.958
	Individual IE	1.031	0.031	0.266	0.258	0.926
	DE	0.938	-0.062	0.321	0.404	0.981
$\mathcal{M}_Y$	Grouped IE	2.270	0.270	0.442	0.436	0.922
	Individual IE	1.130	0.130	0.276	0.258	0.919
	DE	1.007	0.007	0.304	0.375	0.971
None	Grouped IE	2.425	0.425	0.461	0.449	0.855
	Individual IE	1.247	0.247	0.294	0.271	0.868
	DE	0.863	-0.137	0.313	0.373	0.954

Table 4.1: Performance of the WR approach under sample size $n = 100$

model	causal effect	estimate	bias	ESE	ASE	95% CI coverage
$\mathcal{M}_P\&\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	2.008	0.008	0.122	0.124	0.955
	Individual IE	1.003	0.003	0.083	0.082	0.947
	DE	0.999	-0.001	0.095	0.096	0.949
$\mathcal{M}_P\&\mathcal{M}_M$	Grouped IE	2.086	0.086	0.125	0.129	0.946
	Individual IE	1.041	0.041	0.086	0.086	0.936
	DE	0.945	-0.055	0.098	0.110	0.957
$\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	2.009	0.009	0.124	0.127	0.956
	Individual IE	1.002	0.002	0.083	0.084	0.951
	DE	0.999	-0.001	0.094	0.098	0.954
$\mathcal{M}_P\&\mathcal{M}_Y$	Grouped IE	2.022	0.022	0.127	0.158	0.982
	Individual IE	1.010	0.010	0.084	0.099	0.979
	DE	0.998	-0.002	0.095	0.097	0.951
$\mathcal{M}_P$	Grouped IE	2.175	0.175	0.134	0.168	0.881
	Individual IE	1.122	0.122	0.088	0.106	0.853
	DE	0.740	-0.260	0.098	0.109	0.310
$\mathcal{M}_M$	Grouped IE	2.086	0.086	0.129	0.132	0.909
	Individual IE	1.040	0.040	0.086	0.088	0.939
	DE	0.918	-0.082	0.099	0.104	0.886
$\mathcal{M}_Y$	Grouped IE	2.308	0.308	0.139	0.137	0.410
	Individual IE	1.152	0.152	0.091	0.092	0.625
	DE	0.998	-0.002	0.093	0.097	0.955
None	Grouped IE	2.480	0.480	0.147	0.143	0.072
	Individual IE	1.279	0.279	0.096	0.095	0.154
	DE	0.845	-0.155	0.097	0.100	0.664

Table 4.2: Performance of the WR approach under sample size $n = 1000$

4.5 Data application

The data application in this chapter continues the study from Chapter 3. In Chapter 3, applying the proposed method, we identify 15 DNA methylation loci that mediate the process between childhood trauma and long-term psychiatric disorder. Based on the correlation selection results, 13 of the 15 identified loci are grouped together, while 2 loci are considered separately. In this section, we estimate the causal effects of interest using the multiple robust methods proposed in this chapter.

However, there is an issue when applying the proposed multiply robust estimation

model	causal effect	estimate	bias	ESE	ASE	95% CI coverage
$\mathcal{M}_P\&\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	1.996	-0.004	0.431	0.426	0.947
	Individual IE	0.998	-0.002	0.26	0.264	0.943
	DE	1.008	0.008	0.473	0.380	0.933
$\mathcal{M}_P\&\mathcal{M}_M$	Grouped IE	2.006	0.006	0.436	0.456	0.954
	Individual IE	1.003	0.003	0.264	0.287	0.959
	DE	1.014	0.014	0.575	0.416	0.943
$\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	1.993	-0.007	0.431	0.424	0.944
	Individual IE	0.997	-0.003	0.261	0.265	0.940
	DE	1.011	0.011	0.5	0.384	0.933
$\mathcal{M}_P\&\mathcal{M}_Y$	Grouped IE	2.061	0.061	0.465	0.485	0.953
	Individual IE	1.028	0.028	0.287	0.299	0.962
	DE	0.993	-0.007	0.427	0.431	0.953
$\mathcal{M}_P$	Grouped IE	2.203	0.203	0.482	0.504	0.941
	Individual IE	1.136	0.136	0.306	0.314	0.941
	DE	0.775	-0.225	0.450	0.453	0.920
$\mathcal{M}_M$	Grouped IE	2.004	0.004	0.439	0.463	0.951
	Individual IE	1.002	0.002	0.264	0.294	0.961
	DE	1.098	0.098	0.637	0.427	0.934
$\mathcal{M}_Y$	Grouped IE	2.274	0.274	0.466	0.460	0.912
	Individual IE	1.134	0.134	0.281	0.282	0.921
	DE	0.998	-0.002	0.380	0.359	0.940
None	Grouped IE	2.430	0.43	0.488	0.489	0.861
	Individual IE	1.251	0.251	0.300	0.305	0.887
	DE	0.862	-0.138	0.406	0.379	0.918

Table 4.3: Performance of the AIPW approach under sample size $n = 100$

methods to the DNA methylation data. Currently, the proposed method requires the exposure variable to be binary, but in the data, the exposure (childhood trauma) is recorded as the total score of the Childhood Trauma Questionnaire (CTQ) and is measured on a continuous scale. To tackle the issue, we use the Bernstein and Fink [5] cut-off points to categorize continuous exposures into binary ones. For the Childhood Trauma Questionnaire, Bernstein and Fink's cut-off points classify "Low to Moderate" severity of childhood trauma as having: physical abuse ≥ 8; sexual abuse ≥ 6; emotional abuse ≥ 9; physical neglect ≥ 8; emotional neglect ≥ 10). Thus, the presence of maltreatment was considered if a participant had a CTQ score equal to or higher than the low to moderate cut-off point

model	causal effect	estimate	bias	ESE	ASE	95% CI coverage
$\mathcal{M}_P\&\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	2.009	0.009	0.125	0.125	0.955
	Individual IE	1.002	0.002	0.077	0.078	0.955
	DE	0.994	-0.006	0.126	0.113	0.946
$\mathcal{M}_P\&\mathcal{M}_M$	Grouped IE	2.012	0.012	0.126	0.135	0.960
	Individual IE	1.003	0.003	0.078	0.085	0.957
	DE	1.000	0.000	0.134	0.124	0.944
$\mathcal{M}_M\&\mathcal{M}_Y$	Grouped IE	2.009	0.009	0.124	0.123	0.950
	Individual IE	1.001	0.001	0.077	0.077	0.952
	DE	0.994	-0.006	0.124	0.111	0.936
$\mathcal{M}_P\&\mathcal{M}_Y$	Grouped IE	2.035	0.035	0.136	0.145	0.962
	Individual IE	1.014	0.014	0.083	0.090	0.964
	DE	0.993	-0.007	0.148	0.152	0.965
$\mathcal{M}_P$	Grouped IE	2.195	0.195	0.145	0.155	0.773
	Individual IE	1.133	0.133	0.089	0.095	0.736
	DE	0.709	-0.291	0.203	0.200	0.539
$\mathcal{M}_M$	Grouped IE	2.011	0.011	0.127	0.137	0.973
	Individual IE	1.002	0.002	0.078	0.087	0.969
	DE	1.099	0.099	0.138	0.126	0.897
$\mathcal{M}_Y$	Grouped IE	2.300	0.300	0.139	0.136	0.437
	Individual IE	1.146	0.146	0.083	0.084	0.603
	DE	0.994	-0.006	0.121	0.117	0.949
None	Grouped IE	2.470	0.470	0.147	0.147	0.096
	Individual IE	1.273	0.273	0.090	0.092	0.144
	DE	0.831	-0.169	0.140	0.139	0.705

Table 4.4: Performance of the AIPW approach under sample size $n = 1000$

for each maltreatment type [5]. However, then there comes another issue, in the dataset, there is only a total score for each subject and there is no information on detailed scores within each category of the questionnaire. Therefore, we sum the threshold values and regard each individual with a total higher than 41, which is the sum of the cut-off value for each category, as experienced childhood trauma. We acknowledge that this is a major limitation and further study can be conducted when data including detailed scores is available.

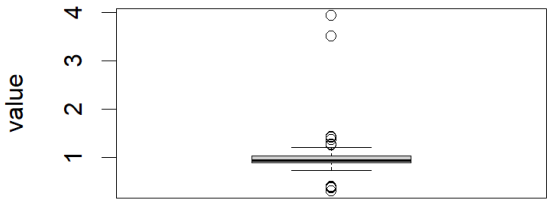
For both MR estimation methods, we need to calculate the weights $\hat{\omega}_M$ and $\hat{\omega}_Y$. In this data application, we use the stabilized weights (4.12) for both methods. To avoid extreme weights, we truncate the weights at the value 3 (i.e. for either $\hat{\omega}_M > 3$ or $\hat{\omega}_Y > 3$, we let $\hat{\omega}_M = 3$ or $\hat{\omega}_Y = 3$). Figure 4.1 presents boxplots illustrating the distribution of calculated weights before and after truncation.

The results of applying the proposed MR methods to the DNA methylation dataset are presented in Table 4.5. We can see that both methods provide similar estimates on

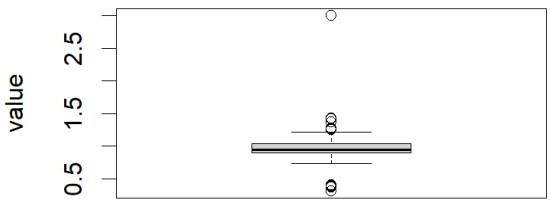
method	causal effects	value	95% CI
AIPW	Grouped IE of cg18634806, cg25626453, cg00096307, cg06992213, cg25448067, cg26657045, cg03643137, cg01696984, cg05292310, cg06001786, cg09211256, cg05051734, cg16180796	-242.776	(-379.519, -106.033)
	Individual IE of cg05608730	-72.994	(-126.042, -19.946)
	Individual IE of cg00578039	55.814	(-21.822, 133.45)
	DE	-90.461	(-192.781, 11.859)
WR	Grouped IE of cg18634806, cg25626453, cg00096307, cg06992213, cg25448067, cg26657045, cg03643137, cg01696984, cg05292310, cg06001786, cg09211256, cg05051734, cg16180796	-289.166	(-480.655, -97.678)
	Individual IE of cg05608730	-100.808	(-179.235, -22.38)
	Individual IE of cg00578039	66.375	(-2.86, 135.61)
	DE	-136.295	(-263.222, 90.632)

Table 4.5: Estimated grouped and individual indirect effects of DNA methylation loci using both the AIPW and WR approaches

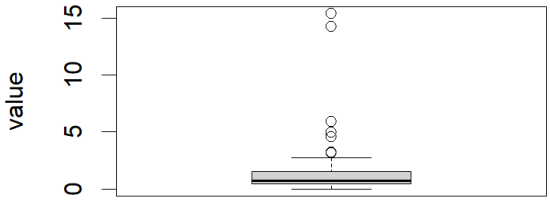
the effects of interest and they are also similar to the ones obtained from the conven-



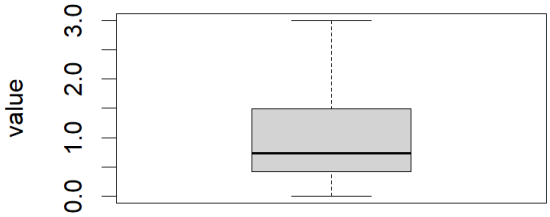
ω_M before truncation



ω_M after truncation



ω_Y before truncation



ω_Y after truncation

Figure 4.1: Calculated Weights applying the proposed MR methods to the DNA methylation dataset

tional method (without multiple robustness properties) that are presented and discussed in Chapter 2.

We discuss the results under the 95% significance level. The results from both methods show that, the grouped gene loci of cg18634806, cg25626453, cg00096307, cg06992213, cg25448067, cg26657045, cg03643137, cg01696984, cg05292310, cg06001786, cg09211256, cg05051734, cg16180796 impose significant indirect effects on the pathway from childhood trauma to long term psychiatric disorder. The single locus cg05608730 shows a significant indirect effect mediating the process also, but cg00578039 does not. Childhood trauma itself shows a negative direct effect leading to long-term psychiatric disorders, though the effect is insignificant from both methods.

Additionally, Table 3.24 in Chapter 3 presents the analysis results for the same data without using the MR estimation methods. The results obtained from either the MR or the non-MR methods are comparable: both the direction and the level of significance for each estimated causal effect remain the same, with only the magnitude slightly changed.

4.6 Discussion

In this chapter, we address the common challenge of model misspecification in causal inference and causal mediation analysis. Aiming to enhance robustness in estimations, we propose two approaches for constructing multiply robust estimators. The first way of constructing such a multiply robust estimator is rooted in the idea of weighted regression and the second one is based on the augmented inverse propensity weighting technique. Both of these estimators exhibit the desirable property of multiple robustness, which means that when at least two out of the three required models (the treatment model, the mediator model, and the response model) are correctly specified, the estimations remain consistent.

However, it's important to note that the weighted regression technique has stricter assumptions regarding model specifications. Specifically, it assumes linearity for both the mediator and the response models, although these linear assumptions can be relaxed with appropriate adjustments. In the simple causal inference problem without mediators involved, the weighted regression technique can be extended to accommodate generalized linear models for the outcome while still demonstrating a double robustness property. Future research can focus on extending this technique to handle non-linear settings in causal mediation analysis. Despite its stricter assumptions, the weighted regression approach offers the advantage of being straightforward to implement. It involves fitting two regressions for both the mediator and the response model with appropriate weights, which can be

easily accomplished using existing software packages. In contrast, the multiply robust estimator based on the augmented inverse propensity weighting method requires calculating more complex estimation formulas and involving the fitting of multiple models. However, it provides the advantage of broader applicability.

In our real data application study, we apply the proposed method to a psychiatric study dataset, investigating the mediation effects of multiple DNA methylation loci, considering them both as grouped and individual factors. To accommodate the model requirements, we transform the treatment variable from its original continuous scale to a binary scale. It's worth noting that the proposed method can also handle continuous treatments by replacing the propensity score with a generalized propensity score that still captures the distribution of treatment given baseline covariates. The multiple robustness property remains intact in this extended framework. For further details on this extension, readers may refer to Hirano et al. [45] for details.

We also acknowledge that there are major limitations associated with the proposed multiply robust estimation methods. In addition to the aforementioned linear assumptions imposed by the weighted regression-based estimator, similar to the method proposed in Chapter 2, both estimators require the assumption that the correlation structure among the mediators remains constant across different treatment assignments. We are aware that in real-world scenarios, this assumption may not always hold, and correlation structures can vary. Future research endeavors may focus on addressing this limitation by developing methods capable of accommodating varying correlation structures among mediators, thereby enhancing the applicability and robustness of our approach.

Chapter 5

Discussion and future works

This thesis presents a thorough investigation into the domain of causal mediation analysis in the presence of multiple correlated mediators. Chapter 2 serves as a foundation that conceptualizes the "uncausally related" relationship among the multiple mediators. Based on that, we introduce a general framework for conducting causal mediation analysis within these contexts. Particularly, in this chapter, we introduce a copula-based method that conveniently and effectively models the joint conditional distribution of multiple uncausally related mediators, which constitutes a pivotal component forming the analysis process. Moving forward, Chapter 3 is inspired by real-life problems where mediators are not typically given explicitly but need to be selected from a large set of candidate variables. Additionally, complex dependency structures among the mediators could pose challenges to estimations of causal effects. We therefore propose a novel method that both selects true mediators from the possibly high-dimensional set of candidates and reduces dependency structures among mediators. These selection processes enhance the precision and efficiency of subsequent investigation of causal effects to a great extent. Chapter 5, on the other hand, emphasizes more on the model misspecification issues that arise in the causal effects estimation process. We introduce two ways of constructing multiply robust estimators within the context of causal mediation analysis that are grounded in the concepts of weighted regression and augmented inverse propensity weighting approach respectively. The three chapters investigate the topic of causal mediation analysis with multiple correlated mediators through different yet interconnected aspects.

Although this thesis offers comprehensive insights into the subject, there remain unresolved challenges that future researchers may address. The issue of missing data presents a prominent challenge. Within the framework of causal mediation analysis involving multiple potentially related mediators, missingness on some of the mediators may cause trouble

since it may lead to biased estimations of both the direct effects and the indirect effects if not handled appropriately. Depending on the different missing mechanisms, researchers may address the issue from different aspects. One way to tackle the issue is by introducing another weight that accounts for the missingness, which leads to a weight-based approach. Alternatively, the imputation-based approach imputes the missing mediator values using available information from either other mediators or baseline covariates. However, it's noteworthy that for either approach, tackling the dependencies among the mediators may impose challenges. There are also occasions when missingness may appear in the covariates or outcomes, prompting the exploration of relevant solutions in these domains. As such, comprehensive studies into solutions for missing data problems within the context of causal mediation analysis with multiple potentially related mediators remain a valuable avenue for future research.

Another topic that may be of great interest to researchers is extending the joint modeling framework of mediators to the outcomes. We may consider the scenario where the outcome is multivariate and correlated. Compared with the univariate outcome scenario, one of the difficulties of performing mediation analysis when multivariate outcomes exist is that, it is hard to define and capture the causal effects under this setting. In the univariate case, causal effects can be defined by comparing differences in the potential outcomes under different treatment or mediator values. However, when dealing with multivariate outcomes, though a simple idea could be to consider causal effects for each outcome element separately, doing so may lead to information loss on retrieving the effects of treatment towards correlations of the multiple outcomes. Therefore, for the synthesis of information from mediation analyses on each outcome, some form of adjustment of the correlations among the multivariate outcomes is necessary, especially when conducting statistical inferences. The joint modeling framework discussed in this thesis may be extended to accommodate such settings.

Another possible approach to address the issue of multiple uncausally related mediators is utilizing the mixture effects models. Such an approach draws inspiration from the inherent logic of the formation of these uncausally related mediators. One of the most common reasons for the existence of uncausally related mediators is that there might be some unknown factors that are affecting all or some of the mediators simultaneously. The random effects can be employed to capture such unobservable factors. By modeling each mediator as a function of the exposure, the baseline covariates, and the random effects, we emulate the natural process underlying the formation of such uncausally related mediators, allowing us to capture their inherent correlations. However, a significant challenge in applying this method is defining causal effects when random effects are involved. In the context of longitudinal data modeling with mixture effect models, causal effects can be

defined from both marginal and conditional effects perspectives and similar approaches can be applied here.

Regarding multiply robust estimation, the method proposed in this thesis requires two out of the three models to be correctly specified for consistent estimation of the desired causal effects. However, there is room to further enhance the flexibility of multiply robust estimations. Future research may aim for methods that achieve consistent estimation with only one correctly specified model.

References

[1] Jeffrey M Albert and Suchitra Nelson. Generalized causal mediation analysis. *Biometrics*, 67(3):1028–1038, 2011.

[2] Lauren Y Atlas, Martin A Lindquist, Niall Bolger, and Tor D Wager. Brain mediators of the effects of noxious heat on pain. *PAIN®*, 155(8):1632–1648, 2014.

[3] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[4] 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.

[5] David P Bernstein, Laura Fink, Leonard Handelsman, and Jeffrey Foote. Childhood trauma questionnaire. *Assessment of family violence: A handbook for researchers and practitioners.*, 1998.

[6] David P Bernstein, Judith A Stein, Michael D Newcomb, Edward Walker, David Pogge, Taruna Ahluvalia, John Stokes, Leonard Handelsman, Martha Medrano, David Desmond, et al. Development and validation of a brief screening version of the childhood trauma questionnaire. *Child Abuse & Neglect*, 27(2):169–190, 2003.

[7] Simina M Boca, Rashmi Sinha, Amanda J Cross, Steven C Moore, and Joshua N Sampson. Testing multiple biological mediators simultaneously. *Bioinformatics*, 30(2):214–220, 2014.

[8] Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

[9] Simon J Bond, Vernon T Farewell, Catherine T Schentag, and Dafna D Gladman. Predictors for radiological damage in psoriatic arthritis: results from a single centre. *Annals of the Rheumatic Diseases*, 66(3):370–376, 2007.

[10] Xizhen Cai, Yeying Zhu, Yuan Huang, and Debashis Ghosh. High-dimensional causal mediation analysis based on partial linear structural equation models. *Computational Statistics & Data Analysis*, 174:107501, 2022.

[11] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

[12] Paul J Catalano and Louise M Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419):651–658, 1992.

[13] Viviana Celli. Causal mediation analysis in economics: Objectives, assumptions, models. *Journal of Economic Surveys*, 36(1):214–334, 2022.

[14] Jun Chen, Andrew N Evans, Ying Liu, Masaru Honda, Juan M Saavedra, and Greti Aguilera. Maternal deprivation in rats is associated with corticotrophin-releasing hormone (crh) promoter hypomethylation and enhances crh transcriptional responses to stress in adulthood. *Journal of Neuroendocrinology*, 24(7):1055–1064, 2012.

[15] Oliver Y Chén, Ciprian Crainiceanu, Elizabeth L Ogburn, Brian S Caffo, Tor D Wager, and Martin A Lindquist. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136, 2018.

[16] Rhian M Daniel, Bianca L De Stavola, SN Cousens, and Stijn Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

[17] Alex R de Leon and Bohsiu Wu. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30(2):175–185, 2011.

[18] Louis De Raeymaeker. The metaphysical problem of causality. *Philosophy Today*, 1(4):219–229, 1957.

[19] Bianca L De Stavola, Rhian M Daniel, George B Ploubidis, and Nadia Micali. Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. *American Journal of Epidemiology*, 181(1):64–80, 2015.

[20] Andriy Derkach, Ruth M Pfeiffer, Ting-Huei Chen, and Joshua N Sampson. High dimensional mediation analysis with latent variables. *Biometrics*, 75(3):745–756, 2019.

[21] Fabrizio Durante, Juan Fernandez-Sanchez, and Carlo Sempi. A topological proof of sklar's theorem. *Applied Mathematics Letters*, 26(9):945–948, 2013.

[22] Gabriel Escarela and Jacques F Carriere. Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4):333–349, 2003.

[23] Fredrik Falkenström, Sungho Park, and Cameron N McIntosh. Using copulas to enable causal inference from nonexperimental data: Tutorial and simulation studies. *Psychological Methods*, 28(2):301, 2023.

[24] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[25] Zicheng Fang. One standard error cross validation criterion for tuning a classification prediction model. Master's thesis, 2023.

[26] Marco Fornili, Vittorio Perduca, Agnès Fournier, Allan Jérolon, Marie Christine Boutron-Ruault, Gertraud Maskarinec, Gianluca Severi, and Laura Baglietto. Association between menopausal hormone therapy, mammographic density and breast cancer risk: results from the e3n cohort study. *Breast Cancer Research*, 23(1):1–10, 2021.

[27] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[28] Markus Frölich and Martin Huber. Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1645–1666, 2017.

[29] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

[30] Yuzhao Gao, Haitao Yang, Ruiling Fang, Yanbo Zhang, Ellen L Goode, and Yuehua Cui. Testing mediation effects in high-dimensional epigenetic studies. *Frontiers in genetics*, 10:1195–1208, 2019.

[31] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*, pages 3354–3361. IEEE, 2012.

[32] Irene Gijbels, Marek Omelka, and Noël Veraverbeke. Estimation of a copula when a covariate affects only marginal distributions. *Scandinavian Journal of Statistics*, 42(4):1109–1126, 2015.

[33] Dafna D Gladman, Miriam Ang, Li Su, Brian DM Tom, Catherine T Schentag, and Vernon T Farewell. Cardiovascular morbidity in psoriatic arthritis. *Annals of the Rheumatic Diseases*, 68(7):1131–1135, 2009.

[34] Dafna D Gladman and Vernon T Farewell. The role of hla antigens as indicators of disease progression in psoriatic arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 38(6):845–850, 1995.

[35] DD Gladman, R Shuckett, ML Russell, JC Thorne, and RK Schachter. Psoriatic arthritis (psa)-an analysis of 220 patients. *QJM: An International Journal of Medicine*, 62(2):127–141, 1987.

[36] Ralitza V Gueorguieva and Alan Agresti. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455):1102–1112, 2001.

[37] Douglas Gunzler, Tian Chen, Pan Wu, and Hui Zhang. Introduction to mediation analysis with structural equation modeling. *Shanghai Archives of Psychiatry*, 25(6):390, 2013.

[38] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[39] Peisong Han. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109(507):1159–1173, 2014.

[40] Christine Heim and Charles B Nemeroff. The role of childhood trauma in the neurobiology of mood and anxiety disorders: preclinical and clinical studies. *Biological Psychiatry*, 49(12):1023–1039, 2001.

[41] Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

[42] Miguel A Hernán and James M Robins. Causal inference, 2010.

[43] Raymond Hicks and Dustin Tingley. Causal mediation analysis. *The Stata Journal*, 11(4):605–619, 2011.

[44] Keisuke Hirano and Guido W Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3):259–278, 2001.

[45] Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

[46] Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[47] Fatemeh Saberi Hosnijeh, Pieter M Kolijn, Delphine Casabonne, Alexandra Nieters, Marta Solans, Sabine Naudin, Pietro Ferrari, James D Mckay, Elisabete Weiderpass, Vittorio Perduca, et al. Mediating effect of soluble b-cell activation immune markers on the association between anthropometric and lifestyle factors and lymphoma development. *Scientific Reports*, 10(1):1–12, 2020.

[48] Camelia E Hostinar, Mollie T McQuillan, Heather J Mirous, Kathryn E Grant, and Emma K Adam. Cortisol responses to a group public speaking task for adolescents: Variations by age, gender, and race. *Psychoneuroendocrinology*, 50:155–166, 2014.

[49] Lotte C Houtepen, Christiaan H Vinkers, Tania Carrillo-Roa, Marieke Hiemstra, Pol A Van Lier, Wim Meeus, Susan Branje, Christine M Heim, Charles B Nemeroff, Jonathan Mill, et al. Genome-wide dna methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nature Communications*, 7(1):1–10, 2016.

[50] Yen-Tsung Huang. Causal mediation of semicompeting risks. *Biometrics*, 77(4):1143–1154, 2021.

[51] Yen-Tsung Huang and Wen-Chi Pan. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413, 2016.

[52] Yen-Tsung Huang and Hwai-I Yang. Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(3):370–378, 2017.

[53] JS Huxley. Problems of relative growth (london, 1932). *Nature*, 129:775–777, 1932.

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

[55] Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Causal mediation analysis using r. *Advances in Social Science Research Using R*, 196:129–154, 2010.

[56] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

[57] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[58] Kosuke Imai and Teppei Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171, 2013.

[59] Lawrence James, S Mulaik, and Jeanne M Brett. Causal analysis: Assumptions, models, and data. 1982.

[60] Allan Jérolon, Laura Baglietto, Etienne Birmelé, Flora Alarcon, and Vittorio Perduca. Causal mediation analysis in presence of multiple mediators uncausally related. *The International Journal of Biostatistics*, 17(2):191–221, 2020.

[61] Allan Jerolon. *Analyse causale de médiation multiple et applications*. PhD thesis, Université de Paris, 2020.

[62] Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of multivariate Analysis*, 94(2):401–419, 2005.

[63] Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005.

[64] Charles M Judd and David A Kenny. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5):602–619, 1981.

[65] Vladimir K Kaishev, Dimitrina S Dimitrova, and Steven Haberman. Modelling the joint distribution of competing risks survival times using copula functions. *Insurance: Mathematics and Economics*, 41(3):339–361, 2007.

[66] D Kane, L Stafford, B Bresnihan, and O FitzGerald. A prospective, clinical and radiological study of early psoriatic arthritis: an early synovitis clinic experience. *Rheumatology*, 42(12):1460–1468, 2003.

[67] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.

[68] Kenneth S Kendler. "a gene for...": the nature of gene action in psychiatric disorders. *American Journal of Psychiatry*, 162(7):1243–1252, 2005.

[69] Ronald C Kessler, Katie A McLaughlin, Jennifer Greif Green, Michael J Gruber, Nancy A Sampson, Alan M Zaslavsky, Sergio Aguilar-Gaxiola, Ali Obaid Alhamzawi, Jordi Alonso, Matthias Angermeyer, et al. Childhood adversities and adult psychopathology in the who world mental health surveys. *The British Journal of Psychiatry*, 197(5):378–385, 2010.

[70] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

[71] Nikolai Kolev and Delhi Paiva. Copula-based regression models: A survey. *Journal of Statistical Planning and Inference*, 139(11):3847–3856, 2009.

[72] Samuel Kotz and Saralees Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[73] Theis Lange, Mette Rasmussen, and Lau Caspar Thygesen. Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology*, 179(4):513–518, 2014.

[74] Theis Lange, Stijn Vansteelandt, and Maarten Bekaert. A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, 176(3):190–195, 2012.

[75] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, 2010.

[76] John Lees-Miller, John Hammersley, and R Wilson. Theoretical maximum capacity as benchmark for empty vehicle redistribution in personal rapid transit. *Transportation Research Record: Journal of the Transportation Research Board*, (2146):76–83, 2010.

[77] Chuanhai Liu. *Robit regression: a simple robust alternative to logistic and probit regression*. Wiley London England, 2004.

[78] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, et al. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, 31(2):142–147, 2013.

[79] Tom Loeys, Beatrijs Moerkerke, Olivia De Smet, Ann Buysse, Johan Steen, and Stijn Vansteelandt. Flexible mediation analysis in the presence of nonlinear relations: beyond the mediation formula. *Multivariate Behavioral Research*, 48(6):871–894, 2013.

[80] Monia Lupparelli and Alessandra Mattei. Joint and marginal causal effects for binary non-independent outcomes. *Journal of Multivariate Analysis*, 178:104609, 2020.

[81] David P MacKinnon and James H Dwyer. Estimating mediated effects in prevention studies. *Evaluation Review*, 17(2):144–158, 1993.

[82] Francis Xavier Meehan. Efficient causality in aristotle and st thomas a dissertation. 1940.

[83] Divya Mehta, Torsten Klengel, Karen N Conneely, Alicia K Smith, André Altmann, Thaddeus W Pace, Monika Rex-Haffner, Anne Loeschner, Mariya Gonik, Kristina B Mercer, et al. Childhood maltreatment is associated with distinct genomic and epigenetic profiles in posttraumatic stress disorder. *Proceedings of the National Academy of Sciences*, 110(20):8302–8307, 2013.

[84] Nadine M Melhem, John G Keilp, Giovanna Porta, Maria A Oquendo, Ainsley Burke, Barbara Stanley, Thomas B Cooper, J John Mann, and David A Brent. Blunted hpa axis activity in suicide attempters compared to those at high risk for suicidal behavior. *Neuropsychopharmacology*, 41(6):1447–1456, 2016.

[85] Andreas Meyer-Lindenberg and Daniel R Weinberger. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, 7(10):818–827, 2006.

[86] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[87] Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[88] Aristidis K Nikoloulopoulos and Dimitris Karlis. Multivariate logit copula model with an application to dental data. *Statistics in Medicine*, 27(30):6393–6406, 2008.

[89] Andrew J Patton. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18, 2012.

[90] Judea Pearl. *Causality*. Cambridge university press, 2009.

[91] Judea Pearl. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–436, 2012.

[92] Judea Pearl. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention science*, 13(4):426–436, 2012.

[93] Judea Pearl. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459–481, 2014.

[94] Chamila Perera, Haixiang Zhang, Yinan Zheng, Lifang Hou, Annie Qu, Cheng Zheng, Ke Xie, and Lei Liu. Hima2: high-dimensional mediation analysis and its application in epigenome-wide dna methylation data. *BMC bioinformatics*, 23(1):1–14, 2022.

[95] Maya L Petersen, Sandra E Sinisi, and Mark J van der Laan. Estimation of direct causal effects. *Epidemiology*, pages 276–284, 2006.

[96] Kristopher J Preacher and Andrew F Hayes. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, 40(3):879–891, 2008.

[97] A Sklar RF. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

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

[99] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[100] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[101] Tracy Robson, MC Joiner, GD Wilson, William McCullough, ME Price, Ian Logan, H Jones, SR McKeown, and DG Hirst. A novel human stress response-related gene with a potential role in induced radioresistance. *Radiation Research*, 152(5):451–461, 1999.

[102] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[103] Andrea Rotnitzky, James M Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, 1998.

[104] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183, 1973.

[105] Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203, 1973.

[106] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[107] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[108] Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26, 1977.

[109] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

[110] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[111] Donald B Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188, 2001.

[112] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[113] Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, 13(4):279–313, 2008.

[114] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[115] Jasjeet S Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford Handbook of Political Methodology*, 2:1–32, 2008.

[116] Joanna H Shih and Thomas A Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4):1384–1399, 1995.

[117] Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

[118] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

[119] M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

[120] Peter X-K Song, Mingyao Li, and Ying Yuan. Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65(1):60–68, 2009.

[121] Yanyi Song, Xiang Zhou, Min Zhang, Wei Zhao, Yongmei Liu, Sharon LR Kardia, Ana V Diez Roux, Belinda L Needham, Jennifer A Smith, and Bhramar Mukherjee. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 76(3):700–710, 2020.

[122] Aris Spanos. *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*. Cambridge University Press, 2019.

[123] B Stamm. *Measurement of stress, trauma, and adaptation.* The Sidran Press, 1996.

[124] Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, 186(2):184–193, 2017.

[125] Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Medflex: an r package for flexible mediation analysis using natural effect models. *Journal of Statistical Software*, 76(11), 2017.

[126] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[127] Francisco Suárez and Alfred J Freddoso. *On Efficient Causality: Metaphysical Disputations 17, 18, and 19*. Yale University Press, 1994.

[128] Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.

[129] Brett D Thombs, David P Bernstein, Jill Lobbestael, and Arnoud Arntz. A validation study of the dutch childhood trauma questionnaire-short form: factor structure, reliability, and known-groups validity. *Child Abuse & Neglect*, 33(8):518–523, 2009.

[130] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[131] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(2):245–266, 2012.

[132] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5), 2014.

[133] Chris Urmson, Joshua Anhalt, Drew Bagnell, Christopher Baker, Robert Bittner, MN Clark, John Dolan, Dave Duggins, Tugrul Galatali, Chris Geyer, et al. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, 25(8):425–466, 2008.

[134] Erik-Jan van Kesteren and Daniel L Oberski. Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(5):710–723, 2019.

[135] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[136] Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115, 2014.

[137] Tyler J VanderWeele. Causal mediation analysis with survival data. *Epidemiology (Cambridge, Mass.)*, 22(4):582–585, 2011.

[138] Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

[139] Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1):196, 2013.

[140] Stijn Vansteelandt, Maarten Bekaert, and Gerda Claeskens. On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1):7–30, 2012.

[141] Stijn Vansteelandt, Maarten Bekaert, and Theis Lange. Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, 1(1):131–158, 2012.

[142] Stijn Vansteelandt and Rhian M Daniel. Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(2):258, 2017.

[143] Christiaan H Vinkers, Aimilia Lydia Kalafateli, Bart PF Rutten, Martien J Kas, Zachary Kaminsky, Jonathan D Turner, and Marco PM Boks. Traumatic stress and human dna methylation: a critical review. *Epigenomics*, 7(4):593–608, 2015.

[144] Christiaan H Vinkers, Jelle V Zorn, Sandra Cornelisse, Susanne Koot, Lotte C Houtepen, Berend Olivier, Joris C Verster, René S Kahn, Marco PM Boks, Tobias Kalenscher, et al. Time-dependent changes in altruistic punishment following stress. *Psychoneuroendocrinology*, 38(9):1467–1475, 2013.

[145] Ingeborg Waernbaum. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine*, 31(15):1572–1581, 2012.

[146] Lei Wang. Multiple robustness estimation in causal inference. *Communications in Statistics-Theory and Methods*, 48(23):5701–5718, 2019.

[147] Wei Wang, Suchitra Nelson, and Jeffrey M Albert. Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Statistics in Medicine*, 32(24):4211–4228, 2013.

[148] Ian CG Weaver, Nadia Cervoni, Frances A Champagne, Ana C D'Alessio, Shakti Sharma, Jonathan R Seckl, Sergiy Dymov, Moshe Szyf, and Michael J Meaney. Epigenetic programming by maternal behavior. *Nature Neuroscience*, 7(8):847–854, 2004.

[149] Larry J Williams, Ernest H O'Boyle, and Jia Yu. Condition 9 and 10 tests of model confirmation: A review of james, mulaik, and brett (1982) and contemporary alternatives. *Organizational Research Methods*, 23(1):6–29, 2020.

[150] Katy Wong, Dafna D Gladman, Janice Husted, Jennifer A Long, Vernon T Farewell, and Jennifer A Long. Mortality studies in psoriatic arthritis. results from a single outpatient clinic. i. causes and risk of death. *Arthritis & Rheumatism*, 40(10):1868–1872, 1997.

[151] ROBERT S WoodWORTH. How emotions are identified and classified. *Feelings and emotions: The Wittenberg Symposium*, 1(1), 1928.

[152] Sewall Wright. The theory of path coefficients a reply to niles's criticism. *Genetics*, 8(3):239–255, 1923.

[153] Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

[154] Fan Xia and Kwun Chuen Gary Chan. Decomposition, identification and multiply robust estimation of natural mediation effects with multiple mediators. *Biometrika*, 02 2022. asac004.

[155] Fei Xue, Xiwei Tang, Grace Kim, Allison Aiello, Karestan Koenen, Sandro Galea, Derek Wildman, Monica Uddin, and Annie Qu. Heterogeneous mediation analysis for causal inference. *BioRxiv*, 2020.

[156] Qingzhao Yu, Ying Fan, and Xiaocheng Wu. General multiple mediation analysis with an application to explore racial disparities in breast cancer survival. *J Biom Biostat*, 5(2):1–9, 2014.

[157] Qingzhao Yu and Bin Li. mma: an r package for mediation analysis with multiple mediators. *Journal of Open Research Software*, 5(1), 2017.

[158] Qingzhao Yu, Richard A Scribner, Claudia Leonardi, Lu Zhang, Chi Park, Liwei Chen, and Neal R Simonsen. Exploring racial disparity in obesity: a mediation analysis considering geo-coded environmental factors. *Spatial and Spatio-temporal Epidemiology*, 21:13–23, 2017.

[159] Qingzhao Yu, Xiaocheng Wu, Bin Li, and Richard A Scribner. Multiple mediation analysis with survival outcomes: with an application to explore racial disparity in breast cancer survival. *Statistics in Medicine*, 38(3):398–412, 2019.

[160] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010.

[161] Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brian Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

[162] Qi Zhang. High-dimensional mediation analysis with applications to causal gene identification. *Statistics in biosciences*, pages 1–20, 2021.

[163] Zhongheng Zhang, Cheng Zheng, Chanmin Kim, Sven Van Poucke, Su Lin, and Peng Lan. Causal mediation analysis in the context of clinical research. *Annals of Translational Medicine*, 4(21):425–430.

[164] Yi Zhao, Martin A Lindquist, and Brian S Caffo. Sparse principal component based high-dimensional mediation analysis. *Computational statistics & data analysis*, 142:106835, 2020.

[165] Yi Zhao and Xi Luo. Pathway lasso: pathway estimation and selection with high-dimensional mediators. *Statistics and Its Interface*, 15(1):39–50, 2022.

[166] Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding. *arXiv preprint arXiv:2102.09412*, 2021.

[167] Ruixuan Rachel Zhou, Liewei Wang, and Sihai Dave Zhao. Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, 107(3):573–589, 2020.

[168] Yiwen Zhu, Franca Centorrino, John W Jackson, Garrett M Fitzmaurice, and Linda Valeri. Meta-analysis of total effect decomposition in the presence of multiple mediators: The example of schizophrenia treatment. *Epidemiology*, 32(1):120–130, 2020.

[169] David M Zimmer and Pravin K Trivedi. Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business & Economic Statistics*, 24(1):63–76, 2006.

APPENDICES

Appendix A

The regularity conditions of MLE

Here we provide the regularity conditions of MLE [11].

1. The parameter space Θ is an open set on $\mathcal{R}$.
2. The probability density function $f_\theta(x)$ has common support set for all $\theta \in \Theta$.
3. The log-likelihood function $l(\theta)$ is continuous and three-times differentiable with respect to any $\theta \in \Theta$.
4. Interchange of differentiation and integration of $f_\theta(x)$ is valid for first and second derivatives with respect to parameter θ, i.e.

   $$\frac{\partial^k}{\partial \theta^k} \int f_\theta(x) dx = \int \frac{\partial^k}{\partial \theta^k} f_\theta(x) dx$$

   for $k = 1, 2$.
5. The third derivative of the log-likelihood function must be bounded. i.e. for every $\theta_0 \in \Theta$, there exists a positive C and a function $M(x)$ (both C and $M(x)$ depend on $\theta + 0$), such that for every $\theta \in (\theta_0 - C, \theta_0 + C)$, we have:

   $$|\frac{\partial^3}{\partial \theta^3} log(f_\theta(x))| < M(x)$$

   for every x in the support set and

   $$E(M(x)) < 0.$$

6. The second moment of the score function is finite, i.e.

$$E[(\frac{\partial}{\partial\theta}log(f_\theta(x)))^2] < \infty$$

7. $\{f_\theta(x) : \theta \in \Theta\}$ is identifiable.

Appendix B

Proofs of Theorems in Chapter 2

B.1 Proof of Theorem 1

We keep the notations. With respect to the mediator model, from Lemma 1,

$$\hat{\boldsymbol{\theta}}_M \xrightarrow{p} \boldsymbol{\theta}_M \quad \text{as} \quad n \to \infty.$$

Now, with the response model, similarly, from Lemma 2, we have,

$$\hat{\boldsymbol{\theta}}_Y \xrightarrow{p} \boldsymbol{\theta}_Y \quad \text{as} \quad n \to \infty.$$

Then, by the continuous mapping theorem [11], provided that the model is correctly specified, we have,

$$\hat{F}(\boldsymbol{m}|\boldsymbol{X}, T; \hat{\boldsymbol{\theta}}_M) \xrightarrow{p} F(\boldsymbol{m}|\boldsymbol{X}, T),$$

and

$$\hat{\mu}_Y(\boldsymbol{X}, T, \boldsymbol{M}; \hat{\boldsymbol{\theta}}_Y) \xrightarrow{p} \mu_Y(\boldsymbol{X}, T, \boldsymbol{M}).$$

Therefore, the expected potential outcome of the form $E\{Y(t_0, \boldsymbol{M}(\boldsymbol{t}))\}$, which is imputed as

$$\int_{\boldsymbol{\mathcal{X}}} \left\{ \int \underset{\boldsymbol{m}}{\ldots} \int \hat{\mu}_Y(\boldsymbol{x}, t_0, \boldsymbol{m}; \hat{\boldsymbol{\theta}}_Y) \, \mathrm{d}\, \hat{F}(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{t}; \hat{\boldsymbol{\theta}}_M) \right\} \, \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}),$$

converges in probability to the true values. Similarly, the estimated causal effects of interest converge in probability to the true values.

If the integration step is done by Monte Carlo, by the Law of Large Numbers, the conclusion still holds. If the integration is done via numerical integration methods, consistency is provided by the respective properties.

B.2 Proof of Theorem 2

We first factorize the joint conditional likelihood of mediators and outcomes as follows:

$$L(\boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{MY}) = L(\boldsymbol{Y}|\boldsymbol{M}, \boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{Y}) \times L(\boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{M}).$$

Taking log on both sides yield the log-likelihood such that,

$$l(\boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{MY}) = l(\boldsymbol{Y}|\boldsymbol{M}, \boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{Y}) + l(\boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{M}).$$

Therefore, the score vectors for estimating $\boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_Y$ are,

$$\boldsymbol{S}(\boldsymbol{\theta}_Y; \boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}) = \frac{\partial l(\boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{MY})}{\partial \boldsymbol{\theta}_Y} = \frac{\partial l(\boldsymbol{Y}|\boldsymbol{M}, \boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{Y})}{\partial \boldsymbol{\theta}_Y};$$

and

$$\boldsymbol{S}(\boldsymbol{\theta}_M; \boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}) = \frac{\partial l(\boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{MY})}{\partial \boldsymbol{\theta}_M} = \frac{\partial l(\boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{M})}{\partial \boldsymbol{\theta}_M}$$

respectively. From here, it's clear that the off-diagonal element of the block information matrix,

$$I_{\boldsymbol{\theta}_Y, \boldsymbol{\theta}_M} = \frac{\partial^2 l(\boldsymbol{Y}, \boldsymbol{M}|\boldsymbol{T}, \boldsymbol{X}; \boldsymbol{\theta}_{MY})}{\partial \boldsymbol{\theta}_Y \partial \boldsymbol{\theta}_M} = \boldsymbol{0}. \tag{B.1}$$

So, we have asymptotically $cor(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_Y|\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{Y}) = 0$. The key point of the proof is that we are assuming the response model and the mediator model does not share parameters.

Appendix C

Derivations of closed forms in Chapter 2

C.1 The mediators and outcome are all continuous with various models

When mediator-treatment interactions are included in the response model and the model is

$$\mu_Y(\boldsymbol{X}, T, \boldsymbol{M}) = \gamma_0 + \sum_{l=1}^{p} \gamma_k X_k + \tau T + \sum_{j=1}^{J} \beta_j M_j + \sum_{j=1}^{J} \tau_j T\, M_j,$$

where the parameter τ_j reflects the interactions among mediators.

Under this scenario, we have:

$$
\begin{aligned}
&E\{Y(t_0, M_1(t_1), \ldots, M_J(t_J))\} \\
=&\int_{\boldsymbol{\mathcal{X}}} \left\{ \int \cdots \int_{\boldsymbol{m}} \mu_Y(\boldsymbol{x}, t_0, \boldsymbol{m})\, d^J F(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{t}) \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}) \\
=&\int_{\boldsymbol{\mathcal{X}}} \left\{ \int \cdots \int_{\boldsymbol{m}} (\gamma_0 + \sum_{l=1}^{p} \gamma_k x_k + \tau t_0 + \sum_{j=1}^{J} \beta_j m_j + \sum_{j=1}^{J} \tau_j t_0 m_j)\, d^J F(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{t}) \right\} \mathrm{d}\, F_{\boldsymbol{X}}(\boldsymbol{x}) \\
=&\mathrm{E}_{\boldsymbol{X}} \left(\gamma_0 + \sum_{l=1}^{p} \gamma_k X_k + \tau t_0 + \sum_{j=1}^{J} \beta_j \mathrm{E}(M_j|t_j, \boldsymbol{X}) + \sum_{j=1}^{J} \tau_j t_0 \mathrm{E}(M_j|t_j, \boldsymbol{X}) \right) \\
=&\mathrm{E}_{\boldsymbol{X}} \left(\gamma_0 + \tau t_0 + \sum_{j=1}^{J} \beta_j \left(\lambda_j t_j + \alpha_{j0} + \sum_{l=1}^{p} \alpha_{jl} X_l \right) + \sum_{j=1}^{J} \tau_j t_0 \left(\lambda_j t_j + \alpha_{j0} + \sum_{l=1}^{p} \alpha_{jl} X_l \right) \right).
\end{aligned}
$$

The statistics of interest are then calculated according to the definitions.

When mediator-mediator interactions are included in the response model, the model changes to,

$$
\mu_Y(\boldsymbol{X}, T, \boldsymbol{M}) = \gamma_0 + \sum_{l=1}^{p} \gamma_k X_k + \tau T + \sum_{j=1}^{J} \beta_j M_j + \sum_{\substack{m=1,\ldots,J \\ n=1,\ldots,J \\ m \le n}} \eta_{m,n} \cdot M_m \cdot M_n,
$$

where the parameters η reflect the interactions among mediators. Under this scenario, we have,

$$
\begin{aligned}
&E\{Y(t_0, M_1(t_1), \ldots, M_J(t_J))\} \\
=&\mathrm{E}_{\boldsymbol{X}} \left(\gamma_0 + \sum_{l=1}^{p} \gamma_k X_k + \tau t_0 + \sum_{j=1}^{J} \beta_j \mathrm{E}\left(M_j|t_j, \boldsymbol{X}\right) + \sum_{\substack{m=1,\ldots,J \\ n=1,\ldots,J \\ m \le n}} \eta_{m,n} \mathrm{E}\left(M_m M_n|t_j, \boldsymbol{X}\right) \right) \\
=&\mathrm{E}_{\boldsymbol{X}} \Bigg\{ \gamma_0 + \tau t_0 + \sum_{j=1}^{J} \beta_j \left(\alpha_{j0} + \sum_{l=1}^{p} \alpha_{jl} X_{il} + \lambda_j t_j \right) \\
&+ \sum_{\substack{m=1,\ldots,J \\ n=1,\ldots,J \\ m \le n}} \eta_{m,n} \left\{ \left(\alpha_{m0} + \sum_{l=1}^{p} \alpha_{ml} X_{ml} + \lambda_m t_m \right) \left(\alpha_{n0} + \sum_{l=1}^{p} \alpha_{nl} X_{nl} + \lambda_n t_n \right) + \rho_{m,n} \sigma_m^2 \sigma_n^2 \right\} \Bigg\}
\end{aligned}
$$

With respect to the statistic of interest, following the definitions respectively, we derive the individual indirect effect for the kth mediator as:

$$IE_k = \mathrm{E}_{\boldsymbol{X}}\left\{\beta_k\lambda_k + \sum_{\substack{n=1,\ldots,J \\ n\neq k}} \eta_{k,n}\left\{\lambda_k\left(\lambda_n t_n + \alpha_{n0} + \sum_{l=1}^{p}\alpha_{nl}x_{nl}\right)\right\}\right\}.$$

C.2 One mediator is log-normally distributed

Here we provide detailed derivations of the variance of the population-level estimator. We first notice that the population level estimator is an average over each individual estimator, i.e. $\widehat{\bar{IE}}_1 = \frac{1}{n}\sum_{i=1}^{n}\widehat{IE}_{i1}$, therefore $Var\{\widehat{\bar{IE}}_1\} = Var\{\frac{1}{n}\sum_{i=1}^{n}\widehat{IE}_{i1}\} = \frac{1}{n^2}Var\{\sum_{i=1}^{n}\widehat{IE}_{i1}\}$. Applying the law of total variance, we decompose the variance of the population level estimator as

$$Var(\widehat{\bar{IE}}_1) = E\{Var(\widehat{\bar{IE}}_1|\boldsymbol{X})\} + Var\{E(\widehat{\bar{IE}}_1|\boldsymbol{X})\}. \tag{C.1}$$

Notice that here the covariate matrix $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, which represent all the covariates used from the data. For $Var\{\widehat{\bar{IE}}_1|\boldsymbol{X}\}$,

$$Var\{\widehat{\bar{IE}}_1|\boldsymbol{X}\} = \frac{1}{n^2}Var\{(\sum_{i=1}^{n}\widehat{IE}_{i1})|\boldsymbol{X}\}.$$

Due to the difficulties of calculating the covariance, we may apply the delta-method to approximate the entire variance of the summation directly, since the inner part $(\sum_{i=1}^{n}\widehat{IE}_{i1})$ is a function of parameters and covariates,

$$\begin{aligned}
&Var\{\widehat{\bar{IE}}_1|\boldsymbol{X}\} \\
=&\frac{1}{n^2}Var\{(\sum_{i=1}^{n}\widehat{IE}_{i1})|\boldsymbol{X}\} \\
=&\frac{1}{n^2}Var\left\{\sum_{i=1}^{n}\hat{\zeta}_1\exp(\hat{\boldsymbol{\alpha}}_1'\boldsymbol{X}_i + \hat{\sigma}_1^2/2)[\exp(\hat{\lambda}_1) - 1)]\middle|\boldsymbol{X}\right\} \\
=&\frac{1}{n^2}Var\left\{\hat{\zeta}_1\exp(\hat{\sigma}_1^2/2)[\exp(\hat{\lambda}_1) - 1)]\sum_{i=1}^{n}\exp(\hat{\boldsymbol{\alpha}}_1'\boldsymbol{X}_i)\middle|\boldsymbol{X}\right\}.
\end{aligned}$$

Denote $h(\theta, \boldsymbol{X}) = \zeta_1 \exp(\sigma_1^2/2)[\exp(\lambda_1) - 1)] \sum_{i=1}^{n} \exp(\boldsymbol{\alpha}_1' \boldsymbol{X}_i)$ representing the items inside the variance. From the delta-method, we approximately,

$$Var\{h(\hat{\theta}, \boldsymbol{X})\} \approx \nabla h'(\theta, \boldsymbol{X})\, \boldsymbol{V}\, \nabla h(\theta, \boldsymbol{X}),$$

where

$$\nabla h(\theta, \boldsymbol{X}) = \begin{pmatrix} \frac{\partial h(\theta, \boldsymbol{X})}{\partial \lambda_1} \\ \frac{\partial h(\theta, \boldsymbol{X})}{\partial \boldsymbol{\alpha}_1} \\ \frac{\partial h(\theta, \boldsymbol{X})}{\partial \sigma_1} \\ \vdots \\ \frac{\partial h(\theta, \boldsymbol{X})}{\partial \zeta_1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \zeta_1 \exp(\sigma_1^2/2) \exp(\lambda_1) \sum_{i=1}^{n} \exp(\boldsymbol{\alpha}_1' \boldsymbol{X}_i) \\ \zeta_1 \exp(\sigma_1^2/2)[\exp(\lambda_1) - 1] \sum_{i=1}^{n} \exp(\boldsymbol{\alpha}_1' \boldsymbol{X}_i) \boldsymbol{X}_i \\ \zeta_1 \exp(\sigma_1^2/2)[\exp(\lambda_1) - 1] \sum_{i=1}^{n} \exp(\boldsymbol{\alpha}_1' \boldsymbol{X}_i) \sigma_1 \\ \vdots \\ \exp(\sigma_1^2/2)[\exp(\lambda_1) - 1] \sum_{i=1}^{n} \exp(\boldsymbol{\alpha}_1' \boldsymbol{X}_i) \\ \vdots \end{pmatrix}$$

Similarly, by replacing the parameters in the above formula by their estimated values, one can obtain an estimation of the conditional variance of the estimated statistic.

Furthermore, back to (C.1), with respect to the second part, the conditional expectation $E(\widehat{\bar{IE}}_1|\boldsymbol{X})$ is also a function of the covariates $\boldsymbol{X}$ and the variances are obtained in a similar manner. From here, we can see that we need information on the distribution of the covariates $\boldsymbol{X}$ in order to fully obtain $Var(\widehat{\bar{IE}}_1)$, though sometimes, it suffices to have the mean and variance information of the covariates $\boldsymbol{X}$. One way to get rid of assuming distributions of $\boldsymbol{X}$ is to use the empirical estimates from the data that are introduced in the main body of the thesis. Another way is assuming the covariates $\boldsymbol{X}$ to be fixed and use the conditional variance $Var\{\widehat{\bar{IE}}_1|\boldsymbol{X}\}$ for inference. The variance estimation of IE_2 and DE are trivial as they only involve the parameters. Estimation procedures are already covered in the main body of the thesis.

In the thesis, we also mention using Taylor's expansion to approximate the expression of the causal effects of interest, here we show the results:

$$\begin{aligned} & IE_{i1} \\ =& \zeta_1[\exp(\boldsymbol{\alpha}_1' \boldsymbol{X}_i + \sigma_1^2/2)(\exp(\lambda_1) - 1] \\ \approx& \zeta_1[e^{\alpha_{10} + \sigma_1^2/2}(\boldsymbol{\alpha}_1' \boldsymbol{X}_i + \frac{1}{2}(\boldsymbol{X}_i)' \boldsymbol{\alpha}_1 \boldsymbol{\alpha}_1' \boldsymbol{X}_i)(\exp(\lambda_1) - 1)] \\ \approx& \zeta_1[e^{\alpha_{10} + \sigma_1^2/2} \boldsymbol{\alpha}_1' \boldsymbol{X}_i (\exp(\lambda_1) - 1)], \end{aligned}$$

with the third line representing approximate the results using second-order Taylor's expansion and the fourth line representing the first-order. The estimation of the other statistics remains the same. Asymptotic variances can also be approximated via the delta-method and the law of total variance.

Appendix D

Additional Real Data Application Results in Chapter 3

D.1 Real data application results under 38 candidates

Here we provide the real data application results under $m_2 = 38$ candidate settings.

	value	95% CI
Grouped IE of cg05781698, cg06992213, cg25448067, cg15663823, cg05051734, cg04000159	-4.299	(-12.821,-4.222)
Grouped IE of cg26801646, cg25626453	-3.972	(-9.828,1.885)
Grouped IE of cg24516147, cg10822172, cg16181903	-5.161	(-11.065,0.742)
Individual IE of cg18634806	-3.477	(-7.955,1.001)
DE	9.312	(-1.197,19.821)

Table D.1: grouped and individual indirect effects of DNA methylation loci under $m_2 = 38$ candidates

Appendix E

Proofs of Theorems in Chapter 4

E.1 Proof of Theorem 4

We only prove the second part of Theorem 3 as the first part is the property of propensity score and is proved in Robin et al. [106]. For any $t_0 \in \mathcal{T}$ and $\{m_1, \ldots, m_J\} \in \mathcal{M}$, we have that,

$$
\begin{aligned}
&P(\boldsymbol{M} \leq \boldsymbol{m} | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)) \\
=&E\left\{I(\boldsymbol{M} \leq \boldsymbol{m}) | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)\right\} \\
=&E\left\{E\left\{I(\boldsymbol{M} \leq \boldsymbol{m}) | T, \boldsymbol{X} F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)\right\} | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)\right\} \\
=&E\left\{E\left\{I(\boldsymbol{M} \leq \boldsymbol{m}) | T, \boldsymbol{X}\right\} | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)\right\} \\
=&E\left\{F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}) | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)\right\} \\
=&F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}).
\end{aligned}
$$

On the other hand, we have,

$$
\begin{aligned}
&P(\boldsymbol{M} \leq \boldsymbol{m} | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X})) \\
=&E\left\{I(\boldsymbol{M} \leq \boldsymbol{m}) | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X})\right\} \\
=&E\left\{E\left\{I(\boldsymbol{M} \leq \boldsymbol{m}) | T, \boldsymbol{X}\right\} | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)\right\} \\
=&F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}).
\end{aligned}
$$

So, we have,

$$
P(\boldsymbol{M} \leq \boldsymbol{m} | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X}), Y(t_0, m_1, \ldots, m_J)) = P(\boldsymbol{M} \leq \boldsymbol{m} | F_{\boldsymbol{M}}(\boldsymbol{m}|T, \boldsymbol{X})).
$$

Therefore, $Y(t_0, m_1, m_2, \ldots, m_J) \perp \{M_1, \ldots, M_J\} | F_{M_1, \ldots, M_J}(m_1', \ldots, m_J'|T, \boldsymbol{X})$ for any $t_0 \in \mathcal{T}$ and $\{m_1, \ldots, m_J\} \in \mathcal{M}$.

E.2 Proof of Theorem 5

From (1.2) and (1.1), we assume a response model under treatment ($T = 1$): $E\{Y(1)\} = \boldsymbol{X}'\boldsymbol{\beta}^1$; and a model under control ($T = 0$): $E\{Y(0)\} = \boldsymbol{X}'\boldsymbol{\beta}^0$. In the following proof, we denote the correct covariates as $\boldsymbol{X}$ and the incorrect ones as $\boldsymbol{Z}$. We denote the regression coefficients associated with the correct covariates $\boldsymbol{X}$ as $\boldsymbol{\beta}$ and the ones associated with the incorrect covariates $\boldsymbol{Z}$ as $\boldsymbol{\beta}^*$. We denote the correct propensity model as $P(T|\boldsymbol{X})$ and the incorrect ones as $P^*(T|\boldsymbol{Z})$.

First, we see that (1.2) and (1.1) yields the following because $\boldsymbol{X}$ includes intercept:

$$P_n\left\{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}(Y - \boldsymbol{X}'\boldsymbol{\beta}^1_{WR})\right\} = 0, \tag{E.1}$$

and

$$P_n\left\{\frac{I(T=0)}{P(T=0|\boldsymbol{X})}(Y - \boldsymbol{X}'\boldsymbol{\beta}^0_{WR})\right\} = 0, \tag{E.2}$$

We consider the following two cases:

1. When the regression model is correctly specified but the propensity model is not. Under this setting, Formula (E.1) becomes:

   $$P_n\left\{\frac{I(T=1)}{P^*(T=1|\boldsymbol{Z})}(Y - \boldsymbol{X}'\boldsymbol{\beta}^1_{WR})\right\} = 0.$$

   Taking expectation inside P_n yields:

   $$\begin{aligned}
   &E\left\{\frac{I(T=1)}{P^*(T=1|\boldsymbol{Z})}(Y - \boldsymbol{X}'\boldsymbol{\beta}^1_{WR})\right\}\\
   =&E\left\{\frac{I(T=1)}{P^*(T=1|\boldsymbol{Z})}E\left\{(Y - \boldsymbol{X}'\boldsymbol{\beta}^1_{WR})|\boldsymbol{X}\right\}\right\}\\
   =&E\left\{\frac{I(T=1)}{P^*(T=1|\boldsymbol{Z})}\cdot 0\right\}\\
   =&0.
   \end{aligned}$$

   Following our assumptions stated at the beginning of the proof, the inner expectation $E\left\{(Y - \boldsymbol{X}'\boldsymbol{\beta}^1)|\boldsymbol{X}\right\} = 0$. Therefore, we have that $\hat{\boldsymbol{\beta}}^1_{WR} \xrightarrow{p} \boldsymbol{\beta}^1$ and similarly $\hat{\boldsymbol{\beta}}^0_{WR} \xrightarrow{p} \boldsymbol{\beta}^0$. Therefore, under our assumption, $\boldsymbol{X}'\hat{\boldsymbol{\beta}}^1 \xrightarrow{p} Y^1$ and $\boldsymbol{X}'\hat{\boldsymbol{\beta}}^0 \xrightarrow{p} Y^0$, so $\hat{\tau}_{WR} \xrightarrow{p} \tau$, which is desired.

2. When the propensity model is correctly specified but the regression model is not. Under this setting, we first need to show that:

$$\begin{aligned}
&\tau_{WR} \\
=&P_n\{\boldsymbol{Z}'\boldsymbol{\beta}^{1*}_{WR} - \boldsymbol{Z}'\boldsymbol{\beta}^{0*}_{WR}\} \\
=&P_n\left\{\underbrace{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}(Y-\boldsymbol{Z}'\boldsymbol{\beta}^{1*}_{WR})}_{\text{(A)}} - \underbrace{\frac{I(T=0)}{P(T=0|\boldsymbol{X})}(Y-\boldsymbol{Z}'\boldsymbol{\beta}^{0*}_{WR})}_{\text{(B)}} + \boldsymbol{Z}'\boldsymbol{\beta}^{1*}_{WR} - \boldsymbol{Z}'\boldsymbol{\beta}^{0*}_{WR}\right\}.
\end{aligned} \tag{E.3}$$

This is because, $\hat{\boldsymbol{\beta}}^{1*}$ and $\hat{\boldsymbol{\beta}}^{0*}$ are solutions to the following incorrect estimating equations:

$$P_n\left\{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}(Y-\boldsymbol{Z}'\boldsymbol{\beta}^{1*}_{WR})\right\} = 0,$$

and

$$P_n\left\{\frac{I(T=0)}{P(T=0|\boldsymbol{X})}(Y-\boldsymbol{Z}'\boldsymbol{\beta}^{0*}_{WR})\right\} = 0.$$

Therefore, though $\hat{\boldsymbol{\beta}}^{1*}_{WR}$ and $\hat{\boldsymbol{\beta}}^{0*}_{WR}$ does not converge in probability to the true $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^0$, $\boldsymbol{\beta}^*$, but part (A) and (B) in (E.7) are always zero. Then, from Formula (E.7), taking expectation inside P_n and re-arranging, we have that:

$$\begin{aligned}
&E\left\{\boldsymbol{Z}'\boldsymbol{\beta}^{1*}_{WR} - \boldsymbol{Z}'\boldsymbol{\beta}^{0*}_{WR}\right\} \\
=&\underbrace{E\left\{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}Y\right\}}_{\text{(C)}} - \underbrace{E\left\{\frac{I(T=0)}{P(T=0|\boldsymbol{X})}Y\right\}}_{\text{(D)}} \\
&+\underbrace{E\left\{\left(1-\frac{I(T=1)}{P(T=1|\boldsymbol{X})}\right)\boldsymbol{Z}'\boldsymbol{\beta}^{1*}_{WR}\right\}}_{\text{(E)}} - \underbrace{E\left\{\left(1-\frac{I(T=0)}{P(T=0|\boldsymbol{X})}\right)\boldsymbol{Z}'\boldsymbol{\beta}^{0*}_{WR}\right\}}_{\text{(F)}}.
\end{aligned}$$

We start with part (E) and (F),

$$\begin{aligned}
\text{(E)} &= E\left\{\left(1-\frac{I(T=1)}{P(T=1|\boldsymbol{X})}\right)\boldsymbol{Z}'\boldsymbol{\beta}_{WR}^{1*}\right\} \\
&= E\left\{E\left\{\left(1-\frac{I(T=1)}{P(T=1|\boldsymbol{X})}\right)|\boldsymbol{X}\right\}\boldsymbol{Z}'\boldsymbol{\beta}_{WR}^{1*}\right\} \\
&= E\left\{0\cdot\boldsymbol{Z}'\boldsymbol{\beta}_{WR}^{1*}\right\} \\
&= 0,
\end{aligned}$$

and similarly, (F) = 0. For (C) and (D), the proof is standard, such that,

$$\begin{aligned}
\text{(C)} &= E\left\{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}Y\right\} \\
&= E\left\{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}Y^1\right\} \\
&= E\left\{E\left\{\frac{I(T=1)}{P(T=1|\boldsymbol{X})}Y^1|\boldsymbol{X}\right\}\right\} \\
&= E\left\{Y^1\right\},
\end{aligned}$$

and similarly, (D) = $E\{Y^0\}$. Therefore, we prove that $E\left\{\boldsymbol{Z}'\boldsymbol{\beta}_{WR}^{1*}-\boldsymbol{Z}'\boldsymbol{\beta}_{WR}^{0*}\right\} = E\left\{Y^1-Y^0\right\}=\tau$. So $\tau_{\hat{WR}}\xrightarrow{p}\tau$ under this setting.

We summarize the two cases and prove that τ_{WR}, constructed by using a weighted least square approach, is a doubly robust estimator for τ.

E.3 Proof of Lemma 3

We start with the first part that states $\alpha_j = E\{M_j(1)-M_j(0)\}$. We notice that under our model assumptions,

$$E\{M_j(1)\} = E\{E\{M_j|T=1,\boldsymbol{X}\}\} = E\{T+\boldsymbol{\eta}_j'\boldsymbol{X}\}$$

and

$$E\{M_j(0)\} = E\{E\{M_j|T=0,\boldsymbol{X}\}\} = E\{\boldsymbol{\eta}_j'\boldsymbol{X}\},$$

where the outer expectation is taken with respect to the covariates $\boldsymbol{X}$. Therefore it is straightforward to show that $\alpha_j = E\{M_j(1) - M_j(0)\}$.

The second part of the lemma is a direct application of Theorem 5. By replacing Y in Theorem 5 with M_j, it is straightforward to show that $\hat{\alpha}_{j,WR}$, which is obtained via solving the weighted regression estimating equation (4.10), is a DR estimator for the causal effect $E\{M_j(1) - M_j(0)\}$.

E.4 Proof of Lemma 4

Similarly, we start with the first part of the lemma. Since the response model is assumed as in (4.7), under our assumptions,

$$E\{Y(t,m)\} = E\{E\{Y|T=t, M=m, \boldsymbol{X}\}\} = E\{\boldsymbol{\gamma}'\boldsymbol{X} + \tau t + \beta m\}.$$

Therefore, it is straightforward to see that $\tau = E\{Y(1,m) - Y(0,m)\}$ and $\beta = (E\{Y(t,m) - Y(t,m')\})/(m-m')$ for any t, m and m'.

For the second part of the lemma, we notice that $\boldsymbol{\theta}_Y$ are estimated by solving (4.11). Therefore, similar to the proof of Theorem 5, solving (4.11) implies solving

$$P_n\left\{\frac{I(T=t, M=m)}{P(T=t|\boldsymbol{X})f(m|T=t,\boldsymbol{X})}\left[Y - \mu_Y(t,m,\boldsymbol{X};\boldsymbol{\theta}_Y)\right]\right\} = 0, \tag{E.4}$$

because $\boldsymbol{X}$ includes a leading $\mathbf{1}$ representing the intercept. Additionally, in a similar manner, we denote the correct covariates as $\boldsymbol{X}$ and the incorrect ones as $\boldsymbol{Z}$. We add a superscript $*$ sign to the coefficients associated with the incorrect covariates. We denote the correct propensity model as $P(T|\boldsymbol{X})$ and the incorrect ones as $P^*(T|\boldsymbol{Z})$, the correct conditional density of M as $f(m|\boldsymbol{X},T)$ and the incorrect ones as $f^*(m|\boldsymbol{Z},T)$.

We consider the following two cases:

1. When the regression model is correctly specified but the weight models are not. Under this setting, Formula (E.4) becomes:

$$P_n\left\{\frac{I(T=t, M=m)}{P^*(T=t|\boldsymbol{Z})f^*(m|\boldsymbol{Z},t)}[Y - \mu_Y(\boldsymbol{X},t,m;\boldsymbol{\theta}_Y)]\right\} = 0, \tag{E.5}$$

and $\hat{\boldsymbol{\theta}}^*_{Y,WR}$ are solutions of $\boldsymbol{\theta}_Y$ to (E.5). Taking expectation inside P_n yields (capital letters are used to denote random variables):

$$
\begin{aligned}
& E\left\{\frac{I(T=t,M=m)}{P^*(T=t|\boldsymbol{Z})f^*(M=m|\boldsymbol{Z},t)}(Y-\mu_Y(\boldsymbol{X},t,m;\boldsymbol{\theta}_Y)\right\} \\
=& E\left\{E\{\frac{I(T=t,M=m)}{P^*(T=t|\boldsymbol{Z})f^*(M=m|\boldsymbol{Z},t)}(Y-\mu_Y(\boldsymbol{X},t,m;\boldsymbol{\theta}_Y)|T=t,M=m,\boldsymbol{X},\boldsymbol{Z}\}\right\} \\
=& E\Bigg\{E\{\frac{I(T=t,M=m)}{P^*(T=t|\boldsymbol{Z})f^*(M=m|\boldsymbol{Z},t)}|T=t,M=m,\boldsymbol{Z}\} \\
& \qquad\qquad E\{(Y-\mu_Y(\boldsymbol{X},t,m;\boldsymbol{\theta}_Y)|T=t,M=m,\boldsymbol{X}\}\Bigg\} \\
=& E\left\{E\{\frac{I(T=t,M=m)}{P^*(T=t|\boldsymbol{Z})f^*(M=m|\boldsymbol{Z},t)}|T=t,M=m,\boldsymbol{Z}\}\cdot 0\right\} \\
=& 0.
\end{aligned}
$$

Since we have zero expected value inside P_n, we have $\hat{\boldsymbol{\theta}}^*_{Y,WR} \xrightarrow{p} \boldsymbol{\theta}_Y$. Therefore, $\hat{\tau}_{WR} \xrightarrow{p} \tau$ and similarly $\hat{\beta}_{WR} \xrightarrow{p} \beta$, which is desired.

2. When the weight models are correctly specified but the regression model is not. Under this setting, the estimating equation (E.4) becomes:

$$
P_n\left\{\frac{I(T=t,M=m)}{P(T=t|\boldsymbol{X})f(m|\boldsymbol{X},t)}[Y-\mu^*_Y(\boldsymbol{Z},t,m;\boldsymbol{\theta}^*_Y)]\right\}=0 \tag{E.6}
$$

and $\hat{\boldsymbol{\theta}}^*_{Y,WR}$ are solutions of $\boldsymbol{\theta}_Y$ to (E.6). We start with β, our goal is to show, for any t, m and m', the solution of β to (E.6) (denoted as $\hat{\beta}^*_{WR}$) still consistently estimate $E\{Y(t,m)-Y(t,m')\}/(m-m')$. First, we see that:

$$
\begin{aligned}
& \mu^*_Y(\boldsymbol{Z},t,m;\boldsymbol{\theta}^*_Y)-\mu^*_Y(\boldsymbol{Z},t,m';\boldsymbol{\theta}^*_Y) \\
=& \underbrace{\left\{\frac{I(T=t,M=m)}{P(T=t|\boldsymbol{X})f(M=m|T=t,\boldsymbol{X})}[Y-\mu^*_Y(\boldsymbol{Z},t,m;\boldsymbol{\theta}^*_Y)]\right\}}_{\text{(A)}} \\
& -\underbrace{\left\{\frac{I(T=t,M=m')}{P(T=t|\boldsymbol{X})f(M=m'|T=t,\boldsymbol{X})}[Y-\mu^*_Y(\boldsymbol{Z},t,m';\boldsymbol{\theta}^*_Y)]\right\}}_{\text{(B)}} \\
& +\mu^*_Y(\boldsymbol{Z},t,m;\boldsymbol{\theta}^*_Y)-\mu^*_Y(\boldsymbol{Z},t,m';\boldsymbol{\theta}^*_Y).
\end{aligned} \tag{E.7}
$$

This is because, $\hat{\boldsymbol{\theta}}_Y^*$ are solutions to (E.4). Therefore, for any values of t, m and $\boldsymbol{Z}$, (A) and (B) are always zero, despite $\hat{\boldsymbol{\theta}}_Y^*$ do not converge to the true $\boldsymbol{\theta}_Y$, but they do converge to some other values $\boldsymbol{\theta}_Y^*$. Then, following Formula (E.7), taking expectation and re-arranging, we have:

$$
\begin{aligned}
& E\left\{\mu_Y^*(\boldsymbol{Z}, t, m; \boldsymbol{\theta}_Y^*) - \mu_Y^*(\boldsymbol{Z}, t, m'; \boldsymbol{\theta}_Y^*)\right\} \\
= & \underbrace{E\left\{\frac{I(T=t, M=m)}{P(T=t|\boldsymbol{X}) f(M=m|T=t, \boldsymbol{X})} Y\right\}}_{\text{(C)}} \\
& \underbrace{- E\left\{\frac{I(T=t, M=m')}{P(T=t|\boldsymbol{X}) f(M=m'|T=t, \boldsymbol{X})} Y\right\}}_{\text{(D)}} \\
& \underbrace{+ E\left\{\left(1 - \frac{I(T=t, M=m)}{P(T=t|\boldsymbol{X}) f(M=m|T=t, \boldsymbol{X})}\right) \mu_Y^*(\boldsymbol{Z}, t, m; \boldsymbol{\theta}_Y^*)\right\}}_{\text{(E)}} \\
& \underbrace{- E\left\{\left(1 - \frac{I(T=t, M=m')}{P(T=t|\boldsymbol{X}) f(M=m'|T=t, \boldsymbol{X})}\right) \mu_Y^*(\boldsymbol{Z}, t, m'; \boldsymbol{\theta}_Y^*)\right\}}_{\text{(F)}}.
\end{aligned}
\tag{E.8}
$$

We start with parts (E) and (F),

$$
\begin{aligned}
\text{(E)} &= E\left\{\left(1 - \frac{I(T=t, M=m)}{P(T=t|\boldsymbol{X}) f(M=m|T=t, \boldsymbol{X})}\right) \mu_Y^*(\boldsymbol{Z}, t, m; \boldsymbol{\theta}_Y^*)\right\} \\
&= E\left\{E\left\{\left(1 - \frac{I(T=t, M=m)}{P(T=t|\boldsymbol{X}) f(M=m|T=t, \boldsymbol{X})}\right) \mu_Y^*(\boldsymbol{Z}, t, m; \boldsymbol{\theta}_Y^*)|\boldsymbol{X}, \boldsymbol{Z}\right\}\right\} \\
&= E\left\{E\left\{\left(1 - \frac{I(T=t, M=m)}{P(T=t|\boldsymbol{X}) f(M=m|T=t, \boldsymbol{X})}\right)|\boldsymbol{X}\right\} E\left\{\mu_Y^*(\boldsymbol{Z}, t, m; \boldsymbol{\theta}_Y^*)|\boldsymbol{Z}\right\}\right\} \\
&= E\left\{\left(1 - \frac{E\left\{I(T=t, M=m)|\boldsymbol{X}\right\}}{f_{T,M}(T=t, M=m|\boldsymbol{X})}\right) E\left\{\mu_Y^*(\boldsymbol{Z}, t, m; \boldsymbol{\theta}_Y^*)|\boldsymbol{Z}\right\}\right\} \\
&= E\left\{0 \cdot \mu_Y^*(\boldsymbol{Z}, t, m; \boldsymbol{\theta}_Y^*)\right\} \\
&= 0,
\end{aligned}
$$

and similarly, (F) = 0. For (C) and (D), we have,

$$\begin{aligned}
\text{(C)} &= E\left\{\frac{I(T=t,M=m)}{P(T=t|\boldsymbol{X})f(M=m|T=t,\boldsymbol{X})}Y\right\} \\
&= E\left\{\frac{I(T=t,M=m)}{f_{T,M}(T=t,M=m|\boldsymbol{X})}Y(t,m)\right\} \\
&= E\left\{E\left\{\frac{I(T=t,M=m)}{f_{T,M}(T=t,M=m|\boldsymbol{X})}|\boldsymbol{X}\right\}Y(t,m)\right\} \\
&= E\left\{\frac{E\left\{I(T=t,M=m)|\boldsymbol{X}\right\}}{f_{T,M}(T=t,M=m|\boldsymbol{X})}Y(t,m)\right\} \\
&= E\left\{Y(t,m)\right\},
\end{aligned}$$

and similarly, (D) = $E\left\{Y(t,m')\right\}$. Therefore, from (E.8), we show that

$$E\left\{\mu_Y^*(\boldsymbol{Z},t,m;\boldsymbol{\theta}_Y^*)-\mu_Y^*(\boldsymbol{Z},t,m';\boldsymbol{\theta}_Y^*)\right\} = E\{Y(t,m)\}-E\{Y(t,m')\}$$

By the unbiasedness property of estimating equations, we have

$$\hat{\beta}_{WR}^* = P_n\{(\mu_Y^*(\boldsymbol{Z},t,m;\boldsymbol{\theta}_Y^*)-\mu_Y^*(\boldsymbol{Z},t,m';\boldsymbol{\theta}_Y^*))/(m-m')\} \xrightarrow{p} \beta,$$

In a similar way, we can also show

$$\hat{\tau}_{WR}^* \xrightarrow{p} \tau = E\{Y(1,m)-Y(0,m)\}.$$

We summarize the two cases and prove that τ_{WR} and β_{WR}, constructed by using the weighted regression approach, are doubly robust estimators for τ and β.