Mediation Analysis with Complex Intermediate Causal Structure

Fan Xia

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Kwun Chuen Gary Chan, Chair

Thomas S. Richardson

James P. Hughes

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Mediation Analysis with Complex Intermediate Causal Structure

Fan Xia

Chair of the Supervisory Committee:
Professor Kwun Chuen Gary Chan
Department of Biostatistics

My doctoral research is oriented around causal inference, specifically causal mediation analysis. Roughly, it can be divided into two parts: (1) understanding and resolving conceptual issues in causal problems, and (2) developing methodology for causal analysis. One goal of my doctoral research is to provide a comprehensive guide to applied statisticians and epidemiologists that can help them navigate the philosophical subtleties and abundant methodology in causal inference. Another goal of my doctoral research is to develop methodology for complex causal mediation structures, including mediation analysis with treatment-induced confounding, and mediation analysis with multiple mediation pathways.

- Clarifying Identification Assumptions in Causal Mediation Analysis
  One of my research projects is to clarify identification assumptions in causal mediation analysis. This project provides a close examination of the definitions of the causal parameters of interest, identification/bounding assumptions and their connections, and widely used tools and statistical methods in mediation analysis.

- Causal Mediation Analysis with Treatment-Induced Confounding
  Treatment–induced confounding is present when some prognostic factors induced by the treatment occur before the mediator and have an effect on it. Sequential ignorability assumptions that are typically used for the identification of the natural direct effect

exclude treatment-induced confounding. Treatment-induced confounding is regarded as a difficult problem in mediaiton analysis. We provide new sets of identification assumptions, including two no-additional-heterogeneity assumptions, to identify the natural direct effect in the presence of treatment-induced confounding. Notably, the identified expression of the natural direct effect is the same as that of the interventional direct effect. We derive the semiparametric efficiency bound for the estimand and propose a multiply robust estimator that remains consistent under four types of possible mis-specification. To ensure model compatibility, we factorize the (conditional) joint distribution of the mediator and the treatment-induced confounder into marginal distributions and a dependence structure using copula.

- Causal Mediation Analysis with Multiple Mediators

  We consider a decomposition of the total indirect effect through multiple mediators, with an unspecified causal ordering, into individual components termed exit indirect effects and a remainder interaction term. We provide a set of identification assumptions for estimating all components. The identified expressions, which are closely related to the interventional indirect effects, continue to have causal interpretations when some identification assumptions are violated, as long as the total indirect effect is identified. We provide four moment-type estimators for each decomposed effect based on different parametrisations and derive the semiparametric efficiency bounds for the effects. The efficient influence functions contain conditional densities that are variation dependent, which is uncommon in existing problems, and we consider a reparameterization based on copulas to avoid model incompatibility and proposed a quadruply robust estimator for each of the decomposed effects that remains consistent and asymptotically normal under four types of possible misspecification and is also locally semiparametric efficient.

TABLE OF CONTENTS

Page

List of Figures iii

Chapter 1: Introduction 1

Chapter 2: Clarifying Identification Assumptions in Causal Mediation Analysis . . 9
2.1 Introduction 9
2.2 Setting and Notation 11
2.3 Direct Effects and Defining Assumptions 13
2.4 Identification Assumptions 17
2.5 Statistical Methods for Mediation Analysis 29
2.6 Discussion 31

Chapter 3: Causal Mediation Analysis with Treatment-Induced Confounding . . . 34
3.1 Introduction 34
3.2 Assumptions and Identification 37
3.3 Semiparametric Inference 40
3.4 Simulation study 45
3.5 Data Example 47
3.6 Discussion 48

Chapter 4: Causal Mediation Analysis with Multiple Mediators 51
4.1 Introduction 51
4.2 Effect Decomposition 54
4.3 Identification and Estimation of the Decomposed Effects 59
4.4 Simulation Studies 67
4.5 Data Application 69
4.6 Discussion 70

Appendix A: Appendix For Chapter 2 79

Appendix B: Appendix for Chapter 3 85
B.1 Proof for Theorem 3.1 and 3.3 88

Appendix C: Appendix For Chapter 4 92
C.1 Proof for Theorem 3.1 92
C.2 Proof for Theorem 3.2 92
C.3 Proof for Theorem 3.3 93
C.4 Proof of Theorem 3.4 96

LIST OF FIGURES

Figure Number Page

1.1 DAG for Randomized Controlled Trials 3
1.2 DAG with Confounding between Treatment and Outcome 4
1.3 The General Steps for Causal Inference 6
1.4 DAG for Mediation 6
1.5 Treatment-induced Confounding 7
2.1 **Relationship Between Assumptions**. Under NPSEM-IE, all three direct effects are identified. Under assumptions of Imai et al., Pearl(Pearl, 2001), Petersen et al., and Hafeman & VanderWeele, both the NDE (PDE) and the CDE are identified. Under the SWIG and MCM assumptions, the CDE is identified. The assumptions of Petersen et al. and Hafeman & VanderWeele are different since they are scale-dependent no-interaction assumptions, while the others are (conditional) independence assumptions. 12
2.2 Causal DAG for the Point-treatment case with a Mediator 12
2.3 Causal DAG with the Treatment-induced Confounding 27
2.4 Causal DAG for the Point-treatment case with a Mediator 31
2.5 Causal DAG of the setting 33
3.1 The Causal Diagram with Treatment-induced Confounding 38
4.1 The Causal DAG with Two Non-ordered Mediators 52
4.2 Possible causal mechanisms 56

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to the University of Washington, the school of Public Health, and especially the Department of Biostatistics for allowing me to pursue my Ph.D. degree here. I am extremely grateful to have learned from and worked with so many brilliant people.

I would like to thank my committee members. I wish to express my deep gratitude to my advisor, Professor. Gary Chan, for his profound belief in my abilities, his professional guidance, endless patience, support, and encouragement throughout my graduate life. This dissertation would not have been possible without him. I would also like to extend my gratitude to Professor James Hughes, who I am lucky to have worked with as a research assistant for five years, for broadening my research and showing me how to be a perfect collaborator. I am also grateful to Professor Thomas Richardson, for his invaluable input for all of my projects and for giving me the opportunity to help organize the causal inference working group. I wish to thank Professor Walter Kukull and Professor Carey Farquhar for helping me see my work from different perspectives.

I would like to extend my sincere thanks to Professor Patrick Heagerty and Professor Deborah Donnell, who I have worked with closely, for their professional insights, and for setting me great examples as researchers. I also want to thank Professor Lurdes Inoue for showing me how to be a good teacher. Special thanks to Gitana Garofalo, our graduate program advisor, for her unwavering support.

I must also thank my family for their emotional support. My mother, Lin Guo, and my father, Jingsong Xia, are always supportive of my choices even when they disagree. My grandfather, Zuze Guo, who I unexpectedly lost in my third year, was my inspiration. He

never had the chance to receive a good education when he was young, but he never stopped learning.

I am blessed to have shared this journey with my friends, especially my cohort and members of the causal inference working group. I very much appreciate Cesar Torres and Phuong Vu, for their helpful advice and practical suggestions throughout my graduate life. I am thankful to Yiran Wang, Junpeng Luo, Junxian Zhu, and Siru Guo, for keeping me sane towards the end of my dissertation. I also had great pleasure working with Dr.Fei Gao, from whom I learned a lot both as a researcher and as a friend.

DEDICATION

to my grandfather, Zuze

Chapter 1

INTRODUCTION

One of the fundamental goals of science is to understand how the world progresses by establishing causation. Questions about causality can be found in numerous fields, including biostatistics, epidemiology, and social science. What are the risk factors for human cancer? Is a healthcare intervention effective in a certain population? Can a given drug help prevent infectious disease? Will a labor market program increase employment? The attempts to answer these questions build up the study of causality, i.e., causal inference.

Although causal thinking is elementary in perceiving events, it is often difficult to reach a consensus on the meaning of causation among people, for the concept is often self-taught by experience. Therefore, the first question that needs to be answered is what causation means. It is widely known that association does not imply causation. A natural question is when an association suggests causation. Once a causal relationship is established, the next question is how to evaluate its strength.

The definition of causation is well-established using the potential outcome framework for causal inference (Neyman 1923, Rubin 1974, 1978). It conceptualizes causation by introducing the idea of potential outcomes that would have occurred had the upstream factors taken certain values that are potentially different from their observed values. An intervention has a causal effect on the outcome if the potential outcomes corresponding to different intervention values are different. This definition captures a subset of causation in which the change in outcome depends on the change in the causes. The potential outcome framework formalizes the definition of causal effects and consequently facilitates the translation of scientific questions into well-defined statistical questions.

Each set of action/exposure values indexes a potential outcome, which forms a set of

potential outcomes corresponding to different actions/exposures. In reality, however, only one potential outcome is observed: the one that corresponds to the actual action or exposure. This potential outcome is the factual outcome, making other potential outcomes "counterfactuals". [1] The naturally embedded missingness of potential outcomes is called the fundamental problem of causal inference. If we view different potential outcomes as factual outcomes from different alternative realities ("worlds") created by different values of actions or exposures, the fundamental problem of causal inference is a result of our inability to travel back in time and experience a different reality. Consequently, causal effects are not directly targetable like usual estimands that are defined using observable data. In this paper, we call estimands defined by functions of potential outcomes causal estimands, distinct from usual estimands that are directly targetable.

One causal estimand that is often of interest is the average causal/treatment effect. It depicts the total effect of an intervention/exposure on the outcome. Consider a generic setting in which we are interested in the causal effect of a binary treatment on an outcome. Denote the treatment as A, the outcome as Y, and the potential outcome as $Y(a)$, which is the value the outcome would have taken had the treatment been set to a. The average causal effect is defined as $E[Y(1)-Y(0)]$, the mean difference between the potential outcomes under treatment and control. The observed outcome for units in the treatment group is $Y(1)$, and the observed outcome for units in the control group is $Y(0)$. Table 1 demonstrates the relationship between the potential outcomes and the observable outcome for each unit. Table 2 demonstrates the fundamental problem of causal inference: observable outcome Y does not recover the joint distribution of $(Y(0), Y(1))$.

Now that the average causal effect is well-defined, the question remains: How to associate this causal estimand to a directly targetable estimand that depends on observable data? One option is through randomized controlled trials. Since the treatment is randomized, there is

[1] We use potential outcomes and counterfactuals interchangeably in this paper.

Unit	A_i	$Y_i(0)$	$Y_i(1)$	Y_i
1	1	0	1	1
2	0	1	0	1
⋮	⋮	⋮	⋮	⋮
n	0	0	1	0

(a) Potential and Observed Outcomes

Unit	A_i	$Y_i(0)$	$Y_i(1)$	Y_i
1	1	?	1	1
2	0	1	?	1
⋮	⋮	⋮	⋮	⋮
n	0	0	?	0

(b) Fundamental Problem of Causal Inference

no confounding between the treatment and the outcome. The absence of confounders implies the independence between the treatment A and the potential outcome $Y(a)$. As illustrated in Figure 1.1, since no other pathways are connecting A and Y, A cannot affect $Y(a)$ with treatment set to a in the potential outcome. With the independence

$$A \perp\!\!\!\perp Y(a), \tag{1.1}$$

we can rewrite $E[Y(a)]$ as $E[Y(a)|A = a]$, which is the mean factual outcome in the treatment group a: $E[Y|A = a]$. Therefore, the causal estimand average causal effect $E[Y(1) - Y(0)]$ is translated into a directly targetable estimand $E[Y|A = 1] - E[Y|A = 0]$.

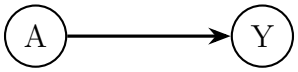


Figure 1.1: DAG for Randomized Controlled Trials

Randomized controlled trials are the gold standard in establishing causal relationships in health-related research. In practice, however, randomization is not always feasible due to practical, logistical, and ethical reasons. Therefore, we need to conduct causal inference using observational data. For example, the U.S. Food & Drug Administration recently started advocating the use of real-world evidence, including electronic health records, disease registries, and health tracking data on mobile devices in regulatory decision-making other

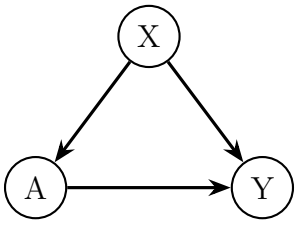


Figure 1.2: DAG with Confounding between Treatment and Outcome

than traditional clinical trials. Without the protection of randomization, confounding cannot be prevented in observational data. Denote the confounder between the treatment A and the outcome Y as X. As illustrated in Figure 1.2, after setting the treatment in $Y(a)$ to value a, A and Y are still connected by X, so A and $Y(a)$ are no longer independent. Suppose X includes all confounders for the relationship between the treatment and the outcome. Conditioning on X, connections between the treatment and the outcome, other than the direct pathway from A to Y, are blocked. In this case, A is independent of $Y(a)$ with treatment set to value a. With this conditional independence

$$A \perp\!\!\!\perp Y(a)|X, \tag{1.2}$$

we can rewrite $E[Y(a)]$ as:

$$E\{E[Y(a)|X]\} = E\{E[Y(a)|A = a, X]\} = E\{E[Y|A = a, X]\}.$$

Therefore, the causal estimand average causal effect $E[Y(1) - Y(0)]$ is translated into an estimable target $E\{E[Y|A = 1, X] - E[Y|A = 0, X]\}$.

The process of associating a causal estimand to an estimable target is called identification. The condition 1.1 needed to identify the average causal effect is satisfied by the study design. In observational studies, condition 1.1 no longer holds. We instead assume the conditional independence 1.2, which prohibits unmeasured confounding, for the identification of the average causal effect. Assumptions needed to identify a causal estimand are called identification assumptions, such as 1.1 and 1.2. The equivalence of the causal estimand

and the identified targetable estimand relies on the validity of identification assumptions. In other words, whether a targetable estimand, derived from observable data, has a causal interpretation is determined by whether the identification assumptions are satisfied.

Identification is answering the question: what conclusions can be drawn from infinite data? Infinite data can recover the true distribution of observable data. In practice, however, we do not have an infinite amount of data. Although the identified targetable estimand does not depend on unobservable potential outcomes, statistical methods are needed for their estimation and inference. In randomized controlled trials, the target $E[Y|A=1]-E[Y|A=0]$ of the average causal effect can be estimated by a difference-in-mean estimator. In observational studies with no unmeasured confounding (assumption 1.2 holds), the targetable estimand $E\{E[Y|A=1,X]-E[Y|A=0,X]\}$ can be estimated by a regression-based estimator.

More generally speaking, we can roughly divide causal inference into the following three steps as illustrated by Figure 1.3:

- Determine a casual estimand based on the scientific question of interest and formalize it into a function of potential outcomes.

- Link the causal estimand to observable data using identification schemes. If reasonable identification assumptions are not enough for point identification, bounds on the causal estimand can be derived instead.

- Use statistical methods to estimate the identified targetable estimand.

The average causal effect depicts the total effect of an intervention/exposure on the outcome. It answers two questions: Does the intervention have a causal effect on the outcome? What is the magnitude of the effect? On many occasions, we are also interested in how the intervention/exposure affects the outcome. When intermediate variables are measured, causal mechanisms can be investigated through the study of direct and indirect effects. The direct effect is the part of the causal effect of an intervention/exposure that does not act through some intermediate variables, known as mediators. Conversely, the part of the causal

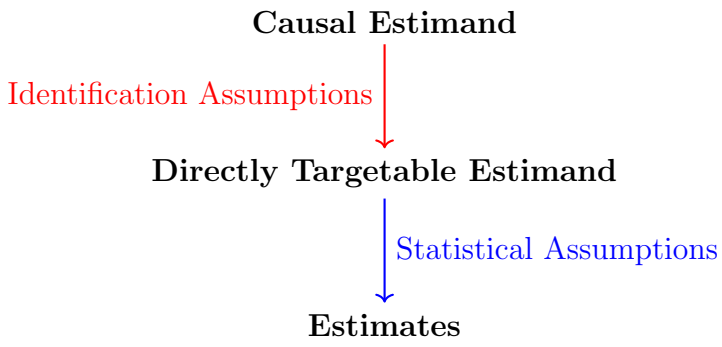


Figure 1.3: The General Steps for Causal Inference

effect that acts through the mediators is called the indirect effect. As illutrated in Figure 1.4, A is the treatment, M is a mediator through which A affects the outcome Y. Mediation analysis often includes the definition, identification, and estimation of a direct effect of an exposure and its indirect effect that goes through some known mediators.

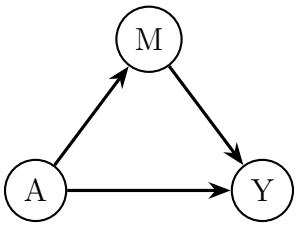


Figure 1.4: DAG for Mediation

Intuitively, to evaluate a direct effect, the mediators need to be somehow fixed. Depending on the scientific question of interest, a variety of direct effects can be defined through different ways of fixing the mediators. Frequently used definitions of direct effects include the controlled direct effect, the pure (natural) direct effect, and the principal stratum direct effect. In the settings with a binary treatment, the controlled direct effect compares the potential outcomes under treatment and control with the mediator fixed at a certain level.

The natural direct effect compares the potential outcomes under treatment and control with mediators fixed to the value they would have taken had there not been any treatment. The principal stratum direct effect compares the potential outcomes with and without treatment in the subgroup where the mediator remains constant under treatment and control. The formal definitions of these direct effects are given in chapter 2.

Each of the direct effect has unique properties in terms of definition, identification, and interpretation. We examine these aspects of different direct effects closely in chapter 2 to facilitate the application of causal mediation analysis methods in practice. Specifically, we provide a comparison of the prerequisites for each direct effect to be well-defined. We also compare the strength of multiple sets of identification assumptions for each direct effect. We provide intuitions to their definitions and identification assumptions and explain their different interpretations. We give a brief summary of the methodological developments in mediation analysis.

In chapter 3, we focus on the pure (natural) direct effect. As will be explained in chapter 2, the pure direct effect is most relevant in studying treatment effect mechanisms. The identification assumptions for the pure direct effect usually prohibit treatment-induced confounding between the mediator and the outcome. A mediator-outcome confounder is treatment-induced if it is affected by the treatment, and subsequently affects the mediator and the outcome, as illustrated by C in Figure 1.5. In practice, however, this assumption is often violated.

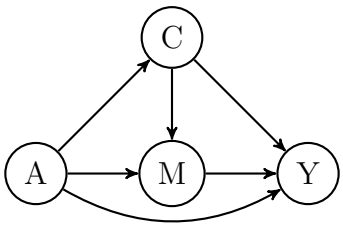


Figure 1.5: Treatment-induced Confounding

To address this fundamental limitation, we provide a set of assumptions that identify the natural direct effect in the presence of treatment-induced confounding, where the identified expression is equivalent to a recently studied interventional direct effect defined for a randomized mediator. We derive the semiparametric efficiency bound for the estimand, which unlike usual expressions, contains conditional densities that are variation dependent. We consider a reparameterization using copula and propose a quadruply robust estimator that remains consistent under four types of possible misspecification and is also locally semiparametric efficient. We use simulation studies to demonstrate the quadruple robustness and apply our method to the 2017 Natality data (CDC (2017)) to investigate the effect of prenatal care on preterm birth mediated by preeclampsia with smoking status during pregnancy being a potential treatment-induced confounder.

In previous chapters, only one mediator is known or of interest and measured. When multiple mediators are measured, the total indirect effect, which is the difference between the average causal effect and the natural direct effect, contains a joint effect of multiple mediators. Apart from the joint effects, the estimation of effects that go through each mediator is also of interest. In chapter 4, we decompose the total indirect effect through multiple mediators, with an unspecified causal ordering, into individual components termed exit indirect effects and a remainder interaction term. We provide a set of identification assumptions for all components. The identified expressions, which are closely related to the interventional indirect effects, continue to have causal interpretations when some identification assumptions are violated, as long as the total indirect effect is identified. We provide four moment-type estimators for each decomposed effect based on different parametrizations and derive the semiparametric efficiency bounds for the effects. The efficient influence functions contain conditional densities that are variation dependent, so we consider a reparameterization based on copulas to avoid model incompatibility. We propose a quadruply robust estimator for each of the decomposed effects that remains consistent and asymptotically normal under four types of possible misspecification and is also locally semiparametric efficient.

Chapter 2

CLARIFYING IDENTIFICATION ASSUMPTIONS IN CAUSAL MEDIATION ANALYSIS

2.1 Introduction

Under the counterfactual framework (Neyman, 1923; Rubin (1974),Rubin (1978), Rubin (1990);), causal contrasts are defined using potential outcomes. Potential outcomes are generally only partially available for an individual in the sense that only the potential outcomes corresponding to the actual treatment assignments can be observed. Consequently, assumptions are needed to derive corresponding effect estimates for the causal contrasts using the empirical data.

Mediators are intermediate variables through which an intervention/exposure affects the outcome. The total effect of an intervention/exposure can be divided into two parts: the part that acts through the mediators, i.e., the indirect effect, and the part that does not, i.e., the direct effect. Intuitively, to get the direct effect of an intervention, the mediators in its causal pathways need to be "fixed". There are several strategies to define the direct and indirect effects using the potential outcome framework depending on different ways the mediator is "fixed". In this paper, we focus on three frequently used definitions for direct effects, including the controlled direct effect (CDE), the pure direct effect[1] (PDE), and the principal stratum direct effect (PSDE) (Robins, 1986; Robins and Greenland, 1992; Pearl, 2001; Frangakis and Rubin, 2002; Rubin, 2004). Different direct effects depend on different potential outcomes that come with different implicit assumptions. These differences are subtle and often not discussed in methodological literature. However, they are essential for the clarification of each direct effect's range of application and interpretation. We make

[1]The pure direct effect is also called the natural direct effect (NDE)

explicit and compare the defining assumptions, and clarify the interpretation and application of the controlled direct effect, the pure direct effect, and the principal stratum direct effect.

Besides the formalization of different direct effects using the potential outcome framework, various sets of identification assumptions are considered by Robins and Greenland (1992), Pearl (2001), Petersen et al. (2006), Hafeman and VanderWeele (2011), Richardson and Robins (2013), Robins and Richardson (2010) , Imai et al. (2010) for the controlled direct effect and the pure direct effect. Apart from the point identification of the pure direct effect, bounds are given by Robins and Richardson (2010), Tchetgen and Phiri (2014), Miles et al. (2015). Similarly, point identification and bounds on the principal stratum direct effect are considered by Zhang and Rubin (2003), Cheng and Small (2006), Frangakis et al. (2007), Imai (2008). A discussion of the relationship between the principal stratum direct effect and the other two direct effects is given by VanderWeele (2008).

Currently, there is a lack of literature comparing identification assumptions for direct effects. We examine the connections between the popular sets of assumptions, including assumptions that arise from Non-parametric Structural Equation Models with Independent Errors (NPSEM-IE) considered in Pearl (2009), sequential ignorability assumptions given by Imai et al. (2010), identification assumptions given by Pearl (2001), Petersen et al. (2006), and Hafeman and VanderWeele (2011), the assumptions encoded in Single World Intervention Graphs (SWIG) developed by Richardson and Robins (2013), and the assumptions of the Minimal Counterfactual Model (MCM) considered by Robins and Richardson (2010). Some extra sets of identification assumptions that are directly inspired by them are also included for the purpose of comparison. We prove the relationships between these assumptions and summarize them in Figure 2.1.

After an appropriate direct effect is chosen and identified, we resort to statistical methods for its estimation. These methods are not unique for causal mediation analysis. Nonetheless, we dedicate a section that reviews the methodological developments in the estimation of mediation effects. In particular, we describe the much used parametric structural equation models, the regression-based outcome/mediator models, the inverse probability weighting

methods, and a triply-robust estimator that has desirable theoretical properties. We also discuss the relationship between these methods. In particular, traditional approaches such as the linear structural equation models (LSEM), which includes the famous Baron and Kenny methods, can be formalized by the potential outcome framework in terms of their underlying assumptions and the causal interpretations of the estimators. The triply-robust estimator is a combination of the outcome/mediator models and the inverse probability weighting methods.

For simplicity, we consider the point-treatment case in our paper. However, the investigation of mediation effects extends to more complicated settings and thus so do the definitions, identification assumptions, and methods. In the discussion section, we briefly discuss mediation analysis in longitudinal cases with time-varying confounding: why confounding adjustment fails, and what are the alternative methods. We also include a discussion on the important idea of the extended DAG (Robins and Richardson, 2010) that reconciles the controversy of the pure direct effect being non-manipulable.

This paper is organized as follows. In section 2, we introduce the point-treatment setting we consider and relevant notations. In section 3, we examine the definitions of the direct effects, the assumptions that come with their definitions, and their implications. In section 4, we review the identification assumptions for the direct effects and explore their relationships. In section 5, we introduce the statistical methods used for mediation analysis. In section 6, we discuss some other developments in causal mediation analysis.

2.2 Setting and Notation

Let A be a binary treatment, M be a mediator, Y be an outcome, and X be baseline covariates that potentially confound the relationship between A and M, the relationship between M and Y, and the relationship between A and Y. The causal DAG is displayed in Figure 2.2.

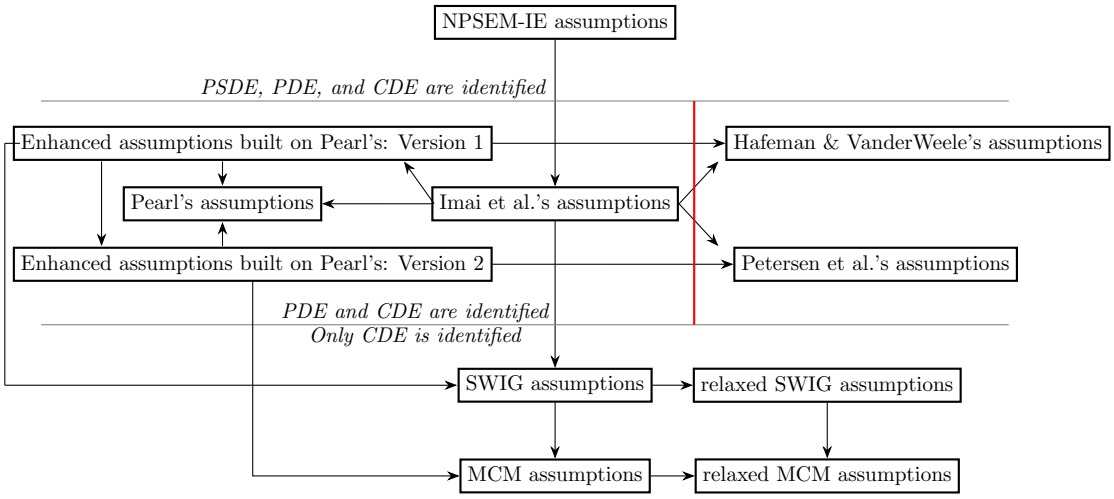


Figure 2.1: **Relationship Between Assumptions**. Under NPSEM-IE, all three direct effects are identified. Under assumptions of Imai et al., Pearl(Pearl, 2001), Petersen et al., and Hafeman & VanderWeele, both the NDE (PDE) and the CDE are identified. Under the SWIG and MCM assumptions, the CDE is identified. The assumptions of Petersen et al. and Hafeman & VanderWeele are different since they are scale-dependent no-interaction assumptions, while the others are (conditional) independence assumptions.

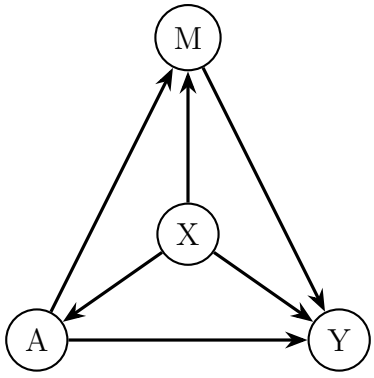


Figure 2.2: Causal DAG for the Point-treatment case with a Mediator

The outcome Y depends on the value of both the treatment A and the mediator M. The potential outcome, denoted as $Y(a, m)$, is the value Y would have taken had A been a and M been m. Each combination of a and m creates an alternative world in which the treatment takes value a and the mediator takes value m, and $Y(a, m)$ would be the observed outcome Y. Similarly, the mediator M depends on the value of the treatment A, and $M(a)$ is the value M would have taken had A been a.

2.3 Direct Effects and Defining Assumptions

The controlled direct effect, the pure direct effect, and the principal stratum direct effect are defined below using the counterfactual framework for a binary treatment in the difference scale. Direct effects can be similarly defined in other scales corresponding to risk ratios or odds ratios.

Definition 2.3.1 *Controlled direct effect*

$$CDE(m) \equiv \mathbb{E}[Y(1, m) - Y(0, m)],$$

Definition 2.3.2 *Pure direct effect*

$$PDE \equiv \mathbb{E}[Y(1, M(0)) - Y(0, M(0))],$$

Definition 2.3.3 *Principle stratum direct effect*

$$PSDE(m) \equiv \mathbb{E}[Y(1, m) - Y(0, m)|M(0) = M(1) = m].$$

The controlled direct effect compares the effect of the combined treatment of A and M on Y between two groups with different levels of A while fixing M at the same level. The pure direct effect compares the effect of the treatment A on Y between the treatment group and the control group while fixing M at the level it would have taken had there not been any treatment. The principal stratum direct effect compares the treatment effect on Y in a subgroup where the mediator remains unchanged with or without treatment.

Throughout our discussion we will assume the suitable unit treatment value assumption (SUTVA). SUTVA allows only one version of the treatment, and it precludes interference between units. A few more technical conditions are needed for the counterfactuals used in mediation analysis to be defined (VanderWeele and Vansteelandt, 2009), specifically, the consistency assumption and the composition assumption. The consistency assumption, which is subsumed by SUTVA, states that when the treatment takes a certain value the observed outcome for each variable equals their corresponding potential outcome with the treatment set to that value. In our case, it implies $Y(a) = Y, M(a) = M$ when $A = a$, and $Y(a, m) = Y$ when $A = a$, and $M(a) = m$. The composition assumption, which is conceptually less challenging than SUTVA or the consistency assumption, can be illustrated by an example: $Y(a, M(a)) = Y(a)$. It implies that the potential outcome with the treatment set to a and the mediator set to the value it would have taken under the same treatment equals the potential outcome with the treatment set to a.

Before choosing one of these direct effects as the causal estimand of interest, we need to examine the assumptions and properties of their definitions:

- **Reliance on a conceivable intervention** For all three direct effects, the intervention on the treatment A needs to be possible, or at least conceivable. For the controlled direct effect and the pure direct effect, the intervention on the mediator M also needs to be possible or conceivable. For the principal stratum direct effect, a plausible intervention on the mediator is not needed.

  We explain this concept through an example. Suppose A is the race of a candidate i, and Y is the result of a job application, it is difficult to envision candidate i's job application outcome had they been a different race, because by changing the race, candidate i is essentially a different person. One can argue that in this case, the potential outcomes are not well-defined and hence the causal contrast is no longer defined. To avoid this conceptual problem, a different question can be asked: what is the effect of the race indicated in a résumé on the job application results? The

potential outcome $Y_i(a)$ is then what the job application result would be had the race information on a résumé been set to a.

The notion of defining counterfactuals based on conceivable interventions is generally accepted in biostatistics and epidemiology studies, and causation can be established without suggesting a mechanism. This notion is closely(Holland, 1986), which advocates against using attributes as causes. The emphasis on manipulability may not apply to other fields, where modeling the mechanism may be more of interest and causes are depicted by the relations between variables.

- **Different contrast for each value of the mediator** The controlled direct effect/the principal stratum direct effect takes a different value when the mediator is set to a different value of m. On the other hand, the pure direct effect does not require setting the mediator to a specific value of m. Therefore it depicts an overall direct effect that averages over the distribution of the potential mediator without treatment.

- **Inference on subgroup** The principal stratum direct effect is defined on a subgroup in which the potential mediator takes the same value with or without the treatment. The principal stratum direct effect is more useful when the mediator is binary or categorical. When the mediator M is continuous, the subgroup defined by M is trivial, and the effect that does not act through a dichotomized version of M is not well-defined (Robins et al., 2007).

- **"Cross-world" definitions and Manipulative effects** Two counterfactuals are from different "worlds" if their shared causes are set to different values. For example, $Y(1, m)$ and $M(0)$ are from different "worlds" because in $Y(1, m)$, the treatment is set to 1, and in $M(0)$, the treatment is set to 0.

  Following this notion, we can see that both the pure direct effect and the principal stratum direct effect are defined using counterfactuals from different "worlds". As

a result, they are not "manipulative" effects. An effect is manipulable if it can be realized by a contrast between treatment regimes from an experiment on an identified subgroup. For the principal stratum direct effect, it is impossible to identify the subset of subjects whose mediators remain unchanged with or without treatment. For the pure direct effect, it is impossible to conduct an experiment (consisting of an intervention on the variables A, M) that simultaneously sets the treatment (A) to 1 for the outcome (Y) and 0 for the mediator (M). On the contrary, the controlled direct effect is manipulable: one can conduct an experiment in which both the treatment and the mediator are randomized.

- **Decompose the total effect** The average causal effect (average treatment effect) is defined as $E[Y(1) - Y(0)]$. It depicts the total effect of the treatment on the outcome, and can be decomposed into the sum of the pure direct effect and the total indirect effect:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1, M(0)) - Y(0, M(0))] + \mathbb{E}[Y(1, M(1)) - Y(1, M(0))].$$

  Since the controlled direct effect depends on the level the mediator is set to, it is not straightforward to define a corresponding indirect effect. In some literature, the indirect effect is defined precisely as the difference between the average treatment effect and the controlled direct effect. Since the principal stratum direct effect is defined on a subgroup in which the mediator remains unchanged, it does not have a corresponding indirect effect.

Different definitions lead to different interpretations. The controlled direct effect can be interpreted as the difference in outcome between two treatment[2] regimes of a randomized experiment, one with $A = 1$ and $M = 1$, the other with $A = 1$ and $M = 0$. The difference is, therefore, attributed to the change in M. The pure direct effect, on the other hand, answers the question: how much would the treatment A affect the outcome if the effect of A on

[2]here treatment represent the combination of A and M.

its mediator M is blocked? When the effect of treatment A on mediator M is blocked, M remains the same (under treatment) as if there had been no treatment: $M(0)$. The principal stratum direct effect (at level m) can be rewritten as:

$$\begin{aligned}&\mathbb{E}[Y(1,m)-Y(0,m)|M(0)=M(1)=m]\\ =&\mathbb{E}[Y(1,M(1))-Y(0,M(0))|M(0)=M(1)=m]\\ =&\mathbb{E}[Y(1)-Y(0)|M(0)=M(1)=m].\end{aligned}$$

Under the composition assumption, the principal stratum direct effect can be interpreted as the average causal effect in the subgroup where the mediator M does not change with the treatment A. We include the properties of these three direct effect definitions in Table 1. Properties set in italics can be favorable. [3]

	Conceivable intervention		**Separate contrast for each level of m**	**Inference on identified population**	**Manipulative effect**	**Indirect effect defined**
	on A	on M				
CDE	Necessary	Necessary	No	*Yes*	*Yes*	No
PDE	Necessary	Necessary	*Yes*	No	No	*Yes*
PSDE	Necessary	*Unnecessary*	No	No	No	No

Table 2.1: Properties of Direct Effect Definitions

2.4 Identification Assumptions

As described in the introduction section, the fundamental problem with causal inference is that counterfactuals used to define causal contrasts cannot be observed simultaneously for

[3]The relevant counterfactuals are $M(a)$ and $Y(a,m)$.

an individual. As a result, identification assumptions are needed to link the causal estimands defined using counterfactuals to the observable data. Identification assumptions are usually imposed on the relationship, such as conditional independence, between counterfactuals. The causal interpretation of the identified effect relies on the validity of identification assumptions. Therefore, it is essential to fathom the strength of identification assumptions and verify their reasonableness before drawing any causal conclusions.

Before comparing different sets of identification assumptions for direct effects, we introduce the concept of "Cross-world" assumptions. In the previous section, we explained how counterfactuals could be from different worlds. The same notion extends to identification assumptions. An identification assumption is "Cross-world" if it is imposed on counterfactuals from different worlds. For example, a "Cross-world" assumption often used is on the independence between the potential outcome $Y(a, m)$ and the potential mediator $M(a^\star)$ where a and $a^\star$ can take different values. "Cross-world" assumptions are non-refutable (without further assumptions or additional interventions) because they cannot be tested by any experiment. In contrast to "Cross-world" assumptions, there are "Single-world" assumptions that only involve counterfactuals from the same world. "Single-world" assumptions can be experimentally tested. Assumptions that cannot be tested by randomized experiments are usually less favorable because their validity depends solely on scientific belief. Sensitivity analyses may help us fathom the effect of assumption violations.

In this section, we examine identification assumptions for the direct effects for the simple setting, as illustrated by Figure 2.2. Since definitions of direct effects have different properties, the strengths of assumptions needed for their identifications vary. In particular, the principal stratum direct effect requires the strongest identification assumptions because it involves the joint distribution of four counterfactuals from different worlds: $(Y(1), Y(0), M(1), M(0))$. The pure direct effect is less restrictive, for it involves the joint distribution of three counterfactuals from different worlds: $(Y(1, m), Y(0), M(0))$. The assumptions needed for the identification of the controlled direct effect is the least restrictive because it only involves counterfactuals from a single world.

As a result, different sets of identification assumptions can be divided into three subgroups:

- the sets that identify all three direct effects;
- the sets that identify the pure direct effect and the controlled direct effect, but not the principal stratum direct effect;
- the sets that only identify the controlled direct effect.

Specifically, we start from the strongest set of assumptions under which all three direct effects are identified: the NPSEM (Non-Parametric Structural Equation Models) with independent errors (sometimes called NPSEM-IE) assume mutual independence between sets of counterfactuals associated with different variables. This Independent Error assumption may not always be reasonable. We move on to the less restrictive sets of assumptions, under which the pure direct effect and the controlled direct effect are identified, including the widely-used sequential ignorability assumptions (Imai et al., 2010), and five other sets of assumptions that are slightly weaker with a narrower range of application. These two groups of assumptions still involve "Cross-world" assumptions that are non-refutable. We then move on to the least restrictive sets of assumptions under which only the controlled direct effect is identified, including the single world intervention graphs (SWIGs) assumptions that are associated with causal graphs. The last group of assumptions only involve "Single-world" assumptions. The shared focus of these assumptions are on three sources of confounding:

- confounding between the treatment and the outcome;
- confounding between the treatment and the mediator;
- confounding between the mediator and the outcome.

Intuitively, the more control we have over the confounding factors, the closer the conditional association is to causation.

In this section, we assume that X includes the set of pre-treatment covariates, meaning that the treatment induces no potential confounding. The identification of the pure direct effect is usually not possible in the presence of treatment-induced confounding under the sequential ignorability assumptions (Avin et al., 2005). We come back to this point at the end of this section.

Assumption 2.4.1 *The Non-Parametric Structural Equation Models independence assumption. The following three sets of variables:*

$$\{X, A(x^{\star\star\star\star})\}, \{M(1,x), M(0,x^{\star})\}, \{Y(1,m,x^{\star\star}), Y(0,m^{\star},x^{\star\star\star})\}\}$$

are mutually independent, where each set includes all the variables obtained by allowing $x^{\star\star\star\star}$, $x^{\star\star\star}$, $x^{\star\star}$, $x^{\star}$, *and* x; m *and* $m^{\star}$ *take different values.*

To simplify this assumption, we omit the covariates X which we usually condition on. Denote the domain of M as $\mathcal{M}$ and rewrite the assumption as:

$$A, \{M(1), M(0)\}, \{Y(1,m), Y(0,m^{\star}), m, m^{\star} \in \mathcal{M}\} \tag{2.1}$$

are mutually independent. It is clear to see that counterfactuals in the second and third sets can be from different worlds: $M(1)$ and $Y(0,m^{\star})$, $M(0)$ and $Y(1,m)$.

Equation (2.1) includes three systems of variables in this setting: the treatment A, the set of potential mediators indexed by different treatment values $\{M(1), M(0)\}$, and the set of potential outcomes indexed by different treatment and mediator values $\{Y(1,m), Y(0,m^{\star})\}$. The NPSEM-IE assumes mutual independence between these three systems. NPSEM-IE implies a list of conditional independences between potential outcomes, potential mediators, and the treatment assignment. As a result, all three direct effects are identified under NPSEM-IE.

To better connect NPSEM-IE to the rest of the assumptions, we present a sequential ignorability implication of it:

Assumption 2.4.2 *NPSEM-IE Assumption Implication*

1. $\{Y(a, m, x^\star), M(a^\star, x^{\star\star\star})\} \perp\!\!\!\perp A(x)|X = x^{\star\star}$,

2. $Y(a, m, x) \perp\!\!\!\perp M(a^\star, x^\star)|A = a^{\star\star}, X = x^{\star\star}$.

The value of A and X in these conditions could be different from their set values in the counterfactuals. The first assumption implies there is no (unmeasured) confounding between the treatment and the downstream variables. The second assumption implies there is no (unmeasured) confounding between the mediator and the outcome. The assumptions are on the "ignorability" of treatment/mediator assignment. [4]

A weaker set of sequential ignorability assumptions from Imai et al. (2010) requires independence only when the confounding variables X are consistently taking the same value in the conditions and the set values of counterfactuals, restricting the space where the independence holds to a smaller one where X is constant. Condition 2 in the Assumption 2.4.2 is further weakened by fixing the value of the treatment in the potential mediator to the treatment value in the condition.

Assumption 2.4.3 *Imai's Assumptions*

1. $\{Y(a, m), M(a^\star)\} \perp\!\!\!\perp A|X = x$,

2. $Y(a^*, m) \perp\!\!\!\perp M(a)|A = a, X = x$.

It is straight-forward to see by fixing X to x, condition 1 in Assumption 2.4.2 implies condition 1 in Assumption 2.4.3. By equating $a^\star$ and $a^{\star\star}$ in condition 2 of Assumption 2.4.2, we have condition 2 of Assumption 2.4.3.

Condition 1 states that the joint distribution of the potential outcome and the potential mediator, possibly with different set treatment values, are independent of the actual treatment given the baseline covariates. Condition 1 implies that covariates X sufficiently control the confounding between the treatment and the downstream variables. In other words, there

[4] "Ignorability" is also referred to as "randomization", "exchangeability", and "endogeneity" in different literatures.

is no unmeasured confounding between the treatment and the mediator, and there is no unmeasured confounding between the treatment and the outcome.

Condition 2 can be restated as $Y(a^\star, m) \perp\!\!\!\perp M|A = a, X = x$ according to the consistency assumption. Similar to condition 1, condition 2 implies there is no unmeasured confounding between the mediator and the outcome. This is a "Cross-world" assumption because a and $a^\star$ can take different values. Consequently, when $a \neq a^\star$, M and $Y(a^\star, m)$ are from different worlds.

A similar but maybe less tangible assumption is to impose independence conditioning only on X, as given by Pearl (2001). In fact, Assumption 2.4.3 implies Pearl's assumption (2.4.4).

Assumption 2.4.4 *Pearl's Assumptions*

There exists a set X of covariates, non-descendants[5] *of neither A nor M, such that, for all values m and a we have:*

1. $\mathbb{P}(M(a^\star) = m|X = x)$ *is identifiable,*

2. $\mathbb{P}(Y(a, m) = y|X = x)$ *is identifiable,*

3. $Y(a, m) \perp\!\!\!\perp M(a^\star)|X$.

Condition 3 is slightly different from condition 2 in Assumption 2.4.3, but the interpretation is similar. It assumes no unmeasured confounding between the mediator and the outcome.

Condition 1 and 2 assume identifiability directly. For the purpose of ordering the strength of assumptions, we give two slightly enhanced versions of Pearl's Assumptions in which $M(a)$ and $Y(a, m)$ are identified. Both of them include the key independence condition 3 in Assumption 2.4.4.

Assumption 2.4.5 *Enhanced Assumptions based on Pearl's: Version 1*

[5]No post-intervention (treatment-induced) confounding. Here mediator is also considered to be a secondary intervention.

1. $\{Y(a,m), M(a^\star)\} \perp\!\!\!\perp A|X = x$,

2. $Y(a,m) \perp\!\!\!\perp M(a^\star)|X = x$.

Assumption 2.4.6 *Enhanced Assumptions based on Pearl's: Version 2*

1. $Y(a,m) \perp\!\!\!\perp A|X = x$,

2. $M(a) \perp\!\!\!\perp A|X = x$,

3. $Y(a,m) \perp\!\!\!\perp M|A, X = x$,

4. $Y(a,m) \perp\!\!\!\perp M(a^\star)|X = x$.

Condition 1 in Assumption 2.4.5 is the same as the condition 1 of Assumption 2.4.3, which implies the independences 1. and 2. imposed by Assumption 2.4.6. Under condition 1 in Assumption 2.4.5, condition 2 in Assumption 2.4.5 implies condition 3 in Assumption 2.4.6.

Both Assumption 2.4.5 and Assumption 2.4.6 would hold in a situation where there is

- no unmeasured confounding between the treatment and the mediator; and
- no unmeasured confounding between the treatment and the outcome; and
- no unmeasured confounding between the mediator and the outcome,

which is similar to Assumption 2.4.3. The differences between these sets of assumptions are technical in practice (Imai, 2008).

Next, we examine two sets of identification assumptions that are weaker with a narrower range of application compared with the previous mentioned sets.

A weakened version of condition $Y(a,m) \perp\!\!\!\perp M|A, X$ is given by Hafeman and Vander-Weele (2011), which is applicable when the mediator is binary,

Assumption 2.4.7 *Hafeman & VanderWeeles's Assumption*

When the mediator M is binary,

1. $\{Y(a,m), M(a^\star)\} \perp\!\!\!\perp A|X = x$,

2. $\mathbb{E}[Y(a, m=0)|A=a, M=1, x] = \mathbb{E}[Y(a, m=0)|A=a, M=0, x]$,

3. $\mathbb{E}[Y(a,m=1) - Y(a,m=0)|A=a, M=1, x] = \mathbb{E}[Y(a,m=1) - Y(a,m=0)|A=a^\star, M=1, x]$

Since $Y(a,m) \perp\!\!\!\perp A|X$, the independence $Y(a,m) \perp\!\!\!\perp M|A, X=x$ implies condition 2 and 3 in Assumption 2.4.7. It can also be shown that this set of assumptions is implied by Assumption 2.4.5.

Condition 1 in Assumption 2.4.7 is the common assumption that implies no unmeasured confounding between the treatment and the mediator/outcome. Condition 2 implies that the value of the observed mediator does not modify the expected potential outcome with m fixed at 0. Condition 3 implies that the controlled effect of the mediator is the same across treatment groups in the subgroup where the observed mediator takes the value 1. Condition 2 and condition 3 are weaker, scale-dependent, "Cross-world" assumptions that replace the "Cross-world" independence between $Y(a,m)$ and $M(a^\star)$.

On the other hand, Petersen et al. (2006) came up with a slightly weaker set of assumptions (compared to Pearl's assumption) in a different direction, replacing the joint conditional independence between $(Y(a,m), M(a^\star))$ and the treatment in condition 1 of Assumption 2.4.3 by two separate marginal independence assumptions, at a cost of an additional direct effect assumption.

Assumption 2.4.8 *Petersen, Sinisi and Van der Laan's Assumptions*

1. $Y(a,m) \perp\!\!\!\perp A|X = x$,

2. $M(a) \perp\!\!\!\perp A|X = x$,

3. $Y(a,m) \perp\!\!\!\perp M|A, X = x$,

4. (Direct effect assumption)
$\mathbb{E}[Y(a,m) - Y(0,m)|M(0) = m, X] = \mathbb{E}[Y(a,m) - Y(0,m)|X]$.

The direct effect assumption implies that the realization of the potential mediator without treatment $M(a = 0)$ does not provide addition heterogeneity on the expected effect of the treatment on the outcome at a controlled level of mediator m. It can be shown that Assumption 2.4.6 implies Assumption 2.4.8. Condition 4 is a weaker, scale-dependent, "Cross-world" assumption that replaces the "Cross-world" independence between $Y(a,m)$ and $M(a^\star)$.

Assumption 2.4.7 and 2.4.8 are fundamentally different from other assumptions because they are imposed in a difference scale, so they are relevant when the direct effects are defined in a difference scale.

The identification results can be summarized by the following theorem:

Theorem 2.4.1 *Identification of the Pure Direct Effect*
The average pure direct effect $\mathbb{E}\{Y(1, M(0)) - Y(0, M(0))\}$ *is non-parametrically identifiable under each of the sets of Assumptions 2.4.2, 2.4.2, 2.4.3, 2.4.5, 2.4.6, 2.4.7, and 2.4.8 with slight abuse of notation by the following formula*[6]*:*

$$\int_{m,x} (\mathbb{E}[Y|m, A = 1, x] - \mathbb{E}[Y|m, A = 0, x])p(M = m|A = 0, x)p(x)dmdx. \tag{2.2}$$

For Assumption 2.4.4, the exact form of identification depends on the form of the identified $\mathbb{E}[Y(a,m)|x]$ and $p(M(a) = m|x)$.

Corollary 1 *Identification under Pearl's assumptions*
The average pure direct effect is non-parametrically identifiable if Pearl's set of assumptions is satisfied:

$$\begin{aligned}&\mathbb{E}\{Y(1, M(0)) - Y(0, M(0))\}\\ &= \int_{m,x} (\mathbb{E}[Y(1,m)|x] - \mathbb{E}[Y(0,m)|x])p(M(0) = m|x)p(x)dmdx.\end{aligned}$$

[6]For 2.4.7, the mediator is binary, thus the integral over m is replaced with a sum.

It is straight-forward to derive the identification of the controlled direct effect based on the assumptions that identify the pure direct effect.

Theorem 2.4.2 *Identification of the Controlled Direct Effect*[7]
The average controlled direct effect $\mathbb{E}\{Y(1,m) - Y(0,m)\}$ *is non-parametrically identifiable under each of the sets of Assumptions 2.4.2, 2.4.2, 2.4.3, 2.4.5, 2.4.6, 2.4.7, and 2.4.8 with slight abuse of notation:*

$$\int_x (\mathbb{E}[Y|m, A = 1, x] - \mathbb{E}[Y|m, A = 0, x])p(x)dx. \tag{2.3}$$

Assumption 2.4.4 directly assumes identification of $\mathbb{E}[Y(a,m)|x]$, which gives the controlled direct effect.

Up until now, we have restricted the confounder set to be pre-treatment covariates. However, it is entirely possible that the ignorability assumption for the mediator-outcome relationship only holds after conditioning on some treatment-induced variables, such as C in Figure 2.3. For example, in Imai's set of assumptions, instead of assuming $Y(a,m) \perp\!\!\!\perp M|A, X$, it may be more reasonable to assume $Y(a,m) \perp\!\!\!\perp M|A, X, C$, where C is the set of treatment-induced confounders for the mediator-outcome relationship. Similar changes can be made to all the previous sets of assumptions. However, the resulting sets are usually not sufficient to identify the pure direct effect without further assumptions, even if C is observed (Avin et al., 2005[8], VanderWeele and Chiba, 2014).

Without "Cross-world" assumptions or the assumption that precludes the treatment-induced confounding, the controlled direct effect can still be identified under the Single World Intervention Graphs assumptions and the Minimal Counterfactual Model Independence Assumptions.

Assumption 2.4.9 *Single World Intervention Graphs Independence Assumption*

1. $\{Y(a,m), M(a)\} \perp\!\!\!\perp A|X = x$,

[7]The identified expression for the PSDE is the same as that of CDE under the NPSEM-IE assumption.

[8]C forms a "recanting witness".

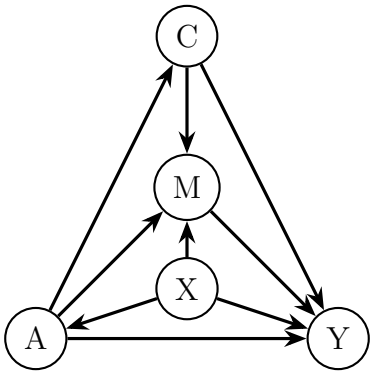


Figure 2.3: Causal DAG with the Treatment-induced Confounding

2. $Y(a,m) \perp\!\!\!\perp M(a)|A=a, X=x, C=c,$

Where X contains pre-treatment variables, and C contains all treatment-induced variables that occur before the mediator, and confounds the mediator-outcome relationship.

Different from Assumption 2.4.2, Assumption 2.4.3, Assumption 2.4.5, and Assumption 2.4.7, Assumption 2.4.9 requires a consistent treatment level for the potential outcome $Y(a,m)$ and the potential mediator $M(a)$ in condition 1, and a consistent treatment level for the treatment A in addition to $Y(a,m)$ and $M(a)$ in condition 2.

Additionally, Assumption 2.4.9 condition 2 imposes independence between $Y(a,m)$ and $M(a)$ given not only the treatment assignment and the baseline covariates, but also the treatment-induced variables that are potential confounders for the relationship between the mediator and the outcome.

An even weaker set of assumption is given as follows:

Assumption 2.4.10 *Minimal Counterfactual Model Independence Assumptions*

$$\{Y(a,m), M(a)\} \perp\!\!\!\perp \mathbb{1}\{A=a\}|X=x,$$
$$Y(a,m) \perp\!\!\!\perp \mathbb{1}\{M(a)=m\}|A=a, X=x, C=c.$$

Minimal Counterfactual Model Independence Assumptions imply that given the past where $A=a$ and $X=x$, the event $M(a)=m$ is independent of the counterfactual outcome

$Y(a, m)$, which is consistent with the past where $M = m$. Assumption 2.4.9 and 2.4.10 are equivalent when the mediator is also binary.

Instead of assuming independence on the joint distribution of the potential outcome and the potential mediators, assumption 2.4.9 and 2.4.10 can be relaxed as follows:

Assumption 2.4.11 *Relaxed SWIG assumption*

1. $Y(a, m) \perp\!\!\!\perp A|X = x$,

2. $Y(a, m) \perp\!\!\!\perp M(a)|A = a, X = x, C = c$,

Assumption 2.4.12 *Relaxed MCM Independence Assumptions*

$$Y(a, m) \perp\!\!\!\perp \mathbb{1}\{A = a\}|X = x,$$
$$Y(a, m) \perp\!\!\!\perp \mathbb{1}\{M(a) = m\}|A = a, X = x, C = c.$$

The relaxed assumptions implies that only two no unmeasured confounding assumptions are needed for the identification of the controlled direct effect:

- no unmeasured confounding between the treatment and the outcome;

- no unmeasured confounding between the mediator and the outcome.

The controlled direct effect is identified under these four sets of assumptions despite the fact that there is treatment-induced confounding and that the pure direct effect is not identified.

Theorem 2.4.3 *Identification of the Controlled Direct effect*
The controlled direct effect is identified under assumption 2.4.9 and 2.4.10 as follows:

$$\mathbb{E}_X\mathbb{E}_{C|A=1,X}\mathbb{E}[Y|M = m, A = 1, X, C] - \mathbb{E}_X\mathbb{E}_{C|A=0,X}\mathbb{E}[Y|M = m, A = 0, X, C].$$

Remark 1 *With Robins No-interaction assumptions (Robins, 2003): For each unit,*

$$Y(1,m) - Y(0,m) = B,$$

where B is a random variable that is independent of the realization of the mediator m, the average natural direct effect under binary treatment $E[Y(1, M(0)) - Y(0, M(0))]$ is non-parametrically identifiable under assumption 2.4.9 or 2.4.10 as follows:

$$\mathbb{E}_X\{\int_c \mathbb{E}[Y|m, A=1, x, c]\mathbb{P}(C=c|A=1,x) - \int_c \mathbb{E}[Y|m, A=0, x, c]\mathbb{P}(C=c|A=0,x)\}.$$

Since $Y(1,m) - Y(0,m)$ is independent of m, the pure direct effect is the same for all levels of m. This no-interaction assumption would not be plausible if the outcome is binary.

2.5 Statistical Methods for Mediation Analysis

In this section, we review some popular statistical models used in mediation analysis and their connections.

We start from the linear structural equation models. Baron and Kenny (1986) proposed an approach to estimating the direct and indirect effect in the point-treatment setting with a system of linear models[9]:

1. $Y_i = \alpha_0 + \alpha_1 A_i + \alpha_2 X_i + \epsilon_{Y1i}$,

2. $M_i = \beta_0 + \beta_1 A_i + \beta_2 X_i + \epsilon_{Mi}$,

3. $Y_i = \gamma_0 + \gamma_1 A_i + \gamma_2 M_i + \gamma_3 X_i + \epsilon_{Y2i}$.

Baron and Kenny suggested that the following conditions need to hold for M to be a mediator:

1. The treatment must affect the outcome in the first equation, i.e. $\alpha_1 \neq 0$;

2. The treatment must affect the mediator in the second equation, i.e. $\beta_1 \neq 0$;

[9]Covariates are not included in the original models.

3. The mediator must affect the outcome in the third equation, i.e. $\gamma_2 \neq 0$.

Under these conditions, the product method use $\beta_1\gamma_2$ as an estimator of the indirect effect, and the difference method use $\alpha_1 - \gamma_1$ as an estimator of the indirect effect. The first condition is not necessary because the overall effect of the treatment on the outcome can be 0 when the direct and indirect effects are in different directions.

There are two major issues with this method:

- the assumptions needed for the estimators to have a causal interpretation are vague: the no unmeasured confounding assumptions are not made explicit.

- the estimators rely on the functional form of the statistical models: the estimators are no longer valid when the outcome is nonlinear or when interaction between the mediator and the treatment is present.

We then move on to direct and indirect effects defined using the counterfactual framework. From the identification results given in the previous section, one approach is to estimate the components in the identified expressions using regression:

- build a regression model for $\mathbb{E}[Y|M, A, X]$;

- build a regression model for $\mathbb{E}[M|A, X]$.

The controlled direct effect and the pure direct effect can then be estimated using the model coefficients (VanderWeele, 2015). The estimators from the LSEM approach can be viewed as special cases. A similar approach can be applied to alternative representations of the identified effects.

In general, we can estimate the (controlled and pure) direct and indirect effects if we specify two out of the three models (VanderWeele, 2015):

- an outcome model conditioning on the treatment, the mediator, and the covariates.

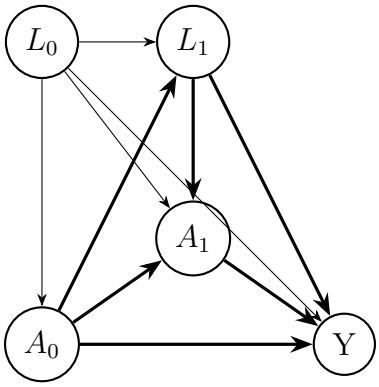


Figure 2.4: Causal DAG for the Point-treatment case with a Mediator

- a model for the mediator conditioning on the treatment and the covariates.
- a propensity score model (treatment model) conditioning on the covariates.

In fact, the triply-robust estimator (Tchetgen Tchetgen and Shpitser, 2012) for the pure direct effect is consistent and asymptotically normal when two out of the three models are correctly specified, and it is locally efficient when all three models are correctly specified.

The statistical methods developed for the estimation of direct and indirect effects are based on the identified expressions, so they are not unique to causal mediation analysis or causal inference in general.

2.6 Discussion

2.6.1 Mediation in Longitudinal Settings

A complex causal relationship in the longitudinal setting is the presence of "treatment-confounder feedback", illustrated in Figure 2.4. In this example, L_0 is a baseline confounder (for all relationships in the system) and, therefore, can be adjusted for in the model. On the other hand, L_1 is a confounder for the association between the current treatment A_1 and the outcome Y. It is also a mediator of the past treatment A_0 on the outcome. In this case, L_1 cannot simply be adjusted for or ignored in the model. G-methods, such as the marginal

structural models, should be used (VanderWeele, 2009; Lin et al. 2017; VanderWeele and Tchetgen Tchetgen, 2017.).

2.6.2 Extended DAG

The term $E[Y(1, M(0))]$ in the pure direct effect is a composite (nested) counterfactual that depends on the joint distribution of $M(0)$ and $Y(1, m)$, where $Y(1, m)$ fixes the treatment A at level 1, and $M(0)$ fixes the treatment at level 0. As mentioned in the previous section, there is no randomized experiment with interventions on the X, A, M or Y can produce a contrast between treatment regimen that corresponds to the pure direct effect. For this reason, the pure direct effect is said to be non-refutable or non-manipulable. This unfavorable property of the pure direct effect makes it difficult to interpret.

Robins and Richardson (2010) provide a new perspective that explains the pure direct effect through a hypothetical randomized experiment. By using the extended DAG, they reconcile the apparent contradiction of the pure direct effect being both explainable through randomized experiments (as given by Pearl) and being non-manipulable .

Take the DAG in Figure 2.2 as an example (the baseline covariates X are ignored for simplicity). Suppose that scientific knowledge indicates that the treatment A affects the mediator M through A', the outcome Y through A'', and that $A'(a) = A''(a) = a$ with probability 1, then $\mathbb{E}[Y(1, M(0))]$ is $\mathbb{E}[Y(A' = 0, A'' = 1)]$. The contrast between $\mathbb{E}[Y(1, M(0))]$ and $\mathbb{E}[Y(0)]$ corresponds to the contrast between the treatment regime in a randomized experiment via interventions Pearl's identification from the previous section, using only the observed data from Figure 2.2.

The extended DAG provides a connection between the pure direct effect and randomized experiments, which not only helps interpretation of the pure direct effect, but also provides insights on the assumptions needed for its identification.

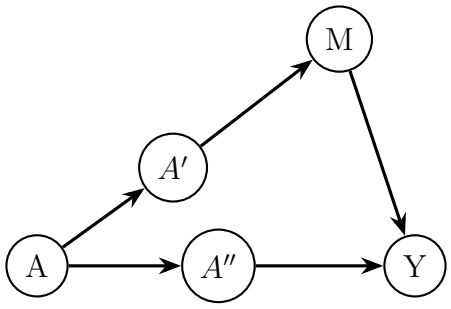


Figure 2.5: Causal DAG of the setting

Chapter 3

CAUSAL MEDIATION ANALYSIS WITH TREATMENT-INDUCED CONFOUNDING

3.1 Introduction

When a treatment has an aggregated effect on an outcome, its effect mechanism is often of interest when variables on causal pathways are observed. The study of the treatment effect mechanism involves the estimation of direct and indirect effects. An effect is called direct when it does not act through some intermediate variables, known as mediators. Conversely, the effect that acts through the mediators is called the indirect effect. Intuitively, to evaluate a direct effect, the mediators need to be somehow fixed. Depending on the scientific question of interest, a variety of direct effects can be defined through different ways of fixing the mediators. The natural (pure) effects are most relevant in studying treatment effect mechanisms. The natural direct effect compares potential outcomes under treatment and control with mediators fixed to the value they would have taken had there not been any treatment. The natural indirect effect is thereby defined by subtracting the natural direct effect from the total treatment effect.

The definition of the natural direct effect is formalized under the potential outcome framework (Robins and Greenland, 1992; Pearl, 2001). A considerable number of methods have been developed for the identification and inference of the natural direct effect (Pearl, 2001; Pearl, 2009; Petersen et al., 2006; Imai et al., 2010; Hafeman and VanderWeele, 2011; Tchetgen Tchetgen and Shpitser, 2012). Common identification assumptions, such as sequential ignorability assumptions (Imai et al., 2010), usually prohibit treatment-induced confounding between the mediator and the outcome.

In practice, the assumption that rules out the treatment-induced confounding can often

be violated, particularly when the mediators come much later than the treatment. In this case, some immediate prognostic factors affected by the treatment can be related to both the mediator and the outcome (Robins, 1999). One example that was given by Vansteelandt & VanderWeele (2012) considered the effect of adequate prenatal care on preterm birth that mediates through preeclampsia. On one hand, smoking status during pregnancy confounds the relationship between preeclampsia and preterm birth because it reduces the risk of preeclampsia while increasing the likelihood of preterm birth. On the other hand, adequate prenatal care may decrease or eliminate smoking. Therefore, smoking status during pregnancy is a potential treatment-induced confounder between the mediator preeclampsia and the outcome preterm birth.

There has been limited methodological development addressing such limitations. Robins (2003) provides a no-interaction assumption between treatment and mediator at the individual level to identify the natural direct effect, but Petersen et al. (2006) suggest that this assumption is unlikely to hold in practice. Robins and Richardson (2010) and Tchetgen Tchetgen and VanderWeele (2014) each provide identification assumptions in the framework of nonparametric structural equation models with independent errors, which as pointed out in Robins and Richardson (2010), imposes many independence assumptions between counterfactuals from different worlds. These "Cross-world" assumptions are strong in the sense that they are experimentally untestable. Moreover, the natural direct effect is not identified in the presence of treatment-induced confounding without additional "Cross-world" assumptions (beyond even those implied by NPSEM) in the framework. When these experimentally untestable assumptions are violated, the identified expression is no longer the natural direct effect, and may not even have a causal interpretation. When point identification is impossible for a set of assumptions, bounds can often be developed. Bounds on the natural direct effect for a binary mediator are given by Robins and Richardson (2010) under Single World Intervention Graphs (SWIG) independence assumptions, which are extended by Tchetgen and Phiri (2014) in the presence of treatment-induced confounding, and are further extended to the polytomous mediator by Miles et al. (2015). Vansteelandt and VanderWeele

(2012) consider a slightly different estimand, the natural direct effect on the treated, whose identification relies on the knowledge of a selection-bias function.

Estimands different from the natural direct effect are also considered to quantify certain direct and indirect effects in the presence of treatment-induced confounding. VanderWeele et al. (2014) summarize three such approaches to decompose the effect of a treatment when there exists treatment-induced confounding: joint effect of mediators and other treatment-induced confounders, path-specific effects, and interventional effects. Avin et al. (2005) and Shpitser (2013) provide identification conditions for path-specific effects. Miles et al. (2017) provide semiparametric inference of a path-specific effect that goes through a mediator without going through its treatment-induced confounders. The interventional direct effect is an analog of the natural direct effect that replaces the potential mediator with a random draw, which is independent of the potential outcome, from the distribution of the mediator among the non-treated. VanderWeele and Tchetgen Tchetgen (2017) define interventional effects for mediation analysis with time-varying exposures and mediators. The estimand is also used in mediation analysis with multiple mediators (VanderWeele and Vansteelandt, 2014; Daniel et al., 2015).

The estimation of the natural direct effect in the presence of treatment-induced confounding is fundamentally different from the usual settings, where all confounders are pre-treatment. Although weighting methods for handling pre-treatment confounders are well studied, none of them can be directly applied with treatment-induced confounders, since they are in the usual pathway between the exposure and outcome of interest. Moreover, if all treatment-induced confounders are observed, sequential ignorability assumptions are not sufficient to identify the natural direct effect (Avin et al., 2005; VanderWeele and Vansteelandt, 2009). Therefore, both the identification assumptions and estimation in the presence of treatment-induced confounding will be substantially different from the usual case. In addition to proposing identification assumptions, we found that, unlike usual expressions, the efficient influence function contains conditional densities that are not variation independent. We consider a reparameterization based on copulas to address the problem of model incom-

patibility. The corresponding estimator is quadruply robust, that is, consistent under four types of misspecification of the nuisance models.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed identification assumptions and the expression of the identified natural direct effect. We explain the connection between our identification results and that of the interventional direct effect. In Section 3, we propose four moment-type estimators and a quadruply robust estimator. In the process, we derive the efficient influence function (hence the semiparametric efficiency bound) of the identified natural direct effect, propose a variation independent parameterization, and prove the quadruple robustness of the estimator. In Section 4, we use numerical simulations to demonstrate the theoretical results derived in Section 3 for both continuous and binary mediators. In Section 5, we apply our method to the 2017 Natality data to estimate the effect of prenatal care on preterm birth mediated by preeclampsia with smoking status during pregnancy being a potential treatment-induced confounder. In Section 6, we discuss some concluding remarks, including the estimation of the natural indirect effect, and sensitivity analysis for identification assumption violations.

3.2 Assumptions and Identification

We denote the treatment as A, the outcome as Y, the mediator as M, the set of treatment-induced confounders as C, and the set of pre-treatment or baseline covariates as X. All variables may be multivariate. Figure 3.1 demonstrates the causal diagram. The set of covariates X is omitted for simplicity because it has arrows to all other variables in the causal diagram.

When there are well-defined interventions for A, C, and M, the potential outcome Y_{acm} is the value the outcome would have taken had the treatment been a, the treatment-induced confounder been c, and the mediator been m. We assume the composition (also called recursive substitution) holds such that $Y_a = Y_{aC_aM_a}$, $Y_{am} = Y_{aC_am}$, and $M_a = M_{aC_a}$. Other conditions are needed as preliminaries of identification. The consistency assumption that implies $C_a = C$ and $M_a = M$ when $A = a$; $Y_{acm} = Y$ when $A = a, C = c, M = m$,

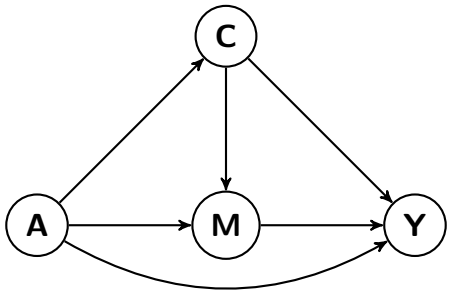


Figure 3.1: The Causal Diagram with Treatment-induced Confounding

and the positivity assumption that implies $f(m|A,C,X) > 0, f(c|A,X) > 0, f(a|X) > 0$ with probability 1 for all m, c, a. Consistency assumption is axiomatic, for the definition of potential outcomes relies on it. The positivity condition is techical, and could be violated in practice.

When the treatment is binary, the average natural direct effect on a difference scale is defined as $E[Y_{1M_0} - Y_{0M_0}]$. It depicts the expected effect of the treatment when the mediator is fixed at the value it would have taken had there not been any treatment. A set of assumptions sufficient to identify the natural direct effect in the presence of treatment-induced confounders are given as follows:

Assumption 3.2.1 $\{Y_{am}, C_a, M_a\} \perp\!\!\!\perp A|X$.

Assumption 3.2.2 $Y_{am} \perp\!\!\!\perp M_a|A = a, C_a = c, X$

Assumption 3.2.3 $E[Y_{1C_1m} - Y_{0C_1m}|M_0 = m, X] = E[Y_{1C_1m} - Y_{0C_1m}|X]$

Assumption 3.2.4 $E[Y_{0C_1m} - Y_{0C_0m}|M_0 = m, X] = E[Y_{0C_1m} - Y_{0C_0m}|X]$

Assumptions 3.2.1 and 3.2.2 are ignorability assumptions that are implied by the assumptions of no unmeasured confounding between the treatment and post-treatment variables, and between the mediator and the outcome respectively. Both of them are "single-world"

assumptions (Richardson and Robins, 2013) and are weaker than usual identification assumptions for the natural direct effect, such as the sequential ignorability assumptions with "cross-world" independence given by Imai et al. (2010). Assumptions 3.2.3 and 3.2.4 imply that there is no additional heterogeneity in a direct effect of A, or in a pure indirect effect of A that goes through C across levels of M_0. These two assumptions are scale-specific.

Theorem 3.2.1 *Under assumptions 3.2.1–3.2.4, the natural direct effect* $E[Y_{1M_0} - Y_{0M_0}]$ *is identified as follows:*

$$\Delta \equiv E_X(E_{M=m|A=0,X}\{E_{C|A=1,X}E[Y|A=1,C,M=m,X]- \\ E_{C|A=0,X}E[Y|A=0,C,M=m,X]\}). \tag{3.1}$$

Our identification result gives the same empirical expression as the interventional effect (VanderWeele et al., 2014; VanderWeele and Tchetgen Tchetgen, 2017). The interventional direct effect is defined by replacing the potential mediator with a random draw from the distribution of the potential mediator M_0 that is independent of the potential outcomes, and requires only Assumptions 3.2.1 and 3.2.2 for identification. In fact, when M_0 is being replaced by a random draw $M_0^\star$, Assumptions 3.2.3 and 3.2.4 are satisfied because $M_0^\star$ is independent of $\{Y_{acm}, C_{a'}\}$.

When there is no treatment-induced confounder, C becomes part of X, assumption 3.2.4 becomes redundant, and assumption 3.2.3 reduces to that in Petersen et al. (2006). The identification of the natural direct effect becomes

$$E_X[E_{M=m|A=0,X}\{E(Y|A=1,M=m,X) - E(Y|A=0,M=m,X)\}],$$

which is the same empirical expression as the natural direct effect identified by the sequential ignorability assumptions (Pearl, 2001; Petersen et al., 2006; Imai et al., 2010).

3.3 Semiparametric Inference

3.3.1 Moment-type Estimators

Denote the identified expression of the natural direct effect in Theorem 3.2.1 as Δ, which is the estimand of interest for the remaining sections. The observed independent samples are $(X_i, A_i, C_i, M_i, Y_i), i = 1, \ldots, n$. With slight abuse of notation, the density (mass) functions are denoted by f. The estimand Δ can be represented in four alternative ways, each leading to a possible estimator.

Theorem 3.3.1 $\Delta = \Delta_1 = \Delta_2 = \Delta_3 = \Delta_4$, *where*

$$\begin{aligned}
\Delta_1 &= E_{X,A,C,M,Y}\left\{\frac{2A-1}{f(A \mid X)}\frac{f(M \mid A=0,X)}{f(M \mid A,C,X)}Y\right\},\\
\Delta_2 &= E_{X,A,C}\left\{\frac{2A-1}{f(A \mid X)}\eta_{C,X}(A)\right\},\\
\Delta_3 &= E_{X,A,M}\left\{\frac{1-A}{f(A=0 \mid X)}(\gamma_{M,X}(1)-\gamma_{M,X}(0))\right\},\\
\Delta_4 &= E_X\Big\{\tau_X(1)-\tau_X(0)\Big\},
\end{aligned}$$

where

$$\begin{aligned}
\eta_{C,X}(a) &= \int E(Y \mid A=a,m,C,X)f(m \mid A=0,X)\,dm,\\
\tau_X(a) &= \int E(Y \mid A=a,m,c,X)f(m \mid A=0,X)f(c \mid A=a,X)\,dm\,dc,\\
\gamma_{M,X}(a) &= \int E(Y \mid A=a,M,c,X)f(c \mid A=a,X)\,dc.
\end{aligned}$$

Based on different representations of Δ, we consider four estimators $\hat{\Delta}_1$, $\hat{\Delta}_2$, $\hat{\Delta}_3$, and $\hat{\Delta}_4$ that replace conditional expectations or densities in Δ_1, Δ_2, Δ_3, and Δ_4 with their estimates and the outer expectation by the empirical average. When Y, M, and C are discrete and low dimensional, $\hat{f}(Y \mid A, M, C, X)$, $\hat{f}(M \mid A, C, X)$, $\hat{f}(C \mid A, X)$, and $\hat{f}(A \mid X)$ can be empirical probability mass functions, and $\hat{E}(Y|A, M, C, X)$ is the expectation under $\hat{f}(Y \mid A, M, C, X)$. The integrals in the estimators become finite sums, and the four estimators

are nonparametric. In practice, however, M and C are likely to be high-dimensional and continuous, thus we use parametric models for the purpose of dimension reduction. The four estimators are consistent when nuisance parameters for each part of them are consistently estimated. In particular, with the rest of the models unrestricted, $\hat{\Delta}_1$ is consistent and asymptotically normal when $\hat{f}(A \mid X)$, $\hat{f}(M \mid A = 0, X)$ and $\hat{f}(M \mid A, C, X)$ are correctly specified, $\hat{\Delta}_2$ is consistent and asymptotically normal when $\hat{f}(A \mid X)$, $E(Y \mid A, M, C, X)$, and $\hat{f}(M \mid A = 0, X)$ are correctly specified, $\hat{\Delta}_3$ is consistent and asymptotically normal when $\hat{f}(A \mid X)$, $E(Y \mid A, M, C, X)$, and $\hat{f}(C \mid A, X)$ are correctly specified, and $\hat{\Delta}_4$ is consistent and asymptotically normal when $E(Y \mid A, M, C, X)$, $\hat{f}(M \mid A = 0, X)$, and $\hat{f}(C \mid A, X)$ are correctly specified.

3.3.2 Efficient Influence Function and the Quadruply Robust Estimator

Next, we derive the efficient influence function of Δ under a nonparametric model $\mathcal{M}_{non}$, which does not impose constraints on the observed data.

Theorem 3.3.2 *The efficient influence function of Δ in $\mathcal{M}_{non}$ is:*

$$\begin{aligned} S_{\Delta}^{eff} =& \frac{2A-1}{f(A \mid X)} \frac{f(M \mid A=0, X)}{f(M \mid A, C, X)} (Y - E[Y \mid A, M, C, X]) + \\ & \frac{2A-1}{f(A \mid X)} \eta_{C,X}(A) - \frac{2A-1}{f(A \mid X)} \tau_X(A) + \frac{1-A}{f(A \mid X)} \{\gamma_{M,X}(1) - \gamma_{M,X}(0)\} \\ & + \Big(1 - \frac{1-A}{f(A \mid X)}\Big) \{\tau_X(1) - \tau_X(0)\} - \Delta. \end{aligned}$$

Hence, the semiparametric efficiency bound for the estimation of Δ in $\mathcal{M}_{non}$ is $E[S_{\Delta}^{eff} S_{\Delta}^{eff^T}]$, and the asymptotic variance of any regular asymptotic linear estimator of Δ in $\mathcal{M}_{non}$ must be greater than or equal to the bound.

The efficient influence function is a function of $f(A \mid X)$, $f(C \mid A, X)$, $f(M \mid A, C, X)$ and $E(Y \mid A, M, C, X)$. While we may posit parametric working models for these functions, a complication arises because $f(C \mid A, X)$, $f(M \mid A, X)$, and $f(M \mid A, C, X)$ are not variation independent, and therefore model incompatibility may occur. Richardson et al. (2017) point

out that the multiple robustness property is relevant only when model incompatibility can be avoided.

We consider reparameterizing the joint distribution $f(M, C \mid A, X)$ into three parts: the two margins conditioned on A and X: $f(M \mid A, X), f(C \mid A, X)$ and their dependence structure modeled using a copula condition on A and X.

A copula is a multivariate cumulative distribution function with uniformly distributed margins on $[0, 1]$. A more detailed discussion on copulas is given by Joe (1997), Nelsen (2007), and Jaworski et al. (2010). For notational simplicity, we consider univariate M and C, and a bivariate conditional copula with support contained in $[0, 1]^2$:

$$\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X) = F(M = m, C = c \mid A, X).$$

Sklar's theorem (Sklar, 1959) allows separate modeling of these three parts. In other words, the joint distribution $F(M, C \mid A, X)$ is uniquely determined by $f(M \mid A, X), f(C \mid A, X)$, and $\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ that can be modeled independently. For continuous margins, when marginal and joint densities exist, Sklar's theorem implies that

$$f(M, C \mid A, X) = \boldsymbol{c}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X) f(M \mid A, X) f(C \mid A, X),$$

where $\boldsymbol{c}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ denotes the copula density as computed by taking the derivative of the copula with respect to $F_{M|A,X}(m)$ and $F_{C|A,X}(c)$. For discrete margins, the probability mass function $f(M = m, C = c \mid A, X)$ is computed using:

$$\begin{aligned}&\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X) - \mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c-1) \mid A, X) \\ &- \mathcal{C}(F_{M|A,X}(m-1), F_{C|A,X}(c) \mid A, X) + \mathcal{C}(F_{M|A,X}(m-1), F_{C|A,X}(c-1) \mid A, X).\end{aligned}$$

For example,

- when M and C are continuous, we may use linear models for the margins, and a bivariate Gaussian copula with correlation $\theta(A, X)$ for the dependence structure:

$$\mathcal{C}_{\theta(A,X)}(F_{M|A,X}(m), F_{C|A,X}(c)) = \boldsymbol{\Phi}_{\theta(A,X)}(\Phi^{-1}(F_{M|A,X}(m)), \Phi^{-1}(F_{C|A,X}(c))),$$

where Φ^{-1} is the inverse cumulative distribution function for a standard normal, and $\mathbf{\Phi}_{\theta(A,X)}$ is the joint cumulative distribution function of a bivariate normal distribution with mean zero and covariance matrix

$$\begin{bmatrix} 1 & \theta(A,X) \\ \theta(A,X) & 1 \end{bmatrix}.$$

- when M and C are binary, we may use logistic models for the margins, and a bivariate Plackett copula $\mathcal{C}_{\theta(A,X)}(F_{M|A,X}(m), F_{C|A,X}(c))$ with an odds ratio $\theta(A,X)$ for the dependence structure:

$$\begin{cases} \dfrac{\{1+(\theta(A,X)-1)(F_{M|A,X}(m)+F_{C|A,X}(c))\}-\mathcal{S}^{1/2}}{2(\theta(A,X)-1)}, \text{ when } \theta \neq 1 \\ F_{M|A,X}(m)F_{C|A,X}(c), \text{ when } \theta = 1. \end{cases}$$

where

$$\mathcal{S} = [\{1+(\theta(A,X)-1)(F_{M|A,X}(m)+F_{C|A,X}(c)\}^2 - 4F_{M|A,X}(m)F_{C|A,X}(c)\theta(A,X)(\theta(A,X)-1)].$$

In multivariate cases, the vine pair copula construction (Panagiotelis et al., 2012) can be used to construct the joint distribution.

Let p_n be the empirical measure. With the variation independent parameterization, we construct a locally efficient estimator based on the following estimating equation:

$$p_n(\hat{S}^{\text{eff}}_{\Delta}(\hat{\Delta}_{quad})) = 0.$$

$\hat{S}^{\text{eff}}_{\Delta}$ is evaluated where all components of the influence function are replaced by their parametric working model: $f(a \mid X)$ is replaced by $f^{\text{par}}(a \mid X)$, $f(c \mid A,X)$ is replaced by $f^{\text{par}}(c \mid A,X)$, $f(m \mid A,X)$ is replaced by $f^{\text{par}}(m \mid A,X)$, and $E(Y \mid A,M,C,X)$ is replaced by $E^{\text{par}}(Y \mid A,M,C,X)$. In particular, $f(m,c \mid A,X)$ is replaced by $f^{\text{par}}(m,c \mid A,X)$, which is modeled by the two marginal distributions $f^{\text{par}}(m \mid A,X)$, $f^{\text{par}}(c \mid A,X)$, and the copula

$\mathcal{C}^{\text{par}}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$. Therefore, $\hat{\Delta}_{quad}$ takes the following form:

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{2A_i-1}{\hat{f}^{\text{par}}(A_i \mid X_i)}\frac{\hat{f}^{\text{par}}(M_i \mid A_i=0, X_i)\hat{f}^{\text{par}}(C_i \mid A_i, X_i)}{\hat{f}^{\text{par}}(M_i, C_i \mid A_i, X_i)}\{Y_i - \hat{E}^{\text{par}}(Y_i \mid A_i, M_i, C_i, X_i)\}+$$
$$\frac{2A_i-1}{\hat{f}^{\text{par}}(A_i \mid X_i)}\hat{\eta}^{\text{par}}_{C_i,X_i}(A_i) - \frac{2A_i-1}{\hat{f}^{\text{par}}(A_i \mid X_i)}\hat{\tau}^{\text{par}}_{X_i}(A_i) + \frac{1-A_i}{\hat{f}^{\text{par}}(A_i=0 \mid X_i)}\{\hat{\gamma}^{\text{par}}_{M_i,X_i}(1) - \hat{\gamma}^{\text{par}}_{M_i,X_i}(0)\}+$$
$$(1 - \frac{1-A_i}{\hat{f}^{\text{par}}(A_i=0 \mid X_i)})\{\hat{\tau}^{\text{par}}_{X_i}(1) - \hat{\tau}^{\text{par}}_{X_i}(0)\}\Big].$$

This estimator is quadruply robust in the sense that only one out of four sets of models needs to be correctly specified for it to be consistent and asymptotically normal as given in Theorem 3.3.3.

Theorem 3.3.3 *The estimator $\hat{\Delta}_{quad}$ is consistent and asymptotically normal under some mild regularity conditions discussed in the supplementary material if one of the following four conditions holds:*

1. $\mathcal{M}_1$: $f^{par}(A \mid X)$, $f^{par}(C \mid A, X)$, $f^{par}(M \mid A, X)$, $\mathcal{C}^{par}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ *are correctly specified.*

2. $\mathcal{M}_2 : f^{par}(A \mid X)$, $f^{par}(M \mid A, X)$, $E^{par}(Y \mid A, M, C, X)$ *are correctly specified.*

3. $\mathcal{M}_3 : f^{par}(A \mid X)$, $f^{par}(C \mid A, X)$, $E^{par}(Y \mid A, M, C, X)$ *are correctly specified.*

4. $\mathcal{M}_4$: $f^{par}(M \mid A, X)$, $f^{par}(C \mid A, X)$, $E^{par}(Y \mid A, M, C, X)$ *are correctly specified.*

It is locally semiparametric efficient in the sense that it achieves the semiparametric efficiency bound at the intersection of the submodels where all four conditions hold, that is, at $\mathcal{M}_{intersection} = \mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3 \cap \mathcal{M}_4$.

Notice that the estimators proposed in section 3.1 are such that $\hat{\Delta}_1$, whose estimation can be conducted using the copula parameterization, is only consistent under $\mathcal{M}_1$, $\hat{\Delta}_2$ is only consistent under $\mathcal{M}_2$, $\hat{\Delta}_3$ is only consistent under $\mathcal{M}_3$, and $\hat{\Delta}_4$ is only consistent under $\mathcal{M}_4$. In contrast, the quadruply robust estimator $\hat{\Delta}_{quad}$ remains consistent under four types of misspecification, which offers more modeling flexibility. In other words, $\hat{\Delta}_{quad}$ is consistent and asymptotically normal at the intersection submodel.

3.4 Simulation study

We use numerical simulations to demonstrate the theoretical results derived in the previous section. We compare the finite sample performance of the moment-based estimators given in section 3.1 to the proposed quadruply robust estimator. We generate 1000 samples, each with 1500 independent observations, for both continuous and binary treatment-induced confounder and mediator. We consider the moment estimators $\hat{\Delta}_1$, $\hat{\Delta}_2$, $\hat{\Delta}_3$, $\hat{\Delta}_4$ and the quadruply robust estimator $\hat{\Delta}_{quad}$. Let *expit* denote the function $expit(x) = exp(x)/(1 + exp(x))$. The data are generated as follows:

Continuous C and M:

$$X \sim N(0,1); P(A = 1 \mid X) = expit(-0.4 + 0.6X);$$

$\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ is a Gaussian Copula with correlation 0.2,

where $F_{M|A,X}(m) = \Phi(\frac{m - \mu_m}{\sigma_m}), \mu_m = 3 + 2A + 4X, \sigma_m = 5,$

$$F_{C|A,X}(c) = \Phi(\frac{c - \mu_c}{\sigma_c}), \mu_c = 1 + 2A + 2X, \sigma_c = 4,$$

$$Y \sim 1 + 2A + 2M + 3C + 5X + 4AC + 2AM + N(0, 4^2).$$

Binary C and M:

$$X \sim N(0,1); P(A = 1 \mid X) = expit(-0.2 + 0.3X);$$

$\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ is a Plackett Copula with Odds-Ratio $exp(1 - 2A + 3X)$;

where $F_{M|A,X}(m) = p_m^m(1 - p_m)^{1-m}, p_m = expit(-0.3 - 0.2A + 0.5X),$

$$F_{C|A,X}(c) = p_c^c(1 - p_c)^{1-c}, p_c = expit(-0.2 - 0.1A + 0.3X),$$

$$Y \sim 1 + 3A + 6M + 3C + 6X + 4AC + 2AM + N(0, 4^2).$$

We compare the five estimators under a series of model misspecifications by replacing the baseline covariates X with an independent normally distributed continuous variable X_2 with mean 0 and variance 1. Table 3.1 shows that the simulation results are consistent with the theoretical results derived in the previous sections: when the entire likelihood is correctly

specified, all five estimators are consistent; when the conditional expectation of Y is mis-specified, only Δ_1 and Δ_{quad} are consistent; when the parametric model for $f(C \mid A, X)$ is mis-specified, only Δ_2 and Δ_{quad} are consistent; when the parametric model for $f(M \mid A, X)$ is mis-specified, only Δ_3 and Δ_{quad} are consistent; when the propensity score $f(A = 1 \mid X)$ is mis-specified, only Δ_4 and Δ_{quad} are consistent. The loss in efficiency for the quadruply robust estimator is relatively small compared to other estimators in all cases. Since Δ_1 consists of a density ratio, it is more variable when the mediator M is continuous, which makes it less preferred even when $\mathcal{M}_1$ is correct. We only present one scenario here, but we ran simulations under different settings and they all gave similar results.

Table 3.1: Simulation Results: 100×Bias (100×Standard Error)

	Continuous C, M				
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$
All correct	-5 (191)	-3 (151)	-4 (141)	-3 (139)	-4 (144)
$\mathcal{M}_1$ is correct	10 (180)	87 (141)	89 (137)	88 (135)	2 (138)
$\mathcal{M}_2$ is correct	77 (176)	3 (149)	599 (141)	600 (140)	4 (144)
$\mathcal{M}_3$ is correct	-1390 (369)	-189 (135)	-6 (134)	-187 (133)	-6 (135)
$\mathcal{M}_4$ is correct	1589 (220)	1587 (187)	-359 (143)	4 (134)	4 (134)
	Binary C, M				
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$
All correct	1 (154)	-1 (26)	-1 (26)	-1 (26)	1 (49)
$\mathcal{M}_1$ is correct	-2 (115)	176 (40)	175 (40)	176 (40)	3 (115)
$\mathcal{M}_2$ is correct	44 (27)	1 (26)	12 (25)	12 (26)	1(26)
$\mathcal{M}_3$ is correct	22 (30)	-4 (24)	-1 (24)	-4 (24)	-1 (25)
$\mathcal{M}_4$ is correct	245 (44)	210 (46)	-9 (27)	-2 (27)	-2 (27)

3.5 Data Example

We use the 2017 Natality data (https://wonder.cdc.gov/natality.html) for births occurring within the United States to U.S. residents to illustrate our method. We focus our analysis on the subset of participants that are AIAN (American Indians or Alaskan Native). Subjects with missing data ($< 9.5\%$ of the sample) are excluded. The total number of observations is 27,138.

As pointed out in the introduction, in the question of how much the effect of prenatal care on preterm birth is mediated by preeclampsia, smoking status during pregnancy is a potential treatment-induced confounder. We are interested in estimating the direct effect of prenatal care (A) on preterm birth (Y) not through preeclampsia (M), in the presence of smoking status during pregnancy (C) that is affected by prenatal care.

The adequacy of prenatal care is determined by the Adequacy of Prenatal Care Utilization Index (Kotelchuck, 1994), which depends on the month prenatal care began, the number of prenatal visits, and the gestational age at the time of delivery. In the AIAN sample, the level of prenatal care is either inadequate or intermediate. Preterm birth is defined using the Obstetric Estimate (OE) (Martin et al., 2015) of the gestational age. The baseline covariates (X) that are potential confounders include maternal demographics: age, education level, and marital status. Assumption 3.2.3 implies that the direct effect of prenatal care on preterm birth (that goes through neither smoking nor preelampsia) is the same among those who would get preelampsia without adequate prenatal care, and those who would not. Similarly, Assumption 3.2.4 implies that the mediated effect of prenatal care through smoking is the same among those who would get preelampsia without adequate prenatal care, and those who would not. If these two assumptions are violated, meaning that the potential preelampsia status without adequate prenatal care modifies either the direct effect of prenatal care or its mediated effect through smoking, then the estimated effects can be interpreted as interventional effects, as explained in section 3.2.

Since both the smoking status and the preeclampsia status are binary, we use the Plackett

copula with a cross-ratio (odds ratio) specified using a log link. Logistic regression models are used for the binary treatment and outcome, as well as the distributions of C and M given A and X. The parameters of the copula are estimated by the maximum likelihood method. The bootstrap confidence intervals are computed for the purpose of inference.

The estimated direct effect of better prenatal care (intermediate care versus inadequate care) not through preeclampsia decreases the risk of preterm birth by 2.5% (1.6%, 3.4%), leaving a tiny indirect effect through preeclampsia that increases the risk of preterm birth by 0.15% (0.07%, 0.23%). The moment-type estimators give similar results (Table 3.2). This is consistent with VanderWeele et al. (2014) who studied this problem on a different population.

Table 3.2: Estimation of Direct Effect of Better Prenatal Care on Preterm Birth

Estimator	Direct Effect Estimate	Bootstrap 95% CI
$\hat{\Delta}_1$	0.026	(0.016, 0.036)
$\hat{\Delta}_2$	0.028	(0.018, 0.037)
$\hat{\Delta}_3$	0.027	(0.018, 0.036)
$\hat{\Delta}_4$	0.027	(0.018, 0.036)
$\hat{\Delta}_{quad}$	0.025	(0.016, 0.034)

3.6 Discussion

In this paper, we identify the natural direct effect in the presence of treatment-induced confounding, and derive semiparametric bounds and propose a quadruply robust estimator. Our method can be applied to continuous, categorical, and multivariate outcomes, and to mediators and treatment-induced confounders.

One favorable feature of the natural direct effect is that the average treatment effect, defined as $E[Y_1 - Y_0]$, can be decomposed into the sum of the average natural direct effect and the natural indirect effect: $E[Y_{1M_0} - Y_{1M_0}]$. While the natural indirect effect is not

the focus of this paper, similar results can be applied to it since the average treatment effect is identified under assumption 3.2.1. The natural indirect effect is then identified as the difference between the identified average treatment effect and the natural direct effect identified in Theorem 3.2.1. The semiparametric estimation theory can also be extended for the natural indirect effect. Specifically, we can construct a quadruply robust estimator for the natural indirect effect by the difference between the doubly robust estimator (augmented inverse propensity weighted estimator) for the average treatment effect (Robins et al., 1994; Robins, 2000, Tsiatis, 2007), and our proposed quadruply robust estimator. The augmented inverse propensity weighted estimator is consistent if either the model for the propensity score or the regression model for the mean outcome is correct. Notice that for each of $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$, the condition for the average treatment effect estimator being consistent is satisfied. Therefore the quadruple robustness extends to the natural indirect effect. We should note, however, that the identified natural indirect effect is different from the interventional indirect effect. This is consistent with the fact that the interventional direct effect and indirect effect do not sum up to be the average treatment effect (Vansteelandt and Daniel, 2017).

Although we studied the natural direct and indirect effect with it defined as the difference in expectation, it can also be defined on other scales, such as a ratio scale since $E[Y_{0M_0}] = E[Y_0]$ is identified. The semiparametric estimation theory can be applied, and the asymptotic variance can be derived using the delta method. However, since the identification assumptions are given on the difference scale, extra care is needed when interpreting the natural direct and indirect effect defined on other scales.

Sensitivity analysis can be used to assess how vulnerable the estimator is under assumption violations. Inspired by Vansteelandt and VanderWeele (2012) and VanderWeele and Chiba (2014), we propose the following two sensitivity functions:

$$q_m(M_0, X) = E[Y_{1C_1m} - Y_{0C_1m} \mid M_0 = m, X] - E[Y_{1C_1m} - Y_{0C_1m} \mid M_0, X],$$
$$l_m(M_0, X) = E[Y_{0C_1m} - Y_{0C_0m} \mid M_0 = m, X] - E[Y_{1C_1m} - Y_{0C_1m} \mid M_0, X].$$

The former captures the heterogeneity in the direct effect of the treatment across differ-

ent mediator subgroups within the control group conditional on X, and the latter captures the heterogeneity in the indirect effect of the treatment through the treatment-induced confounder across different mediator subgroups within the control group conditional on X. With the knowledge of the sensitivity functions, the natural direct effect can be identified as:

$$\Delta + \int (E[q_m(M, X) + l_m(M, X) \mid A = 0, X]) f(M = m \mid A = 0, X) dm.$$

As Robins and Richardson (2010) point out, different assumptions give different identifying expressions. It is sometimes not clear how scientists can choose an identification assumption when it lacks scientific justification, because they are not refutable even by experiments. Our identified expression has the advantage that even when the no additional effect heterogeneity assumptions are inappropriate, it can still be interpreted as the interventional effect, to which the semiparametric theory and the quadruply robust estimator are still applicable.

Chapter 4

CAUSAL MEDIATION ANALYSIS WITH MULTIPLE MEDIATORS

4.1 Introduction

The development of causal inference in the past decade promotes the study of mediation by providing well-defined causal direct and indirect effects, by clarifying causal assumptions needed for effect identification, and by cultivating methodology development for the effect estimation. In particular, mediation analysis often includes the definition, identification, and estimation of a direct effect of an exposure and its indirect effect that goes through some known mediators (Robins and Greenland, 1992; Petersen et al., 2006; Imai et al., 2010; Pearl, 2013).

The total effect of an exposure called the average treatment effect can be decomposed into a particular indirect effect through a set of mediators called the total indirect effect and a corresponding pure direct effect (Robins and Greenland, 1992; VanderWeele, 2013). This decomposition is widely used, and is of particular interest when causal mechanism discovery is the primary goal. Most existing methods of mediation analysis focus on applications when only one mediator is known or of interest and measured. When multiple mediators are measured, the total indirect effect contains a joint effect of the multiple mediators, see VanderWeele and Vansteelandt (2014) for a regression-based estimation of joint direct and indirect effects of multiple mediators. Apart from the joint effects, the estimation of effects that go through each mediator would be of interest. For example, the effect of ethnicity on cardiovascular disease risk is mediated through diet and exercise, and how much of the effect is mediated through diet and how much is through exercise may inform different interventions or policies.

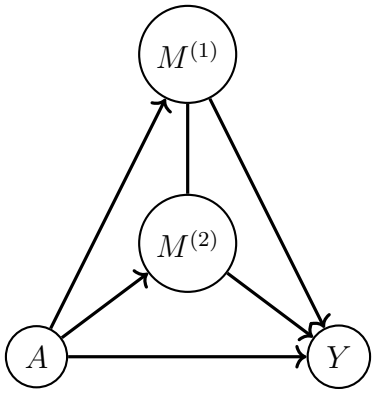


Figure 4.1: The Causal DAG with Two Non-ordered Mediators

When the causal ordering of the mediators is known, decomposition of the total effect and identification/estimation strategies for mediated effects through each mediator can be developed analogous to scenarios with a single mediator (Daniel et al., 2015; Bellavia and Valeri, 2018; Steen et al., 2017). However, the causal structure between the mediators is often unknown in practice when the temporal order is unclear. In other words, it is not always clear whether $M^{(1)}$ affects $M^{(2)}$ or vice versa, as illustrated in Figure 1. Moreover, there could be (unmeasured) confounding between the two mediators. Therefore, it is preferable to estimate the mediated effect without having to know the causal structure. In this paper, we focus on further decomposing the total indirect effect in the presence of multiple mediators, without assuming a particular causal direction among mediators, and allowing for confounding between the two mediators is not induced by the treatment.

Distinguishing the indirect effects through each mediator is complicated in general. One mediator can serve as a treatment-induced confounder for another, which leads to difficulties when isolating the mediated effect through each mediator. Specifically, the commonly used sequential ignorability assumptions (Imai et al., 2010), which are crucial to the identification of the total indirect effect, are often violated in the presence of treatment-induced confounders. Even when the order of the mediators is known, strong identification assump-

tions are often required for mediation analysis (Steen et al., 2017, Daniel et al., 2015). There has been relatively little literature on mediation analysis for multiple mediators without the knowledge of the causal structure between the mediators. In such a setting, Vansteelandt and Daniel (2017) proposed a new type of indirect effect called interventional effects that capture the effect of an exposure on an outcome through specific pathways in the presence of multiple mediators between which the causal structure is unknown. The interventional effects do not arise from the decomposition of the total indirect effect, so it is not directly applicable when the total indirect effect is used to capture the aggregated indirect effect of mediators. On the other hand, Taguri et al. (2018) provide a decomposition of the total indirect effect for cases with two or three non-ordered mediators. However, their identification assumptions include a strong "extended cross-world independence assumption", and the interpretation of the estimands is unclear when such an assumption is violated.

In this paper, we provide a decomposition of the total indirect effect that does not depend on the knowledge of the causal relationship between multiple mediators, along with a set of identification assumptions that are weaker than the ones proposed in Taguri et al. (2018). In addition, we study semiparametric inference for the estimation problem. We first discuss the properties of the decomposed effects and identification assumptions. We show that the formulae identifying the effects may be interpreted as interventional effects when some identification assumptions are violated, as long as the total indirect effect is identified. Moreover, when the mediators are not causally associated with each other, *i.e.*, when the mediators are not causes of one another, the decomposed effect that goes through a mediator is consistent with the total indirect effect when it is the only mediator considered. To facilitate inference, we derive the semiparametric bound for the effects. Unlike usual expressions, the efficient influence functions for the mediated effects contain conditional densities that are not variation independent. We consider a reparameterization based on copulas to address the problem of model incompatibility. The corresponding estimator is quadruply robust, that is, consistent under four types of misspecification of the nuisance models.

The rest of the paper is organized as follows. In Section 2, we provide decompositions for the average treatment effect, the total indirect effect and define the indirect effects that exit through each of the mediators. In Section 3, we discuss the identification assumptions needed for the decomposed effects, and develop the semi-parametric theory for robust estimation of the decomposed effects. In Section 3, we use simulation to demonstrate the quadruple robustness of the estimators derived in the previous section. In Section 4, we apply the method to a political framing data set that investigates the effect of media framing on people's attitudes towards immigration. Several concluding remarks are given in Section 6.

4.2 Effect Decomposition

4.2.1 Review of Decomposition of Average Treatment Effect

Let A denote the treatment, Y denote the outcome, M denote the mediators, and X denote all the baseline confounders. Before we define the direct and indirect effects, we first introduce the potential outcome framework. With well-defined interventions on A and M, the potential outcome Y_{am} is the value the outcome Y would have taken had the treatment been set to a and the mediator been set to m. Similar definitions can be given to any variable with some known causes. For example, the potential mediator M_a is the value the mediator M would have taken had the treatment been set to a. The potential outcomes are sometimes called counterfactuals because we can only observe a single version of potential outcomes for a subject, that is the potential outcome with causes set to the factual value they indeed take.

Some technical assumptions are generally needed for the definition of direct and indirect effects. We assume consistency holds, that is, an observed variable equals its corresponding counterfactual with its causes set to the factual values. For example, consistency implies $Y_{am} = Y$ when $A = a$ and $M = m$. More generally, we also assume composition holds such that when a downstream cause is not set to a level, it is equivalent to setting it to its counterfactual value with its upstream causes set to the same level they are set to in

the potential outcome, such as $Y_{a=1} = Y_{1M_1}$. We also require the positivity assumptions $f(a \mid X) > 0$ and $f(m \mid A, X) > 0$ to hold for all a and m.

An overall effect of the exposure on the outcome, regardless of causal pathways, is captured by the average treatment effect, defined as $E(Y_{a=1} - Y_{a=0})$, where Y_a is the potential outcome had the treatment A been a. The average treatment effect can be decomposed into two parts: the pure (natural) direct effect of A (PDE) and the total (natural) indirect effect of A that goes through M (TIE). They can be defined rigorously using the potential outcome framework, where M_a denotes the value the mediator would have taken had the treatment been a:

$$\text{PDE} = E(Y_{1M_0} - Y_{0M_0});$$
$$\text{TIE} = E(Y_{1M_1} - Y_{1M_0}).$$

The pure direct effect is the part of treatment that does not go through the downstream mediators, so the mediator is fixed at the control ($a = 0$) level. The total indirect effect is the part of treatment that goes through the downstream mediators when the treatment is presented.

Besides the direct effect of the treatment that does not go through any mediators, we are often interested in the indirect (mediated) effect that goes through some known mediation mechanisms. When there are multiple mediation pathways, further decomposition of the total indirect effect is needed.

4.2.2 Further Decomposition of the Total Indirect Effect

To simplify the exposition, we focus on the scenario with two mediators, which is often of scientific interest. Three or more mediators can be handled similarly, see Section 6. Let $M = (M^{(1)}, M^{(2)})$, where $M^{(1)}$ and $M^{(2)}$ denote two possibly multivariate mediators of interest. Without assuming a causal structure between $M^{(1)}$ and $M^{(2)}$, we consider the

following decomposition:

$$TIE = EIE_{M^{(1)}} + EIE_{M^{(2)}} - INT, \tag{4.1}$$

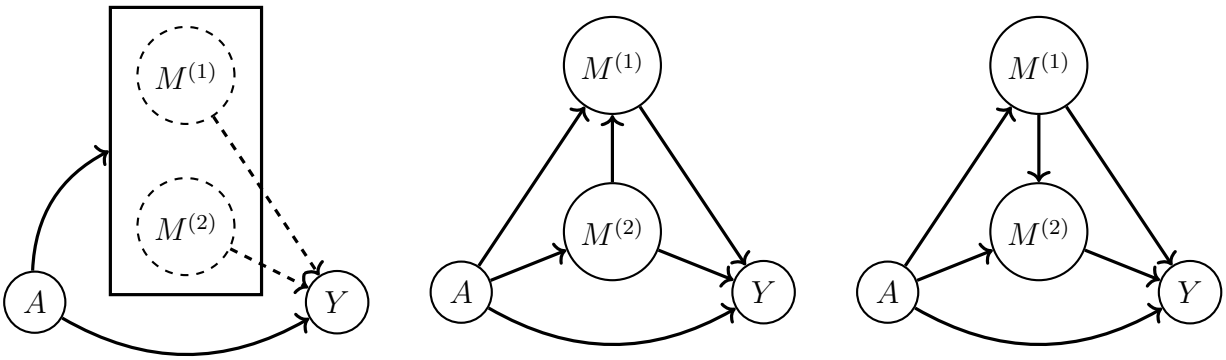
where

1. The exit indirect effect through $M^{(1)}$: $EIE_{M^{(1)}} \equiv E(Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}})$.

2. The exit indirect effect through $M^{(2)}$: $EIE_{M^{(2)}} \equiv E(Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_1^{(1)}M_0^{(2)}})$.

3. An interaction effect between $M^{(1)}$ and $M^{(2)}$:

$$INT \equiv E(Y_{1M_1^{(1)}M_1^{(2)}} + Y_{1M_0^{(1)}M_0^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}} - Y_{1M_1^{(1)}M_0^{(2)}}).$$



a: Unknown Structure b: $M^{(2)}$ Causes $M^{(1)}$ c: $M^{(1)}$ Causes $M^{(2)}$

Figure 4.2: Possible causal mechanisms

The exit indirect effect $EIE_{M^{(i)}}$ picks up the indirect effect of the exposure that exits the mediator set through $M^{(i)}$ and goes straight to the outcome. In other words, any indirect effects of the exposure whose last stop is $M^{(i)}$ before going into the outcome, regardless of the exact path, is captured by $EIE_{M^{(i)}}$. Dotted arrows in Figure 4.2a represent the exit indirect effects through $M^{(1)}$ and $M^{(2)}$, which illustrates how the indirect effect enters each

of the mediators is irrelevant for the definition of Exit Indirect Effect. Figure 4.2b and 4.2c demonstrate two cases when the causal ordering is known. In Figure 4.2b, $M^{(2)}$ is the cause of $M^{(1)}$, $EIE_{M^{(1)}}$ picks up the indirect effect that exits the mediator set ($\{M^{(1)}, M^{(2)}\}$) through $M^{(1)}$ by pathways $A \to M^{(2)} \to M^{(1)} \to Y$ and $A \to M^{(1)} \to Y$, and $EIE_{M^{(2)}}$ picks up the indirect effect that exit through the cause set through $M^{(2)}$ by pathway $A \to M^{(2)} \to Y$. Similarly, in Figure 4.2c, $M^{(1)}$ is the cause of $M^{(2)}$, $EIE_{M^{(1)}}$ picks up the indirect effect that exits the cause set through $M^{(1)}$ by pathway $A \to M^{(1)} \to Y$, and $EIE_{M^{(2)}}$ picks up the indirect effects that exit through the mediator set through $M^{(2)}$ by pathways $A \to M^{(1)} \to M^{(2)} \to Y$ and $A \to M^{(2)} \to Y$.

This decomposition has the following properties:

Property 1 *Exit indirect effects through $M^{(1)}$ and $M^{(2)}$ are symmetric in the sense that they stay invariant when exchanging the labels of mediators. When the causal structure is unknown between mediators, the labeling of the mediators is entirely arbitrary. Therefore, it is desirable to define and consequently interpret indirect effects in a manner that is invariant to the labeling of mediators.*

Property 2 *When an intervention on $M^{(2)}$ does not change the potential values of $M^{(1)}$ and vice versa, by composition assumption,*

$$Y_{1M_1^{(1)}M_0^{(2)}} = Y_{1M_0^{(2)}},\ Y_{1M_0^{(1)}M_1^{(2)}} = Y_{1M_0^{(1)}},\ Y_{1M_1^{(1)}M_1^{(2)}} = Y_{1M_1^{(2)}} = Y_{1M_1^{(1)}}, \tag{4.2}$$

the exit indirect effects reduce to the well-studied total (natural) indirect effects:

- $EIE_{M^{(1)}} = E(Y_{1M_1^{(1)}} - Y_{1M_0^{(1)}}) \equiv TIE_{M^{(1)}}$,
- $EIE_{M^{(2)}} = E(Y_{1M_1^{(2)}} - Y_{1M_0^{(2)}}) \equiv TIE_{M^{(2)}}$.

Property 3 *When causal ordering between mediators is known, the exit indirect effect for the mediator closest to the outcome in the causal chain is the total indirect effect for that respective variable. Under the causal DAG in Figure 4.2b, $EIE_{M^{(1)}} = TIE_{M^{(1)}}$, and $M^{(2)}$ is a*

treatment-induced confounder for the relationship between $M^{(1)}$ and Y. In this case, $EIE_{M^{(2)}}$ is the effect of A on Y not through $M^{(1)}$. Similarly, under the causal DAG in Figure 4.2c, $EIE_{M^{(2)}} = TIE_{M^{(2)}}$, and $M^{(1)}$ is a treatment-induced confounder for the relationship between $M^{(2)}$ and Y. Such equivalences are favorable because when the causal ordering is known, the interpretation and estimation of the total indirect effects reduces to previous works on mediation analysis in the presence of treatment-induced confounding.

Property 4 *The remainder term INT can be written as a difference between two differences*

$$Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}},\ Y_{1M_1^{(1)}M_0^{(2)}} - Y_{1M_0^{(1)}M_0^{(2)}}.$$

The former term can be seen as an indirect effect through $M^{(1)}$ with $M^{(2)}$ fixed at the level it would have taken had the treatment been 1, and the latter term can be seen as an indirect effect through $M^{(1)}$ with $M^{(2)}$ fixed at the level it would have taken had the treatment been 0. The difference between these two terms can be seen as the indirect effect through $M^{(1)}$ modified by $M^{(2)}$. Therefore INT depicts the effect of the interaction between $M^{(1)}$ and $M^{(2)}$ on the mean outcome. When the effect of $M^{(1)}$ and $M^{(2)}$ on the mean outcome do not interact with each other, meaning that $M^{(1)}$ do not modify the effect of $M^{(2)}$ on the mean outcome and vice versa, the remainder term $INT = 0$. Under a linear structural equation model for the potential outcome, this is reflected as the coefficient of the interaction term between $M^{(1)}$ and $M^{(2)}$ being 0.

Remark 2 *A decomposition advocated in Taguri et al. (2018) is similar to our proposal:*

$$TIE = PSE_{M^{(1)}} + PSE_{M^{(2)}} + INT, \tag{4.3}$$

where $PSE_{M^{(1)}} = E(Y_{1M_1^{(1)}M_0^{(2)}} - Y_{1M_0^{(1)}M_0^{(2)}})$ and $PSE_{M^{(2)}} = E(Y_{1M_0^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_0^{(2)}})$. Taguri et al. (2018) called $PSE_{M^{(1)}}$ and $PSE_{M^{(2)}}$ path-specific effects but they are not the same as the conventional definition of path-specific effects in Daniel et al. (2015). When considering the mediated effect that goes through each mediator, instead of setting the mediators to the value they would have taken had the treatment been set to 1 as in our decomposition, Taguri

et al. (2018) set the mediator that is not of interest to the value it would have taken had the treatment been set to 0. However, the decomposition in (4.3) does not enjoy Properties 2.2 and 2.3, so the interpretation remains rather complicated even when the causal structure is known between $M^{(1)}$ and $M^{(2)}$. Nonetheless, the identification and estimation results in section 3 and 4 can be easily adapted to this decomposition.

4.3 Identification and Estimation of the Decomposed Effects

4.3.1 Identification

The decomposition of the total indirect effect is only reasonable when the average treatment effect and the total indirect effect (hence also the pure direct effect) are identified. Consequently, the set of identification assumptions needed for the average treatment effect and the total indirect effect are first made and denoted as **Set** I. The additional set of assumptions needed for the identification of the decomposed effects is denoted as **Set** II.

Set I: Identification Assumptions for the average treatment effect and the total indirect effect:

I.1 $\{Y_{am^{(1)}m^{(2)}}, M_a^{(1)}, M_a^{(2)}\} \perp\!\!\!\perp A \mid X,$

I.2 $Y_{am^{(1)}m^{(2)}} \perp\!\!\!\perp \{M_a^{(1)}, M_a^{(2)}\} \mid A = a, X,$

I.3 $Y_{am^{(1)}m^{(2)}} \perp\!\!\!\perp \{M_{a^\star}^{(1)}, M_{a^\star}^{(2)}\} \mid X.$

Set II: Additional Identification Assumptions for $EIE_{M^{(1)}}$, $EIE_{M^{(2)}}$, and INT:

II.1 $E(Y_{1M_1^{(1)}m^{(2)}} - Y_{1M_0^{(1)}m^{(2)}} \mid M_1^{(2)} = m^{(2)}, X) = E(Y_{1M_1^{(1)}m^{(2)}} - Y_{1M_0^{(1)}m^{(2)}} \mid X),$

II.2 $E(Y_{1m^{(1)}M_1^{(2)}} - Y_{1m^{(1)}M_0^{(2)}} \mid M_1^{(1)} = m^{(1)}, X) = E(Y_{1m^{(1)}M_1^{(2)}} - Y_{1m^{(1)}M_0^{(2)}} \mid X).$

Set I is essentially the sequential ignorability assumptions of Imai et al. (2010). Assumption I.1 is implied by the assumption that there is no unmeasured confounding between the treatment and the downstream variables. Assumption I.2 implies that there is no unmeasured confounding between the mediator and the outcome. Assumption I.3 precludes

any confounders between the mediator and the outcome that is affected by the treatment. Note that confounders not affected by the treatment that are causes of both mediators are allowed. **Set** II contains two assumptions that can be seen as that one mediator does not induce additional effect heterogeneity of the treatment that is mediated through the other mediator. Under **Set** I and **Set** II, we have the following identification results:

Theorem 4.3.1 *Under assumption* ***Set*** I *and* ***Set*** II*, the decomposed effects in (2) are identified as*

1. *The Exit Indirect Effect of* $M^{(1)}$*:*

$$\begin{aligned}\Delta^{M^{(1)}} \equiv\ & E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M^{(1)}=m^{(1)}|A=1,X} E(Y \mid A=1, m^{(1)}, m^{(2)}, X) \\ & - E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M^{(1)}=m^{(1)}|A=0,X} E(Y \mid A=1, m^{(1)}, m^{(2)}, X),\end{aligned}$$

2. *The Exit Indirect Effect of* $M^{(2)}$*:*

$$\begin{aligned}\Delta^{M^{(2)}} \equiv\ & E_X E_{M^{(1)}=m^{(1)}|A=1,X} E_{M^{(2)}=m^{(2)}|A=1,X} E(Y \mid A=1, m^{(1)}, m^{(2)}, X) \\ & - E_X E_{M^{(1)}=m^{(1)}|A=1,X} E_{M^{(2)}=m^{(2)}|A=0,X} E(Y \mid A=1, m^{(1)}, m^{(2)}, X).\end{aligned}$$

3. *Effect through interaction between* $M^{(1)}$ *and* $M^{(2)}$*:*

$$\begin{aligned}\Delta^{INT} \equiv\ & E_X E_{M^{(1)}=m^{(1)}, M^{(2)}=m^{(2)}|A=1,X} E(Y|A=1, m^{(1)}, m^{(2)}, X) \\ & - E_X E_{M^{(1)}=m^{(1)}, M^{(2)}=m^{(2)}|A=0,X} E(Y|A=1, m^{(1)}, m^{(2)}, X) \\ & - 2 \times E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M^{(1)}=m^{(1)}|A=1,X} E(Y \mid A=1, m^{(1)}, m^{(2)}, X) \\ & + E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M^{(1)}=m^{(1)}|A=0,X} E(Y \mid A=1, m^{(1)}, m^{(2)}, X) \\ & + E_X E_{M^{(1)}=m^{(1)}|A=1,X} E_{M^{(2)}=m^{(2)}|A=0,X} E(Y \mid A=1, m^{(1)}, m^{(2)}, X).\end{aligned}$$

Assumptions I.3, II.1 and II.2 involve counterfactuals from different worlds, meaning that some of their causes are set to different values. These assumptions are called "Cross-World" assumptions. "Cross-World" assumptions are strong in the sense that they are not testable

by experiments on the variables in the system without further assumptions, which makes them impossible to be confirmed or disputed (Richardson & Robins, 2010). On the contrary, Assumption I.1 and I.2 in **Set** I are "Single-World" assumptions that can be tested by an experiment that randomizes both the treatment and the mediators.

When only Assumption I.1 and I.2 hold, the identified expression $\Delta^{M^{(2)}}$ of $EIE_{M^{(2)}}$ equals the interventional indirect effect of exposure on outcome via $M^{(2)}$ defined by VanderWeele and Tchetgen Tchetgen (2017), which extended the concept of interventional effects (VanderWeele et al., 2014) to the non-ordered multiple mediators setting. Therefore the identified expression for $EIE_{M^{(2)}}$ in Theorem 4.3.1 continues to have a causal interpretation as an interventional effect even when all the strong "Cross-World" assumptions in **Set** II fail to hold. The interventional indirect effects were not symmetrically defined for the two mediators in Vansteelandt and Daniel (2017) so that the remainder term in a decomposition of the combined interventional effect would have a simple form. Since $EIE_{M^{(1)}}$ and $EIE_{M^{(2)}}$ are defined symmetrically, the identifying formula $\Delta_{M^{(1)}}$ of $EIE_{M^{(1)}}$ is not exactly the same as the interventional indirect effect of exposure on outcome via $M^{(2)}$. Another notable difference is that interventional effects are not defined for decomposing the total indirect effect. In fact, they do not add up to the total indirect effect.

Remark 3 *Under the identification assumptions* I *and* II*, when $M^{(1)}$ is independent of $M^{(2)}$ given A and X, the identifying expression for the total indirect effect $E[Y_{1M_1^{(j)}} - Y_{1M_0^{(j)}}]$ coincides with that of $EIE_{M^{(j)}}$, when only mediator j is considered, where $j = 1, 2$. The result holds under different assumptions than Property 2, which requires a stronger assumption that no effect of $M^{(1)}$ on $M^{(2)}$ and vice versa at the individual level, as listed in equation (4.2), but not assumptions* I *and* II*.*

Remark 4 *In addition to* ***Set*** I*, Taguri et al. (2018) makes the following "extended cross-world independence assumptions":*

$$Y_{am^{(1)}m^{(2)}} \perp\!\!\!\perp (M^{(1)}_{a^*}, M^{(2)}_{a^{**}}) \mid X, \qquad M^{(1)}_{a^*} \perp\!\!\!\perp M^{(2)}_{a^{**}} \mid X,$$

for all $a, m^{(1)}, m^{(2)}, a^*, a^{**}$. *This assumption is much stronger than Assumption* ***Set*** II. *In particular, we do not require the joint distribution of the "Cross-World" potential mediators to be independent of every potential outcome, nor do we require the mediators to be conditionally independent of each other.*

4.3.2 Moment-type Estimators

Denote the identified exit indirect effect through $M^{(j)}$ as $\Delta^{M^{(j)}}$, where $j = 1, 2$. The following theorem gives four equivalent forms of $\Delta^{M^{(j)}}$, $j = 1, 2$.

Theorem 4.3.2 $\Delta^{M^{(j)}} = \Delta_1^{M^{(j)}} = \Delta_2^{M^{(j)}} = \Delta_3^{M^{(j)}} = \Delta_4^{M^{(j)}}$, $j = 1, 2$, *where*

$$\begin{aligned}
\Delta_1^{M^{(j)}} &= E\left(\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid A = 1, X) - f(M^{(j)} \mid A = 0, X)}{f(M^{(j)} \mid A = 1, M^{(3-j)}, X)} Y\right), \\
\Delta_2^{M^{(j)}} &= E\left(\frac{A}{f(A \mid X)} \{\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)\}\right), \\
\Delta_3^{M^{(j)}} &= E\left(\frac{2A-1}{f(A \mid X)} \eta_{3-j}(1, M^{(j)}, X)\right), \\
\Delta_4^{M^{(j)}} &= E\left(\gamma_j(1, X) - \gamma_j(0, X)\right),
\end{aligned}$$

and

$$\begin{aligned}
\eta_j(a, M^{(3-j)}, X) &= \int E[Y \mid A = 1, m^{(j)}, M^{(3-j)}, X] f(m^{(j)} \mid A = a, X) dm^{(j)}, \\
\gamma_j(a, X) &= \iint E[Y \mid A = 1, m^{(3-j)}, m^{(j)}, X] f(m^{(j)} \mid A = a, X) f(m^{(3-j)} \mid 1, X) dm^{(j)} dm^{(3-j)}.
\end{aligned}$$

Theorem 3.2 presents four different representations of the target estimand $\Delta^{M^{(j)}}$. Moment-type estimators $\hat{\Delta}_1^{M^{(j)}}$, $\hat{\Delta}_2^{M^{(j)}}$, $\hat{\Delta}_3^{M^{(j)}}$, and $\hat{\Delta}_4^{M^{(j)}}$ can be derived from these representations by replacing conditional densities or expectations with their estimates and the outer expectation by the empirical average. To reduce the burden of the possible high-dimensionality of $M^{(1)}$ and $M^{(2)}$, we consider parametric models for the estimation of nuisance parameters for the components of the moment-type estimators. Each estimator is consistent when its components are consistently estimated. That is to say, the validity of $\Delta_1^{M^{(j)}}$ relies on correctly specified models for $f(A \mid X)$, $f(M^{(j)} \mid A, X)$, and $f(M^{(j)} \mid A = 1, M^{(3-j)})$; the

validity of $\Delta_2^{M^{(j)}}$ relies on correctly specified models for $f(A \mid X)$, $E(Y \mid A, M^{(j)}, M^{(3-j)}, X)$, and $f(M^{(j)} \mid A, X)$; the validity of $\Delta_3^{M^{(j)}}$ relies on correctly specified models for $f(A \mid X)$, $E(Y \mid A, M^{(j)}, M^{(3-j)}, X)$, and $f(M^{(3-j)} \mid A, X)$; and the validity of $\Delta_4^{M^{(j)}}$ relies on correctly specified models for $E(Y \mid A, M^{(j)}, M^{(3-j)}, X)$, $f(M^{(j)} \mid A, X)$, and $f(M^{(3-j)} \mid A, X)$.

The estimators for Δ^{INT} can be constructed by subtracting any estimator of the total indirect effect, denoted by Δ^{TIE} under assumptions in **Set** I, by the moment-type estimators. Possible choices of $\hat{\Delta}^{TIE}$ include moment-type estimators and the triply robust estimator $\hat{\Delta}_{tri}^{TIE}$ for the total indirect effect proposed in the seminal work of Tchetgen Tchetgen and Shpitser (2012).

4.3.3 Robust Estimation

We derive the efficient influence function of $\Delta^{M^{(j)}}$ under a nonparametric model $\mathcal{M}_{non}$ that does not impose constraints on the observed data $\mathcal{O}$.

Theorem 4.3.3 *The efficient influence function in $\mathcal{M}_{non}$ for $EIE_{M^{(j)}}$ is*

$$
\begin{aligned}
&S_{eff}^{M^{(j)}}(\mathcal{O}, \Delta^{M^{(j)}}) \\
=&\frac{A}{f(A|X)} R_{\boldsymbol{M},X}^{(j)}(Y - E[Y|A, M^{(3-j)}, M^{(j)}, X]) + \frac{A}{f(A|X)}(\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)) \\
&+ \frac{2A-1}{f(A|X)}\eta_{3-j}(1, M^{(j)}, X) + \left(1 - \frac{2A}{f(A|X)}\right)\gamma_j(1, X) - \left(1 - \frac{1}{f(A|X)}\right)\gamma_j(0, X) - \Delta^{M^{(j)}},
\end{aligned}
$$

where

$$
\begin{aligned}
R_{\boldsymbol{M},X}^{(j)} &= \frac{\{f(M^{(j)}|A=1, X) - f(M^{(j)}|A=0, X)\} f(M^{(3-j)}|A=1, X)}{f(M^{(j)}, M^{(3-j)}|A=1, X)}, \\
\eta_j(a, M^{(3-j)}, X) &= \int E[Y \mid A=1, m^{(j)}, M^{(3-j)}, X] f(m^{(j)} \mid A=a, X) dm^{(j)}, \\
\gamma_j(a, X) &= \iint E[Y \mid A=1, m^{(3-j)}, m^{(j)}, X] f(m^{(j)} \mid A=a, X) f(m^{(3-j)} \mid 1, X) dm^{(j)} dm^{(3-j)},
\end{aligned}
$$

The efficient influence function in $\mathcal{M}_{non}$ for the effect through mediator interaction INT is:

$$
S_{eff}^{INT}(\mathcal{O}, \Delta^{INT}) = S_{eff}^{TIE}(\mathcal{O}, \Delta^{TIE}) - S_{eff}^{M^{(1)}}(\mathcal{O}, \Delta^{M^{(1)}}) - S_{eff}^{M^{(2)}}(\mathcal{O}, \Delta^{M^{(2)}}),
$$

where $S_{eff}^{TIE}(\mathcal{O}, \Delta^{TIE})$ is the efficient score function in $\mathcal{M}_{non}$ of the total indirect effect proposed in the seminal work of Tchetgen Tchetgen and Shpitser (2012).

Hence, the semiparametric efficiency bounds for the estimation of $\Delta^{M^{(j)}}$ and Δ^{INT} in $\mathcal{M}_{non}$ are $E(S_{eff}^{M^{(j)}} S_{eff}^{M^{(j)}{}^T})$ and $E(S_{eff}^{INT} S_{eff}^{INT^T})$, and the asymptotic variance of any regular and asymptotically linear estimator of $\Delta^{M^{(j)}}$ and Δ^{INT} in $\mathcal{M}_{non}$ are greater than or equal to these bounds. Here we allow $M^{(j)}$ to be multivariate and the superscript ^{T} denotes the vector transpose.

Construction of an estimator for $\Delta^{M^{(j)}}$ based on the efficient influence function can be implemented by the estimating equation:

$$p_n(\hat{S}_{\text{eff}}^{M^{(j)}}(\mathcal{O}, \hat{\Delta}_{\text{quad}}^{M^{(j)}})) = 0, \tag{4.4}$$

where $\hat{S}_{\text{eff}}^{M^{(j)}}$ is an estimated efficient influence function with components replaced by their parametric working models, and p_n is the empirical measure. The expression of $S_{\text{eff}}^{M^{(j)}}$ includes variation dependent components:

$$f(M^{(j)}|A, X), f(M^{(3-j)}|A, X), f(M^{(3-j)}|A, M^{(j)}, X), f(M^{(j)}|A, M^{(3-j)}, X).$$

However, Richardson et al. (2017) pointed out that the multiple robustness property to be defined precisely later is relevant only when the components of $\hat{\Delta}_{\text{quad}}^{M^{(j)}}$ can be compatibly modeled, in other words, when the components are variation independent.

To address this issue, we reparameterize the joint distribution $f(M^{(j)}, M^{(3-j)} \mid A, X)$ in to three parts:

- $f(M^{(j)}|A, X)$,
- $f(M^{(3-j)}|A, X)$,
- $\mathcal{C}(F_{M^{(1)}|A,X}(M^{(1)}), F_{M^{(2)}|A,X}(M^{(2)}) \mid A, X) = F(M^{(1)}, M^{(2)} \mid A, X)$.

The first two terms are marginal distributions of the mediators, and the third term is a copula that captures the dependence structure between $M^{(1)}$ and $M^{(2)}$. Here F denotes the cumulative distribution function.

Sklar's Theorem (Sklar, 1959) implies that the three components uniquely determine the joint distribution, and that they can be modeled independently. Using the new parameterization, we can rewrite $S_{\text{eff}}^{M^{(j)}}$ by changing the form of $R_{\mathbf{M},X}^{(j)}$ using the copula. For example, for univariate $M^{(1)}$ and $M^{(2)}$, when both $M^{(1)}$ and $M^{(2)}$ are continuous,

$$R_{\mathbf{M},X}^{(j)} = \frac{f(M^{(j)} \mid A=1,X) - f(M^{(j)} \mid A=0,X)}{f(M^{(j)} \mid A=1,X)\boldsymbol{c}(F_{M^{(1)}|A=1,X}(M^{(1)}), F_{M^{(2)}|A=1,X}(M^{(2)}) \mid A=1,X)},$$

where $\boldsymbol{c}(F_{M^{(1)}|A,X}(m^{(1)}), F_{M^{(2)}|A,X}(m^{(2)}) \mid A,X)$ is the copula density computed by taking derivatives of the copula $\mathcal{C}$ with respect to its arguments and evaluated at $F_{M^{(1)}|A,X}(m^{(1)})$ and $F_{M^{(2)}|A,X}(m^{(2)})$. When both $M^{(1)}$ and $M^{(2)}$ are binary,

$$R_{\mathbf{M},X}^{(j)} = \frac{\{f(M^{(j)} \mid A=1,X) - f(M^{(j)}) \mid A=0,X)\}f(M^{(3-j)} \mid A=1,X)}{\mathcal{C}(F_{M^{(1)}|A=1,X}(M^{(1)}), F_{M^{(2)}|A=1,X}(M^{(2)}) \mid A=1,X)}.$$

With the variation independent parameterization, we can construct a locally efficient estimator $\hat{\Delta}_{\text{quad}}^{M^{(j)}}$ of $\Delta^{M^{(j)}}$using estimating equation (3). The components of $S_{\text{eff}}^{M^{(j)}}$ are replaced by their parametric working models that are denoted by f^{par} and E^{par}. In particular, $f^{\text{par}}(M^{(j)}, M^{(3-j)}|A=1,X)$ is modeled by $f^{\text{par}}(M^{(j)}|A=1,X)$, $f^{\text{par}}(M^{(3-j)}|A=1,X)$, and $\mathcal{C}^{\text{par}}(F_{M^{(1)}|A,X}(m^{(1)}), F_{M^{(2)}|A,X}(m^{(2)}) \mid A,X)$. Examples of copula models are given by Joe (1997), Nelsen (2007), and Jaworski et al. (2010).

In summary, the estimator $\hat{\Delta}_{\text{quad}}^{M^{(j)}}$ that takes the form:

$$\begin{aligned}
\hat{\Delta}_{\text{quad}}^{M^{(j)}} =& \frac{A}{\hat{f}^{\text{par}}(A|X)}\hat{R}_{\mathbf{M},X}^{(j),\text{par}}(Y - \hat{E}^{\text{par}}[Y|A, M^{(3-j)}, M^{(j)}, X]) \\
&+ \frac{A}{\hat{f}^{\text{par}}(A|X)}\{\hat{\eta}_j^{\text{par}}(1, M^{(3-j)}, X) - \hat{\eta}_j^{\text{par}}(0, M^{(3-j)}, X)\} \\
&+ \frac{2A-1}{\hat{f}^{\text{par}}(A|X)}\hat{\eta}_{3-j}^{\text{par}}(1, M^{(j)}, X) + \left(1 - \frac{2A}{\hat{f}^{\text{par}}(A|X)}\right)\hat{\gamma}_j^{\text{par}}(1,X) - \left(1 - \frac{1}{\hat{f}^{\text{par}}(A|X)}\right)\hat{\gamma}_j^{\text{par}}(0,X).
\end{aligned}$$

where

$$\hat{R}^{(j),\text{par}}_{\mathbf{M},X} = \frac{\{\hat{f}^{\text{par}}(M^{(j)}|A=1,X) - \hat{f}^{\text{par}}(M^{(j)}|A=0,X)\}\hat{f}^{\text{par}}(M^{(3-j)}|A=1,X)}{\hat{f}^{\text{par}}(M^{(j)},M^{(3-j)}|A=1,X)},$$

$$\hat{\eta}^{\text{par}}_j(a,M^{(3-j)},X) = \int \hat{E}^{\text{par}}[Y \mid A=1,m^{(j)},M^{(3-j)},X]\hat{f}^{\text{par}}(m^{(j)} \mid A=a,X)dm^{(j)},$$

$$\hat{\gamma}^{\text{par}}_j(a,X) = \iint \hat{E}^{\text{par}}[Y \mid A=1,m^{(3-j)},m^{(j)},X]\hat{f}^{\text{par}}(m^{(j)} \mid A=a,X)\hat{f}^{\text{par}}(m^{(3-j)} \mid 1,X)dm^{(j)}dm^{(3-j)}.$$

The estimator for Δ^{INT} is

$$\hat{\Delta}^{\text{INT}}_{\text{quad}} = \hat{\Delta}^{TIE}_{tri} - \hat{\Delta}^{M^{(1)}}_{\text{quad}} - \hat{\Delta}^{M^{(2)}}_{\text{quad}},$$

where $\hat{\Delta}^{TIE}_{tri}$ is the triply robust estimator of the total indirect effect proposed by Tchetgen Tchetgen and Shpitser (2012).

Theorem 4.3.4 *The estimators $\hat{\Delta}^{M^{(j)}}_{quad}$ and $\hat{\Delta}^{INT}_{quad}$ are consistent and asymptotically normal under some mild regularity conditions discussed in the supplementary material if one of the following four conditions holds. Here each line lists the set of models that are assumed to be correctly specified.*

$\mathcal{M}_1 : f^{par}(A \mid X), f^{par}(M^{(1)} \mid A,X), f^{par}(M^{(2)} \mid A,X), \mathcal{C}^{par}(F_{M^{(1)}|A,X}(m^{(1)}), F_{M^{(2)}|A,X}(m^{(2)}) \mid A,X);$

$\mathcal{M}_2 : f^{par}(A \mid X), f^{par}(M^{(1)} \mid A,X), E^{par}[Y \mid A,M^{(1)},M^{(2)},X];$

$\mathcal{M}_3 : f^{par}(A \mid X), f^{par}(M^{(2)} \mid A,X), E^{par}[Y \mid A,M^{(1)},M^{(2)},X];$

$\mathcal{M}_4 : f^{par}(M^{(1)} \mid A,X), f^{par}(M^{(2)} \mid A,X), \mathcal{C}^{par}(F_{M^{(1)}|A,X}(m^{(1)}), F_{M^{(2)}|A,X}(m^{(2)}) \mid A,X),$
$E^{par}[Y \mid A,M^{(1)},M^{(2)},X].$

The estimators are locally semiparametric efficient in the sense that they achieve the semiparametric efficiency bounds at the intersection of the submodels $\mathcal{M}_{intersection} = \mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3 \cap \mathcal{M}_4$ where all four conditions hold.

Compared with the moment-type estimators proposed in section 3.1, while the quadruply robust estimator $\hat{\Delta}^{M^{(j)}}_{\text{quad}}$ is consistent and asymptotically normal under any of the four conditions $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$, the moment-based estimator $\hat{\Delta}^{M^{(j)}}_1$ is only consistent under $\mathcal{M}_1$, $\hat{\Delta}^{M^{(j)}}_2$ is only consistent under $\mathcal{M}_2$, $\hat{\Delta}^{M^{(j)}}_3$ is only consistent under $\mathcal{M}_3$, and $\hat{\Delta}^{M^{(j)}}_4$ is only consistent under $\mathcal{M}_4$. Therefore, $\hat{\Delta}^{M^{(j)}}_{\text{quad}}$ is more robust against model misspecification.

4.4 Simulation Studies

We use numerical simulations to demonstrate the performance of the proposed estimators and the theoretical results derived in the previous section. We compare the finite sample performance of the moment-based estimators $\hat{\Delta}_1^{M^{(j)}}$, $\hat{\Delta}_2^{M^{(j)}}$, $\hat{\Delta}_3^{M^{(j)}}$, $\hat{\Delta}_4^{M^{(j)}}$ given in section 3.2 to the proposed quadruply robust estimator $\hat{\Delta}_{\text{quad}}^{M^{(j)}}$. We generate 1000 samples, each with 500 independent observations, for binary mediators. Let *expit* denote the function $expit(x) = \exp(x)/(1+\exp(x))$. The data are generated as follows:

$X \sim N(0,1); P(A=1 \mid X) = expit(-0.6 + 1.2X);$

$\mathcal{C}(F_{M^{(1)}|A,X}(m^{(1)}), F_{M^{(2)}|A,X}(m^{(2)}) \mid A, X)$ is a Plackett Copula with Odds-Ratio $exp(1 - 2A + 5X)$;

where $F_{M^{(1)}|A,X}(m^{(1)}) = p_1^{m^{(1)}}(1-p_1)^{1-m^{(1)}}, p_1 = expit(-0.2 - 0.3A + 1.5X)$,

$F_{M^{(2)}|A,X}(m^{(2)}) = p_2^{m^{(2)}}(1-p_2)^{1-m^{(2)}}, p_2 = expit(-0.1 - 0.4A + 1.2X)$,

$Y \sim 1 + 2A + 2M^{(1)} + 4M^{(2)} + 3X + 4AM^{(2)} + 2AM^{(1)} + 4M^{(1)}M^{(2)} + N(0, 3^2).$

We compare the five estimators under a series of model misspecification by replacing the baseline covariates X with an independent normally distributed continuous variable X_2 with mean 0 and variance 1. Table 4.1 and Table 4.2 show that the simulation results are consistent with the theoretical results derived in the previous sections: when the entire likelihood is correctly specified, all five estimators are consistent; when the conditional expectation of Y is misspecified, only $\hat{\Delta}_1^{M^{(j)}}$ and $\hat{\Delta}_{quad}^{M^{(j)}}$ are consistent; when the parametric model for $f(M^{(3-j)} \mid A, X)$ is misspecified, only $\hat{\Delta}_2^{M^{(j)}}$ and $\hat{\Delta}_{quad}^{M^{(j)}}$ are consistent; when the parametric model for $f(M^{(j)} \mid A, X)$ is misspecified, only $\hat{\Delta}_3^{M^{(j)}}$ and $\hat{\Delta}_{quad}^{M^{(j)}}$ are consistent; when the propensity score $f(A=1 \mid X)$ is misspecified, only $\hat{\Delta}_4^{M^{(j)}}$ and $\hat{\Delta}_{quad}^{M^{(j)}}$ are consistent. The loss in efficiency for the quadruply robust estimator is relatively small compared to the other consistent estimators in all cases. Due to a limitation in space, we do not present results from other simulation settings, as they all gave similar quantitative conclusions.

Table 4.1: Simulation Results for $EIE_{M^{(1)}}$: 100×Bias (100×Standard Error)

	Binary $M^{(1)}$, $M^{(2)}$				
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$
All correct	0 (50)	1 (24)	3 (53)	2 (24)	1 (27)
$\mathcal{M}_1$ is correct	0 (33)	-9 (31)	-9 (54)	-9 (31)	-1 (29)
$\mathcal{M}_2$ is correct	12 (28)	-1 (23)	-2 (51)	-3 (24)	-1 (24)
$\mathcal{M}_3$ is correct	217 (145)	144 (29)	1 (52)	144 (29)	1 (39)
$\mathcal{M}_4$ is correct	-7 (40)	-3 (26)	646 (65)	0 (23)	0 (24)

Table 4.2: Simulation Results for $EIE_{M^{(2)}}$: 100×Bias (100×Standard Error)

	Binary $M^{(1)}$, $M^{(2)}$				
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$
All correct	4(53)	1 (42)	2 (63)	2 (42)	1 (44)
$\mathcal{M}_1$ is correct	5 (47)	-6 (47)	-6 (66)	-6 (47)	2 (46)
$\mathcal{M}_2$ is correct	77(135)	2(42)	-3 (65)	-3 (44)	2 (53)
$\mathcal{M}_3$ is correct	204 (52)	213 (45)	2 (67)	213 (45)	0 (47)
$\mathcal{M}_4$ is correct	-4 (49)	-7 (45)	648 (72)	0 (41)	0 (41)

Table 4.3: Simulation Results for Δ^{INT}: 100×Bias (100×Standard Error)

	Binary $M^{(1)}$, $M^{(2)}$				
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$
All correct	-1 (70)	0 (19)	-2 (97)	0 (19)	0 (12)
$\mathcal{M}_1$ is correct	-3 (59)	17 (30)	16 (95)	17 (30)	1 (17)
$\mathcal{M}_2$ is correct	-291 (272)	0 (43)	-139 (46)	-139 (30)	-1 (56)
$\mathcal{M}_3$ is correct	-217 (58)	-211 (51)	-1 (41)	-210 (33)	1 (13)
$\mathcal{M}_4$ is correct	12 (32)	10 (14)	-1294 (126)	1 (12)	1 (10)

4.5 Data Application

We demonstrate the use of our proposed estimators using data from a framing experiment (Brader et al., 2008). The question of interest is to understand the mechanism that triggers public opposition to immigration. The exposure is a racial group cue in a news article about white or non-white immigrants, and the outcome is an individual action in response to immigration information. The authors suggest different mechanisms that racial cues may affect the response of individuals through emotional reactions and perceptions about potential negative consequences.

A total of 265 individuals are randomized to receive different news stories about European and Latino immigrants. As in Imai and Yamamoto (2013), we define treatment $A = 1$ as a negative news story featuring Latino immigrants. The outcome Y is whether or not the participant agreed to send a letter about immigration policy to their member of Congress. The two mediators of interest are emotion $M^{(1)}$, which is based on a post-test questionnaire asking how they feel about increased immigration, and perceived harm $M^{(2)}$, which is calculated from participants' self-reported views on immigration. The baseline covariates X include gender, age, education, and income.

We pose similar parametric models on the components of proposed estimators as in Tingley et al. (2014), who studied different estimators in a single-mediator setting:

- A logistic regression $f(A \mid X)$ for binary treatment A.
- Linear regressions $f(M^{(1)} \mid A, X)$, $f(M^{(1)} \mid A, X)$, and a Gaussian copula for continuous mediators $M^{(1)}$ and $M^{(2)}$.
- A probit regression $f(Y \mid A, M^{(1)}, M^{(2)}, X)$ for the binary outcome Y.

We use the augmented inverse propensity weighted (AIPW) estimator (Robins et al., 1994) to estimate the average treatment effect, and the triply robust estimator proposed by Tchetgen Tchetgen and Shpitser (2012) to estimate the total indirect effect. Assumption Set II

requires that the effect of the negative news story mediated through emotion is not modified by the level of perceived harm under treatment, and the effect of the negative news story mediated through perceived harm is not modified by the level of emotion under treatment. If either of the assumptions fails to hold, the estimated effects can be interpreted as interventional indirect effects. Using the proposed quadruply robust estimator, we show that a negative news story featuring Latino immigrants increases the probability of a participant agreeing to send a letter about immigration policy to his or her member of Congress by 0.097 (-0.027, 0.251) where the parenthesis indicates the 95% confidence interval, of which 0.077 (0.005, 0.158) is mediated through emotion or perceived harm. The part of indirect effect on the probability of a participant agreed to send a letter about immigration policy to his or her member of Congress that comes straight from the difference in emotion is 0.030 (-0.139, 0.118), and the part that comes straight from the difference in perceived harm is 0.053 (-0.066, 0.195). The effect of the interaction between emotion and perceived harm is minimal: -0.006 (-0.115, 0.213).

4.6 Discussion

Compared with some other decompositions of the total indirect effect, our decomposed effects are not "pure" or "path-specific" in the sense that they consist of more than the amount of the exposure effect that only goes directly through one mediator, but all the effect that eventually leaves the mediator to enter the outcome. Knowing the causal ordering makes clear which pathways are included in the exit indirect effects. However, the identification of each "path-specific" effect usually requires additional assumptions such as "Cross-World" independence between the mediators.

We consider cases with two mediation pathways, but the idea can be applied to cases with more mediators. For example, in the case with three mediators, denoted as $(M^{(1)}, M^{(2)}, M^{(3)})$, the total indirect effect is $E[Y_{1M_1^{(1)}M_1^{(2)}M_1^{(3)}} - Y_{1M_0^{(1)}M_0^{(2)}M_0^{(3)}}]$, which can be decomposed as the sum of exit indirect effects $E[Y_{1M_1^{(1)}M_1^{(2)}M_1^{(3)}} - Y_{1M_0^{(1)}M_1^{(2)}M_1^{(3)}}]$, $E[Y_{1M_1^{(1)}M_1^{(2)}M_1^{(3)}} - Y_{1M_1^{(1)}M_0^{(2)}M_1^{(3)}}]$, $E[Y_{1M_1^{(1)}M_1^{(2)}M_1^{(3)}} - Y_{1M_1^{(1)}M_1^{(2)}M_0^{(3)}}]$, and an interaction term that can be further written as the

sum of two-way interactions minus a three-way interaction. The interpretation of the exit indirect effects remain the same as when there are two mediators.

Given its close connection to the interventional effects, which are extended to longitudinal settings in VanderWeele and Tchetgen Tchetgen (2017), a future direction of extension is to longitudinal settings where the causal structures are more complicated than the point-treatment settings. Another future direction is to investigate other variation independent parameterizations that may not involve copulas.

BIBLIOGRAPHY

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects.

Bellavia, A. and Valeri, L. (2018). Decomposition of the total effect in the presence of multiple mediators and interactions. *American journal of epidemiology*, 187(6):1311–1318.

Brader, T., Valentino, N. A., and Suhay, E. (2008). What triggers public opposition to immigration? anxiety, group cues, and immigration threat. *American Journal of Political Science*, 52(4):959–978.

CDC (2017). *Natality Information*.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):815–836.

Daniel, R., De Stavola, B., Cousens, S., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14.

Didelez, V., Dawid, A. P., and Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 138–146. AUAI Press.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Frangakis, C. E., Rubin, D. B., An, M.-W., and MacKenzie, E. (2007). Principal stratification designs to estimate input data missing due to death. *Biometrics*, 63(3):641–649.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):199–215.

Hafeman, D. M. and VanderWeele, T. J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, pages 753–764.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statistics & probability letters*, 78(2):144–149.

Imai, K., Keele, L., Yamamoto, T., et al. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 25(1):51–71.

Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171.

Jaworski, P., Durante, F., Hardle, W. K., and Rychlik, T. (2010). *Copula theory and its applications*, volume 198. Springer.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kotelchuck, M. (1994). An evaluation of the kessner adequacy of prenatal care index and a proposed adequacy of prenatal care utilization index. *American journal of public health*, 84(9):1414–1420.

Manski, C. F. (1988). *Analog Estimation Methods in Econometrics: Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Chapman and Hall.

Martin, J. A., Osterman, M., Kirmeyer, S., and Gregory, E. (2015). Measuring gestational age in vital statistics data: transitioning to the obstetric estimate. *National Vital Statistics*

Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 64(5):1–20.

Miles, C. H., Kanki, P., Meloni, S., and Tchetgen, E. J. T. (2015). On partial identification of the pure direct effect. *arXiv preprint arXiv:1509.01652*.

Miles, C. H., Shpitser, I., Kanki, P., Meloni, S., and Tchetgen, E. J. T. (2017). On semi-parametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *arXiv preprint arXiv:1710.02011*.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux expereince agaricales: Essay des principles.(excerpts reprinted and translated to english, 1990). *Statistical Science*, 5:463–472.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2013). Direct and indirect effects. *arXiv preprint arXiv:1301.2300*.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, pages 276–284.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128.

Richardson, T. S., Robins, J. M., and Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, 112(519):1121–1130.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J., Rotnitzky, A., Vansteelandt, S., Have, T. T., Xie, Y., and Murphy, S. (2007). Discussions on" principal stratification designs to estimate input data missing due to death". *Biometrics*, 63(3):650–658.

Robins, J. M. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. *Computation, causation, and discovery*, pages 349–405.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. *Oxford Statistical Science Series*, pages 70–82.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155.

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170.

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035.

Sklar, A. (1959). Fonctions de reprtition an dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris*, 8:229–231.

Steen, J., Loeys, T., Moerkerke, B., and Vansteelandt, S. (2017). Flexible mediation analysis with multiple mediators. *American journal of epidemiology*, 186(2):184–193.

Taguri, M., Featherstone, J., and Cheng, J. (2018). Causal mediation analysis with multiple causally non-ordered mediators. *Statistical methods in medical research*, 27(1):3–19.

Tchetgen, E. J. T. and Phiri, K. (2014). Bounds for pure direct effect. *Epidemiology (Cambridge, Mass.)*, 25(5):775–776.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40(3):1816–1845.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2014). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology (Cambridge, Mass.)*, 25(2):282–291.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59:Issue 5.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters*, 78(17):2957–2962.

VanderWeele, T. J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224–232.

VanderWeele, T. J. and Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, biostatistics, and public health*, 11(2):e9027.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B*, 79(3):917–938.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2(4):457–468.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306.

Vansteelandt, S. and Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258–265.

Vansteelandt, S. and VanderWeele, T. J. (2012). Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics*, 68(4):1019–1027.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Appendix A

APPENDIX FOR CHAPTER 2

To make the proofs more compact, we use subscripts to denote the potential outcomes. For example, $Y(a, m)$ is denoted as Y_{am}.

Proof 1 *Assumption 2.4.3 → 2.4.4*

$$\begin{aligned}&\mathbb{P}(M_{a^\star} = m|X = x) = \mathbb{P}(M_{a^\star} = m|A = a^\star, X = x) \text{ by 2.4.3 assumption 1}\\ =&\mathbb{P}(M = m|A = a^\star, X = x) \text{ by the consistency assumption.}\end{aligned}$$

Hence, $\mathbb{P}(M_{a^\star} = m|X = x)$ *is identifiable.*

$$\begin{aligned}&\mathbb{P}(Y_{am} = y|X = x) = \mathbb{P}(Y_{am} = y|A = a, X = x) \text{ by 2.4.3 assumption 1}\\ =&\mathbb{P}(Y_{am} = y|M = m, A = a, X = x) \text{ by 2.4.3 assumption 2}\\ =&\mathbb{P}(Y|A = a, M = m, X = x) \text{ by the consistency assumption.}\end{aligned}$$

Hence, $\mathbb{P}(Y_{am} = y|X = x)$ *is identifiable.*

$$\begin{aligned}&\mathbb{P}(Y_{am}|M_{a^\star}, X) = \mathbb{P}(Y_{am}|M_{a^\star}, A = a^\star, X) \text{ by 2.4.3 assumption 1}\\ =&\mathbb{P}(Y_{am}|A = a^\star, X) \text{ by 2.4.3 assumption 2}\\ =&\mathbb{P}(Y_{am}|X) \text{ by 2.4.3 assumption 1}\end{aligned}$$

Hence, $Y_{am} \perp\!\!\!\perp M_{a^\star}|X$.

Proof 2 *Assumption 2.4.5 → Assumption 2.4.4*

$$\begin{aligned}&\mathbb{P}(Y_{am}|X) = \mathbb{P}(Y_{am}|M_a = m, X) \text{ by 2.4.5 assumption 2}\\ =&\mathbb{P}(Y_{am}|M_a = m, A = a, X) \text{ by 2.4.5 assumption 1}\\ =&\mathbb{P}(Y|M = m, A = a, X) \text{ by the consistency assumption.}\end{aligned}$$

$$\mathbb{P}(M_a|X) = \mathbb{P}(M_a|A=a,X) \text{ by 2.4.5 assumption 1}$$
$$=\mathbb{P}(M|A=a,X) \text{ by the consistency assumption.}$$

Hence both $\mathbb{P}(M_a = m|X = x)$ *and* $\mathbb{P}(Y_{am} = y|X = x)$ *are identified. The last assumption is identical to assumption 3 in Assumption 2.4.4.*

Proof 3 *Assumption 2.4.6* $\rightarrow$ *Assumption 2.4.4*

$$\mathbb{P}(Y_{am}|X) = \mathbb{P}(Y_{am}|A,X) \text{ by 2.4.6 assumption 1}$$
$$=\mathbb{P}(Y_{am}|M,A,X) \text{ by 2.4.6 assumption 3}$$
$$=\mathbb{P}(Y|M=m,A=a,X) \text{ by the consistency assumption.}$$

$$\mathbb{P}(M_a|X) = \mathbb{P}(M_a|A=a,X) \text{ by 2.4.6 assumption 2}$$
$$=\mathbb{P}(M|A=a,X) \text{ by the consistency assumption.}$$

Hence both $\mathbb{P}(M_a = m|X = x)$ *and* $\mathbb{P}(Y_{am} = y|X = x)$ *are identified. The last assumption is identical to assumption 3 in Assumption 2.4.4.*

Proof 4 *Condition 3 in 2.4.4* $\rightarrow$ *Condition 4 in 2.4.8*

$$\mathbb{E}[Y_{am} - Y_{0m}|M_0 = m, X]$$
$$=\mathbb{E}[Y_{am}|M_0 = m, X] - \mathbb{E}[Y_{0m}|M_0 = m, X]$$
$$=\mathbb{E}[Y_{am}|X] - \mathbb{E}[Y_{0m}|X] = \mathbb{E}[Y_{am} - Y_{0m}|X]$$

Proof 5 *Identification Under Pearl's assumptions*

$$\mathbb{E}[Y_{aM_{a^\star}}] = \mathbb{E}\{\mathbb{E}[Y_{aM_{a^\star}}|X]\} = \mathbb{E}\{\mathbb{E}_{M_{a^\star}}(\mathbb{E}[Y_{aM_{a^\star}}|M_{a^\star} = m, X]|X)\}$$
$$=\mathbb{E}\{\mathbb{E}_{M_{a^\star}}(\mathbb{E}[Y_{am}|M_{a^\star} = m, X]|X)\}$$
$$=\mathbb{E}\{\mathbb{E}_{M_{a^\star}=m}(\mathbb{E}[Y_{am}|X]|X = x)\}$$
$$=\Sigma_{m,x}\mathbb{E}[Y_{am}|x]\mathbb{P}(M_{a^\star} = m|x)\mathbb{P}(x).$$

The fifth equality comes from assumption 3.

Similarly, $\mathbb{E}[Y_{a^\star M_{a^\star}}] = \Sigma_{m,x}\mathbb{E}[Y_{a^\star m}|X=x]\mathbb{P}(M_{a^\star}=m|X=x)\mathbb{P}(x)$.

Hence $\mathbb{E}[Y_{aM_{a^\star}} - Y_{a^\star M_{a^\star}}] = \Sigma_{m,x}\{\mathbb{E}[Y_{am}|x] - \mathbb{E}[Y_{a^\star m}|x]\}\mathbb{P}(M_{a^\star}=m|x)\mathbb{P}(x)$.

Proof 6 *Identification Under Vanderweele's assumptions*

$$\begin{aligned}
&\mathbb{E}[Y_{aM_{a^\star}}|x] = \mathbb{E}[Y_{a1}M_{a^\star} + Y_{a0}(1-M_{a^\star})|x] \\
=&\mathbb{E}[(Y_{a1}-Y_{a0})M_{a^\star})|x] + \mathbb{E}[Y_{a0}|x] \\
=&\mathbb{E}[(Y_{a1}-Y_{a0})|M_{a^\star}=1,x]\mathbb{P}(M_{a^\star}=1|x) + \mathbb{E}[Y_{a0}|A=a,x] \\
=&\mathbb{E}[(Y_{a1}-Y_{a0})|M_{a^\star}=1,A=a^\star,x]\mathbb{P}(M_{a^\star}=1|x) + \mathbb{E}[Y_{a0}|A=a,x] \\
=&\mathbb{E}[(Y_{a1}-Y_{a0})|M=1,A=a^\star,x]\mathbb{P}(M_{a^\star}=1|x) + \mathbb{E}[Y_{a0}|A=a,x] \\
=&\mathbb{E}[(Y_{a1}-Y_{a0})|M=1,A=a,x]\mathbb{P}(M_{a^\star}=1|x) + \mathbb{E}[Y_{a0}|A=a,x].
\end{aligned}$$

The third equation used assumption 1, the fifth equation used assumption 3. By assumption 2, $\mathbb{E}[Y_{m=0,a}|A=a,M=1,x] = \mathbb{E}[Y_{m=0,a}|A=a,M=0,x]$,

$$\begin{aligned}
&\mathbb{E}[Y_{a0}|a,x] = \mathbb{E}[Y_{a0}|a,M=1,x]\mathbb{P}(M=1|a,x) + \mathbb{E}[Y_{a0}|a,M=0,x]\mathbb{P}(M=0|a,x) \\
=&\mathbb{E}[Y_{a0}|a,M=1,x] = \mathbb{E}[Y_{a0}|a,M=0,x].
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}[Y_{aM_{a^\star}}|x] \\
=&\mathbb{E}[(Y_{a1}-Y_{a0})|M=1,A=a,x]\mathbb{P}(M_{a^\star}=1|x) + \\
&\mathbb{E}[Y_{a0}|a,M=1,x]\mathbb{P}(M_{a^\star}=1|x) + \mathbb{E}[Y_{a0}|a,M=0,x]\mathbb{P}(M_{a^\star}=0|x) \\
=&\mathbb{E}[Y_{a1}|M=1,A=a,x]\mathbb{P}(M_{a^\star}=1|x) + \mathbb{E}[Y_{a0}|a,M=0,x]\mathbb{P}(M_{a^\star}=0|x) \\
=&\mathbb{E}[Y|M=1,A=a,x]\mathbb{P}(M_{a^\star}=1|a^\star,x) + \mathbb{E}[Y|a,M=0,x]\mathbb{P}(M_{a^\star}=0|a^\star,x) \\
=&\mathbb{E}[Y|M=1,A=a,x]\mathbb{P}(M=1|a^\star,x) + \mathbb{E}[Y|a,M=0,x]\mathbb{P}(M=0|a^\star,x).
\end{aligned}$$

The third equation used assumption 1.

$$\begin{aligned}
&\mathbb{E}[Y_{a^\star M_{a^\star}}|x] \\
=&\mathbb{E}[Y_{a^\star 1}M_{a^\star} + Y_{a^\star 0}(1-M_{a^\star})|x] \\
=&\mathbb{E}[Y_{a^\star 1}|M_{a^\star}=1,x]\mathbb{P}(M_{a^\star}=1|x) + \mathbb{E}[Y_{a^\star 0}|M_{a^\star}=0,x]\mathbb{P}(M_{a^\star}=0|x) \\
=&\mathbb{E}[Y_{a^\star 1}|M_{a^\star}=1,a^\star x]\mathbb{P}(M_{a^\star}=1|a^\star,x) + \mathbb{E}[Y_{a^\star 0}|M_{a^\star}=0,a^\star,x]\mathbb{P}(M_{a^\star}=0|a^\star,x) \\
=&\mathbb{E}[Y|M=1,a^\star,x]\mathbb{P}(M=1|a^\star,x) + \mathbb{E}[Y|M=0,a^\star,x]\mathbb{P}(M=0|a^\star,x).
\end{aligned}$$

The third equation used assumption 1.

Hence,

$$\begin{aligned}
&\mathbb{E}[Y_{aM_{a^\star}} - Y_{a^\star M_{a^\star}}] = \Sigma_x[\mathbb{E}[Y_{aM_{a^\star}}|x] - \mathbb{E}[Y_{a^\star M_{a^\star}}|x]]\mathbb{P}(x) \\
=&\Sigma_x[\mathbb{E}[Y|M=1,a,x]\mathbb{P}(M=1|a^\star,x) + \mathbb{E}[Y|a,M=0,x]\mathbb{P}(M=0|a^\star,x) - \\
&\mathbb{E}[Y|M=1,a^\star,x]\mathbb{P}(M=1|a^\star,x) - \mathbb{E}[Y|M=0,a^\star,x]\mathbb{P}(M=0|a^\star,x)]\mathbb{P}(x).
\end{aligned}$$

Proof 7 *Identification Under Petersen, Sinisi and Van der Laan's assumptions*

$$\begin{aligned}
&\mathbb{E}[Y_{aM_0}|x] - \mathbb{E}[Y_{0M_0}|x] = \mathbb{E}_{M_0=m}[\mathbb{E}[(Y_{am}-Y_{0m})|M_0=m,x]|x] \\
=&\mathbb{E}_{M_{0=m}}[\mathbb{E}[(Y_{am}-Y_{0m})|x]|x] \\
=&\Sigma_{x,m}\mathbb{E}[(Y_{am}-Y_{0m})|x]\mathbb{P}(M_0=m|x)\mathbb{P}(x) \\
=&\Sigma_{x,m}\mathbb{E}[Y_{am}|x]\mathbb{P}(M_0=m|x)\mathbb{P}(x) - \Sigma_{x,m}\mathbb{E}[Y_{0m}|x]\mathbb{P}(M_0=m|x)\mathbb{P}(x) \\
=&\Sigma_{x,m}\mathbb{E}[Y_{am}|a,x]\mathbb{P}(M_0=m|a,x)\mathbb{P}(x) - \Sigma_{x,m}\mathbb{E}[Y_{0m}|A=0,x]\mathbb{P}(M_0=m|A=0,x)\mathbb{P}(x) \\
=&\Sigma_{x,m}\mathbb{E}[Y_{am}|m,a,x]\mathbb{P}(M_0=m|a,x)\mathbb{P}(x) - \Sigma_{x,m}\mathbb{E}[Y_{0m}|m,A=0,x]\mathbb{P}(M_0=m|A=0,x)\mathbb{P}(x) \\
=&\Sigma_{x,m}\mathbb{E}[Y|m,a,x]\mathbb{P}(M=m|a,x)\mathbb{P}(x) - \Sigma_{x,m}\mathbb{E}[Y|m,A=0,x]\mathbb{P}(M=m|A=0,x)\mathbb{P}(x)
\end{aligned}$$

The second equation used assumption 4. The fourth equation used assumption 1 and 2. The fifth equation used assumption 3.

Proof 8 *Identification Under Robins' No-interaction assumptions*

$$\begin{aligned}
&\mathbb{E}[Y_{1M_0} - Y_{0M_0}] = \mathbb{E}_{X,C}\mathbb{E}[Y_{1M_0} - Y_{0M_0}|X, C] \\
=&\mathbb{E}_{X,C}\mathbb{E}_{M_0|X,C}\mathbb{E}[Y_{1M_0} - Y_{0M_0}|M_0 = m, X, C] \\
=&\mathbb{E}_{X,C}\mathbb{E}_{M_0|X,C}\mathbb{E}[Y_{1m} - Y_{0m}|M_0 = m, X, C] \\
=&\mathbb{E}_{X,C}\mathbb{E}_{M_0|X,C}\mathbb{E}[B|M_0 = m, X, C]
\end{aligned}$$

Since B is a random variable that is independent of the realization of the mediator m, the expression becomes

$$\begin{aligned}
&\mathbb{E}_{X,C}\mathbb{E}[B|X, C] \\
=&\mathbb{E}_X\mathbb{E}[Y_{1m} - Y_{0m}|X] \\
=&\mathbb{E}_X\mathbb{E}[Y_{1m}|A = 1, X] - \mathbb{E}_X\mathbb{E}[Y_{0m}|A = 0, X] \\
=&\Sigma_c\{\mathbb{E}_X\mathbb{E}[Y_{1m}|A = 1, x, c]\mathbb{P}(C = c|A = 1, x) - \mathbb{E}_X\mathbb{E}[Y_{0m}|A = 0, x, c]\mathbb{P}(C = c|A = 0, x)\} \\
=&\Sigma_c\{\mathbb{E}_X\mathbb{E}[Y_{1m}|m, A = 1, x, c]\mathbb{P}(C = c|A = 1, x) - \mathbb{E}_X\mathbb{E}[Y_{0m}|m, A = 0, x, c]\mathbb{P}(C = c|A = 0, x)\} \\
=&\Sigma_c\{\mathbb{E}_X\mathbb{E}[Y|m, A = 1, x, c]\mathbb{P}(C = c|A = 1, x) - \mathbb{E}_X\mathbb{E}[Y|m, A = 0, x, c]\mathbb{P}(C = c|A = 0, x)\}.
\end{aligned}$$

The second equation used assumption 1 and the fourth equation used assumption 2. The proof under MCM is similar for the first part, the second part (identification of $\mathbb{E}_{X,C}\mathbb{E}[B|X, C]$) is given in proof 9.

Proof 9 *Identification of CDE under MCM assumptions*

$$
\begin{aligned}
&\mathbb{E}[Y(1,m) - Y(0,m)] \\
=&\mathbb{E}_X\mathbb{E}[Y(1,m) - Y(0,m)|X] \\
=&\mathbb{E}_X\mathbb{E}[Y(1,m)|A=1,X] - \mathbb{E}_X\mathbb{E}[Y(0,m)|A=0,X] \\
=&\mathbb{E}_X\mathbb{E}_{C|A=1,X}\mathbb{E}[Y(1,m)|A=1,X,C] - \mathbb{E}_X\mathbb{E}_{C|A=0,X}\mathbb{E}[Y(0,m)|A=0,X,C] \\
=&\mathbb{E}_X\mathbb{E}_{C|A=1,X,C}\mathbb{E}[Y(1,m)|M(1)=m,A=1,X] - \mathbb{E}_X\mathbb{E}_{C|A=0,X}\mathbb{E}[Y(0,m)|M(0)=m,A=0,X,C] \\
=&\mathbb{E}_X\mathbb{E}_{C|A=1,X}\mathbb{E}[Y|M=m,A=1,X,C] - \mathbb{E}_X\mathbb{E}_{C|A=0,X}\mathbb{E}[Y|M=m,A=0,X,C].
\end{aligned}
$$

Appendix B

APPENDIX FOR CHAPTER 3

Proof for Theorem 3.2.1

$$E(Y_{1M_0} - Y_{0M_0} \mid X) = E(Y_{1C_1M_0} - Y_{0C_0M_0} \mid X) = E_{M_0=m|X}E(Y_{1C_1m} - Y_{0C_0m} \mid M_0 = m, X)$$
$$= E_{M_0=m|X}E(Y_{1C_1m} - Y_{0C_1m} \mid M_0 = m, X) + E_{M_0=m|X}E(Y_{0C_1m} - Y_{0C_0m} \mid M_0 = m, X).$$

Under Assumption 3.2.3, the first term can be written as

$$E_{M_0=m|X}E(Y_{1C_1m} - Y_{0C_1m} \mid X) = E_{M_0=m|X}E_{C_1=c_1|X}\{E(Y_{1c_1m} \mid c_1, X) - Y_{0c_1m} \mid c_1, X)\}$$

and under Assumption 3.2.4, the second term can be written as:

$$E_{M_0=m|X}E(Y_{0C_1m} - Y_{0C_0m} \mid X) =$$
$$E_{M_0=m|X}E_{C_1=c_1|X}E(Y_{0c_1m} \mid c_1, X) - E_{M_0=m|X}E_{C_0=c_0|X}E(Y_{0c_0m} \mid c_0, X).$$

Hence, the sum of (1) and (2) can be written as:

$$E_{M_0=m|X}E_{C_1=c_1|X}E(Y_{1c_1m} \mid C_1 = c_1, X) - E_{M_0=m|X}E_{C_0=c_0|X}E(Y_{0c_0m} \mid C_0 = c_0, X),$$

which can be identified as follows under Assumption 3.2.1 and Assumption 3.2.2:

$$E_{M_0=m|A=0,X}E_{C_1=c_1|A=1,X}E(Y_{1c_1m} \mid C_1 = c_1, A = 1, M_1 = m, X)$$
$$- E_{M_0=m|A=0,X}E_{C_0=c_0|A=0,X}E(Y_{0c_0m} \mid C_0 = c_0, A = 0, M_0 = m, X)$$
$$= E_{M=m|A=0,X}E_{C=c_1|A=1,X}E(Y \mid C = c_1, A = 1, M = m, X)$$
$$- E_{M=m|A=0,X}E_{C=c_0|A=0,X}E(Y \mid C = c_0, A = 0, M = m, X).$$

Proof for Theorem 3.2

The parameter of interest is: $\Delta = \delta_1 - \delta_0$, where

$$\delta_1 \equiv \iiint E(Y \mid A=1, M=m, C=c, X) f_{M|A,X}(m \mid A=0, X) f_{C|A,X}(c \mid A=1, X) dm\ dc\ dx,$$
$$\delta_0 \equiv \iiint E(Y \mid A=0, M=m, C=c, X) f_{M|A,X}(m \mid A=0, X) f_{C|A,X}(c \mid A=0, X) dm\ dc\ dx.$$

For simplicity, δ_1 and δ_0 are abbreviated to:

$$\delta_1 \equiv \iiint E(Y \mid 1, m, c, X) f(m \mid 0, X) f(c \mid 1, X) dm\ dc\ dx,$$
$$\delta_0 \equiv \iiint E(Y \mid 0, m, c, X) f(m \mid 0, X) f(c \mid 0, X) dm\ dc\ dx.$$

Denote the likelihood of $Z \equiv (Y, A, M, C, X)$ as $f(Y, A, M, C, X)$, and consider the one-dimensional parametric submodel

$$\begin{aligned} & f_t(A, C, M, Y, X) \\ =& f_t(Y \mid A, M, C, X) f_t(M \mid A, C, X) f_t(C \mid A, X) f_t(A \mid X) f_t(X), \end{aligned}$$

where $f_0 = f$. To simplify notation, we only show the results for continuous C and M with a density. By the definition of the copula density, the likelihood can be factorized differently as follows:

$$\begin{aligned} & f_t(A, C, M, Y, X) \\ =& f_t(Y \mid A, M, C, X) \boldsymbol{c}_t(F_{tM|A,X}, F_{tC|A,X}) f_t(M \mid A, X) f_t(C \mid A, X) f_t(A \mid X) f_t(X). \end{aligned}$$

Denote its score function as $S_t(A, C, M, Y, X)$, which can be written as a sum of scores with respect to conditional densities. First we look at δ_1. The influence function $\phi(Z)$ of any regular asymptotically linear (RAL) estimator of δ_1 satisfies:

$$E[\phi_1(Z) S_{t=0}(Z)] = \frac{\partial \delta_1(t)}{\partial t}\Big|_{t=0}$$

Take derivatives with respect to t at $t = 0$:

$$\frac{\partial \delta_1(t)}{\partial t}\Big|_{t=0} = \iiint \nabla_{t=0} E_t(Y \mid 1, m, c, x) f(m \mid 0, x) f(c \mid 1, x) f(x) dm dc dx \tag{B.1}$$

$$+ \iiint E(Y \mid 1, m, c, x) \nabla_{t=0} f_t(m \mid 0, x) f(c \mid 1, x) f(x) dm dc dx \tag{B.2}$$

$$+ \iiint E(Y \mid 1, m, c, x) f(m \mid 0, x) \nabla_{t=0} f_t(c \mid 1, x) f(x) dm dc dx \tag{B.3}$$

$$+ \iiint E(Y \mid 1, m, c, x) f(m \mid 0, x) f(c \mid 1, x) \nabla_{t=0} f_t(x) dm dc dx. \tag{B.4}$$

where $(B.1)$

$$= \iiint \nabla_{t=0} E_t(Y \mid 1, m, c, x) f(m \mid 0, x) f(c \mid 1, x) f(x) dm dc dx$$

$$= \iiint \nabla_{t=0} E_t(Y \mid 1, m, c, x) \frac{f(m \mid 0, x)}{f(m \mid 1, x) \boldsymbol{c}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)} f(m, c \mid 1, x) f(x) dm dc dx$$

$$= E[S_{t=0}(Y \mid A, M, C, X) \frac{A}{f(A \mid X)} \frac{f(M \mid 0, X)}{f(M \mid A, X) \boldsymbol{c}(F_{M|A,X}(M), F_{C|A,X}(C) \mid A, X)} Y]$$

$$= E[S_{t=0}(Y \mid A, M, C, X) \frac{A}{f(A \mid X)} \frac{f(M \mid 0, X)}{f(M \mid A, X) \boldsymbol{c}(F_{M|A,X}(M), F_{C|A,X}(C) \mid A, X)} (Y - E[Y \mid A, M, C, X])]$$

$$= E[S_{t=0}(Y, A, M, C, X) \frac{A}{f(A \mid X)} \frac{f(M \mid A = 0, X)}{f(M \mid A, X) \boldsymbol{c}(F_{M|A,X}(M), F_{C|A,X}(C) \mid A, X)} (Y - E[Y \mid A, M, C, X])].$$

The second to last equation results from the property of the score function of the conditional density for Y. The last equation results from $S_{t=0}(A, M, C, X)$ being a function of (A, M, C, X), and $E[g(A, M, C, X)\{Y - E(Y \mid A, M, C, X)\}] = 0$ for any function g. By similar arguments, we have

$$(B.2) = \iiint E(Y \mid 1, m, c, x) \nabla_{t=0} f(m \mid 0, x) f_t(c \mid 1, x) f(x) dm dc dx$$

$$= E\{S_{t=0}(M \mid A, X) \frac{1 - A}{f(A \mid X)} \int E(Y \mid 1, M, c, X) f(c \mid 1, X) dc\}$$

$$= E[S_{t=0}(A, C, M, Y, X) \frac{1 - A}{f(A \mid X)} \{\gamma_{M,X}(1) - \tau_X(1)\}],$$

$$
\begin{aligned}
(B.3) =& \iiint E(Y \mid 1, m, c, x) f(m \mid 0, x) \nabla_{t=0} f_t(c \mid 1, x) f(x) dm dc dx \\
=& E\{S_{t=0}(C \mid A, X) \frac{A}{f(A \mid X)} \int E(Y \mid A, m, C, X) f(m \mid 0, X) dm\} \\
=& E[S_{t=0}(A, C, M, Y, X) \frac{A}{f(A \mid X)} \{\int E(Y \mid A, m, C, X) f(m \mid 0, X) dm - \tau_X(1)\}], \\
(B.4) =& \iiint E(Y \mid 1, m, c, x) f(m \mid 0, x) f(c \mid 1, x) \nabla_{t=0} f_t(x) dm dc dx \\
=& E[S_{t=0}(X) \tau_X(1)] = E[S_{t=0}(X)(\tau_X(1) - \delta_1)] \\
=& E[S_{t=0}(A, C, M, Y, X)(\tau_X(1) - \delta_1)].
\end{aligned}
$$

We therefore find and influence function for δ_1:

$$
\begin{aligned}
\phi_1(Z) =& \frac{A}{f(A \mid X)} \frac{f(M \mid A = 0, X)}{f(M \mid A, X) \boldsymbol{c}(F_{M|A,X}(M), F_{C|A,X}(C) \mid A, X)} (Y - E[Y \mid A, M, C, X]) + \\
& \frac{1 - A}{f(A \mid X)} \{\gamma_{M,X}(1) - \tau_X(1)\} + \\
& \frac{A}{f(A \mid X)} \{\int E(Y \mid A, m, C, X) f(m \mid 0, X) dm - \tau_X(1)\} + \\
& (\tau_X(1) - \delta_1).
\end{aligned}
$$

An influence function $\phi_0(Z)$ for δ_0 can be constructed similarly and the derivation is omitted. Their difference $\phi_1(Z) - \phi_0(Z)$ is then an influence function for Δ. It is the only influence function (therefore efficient) in $\mathcal{M}_{non}$ of a RAL estimator of Δ. It takes the following form:

$$
\begin{aligned}
S_{\Delta}^{\text{eff}} =& \frac{2A - 1}{f_{A|X}(A \mid X)} \frac{f_{M|A=0,X}(M \mid A = 0, X)}{f_{M|A,C,X}(M \mid A, C, X)} \{Y - E(Y \mid A, M, C, X)\} - \frac{2A - 1}{f_{A|X}(A \mid X)} \tau_X(A) + \\
& \frac{1 - A}{f_{A|X}(A = 0 \mid X)} \{\gamma_{M,X}(1) - \gamma_{M,X}(0)\} + \left\{1 - \frac{1 - A}{f_{A|X}(A = 0 \mid X)}\right\} \{\tau_X(1) - \tau_X(0)\} + \\
& \frac{2A - 1}{f_{A|X}(A \mid X)} \eta_{C,X}(A) - \Delta.
\end{aligned}
$$

B.1 Proof for Theorem 3.1 and 3.3

We prove Theorem 3.3, and Theorem 3.1 will follow from similar arguments. Again, for notational simplicity, we consider continuous M and C for which the joint density exists.

The proof for the general case is similar. We denote the incorrectly specified components with a superscript $\star$.

First we prove that $E\{S_{\Delta}^{\text{eff}}(\Delta^*_{quad})\} = 0$ under four misspecification conditions: $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$.

- $\mathcal{M}_1$: $f(A \mid X)$, $f(C \mid A, X)$, $f(M \mid A, X)$, $\boldsymbol{c}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ are correctly specified, and $E(Y \mid M, A, C, X)$ is mis-specified by a parametric model $E^{*par}[Y \mid A, M, C, X]$:

$$
\begin{aligned}
& E\{S_{\Delta}^{\text{eff}}(f_{A|X}, f_{C|A,X}, f_{M|A,X}, \boldsymbol{c}_{M,C|A,X})\} \\
=& E\{\frac{2A-1}{f(A \mid X)} \frac{f(M \mid A=0, X)}{f(M \mid A, X)\boldsymbol{c}(F_{M|A,X}(M), F_{C|A,X}(C))}(Y - E^{*par}[Y \mid A, M, C, X]) + \\
& \frac{2A-1}{f(A \mid X)}\eta^{*par}_{C,X}(A) - \frac{2A-1}{f(A \mid X)}\tau^{*par}_{X}(A) + \frac{1-A}{f(A \mid X)}\{\gamma^{*par}_{M,X}(1) - \gamma^{*par}_{M,X}(0)\} \\
& + (1 - \frac{1-A}{f(A \mid X)})\{\tau^{*par}_{X}(1) - \tau^{*par}_{X}(0)\} - \Delta\} \\
=& E\{\frac{2A-1}{f(A \mid X)} \frac{f(M \mid A=0, X)}{f(M \mid A, X)\boldsymbol{c}(F_{M|A,X}(M), F_{C|A,X}(C))} Y\} - E\{\frac{2A-1}{f(A \mid X)}\eta^{*par}_{C,X}(A)\} + \\
& E\{\frac{2A-1}{f(A \mid X)}\eta^{*par}_{C,X}(A)\} - E\{\frac{2A-1}{f(A \mid X)}\tau^{*par}_{X}(A)\} + E\{\frac{1-A}{f(A \mid X)}\{\gamma^{*par}_{M,X}(1) - \gamma^{*par}_{M,X}(0)\}\} \\
& + E\{(1 - \frac{1-A}{f(A \mid X)})\{\tau^{*par}_{X}(1) - \tau^{*par}_{X}(0)\}\} - \Delta = 0, \\
& \text{for } E\{\frac{2A-1}{f(A \mid X)}\tau^{*par}_{X}(A)\} = E\{\tau^{*par}_{X}(1) - \tau^{*par}_{X}(0)\}, \\
& \text{and } E\{\frac{2A-1}{f(A \mid X)} \frac{f(M \mid A=0, X)}{f(M \mid A, X)\boldsymbol{c}(F_{M|A,X}(M), F_{C|A,X}(C))} Y\} = \Delta.
\end{aligned}
$$

- $\mathcal{M}_2$: $f(A \mid X)$, $f(M \mid A, X)$, $E(Y \mid A, M, C, X)$ are correctly specified, and $f(C \mid$

$A, X)$ and $\boldsymbol{c}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ are mis-specified by $f^{\star\text{par}}(C \mid A, X)$ and $\boldsymbol{c}^{*par}$:

$$
\begin{aligned}
&E\{S_{\Delta}^{\text{eff}}(f_{A|X}, f_{C|A,X}^{*par}, f_{M|A,X}, \boldsymbol{c}^{*par}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X), E_{Y|M,A,C,X})\} \\
=&E\{\frac{2A-1}{f(A \mid X)} \frac{f(M \mid A=0, X)}{f(M \mid A, X)\boldsymbol{c}^{*par}(F_{M|A,X}(m), F_{C|A,X}^{*par}(c) \mid A, X)}(Y - E[Y \mid A, M, C, X]) + \\
&\frac{2A-1}{f(A \mid X)}\eta_{C,X}(A) - \frac{2A-1}{f(A \mid X)}\tau_X^{*par}(A) + \frac{1-A}{f(A \mid X)}\{\gamma_{M,X}^{*par}(1) - \gamma_{M,X}^{*par}(0)\} \\
&+ (1 - \frac{1-A}{f(A \mid X)})\{\tau_X^{*par}(1) - \tau_X^{*par}(0)\} - \Delta\} = 0,
\end{aligned}
$$

for $E\{\frac{2A-1}{f(A \mid X)}\eta_{C,X}(A)\} = \Delta$.

- $\mathcal{M}_3$: $f(A \mid X)$, $f(C \mid A, X)$, $E(Y \mid A, M, C, X)$ are correctly specified, and $f(M \mid A, X)$and $\boldsymbol{c}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ are mis-specified by $f^{\star\text{par}}(M \mid A, X)$ and $\boldsymbol{c}^{*par}$:

$$
\begin{aligned}
&E\{S_{\Delta}^{\text{eff}}(f_{A|X}, f_{C|A,X}, f_{M|A,X}^{*par}, \boldsymbol{c}^{*par}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X), E_{Y|M,A,C,X})\} \\
=&E\{\frac{2A-1}{f(A \mid X)} \frac{f^{*par}(M \mid A=0, X)}{f^{*par}(M \mid A, X)\boldsymbol{c}^{*par}(F_{M|A,X}^{*par}(m), F_{C|A,X}(c) \mid A, X)}(Y - E[Y \mid A, M, C, X]) + \\
&\frac{2A-1}{f(A \mid X)}\eta_{C,X}^{*par}(A) - \frac{2A-1}{f(A \mid X)}\tau_X^{*par}(A) + \frac{1-A}{f(A \mid X)}\{\gamma_{M,X}(1) - \gamma_{M,X}(0)\} \\
&+ (1 - \frac{1-A}{f(A \mid X)})\{\tau_X^{*par}(1) - \tau_X^{*par}(0)\} - \Delta\} = 0,
\end{aligned}
$$

for $E\{\frac{1-A}{f(A \mid X)}\{\gamma_{M,X}(1) - \gamma_{M,X}(0)\}\} = \Delta$.

- $\mathcal{M}_4$: $f(M \mid A=0, X)$, $f(C \mid A, X)$, $E(Y \mid A, M, C, X)$ are correctly specified, $f(A \mid X)$ and $\boldsymbol{c}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ are mis-specified by $f^{\star\text{par}}(A \mid X)$ and $\boldsymbol{c}^{*par}$:

$$
\begin{aligned}
&E\{S_{\Delta}^{\text{eff}}(f_{A|X}^{*par}, f_{C|A,X}, f_{M|A,X}, \boldsymbol{c}^{*par}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X))\} \\
=&E\{\frac{2A-1}{f^{*par}(A \mid X)} \frac{f(M \mid A=0, X)}{f(M \mid A, X)\boldsymbol{c}^{*par}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)}\}(Y - E[Y \mid A, M, C, X]) + \\
&\frac{2A-1}{f^{*par}(A \mid X)}\eta_{C,X}(A) - \frac{2A-1}{f^{*par}(A \mid X)}\tau_X(A) + \frac{1-A}{f^{*par}(A \mid X)}\{\gamma_{M,X}(1) - \gamma_{M,X}(0)\} \\
&+ (1 - \frac{1-A}{f^{*par}(A \mid X)})\{\tau_X(1) - \tau_X(0)\} - \Delta\} = 0,
\end{aligned}
$$

for $E\{\tau_X(1) - \tau_X(0)\} = \Delta$ by definition.

Under suitable regularity conditions (Manski, 1988) hold for S_{Δ}^{eff} and the score of the parameters in the nuisance models. By a Taylor expansion we have (denoting the probability limits of the nuisance parameters as θ^*, and its score as $S_\theta(\theta)$):

$$\sqrt{n}(\hat{\Delta}_{quad} - \Delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_{\Delta}^{\text{intersection}}(O_i, \Delta, \theta^*) + o_p(1),$$

where O_i represents the data pertaining to the i^{th} subject, and

$$\begin{aligned} & S_{\Delta}^{\text{intersection}}(\Delta, \theta^*) \\ =& S_{\Delta}^{\text{eff}}(\Delta, \theta^*) - \frac{\partial E[S_{\Delta}^{\text{eff}}(\Delta, \theta^*)]}{\partial \theta^T} E^{-1}[\frac{\partial S_\theta(\theta^*)}{\partial \theta^T}] S_\theta(\theta^*). \end{aligned}$$

The asymptotic normality of the estimator can be proven directly using the Central Limit Theorem (and Slutsky's Theorem).

Appendix C

APPENDIX FOR CHAPTER 4

C.1 Proof for Theorem 3.1

We prove the identification for $EIE_{M^{(1)}}$, and the proof is similar for $EIE_{M^{(2)}}$.

$$
\begin{aligned}
&E(Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}}) \\
=&E_X E(Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}} \mid X) \\
=&E_X E_{M^{(2)}=m^{(2)}|A=1,X} E(Y_{1M_1^{(1)}m^{(2)}} - Y_{1M_0^{(1)}m^{(2)}} \mid M^{(2)}=m^{(2)}, X) \\
=&E_X E_{M^{(2)}=m^{(2)}|A=1,X} E(Y_{1M_1^{(1)}m^{(2)}} - Y_{1M_0^{(1)}m^{(2)}} \mid X) \\
=&E_X E_{M^{(2)}=m^{(2)}|A=1,X} E(Y_{1M_1^{(1)}m^{(2)}} \mid X) - E_X E_{M^{(2)}=m^{(2)}|A=1,X} E(Y_{1M_0^{(1)}m^{(2)}} \mid X) \\
=&E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M_1^{(1)}=m^{(1)}|X} E(Y_{1m^{(1)}m^{(2)}} \mid M_1^{(1)}=m^{(1)}, X) \\
&- E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M_0^{(1)}=m^{(1)\prime}|X} E(Y_{1m^{(1)\prime}m^{(2)}} \mid M_0^{(1)}=m^{(1)\prime}, X) \\
=&E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M^{(1)}=m^{(1)}|A=1,X} E(Y \mid A=1, M^{(1)}=m^{(1)}, M^{(2)}=m^{(2)}, X) \\
&- E_X E_{M^{(2)}=m^{(2)}|A=1,X} E_{M^{(1)}=m^{(1)\prime}|A=0,X} E(Y \mid A=1, M^{(1)}=m^{(1)\prime}, M^{(2)}=m^{(2)}, X)
\end{aligned}
$$

The third equality is implied by assumption **Set** II, and the rest are implied by assumption **Set** I and iterative expectation.

C.2 Proof for Theorem 3.2

By **Theorem 3.1**, $\Delta^{M^{(j)}} = \Delta_4^{M^{(j)}}$. Moreover,

- $\Delta_1^{M^{(j)}} = \Delta_4^{M^{(j)}}$:

$$
\begin{aligned}
&\Delta_1^{M^{(j)}} \\
=&E\left(\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid A=1, X)-f(M^{(j)} \mid A=0, X)}{f(M^{(j)} \mid A=1, M^{(3-j)}, X)} Y\right) \\
=&\iiiint \int \frac{a}{f(a \mid x)} \frac{f(m^{(j)} \mid A=1, x)-f(m^{(j)} \mid A=0, x)}{f(m^{(j)} \mid A=1, m^{(3-j)}, x)} y f(y, a, m^{(j)}, m^{(3-j)}, x) dy\ da\ dm^{(j)}\ dm^{(3-j)}\ dx \\
=&E\left(\iint\{f(m^{(j)} \mid 1, X)-f(m^{(j)} \mid 0, X)\} f(m^{(3-j)} \mid 1, X) E(Y \mid 1, m^{(j)}, m^{(3-j)}, X) dm^{(j)}\ dm^{(3-j)}\right) \\
=&\Delta_4^{M^{(j)}}.
\end{aligned}
$$

- $\Delta_2^{M^{(j)}} = \Delta_4^{M^{(j)}}$:

$$
\begin{aligned}
\Delta_2^{M^{(j)}} =&E\left(\frac{A}{f(A \mid X)}\{\eta_j(1, M^{(3-j)}, X)-\eta_j(0, M^{(3-j)}, X)\}\right) \\
=&\iint \frac{a}{f(a \mid x)}\{\eta_j(1, m^{(3-j)}, x)-\eta_j(0, m^{(3-j)}, x)\} f(a, m^{(3-j)}, x) da\ dm^{(3-j)}\ dx \\
=&\iint\{\eta_j(1, m^{(3-j)}, x)-\eta_j(0, m^{(3-j)}, x)\} f(m^{(3-j)} \mid A=1, x) da\ dm^{(3-j)}\ dx = \Delta_4^{M^{(j)}}.
\end{aligned}
$$

- $\Delta_3^{M^{(j)}} = \Delta_4^{M^{(j)}}$:

$$
\begin{aligned}
\Delta_3^{M^{(j)}} =&E\left(\frac{2A-1}{f(A \mid X)} \eta_{3-j}(1, M^{(j)}, X)\right) \\
=&\iint \frac{2a-1}{f(a \mid X)} \eta_{3-j}(1, m^{(j)}, x) f(a, m^{(j)}, x) da\ dm^{(j)}\ dx \\
=&\iint (2a-1) \eta_{3-j}(1, m^{(j)}, X) f(M^{(j)} \mid a, x) da\ dm^{(j)}\ dx \\
=&\iint \eta_{3-j}(1, m^{(j)}, x) f(M^{(j)} \mid A=1, x) dm^{(j)}\ dx - \iint \eta_{3-j}(1, m^{(j)}, x) f(M^{(j)} \mid A=0, x) dm^{(j)} dx = \Delta_4^{M^{(j)}}.
\end{aligned}
$$

Hence $\Delta^{M^{(j)}} = \Delta_1^{M^{(j)}} = \Delta_2^{M^{(j)}} = \Delta_3^{M^{(j)}} = \Delta_4^{M^{(j)}}$.

C.3 Proof for Theorem 3.3

We prove Theorem 3.3 for continuous mediators. The proofs for other cases are similar. Denote $\int \gamma_j(a) f(x) dx$ as $\theta_a^{(j)}$, then $\Delta^{M^{(j)}} = \theta_1^{(j)} - \theta_0^{(j)}$. Denote the observed data $(Y, A, M^{(1)}, M^{(2)}, X)$ as $\mathcal{O}$. An influence function $\phi_j(\mathcal{O})$ of a regular and asymptotically linear (RAL) estimator of $\theta_a^{(j)}$ satisfies:

$$
E\left[\phi_j(\mathcal{O}) S_{t=0}(\mathcal{O})\right] = \frac{\partial \theta_a^{(j)}(t)}{\partial t}|_{t=0},
$$

where $S_{t=0}(\mathcal{O})$ is the score function of the likelihood for a one-dimensional parametric submodel $f_t(\mathcal{O})$, which satisfies $f_0 = f$, with the parameter t set to 0.
The right hand side of the equation can be written as:

$$\iiint \nabla_{t=0} E_t(Y \mid A=1, m^{(3-j)}, m^{(j)}, X)|_{t=0} f(m^{(j)} \mid A=a, X) f(m^{(3-j)} \mid A=1, X) f(X) dm^{(j)}\ dm^{(3-j)}\ dx \tag{C.1}$$

$$+ \iiint E(Y \mid A=1, m^{(3-j)}, m^{(j)}, X) \nabla_{t=0} f_t(m^{(j)} \mid A=a, X)|_{t=0} f(m^{(3-j)} \mid A=1, X) f(X) dm^{(j)}\ dm^{(3-j)}\ dx \tag{C.2}$$

$$+ \iiint E(Y \mid A=1, m^{(3-j)}, m^{(j)}, X) f(m^{(j)} \mid A=a, X) \nabla_{t=0} f_t(m^{(3-j)} \mid A=1, X)|_{t=0} f(X) dm^{(j)}\ dm^{(3-j)}\ dx \tag{C.3}$$

$$+ \iiint E(Y \mid A=1, m^{(3-j)}, m^{(j)}, X) f(m^{(j)} \mid A=a, X) f(m^{(3-j)} \mid A=1, X) \nabla_{t=0} f_t(X)|_{t=0} dm^{(j)}\ dm^{(3-j)}\ dx. \tag{C.4}$$

We utilize the following facts about scores in the proof:

1. $E[g(X)(Y - E[Y \mid X])] = 0$ for all g,

2. $E[S(Y \mid X) g(X)] = 0$ for all g,

3. $S(Y, X) = S(Y \mid X) + S(X)$.

Define the following term to simplify the notation:

$$R_a^{(j)}(m^{(j)}, m^{(3-j)}, X) \equiv \frac{f(m^{(j)} \mid A=a, X)}{f(m^{(j)} \mid A=1, X) \boldsymbol{c}(F_{m^{(j)}|A=1,X}(m^{(j)}), F_{m^{(3-j)}|A=1,X}(m^{(3-j)}) \mid A=1, X)}.$$

We look at each term of the expression of $\theta_a^{(j)}$:

$$
\begin{aligned}
(1) &= \iiint \nabla_{t=0} E_t(Y \mid A=1, m^{(3-j)}, m^{(j)}, X)|_{t=0} R_a^{(j)}(m^{(j)}, m^{(3-j)}, X) f(m^{(3-j)} \mid A=1, X) f(X) dm^{(j)} dm^{(3-j)} dx \\
&= E[S_{t=0}(Y \mid A, M^{(3-j)}, M^{(j)}, X) \frac{A}{f(A \mid X)} R_a^{(j)}(M^{(j)}, M^{(3-j)}, X) Y] \\
&= E[S_{t=0}(Y, A, M^{(3-j)}, M^{(j)}, X) \frac{A}{f(A \mid X)} R_a^{(j)}(M^{(j)}, M^{(3-j)}, X)(Y - E[Y \mid A=1, M^{(3-j)}, M^{(j)}, X])], \\
(2) &= E[S_{t=0}(M^{(j)} \mid A, X) \frac{1(A=a)}{f(A=a \mid X)} \int E(Y \mid A=1, m^{(3-j)}, m^{(j)}, X) f(m^{(3-j)} \mid A=1, X) dm^{(3-j)}] \\
&= E[S_{t=0}(M^{(j)} \mid A, X) \frac{1(A=a)}{f(A=a \mid X)} \eta_{3-j}(1, M^{(j)}, X)] \\
&= E[S_{t=0}(M^{(j)} \mid A, X) \frac{1(A=a)}{f(A=a \mid X)} (\eta_{3-j}(1, M^{(j)}, X) - \int \eta_{3-j}(1, M^{(j)}, X) f(m^{(j)} \mid A=a, X) dm^{(j)})] \\
&= E[S_{t=0}(M^{(j)}, A, X) \frac{1(A=a)}{f(A=a \mid X)} (\eta_{3-j}(1, M^{(j)}, X) - \gamma_j(a, X))] \\
&= E[S_{t=0}(Y, M^{(3-j)}, M^{(j)}, A, X) \frac{1(A=a)}{f(A=a \mid X)} (\eta_{3-j}(1, M^{(j)}, X) - \gamma_j(a, X))], \\
(3) &= E[\int E(Y \mid A=1, m^{(3-j)}, m^{(j)}, X) f(m^{(j)} \mid A=a, X) dm^{(j)} \frac{A}{f(A \mid X)} S_{t=0}(m^{(3-j)} \mid A=1, X)] \\
&= E[S_{t=0}(M^{(3-j)} \mid A, X) \frac{A}{f(A \mid X)} \eta_j(a, M^{(3-j)}, X)] \\
&= E[S_{t=0}(M^{(3-j)} \mid A, X) \frac{A}{f(A \mid X)} (\eta_j(a, M^{(3-j)}, X) - \int \eta_j(a, M^{(3-j)}, X) f(m^{(3-j)} \mid A=1, X) dm^{(3-j)})] \\
&= E[S_{t=0}(M^{(3-j)} \mid A, X) \frac{A}{f(A \mid X)} (\eta_j(a, M^{(3-j)}, X) - \gamma_j(a, X))] \\
&= E[S_{t=0}(M^{(3-j)}, A, X) \frac{A}{f(A \mid X)} (\eta_j(a, M^{(3-j)}, X) - \gamma_j(a, X))] \\
&= E[S_{t=0}(Y, M^{(j)}, M^{(3-j)}, A, X) \frac{A}{f(A \mid X)} (\eta_j(a, M^{(3-j)}, X) - \gamma_j(a, X))] \\
(4) &= \int \gamma_j(a, X) \nabla_{t=0} f_t(X)|_{t=0} dx \\
&= E[S_{t=0}(X) \gamma_j(a, X)] \\
&= E[S_{t=0}(Y, M^{(3-j)}, M^{(j)}, A, X)(\gamma_j(a, X) - \theta_a^{(j)})]
\end{aligned}
$$

Hence an influence function of a RAL of $\theta_a^{(j)}$ is:

$$
\begin{aligned}
S_{\text{eff},a}^{M^{(j)}}(\mathcal{O}) &= \frac{A}{f(A \mid X)} R_a^{(j)}(M^{(j)}, M^{(3-j)}, X)(Y - E[Y \mid 1, M^{(3-j)}, M^{(j)}, X]) \\
&+ \frac{1(A=a)}{f(A=a \mid X)} (\eta_{3-j}(1, M^{(j)}, X) - \gamma_j(a, X)) \\
&+ \frac{A}{f(A \mid X)} (\eta_j(a, M^{(3-j)}, X) - \gamma_j(a, X)) + \gamma_j(a, X) - \theta_a^{(j)}.
\end{aligned}
$$

Under $\mathcal{M}_{\text{non}}$, $S_{\text{eff},a}^{M^{(j)}}(\mathcal{O})$ is the efficient influence function. Hence the efficient influence func-

tion for Δ is the difference between the EIF for θ_1 and θ_0, which is $S^{M^{(j)}}_{\text{eff},1}(\mathcal{O}) - S^{M^{(j)}}_{\text{eff},0}(\mathcal{O})$.

C.4 Proof of Theorem 3.4

We prove Theorem 3.4 for continuous mediators. The proofs for other cases are similar. First we prove the unbiasedness of the quadruple robust estimator $\hat{\Delta}^{(j)}_{\text{quad}}$ under $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$. The limits of the incorrect components are indexed by an asterisk sign. Since the robustness conditions of the triply robust estimator for the total indirect effect given in Tchetgen Tchetgen and Shpitser (2012) are implied from our quadruple robustness conditions, the estimator similarly constructed for Δ^{INT} is also quadruply robust.

C.4.1 Under $\mathcal{M}_1$, when $E^\star[Y \mid A, M^{(j)}, M^{(3-j)}, X]$ is incorrectly specified

$$
\begin{aligned}
&E[\hat{\Delta}^{(j)}_{\text{quad}}] \\
=&E[\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f(M^{(j)} \mid 1, X)\boldsymbol{c}(F_{M^{(1)}|1,X}(M^{(1)}), F_{M^{(2)}|1,X}(M^{(2)}) \mid 1, X)}(Y - E^\star[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f(A \mid X)}\{\eta^\star_j(1, M^{(3-j)}, X) - \eta^\star_j(0, M^{(3-j)}, X)\} + \frac{2A-1}{f(A \mid X)}\eta^\star_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f(A \mid X)}\right)\gamma^\star_j(1, X) - \left(1 - \frac{1}{f(A \mid X)}\right)\gamma^\star_j(0, X)] \\
=&E[\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f(M^{(j)} \mid 1, M^{(3-j)}, X)}(Y - E^\star[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f(A \mid X)}\{\eta^\star_j(1, M^{(3-j)}, X) - \eta^\star_j(0, M^{(3-j)}, X)\} + \frac{2A-1}{f(A \mid X)}\eta^\star_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f(A \mid X)}\right)\gamma^\star_j(1, X) - \left(1 - \frac{1}{f(A \mid X)}\right)\gamma^\star_j(0, X)] \\
-&\Delta^{M^{(j)}},
\end{aligned}
\tag{C.5}
$$

where

$$
\begin{aligned}
&\eta^\star_j(a, M^{(3-j)}, X) = \int E^\star[Y \mid A = 1, m^{(j)}, M^{(3-j)}, X] f(m^{(j)} \mid A = a, X) dm^{(j)}, \\
&\gamma^\star_j(a, X) = E[\eta^\star_j(a, M^{(3-j)}, X) \mid A = 1, X].
\end{aligned}
$$

Hence

$$
\begin{aligned}
(5) &= E[\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f(M^{(j)} \mid 1, M^{(3-j)}, X)} Y] - E[\gamma_j^\star(1, X) - \gamma_j^\star(0, X)] \\
&+ E[\frac{A}{f(A \mid X)} \{\eta_j^\star(1, M^{(3-j)}, X) - \eta_j^\star(0, M^{(3-j)}, X)\}] + E[\frac{2A-1}{f(A \mid X)} \eta_{3-j}^\star(1, M^{(j)}, X)] \\
&+ E[\left(1 - \frac{2A}{f(A \mid X)}\right) \gamma_j^\star(1, X) - \left(1 - \frac{1}{f(A \mid X)}\right) \gamma_j^\star(0, X)] - \Delta^{M^{(j)}} \\
&= E[\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f(M^{(j)} \mid 1, M^{(3-j)}, X)} Y] - E[\gamma_j^\star(1) - \gamma_j^\star(0)] \\
&+ E[\gamma_j^\star(1, X) - \gamma_j^\star(0, X)] + E[\gamma_j^\star(1, X) - \gamma_j^\star(0, X)] + E[-\gamma_j^\star(1, X) + \gamma_j^\star(0, X)] - \Delta^{M^{(j)}} \\
&= \Delta^{M^{(j)}} - \Delta^{M^{(j)}} = 0.
\end{aligned}
$$

C.4.2 Under $\mathcal{M}_2$, when $f(M^{(j)} \mid A, X)$ and $\boldsymbol{c}(F_{M^{(1)}|A,X}(M^{(1)}), F_{M^{(2)}|A,X}(M^{(2)}) \mid A, X)$ are incorrectly specified

$$
\begin{aligned}
&E[\hat{\Delta}_{\text{quad}}^{(j)}] \\
=&E[\frac{A}{f(A \mid X)} \frac{f^\star(M^{(j)} \mid 1, X) - f^\star(M^{(j)} \mid 0, X)}{f^\star(M^{(j)} \mid 1, X) \boldsymbol{c}^\star(F_{M^{(1)}|1,X}(M^{(1)}), F_{M^{(2)}|1,X}(M^{(2)}) \mid 1, X)} (Y - E[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f(A \mid X)} \{\eta_j^\star(1, M^{(3-j)}, X) - \eta_j^\star(0, M^{(3-j)}, X)\} + \frac{2A-1}{f(A \mid X)} \eta_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f(A \mid X)}\right) \gamma_j^\star(1, X) - \left(1 - \frac{1}{f(A \mid X)}\right) \gamma_j^\star(0, X)] \\
=&E[\frac{A}{f(A \mid X)} \frac{f^\star(M^{(j)} \mid 1, X) - f^\star(M^{(j)} \mid 0, X)}{f^\star(M^{(j)} \mid 1, M^{(3-j)}, X)} (Y - E[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f(A \mid X)} \{\eta_j^\star(1, M^{(3-j)}, X) - \eta_j^\star(0, M^{(3-j)}, X)\} + \frac{2A-1}{f(A \mid X)} \eta_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f(A \mid X)}\right) \gamma_j^\star(1, X) - \left(1 - \frac{1}{f(A \mid X)}\right) \gamma_j^\star(0, X)] \\
-&\Delta^{M^{(j)}},
\end{aligned}
\tag{C.6}
$$

where

$$
\begin{aligned}
&\eta_j^\star(a, M^{(3-j)}, X) = \int E[Y \mid A = 1, m^{(j)}, M^{(3-j)}, X] f^\star(m^{(j)} \mid A = a, X) dm^{(j)}, \\
&\gamma_j^\star(a) = E[\eta_j^\star(a, M^{(3-j)}, X) \mid 1, X].
\end{aligned}
$$

Hence

$$
\begin{aligned}
(6) &= 0 + E[\gamma_j^\star(1,X) - \gamma_j^\star(0,X)] + E[\frac{2A-1}{f(A \mid X)}\eta_{3-j}(1, M^{(j)}, X)] + E[-\gamma_j^\star(1,X) + \gamma_j^\star(0,X)] - \Delta^{M^{(j)}} \\
&= E[\frac{2A-1}{f(A \mid X)}\eta_{3-j}(1, M^{(j)}, X)] - \Delta^{M^{(j)}} \\
&= \Delta^{M^{(j)}} - \Delta^{M^{(j)}} = 0
\end{aligned}
$$

C.4.3 Under $\mathcal{M}_3$, when $f(M^{(3-j)} \mid A, X)$ and $\boldsymbol{c}(F_{M^{(1)}|A,X}, F_{M^{(2)}|A,X} \mid A, X)$ are incorrectly specified

$$
\begin{aligned}
&E[\hat{\Delta}^{(j)}_{\text{quad}}] \\
=&E[\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f(M^{(j)} \mid 1, X)\boldsymbol{c}^\star(F_{M^{(1)}|A=1,X}(M^{(1)}), F_{M^{(2)}|1,X}(M^{(2)}) \mid 1, X)}(Y - E[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f(A \mid X)}\{\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)\} + \frac{2A-1}{f(A \mid X)}\eta^\star_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f(A \mid X)}\right)\gamma_j^\star(1,X) - \left(1 - \frac{1}{f(A \mid X)}\right)\gamma_j^\star(0,X)] \\
=&E[\frac{A}{f(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f^\star(M^{(j)} \mid 1, M^{(3-j)}, X)}(Y - E[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f(A \mid X)}\{\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)\} + \frac{2A-1}{f(A \mid X)}\eta^\star_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f(A \mid X)}\right)\gamma_j^\star(1,X) - \left(1 - \frac{1}{f(A \mid X)}\right)\gamma_j^\star(0,X)] \\
-&\Delta^{M^{(j)}},
\end{aligned}
\tag{C.7}
$$

where

$$
\begin{aligned}
&\eta_j^\star(a, M^{(3-j)}, X) = \int E[Y \mid A = 1, m^{(j)}, M^{(3-j)}, X] f^\star(m^{(j)} \mid A = a, X) dm^{(j)}, \\
&\gamma_j^\star(a, X) = \iint E[Y \mid A = 1, m^{(j)}, m^{(3-j)}, X] f^\star(m^{(3-j)} \mid A = a, X) f(m^{(j)} \mid A = a, X) dm^{(j)}\ dm^{(3-j)}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
(7) &= 0 + E[\frac{A}{f(A \mid X)}\{\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)\}] + E[\gamma_j^\star(1,X) - \gamma_j^\star(0,X)] \\
&\quad + E[-\gamma_j^\star(1,X) + \gamma_j^\star(0,X)] - \Delta^{M^{(j)}} \\
&= E[\frac{A}{f(A \mid X)}\{\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)\}] - \Delta^{M^{(j)}} \\
&= \Delta^{M^{(j)}} - \Delta^{M^{(j)}} = 0.
\end{aligned}
$$

C.4.4 Under $\mathcal{M}_4$, when $f(A \mid X)$ is incorrectly specified

$$
\begin{aligned}
&E[\hat{\Delta}^{(j)}_{\text{quad}}] \\
=&E[\frac{A}{f^\star(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f(M^{(j)} \mid 1, X)\boldsymbol{c}(F_{M^{(1)}|1,X}(M^{(1)}), F_{M^{(2)}|1,X}(M^{(2)}) \mid 1, X)}(Y - E[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f^\star(A \mid X)}\{\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)\} + \frac{2A-1}{f^\star(A \mid X)}\eta_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f^\star(A \mid X)}\right)\gamma_j^\star(1, X) - \left(1 - \frac{1}{f^\star(A \mid X)}\right)\gamma_j^\star(0, X)] \\
=&E[\frac{A}{f^\star(A \mid X)} \frac{f(M^{(j)} \mid 1, X) - f(M^{(j)} \mid 0, X)}{f(M^{(j)} \mid 1, M^{(3-j)}, X)}(Y - E[Y \mid 1, M^{(j)}, M^{(3-j)}, X])] \\
+&E[\frac{A}{f^\star(A \mid X)}\{\eta_j(1, M^{(3-j)}, X) - \eta_j(0, M^{(3-j)}, X)\} + \frac{2A-1}{f^\star(A \mid X)}\eta_{3-j}(1, M^{(j)}, X)] \\
+&E[\left(1 - \frac{2A}{f^\star(A \mid X)}\right)\gamma_j(1, X) - \left(1 - \frac{1}{f^\star(A \mid X)}\right)\gamma_j(0, X)] - \Delta^{M^{(j)}},
\end{aligned} \tag{C.8}
$$

where

$$
\begin{aligned}
&\eta_j(a, M^{(3-j)}, X) = \int E[Y \mid A = 1, m^{(j)}, M^{(3-j)}, X] f(m^{(j)} \mid A = a, X) dm^{(j)}, \\
&\gamma_j(a, X) = \iint E[Y \mid A = 1, m^{(j)}, m^{(3-j)}, X] f(m^{(3-j)} \mid A = a, X) f(m^{(j)} \mid A = a, X) dm^{(j)}\ dm^{(3-j)}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
(8) &= 0 + E[\frac{f(A = 1 \mid X)}{f^\star(A = 1 \mid X)}\{\gamma_j(1, X) - \gamma_j(0, X)\}] + E[\frac{f(A = 1 \mid X)}{f^\star(A = 1 \mid X)}\gamma_j^\star(1, X) - \frac{f(A = 0 \mid X)}{f^\star(A = 0 \mid X)}\gamma_j^\star(0, X)] \\
&\quad + E[-\frac{2f(A = 1 \mid X)}{f^\star(A = 1 \mid X)}\gamma_j^\star(1, X) + \frac{f(A \mid X)}{f^\star(A \mid X)}\gamma_j^\star(0, X)] + E[\gamma_j(1, X) - \gamma_j(0, X)] - \Delta^{M^{(j)}} \\
&= \Delta^{M^{(j)}} - \Delta^{M^{(j)}} = 0.
\end{aligned}
$$

Under suitable regularity conditions (Manski, 1988) for $S^{M^{(j)}}_{\text{eff},a}(\mathcal{O})$ and the score of the parameters in the nuisance models, by Taylor expansion we have:

$$
\sqrt{n}(\hat{\Delta}^{(j)}_{\text{quad}} - \Delta^{M^{(j)}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S^{\text{intersection}}_{\Delta^{M^{(j)}}}(\mathcal{O}_i, \Delta^{M^{(j)}}, \tau^\star) + o_p(1),
$$

where τ denotes the nuisance parameters, with $\tau^\star$ being its probability limit, and S_τ being its score, and

$$
S^{\text{intersection}}_{\Delta^{M^{(j)}}}(\mathcal{O}_i, \Delta^{M^{(j)}}, \tau^\star) = S^{M^{(j)}}_{\text{eff},a}(\mathcal{O}_i, \Delta^{M^{(j)}}, \tau^\star) - \frac{\partial E[S^{M^{(j)}}_{\text{eff},a}(\mathcal{O}_i, \Delta^{M^{(j)}}, \tau^\star)]}{\partial \tau^T} E^{-1}[\frac{\partial S_\tau(\mathcal{O}_i, \tau^\star)}{\partial \tau^T}] S_\tau(\mathcal{O}_i, \tau^\star).
$$

Then the asymptotic normality of the quadruply robust estimator can be proven using Slutsky's Theorem and the Central Limit Theorem.