



University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations


Fall 12-23-2009

Causal Inference in Discretely Observed Continuous Time Processes

Mingyuan Zhang

University of Pennsylvania, zhangmi@wharton.upenn.edu

Follow this and additional works at: <http://repository.upenn.edu/edissertations>

 Part of the [Biostatistics Commons](#), [Longitudinal Data Analysis and Time Series Commons](#), and the [Statistical Methodology Commons](#)

Recommended Citation

Zhang, Mingyuan, "Causal Inference in Discretely Observed Continuous Time Processes" (2009). *Publicly Accessible Penn Dissertations*. 44.

<http://repository.upenn.edu/edissertations/44>

This paper is posted at ScholarlyCommons. <http://repository.upenn.edu/edissertations/44>

For more information, please contact libraryrepository@pobox.upenn.edu.

Causal Inference in Discretely Observed Continuous Time Processes

Abstract

In causal inference for longitudinal data, standard methods usually assume that the underlying processes are discrete time processes, and that the observational time points are the time points when the processes change values. The identification of these standard models often relies on the sequential randomization assumption, which assumes that the treatment assignment at each time point only depends on current covariates and the covariates and treatment that are observed in the past. However, in many real world data sets, it is more reasonable to assume that the underlying processes are continuous time processes, and that they are only observed at discrete time points. When this happens, the sequential randomization assumption may not be true even if it is still a reasonable abstraction of the treatment decision mechanism at the continuous time level. For example, in a multi-round survey study, the decision of treatment can be made by the subject and the subject's physician in continuous time, while the treatment level and covariates are only collected in discrete times by a third party survey organization. The mismatch in the treatment decision time and the observational time makes the sequential randomization assumption false in the observed data. In this dissertation, we show that the standard methods could produce severely biased estimates, and we would explore what further assumptions need to be made to warrant the use of standard methods. If these assumptions are false, we advocate the use of controlling-the-future method of Joffe and Robins (2009) when we are able to reconstruct the potential outcomes from the discretely observed data. We propose a full modeling approach and demonstrate it by an example of estimating the effect of vitamin A deficiency on children's respiratory infection, when we are not able to do so. We also provide a semi-parametric analysis of the controlling-the-future method, giving the semi-parametric efficient estimator.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Dylan Small

Second Advisor

Marshall Joffe

Keywords

Causal inference, Longitudinal data, Continuous time process, Semi-parametric model, Diarrhea, Vitamin A deficiency

Subject Categories

Biostatistics | Longitudinal Data Analysis and Time Series | Statistical Methodology

CAUSAL INFERENCE IN DISCRETELY OBSERVED CONTINUOUS TIME
PROCESSES

Mingyuan Zhang

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2009

Supervisor of Dissertation

Co-Supervisor

Dylan Small
Associate Professor, Statistics

Marshall Joffe
Associate Professor, Biostatistics

Graduate Group Chairperson

Eric Bradlow, Professor of Marketing, Statistics, and Education

Dissertation Committee

Dylan Small, Associate Professor, Statistics
Marshall Joffe, Associate Professor, Biostatistics
Paul Rosenbaum, Professor of Statistics

Dedicated to

My Family

Acknowledgements

First and foremost, I offer my sincerest gratitude to my advisor, Dylan Small, who taught me the theory of exponential family four years ago, and has helped and guided me in every aspect of study, research and professional pursuit ever since. My deepest gratitude also goes to my co-advisor, Marshall Joffe, who has inspired me many times throughout my dissertation research with his broad knowledge and wonderful intuitions, and has always been nice and patient when I encountered with difficulties. It is impossible to over express how thankful I am to have them as my advisers. They are the real *Liang-Shi-Yi-You* (mentor and friend in Chinese), who always consider and suggest what is best for me.

I thank Paul Rosenbaum for kindly agreeing to serve in my committee and for his insightful comments on my dissertation. I would also like to thank Andreas Buja, Tony Cai, Shane Jensen, Robin Pemantle, Alexander Rakhlin, Paul Shaman, Mike Steele, Linda Zhao and all other the professors from the Department of Statistics, who educated me and brought me to the fascinating world of statistics. I am also lucky and grateful to have worked with all the staff and my fellow student colleagues at the statistics department. They made my PHD life easy and enjoyable.

I thank Tom Ten Have, Kevin Lynch, Jinbo Chen and other the members in the causal research group and the semi-parametric reading group at Penn, who listened to my research presentation several times and always provided constructive feedbacks. I also appreciate several discussions with Judith Lok and James Robins from Harvard University. It was very helpful to talk to them, as I spent days and nights studying their work.

Finally, I would like to thank my parents and Min for their support and love. Without them, no accomplishment is meaningful.

Mingyuan Zhang

November 30th, 2009

ABSTRACT

CAUSAL INFERENCE FOR DISCRETELY OBSERVED CONTINUOUS TIME PROCESSES

Mingyuan Zhang

Supervisors: Dylan Small and Marshall Joffe

In causal inference for longitudinal data, standard methods usually assume that the underlying processes are discrete time processes, and that the observational time points are the time points when the processes change values. The identification of these standard models often relies on the sequential randomization assumption, which assumes that the treatment assignment at each time point only depends on current covariates and the covariates and treatment that are observed in the past. However, in many real world data sets, it is more reasonable to assume that the underlying processes are continuous time processes, and that they are only observed at discrete time points. When this happens, the sequential randomization assumption may not be true even if it is still a reasonable abstraction of the treatment decision mechanism at the continuous time level. For example, in a multi-round survey study, the decision of treatment can be made by the subject and the subject's physician in continuous time, while the treatment level and covariates are only collected in discrete times by

a third party survey organization. The mismatch in the treatment decision time and the observational time makes the sequential randomization assumption false in the observed data. In this dissertation, we show that the standard methods could produce severely biased estimates, and we would explore what further assumptions need to be made to warrant the use of standard methods. If these assumptions are false, we advocate the use of controlling-the-future method of Joffe and Robins (2009) when we are able to reconstruct the potential outcomes from the discretely observed data. We propose a full modeling approach and demonstrate it by an example of estimating the effect of vitamin A deficiency on children's respiratory infection, when we are not able to do so. We also provide a semi-parametric analysis of the controlling-the-future method, giving the semi-parametric efficient estimator.

Contents

Title Page	i
Dedication	ii
Acknowledgements	iii
Abstract	v
Table of Contents	vii
List of Tables	xii
List of Figures	xiii
1 Introduction	1
1.1 Standard Causal Inference in Longitudinal Data	1
1.2 When the Longitudinal Data are Discrete Time Observations from Continuous Time Processes	4
2 Causal Inference for Continuous Time Processes when Covariates	

Are Observed Only at Discrete Times	10
2.1 Motivations and the Basic Setup	11
2.1.1 Examples of Treatments Varying in Continuous Time where Covariates are Observed Only at Discrete Times	11
2.1.2 An Model Data Generating Process	13
2.1.3 Difficulties Posed by Treatments Varying in Continuous Time when Covariates Are Observed Only at Discrete Times	14
2.2 Standard Discrete Time Structural Nested Model and G-estimation and Their Versions with Underlying Continuous Time Processes	17
2.2.1 Discrete Time Structural Nested Model and G-estimation	17
2.2.2 A Continuous Time Deterministic Model and Continuous Time Sequential Randomization	20
2.2.3 Modified G-estimation on Discrete Time Observed Data from the Continuous Time Model	22
2.3 Justification of the Use of the Modified Discrete Time G-estimation	25
2.3.1 Sequential Randomization at Any Finite Subset of Time Points	26
2.3.2 A Markovian Condition	28
2.4 The Controlling-the-future Method	32
2.4.1 Modified Assumption and Estimation of Parameters	33
2.4.2 The Controlling-the-future Method and the Markovian Condition	35
2.5 Simulation Study	40

2.5.1	The Simulation Models	41
2.5.2	Estimations and Results under M1	44
2.5.3	Simulation Results under M2 and M3	46
2.5.4	Estimations and Results under M4	48
2.6	Application To The Diarrhea Data	50
2.7	Conclusion	54
3	Causal Inference for a Discretely Observed Continuous Time Non-stationary Markov Process	57
3.1	Introduction	58
3.2	A Markov Model	62
3.2.1	The Causal Model	63
3.2.2	Continuous Time Markov Process	63
3.3	Estimation: MCMC with Data Augmentation	69
3.3.1	Discretization Scheme	70
3.3.2	The MCMC Algorithm	71
3.3.3	Proposal Distributions for the Augmented Data	73
3.4	Simulation	78
3.5	Application	80
3.5.1	Estimation Result	82
3.5.2	Simulation Based Causal Interpretation	84
3.5.3	Model Assumptions Revisited	86

3.6	Conclusion	94
4	Controlling-the-future Revisited: the Optimal Estimating Equation	97
4.1	A Single Period Semi-parametric Model Under Relaxed Ignorability	
	Assumption	98
4.1.1	The Single Period Model	98
4.1.2	Characterization of the Nuisance Tangent Space	99
4.1.3	The Space that is Orthogonal to the Nuisance Tangent Space	101
4.1.4	The Efficient Score	103
4.2	Locally Efficient Doubly Robust RAL Estimator	107
4.2.1	The Estimator that Achieves Semi-parametric Efficiency Bound	107
4.2.2	Construction of a Locally Efficient Doubly Robust RAL Estimator	108
4.2.3	Locally Efficiency	109
4.2.4	Double Robustness	115
4.3	Important Special Cases	118
4.3.1	When the Treatment is Binary	118
4.3.2	Special Blip-down Functions	120
4.3.3	Identification Issue	121
4.4	Extension to Multi-period Case	122
4.4.1	Likelihood Function	124
4.4.2	Nuisance Tangent Space	125

4.4.3	The Efficient Score	128
4.5	Conclusion	134
5	Appendices	136
5.1	Estimating Covariance Matrix of Estimated Parameters	136
5.2	Definition of N_t and Explicit Formula of λ_t	137
5.3	Proof of FTSR Implying CTSR	139
5.4	Proof of Theorem 2.3.4	150
5.5	Simulation Parameters	155
5.6	Continuous Time Ignorability	156
5.7	Simulation of Endpoint-Conditioned Bounded Simple Random Walk .	160

List of Tables

2.1	Estimated Causal Parameters from Data Generated by M1-4	47
2.2	Verification of Observational Time Sequential Randomization Under M4	49
2.3	Estimation of Ψ from the Diarrhea Data Set	53
3.1	Simulation Result for the MCMC Algorithm	81
3.2	Estimation Result from the Vitamin A Deficiency Data	83
3.3	Simulation Based Causal Interpretation for the Example	86
3.4	Estimates of $\tilde{\delta}$ from Different Models	88
3.5	Estimates of the Average Causal Difference for Different Models . . .	89
3.6	Estimation Result from the Vitamin A Deficiency Data without Continuous Time Ignorability	91
3.7	Estimation Result from the Vitamin A Deficiency Data with a Different Q	93

List of Figures

2.1	Directed Acyclic Graph	32
2.2	Directed Acyclic Graph with Non-Markovian Y_t^0	36
2.3	Directed Acyclic Graph with Leading Indicator in L_t	38
2.4	Example of Continuous Time Paths Under M1	44
3.1	Discretized Example of Data Generating Process	64
3.2	MCMC with Data Augmentation	72
3.3	Full MCMC Algorithm with Data Augmentation	78

Chapter 1

Introduction

1.1 Standard Causal Inference in Longitudinal Data

In a cross-sectional observational study of the effect of a treatment on an outcome, a usual assumption for making causal inferences is that there are no unmeasured confounders, i.e., that conditional on the measured confounders, the data is generated as if the treatment was assigned randomly. Under this assumption, a consistent estimate of the average causal effect of the treatment can be obtained from a correct model of the association between the treatment and the outcome conditional on the measured confounders (Cochran, 1965). In a longitudinal study, the analogue of the no unmeasured confounders assumption is that at the time of each treatment assignment, there are no unmeasured confounders; this is called the *sequential randomization* or *sequential ignorability* assumption:

(A1) The longitudinal data of interest are generated as if the treatment is randomized in each period, conditional on the current values of measured covariates and the

history of the measured covariates and the treatment.

The sequential randomization assumption implies that decision on treatment assignment is based on observable history and contemporaneous covariates and that people have no ability to peek into the future. Robins (1986) has shown that for a longitudinal study, unlike for a cross-sectional study, even if the sequential randomization assumption holds, the standard method of estimating the causal effect of the treatment by the association between the outcome and the treatment history conditional on the confounders can provide a biased and inconsistent estimate. This bias can occur when we are interested in estimating the joint effects of all treatment assignments and when the following conditions hold:

- (c1) conditional on past treatment history, a time-dependent variable is a predictor of the subsequent mean of the outcome and also a predictor of subsequent treatment;
- (c2) past treatment history is an independent predictor of the time-dependent variable.

An example in which the standard methods are biased is the estimation of the causal effect of the drug AZT (zidovudine) on CD4 counts in AIDS patients. Past CD4 count is a time-dependent confounder for the effect of AZT on future CD4 count, since it not only predicts future CD4 count but also subsequent initiation of AZT therapy. Also, past AZT history is an independent predictor of subsequent CD4 count (Hernán, Brumback and Robins, 2002).

To eliminate the bias of standard methods for estimating the causal effect of

treatment in longitudinal studies where sequential randomization holds but there are time-dependent confounders satisfying conditions (c1) and (c2) (e.g., past CD4 counts), Robins (1992, 1994, 1998, 1999) developed a number of innovative methods, including g-computation algorithm, structural nested models (SNMs) with g-estimation, and marginal structural models (MSMs) with IPTW (inverse probability of treatment weighted) estimation.

A significant portion of this thesis focuses on structural nested models (SNMs) and their associated methods of g-testing and g-estimation. The basic idea of the g-test is the following. Given a hypothesized treatment effect and a deterministic model of the treatment effect, we can calculate the potential outcome that a subject would have received if she never received the treatment; if the hypothesized treatment effect is the true treatment effect, then this potential outcome will be independent of the actual treatment the subject received conditional on the confounder and treatment history, under the sequential randomization assumption (A1). G-estimation involves finding the treatment effect that makes the g-test statistic have its expected null value. A formal description of g-estimation is given in Section 2.2.1 of Chapter 2.

G-estimation is very attractive because researchers are usually only required to model the propensity score for the treatment. The estimate is consistent when the model of the treatment assignment is correct. It is also possible to construct a g-estimator that is locally efficient doubly robust. As long as we can correctly model either the propensity score for the treatment or the conditional distribution for the

potential outcomes conditional on the covariates, the estimator will be consistent, asymptotically normal and regular. The researchers get “two shots” rather than one to get a consistent estimate. Moreover, if we can correctly model the both the propensity score and the conditional distribution of the potential outcomes simultaneously, the estimator achieves the semi-parametric efficiency bound under the sole restriction that the assumption (A1) is true. In another word, there exists no other estimator that is more efficient than the locally efficient doubly robust g-estimator.

1.2 When the Longitudinal Data are Discrete Time Observations from Continuous Time Processes

Intriguing as the standard methods are, they are mostly designed for discrete time models. The discrete time models assume that the treatment, covariates and outcome processes can only change values at discrete time points, and that these discrete time points are fully observed. If at these observational time points we are able to measure all the covariates that is needed for the input of the decision making mechanism, we could reasonably assume assumption (A1) in the data, and in principle we might be able to model the decision making mechanism correctly given enough data. The standard methods would apply. Lok (2007) has extended the theory to continuous time processes, under the condition that the full continuous time paths of treatment and covariates are observed.

In this thesis, we consider causal inference in longitudinal data, but we assume

that the longitudinal data are discrete time observations of continuous time processes. In many real world problems, this is a more realistic assumption. For example, in the example of the effect of Zidovudine (AZT) on CD4 count, a data set from the Multi-center AIDS Cohort Study (Kaslow et al., 1987) can be used. In the study, the patients might visit the doctors any time during the years, and the doctors would decide the use of AZT at the time of the visits. However, the organization who collected the data only regularly collected them once every six months. It is more reasonable to assume that the treatment, covariates and outcomes processes are in continuous time, but only observed at discrete time points. More details on this study and other real world examples are in Chapter 2 and Chapter 3.

When this setting is true, standard methods that ignore the continuous time structure could produce severely biased estimates. In particular, we identify two important sources of the bias:

- Unmeasured Confounders:

The identification of the standard methods rely crucially on sequential randomization assumption in the data. In the setting that is of interest in this dissertation, the sequential randomization assumption could be a reasonable abstraction at the treatment decision level, i.e., the continuous time level. However, It is unlikely to be true in the observed data in discrete times. The covariates in between two consecutive time points are not measured and they could be important unmeasured confounders that associated with both the treatment

and the outcomes. As a result, the standard discrete time methods could be biased.

- Treatment Measurement Error:

When the treatment process is only observed at discrete time points, the amount of treatment that the subject receives may not be known precisely. For example, when the treatment affects the outcome cumulatively, discretely observed treatment process will not give us exact amount of cumulative treatment; when the treatment has direct or indirect effect on future outcomes, the past treatment affecting the outcomes at the discrete observational time points may not be observed. The treatment measurement error problem also includes the case when we do observe the whole continuous time treatment process, but the treatment effect varies with covariate processes, which might be unobserved in between two consecutive observational time points. In all these cases, even given the true parameters in the causal model, we will not be able to accurately reconstruct the counterfactual outcomes, or the mimicking counterfactual outcomes in case of non-rank preserving models (see Lok 2004), or the mean of counterfactual outcomes in case of mean models (see Hernán, Brumack and Robins, 2002). Any comparison among the counterfactuals might be biased.

This dissertation discusses conditions and methods that could eliminate or reduce these two sources of bias in several scenarios. In particular, it is organized as follows.

Chapter 2 focuses on the problem of unmeasured confounder. We assume that the

covariate and outcome processes are observed in discrete time points while the full continuous time treatment process is observed. This is practical when the treatment is a binary process and at the observational time points we are able to ask subjects about their treatment history (e.g., time of initiation and time of cessation). Under certain rank preserving models, we are free of the treatment measurement error problem. We discuss the use of standard methods in this scenario, and focus especially on SNMs and their associated g-test and g-estimation. Conditions on the continuous time processes are given in the chapter to warrant the use of a modified g-estimation. When these conditions fails, we propose the controlling-the-future method of Joffe and Robins (2009), which are based on a relaxed discrete time sequential randomization assumption that allows the treatment to depend on future potential outcomes given the past treatment and covariates. We show that the method can be used to correct or reduce bias from the unmeasured confounders in our continuous time setting. The content of this chapter is based on a working paper by Zhang et al. (2009).

Chapter 3 considers a case when we do not have the full continuous time treatment process and both the unmeasured confounder and the treatment error problems arise. We propose a full modeling approach for causal inference, demonstrated by an example of analyzing causal effect of vitamin A deficiency on children's respiratory infection from a longitudinal data collected in Indonesia in 1983 (see Sommer et al., 1983). The level of vitamin A deficiency could change any time in the years, while

the data were only collected once every season. Important covariates that predict the change in the levels of vitamin A deficiency and are related to the outcomes may be unobserved causing the unmeasured confounder problem. The treatment in between two observational time points are unobserved, causing the treatment error problem. The treatment error problem is worsened by the fact that we only observe a coarsened vitamin A deficiency level. In this chapter, we model the data generating process as a continuous time Markov process observed at discrete time points. We design an MCMC algorithm to estimate the Markov model. The content of this chapter is based on a working paper by Zhang and Small (2009).

Chapter 4 revisits the controlling-the-future method we used in Chapter 2. We view the relaxed sequential randomization assumption and the controlling-the-future method as a powerful extension of the standard g-estimation for dealing with unmeasured confounders. In this chapter, we provide a theoretical analysis of this method using the semi-parametric theory. In parallel to the standard theory for g-estimation, we derive the nuisance tangent space under the sole restriction of the relaxed sequential randomization assumption, and we calculate the efficient score under a single period semi-parametric model. We also propose a locally efficient doubly robust estimator for the controlling-the-future method. The calculation of nuisance tangent space and the efficient score are then extended to multi-period model with repeated outcomes. Chapter 4 can be viewed as a theoretical supplement for Joffe and Robins (2009).

Chapter 5 is the appendices, which include related the technical proofs in previous chapters.

Chapter 2

Causal Inference for Continuous Time Processes when Covariates Are Observed Only at Discrete Times

In this chapter, we study assumptions and methods for making causal inferences about the effect of a treatment that varies in continuous time when its full history of continuous time treatment process is observed but its time-dependent confounders are observed only at discrete times. In the framework of Section 1.2, we have unmeasured confounders, and under the causal models we will assume for this chapter (see Section 2.7), we are free of the measurement error problems. In such settings, standard discrete time g-estimation usually do not work, except when certain conditions are assumed for the continuous time process. In this chapter, we formulate such conditions. When these conditions do not hold, we propose a controlling-the-future method that can produce consistent estimates when g-estimation is consistent, and

is still consistent in some cases when g-estimation is severely inconsistent.

2.1 Motivations and the Basic Setup

First, we describe two motivating examples when treatment changes values in continuous time while covariates are only observed at discrete times.

2.1.1 Examples of Treatments Varying in Continuous Time where Covariates are Observed Only at Discrete Times

Example 1: *The effect of AZT (Zidovudine) on CD4 counts.* The Multicenter AIDS Cohort Study (Kaslow et al., 1987) has been used to study the effect of AZT on CD4 counts (Hernán, Brumback and Robins, 2002; Brumback et al. 2004). Participants in the study are asked to come semi-annually for visits at which they are asked to complete a detailed interview including a complete history of AZT use, as well as take a physical examination and provide blood samples from which CD4 counts are obtained. Decisions on AZT use are made by subjects and their physicians, and switches of treatment might happen any time between two visits. These decisions are based on the values of diagnostic variables, possibly including CD4 and CD8 counts, presence of certain symptoms, and other related time-dependent covariates. However, these covariates are only measured by MACS at the time of visits; the values of these covariates at the exact times that treatment decisions are made between visits are not available.

Example 2: *The effect of diarrhea on children's height.* Diarrheal disease is one of the leading causes of childhood illness in developing regions of the world (Kosek, Bern and Guerrant, 2003). Consequently, there is considerable concern about the effects of diarrhea on a child's physical and cognitive development (Moore et al., 2001; Guerrant et al., 2002). A data set which provides the opportunity to study the impact of diarrhea on a child's height is a longitudinal household survey conducted in Bangladesh in 1998-1999 after Bangladesh was struck by its worst flood in over a century in the summer of 1998 (del Ninno et al., 2001; del Ninno and Lundberg, 2005). The survey was fielded in three waves from a sample of 757 households: round 1 in November, 1998; round 2 in March-April, 1999; and round 3 in November, 1999. The survey recorded all episodes of diarrhea for each child in the household in the past six months or since the last interview by asking the families at the time of each interview. In addition, the survey recorded at each of the three interview times several important time-dependent covariates for the effect of diarrhea on a child's future height: the child's current height and weight, the amount of flooding in the child's home and village; and the household's economic and sanitation status. The child's current height and weight in particular are time-dependent confounders that satisfy conditions (c1) and (c2) in Section 1.1, making standard longitudinal data analysis methods biased (see Martorell and Ho, 1984 and Moore et al., 2001 for discussion of evidence for and reasons why current height and weight satisfy conditions (c1) and (c2)). The time-dependent confounders of current height and weight are available

only at the time of the interview, and changes in their value that might affect the exposure of the child to the “treatment” of diarrhea, which varies in continuous time, are not recorded in continuous time.

2.1.2 An Model Data Generating Process

In both the examples of AZT and diarrhea, the exposure or treatment process happens continuously in time and a complete record of the process is available, but the time-dependent confounders are only observed at discrete times. There could be various interpretations of the relationship between the data at the treatment decision level and the data at the observational time level. To clarify the problem of interest in this chapter, we consider the following model data generating process:

- (a1) A patient takes a certain medicine under the advice of a doctor.
- (a2) A doctor continuously monitors and records a list of health indicators of her patient, and decides the initiation and cessation of the medicine solely based on current and historical records of these conditions, the historical use of the medicine, as well as possibly random factors unrelated to the patient’s health.
- (a3) A third party organization asks a collection of patients from various doctors to visit the organization’s office semi-annually. The organization measures the same list of health indicators for the patients at the time of the visits, and asks the patients to report the detailed history of the use of the medicine between two visits.

(a4) We are only provided with the third party’s data.

Note that in (a2), we assume the sequential randomization assumption (A1) at the treatment decision level.

The AZT example can be approximated by the above data generating process. In the AZT example, (a1) and (a2) approximately describes the joint decision making process by the patient and the doctor in the real world. (a3) can be justified by reasonably assuming that the staffs at the MACS receive similar medical training and use similar medical equipment as the patients’ doctors. In the diarrhea example, the patient’s body, rather than a doctor determines whether the patient gets diarrhea. Assumption (a3) then is saying that the third party organization (the survey organization) collects enough health data, and that if all the history of such health data are available, the organization will be able to predict as well as is possible with current medical knowledge whether a patient gets diarrhea at the time of the survey.

2.1.3 Difficulties Posed by Treatments Varying in Continuous Time when Covariates Are Observed Only at Discrete Times

Suppose our data are generated as in the previous section, and we apply discrete time g-estimation at the discrete times at which the time-dependent covariates are observed; we will denote these observation times by $0, \dots, K$. In discrete time g-estimation, we are testing whether the observed treatment at time t ($t = 0, \dots, K$) is, conditional on the observed treatments at times $0, 1, \dots, t - 1$ and observed co-

variates at times $0, \dots, t$, independent of the putative potential outcomes at times $t + 1, \dots, K$ calculated under the hypothesized treatment effect, where the putative potential outcomes considered are what the subject's outcome would be at times $t + 1, \dots, K$ if the subject never received treatment at any time point (see Section 1.1). The difficulty with this procedure is that even if sequential randomization holds when the measured confounders are measured in continuous time (as is assumed in (a2)), it may not hold when the measured confounders are measured only at discrete times. For the discrete time data, there can be *unmeasured confounders*. In the MACS example, the diagnostic measures at the time of AZT initiation are missing unless the start of AZT initiation occurred exactly at one of the discrete times that the covariates are observed; the diagnostic measures at the initiation time are clearly important confounders for the treatment status at the subsequent observational time. In the diarrhea example, the nutrition status of the child before the start of a diarrhea episode is missing unless the start of the diarrhea episode occurred exactly at one of the discrete times that covariates are observed; this nutrition status is also an important confounder for the diarrhea status at the subsequent observational time. Continuous time sequential randomization does not in general justify sequential randomization holding for the discrete time data, meaning that discrete time g-estimation can produce inconsistent estimates even when continuous time sequential randomization holds.

In this chapter, we approach this problem from two perspectives. First, we give

conditions on the underlying continuous time processes under which discrete time sequential randomization is implied, warranting the use of discrete time g-estimation. Second, we propose a new estimation method called the controlling-the-future method that can produce consistent estimates whenever discrete time g-estimation is consistent and produce consistent estimates in some cases when discrete time g-estimation is inconsistent.

Our discussion focuses on a binary treatment and repeated continuous outcomes. We also assume that the cumulative amount of treatment between two visits is observed. This is true for Examples 1 and 2, the AZT and diarrhea studies respectively. If cumulative treatment is not observed, there will often be a measurement error problem in the amount of treatment, as is discussed in Section 1.2. Chapter 3 will present an example dealing with measurement error problems.

The organization of the chapter is as follows: Section 2.2 reviews the standard discrete time structural nested model and g-estimation, introduces settings for continuous time studies, and shows what we mean by applying discrete time g-estimation on discrete time observed data when the underlying process is in continuous time; Section 2.3 proposes conditions on the continuous time processes such that discrete time g-estimation works, and discusses their interpretability and usefulness; Section 2.4 describes our controlling-the-future method; Section 2.5 presents a simulation study; Section 2.6 provides an application to the diarrhea study discussed in Example 2; Section 2.7 concludes the chapter.

2.2 Standard Discrete Time Structural Nested Model and G-estimation and Their Versions with Underlying Continuous Time Processes

In this section, we review the discrete time SNM and g-estimation, introduce the notation for a continuous time study, formally define continuous time sequential randomization and explain the application of discrete g-estimation on the discrete time observations from the continuous time process.

2.2.1 Discrete Time Structural Nested Model and G-estimation

We describe a deterministic structural nested model, assuming all variables can only change at discrete, observable times. To save notation for the continuous time model, we use a star superscript on every variable in this section.

We assume that the study starts at time 0 and ends at time K . All variables can only change values at time $0, 1, 2, \dots, K$. We use A_k^* to denote the binary treatment decision at time k . Under the discrete time setup, A_k^* is assumed to be the constant level of treatment between time k and time $(k+1)$. We use Y_k^{0*} to denote the baseline counterfactual outcome of the study at time k , if the subject does not receive any treatment throughout the study, and use Y_k^* to denote the actual outcome at time k . In this chapter, we assume that all Y_k^{0*} 's and Y_k^* 's are continuous variables. Let L_k^* be the vector of covariates collected at time k . As a convention, Y_k^* is included in L_k^* .

We consider a simple deterministic model for illustration purpose. Generalizations of the results in this chapter for more complicated models are in Section 2.7. The model we consider is

$$Y_k^* = Y_k^{0*} + \Psi \sum_{i=0}^{k-1} A_i^* \quad (2.2.1)$$

where Ψ is the causal parameter of interest, and can be interpreted as the effect of one unit of the treatment on the outcome. In this model, the treatment affects the outcome cumulatively.

Model (2.2.1) is a rank-preserving model. It assumes that for subject i and j , with the same observed treatment history up to time k , if we observe $Y_{k,i} < Y_{k,j}$, we must have $Y_{k,i}^{0*} < Y_{k,j}^{0*}$.

The general purpose of causal inference is to estimate Ψ from the observables, the A_k^* 's and L_k^* 's (note that the Y_k^* 's are included in L_k^* 's). One way to achieve the identification of Ψ is to assume sequential randomization (A1).

Given this notation and model (2.2.1), a mathematical formulation of (A1) is

$$pr(A_k^* | \bar{L}_k^*, \bar{A}_{k-1}^*, \underline{Y}_{k+}^{0*}) = pr(A_k^* | \bar{L}_k^*, \bar{A}_{k-1}^*) \quad (2.2.2)$$

where $\bar{L}_k^* = (L_0, L_1, \dots, L_k)$, $\bar{A}_{k-1}^* = (A_0, A_1, \dots, A_{k-1})$, and $\underline{Y}_{k+}^{0*} = (Y_{k+1}^{0*}, Y_{k+2}^{0*}, \dots, Y_K^{0*})$.

For any hypothesized value of Ψ , we define a putative counterfactual

$$Y_k^{0*}(\Psi) = Y_k^* - \Psi \sum_{i=0}^{k-1} A_i$$

Then under (2.2.1) and (2.2.2), the correct Ψ should solve

$$E[U(\Psi)] \equiv E\left\{ \sum_{\substack{k < m \leq K \\ 1 \leq i \leq N}} [A_{i,k}^* - p_k(X_{i,k}^*)] g(Y_{i,m}^{0*}(\Psi), X_{i,k}^*) \right\} = 0 \quad (2.2.3)$$

where i is the index for each subject and there are N subjects, $X_{i,k}^* = (\bar{L}_{i,k}^*, \bar{A}_{i,k-1}^*)$, $p_k(X_{i,k}^*) = Pr(A_{i,k}^* = 1 | X_{i,k}^*)$ is the propensity score for subject i at time k , and g is any function. This estimating equation can be generalized with g being a function of any number of future $Y_{i,m}^{0*}(\Psi)$'s and $X_{i,k}^*$.

To estimate Ψ , we solve the empirical version of (2.2.3):

$$U(\Psi) \equiv \sum_{\substack{k < m \leq K \\ 1 \leq i \leq N}} [A_{i,k}^* - p_k(X_{i,k}^*)] g(Y_{i,m}^{0*}(\Psi), X_{i,k}^*) = 0 \quad (2.2.4)$$

The method is known as g -estimation. The efficiency of the estimate depends on the functional form of g . The optimal g function that produces the most efficient estimation can be derived (Robins, 1992).

In real applications, the treatment assignment scheme, or the true propensity score $p_k(X_k^*)$ is usually unknown, and is parameterized as $p_k(X_k^*, \beta)$. Then additional estimating equations are needed to identify β . One may use various g functions to construct these estimating equations, as long as these equations are not linearly correlated. For example, the following estimating equations could be used:

$$U(\Psi, \beta) = \sum_{\substack{k < m \leq K \\ 1 \leq i \leq N}} [A_{i,k}^* - p_k(X_{i,k}^*)] [Y_{i,m}^{0*}(\Psi), X_{i,k}^*]^T = 0 \quad (2.2.5)$$

The formulas for estimating the covariance matrix of $(\hat{\Psi}, \hat{\beta})$ are given in the appendices in Section 5.1.

2.2.2 A Continuous Time Deterministic Model and Continuous Time Sequential Randomization

We now extend the model in Section 2.2.1 to a continuous time model, and define a continuous time version of the sequential randomization assumption (A1) as a counterpart of (2.2.2).

Same as before, we assume that the continuous time study starts at time 0 and ends at an integer time K , but now the variables can change their values at any real time between 0 and K . The model in Section 2.2.1 is then extended as follows:

- $\{Y_t; 0 \leq t \leq K\}$ is the continuous time continuously valued outcome process.
- $\{L_t; 0 \leq t \leq K\}$ is the continuous time covariate process. It can be multi-dimensional, and Y_t is an element of L_t .
- $\{A_t; 0 \leq t \leq K\}$ is the continuous time binary treatment process.
- $\{Y_t^0; 0 \leq t \leq K\}$ is the continuous time continuously valued potential outcome process if the subject does not receive any treatment from time 0 to time K . It can be thought of as the *natural process* of the subject, free of treatment/intervention.

A natural extension of model (2.2.1) is

$$Y_t = Y_t^0 + \Psi \int_0^t A_s ds \tag{2.2.6}$$

where Ψ is the causal parameter of interest. Ψ can be interpreted as the effect rate of the treatment on the outcome.

In this continuous time model, a continuous time version of the sequential randomization assumption (A1) can be formalized, though it does not have the simple form similar to Equation (2.2.2). It was noted by Lok (2007) that a direct extension of the formula (2.2.2) involves “conditioning null events on null events”.

Lok (2007) formally defined continuous time sequential randomization when there is only one outcome at the end of the study. We propose a similar definition for studies with repeated outcomes under the deterministic model (2.2.6).

Following Lok (2007), we assume that all the continuous time stochastic processes are *càdlàg* processes (*continue à droite, limitée à gauche*, i.e., continuous from the right, having limit from the left), throughout this chapter. Let $Z_t = (L_t, A_t, Y_t^0)$. Let $\sigma(Z_t)$ be the σ -field generated by Z_t , i.e., the smallest σ -field that makes Z_t measurable. Let $\sigma(\bar{Z}_t)$ be the σ -field generated by $\bigcup_{u \leq t} \sigma(Z_u)$. Similarly, $\sigma(\bar{Z}_t, \underline{Y}_{t+}^0)$ is the σ -field generated by $\sigma(\bar{Z}_t) \cup \sigma(\underline{Y}_{t+}^0)$, where $\sigma(\underline{Y}_{t+}^0)$ is the σ -field generated by $\bigcup_{u > t} \sigma(Y_u^0)$. By definition, the sequence of $\sigma(\bar{Z}_t)$, $0 \leq t \leq K$, forms a filtration. The sequence of $\sigma(\bar{Z}_t, \underline{Y}_{t+}^0)$, $0 \leq t \leq K$, also forms a filtration, because $\sigma(\bar{Z}_t, \underline{Y}_{t+}^0) \subset \sigma(\bar{Z}_s, \underline{Y}_{s+}^0)$, for $t < s$ (note that this is true under the deterministic model (2.2.6) but not in general).

Let N_t be a counting process determined by A_t . It counts the number of jumps in the A_t process. Let λ_t be the intensity process of N_t with respect to $\sigma(\bar{Z}_t)$.

(The explicit definition of N_t and an explicit formula for λ_t is in the appendices in Section 5.2.) $M_t = N_t - \int_0^t \lambda_s ds$ will be a martingale w.r.t. $\sigma(\bar{Z}_t)$.

Definition 2.2.1. With N_t , λ_t and M_t defined as above, the *càdlàg* process $Z_t \equiv (L_t, A_t, Y_t^0)$, $0 \leq t \leq K$ is said to satisfy the **continuous time sequential randomization** assumption, or **CTSR**, if M_t is also a martingale w.r.t. $\sigma(\bar{Z}_t, \underline{Y}_{t+}^0)$. Or, equivalently, λ_t is also the intensity of N_t , w.r.t. the filtration of $\sigma(\bar{Z}_t, \underline{Y}_{t+}^0)$.

In this definition, given A_0 , the counting process $\{N_t\}_0^T$ offers an alternative description of the treatment process $\{A_t\}_0^T$. The intensity process λ_t , which models the jumping rate of N_t , plays the same role as the propensity scores in discrete time model, which models the switching of the treatment process. Definition 2.2.1 formalizes the assumption (A1) in the continuous time model, by stating that λ_t does not depend on future potential outcomes. It is worth noting that the definition here is only for a rank preserving model. A generalization of this definition is given in Section 5.6.

2.2.3 Modified G-estimation on Discrete Time Observed Data from the Continuous Time Model

In this chapter, we assume that the continuous process defined in Section 2.2.2 can only be observed at integer times, namely, times $0, 1, 2, \dots$, and K . We use the same set of starred notation as in Section 2.2.1, but interpret them as discrete time observations from the model in Section 2.2.2. Specifically,

- $\{A_k^*, k = 0, 1, 2, \dots, K\}$ denotes the set of treatment assignments observable at times $0, 1, 2, \dots, K$. We use \bar{A}_k^* to denote the observed history of observed discrete time treatment up to time k , i.e., $(A_0^*, A_1^*, \dots, A_k^*)$. Additionally, we use $cumA_k^* = \int_0^{k-} A_s ds$ to denote the cumulative amount of treatment up to time k . Note that in the continuous time model, $cumA_k^* \neq \sum_{k'=0}^{k-1} A_{k'}^*$, as it would in discrete time models. We let $\overline{cumA}_k^* = (cumA_1^*, cumA_2^*, \dots, cumA_k^*)$. We note that in practice, people sometimes use $\tilde{A}_k^* = cumA_{k+1}^* - cumA_k^*$ as the treatment at time k , when applying discrete time g-estimation on discrete time observational data. Such use of g-estimation usually requires stronger conditions than the conditions discussed in this chapter. Throughout this chapter, we define the treatment at time k as A_k^* .

- We define L_k^* , the observed covariates at time k to be L_{k-} , the left limit of L at time k , following the convention that in discrete model, people usually assume that the covariates are measured before the treatment decision. Y_k^* and Y_k^{0*} are also defined as Y_{k-} and Y_{k-}^0 respectively, following the same convention. \bar{L}_k^* denotes $(L_0^*, L_1^*, \dots, L_k^*)$, and \bar{Y}_k^* and \bar{Y}_k^{0*} are defined accordingly. $\underline{Y}_{k+}^{0*} = (Y_{k+1}^{0*}, Y_{k+2}^{0*}, \dots, Y_K^{0*})$.

With this notation, following the spirit of g-estimation, which controls all observed history in the propensity score model for the treatment, we propose the following

estimating equation:

$$U(\Psi) \equiv \sum_{\substack{k < m \leq K \\ 1 \leq i \leq N}} [A_{i,k}^* - p_k(X_{i,k}^*)] g(Y_{i,m}^{0*}(\Psi), X_{i,k}^*) = 0 \quad (2.2.7)$$

where $X_{i,k}^*$ is the collection of $\bar{L}_{i,k}^*$, $\bar{A}_{i,k-1}^*$ and $\overline{cumA}_{i,k}^*$, $p_k(X_{i,k}^*) = Pr(A_{i,k}^* = 1 | X_{i,k}^*)$, and $Y_{i,m}^{0*}(\Psi) = Y_{i,m}^* - \Psi cumA_{i,k}^*$.

In practice, $p_k(X_{i,k}^*)$ is unknown and has to be parametrized as $p_k(X_{i,k}^*; \beta)$, and we use different functions g to identify the all the parameters, as in Section 2.2.1. The covariance matrix of estimated parameters can be estimated as in Appendix A.

The estimating equation has the same form as (2.2.4), except for two important differences. First, the propensity score model in this section conditions on additional $\overline{cumA}_{i,k}^*$. In the discrete time model of Section 2.2.1, $\overline{cumA}_{i,k}^*$ would be a transformed version of $\bar{A}_{i,k-1}^*$, and was redundant information. However, with continuous time underlying processes, $\overline{cumA}_{i,k}^*$ is new information on the treatment history. Second, the putative counterfactual $Y_{i,m}^{0*}(\Psi)$ is calculated by subtracting the $cumA_{i,k}^*$ from $Y_{i,m}^*$, instead of $\sum_{l=0}^{k-1} A_{i,l}^*$. We will refer later to the g-estimation in this section as the modified g-estimation (although it is the true spirit of g-estimation). The justification and limitation of using the modified g-estimation will be discussed in Section 2.3.

We refer to the g-estimation in Section 2.2.1 as naive g-estimation, when it is applied to data from continuous time model. When the data come from a continuous time model, the naive g-estimation can be severely biased, as we will show in our simulation study and the diarrhea application. One source of bias is a measurement

error problem, $\sum_{l=0}^{k-1} A_{i,l}^*$ is not the correct measure of the treatment; another source of bias is that important information $\overline{cumA}_{i,k}$ is not conditioned on in the propensity score. Although we would not expect researchers to use the naive g-estimation when the true cumulative treatments are available, we present the simulation and real application results using this method as a reference to show how severely biased the estimates would be had we not known the true cumulative treatments and the measurement error problem had dominated.

2.3 Justification of the Use of the Modified Discrete Time G-estimation

Given discrete time observational data from continuous time underlying processes, solving equation (2.2.7) provides an estimate for Ψ . For this Ψ estimate to be consistent, an analogue to condition (2.2.2) is needed:

$$pr(A_k^* | \bar{L}_k^*, \bar{A}_{k-1}^*, \overline{cumA}_k^*, \underline{Y}_{k+}^{*0}) = pr(A_k^* | \bar{L}_k^*, \bar{A}_{k-1}^*, \overline{cumA}_k^*) \quad (2.3.1)$$

Condition (2.3.1) is a requirement on the observed variables at discrete times. It is not a condition at the level of the data-generating process, and cannot be easily used to determine the appropriateness of using the modified g-estimation by domain knowledge. This contrasts with the case of discrete time ignorability for the discrete time data. In this section, we will seek conditions at the data generating process level that can justify the use of g-estimation.

2.3.1 Sequential Randomization at Any Finite Subset of Time Points

Recall the data generating process described in Section 2.1.2. The third party organization periodically (e.g., semi-annually) collects the health data and treatment records of the patients. Suppose a researcher thinks (2.3.1) holds for the time points at which the third party organization collects these data. If the time points have not been chosen in a special way to make (2.3.1) hold, then the researcher will often be willing to make the stronger assumption that (2.3.1) would hold for any finite subset of time points at which the third party organization chose to collect data. For example, for the diarrhea study, the survey was actually conducted in November, 1998, March-April, 1999 and November, 1999. If a researcher thought (8) held for these three time points, then she might be willing to assume that (2.3.1) should also hold if instead the survey was conducted in December, 1998, February, 1999, May, 1999 and October, 1999.

Before formalizing the researcher's assumption on any finite subset of time points, we observe that

Proposition 2.3.1. *Under the deterministic model assumption (2.2.6), the true model for the propensity score has the following property:*

$$pr(A_k^* = 1 | \bar{L}_k^*, \bar{A}_{k-1}^*, \overline{cumA_k^*}) = pr(A_k^* = 1 | \bar{L}_k^*, \bar{A}_{k-1}^*, \bar{Y}_k^{0*}) \quad (2.3.2)$$

Proof. Under the deterministic assumption (2.2.6) and the correct Ψ , $(\bar{L}_k^*, \bar{A}_{k-1}^*, \overline{cumA_k^*})$ is a one-to-one transformation of $(\bar{L}_k^*, \bar{A}_{k-1}^*, \bar{Y}_k^{0*})$. \square

Using Proposition 2.3.1, we state sequential randomization assumption at any finite subset of time points as

Definition 2.3.2. A càdlàg process $Z_t \equiv (L_t, A_t, Y_t^0)$, $0 \leq t \leq K$ is said to satisfy the **finite time sequential randomization** assumption, or **FTSR**, if for any finite subset of time points, $0 \leq t_1 < t_2 < \dots < t_n < t_{n+1} < \dots < t_{n+l} \leq K$, we have

$$pr(A_{t_n} | \bar{L}_{t_n-}, \bar{A}_{t_{n-1}}, \bar{Y}_{t_n-}^0, \underline{Y}_{t_n+}^0) = pr(A_{t_n} | \bar{L}_{t_n-}, \bar{A}_{t_{n-1}}, \bar{Y}_{t_n-}^0) \quad (2.3.3)$$

where $\bar{L}_{t_n-} = (L_{t_1-}, L_{t_2-}, \dots, L_{t_n-})$, $\bar{A}_{t_{n-1}} = (A_{t_1}, A_{t_2}, \dots, A_{t_{n-1}})$, $\bar{Y}_{t_n-}^0 = (Y_{t_1-}^0, Y_{t_2-}^0, \dots, Y_{t_n-}^0)$, and $\underline{Y}_{t_n+}^0 = (Y_{t_{n+1}-}^0, Y_{t_{n+2}-}^0, \dots, Y_{t_{n+l}-}^0)$.

Finite time sequential randomization assumption obviously justifies the use of g-estimation in the settings described in Section 2.2.2. It does not refer to continuous time sequential randomization directly. However, the following theorem shows that it is a stronger assumption than the CTSR assumption.

Theorem 2.3.3. *If a continuous time càdlàg process Z_t satisfies finite time sequential randomization, under some regularity conditions, it will also satisfy continuous time sequential randomization.*

Proof. See the appendices in Section 5.3. The regularity conditions are also stated in Section 5.3. □

The result of Theorem 2.3.3 is natural. As mentioned in Section 2.1.3, the continuous time sequential randomization does not imply FTSR, because in discrete time

observations, we do not have the full continuous time history to control. To compensate for the incomplete data problem, some stronger assumption on the continuous time processes has to be made if identification is to be achieved.

2.3.2 A Markovian Condition

Given the finite time sequential randomization assumption described above, two important questions arise. First, Theorem 2.3.3 shows that the FTSR assumption is stronger than the continuous time sequential randomization assumption. It is natural to ask how much stronger it is than the CTSR assumption. Secondly, the FTSR assumption has a descriptive nature, and unlike the usual sequential randomization assumption (A1), it is not an assumption on the data generating process and thus is not useful for incorporating domain knowledge to justify itself. A condition at the data generating process level will be more helpful for researchers in deciding whether g-estimation is valid.

We answer both questions partially in the following theorem.

Theorem 2.3.4. *Assuming that the process (Y_t^0, L_t, A_t) follows the continuous time sequential randomization assumption, and that the process (Y_{t-}^0, L_{t-}, A_t) is Markovian, for any time t and $t + s$, $s > 0$, we have*

$$pr(A_t | L_{t-}, Y_{t-}^0, Y_{t+s}^0) = pr(A_t | L_{t-}, Y_{t-}^0), \quad (2.3.4)$$

which implies the finite time sequential randomization assumption.

Proof. The proof can be found in the Section 5.4. □

We make the following comments on the theorem.

- The theorem partially answers our first question - the FTSR assumption is stronger than the CTSR assumption, but the gap between the two assumptions is less than a Markovian assumption. The result is not surprising, as with missing covariates between observational time points, we would hope that the variables at the observational time points well summarize the missing information. The Markovian assumption guarantees that variables at a observational time point summarize all information prior to that time point.
- The theorem also partially answers our second question. The CTSR assumption is usually justified by domain knowledge of how treatments are decided. Theorem 2.3.4 suggests that the researchers could further look for biological evidence that the process is Markovian to validate the use of g-estimation. The Markovian assumption can also be tested. One could first use the modified g-estimation to estimate the causal parameter, construct the Y^0 process at the observational time points, and then test whether the full observational data of A, L, Y^0 come from a Markov process. A strict test of whether the discretely observed panel data come from a continuous time (usually non-stationary) Markov process could be difficult and is beyond the scope of this chapter. As a starting point, we suggest Singer's trace inequalities (Singer, 1981) as a criterion to test for the Markovian property. A weaker test of the Markovian property is to test conditional independence of past observed values and future observed values

conditioning on current observed values.

- In the theorem, equation (2.3.4) looks like an even stronger version of continuous time sequential randomization assumption - the treatment decision seems to be based only on current covariates and current potential outcomes. One could of course directly assume this stronger version of randomization and apply g-estimation. However, Theorem 2.3.4 is more useful as we are assuming a weaker untestable CTSR assumption and a Markovian assumption that is testable in principle.
- The theorem suggests that it is sufficient to control for current covariates and current potential outcomes for g-estimation to be consistent. In practice, we advise controlling for necessary past covariates and treatment history. The estimate would still be consistent if the Markovian assumption is true, and it might reduce bias when the Markovian assumption is not true. As a result, we do control for previous covariates and treatments in our simulation and application to the diarrhea data.
- It is worth noting that the labeling of time is arbitrary. In practice, researchers can label whatever they have controlled for in their propensity score as the “current” covariates, which could include covariates and treatments that are measured or assigned previously in physical time. In this case the dimension of the process that needs to be tested for the Markovian property should also

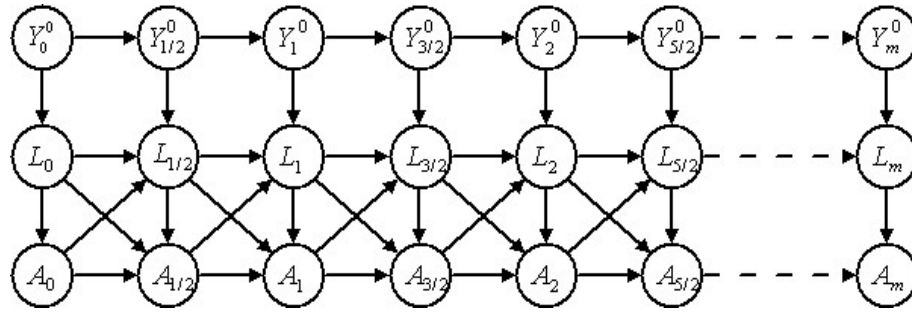
be expanded to include older covariates and treatments in physical time.

- Finally, we note that a discrete time version of the theorem is implied by Corollary 4.2 of Robins (1997), if we set, in his notation, U_{ak} to be the covariates between two observational time points and U_{bk} to be the null set.

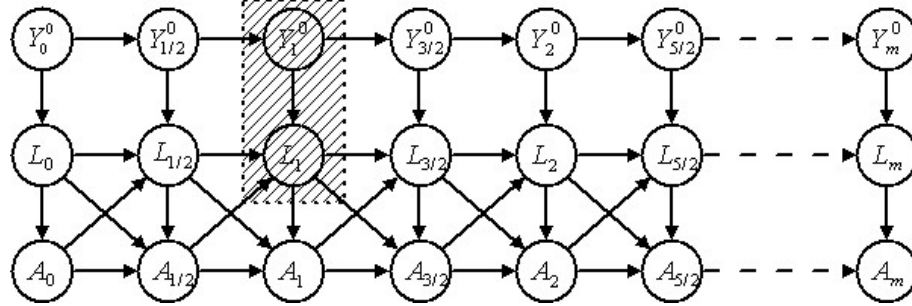
As a discretized example, we illustrate the idea of Theorem 2.3.4 by a directed acyclic graph (DAG) in part (a) of Figure 2.1, which assumes that all variables can only change values at time points $0, 1/2, 1, 3/2, 2, \dots, m$. Note that we do not distinguish the left limit of variables and themselves in all DAGs of this chapter, for reasons discussed in Appendix C. We also assume that the process can only be observed at time $0, 1, 2, \dots, m$. It is easy to verify that the DAG satisfies sequential randomization at the $0, 1/2, 1, 3/2, 2, \dots, m$ time level. The DAG is also Markovian in time. For example, if we control A_1, L_1, Y_1^0 , any variable prior to time 1 will be d-separated from any other variable after time 1.

Part (b) of Figure 2.1 verifies that A_1 is d-separated from $Y_m^0, m > 1$ by the shaded variables, namely, L_1 and Y_1^0 , which justifies equation (2.3.2). As implied by Theorem 2.3.4, the modified g-estimation works for data observed at the integer times if they are generated by the model defined by this DAG.

It is true that the Markovian condition that justifies the g-estimation equation (2.2.7) is restrictive as will be discussed in the following section. However, our simulation study shows that g-estimation has some level of robustness when the Markovian assumption is not seriously violated.



(a) DAG of a Markovian Process



(b) Verification of Equation (2.3.2)

Figure 2.1: Directed Acyclic Graph

2.4 The Controlling-the-future Method

In this section, we consider situations in which the observational time sequential randomization fails and seek methods that are more robust to this failure than the modified g-estimation given in Section 2.2.3. The method we are going to introduce is proposed by Joffe and Robins (2009), which deals with a more general case of the existence of unmeasured confounders. It can be applied to deal with unmeasured confounders coming from either a subset of contemporaneous covariates or a subset of covariates that represent past time, the latter case being of interest for this chap-

ter. The method, which we will refer to as the controlling-the-future method (the reason for the name will be more clear later on), gives consistent estimates when g-estimation is consistent, and it produces consistent estimates in some cases even when g-estimation is severely inconsistent.

In what follows, we will first describe an illustrative application of the controlling-the-future method, and then discuss its relationship with our framework of g-estimation in continuous time processes with covariates observed at discrete times.

2.4.1 Modified Assumption and Estimation of Parameters

We assume the same continuous time model as in Section 2.2.2. Following Joffe and Robins (2009), we consider a revised sequential randomization assumption on variables at the observational time points

$$pr(A_k^* | \bar{L}_k^*, \bar{A}_{k-1}^*, \overline{cumA}_k^*, \underline{Y}_{k+}^{0*}) = pr(A_k^* | \bar{L}_k^*, \bar{A}_{k-1}^*, \overline{cumA}_k^*, Y_{k+1}^{0*}) \quad (2.4.1)$$

This assumption relaxes (2.3.1). At each time point, conditioning on previous observed history, the treatment can depend on future potential outcomes, but only on the next period's potential outcome. In Joffe and Robins' extended formulation, this can be further relaxed to allow for dependence on more than one period of future potential outcomes, as well as other forms of dependence on the potential outcomes..

If the revised assumption (2.4.1) is true, we obtain a similar estimating equation as (2.2.7). For each putative Ψ , we map Y_k^* to

$$Y_k^{0*}(\Psi) = Y_k^* - \Psi cumA_k^*$$

the potential outcome if the subject never received any treatment under the hypothesized treatment effect Ψ .

Define the putative propensity score as

$$p_k(\Psi) \equiv pr(A_k^* = 1 | \bar{L}_k^*, \bar{A}_{k-1}^*, \overline{cumA}_k^*, Y_{k+1}^{0*}(\Psi)) \quad (2.4.2)$$

Under assumption (2.4.1), the correct Ψ should solve

$$U(\Psi) = E\left\{ \sum_{\substack{1 \leq i \leq n \\ k+1 < m \leq K}} [A_{i,k}^* - p_{i,k}(\Psi)] g(Y_{i,m}^{0*}(\Psi), X_{i,k}^*, h_{i,k}) \right\} = 0 \quad (2.4.3)$$

where $X_{i,k}^* = (\bar{L}_{i,k}^*, \bar{A}_{i,k-1}^*, \overline{cumA}_{i,k}^*)$, $h_{i,k} = Y_{i,k+1}^{0*}(\Psi)$, and g is any function and can be generalized to functions of $X_{i,k}^*$, $h_{i,k}$ and any number of future potential outcomes that are later than time $k + 1$, e.g. $g(Y_{i,k+2}^{0*}(\Psi), Y_{i,k+3}^{0*}(\Psi), X_{i,k}^*, h_{i,k})$. In most real applications, the model for $p_k(\Psi) = E[A_k^* | X_k^*, h_k]$ is unknown, and is usually estimated by a parametric model

$$p_{i,k}(\Psi; \beta_X, \beta_h) = E[A_{i,k}^* | X_{i,k}^*, h_{i,k}; \beta_X, \beta_h].$$

We can solve the following set of estimating equations to obtain the estimates of Ψ , β_X and β_h

$$\begin{aligned} & U(\Psi, \beta_X, \beta_h) \\ &= \sum_{\substack{1 \leq i \leq n \\ k+1 < m \leq K}} (A_{i,k}^* - p_{i,k}(\Psi; \beta_X, \beta_h)) [g(Y_{i,m}^{0*}(\Psi), X_{i,k}^*, h_{i,k}), X_{i,k}^*, h_{i,k}]^T = 0 \end{aligned} \quad (2.4.4)$$

The estimation of the covariance matrix of Ψ , β_X and β_h is similar to the usual standard g-estimation, which is described in Appendix A.

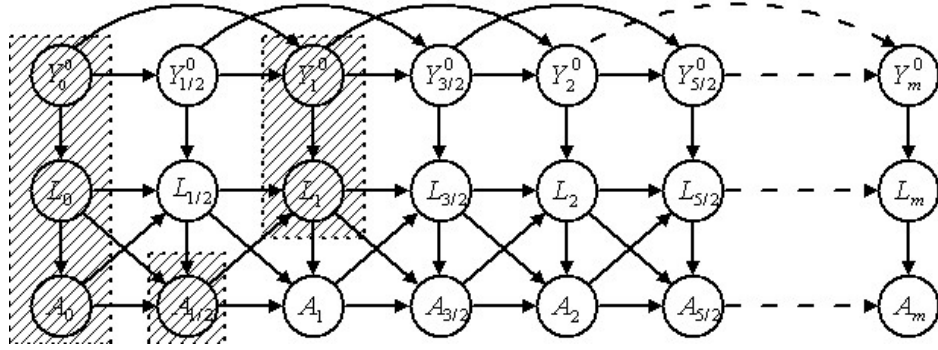
Two important features of estimating equation (2.4.4) distinguish it from estimating equation (2.2.7). First, in (2.4.4) there is a common parameter Ψ in both p_k 's model and $Y_m^{0*}(\Psi)$, caused by the fact that the treatment depends on future potential outcome. Second, in (2.4.4), the sum over m and k is restricted to $m > k + 1$, while in (2.2.7) we only need $m > k$. If we use $m = k + 1$ in (2.4.4), $E\{[A_{i,k}^* - p_{i,k}(\Psi)]g(Y_{i,k+1}^{0*}(\Psi), X_{i,k}^*, h_{i,k})\} = 0$ usually does not lead to the identification of Ψ , unless certain functional forms of the propensity score model are assumed to be true (see Joffe and Robins 2009).

2.4.2 The Controlling-the-future Method and the Markovian Condition

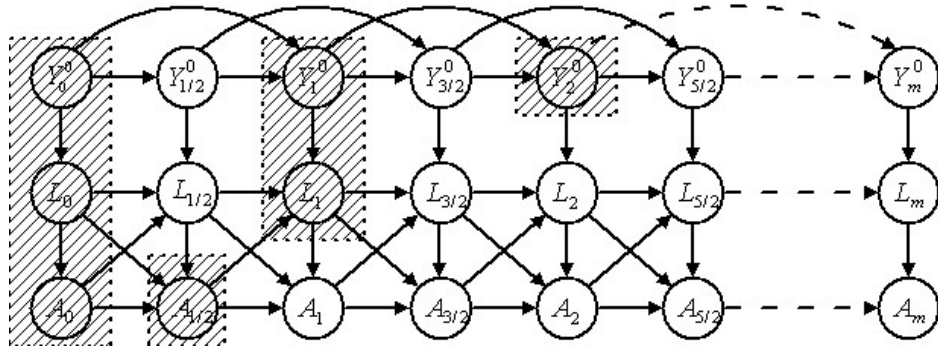
Joffe and Robins' revised assumption (2.4.1) is an assumption on the discrete time observational data. It relaxes the observational time sequential randomization (2.3.1) because (2.3.1) always implies (2.4.1). At the continuous time data generating level, (2.4.1) allows less stringent underlying stochastic processes than the Markovian process in Theorem 2.3.4.

In particular, we identify two important scenarios where the relaxation happens. One scenario is to allow for more direct temporal dependence for the Y^0 process, which we will refer to as the *non-Markovian- Y^0* case. The other scenario is to allow colliders in L , which we will refer to as the *leading-indicator-in- L* case. We illustrate both cases by modifying the directed acyclic graph (DAG) example in Figure 2.1.

The Non-Markovian- Y^0 Case



(a) Not control for future Y_t^0



(b) Control for future Y_t^0

Figure 2.2: Directed Acyclic Graph with Non-Markovian Y_t^0

Assume for example, our data is generated from the DAG in Figure 2.2 where we allow the dependence of Y_2^0 on Y_1^0 , even if $Y_{3/2}^0$ is controlled. In part (a) of Figure 2.2, we control for observed covariates (L_0, L_1) , treatment $(A_0, A_{1/2})$ and current and historical potential outcome (Y_0^0, Y_1^0) for treatment at time 1 (A_1), i.e., we have controlled for all historically observed covariates, treatment and cumulative treatment as suggested in the comments for Theorem 2.3.4. In this case, the modified g-estimation fails, because the paths like $A_1 \leftarrow L_{1/2} \leftarrow Y_{1/2}^0 \rightarrow Y_{3/2}^0 \rightarrow Y_2^0 \rightarrow \dots \rightarrow Y_m^0$

are not blocked by the shaded variables. In part (b) of Figure 2.2, we control for the additional Y_2^0 . A_1 is not completely blocked from Y_m^0 , but some paths that are not blocked in part (a) are now blocked, for example, the path of $A_1 \leftarrow L_{1/2} \leftarrow Y_{1/2}^0 \rightarrow Y_{3/2}^0 \rightarrow Y_2^0 \rightarrow Y_{5/2}^0 \rightarrow \dots \rightarrow Y_m^0$. Also, no additional paths are opened by conditioning on Y_2^0 . We would usually expect that the correlation between A_1 and Y_m^0 is weakened. Under the framework of Joffe and Robins (2009), we can control for more than one period of future potential outcomes, and expect to further weaken the correlation between A_1 and Y_m^0 . A modification of assumption (2.4.1) that conditions on more future potential outcomes may be approximately true.

The scenario relates to real world problems. For instance, in the diarrhea example, Y_t^0 is the natural height growth of a child without any occurrence of diarrhea. Height in the next month not only depends on current month's height, but also depends on previous month's height: the complete historical growth curve of the child provides information on genetics and nutritional status, and provides information about future natural height beyond that of just current natural height. Therefore, the potential height process for the child is not Markovian. (For a formal argument why children's height growth is not Markovian, see Gasser, T. et al. 1984.) By the reasoning we discussed above, g-estimation fails. However, if we assume that the delayed dependence of natural height wanes out after a period of time (like in Figure 2.2), controlling for the next period potential height in the propensity score model might weaken the relationship between current diarrhea exposure and future potential

height later than the next period, and the assumptions of the controlling-the-future method might hold approximately.

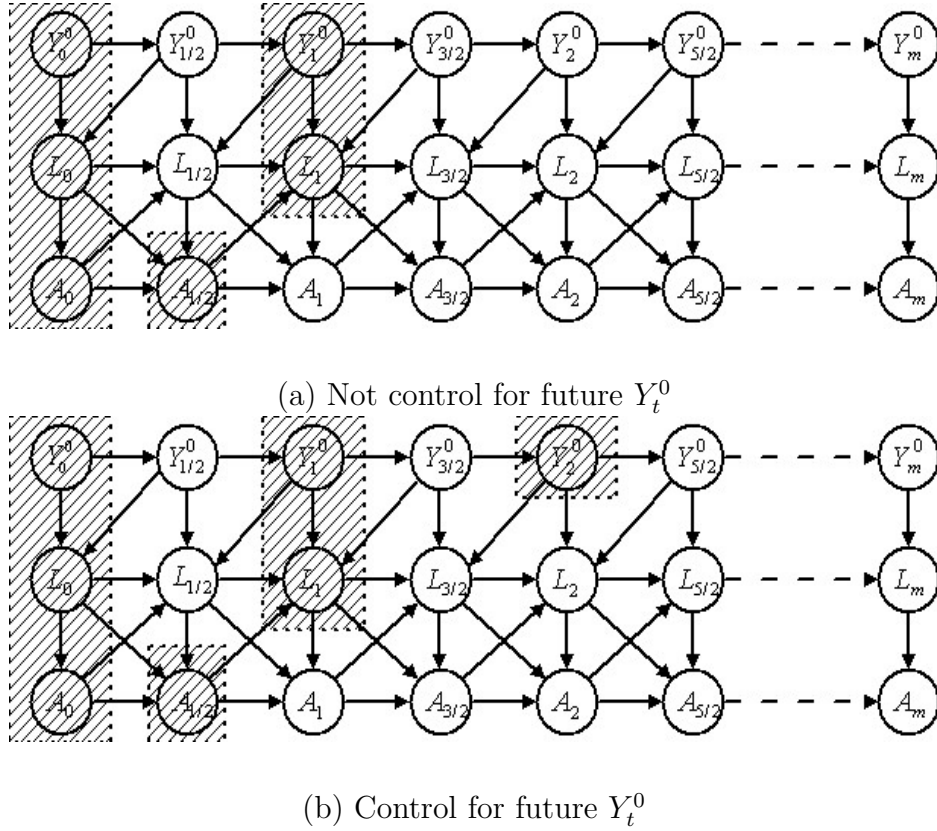


Figure 2.3: Directed Acyclic Graph with Leading Indicator in L_t

The Leading-indicator-in- L Case

In Figure 2.1, we do not allow any arrows from future Y^0 to previous L , which means that among all measures of the subject, there is no elements in L that contain any leading information about future Y^0 . This means that Y^0 is a measure that is ahead of all other measures. This is not realistic in many real world problems. In the example of the effect of the diarrhea on height, weight is an important covariate.

While both height and weight reflect the nutritional status of a child, malnutrition usually affects weight more quickly than height, i.e., the weight contains leading information for the natural height of the child. Figure 2.1 is then not an appropriate model for studying the effect of the diarrhea on height.

In Figure 2.3, we allow arrows from $Y_{1/2}^0$ to L_0 , from Y_1^0 to $L_{1/2}$ and so on, which assumes that L contains leading indicators of Y^0 , but the leading indicators are only ahead of Y^0 for less than one unit of time. Part (a) of Figure 2.3 shows that controlling for history of covariates, treatment and potential outcomes does not block A_1 from Y_m^0 . On the path of $A_1 \leftarrow L_{1/2} \rightarrow L_1 \leftarrow Y_{3/2}^0 \rightarrow Y_2^0 \rightarrow Y_{5/2}^0 \rightarrow \dots \rightarrow Y_m^0$, L_1 is a controlled collider. However, in part (b), if we do control for Y_2^0 additionally, the same path will be blocked. In general, if we assume that there exist leading indicators in covariates and that the leading indicators are not ahead of potential outcomes for more than one time unit, g-estimation will fail, but the controlling-the-future method will produce consistent estimates.

The fact that the controlling-the-future method can work in the leading information scenario can also be related to the discussion of Section 3.6 of Rosenbaum (1984). The main reason for g-estimation's failure in the DAG example is that $L_{1/2}$ is not observable and cannot be controlled. If $L_{1/2}$ is observed, it is easy to verify that the DAG in Figure 2.3 satisfies sequential randomization on the finest time grid. The idea behind the controlling-the-future method is to condition on a "surrogate" for $L_{1/2}$. The surrogate should satisfy the property that Y_m^0 is independent of the un-

observed $L_{1/2}$ given the surrogate and other observed covariates, (similar to formula 3.17 in Rosenbaum (2007)). In the leading information case, when $m > k + 1$ and we have covariates \bar{L}_k that are only ahead of the potential outcome until time at most $k + 1$, then the future potential outcome Y_{k+1}^0 is a surrogate. It is easy to check that in Figure 2.3, $L_{1/2}$ is independent of Y_m^0 , given Y_2^0 , L_1 , A_0 , and $cumA_1$ (equivalently Y_1^0).

It is worth noting that we do not need to control for anything except Y_2^0 in Figure 2.3 to get a consistent estimate. It is possible to construct more complicated DAGs in which controlling for additional past and current covariates is necessary, which involves more model specifications for the relationships among different covariates and deviates from the main point of this chapter.

In Section 2.5, we will simulate data in cases of non-Markovian- Y_t^0 and leading-indicator-in- L_t respectively, and show that the controlling-the-future method does produce better estimates than the g-estimation. However, it worth noting that when the modified g-estimation in Section 2.2.3 is consistent, the controlling-the-future estimation is considerably less efficient, because it uses less data and estimates more parameters.

2.5 Simulation Study

We set up a simple continuous time model that satisfies sequential ignorability in continuous time, and simulate and record discrete time data from variations of the

simple model. We estimate causal parameters from both the modified g-estimation and the controlling-the-future estimation. We also present the estimates from naive g-estimation in Section 2.2.1, where we ignore the continuous time information of the treatment processes, as a reference to show the severity of the bias in presence of the measurement error problem. The results support the discussions in Section 2.3 and Section 2.4

In the simulation models below, M1 satisfies the Markovian condition in Theorem 2.3.4. It also serves as a proof that there exist processes satisfying the conditions in Theorem 2.3.4.

2.5.1 The Simulation Models

We first consider a continuous time Markov model, which satisfies the CTSR assumption.

- Y_t^0 is the potential outcome process if the patient is not receiving any treatment.

We assume that

$$Y_t^0 = g(V, t) + e_t$$

where $g(V, t)$ is a function of baseline covariates V and time t . Let $g(V, t)$ be continuous in t , and e_t follows an Ornstein-Uhlenbeck process, i.e.

$$de_t = -\theta e_t dt + \sigma dW_t$$

where W_t is the standard Brownian motion.

- Y_t is the actual outcome process and follows the deterministic model (2.2.6):

$$Y_t = Y_t^0 + \Psi \int_0^t A_s ds$$

- A_t is the treatment process, taking binary values. The jump of the A_t process follows the following formula:

$$pr(A_s \text{ jumps once from } (t, t+h] | \bar{A}_t, \bar{Y}_t, \bar{Y}^0) = s(A_t, Y_t)h + o(h)$$

$$pr(A_s \text{ jumps more than once from } (t, t+h] | \bar{A}_t, \bar{Y}_t, \bar{Y}^0) = o(h)$$

where \bar{A}_t and \bar{Y}_t are the full continuous time history of treatment and outcome up to time t , and \bar{Y}^0 is the full continuous time path of potential outcome from time 0 to time K . By making $s(\cdot)$ independent of \bar{Y}^0 , we make our model satisfy the continuous time sequential randomization assumption.

In this model, the only time-dependent confounder is the outcome process itself.

We also consider several variations of the above model (denoted as M1 in below):

- Model (M2) extends (M1) to the non-Markovian- Y_t^0 case. Specifically, we consider e_t in the model of Y_t^0 follows a non-Markovian process, namely an Ornstein-Uhlenbeck process in random environments, which is defined as the following:

1. J_t is a continuous time Markov process taking values in a finite set $\{1, \dots, m\}$, which is the environment process.
2. we have $m > 1$ sets of parameters $\theta_1, \sigma_1, \dots, \theta_m, \sigma_m$.

3. e_t follows an Ornstein-Uhlenbeck process with parameters θ_j, σ_j , when $J_t = j$; the starting point of each diffusion is chosen to be simply the end point of the previous one.

- Model (M3) extends (M1) to another setting of non-Markovian Y_t^0 process, where

$$Y_t^0 = g(V, t) + 0.8e_{t-1} + 0.2e_t$$

e_t follows the same Markovian Ornstein-Uhlenbeck process as in M1. Every other variable is the same as in M1.

- Model (M4) considers the case with more than one covariate. In M4, we keep the assumptions on Y_t^0 as in (M1) and the deterministic model of Y_t . We add one more covariate, which is generated as follows

$$L_t^- = 0.2Y_t + 0.8Y_{t+0.5}^0 + 0.5\eta_t$$

η_t follows an Ornstein-Uhlenbeck process independent of the Y_t^0 process. In this specification, the covariate L_t^- contains some leading information about Y^0 , but it is only ahead of Y^0 for 0.5 length of a time unit. Here we use L_t^- instead of L_t to denote that it is the covariate other than Y_t itself. The simulation model for A_t process is given in Appendix E.

In all these models, to simulate data, we use $g(V, t) = C$ a constant, $\Psi = 1$, a time span from 0 to 5, and a sample size of 5000. Details of other parameter specifications can be found in Section 5.5. We generate 5000 continuous paths of Y_t and A_t

(and L_t^- in M4), from time 0 to time 5, and record $Y_0^*, A_0^*, Y_1^*, A_1^*, \dots, Y_4^*, A_4^*, Y_5^*$ and $cumA_1^*, \dots, cumA_5^*$ (and $L_0^{-*}, \dots, L_4^{-*}$ in M4) as the observed data.

2.5.2 Estimations and Results under M1

Figure 2.4 shows a typical continuous time path of Y_t^0, Y_t and A_t . The treatment switches around time 0.7 and time 2.8.

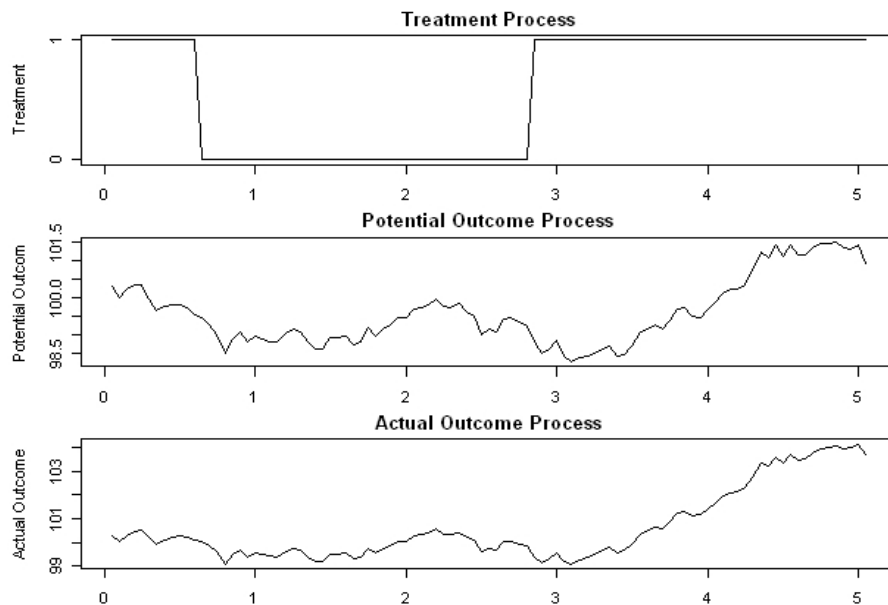


Figure 2.4: Example of Continuous Time Paths Under M1

We apply three estimating methods on data simulated from M1: the naive discrete time g-estimation described in Section 2.2.1, which ignores the underlying continuous time processes; the modified g-estimation described in Section 2.2.3, which controls for the all the observed discrete time history; and the controlling-the-future method in Section 2.4.1 of controlling for the next period's potential outcome in addition to

the discrete time history.

For estimation, even though we know the data generating process, it is too complicated to use the correct model for the propensity score, i.e., the correct functional form for $p_k(\Psi) \equiv pr(A_k^* | \bar{L}_k^*, \bar{Y}_k^*, \bar{A}_{k-1}^*, \overline{cumA}_k^*, Y_{k+}^{0*}(\Psi))$. Therefore, we use the following approximations (Note that we control for past treatment and covariates as well. See comments for Theorem 2.3.4):

1. Standard g-estimation ignoring continuous time processes (naive g-estimation)

$$\text{logit}(p_k(\Psi)) = \beta_0 + \beta_1 A_{k-1}^* + \beta_2 Y_{k-1}^* + \beta_3 Y_k^*$$

2. G-estimation controlling for all observed history (modified g-estimation)

$$\text{logit}(p_k(\Psi)) = \beta_0 + \beta_1 A_{k-1}^* + \beta_2 Y_{k-1}^* + \beta_3 Y_k^* + \beta_4 cumA_k^*$$

3. The controlling-the-future method, controlling for next period potential outcomes (controlling-the-future estimation)

$$\text{logit}(p_k(\Psi)) = \beta_0 + \beta_1 A_{k-1}^* + \beta_2 Y_{k-1}^* + \beta_3 Y_k^* + \beta_4 cumA_k^* + \beta_5 Y_{k+1}^{0*}(\Psi)$$

We plug these models for the propensity scores in estimation equations (2.2.5), (2.2.7) and (2.4.4) respectively. (Note in equation (2.2.5), $Y_k^{0*}(\Psi) = Y_k^* - \Psi \sum_{l=0}^{k-1} A_l^*$, while in the other two, $Y_k^{0*}(\Psi) = Y_k^* - \Psi cumA_k^*$.)

The first panel of Table 2.1 shows a summary of the estimates of causal parameters for 1000 simulations from M1. The naive g-estimation gives severely biased estimates. Controlling for all observed history and controlling for additional next

period potential outcome both give us unbiased estimates. As we discussed at the end of Section 2.4.2, the controlling-the-future method has lower efficiency.

The last row of the first panel in Table 2.1 shows the coverage rate of 95% confidence interval estimated from the 1000 independent simulations. Naive g-estimation has a zero coverage rate, while the other two methods have coverage rates around 95%.

2.5.3 Simulation Results under M2 and M3

In M2, we generate data from a non-Markovian Y_t^0 , namely the Ornstein-Uhlenbeck process in random environment. The results in the second panel of Table 2.1 are typical for different values of parameters under M2. The naive g-estimation performs badly, while both the other methods still work fine with the data generated from M2. This shows that the modified g-estimation and the controlling-the-future method have some level of robustness to mild violations of the Markovian assumption.

The third part of Table 2.1 shows the results of simulation from M3, where Y^0 violates Markov property more substantially. In this case, we can see that the mean of the modified g-estimates is biased, but the mean of the controlling-the-future estimates is almost unbiased. In last row of the third panel, the coverage rate for the modified g-estimation drops to 0.855, while the controlling-the-future method still has a coverage rate of 0.956.

Table 2.1: Estimated Causal Parameters from Data Generated by M1-4
Simulation Results from M1

<i>True Parameter = 1</i>			
	Naive g-est.	mod. g-est.	ctr-future est.
Mean Estimate [†]	0.7728	1.0005	0.9988
S.D. of Estimates [‡]	0.0183	0.0191	0.0403
S.D. of the Mean Estimate [*]	0.0005	0.0006	0.0013
Absolute Bias ^{**}	0.2272	0.0005	0.0012
Coverage [◇]	0	0.946	0.956

Simulation Results from M2

<i>True Parameter = 1</i>			
	Naive g-est.	mod. g-est.	ctr-future est.
Mean Estimate [†]	0.7651	1.0016	1.0000
S.D. of Estimates [‡]	0.0132	0.0158	0.0371
S.D. of the Mean Estimate [*]	0.0004	0.0005	0.0012
Absolute Bias ^{**}	0.2349	0.0016	0.0000
Coverage [◇]	0	0.953	0.950

Simulation Results from M3

<i>True Parameter = 1</i>			
	Naive g-est.	mod. g-est.	ctr-future est.
Mean Estimate [†]	0.7580	0.9845	1.0026
S.D. of Estimates [‡]	0.0149	0.0180	0.0487
S.D. of the Mean Estimate [*]	0.0005	0.0006	0.0015
Absolute Bias ^{**}	0.2420	0.0155	0.0026
Coverage [◇]	0	0.855	0.956

Simulation Results from M4

<i>True Parameter = 1</i>			
	Naive g-est.	mod. g-est.	ctr-future est.
Mean Estimate [†]	0.7816	1.0853	1.0085
S.D. of Estimates [‡]	0.0201	0.0289	0.0806
S.D. of the Mean Estimate [*]	0.0006	0.0009	0.0025
Absolute Bias ^{**}	0.2184	0.0853	0.0085
Coverage [◇]	0	0.115	0.948

[†] Averaged Over Estimates from 1000 Independent Simulations of Sample Size 5000.

[‡] Sample Standard Deviation of the 1000 Estimates.

^{*} Sample S.D./ $\sqrt{1000}$.

^{**} Absolute Value of (1-Mean Estimates).

[◇] Coverage Rate of 95% Confidence Intervals for 1000 Simulations.

2.5.4 Estimations and Results under M4

In M4, we create a covariate L_t^- that has leading information about Y_t^0 . In the data simulated from M4, the observational time sequential randomization (2.3.1) no longer holds, although the data are generated following continuous time sequential randomization. This simulation serves as a numerical proof to the claim that continuous time sequential randomization does not imply discrete time sequential randomization.

To show this, we run the following two logistic regression model for $k = 2$ and $m = 4$

- Not controlling for the next period potential outcome (used in modified g-estimation)

$$\begin{aligned} \text{Logit}(P(A_k^* = 1)) = & \beta_0 + \beta_1 \text{cum}A_k^* + \beta_2 L_{k-1}^{-*} + \beta_3 L_k^{-*} + \beta_4 A_{k-1}^* \\ & + \beta_5 Y_k^* + \beta_6 Y_{k-1}^* + \beta_8 Y_m^{0*} \end{aligned} \quad (2.5.1)$$

- Controlling for the next period potential outcome (used in controlling-the-future estimation)

$$\begin{aligned} \text{Logit}(P(A_k^* = 1)) = & \beta_0 + \beta_1 \text{cum}A_k^* + \beta_2 L_{k-1}^{-*} + \beta_3 L_k^{-*} + \beta_4 A_{k-1}^* \\ & + \beta_5 Y_k^* + \beta_6 Y_{k-1}^* + \beta_7 Y_{k+1}^{0*} + \beta_8 Y_m^{0*} \end{aligned} \quad (2.5.2)$$

We can use the true values of Y_{k+1}^{0*} and Y_m^{0*} in the regression to test the discrete time ignorability, since the data are simulated by us. Table 2.2 shows the estimates of β_7 and β_8 in both regression models. The result shows that the coefficient of Y_m^{0*} , β_8 , is

Table 2.2: Verification of Observational Time Sequential Randomization Under M4

	Reg. Model 2.5.1*	Reg. Model 2.5.2*
β_7		0.1868
p-value		5.56e-05
β_8	0.0936	0.0134
p-value	0.0006	0.691

* Simulation sample size = 10000

significant if we do not control for the future potential outcome, and is not significant if we control for the future potential outcome. This shows that observational time sequential randomization (2.3.1) does not hold, while the revised assumption (2.4.1) holds.

The estimation results from M4 are in the fourth panel of Table 2.1. In applying these methods we use the following propensity score models separately.

1. G-estimation ignoring the underlying continuous time processes (naive g-estimation)

$$\text{logit}(p_k(\Psi)) = \beta_0 + \beta_1 A_{k-1}^* + \beta_2 Y_{k-1}^* + \beta_3 Y_k^* + \beta_5 L_{k-1}^{-*} + \beta_6 L_k^{-*}$$

2. G-estimation controlling for all observed history (modified g-estimation)

$$\text{logit}(p_k(\Psi)) = \beta_0 + \beta_1 A_{k-1}^* + \beta_2 Y_{k-1}^* + \beta_3 Y_k^* + \beta_4 \text{cum} A_k^* + \beta_5 L_{k-1}^{-*} + \beta_6 L_k^{-*}$$

3. The controlling-the-future method controlling for next period potential outcomes (controlling-the-future estimation)

$$\begin{aligned} \text{logit}(p_k(\Psi)) = & \beta_0 + \beta_1 A_{k-1}^* + \beta_2 Y_{k-1}^* + \beta_3 Y_k^* + \beta_4 \text{cum} A_k^* + \beta_5 L_{k-1}^{-*} \\ & + \beta_6 L_k^{-*} + \beta_7 Y_{k+1}^{0*}(\Psi) \end{aligned}$$

Both the naive g-estimation and the modified g-estimation give us estimates with severe bias, and have coverage rate of 0 and 0.115 separately, for the 95% confidence interval constructed from them. It is worth noting that model 3 is misspecified but, nevertheless, leads to much less biased estimates and the controlling-the-future method has a coverage rate of 0.948.

2.6 Application To The Diarrhea Data

In this section, we apply the different approaches to the diarrhea example mentioned in Section 2.1 (Example 2). We use a set of 224 children with complete records between age 3 and age 6 from 757 households in Bangladesh around 1998. The outcomes, Y_k^* , are the heights of the children in centimeters measured at round k of the interviews, for $k = 1, 2, 3$. The treatment A_k^* at the interview k is defined as $A_k^* = 1$ if the child was sick with diarrhea during the past two weeks of the interview, and $A_k^* = 0$ otherwise. The cumulative treatment $cumA_k^*$ is the number of days the child suffered from diarrhea from four months before the first interview (July 15th, 1998) to the k^{th} interview. Baseline covariates V include age in months, mother's height and whether the household was exposed to the flood. Time dependent covariates other than the outcome, i.e., L_k^{-*} , include mid-upper arm circumference, weight for age z-score, type of toilet (open place, fixed place, unsealed toilet, water sealed toilet or other), garbage disposal method (throwing away in own fixed place, throwing away in own non-fixed place, disposing anywhere or other method), water

purifying process (filter, filter and broil, or other), and source of cooking water (from pond or river/canal, or from tube well, ring well or supply water).

We apply naive g-estimation, modified g-estimation and the controlling-the-future method to this data set. Since we only have three rounds, the actual propensity score models and the estimating equations for the three methods are

- Naive g-estimation uses the following propensity score model

$$\text{logit}\{pr[A_k^* = 1|V, L_k^{-*}, Y_k^*]\} = \beta_0 + \beta_V V + \beta_L L_k^{-*} + \beta_Y Y_k^*$$

where $k = 1, 2$.

The estimating equations follow the form of (2.2.5) in Section 2.2.1:

$$\sum_{\substack{1 \leq k < m \leq 3 \\ 1 \leq i \leq n}} [A_{k,i}^* - pr(A_{k,i}^* = 1|V_i, L_{k,i}^{-*}, Y_{k,i}^*)] \begin{pmatrix} Y_{m,i}^{0*}(\Psi) \\ V_i \\ L_{k,i}^{-*} \\ Y_{k,i}^* \end{pmatrix} = 0$$

where $Y_{m,i}^{0*}(\Psi) = Y_{m,i}^* - \Psi \sum_{l=1}^{m-1} A_l$.

- Modified g-estimation uses this propensity score model

$$\begin{aligned} & \text{logit}\{pr[A_k^* = 1|V, L_k^{-*}, Y_k^*, cumA_k^*]\} \\ & = \beta_0 + \beta_V V + \beta_L L_k^{-*} + \beta_Y Y_k^* + \beta_{cumA} cumA_k^* \end{aligned}$$

where $k = 1, 2$.

The estimating equations follow the form of (2.2.7) in Section 2.2.3.

$$\sum_{\substack{1 \leq k < m \leq 3 \\ 1 \leq i \leq n}} [A_{k,i}^* - pr(A_{k,i}^* = 1 | V_i, L_{k,i}^{-*}, Y_{k,i}^*, cumA_{k,i}^*)] \begin{pmatrix} Y_{m,i}^{0*}(\Psi) \\ V_i \\ L_{k,i}^{-*} \\ Y_{k,i}^* \\ cumA_{k,i}^* \end{pmatrix} = 0$$

where $Y_{m,i}^{0*}(\Psi) = Y_{m,i}^* - \Psi cumA_m$.

- Controlling-the-future estimation uses the following propensity score model

$$\begin{aligned} & \text{logit}\{pr[A_1^* = 1 | V, L_1^{-*}, Y_1^*, cumA_1^*, Y_2^{0*}(\Psi)]\} \\ & = \beta_0 + \beta_V V + \beta_L L_1^{-*} + \beta_Y Y_1^* + \beta_{cumA} cumA_1^* + \beta_{Y_0} Y_2^{0*}(\Psi) \end{aligned}$$

The estimating equations follow (2.4.4) in Section 2.4

$$\sum_{1 \leq i \leq n} [A_{1,i}^* - pr(A_{1,i}^* | V_i, L_{1,i}^{-*}, Y_{1,i}^*, cumA_{1,i}^*, Y_{2,i}^{0*}(\Psi))] \begin{pmatrix} Y_{3,i}^{0*}(\Psi) \\ V_i \\ L_{1,i}^{-*} \\ Y_{1,i}^* \\ cumA_{1,i}^* \\ Y_{2,i}^{0*}(\Psi) \end{pmatrix} = 0$$

where $Y_{3,i}^{0*}(\Psi) = Y_{3,i}^* - \Psi cumA_3$.

The interpretation of Ψ in the last two models is that one day of suffering from diarrhea reduces the height of the child by Ψ centimeters. For naive g-estimation,

Table 2.3: Estimation of Ψ from the Diarrhea Data Set

Method	Estimate	Std Err
Naive g-est.	-0.3991	0.2469
Modified g-est.	-0.3481	0.2832
controlling-the-future est.	-0.0840	0.1894

the underlying data generating model treats the exposure at the observational time as the constant exposure level for the next six months, which does not make sense in the context. Ψ in this model should be interpreted as the effect of having diarrhea at the time of visits, as oppose to having diarrhea at any time, on the height of the child, which does not make too much sense either.

The estimated Ψ and its standard deviation are reported in Table 2.3. Modified g-estimation estimates $\hat{\Psi} = -0.3481$, which means that the height of the child is reduced by 0.35cm if the child has one day of diarrhea. Our controlling-the-future method produces an estimate of $\hat{\Psi} = -0.0840$. Although all the estimates are not significant because of the small sample size, the sign and magnitude of the estimate from the controlling-the-future method are consistent with other research on diarrhea’s effect on height (e.g. Moore et al. 2001).

In addition, we notice that the standard deviation of the modified g-estimate is higher than that of the controlling-the-future estimate. As discussed at the end of Section 2.4.2, if the modified g-estimation is consistent, we would expect that the controlling-the-future estimation will have larger standard deviation. The standard deviations in Table 2.3 provide evidence that the modified g-estimation is not consistent.

2.7 Conclusion

In this chapter, we have studied causal inference from longitudinal data when the underlying processes are in continuous time but the covariates are only observed at discrete times. We have investigated two aspects of the problem. One is the validity of the discrete time g-estimation. Specifically, we investigate a version of g-estimation that follows the spirit of standard discrete time g-estimation but is modified to incorporate the information of the underlying continuous time treatment process, which we referred to as modified g-estimation throughout the chapter. We have shown that an important condition that justifies this g-estimation is the finite time sequential randomization assumption at any subset of time points, which is strictly stronger than the continuous time sequential randomization. We have also shown that a Markovian assumption and the continuous time sequential randomization would imply the FTSR assumption. The Markovian condition is more useful than the FTSR assumption, in the sense that it can potentially help researchers decide whether the application of g-estimation is appropriate. The other aspect is the controlling-the-future method that we propose to use when the condition to warrant g-estimation does not hold. Controlling-the-future method can produce consistent estimates when g-estimation is inconsistent and is less biased in other scenarios. In particular, we identified two important cases in which controlling the future is less biased, namely, when there is delayed dependence in the baseline counterfactual process and when there are leading indicators of the counterfactual process in the covariate process.

In our simulation study, we have shown the performance of the modified g-estimation and the controlling-the-future estimation. The results confirm our discussion in earlier sections. The simulation results have also warned about the danger of applying naive g-estimation, which is usually severely biased and inconsistent when its underlying assumptions are violated, as in the situations considered.

We have applied the g-estimations and the controlling-the-future method to estimating the effect of diarrhea on a child's height, and estimated that its effect is negative but not significant. The real application also provides some evidence that the modified g-estimation is not consistent.

All the discussion in this chapter are based on a particular form of causal model - equation (2.2.6). However, all the arguments could apply to a class of more general rank preserving models, with necessary adjustments in various equations. If we assume a generic rank preserving model with $Y_t = f(Y_t^0, h(\bar{A}_{t-}); \Psi)$, where \bar{A}_{t-} is the continuous time path of A from time 0 to $t-$, h is some functional (e.g., in our chapter $h(\bar{A}_{t-}) = \int_0^t A_s ds$), and f is some strictly monotonic function with respect to the first argument (e.g., in our chapter, $f(x, y; \Psi) = x + \Psi y$), we map Y_k^* to $Y_k^{0*} = f^{-1}(Y_k^*, h(\bar{A}_{k-}); \Psi)$, where f^{-1} is the inverse of $f(x, y; \Psi)$ with respect to x for any given y . We can then substitute all $cumA_k^*$'s in this chapter by the $h(\bar{A}_{k-})$'s. All the discussions and formulas in the chapter would still work, under the assumption that we observe all $h(\bar{A}_{k-})$'s, which can be easily satisfied with detailed continuous time records of the treatment. It should be noted that the argument does

not work if a time-varying covariate modifies the effect of treatment. For example, if $Y_t = Y_t^0 + \Psi \int_0^t L_s^2 A_s ds$, where L_s is a time varying covariate, observing the full continuous time treatment process is not enough. Some imputation for the L_s process is necessary.

The methods considered here have several limitations. These include rank preservation, a strong assumption that the effects of treatment are deterministic. This assumption facilitates interpretation of models. In other work on structural nested distribution and related models (e.g., Robins, 2008), rank preservation has been shown to be unnecessary in settings in which one is not modeling the joint distribution of potential outcomes under different treatments. We expect that this is the case here as well, and work justifying this more formally is in progress.

In this chapter, we also require that cumulative amount of treatment (or the full continuous time treatment process, if using other causal models mentioned above) between the discrete time points when the covariates are observed is known. In a word, we are free of the treatment measurement error problem mentioned in Section 1.2. The next chapter presents an example of getting causal inference when we do have the treatment measurement error problem.

Chapter 3

Causal Inference for a Discretely Observed

Continuous Time Non-stationary Markov Process

In this chapter, we consider an example when both the treatment process and the covariate process are only observed at discrete time points. The problems of unmeasured confounders and the treatment measurement error both arise. In order to identify causal effect in this scenario, it is natural to expect that more modeling assumptions are needed. In this chapter, we propose a continuous time non-stationary Markov model to infer the effect of vitamin A deficiency on respiratory infection among young children. An MCMC algorithm is developed to estimate the model from the discretely observed data. Our simulation and real application show that the model and the estimation algorithm work reasonably well, and we are able to infer a strong effect of vitamin A deficiency on respiratory infection.

3.1 Introduction

Vitamin A deficiency has been reported to have significant consequences in developing countries in terms of child's mortality and morbidity (Vijayaraghavan et al., 1990; West et al., 1991; Daulaire et al., 1992). Severe vitamin A deficiency is usually identified by its ocular manifestation, xerophthalmia, the signs of which include night blindness, conjunctival or corneal xerosis. Beginning in the early 1980s, it has been revealed that even subclinical vitamin A deficiency has broad consequences in child's mortality and morbidity (Humphrey, 1992). In this chapter, we use a subset of the cohort studied by Somer, Katz and Tarwotjo (see Sommer et al., 1983) to study the causal effect of vitamin A deficiency on the occurrence of respiratory disease for young children. In this longitudinal data set, 250 preschool children were examined up to six consecutive quarters for the occurrence of respiratory infection and the presence of xerophthalmia. The covariates of interest include ages in months (centered at 36), gender, cosine and sine terms for annual cycle, and presence of stunting (defined as being below the 85th percentile in height for age of the National Center for Health Statistics (NCHS) standard), which indicates longer term nutritional status.

The same data set has been analyzed by Zeger and Karim (1991) using a logistic model with random effect (we have dropped the variable height for age to simplify computation) and by Zeger and Liang (1991) with a feedback time series model, and important discoveries have been found from this data set. In Zeger and Karim's work, they found significant association between xerophthalmia and respiratory infection

conditional on other covariates, by carefully incorporating within-subject correlations with a random effect term in their logistic regression model. In Zeger and Liang's work, they studied the feedback relationship among xerophthalmia, respiratory infection and diarrheal disease using a multivariate time series model, and found significant evidence for a feedback cycle between xerophthalmia and diarrheal disease but not for a feedback cycle between xerophthalmia and respiratory infection. Both analysis are associational (see Robins et al., 1999).

In this chapter, we would like to make causal inferences of vitamin A deficiency on respiratory infection, while incorporating the dynamic nature of the subject evolving with the time. In particular, we hope to properly adjust for baseline covariates, seasonality and possible feedback cycles among time dependent covariates simultaneously, and answer questions like, what is the probability of a child suffering from respiratory disease a year later, if the child starts taking vitamin A supplement that effectively eliminates any vitamin A deficiency, as compared with the same probability if the child does not take any vitamin A supplement and grows naturally.

To achieve this goal, we model the longitudinal data under Rubin's counterfactual framework (Rubin, 1974), and assume that the covariates, vitamin A deficiency levels, actual outcomes and counterfactual outcomes for the occurrence of respiratory infection follow a continuous time Markov process that is only partially observed at the six discrete follow-up times. Two features of this model make our study novel and of particular interest for its generalization to other problems. First, we try to

infer the causal effect of vitamin A deficiency level, which is latent, and the symptom of xerophthalmia is only a surrogate for vitamin A deficiency, indicating whether vitamin A deficiency level is above some threshold (see the discussion section of Zeger and Liang (1991)). Secondly, we only observe this surrogate and other time dependent covariates at discrete time points, even though everything changes in continuous time.

These two features distinguish our model from the established semi-parametric methods for longitudinal data developed by Robins and his collaborators (Robins, 1986, 1992, 1994, 1998; Robins et al., 1999). In most of Robins' work, data are assumed to be generated from a discrete time process, and at each time point, all the confounders are observed such that the treatment looks as if it is randomly assigned conditional on all the covariates and historical treatment levels (the *ignorability assumption*, see assumption (A1) in Section 1.1). In our data set, the presence/absence of respiratory disease, the presence/absence of stunting and the level of vitamin A deficiency can all switch at any time between two consecutive time points, and it is more reasonable to assume that the whole process is in continuous time and only observed at discrete times. In Chapter 2, we have shown that even if the ignorability assumption is true in continuous time, it may not hold at discrete observational times and thus the standard g-estimation of Robins may be biased.

The model in this chapter also differs from the ones discussed in Chapter 2. In Chapter 2, the full continuous time history of the exact amount of treatment is

assumed to be known, in which case a modified ignorability assumption and the controlling-the-future method of Joffe and Robins (Joffe and Robins, 2009) can be used to identify the causal effect semi-parametrically in some scenarios. However, the luxury of observing the exact amount of treatment is not available in our data set, as the treatment level, the real vitamin A deficiency level, is unobserved. Only when the deficiency reached a level high enough can xerophthalmia be observed and recorded in the data, and xerophthalmia itself is not recorded in continuous time - it is observable only at the time of the visits. In a word, we are suffering from the measurement error problem discussed in Section 1.2. While standard semi-parametric approach fails to adjust for the measurement error, more assumptions are needed for identification. In this chapter, we choose to use a full parametric model.

This chapter also contributes to the literature of estimating a discretely observed continuous time Markov process. Vast amount of research work has been done on various cases of the problem, including discretely observed continuous time diffusions, e.g., (Johannes et al., 2009; Elerian et al., 2001; Blackwell, 2003; Aït-Sahalia, 2002), discretely observed stationary continuous time Markov chain, e.g., (Bladt and Sørensen, 2005), and discretely observed two-state non-stationary continuous time Markov chain, e.g., (Singer, 1981). In our model, we are facing a discretely observed 32-state non-stationary continuous time Markov chain. We are able to find an MCMC algorithm to estimate the parameters for our particular problem. It will be of great interest to search for more general and more efficient algorithms for esti-

mating discretely observed non-stationary continuous time Markov process in future research.

The organization of the chapter is as follows: Section 3.2 describes our continuous time Markov model under the counterfactual framework; Section 3.3 describes our MCMC algorithm for estimating the model; Section 3.4 is a simulation study of our computational algorithm; Section 3.5 reports the results for applying our model to the vitamin A deficiency data; Section 3.6 concludes the chapter.

3.2 A Markov Model

In this chapter, we denote the level of vitamin A deficiency at time t by an ordinal latent variable A_t^* with d states (e.g., $d = 4$ and the states are 0, 1, 2, 3). We define A_t to be a binary indicator of xerophthalmia, which will take value 1 if $A_t^* \geq c$ and 0 if $A_t^* < c$ (e.g., $c = 2$ if $d = 4$). Let L_t be the binary indicator of whether the child is stunted at time t , and Y_t be the binary variable that indicates whether the child is suffering from respiratory infection at time t . Consider the counterfactual outcome $Y_t^{\bar{A}_s^*, 0}$, $t \geq s$, which is the potential outcome for respiratory infection status at time t , if the subject receives the realized treatment A_l^* from time 0 to just prior to time s and keeps the treatment level at the lowest (i.e., 0) from time t and on.

3.2.1 The Causal Model

We assume the following model for the causal relationship between A_t^* and Y_t .

$$P(Y_t = 1 | Y_t^{\bar{A}_t^*, 0}, A_t^*) = \begin{cases} 1 & \text{if } Y_t^{\bar{A}_t^*, 0} = 1 \\ 0 & \text{if } Y_t^{\bar{A}_t^*, 0} = 0, A_t^* = 0 \\ \delta_j & \text{if } Y_t^{\bar{A}_t^*, 0} = 0, A_t^* = j \quad (j > 0) \end{cases} \quad (3.2.1)$$

The model describes how the treatment levels affect the realized outcome Y_t , given the baseline potential outcome at time t . The δ 's are the causal effect of A_t on Y_t .

In (3.2.1), we assume that if $Y_t^{\bar{A}_t^*, 0} = 1$, $Y_t = 1$, i.e., if the child would suffer from respiratory infection even without any vitamin A deficiency, any vitamin A deficiency could only make the child worse. If $Y_t^{\bar{A}_t^*, 0} = 0$, different levels of vitamin A deficiency results in different levels of risk for the child to get respiratory infection, and we would expect that the higher the level of vitamin A deficiency is, the higher risk of respiratory infection the child has. It is worth noting that in this model, $Y_t = 0$ if $Y_t^{\bar{A}_t^*, 0} = 0$ and $A_t^* = 0$, which is a consistency assumption that is often assumed in most causal research work (e.g., Robins et al. (2000)). For example, with a single outcome $Y^{observe}$, people usually assume that $Y^{observe} = Y^a$ if $A = a$, where Y^a is the counterfactual outcome under treatment level a .

3.2.2 Continuous Time Markov Process

We assume that $(A_t^*, Y_t^{\bar{A}_t^*, 0}, L_t, Y_t)$ follows a continuous time Markov process, which will be described below. Note that A_t is only a coarsened observation of A_t^* . We

also assume that the observational times are $0, 1, \dots, K$. At each time point k , only (A_k, Y_k, L_k) are observed. Any variables in between two consecutive time points and any potential outcomes are not observable. Figure 3.1 shows a discretized example of the data generating process from this model. In the example, $(A_t^*, Y_t^{\bar{A}_t^*, 0}, L_t, Y_t)$ can change values at time 1, 1.5, 2, 2.5, 3 and 3.5, but only (A_k, Y_k, L_k) can be observed at time 1, 2, and 3.

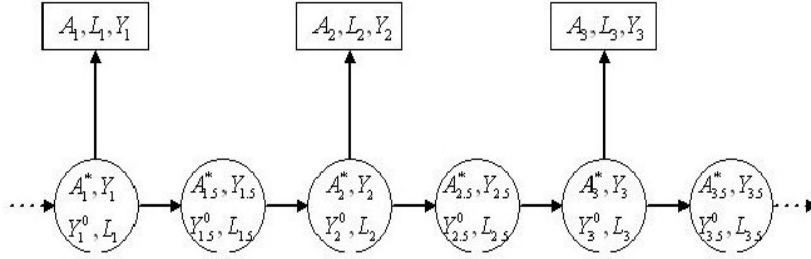


Figure 3.1: Discretized Example of Data Generating Process

The Transition Rate Matrix Q

The continuous time Markov process $(A_t^*, Y_t^{\bar{A}_t^*, 0}, L_t, Y_t)$ has 32 states (4 by 2 by 2 by 2), each state being denoted by a vector $s = (a^*, y^0, l, y)^T$, where $a^* \in \{0, 1, 2, 3\}$, $y^0 \in \{0, 1\}$, $l \in \{0, 1\}$ and $y \in \{0, 1\}$. We model the Markov process by specifying how we construct the transition rate matrix $Q(t)$. $Q(t)$ is a 32 by 32 matrix, which may depend on covariates like time, age and gender. Note that unlike stunting, time, age and gender are not time-dependent confounders. They are not modeled as the state of the Markov process, but are conditioned on when calculating the transition

rate matrix.

We denote $X_t = (A_t^*, Y_t^{\bar{A}_t^*, 0}, L_t, Y_t)$ and consider the elements in $Q(t)$:

$$q_t(s_1, s_2) = \lim_{h \downarrow 0} \frac{Pr(X_{t+h} = s_2 | X_t = s_1, age, sex)}{h}$$

for $s_1 \neq s_2$, and

$$q_t(s, s) = - \sum_{s' \neq s} q_t(s, s')$$

We factorize

$$\begin{aligned} & P(X_{t+h} | X_t, age, sex) \\ &= P(A_{t+h}^*, Y_{t+h}^{\bar{A}_{t+h}^*, 0} | X_t, age, sex) P(L_{t+h} | X_t, age, sex, A_{t+h}^*, Y_{t+h}^{\bar{A}_{t+h}^*, 0}) \\ & \quad \times P(Y_{t+h} | X_t, age, sex, A_{t+h}^*, Y_{t+h}^{\bar{A}_{t+h}^*, 0}, L_{t+h}) \\ &= P(A_{t+h}^* | X_t, age, sex) P(Y_{t+h}^{\bar{A}_{t+h}^*, 0} | X_t, A_{t+h}^*, age, sex) \\ & \quad \times P(L_{t+h} | X_t, age, sex, A_{t+h}^*, Y_{t+h}^{\bar{A}_{t+h}^*, 0}) \times P(Y_{t+h} | A_{t+h}^*, Y_{t+h}^{\bar{A}_{t+h}^*, 0}). \end{aligned}$$

Note that with this factorization, we have assumed that Y_{t+h} is independent of X_t, sex, age and L_{t+h} conditional on A_{t+h}^* and $Y_{t+h}^{\bar{A}_{t+h}^*, 0}$. This assumption means that causal effect of A_{t+h} on Y_{t+h} does not depend on any pre-treatment covariate. Similar assumptions have been assumed in many causal researches (e.g. the model of additive effect discussed in the Section 2 of Rosenbaum (2002) is such a model).

We model each component in the factorization as

- Model for the jump of A_t^*

$$P(A_{t+h}^* = j | X_t, age, sex) \tag{3.2.2}$$

$$= \begin{cases} h\alpha(A_{t+h}^*; A_t^*, L_t, Y_t, \text{age}, \text{sex}, t) + o(h), & |A_t^* - j| = 1 \\ o(h), & |A_t^* - j| > 1 \end{cases}$$

The model assumes that the level of vitamin A deficiency can only switch to an adjacent level when it switches.

- Model for the jump of $Y_t^{\bar{A}_t^*, 0}$

$$\begin{aligned} P(Y_{t+h}^{\bar{A}_{(t+h)}^*, 0} = j | X_t, A_{t+h}^*, \text{age}, \text{sex}) & \quad (3.2.3) \\ = h\beta(Y_{t+h}^{\bar{A}_{(t+h)}^*, 0}; X_t, \text{age}, \text{sex}, t) + o(h), & \quad Y_t^{\bar{A}_t^*, 0} \neq j; \end{aligned}$$

- Model for the jump of L_t

$$\begin{aligned} P(L_{t+h} = j | X_t, \text{age}, \text{sex}, A_{t+h}^*, Y_{t+h}^{\bar{A}_{(t+h)}^*, 0}) & \quad (3.2.4) \\ = h\gamma(L_{t+h}; X_t, \text{age}, \text{sex}, A_{t+h}^*, Y_{t+h}^{\bar{A}_{(t+h)}^*, 0}, t) + o(h), & \quad L_t \neq j. \end{aligned}$$

- $P(Y_{t+h} | A_{t+h}^*, Y_{t+h}^{\bar{A}_{(t+h)}^*, 0})$ is defined in (3.2.1). We denote

$$p_y(y; a, y^0) = P(Y_{t+h} = y | A_{t+h}^* = a, Y_{t+h}^{\bar{A}_{(t+h)}^*, 0} = y^0).$$

Remark 3.2.1. By assuming that the α function is not a function of $Y_t^{\bar{A}_t^*, 0}$ and is a function of Y_t and L_t , and that the β function is not a function of A_{t+h}^* , we have assumed continuous time ignorability for our model, i.e., the treatment only depends on realized past covariates and past treatments and does not depend on future potential outcomes. For a formal definition of continuous time ignorability assumption and the proof that the model conforms with the ignorability assumption, see Section 5.6.

Given (3.2.2), (3.2.3) and (3.2.4), the elements of the transition rate matrix $Q(t)$ are, for $s_1 \neq s_2$

$$q_t(s_1, s_2) = \begin{cases} \alpha_t(a_2; a_1, l_1, y_1) p_y(y_2; a_2, y_2^0) & \text{if } |a_1 - a_2| = 1, l_1 = l_2, y_1^0 = y_2^0 \\ \beta_t(y_2^0; s_1) p_y(y_2; a_2, y_2^0) & \text{if } y_1^0 \neq y_2^0, a_1 = a_2, l_1 = l_2 \\ \gamma_t(l_2; s_1, a_2, y_2^0) p_y(y_2; a_2, y_2^0) & \text{if } l_1 \neq l_2, a_1 = a_2, y_1^0 = y_2^0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\alpha_t(a_2; a_1, l_1, y_1) = \alpha(A_{t+h}^* = a_2; A_t^* = a_1, L_t = l_1, Y_t = y_1, \text{age}, \text{sex}, t)$$

$$\beta_t(y_2^0; s_1) = \beta(Y_{t+h}^{\bar{A}^*(t+h)^-, 0} = y_2^0; X_t = s_1, \text{age}, \text{sex}, t)$$

$$\gamma_t(l_2; s_1, a_2, y_2^0) = \gamma(L_{t+h} = l_2; X_t = s_1, \text{age}, \text{sex}, A_{t+h}^* = a_2,$$

$$Y_{t+h}^{\bar{A}^*(t+h)^-, 0} = y_2^0, t).$$

Initial Conditions and the Conditional Likelihood Function

For a complete model, we also need to give the initial probability distribution for $(A_0^*, Y_0^{\bar{A}_0^-, 0}, L_0, Y_0)$. The causal relationship and the feedback cycles are already encoded in (3.2.1) and the definition of $Q(t)$. It would be reasonable to make inference from a conditional likelihood that conditions on the initial states. However, $A_0^*, Y_0^{\bar{A}_0^-, 0}$ are unobserved, and only (A_0, L_0, Y_0) are observed. We hereby assume that

$$P(A_0^*, Y_0^{\bar{A}_0^-, 0} | A_0, L_0, \text{age}, \text{sex}) = \frac{1}{2} \times \frac{1}{2} \times I_{A_0=(A_0^* \geq c)} \quad (3.2.5)$$

This initial distribution claims that A_0^* is uniform on $\{0, 1\}$ if $A_0 = 0$, and uniform on $\{2, 3\}$ if $A_0 = 1$, and that $Y_0^{\bar{A}_0^-, 0}$ is uniform on $\{0, 1\}$, conditional on the value of A_0

and L_0 . This assumption gives the levels of A^* certain physical meaning. If we view the *true* level of vitamin A deficiency as a continuous measure and xerophthalmia as an indicator whether the *true* level is above a threshold, (3.2.5) assumes that the cut-point between $A^* = 1$ and $A^* = 0$ is the median of the conditional distribution of the *true* level, conditioning on $A = 0$, and that the cut-point between $A^* = 2$ and $A^* = 3$ is the median of the conditional distribution of the *true* level, conditioning on $A = 1$.

With equation (3.2.5) and model (3.2.1), $P(A_0^*, Y_0^{\bar{A}_0^*, 0} | A_0, L_0, Y_0, age, sex)$ can be calculated. Given $P(A_0^*, Y_0^{\bar{A}_0^*, 0} | A_0, L_0, Y_0, age, sex)$, a conditional likelihood function of the observed data can be calculated as

$$\begin{aligned} & f(A_1, \dots, A_K, L_1, \dots, L_K, Y_1, \dots, Y_K | A_0, L_0, Y_0, age, sex) \\ &= \int \int P(A_0^*, Y_0^{\bar{A}_0^*, 0} | A_0, L_0, Y_0, age, sex) \\ & \quad \times P(A_1, \dots, A_K, L_1, \dots, L_K, Y_1, \dots, Y_K | A_0^*, Y_0^{\bar{A}_0^*, 0}, L_0, Y_0, age, sex) dA_0^* dY_0^{\bar{A}_0^*, 0} \end{aligned}$$

$P(A_1, \dots, A_K, L_1, \dots, L_K, Y_1, \dots, Y_K | A_0^*, Y_0^{\bar{A}_0^*, 0}, L_0, Y_0, age, sex)$ can be determined by $Q(t)$ and model (3.2.1), even though a direct computation is almost infeasible. We will discuss our approach in Section 3.3.

To summarize, our model assumes that $(A_t^*, Y_t^{\bar{A}_t^*, 0}, L_t, Y_t)$ follows a continuous time non-stationary Markov process defined by the transition rate matrix $Q(t)$, and that A_t is determined deterministically by A_t^* . We only observe the (A_t, Y_t, L_t) at time $0, 1, \dots, K$. Different subjects are assumed to be independent realizations of the process conditional on their baseline covariates and initial conditions.

3.3 Estimation: MCMC with Data Augmentation

To estimate the model defined in Section 3.2, we parametrize $\delta_i = \frac{\exp(\tilde{\delta}_i)}{1 + \exp(\tilde{\delta}_i)}$, $\alpha = \alpha(\tilde{\alpha})$, $\beta = \beta(\tilde{\beta})$ and $\gamma = \gamma(\tilde{\gamma})$. Denote $\Theta = (\tilde{\delta}_1, \tilde{\delta}_2, \tilde{\delta}_3, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$. Θ is the collection of parameters that needs to be estimated.

Even if we know the true value of Θ , it is quite difficult to evaluate the likelihood of the observed data in our model. The difficulty lies in the fact that the Markov process is non-stationary and that it depends on baseline covariates. For a stationary continuous time Markov process with finite number of states and without covariates, transition matrix from time t to time $t + 1$ is simply e^Q , where Q is the transition rate matrix. With e^Q calculated once, the problem becomes a standard discrete time hidden Markov model, where the transition matrix between states is e^Q and the emission matrix easily defined since A_t, Y_t, L_t are deterministic observations from the states of the Markov model. With a non-stationary Markov process, the transition matrix does not usually have a simple form (for two-state process, there is one, see Singer (1981); for a more general process, see discussions in Wei and Norman (1963, 1964)). One practical approach may be discretizing the process and approximate the transition matrix from time t to time $t + 1$ by $\prod_{i=1}^n [\frac{1}{n}Q(t + \frac{i-1}{n}) + I]$, which will be quite time consuming, and considering that fact that we have covariates and we need to do this computation for every subject in every time interval, it quickly becomes computationally infeasible.

In this section, we propose a Monte Carlo Markov Chain approach with data

augmentation for estimation (see more about data augmentation in van Dyk and Meng (2001)). In our algorithm, we also need to discretize the continuous time Markov process and approximate the process by an embedded Markov chain.

3.3.1 Discretization Scheme

Denote $Z_t = (A_t^*, Y_t^{\bar{A}_t^*, 0}, L_t, A_t, Y_t)$ (notice that Z_t is also a Markov process under the model defined in Section 3.2). We discretize the continuous time Markov process at the time grid of $\frac{1}{n}$, i.e., we consider the embedded discrete time Markov chain $Z_0, Z_{\frac{1}{n}}, Z_{\frac{2}{n}}, \dots, Z_{\frac{n-1}{n}}, Z_1, Z_{\frac{n+1}{n}}, \dots, Z_K$. We approximate the transition probability of the discretized chain by (to simplify notation, we define $Y_t^0 = Y_t^{\bar{A}_t^*, 0}$)

$$\begin{aligned}
& Pr(Z_{\frac{m+1}{n}} | Z_{\frac{m}{n}}) \\
& \approx \left[\frac{1}{n} \alpha(A_{\frac{m+1}{n}}^*; A_{\frac{m}{n}}^*, L_{\frac{m}{n}}, age, sex, Y_{\frac{m}{n}}, \frac{m}{n}) \right]^{I_{A_{\frac{m+1}{n}}^*} \neq A_{\frac{m}{n}}^*} \\
& \quad \times \left[1 - \frac{1}{n} (1 + I_{A_{\frac{m}{n}}^* \notin \{0, d-1\}}) \alpha(A_{\frac{m+1}{n}}^*; A_{\frac{m}{n}}^*, L_{\frac{m}{n}}, age, sex, Y_{\frac{m}{n}}, \frac{m}{n}) \right]^{I_{A_{\frac{m+1}{n}}^*} = A_{\frac{m}{n}}^*} \\
& \quad \times \left[\frac{1}{n} \beta(Y_{\frac{m+1}{n}}^0; X_{\frac{m}{n}}, age, sex, Y_{\frac{m}{n}}, \frac{m}{n}) \right]^{I_{Y_{\frac{m+1}{n}}^0} \neq Y_{\frac{m}{n}}^0} \\
& \quad \times \left[1 - \frac{1}{n} \beta(Y_{\frac{m+1}{n}}^0; X_{\frac{m}{n}}, age, sex, Y_{\frac{m}{n}}, \frac{m}{n}) \right]^{I_{Y_{\frac{m+1}{n}}^0} = Y_{\frac{m}{n}}^0} \\
& \quad \times \left[\frac{1}{n} \gamma(L_{\frac{m+1}{n}}; X_{\frac{m}{n}}, age, sex, Y_{\frac{m}{n}}, A_{\frac{m+1}{n}}^*, Y_{\frac{m+1}{n}}^0, \frac{m}{n}) \right]^{I_{L_{\frac{m+1}{n}}} \neq L_{\frac{m}{n}}} \\
& \quad \times \left[1 - \frac{1}{n} \gamma(L_{\frac{m+1}{n}}; X_{\frac{m}{n}}, age, sex, Y_{\frac{m}{n}}, A_{\frac{m+1}{n}}^*, Y_{\frac{m+1}{n}}^0, \frac{m}{n}) \right]^{I_{L_{\frac{m+1}{n}}} = L_{\frac{m}{n}}} \\
& \quad \times Pr(A_{\frac{m+1}{n}} | A_{\frac{m+1}{n}}^*) \times Pr(Y_{\frac{m+1}{n}} | Y_{\frac{m+1}{n}}^0, A_{\frac{m+1}{n}}^*)
\end{aligned}$$

where $Pr(A_{\frac{m+1}{n}} | A_{\frac{m+1}{n}}^*) = I_{A_{\frac{m+1}{n}} = A_{\frac{m+1}{n}}^*} \geq c$, and $Pr(Y_{\frac{m+1}{n}} | Y_{\frac{m+1}{n}}^0, A_{\frac{m+1}{n}}^*)$ is decided by equation (3.2.1).

3.3.2 The MCMC Algorithm

Notice that after discretizing the process, if we do observe full Z_t process, the evaluation of the full likelihood is relatively easy, as we would avoid matrix multiplications. Motivated by this fact, we consider an MCMC algorithm with data augmentation. Denote the observed data by O , where $O = (A_0, Y_0, L_0, A_1, Y_1, L_1, \dots, A_K, Y_K, L_K)$, the full data by Z , where $Z = (Z_0, Z_{\frac{1}{n}}, Z_{\frac{2}{n}}, \dots, Z_K)$, and the missing data by U , where $U = Z - O$.

To fit into the Bayesian framework, we assume a proper prior distribution for Θ . In our implementation, the prior distribution π is assumed to be a multi-normal distribution, with mean $(0, 0, \dots, 0)$ and covariance matrix σI , where I is the identity matrix. We view the proper prior distribution as a regularization in estimating the parameters. It has been recognized for a long time that maximum likelihood estimates for discretely observed continuous time Markov process could be very unstable (see Kalbfleisch and Lawless, 1985). For example, in the case of stationary Markov process, if the transition rate matrix has large elements and the time interval between two observations is too long, the transition probability between two observational time points could be close to the stationary distribution. Then the transition rate matrix multiplied by any large positive number would explain the data very well. MLE could have huge variance or not even exist. See Bladt and Sørensen (2005) for a concrete example when MLE does not exist. In their example, the Markov model is saturated, i.e., there is no constraint on the transition rate matrix. Although we

have assumed structures for our transition rate matrix, we suspect that our likelihood function is not well behaved either, as our experiments on MCMC with a flat prior distribution fails to converge even after a huge number of iterations.

With the setup, a sketch of a general MCMC algorithm with data augmentation is in Figure 3.2.

Figure 3.2: MCMC with Data Augmentation

1. Initialize Θ and U .
 2. Given U , we update Θ by simulating from $P(\Theta|U, O)$.
 3. Given Θ , we update U by simulating from $P(U|\Theta, O)$.
 4. Repeat 2-3.
-

Here $P(\Theta|U, O) \propto \pi(\Theta)P(U, O|\Theta)$ and $P(U|\Theta, O) \propto P(U, O|\Theta)$. $P(U, O|\Theta)$ can be easily calculated using the formula in Section 3.3.1. Under certain regularity conditions, the limiting marginal distribution of Θ will be the posterior distribution $P(\Theta|O)$.

With our model, it is quite difficult to simulate directly from $P(\Theta|U, O)$ and $P(U|\Theta, O)$. We therefore substitute step 2 and 3 by Metropolis-Hasting steps.

- *Step 2* Given U and Θ_{old} , we simulate Θ_{new} from a proposal distribution $q(\Theta|\Theta_{old})$, and calculate

$$r_1 = \min\left(1, \frac{P(\Theta_{new}|U, O)q(\Theta_{old}|\Theta_{new})}{P(\Theta_{old}|U, O)q(\Theta_{new}|\Theta_{old})}\right).$$

We update Θ by Θ_{new} with probability r_1 , and by Θ_{old} with probability $1 - r_1$.

- *Step 3* Given Θ , and U_{old} , we simulate U_{new} from a proposal distribution $q(U|U_{old})$, and calculate

$$r_2 = \min\left(1, \frac{P(U_{new}|\Theta, O)q(U_{old}|U_{new})}{P(U_{old}|\Theta, O)q(U_{new}|U_{old})}\right).$$

We update U by U_{new} with probability r_2 , and by U_{old} with probability $1 - r_2$.

In our implementation, the proposal distribution $q(\Theta|\Theta_{old})$ is taken to be the multi-normal distribution with mean Θ_{old} and covariance matrix being a diagonal matrix.

In this case, the calculation of r_1 simplifies to

$$r_1 = \min\left(1, \frac{P(\Theta_{new}|U, O)}{P(\Theta_{old}|U, O)}\right).$$

The proposal distribution $q(U|U_{old})$ is more complicated, which we describe in the following section.

3.3.3 Proposal Distributions for the Augmented Data

To find a good proposal distribution for U , we first need to be able to simulate a whole path of the discretized Markov chain, such that the values of the path match the observed values at the observational times, namely $0, 1, \dots, K$. Secondly, the path we simulate has to have a significant probability to occur under the Markov model. There are available methods for simulating paths that match the end points in continuous time stationary Markov process (see Blackwell, 2003; Bladt and Sørensen, 2005; Nielsen, 2002; Hobolth, 2008). In this chapter, we have designed a proposal

distribution that works well for our model, which is a continuous time non-stationary Markov process.

Rather than simulating the paths of states for the whole chain, our proposal distribution simulates the A_t^* , L_t , Y_t^0 , Y_t and A_t separately. Also, we do not simulate the whole path altogether, but rather we update the path piece by piece, i.e., for one subject, we update $(A_i^*, Y_i^0, Z_{i+\frac{1}{n}}, \dots, Z_{i+\frac{n-1}{n}})$ once at a time, with the sequence of $i = 0, 1, 2, \dots, K - 1$, and then update (A_K^*, Y_K^0) .

Specifically, for any $i = 0, 1, 2, \dots, K - 1$, we simulate the missing data as follows:

- If $A_i = 1$, simulate $A_i^* = 2 + \text{Bernoulli}(0.5)$; if $A_i = 0$, simulate $A_i^* = \text{Bernoulli}(0.5)$. Let $q_1 = 0.5$.
- If $Y_i = 0$, simulate $Y_i^0 = 0$, and let $q_2 = 1$; if $Y_i = 1$, simulate $Y_i^0 = \text{Bernoulli}(0.5)$. Let $q_2 = 0.5$.
- To simulate the path for L , we consider $x_j = L_{i+\frac{j}{n}} - L_{i+\frac{j-1}{n}}$, i.e., $\{x_j\}_{j=1}^n$ is the first order difference of $\{L_{i+\frac{j}{n}}\}_{j=0}^n$. Each x_j must be $+1$, 0 or -1 , $L_i + \sum_{j=1}^m x_j$ must be either 1 or 0 , and $\sum_{j=1}^n x_j = L_{i+1} - L_i$. We would like to simulate the x_j 's satisfying these constraints with certain control over the number of non-zeros in them (or equivalently, the number of jumps in L), as we are expecting the number of jumps to be small for our particular problem.

If $L_i = L_{i+1}$, the number of jumps for the L process must be $0, 2, 4, \dots, 2 \times \lfloor \frac{n}{2} \rfloor$.

Define $q_L^*(x)$ to be a probability distribution on $0, 2, 4, \dots, 2 \times \lfloor \frac{n}{2} \rfloor$, where $\lfloor x \rfloor$ is the largest integer smaller than x . Simulate M from q_L^* . Randomly pick M

positions from n positions. Those M positions will be the non-zero x_j 's. With M and those M positions, a x sequence and the L sequence are determined. We denote the proposal probability by

$$q_3 = q_L^*(M) \frac{1}{\binom{n}{M}}.$$

If $L_i \neq L_{i+1}$, the number of jumps for the L process must be 1, 3, 5, \dots , $2 \times \lfloor \frac{n}{2} \rfloor - 1$. A similar procedure can be adopted.

- Simulating Y_t^0 uses exactly the same procedure as for simulating L_t . Let q_4 be the probability from the proposal distribution.
- To simulate the path for A^* , let $x_j = A_{i+\frac{j}{n}}^* - A_{i+\frac{j-1}{n}}^*$. Each x_j must be +1, 0 or -1 (A^* are only allowed to switch to adjacent levels), $A_i^* + \sum_{j=1}^m x_j$ must be between 0 and 3 inclusively, and $\sum_{j=1}^n x_j = A_{i+1}^* - A_i^*$. We first simulate the number of switches in the A^* process. If $|A_i^* - A_{i+1}^*| = k$, the number of non-zeros in x_j must be one of $(k, k + 2, \dots, 2 \times \lfloor \frac{n-k}{2} \rfloor + k)$. Define q_A^* to be a probability distribution on them. We simulate M from q_A^* , and randomly sample M positions from n for the non-zero x_j 's. However, since A^* is not binary, there could be many paths that has M jumps and jump at the sampled positions. We randomly sample one from the qualified paths. We denote the

number of qualified paths by n_q , and let

$$q_5 = q_A^*(M) \frac{1}{\binom{n}{M}} \frac{1}{n_q}.$$

Conditional on M and the M positions, randomly sampling one path from all qualified paths is not so straightforward. We describe our solution to this problem Section 5.7.

- The simulation of Y follows the causal model (3.2.1). Let

$$q_6 = \prod_{j=1}^{n-1} p_y(Y_{i+\frac{j}{n}}; Y_{i+\frac{j}{n}}^0, A_{i+\frac{j}{n}}^*).$$

- The simulation of A is deterministic: $A_{i+\frac{j}{n}} = I_{A_{i+\frac{j}{n}} > c}$. Let $q_7 = 1$.

Let $q_{new} = \prod_{j=1}^7 q_j$. We have simulated a set of $U_i = (A_i^*, Y_i^0, Z_{i+\frac{1}{n}}, \dots, Z_{i+\frac{n-1}{n}})$ from our proposal distribution and calculated the proposal probability as q_{new} . With the proposal distribution, it is also easy to calculate the probability of the old paths under the proposal distribution. We denote it as q_{old} .

We then calculate, for $i > 0$,

$$r_{2,i} = \min\left(1, \frac{P(A_i, Y_i, U_i^{new}, Z_{i+1} | \Theta, Z_{i-\frac{1}{n}}) q_{old}}{P(A_i, Y_i, U_i^{old}, Z_{i+1} | \Theta, Z_{i-\frac{1}{n}}) q_{new}}\right)$$

for $i = 0$

$$r_{2,i} = \min\left(1, \frac{P(Y_{0,new}^0, A_{0,new}^* | A_0, L_0, Y_0) P(U_i^{new}, Z_{i+1} | \Theta, Z_{i,new}) q_{old}}{P(Y_{0,old}^0, A_{0,old}^* | A_0, L_0, Y_0) P(U_i^{old}, Z_{i+1} | \Theta, Z_{i,old}) q_{new}}\right)$$

and keep U_i^{new} with probability $r_{2,i}$.

To update (A_K^*, Y_K^0) , we simulate them as follows,

- If $A_K = 1$, simulate $A_K^* = 2 + \text{Bernoulli}(0.5)$; if $A_K = 0$, simulate $A_K^* = \text{Bernoulli}(0.5)$. Let $q_1 = 0.5$.
- If $Y_K = 0$, simulate $Y_K^0 = 0$, and let $q_2 = 1$; if $Y_K = 1$, simulate $Y_K^0 = \text{Bernoulli}(0.5)$. Let $q_2 = 0.5$.

$q_{new} = q_1 q_2$, and q_{old} is calculated accordingly. Define

$$r_{2,K} = \min\left(1, \frac{P(Z_{K,new}|Z_{K-\frac{1}{n}}, \Theta)q_{old}}{P(Z_{K,old}|Z_{K-\frac{1}{n}}, \Theta)q_{new}}\right)$$

we keep $(A_{K,new}^*, Y_{K,new}^0)$ with probability $r_{2,K}$.

In this proposal distribution, we separate the simulation of the number of jumps and the positions of the jumps, which gives us a better control over the proposal distribution. We can incorporate our intuition of how frequent of the process jumps into the proposal distribution and efficiently walk through the space of all possible U . There are also shortcomings of this approach. As each component is simulated separately, it is possible that several components jump together, which is a rare event in the original model and makes our algorithm less efficient. For the particular study in this chapter, our proposal distribution seems to be working fine.

As a summary, our full MCMC algorithm with data augmentation is in Figure 3.3.

Figure 3.3: Full MCMC Algorithm with Data Augmentation

1. Initialize Θ and U .
2. Given U and Θ_{old} , we simulate Θ_{new} from a proposal distribution $q(\Theta|\Theta_{old})$, and calculate

$$r_1 = \min\left(1, \frac{P(\Theta_{new}|U, O)q(\Theta_{old}|\Theta_{new})}{P(\Theta_{old}|U, O)q(\Theta_{new}|\Theta_{old})}\right).$$

We update Θ by Θ_{new} with probability r_1 , and by Θ_{old} with probability $1 - r_1$.

3. Given Θ , for every subject, we update U by the algorithm described in Section 3.3.3.
 4. Repeat 2-3.
-

3.4 Simulation

In this section, we present one particular parametrization for the α , β and γ functions. The same parametrization will be used in our application to the vitamin A deficiency data in the next section. We simulate a similar data set as the vitamin A deficiency data from our Markov model, i.e., by setting the parameter values to be the estimates from the real data, and the number of subjects to be the number of subjects in the real data. We then estimate the parameters from the simulated data to see if our MCMC algorithm can correctly reconstruct the true values.

The models for α , β and γ are:

$$\alpha(A_{t+h}^*; A_t^*, L_t, Y_t, age, sex, t; \tilde{\alpha})$$

$$\begin{aligned}
&= \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 A_t^* + \tilde{\alpha}_2 L_t + \tilde{\alpha}_3 Y_t + \tilde{\alpha}_4(\text{age}) + \tilde{\alpha}_5(\text{sex}) + \tilde{\alpha}_6 \cos(t) + \tilde{\alpha}_7 \sin(t)) \\
&\quad / \{2[1 + \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 A_t^* + \tilde{\alpha}_2 L_t + \tilde{\alpha}_3 Y_t + \tilde{\alpha}_4(\text{age}) + \tilde{\alpha}_5(\text{sex}) + \tilde{\alpha}_6 \cos(t) \\
&\quad \quad + \tilde{\alpha}_7 \sin(t))]\} \\
&\quad \beta(Y_{t+h}^{\bar{A}_{(t+h)}^*}, 0; X_t, \text{age}, \text{sex}, t; \tilde{\beta}) \\
&= \exp(\tilde{\beta}_0 + \tilde{\beta}_1 A_t^* + \tilde{\beta}_2 L_t + \tilde{\beta}_3 Y_t + \tilde{\beta}_4 Y_t^{\bar{A}_{t-}^*, 0} + \tilde{\beta}_5(\text{age}) + \tilde{\beta}_6(\text{sex}) \\
&\quad \quad + \tilde{\beta}_7 \cos(t) + \tilde{\beta}_8 \sin(t)) \\
&\quad / \{1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 A_t^* + \tilde{\beta}_2 L_t + \tilde{\beta}_3 Y_t + \tilde{\beta}_4 Y_t^{\bar{A}_{t-}^*, 0} + \tilde{\beta}_5(\text{age}) + \tilde{\beta}_6(\text{sex}) \\
&\quad \quad + \tilde{\beta}_7 \cos(t) + \tilde{\beta}_8 \sin(t))\} \\
&\quad \gamma(L_{t+h}; X_t, \text{age}, \text{sex}, A_{t+h}^*, Y_{t+h}^{\bar{A}_{(t+h)}^*}, 0, t; \tilde{\gamma}) \\
&= \exp(\tilde{\gamma}_0 + \tilde{\gamma}_1 A_t^* + \tilde{\gamma}_2 L_t + \tilde{\gamma}_3 Y_t + \tilde{\gamma}_4 Y_t^{\bar{A}_{t-}^*, 0} + \tilde{\gamma}_5 A_{t+h}^* + \tilde{\gamma}_6 Y_{t+h}^{\bar{A}_{(t+h)}^*}, 0 \\
&\quad \quad + \tilde{\gamma}_7(\text{age}) + \tilde{\gamma}_8(\text{sex}) + \tilde{\gamma}_9 \cos(t) + \tilde{\gamma}_{10} \sin(t)) \\
&\quad / \{2[1 + \exp(\tilde{\gamma}_0 + \tilde{\gamma}_1 A_t^* + \tilde{\gamma}_2 L_t + \tilde{\gamma}_3 Y_t + \tilde{\gamma}_4 Y_t^{\bar{A}_{t-}^*, 0} + \tilde{\gamma}_5 A_{t+h}^* + \tilde{\gamma}_6 Y_{t+h}^{\bar{A}_{(t+h)}^*}, 0 \\
&\quad \quad + \tilde{\gamma}_7(\text{age}) + \tilde{\gamma}_8(\text{sex}) + \tilde{\gamma}_9 \cos(t) + \tilde{\gamma}_{10} \sin(t))]\}
\end{aligned}$$

In this formulation, We assume that all of A_t^* , Y_t^0 and L_t could depend on age, sex, and seasonal factor $\cos(t)$ and $\sin(t)$. Also note that the values of function α and γ are always between 0 and 0.5, and that the value of function β is always between 0 and 1. The general model we proposed in Section 3.2 only requires these functions to be positive. The validity of the more stringent models here relies on our belief that vitamin A deficiency, counterfactual respiratory infection and stunting switch states

infrequently. The models also prevents problems in discretization, as it guarantees that, for example, $\frac{1}{n}\alpha$ is always a probability.

Using the set of parameters we estimated from the vitamin A deficiency data, we simulate from the above model for 250 subject from time 1 to time 6, with baseline covariates and baseline variables generated randomly, and record A_t, Y_t and L_t at the integer times. We then estimate the model using our MCMC algorithm with a prior distribution $N(0, I)$, i.e., we assume the prior distribution for every parameter is independent standard normal. We repeat 10 independent simulations, and estimate parameters by posterior mean from the 10 simulated data sets.

The estimation results from our simulated data sets are reported in Table 3.1. The second and the third columns of Table 3.1 show the true parameter values and the mean estimates. While the prior distribution obviously causes bias on some parameter estimates, the fourth column shows that most of the biases are non-significant. Considering the fact that we used a proper prior distribution as regularization and the size of our data set, we believe that our MCMC algorithm is doing a decent job in estimating the parameters.

3.5 Application

This section reports the result of our vitamin A deficiency analysis.

Table 3.1: Simulation Result for the MCMC Algorithm

Parameter	True Value	Average Estimates [◇]	t-statistic [†]
$\tilde{\delta}_1$	-3.3	-3.19	1.25
$\tilde{\delta}_2$	-2.01	-1.69	2.13
$\tilde{\delta}_3$	-0.77	-1.03	-1.94
$\tilde{\alpha}_0$	-3.07	-2.18	4.64*
$\tilde{\alpha}_1$	2.00	1.59	-3.62*
$\tilde{\alpha}_2$	-0.13	-0.02	0.53
$\tilde{\alpha}_3$	0.34	0.12	-1.29
$\tilde{\alpha}_4$	0.05	0.04	-1.71
$\tilde{\alpha}_5$	-1.11	-1.11	-0.04
$\tilde{\alpha}_6$	0.04	0.01	-0.4
$\tilde{\alpha}_7$	-0.35	-0.42	-0.73
$\tilde{\beta}_0$	-2.54	-2.08	7.11*
$\tilde{\beta}_1$	-0.7	-0.61	1.43
$\tilde{\beta}_2$	-0.17	-0.19	-0.24
$\tilde{\beta}_3$	0.91	0.32	-1.49
$\tilde{\beta}_4$	4.17	3.89	-1.11
$\tilde{\beta}_5$	-0.04	-0.04	1.24
$\tilde{\beta}_6$	-0.62	-0.75	-1.21
$\tilde{\beta}_7$	1.07	0.67	-4.59*
$\tilde{\beta}_8$	0.82	0.53	-4.17*
$\tilde{\gamma}_0$	-2.79	-2.62	2.24
$\tilde{\gamma}_1$	-0.32	-0.42	-0.73
$\tilde{\gamma}_2$	2.5	2.27	-1.71
$\tilde{\gamma}_3$	0.09	-0.25	-5.49*
$\tilde{\gamma}_4$	-0.02	0.12	1.25
$\tilde{\gamma}_5$	-0.29	-0.18	0.96
$\tilde{\gamma}_6$	0.15	0.02	-1.15
$\tilde{\gamma}_7$	-0.01	-0.01	-1.3
$\tilde{\gamma}_8$	-0.02	-0.22	-1.56
$\tilde{\gamma}_9$	0.18	0.06	-1.27
$\tilde{\gamma}_{10}$	-0.52	-0.49	0.21

[◇] Average estimates for 10 simulations.

[†] Average bias / SD of the 10 estimated biases $\times \sqrt{10}$, compared with $t_{0.975,9} = 2.26$.

* denotes significance at 5% level.

3.5.1 Estimation Result

We estimate the model described in Section 3.2 and Section 3.4, using the real vitamin A deficiency data. The independent standard normal distribution is also used as the prior distribution for parameters. Note that the original data may have missing visits for particular subjects. We assume that the missing visits are missing at random, i.e., for any subject, which visits are missing is independent of all the covariates, treatment, potential outcomes and outcomes. With such an assumption, the application of our model and estimation algorithm is straightforward, as missing visits simply means that the observational times are at a coarser time grid.

We have run multiple MCMC paths up to two million steps for each path, and calculated the Gelman-Rubin statistics for each parameter as a criteria for judging convergence. The Gelman-Rubin statistics for all parameters are smaller than 1.04. The estimates are in Table 3.2.

In Table 3.2, the first panel are estimates for the $\tilde{\delta}$'s, which are the logit of the δ 's. Recalling equation (3.2.1), the δ 's represent the effect of current vitamin A deficiency on current respiratory infection status. Our estimates show that for $Y_t^0 = 0$, if $A_t^* = 1$, i.e., the subject is suffering from mild vitamin A deficiency, the probability of $Y_t = 1$, i.e., the subject suffers from respiratory infection, is $\frac{\exp(-3.28)}{1+\exp(-3.28)} = 3.6\%$. If $A_t^* = 2$, the probability of $Y_t = 1$ is 11.8%. If $A_t^* = 3$, the probability of $Y_t = 1$ is 31.6%. As expected, the higher level of vitamin A deficiency is causes higher probability of respiratory infection. This shows a strong causal effect of vitamin A deficiency on

Table 3.2: Estimation Result from the Vitamin A Deficiency Data

Parameter	Posterior Mean	95% Credible Set
$\tilde{\delta}_1$	-3.30	(-4.27 , -2.57)
$\tilde{\delta}_2$	-2.01	(-3.39 , -1.00)
$\tilde{\delta}_3$	-0.77	(-2.34 , 0.85)
$\tilde{\alpha}_0$	-3.07	(-4.10 , -2.16)
$\tilde{\alpha}_1$	2.00	(1.20 , 2.93)
$\tilde{\alpha}_2$	-0.13	(-1.42 , 1.45)
$\tilde{\alpha}_3$	0.34	(-1.34 , 1.88)
$\tilde{\alpha}_4$	0.05	(0.02 , 0.08)
$\tilde{\alpha}_5$	-1.11	(-2.16 , -0.11)
$\tilde{\alpha}_6$	0.04	(-0.68 , 0.78)
$\tilde{\alpha}_7$	-0.35	(-1.09 , 0.35)
$\tilde{\beta}_0$	-2.54	(-3.23 , -1.90)
$\tilde{\beta}_1$	-0.70	(-1.84 , 0.27)
$\tilde{\beta}_2$	-0.17	(-1.46 , 0.95)
$\tilde{\beta}_3$	0.91	(-0.64 , 2.59)
$\tilde{\beta}_4$	4.17	(2.53 , 5.64)
$\tilde{\beta}_5$	-0.04	(-0.06 , -0.01)
$\tilde{\beta}_6$	-0.62	(-1.46 , 0.14)
$\tilde{\beta}_7$	1.07	(0.38 , 1.78)
$\tilde{\beta}_8$	0.82	(0.19 , 1.55)
$\tilde{\gamma}_0$	-2.79	(-3.57 , -2.12)
$\tilde{\gamma}_1$	-0.32	(-2.03 , 1.1)
$\tilde{\gamma}_2$	2.50	(1.61 , 3.45)
$\tilde{\gamma}_3$	0.09	(-1.39 , 1.47)
$\tilde{\gamma}_4$	-0.02	(-1.45 , 1.53)
$\tilde{\gamma}_5$	-0.29	(-1.70 , 1.24)
$\tilde{\gamma}_6$	0.15	(-1.81 , 1.93)
$\tilde{\gamma}_7$	-0.01	(-0.03 , 0.02)
$\tilde{\gamma}_8$	-0.02	(-0.88 , 0.77)
$\tilde{\gamma}_9$	0.18	(-0.54 , 0.87)
$\tilde{\gamma}_{10}$	-0.52	(-1.26 , 0.18)

the respiratory infection.

We also notice that $\tilde{\alpha}_0$, $\tilde{\beta}_0$ and $\tilde{\gamma}_0$ are large negative numbers, which means that all of A_t^* , Y_t^0 and L_t switch states infrequently. This in turn reassures us for our particular choice of the α , β and γ models in Section 3.4.

The credible set of $\tilde{\alpha}_1$ does not include zero, which indicates that the effect of current A_t^* level on the jump probability in the next instant is significant. The positive coefficient shows that subject is less stable in higher vitamin A deficiency levels. $\tilde{\alpha}_4$ is significantly positive and $\tilde{\alpha}_5$ is significantly negative, which shows that the older and the female children are less stable in the levels of vitamin A deficiency.

In the model for counterfactual respiratory infection, a significantly positive $\tilde{\beta}_4$ shows that respiratory infection usually do not last long. A significantly negative $\tilde{\beta}_5$ shows that younger children are prone to be on and off with respiratory infections. The significance of $\tilde{\beta}_7$ and $\tilde{\beta}_8$ shows that the respiratory infection is highly seasonal.

In the model for stunting, $\tilde{\gamma}_2$ is significantly positive, which indicates that stunting usually does not last long either.

3.5.2 Simulation Based Causal Interpretation

With the estimated model, we get an estimate of the effect of A_t^* on Y_t from the δ 's. Moreover, equipped with the full parametric model, we can answer many other causal questions by simulation.

Example Assume that Ben is 48-month old at time 1. He is not stunting, not

suffering from respiratory infection, but has mild vitamin A deficiency ($A^* = 1$). What is the probability that Ben will be suffering from respiratory infection at time 4, if he starts taking vitamin A supplements and effectively controls his deficiency level at minimum ($A^* = 0$)? What if he grows naturally without taking any vitamin A supplements? What if his deficiency level is kept constantly at 2?

The answer to the example question can be easily given provided that our model is true and we have the parameter values. We can use the information in the example as the initial condition and simulate what would happen at time 4 based on our model.

For this particular example, we can use posterior mean as our parameter value. The simulation results that answer the example questions are in Table 3.3. The numbers show that if Ben is able to control his vitamin A deficiency at either level 0 or 1, it would benefit him in terms of the probability of suffering from respiratory infection. Note the probabilities of $Y_4 = 1$ for different levels of A^* are close to but different from the δ 's. The numbers in Table 3.3 fully incorporate the dynamic evolution of all variables.

The 95% credible set is obtained by simulation with parameters values being samples from the posterior distribution of the parameters, which are available from our MCMC algorithm. This takes the uncertainty of estimated parameters into account.

For a standard causal comparison, the improvement in $P(Y_4 = 1)$ for controlling A^* at 0 over growing naturally is around 4.8%, with a 95% credible set (1.8%, 7.8%).

Table 3.3: Simulation Based Causal Interpretation for the Example

A^* level	$P(Y_4 = 1)$	95% Credible Set
grow naturally	7.0%	(4.6%, 10.3%)
0	2.2%	(0.9%, 6.3%)
1	5.4%	(3.4%, 9.6%)
2	13.2%	(3.5%, 23.9%)
3	32.2%	(9.9%, 75.5%)

The improvement in $P(Y_4 = 1)$ for controlling A^* at 1 over growing naturally is around 1.6%, with a 95% credible set $(-0.1\%, 4.6\%)$. The credible set is also obtained by simulations from the posterior distribution of the parameters.

3.5.3 Model Assumptions Revisited

We have made many assumptions in our model and our causal interpretation relies heavily on the soundness of these assumptions. In this section, we experiment with a few variations of our model. These results largely confirm that the model we used in Section 3.2 is a reasonable and consistent description of the data.

Number of Levels for Vitamin A Deficiency

We estimated our model by assuming that vitamin A deficiency has four levels. The lower two levels do not exhibit xerophthalmia, while the upper two levels do. It is natural to ask if such a discretization is sufficient or necessary. We could consider two alternatives of our model:

- *2-level A^** We assume that A_t^* only has two levels, 0 and 1. $A_t = 1$ if $A_t^* = 1$, and $A_t = 0$ if $A_t^* = 0$. Assuming other model specifications are the same as

before, we do not have any hidden levels of treatment.

- *6-level A^** We assume that A_t^* has six levels, 0, 1, 2, 3, 4, and 5. $A_t = 1$ if $A_t^* > 2$, and $A_t = 0$ if $A_t^* < 3$, and that other model specifications are the same as before.

Under the *2-level A^** model, we assume that if $A_t = 0$, Y_t is solely decided by Y_t^0 , and that we only need to estimate parameter δ_1 that is associated with $A_t^* = 1$. Under the *6-level A^** model, we need to estimate parameters δ_j , $j = 1, \dots, 5$, which are associated with $A_t^* = 1, \dots, 5$ respectively.

We estimate the both models using our MCMC algorithm and summarize the estimates for $\tilde{\delta}$'s in Table 3.4. In the table we denote our original model as the *4-level A^** model. The 95% credible interval for each parameter is below its estimate. From the estimates in the table, we believe that the *4-level A^** is a reasonable choice. In the middle column, $\tilde{\delta}_2$ and $\tilde{\delta}_3$ are distinct enough from each other, and grouping them together as in the *2-level A^** model is over-simplification. One evidence is that in the *4-level A^** model estimates, $\tilde{\delta}_3$ does not fall into the CI for $\tilde{\delta}_2$. However, using the *6-level A^** model is un-necessary. For example, in the *6-level A^** model, $\tilde{\delta}_4$ and $\tilde{\delta}_5$ are very close to each other, and their CI's cover each other; $\tilde{\delta}_2$ and $\tilde{\delta}_3$ are also so close to each that their CI's cover each other. This shows that the *4-level A^** model has captured more structures than the *2-level A^** model and that the *6-level A^** model is not capturing more structures.

Table 3.4: Estimates of $\tilde{\delta}$ from Different Models

2-level A^*	4-level A^*	6-level A^*
	$\tilde{\delta}_1$ -3.30 (-4.27 , -2.57)	$\tilde{\delta}_1$ -3.39 (-4.29 , -2.59) $\tilde{\delta}_2$ -2.87 (-3.81 , -2.11)
$\tilde{\delta}_1$ -2.08 (-3.27 , -1.12)	$\tilde{\delta}_2$ -2.01 (-3.39 , -1.00) $\tilde{\delta}_3$ -0.77 (-2.34 , 0.85)	$\tilde{\delta}_3$ -1.79 (-2.98 , -0.83) $\tilde{\delta}_4$ -0.88 (-2.86 , 0.99) $\tilde{\delta}_5$ -0.35 (-2.15 , 1.58)

The same phenomenon can be observed in terms of estimated causal effect. It is difficult to directly compare the difference in counterfactual outcomes under different models, as the treatment levels mean different things in these models. We consider the following causal comparison as a fair comparison:

Example Consider all the boys who are 48-month old at time 1, not stunting, not suffering from respiratory infection, but has symptoms of xerophthalmia ($A_1 = 1$). We randomly pick any one of them and randomly pick an A^* level that corresponds to $A = 0$, and treat him with vitamin A supplement such that his vitamin A deficiency level is kept at that A^* level. At time 4, what is the expected difference between the probability of getting respiratory infection if the child grows naturally and the probability of getting respiratory infection if the child's vitamin A deficiency is kept at the randomly picked A^* level?

Note that the expectation is taken over the population of all such boys and the A^* levels that corresponds to $A = 0$. We recall that in Section 3.2.2, we have assumed

Table 3.5: Estimates of the Average Causal Difference for Different Models

	2-level A^*	4-level A^*	6-level A^*
Causal Difference	3.2%	12.9%	20.4%
CI	(1.0%, 9.5%)	(4.6%, 24.8%)	(10.4%, 38.8%)

that the initial distribution of A^* given A is uniform over the corresponding levels. The expected causal effect is the simple average of causal difference for all possible combinations of A^* levels corresponding to $A = 1$ and A^* levels corresponding to $A = 0$.

We summarize our simulation-based estimate of the average causal difference in Table 3.5. The estimated causal differences for the *4-level* A^* model and the *6-level* A^* model are close to each other, as their credible intervals cover each other. The estimated causal difference for the *2-level* A^* model is significantly different from the other two estimates. As a result, we believe that discretizing vitamin A deficiency into four levels in our analysis is appropriate.

Continuous Time Ignorability

We assumed *continuous time ignorability* in our model (see Remark 3.2.1). With a fully parametric model, we do not need the assumption for identification. For example, we can assume that the α and β functions as the follows:

$$\begin{aligned} & \alpha(A_{t+h}^*; A_t^*, L_t, Y_t, Y_t^0, age, sex, t; \tilde{\alpha}) \\ &= \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 A_t^* + \tilde{\alpha}_2 L_t + \tilde{\alpha}_3 Y_t + \tilde{\alpha}_4(age) + \tilde{\alpha}_5(sex) + \tilde{\alpha}_6 \cos(t) \\ & \quad + \tilde{\alpha}_7 \sin(t) + \tilde{\alpha}_8 Y_t^0) \end{aligned}$$

$$\begin{aligned}
& / \{2[1 + \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 A_t^* + \tilde{\alpha}_2 L_t + \tilde{\alpha}_3 Y_t + \tilde{\alpha}_4(\text{age}) + \tilde{\alpha}_5(\text{sex}) + \tilde{\alpha}_6 \cos(t) \\
& \quad + \tilde{\alpha}_7 \sin(t) + \tilde{\alpha}_8 Y_t^0)]\} \\
& \beta(Y_{t+h}^{\bar{A}_{(t+h)}^*}{}^{-,0}; A_{t+h}^*, X_t, \text{age}, \text{sex}, t; \tilde{\beta}) \\
& = \exp(\tilde{\beta}_0 + \tilde{\beta}_1 A_t^* + \tilde{\beta}_2 L_t + \tilde{\beta}_3 Y_t + \tilde{\beta}_4 Y_t^{\bar{A}_t^*}{}^{-,0} + \tilde{\beta}_5(\text{age}) + \tilde{\beta}_6(\text{sex}) \\
& \quad + \tilde{\beta}_7 \cos(t) + \tilde{\beta}_8 \sin(t) + \tilde{\beta}_9 A_{t+h}^*) \\
& / \{1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 A_t^* + \tilde{\beta}_2 L_t + \tilde{\beta}_3 Y_t + \tilde{\beta}_4 Y_t^{\bar{A}_t^*}{}^{-,0} + \tilde{\beta}_5(\text{age}) + \tilde{\beta}_6(\text{sex}) \\
& \quad + \tilde{\beta}_7 \cos(t) + \tilde{\beta}_8 \sin(t) + \tilde{\beta}_9 A_{t+h}^*)\}
\end{aligned}$$

By allowing α depending on Y_t^0 and β depending on A_{t+h}^* , we are no longer assuming the continuous time ignorability assumption (see Remark 3.2.1).

With this model, we summarize our estimates in Table 3.6. For comparison, we include the estimates from the original model. The results show that the new parameters $\tilde{\alpha}_8$ and $\tilde{\beta}_9$ are not significant at all. The new estimates of other parameters are close to the original estimates. This indicates that continuous time ignorability might be a reasonable assumption.

Functional Form of Q

With the similar idea, different functional forms of Q can be experimented to see if we have missed out any relationship in our estimated model. For example, one concern is that our models for α and β should contain interaction terms. One might expect that if the child is having respiratory infection, the influence of vitamin A

Table 3.6: Estimation Result from the Vitamin A Deficiency Data without Continuous Time Ignorability

Parameter	Original Estimate	New Posterior Mean	95% Credible Set
$\tilde{\delta}_1$	-3.30	-3.27	(-4.06 , -2.58)
$\tilde{\delta}_2$	-2.01	-1.85	(-3.11 , -0.85)
$\tilde{\delta}_3$	-0.77	-0.87	(-2.85 , 0.86)
$\tilde{\alpha}_0$	-3.07	-3.02	(-4.02 , -1.93)
$\tilde{\alpha}_1$	2.00	1.95	(1.11 , 2.85)
$\tilde{\alpha}_2$	-0.13	-0.20	(-1.52 , 1.29)
$\tilde{\alpha}_3$	0.34	0.49	(-1.01 , 1.98)
$\tilde{\alpha}_4$	0.05	0.05	(0.02 , 0.08)
$\tilde{\alpha}_5$	-1.11	-1.15	(-2.20 , -0.18)
$\tilde{\alpha}_6$	0.04	0.05	(-0.65 , 0.79)
$\tilde{\alpha}_7$	-0.35	-0.35	(-1.03 , 0.32)
$\tilde{\alpha}_8$		0.04	(-2.1 , 1.74)
$\tilde{\beta}_0$	-2.54	-2.59	(-3.35 , -1.93)
$\tilde{\beta}_1$	-0.70	-0.3	(-1.88 , 1.18)
$\tilde{\beta}_2$	-0.17	-0.06	(-1.33 , 1.24)
$\tilde{\beta}_3$	0.91	2.59	(1.25 , 4.00)
$\tilde{\beta}_4$	4.17	3.19	(1.81 , 4.63)
$\tilde{\beta}_5$	-0.04	-0.04	(-0.07 , -0.02)
$\tilde{\beta}_6$	-0.62	-0.65	(-1.63 , 0.23)
$\tilde{\beta}_7$	1.07	1.18	(0.50 , 1.96)
$\tilde{\beta}_8$	0.82	0.93	(0.18 , 1.69)
$\tilde{\beta}_9$		-0.86	(-2.26 , 0.66)
$\tilde{\gamma}_0$	-2.79	-2.79	(-3.53 , -2.07)
$\tilde{\gamma}_1$	-0.32	-0.21	(-1.64 , 1.24)
$\tilde{\gamma}_2$	2.50	2.51	(1.67 , 3.45)
$\tilde{\gamma}_3$	0.09	-0.04	(-1.6 , 1.57)
$\tilde{\gamma}_4$	-0.02	0.13	(-1.41 , 1.72)
$\tilde{\gamma}_5$	-0.29	-0.37	(-1.85 , 0.97)
$\tilde{\gamma}_6$	0.15	0.02	(-1.54 , 1.78)
$\tilde{\gamma}_7$	-0.01	-0.01	(-0.03 , 0.01)
$\tilde{\gamma}_8$	-0.02	-0.01	(-0.83 , 0.81)
$\tilde{\gamma}_9$	0.18	0.18	(-0.51 , 0.85)
$\tilde{\gamma}_{10}$	-0.52	-0.53	(-1.19 , 0.1)

deficiency on the child's transition probability may be different from that if the child is not having respiratory infection. To see if this is the case, we parametrize α and β as

$$\begin{aligned}
& \alpha(A_{t+h}^*; A_t^*, L_t, Y_t, Y_t^0, age, sex, t; \tilde{\alpha}) \\
&= \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 A_t^* + \tilde{\alpha}_2 L_t + \tilde{\alpha}_3 Y_t + \tilde{\alpha}_4(age) + \tilde{\alpha}_5(sex) + \tilde{\alpha}_6 \cos(t) \\
&\quad + \tilde{\alpha}_7 \sin(t) + \tilde{\alpha}_8 A_t^* Y_t) \\
& / \{2[1 + \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 A_t^* + \tilde{\alpha}_2 L_t + \tilde{\alpha}_3 Y_t + \tilde{\alpha}_4(age) + \tilde{\alpha}_5(sex) + \tilde{\alpha}_6 \cos(t) \\
&\quad + \tilde{\alpha}_7 \sin(t) + \tilde{\alpha}_8 A_t^* Y_t)]\} \\
& \\
& \beta(Y_{t+h}^{\bar{A}_{(t+h)}^*}; A_{t+h}^*, X_t, age, sex, t; \tilde{\beta}) \\
&= \exp(\tilde{\beta}_0 + \tilde{\beta}_1 A_t^* + \tilde{\beta}_2 L_t + \tilde{\beta}_3 Y_t + \tilde{\beta}_4 Y_t^{\bar{A}_{t-}^*, 0} + \tilde{\beta}_5(age) + \tilde{\beta}_6(sex) \\
&\quad + \tilde{\beta}_7 \cos(t) + \tilde{\beta}_8 \sin(t) + \tilde{\beta}_9 A_t^* Y_t) \\
& / \{1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 A_t^* + \tilde{\beta}_2 L_t + \tilde{\beta}_3 Y_t + \tilde{\beta}_4 Y_t^{\bar{A}_{t-}^*, 0} + \tilde{\beta}_5(age) + \tilde{\beta}_6(sex) \\
&\quad + \tilde{\beta}_7 \cos(t) + \tilde{\beta}_8 \sin(t) + \tilde{\beta}_9 A_t^* Y_t)\}
\end{aligned}$$

The estimating results from this model is in Table 3.7. The results suggest that the interaction term, $A_t^* Y_t$, is not significant in either the α model or the β model. Estimates of other parameters are close to the original estimates. We do not find significant evidence against our original plain vanilla model.

Table 3.7: Estimation Result from the Vitamin A Deficiency Data with a Different Q

Parameter	Original Estimate	New Posterior Mean	95% Credible Set
$\tilde{\delta}_1$	-3.30	-3.39	(-4.33 , -2.66)
$\tilde{\delta}_2$	-2.01	-1.73	(-2.72 , -0.86)
$\tilde{\delta}_3$	-0.77	-1.04	(-3.06 , 0.70)
$\tilde{\alpha}_0$	-3.07	-2.97	(-4.09 , -1.94)
$\tilde{\alpha}_1$	2.00	1.78	(0.93 , 2.66)
$\tilde{\alpha}_2$	-0.13	-0.27	(-1.59 , 1.29)
$\tilde{\alpha}_3$	0.34	0.14	(-1.33 , 1.6)
$\tilde{\alpha}_4$	0.05	0.05	(0.02 , 0.08)
$\tilde{\alpha}_5$	-1.11	-1.01	(-2.01 , 0.01)
$\tilde{\alpha}_6$	0.04	0.03	(-0.69 , 0.72)
$\tilde{\alpha}_7$	-0.35	-0.40	(-1.11 , 0.29)
$\tilde{\alpha}_8$		0.89	(-0.31 , 2.29)
$\tilde{\beta}_0$	-2.54	-2.56	(-3.37 , -1.85)
$\tilde{\beta}_1$	-0.70	-1.16	(-2.38 , -0.05)
$\tilde{\beta}_2$	-0.17	-0.18	(-1.67 , 1.14)
$\tilde{\beta}_3$	0.91	2.49	(1.05 , 3.86)
$\tilde{\beta}_4$	4.17	3.05	(1.62 , 4.42)
$\tilde{\beta}_5$	-0.04	-0.04	(-0.07 , -0.02)
$\tilde{\beta}_6$	-0.62	-0.70	(-1.63 , 0.21)
$\tilde{\beta}_7$	1.07	1.26	(0.59 , 1.99)
$\tilde{\beta}_8$	0.82	0.90	(0.21 , 1.66)
$\tilde{\beta}_9$		0.69	(-0.49 , 1.97)
$\tilde{\gamma}_0$	-2.79	-2.86	(-3.67 , -2.18)
$\tilde{\gamma}_1$	-0.32	-0.20	(-2.02 , 1.32)
$\tilde{\gamma}_2$	2.50	2.50	(1.64 , 3.39)
$\tilde{\gamma}_3$	0.09	-0.08	(-1.49 , 1.31)
$\tilde{\gamma}_4$	-0.02	0.24	(-1.33 , 1.90)
$\tilde{\gamma}_5$	-0.29	-0.26	(-1.89 , 1.37)
$\tilde{\gamma}_6$	0.15	0.08	(-1.56 , 1.70)
$\tilde{\gamma}_7$	-0.01	-0.01	(-0.03 , 0.02)
$\tilde{\gamma}_8$	-0.02	-0.02	(-0.81 , 0.80)
$\tilde{\gamma}_9$	0.18	0.17	(-0.49 , 0.89)
$\tilde{\gamma}_{10}$	-0.52	-0.54	(-1.17 , 0.13)

3.6 Conclusion

In this chapter, we have considered a data set where it is reasonable to assume that the treatment process, covariates, counterfactual outcome and the outcome follow a continuous time process. However, only a coarsened indicator of the treatment level, the covariates and the outcome are observed and they are only observed at the discrete observational time points. We are facing problems with

- 1) Unmeasured confounders caused by the missing data in between two consecutive observational time points, and
- 2) Measurement error in treatment levels, as the amount of treatment is not directly observable and the treatment levels in between two observational time points are not observable.

In Chapter 2, we have shown that 1) would cause problems in standard longitudinal estimations that are based on ignorability assumption in the observational data. Even if the continuous time ignorability holds, the observational time ignorability may not hold. 2) usually causes a even more severe problem in standard semi-parametric methods, as one does not even know exactly how much treatment the subject has received.

With both 1) and 2) in our data set, it requires more modeling assumption than standard causal methods. We have chosen to fully model the process by a continuous time non-stationary Markov process and assume that the data are discrete time

observations of the process. We have also designed a MCMC algorithm with data augmentation to estimate the parameters by constructing a reasonable proposal distribution for the augmented data. The Markov model and the MCMC algorithm work well for our vitamin A deficiency data, as we have shown that in Section 3.4 the MCMC algorithm does produce estimates that are close to the true values and that in Section 3.5 our estimates make sense in the context of vitamin A deficiency and respiratory infection.

In the vitamin A deficiency study, we find that the levels of vitamin A deficiency has a strong causal effect on the respiratory infection, as is evidenced by the values of δ_1 , δ_2 and δ_3 . By simulation, we are also able to estimate the causal effect for certain treatment regime through time. For example, we can estimate the difference in the probabilities of having respiratory disease between keeping vitamin A deficiency at the lowest level and keeping vitamin A deficiency at the highest level. The luxury comes with the cost of more extensive modeling assumptions.

This chapter serves as an example for causal inference in longitudinal data with a full model, when standard semi-parametric methods are not applicable. In particular, we have dealt with binary outcomes, binary covariates (it could be generalized to discrete covariates), discrete treatment levels and that the treatment does not affect the outcome in a cumulative way. Many other examples have yet to be worked out, e.g., examples with continuous covariates, examples with continuous outcomes, etc.. When working on a full causal model, domain knowledge, e.g., the biological relation-

ships among different variables, plays the most important role. Still, some modeling considerations and computational techniques in this chapter could be borrowed and generalized to other real world problems.

Chapter 4

Controlling-the-future Revisited: the Optimal Estimating Equation

The relaxed sequential randomization assumption (2.4.1) and the controlling-the-future estimating equation (2.4.3) are powerful generalization of the standard g-estimation. As we have shown in Chapter 2, they are especially useful in correcting the bias caused by the unmeasured confounders in our setting of inferring causal effect from discretely observed continuous time processes. In this chapter, we revisit the controlling-the-future method from a theoretical point of view. We would develop a semi-parametric theory that is parallel to the standard g-estimation, including the derivation of the efficient score function and the locally efficient doubly robust estimator. This chapter justifies that the estimating equation (2.4.3) we used in Chapter 2 would produce \sqrt{n} consistent and asymptotically normal and regular estimators.

The organization of this chapter is as follows: in Section 4.1, we consider a single

period model with relaxed ignorability assumption and derive its nuisance tangent space and the efficient score; Section 4.2 proposes a useful estimator that is motivated by the efficient score function and proves that the estimator is locally efficient and doubly robust; in Section 4.3, we consider two special cases of the general theory; Section 4.4 extends the model to multi-period case; Section 4.5 concludes the chapter.

4.1 A Single Period Semi-parametric Model Under Relaxed Ignorability Assumption

4.1.1 The Single Period Model

We consider a one-period deterministic model, with two outcomes Y_1 and Y_2 . Let (Y_1^0, Y_2^0) denote the potential outcomes if the subject does not receive any treatment, and (Y_1, Y_2) denote the observed outcome. Let A be the treatment assignment and X be the collection of pre-treatment covariates.

We assume that h_1 and h_2 are the blip-down functions, such that $Y_1^0 = h_1(Y_1, A, X; \Psi_0)$, and that $Y_2^0 = h_2(Y_2, Y_1, A, X; \Psi_0)$. Here h_1 and h_2 are parametrized by Ψ . When Ψ takes its true value Ψ_0 , the functions correctly blip Y_1 and Y_2 down to Y_1^0 and Y_2^0 respectively. We denote $Y_1^0(\Psi) = h_1(Y_1, A, X; \Psi)$ and $Y_2^0(\Psi) = h_2(Y_2, Y_1, A, X; \Psi)$ as functions of Ψ . In what follows, we use lower case to denote the realization of these random variables.

We assume the following version of ignorability

$$Y_2^0 \perp\!\!\!\perp A|X, Y_1^0 \quad (4.1.1)$$

With this model, the likelihood function of the observed variables for one subject can be written as

$$\begin{aligned} f_{Y_1, Y_2, A, X}(y_1, y_2, a, x) &= \frac{\partial h_1}{\partial y_1} \frac{\partial h_2}{\partial y_2} f_{Y_1^0, Y_2^0, A, X}(y_1^0, y_2^0, a, x) \\ &= \frac{\partial h_1}{\partial y_1} \frac{\partial h_2}{\partial y_2} f_{X, Y_1^0}(x, y_1^0(\Psi_0); \eta_{10}) \\ &\quad \times f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0); \eta_{20}) f_{A|X, Y_1^0}(a|x, y_1^0(\Psi_0); \eta_{30}) \end{aligned} \quad (4.1.2)$$

Here, the parameters of our model are $\Psi, \eta_1, \eta_2, \eta_3$. Their true values are $\Psi_0, \eta_{10}, \eta_{20}$, and η_{30} . The parameter of interest is Ψ . We assume that Ψ is a q dimensional vector. η_1, η_2, η_3 are nuisance parameters, possibly infinite dimensional. (4.1.2) here describes the semi-parametric model under the sole condition that (4.1.1) is true.

4.1.2 Characterization of the Nuisance Tangent Space

Consider any parametric submodel:

$$\begin{aligned} &f_{Y_1, Y_2, A, X}(y_1, y_2, a, x; \Psi, \gamma_1, \gamma_2, \gamma_3) \\ &= \frac{\partial h_1}{\partial y_1} \frac{\partial h_2}{\partial y_2} f_{X, Y_1^0}(x, y_1^0(\Psi); \gamma_1) \\ &\quad \times f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi)|x, y_1^0(\Psi); \gamma_2) f_{A|X, Y_1^0}(a|x, y_1^0(\Psi); \gamma_3) \end{aligned}$$

where $\gamma_1, \gamma_2, \gamma_3$ are of dimension n_1, n_2 and n_3 respectively.

The nuisance score of this parametric submodel is

$$S_{\gamma_0} = \{S_{\gamma_{10}}^T, S_{\gamma_{20}}^T, S_{\gamma_{30}}^T\}^T$$

where

$$S_{\gamma_{10}} = \frac{\partial \log f_{X, Y_1^0}(x, y_1^0(\Psi_0); \gamma_{10})}{\partial \gamma_1},$$

$$S_{\gamma_{20}} = \frac{\partial \log f_{Y_2^0 | X, Y_1^0}(y_2^0(\Psi_0) | x, y_1^0(\Psi_0); \gamma_{20})}{\partial \gamma_2},$$

and

$$S_{\gamma_{30}} = \frac{\partial \log f_{A | X, Y_1^0}(a | x, y_1^0(\Psi_0); \gamma_{30})}{\partial \gamma_3}.$$

Define the following sub-spaces:

$$\Lambda_{\gamma_0} = \{B_{q \times (n_1 + n_2 + n_3)} S_{\gamma_0}\}$$

$$\Lambda_{\gamma_{10}} = \{B_{q \times n_1} S_{\gamma_{10}}\}$$

$$\Lambda_{\gamma_{20}} = \{B_{q \times n_2} S_{\gamma_{20}}\}$$

$$\Lambda_{\gamma_{30}} = \{B_{q \times n_3} S_{\gamma_{30}}\}$$

It is easy to see that $\Lambda_{\gamma_0} = \Lambda_{\gamma_{10}} \oplus \Lambda_{\gamma_{20}} \oplus \Lambda_{\gamma_{30}}$. We define $\Lambda, \Lambda_1, \Lambda_2, \Lambda_3$ to be the mean-square closure of all $\Lambda_{\gamma_0}, \Lambda_{\gamma_{10}}, \Lambda_{\gamma_{20}}, \Lambda_{\gamma_{30}}$, respectively. It can also be shown that $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$ (see the argument on page 77 of Tsiatis (2006)). Here Λ is the nuisance tangent space.

Λ_1, Λ_2 and Λ_3 can be characterized as follows:

Lemma 4.1.1.

$$\Lambda_1 = \{a_1(X, Y_1^0(\Psi_0)) : E[a_1(X, Y_1^0(\Psi_0))] = 0_{1 \times q}\}$$

$$\Lambda_2 = \{a_2(Y_1^0(\Psi_0), Y_2^0(\Psi), X) : E[a_2(Y_1^0(\Psi_0), Y_2^0(\Psi_0), X)|X, Y_1^0(\Psi_0)] = 0_{1 \times q}\}$$

$$\Lambda_3 = \{a_3(A, X, Y_1^0(\Psi_0)) : E[a_3(A, X, Y_1^0(\Psi_0))|X, Y_1^0(\Psi_0)] = 0_{1 \times q}\}$$

where a_1, a_2, a_3 all have finite variance.

The proof of this lemma is routine and is omitted. See Tsiatis (2006).

We also notice that Λ_1, Λ_2 and Λ_3 are orthogonal to each other, which can be proved as follows.

- For $\Lambda_1 \perp \Lambda_2$, pick any $a_1 \in \Lambda_1$ and $a_2 \in \Lambda_2$.

$$E[a_1 a_2^T] = E[a_1 E[a_2^T | X, Y_1^0(\Psi_0)]] = 0$$

- For $\Lambda_1 \perp \Lambda_3$, pick any $a_1 \in \Lambda_1$ and $a_3 \in \Lambda_3$.

$$E[a_1 a_3^T] = E[a_1 E[a_3^T | X, Y_1^0(\Psi_0)]] = 0$$

- For $\Lambda_2 \perp \Lambda_3$, pick any $a_2 \in \Lambda_2$ and $a_3 \in \Lambda_3$.

$$E[a_2 a_3^T] = E[a_2 E[a_3^T | Y_1^0(\Psi_0), Y_2^0(\Psi_0), X]] = E[a_2 E[a_3^T | Y_1^0(\Psi_0), X]] = 0$$

The second equality is because of the ignorability assumption.

4.1.3 The Space that is Orthogonal to the Nuisance Tangent Space

The Hilbert space $\mathcal{H} = \Lambda \oplus \Lambda^\perp$. Consider any function $g(A, X, Y_1^0, Y_2^0) \in \mathcal{H}$. Its projection on to Λ is

$$\prod(g|\Lambda) = \prod(g|\Lambda_1) + \prod(g|\Lambda_2) + \prod(g|\Lambda_3)$$

We then show the following facts.

Lemma 4.1.2.

$$\prod(g|\Lambda_1) = E[g|X, Y_1^0(\Psi_0)]$$

$$\prod(g|\Lambda_2) = E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] - E[g|X, Y_1^0(\Psi_0)]$$

$$\prod(g|\Lambda_3) = E[g|A, X, Y_1^0(\Psi_0)] - E[g|X, Y_1^0(\Psi_0)]$$

Proof. For $\prod(g|\Lambda_1) = E[g|X, Y_1^0(\Psi_0)]$, we can show that for any $a_1 \in \Lambda_1$

$$\begin{aligned} & E[\{g - E[g|X, Y_1^0(\Psi_0)]\}a_1^T] \\ &= E\{E[(g - E[g|X, Y_1^0(\Psi_0)])a_1^T|X, Y_1^0(\Psi_0)]\} \\ &= E\{E[(g - E[g|X, Y_1^0(\Psi_0)])|X, Y_1^0(\Psi_0)]a_1^T\} \\ &= 0 \end{aligned}$$

For $\prod(g|\Lambda_2) = E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] - E[g|X, Y_1^0(\Psi_0)]$, we can show that for any $a_2 \in \Lambda_2$,

$$\begin{aligned} & E[\{g - E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)]\}a_2^T] \\ &= E\{E[(g - E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)])a_2^T|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)]\} \\ &= E\{a_2(E[(g - E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)])|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)]^T)\} \\ &= E\{a_2(E[g|X, Y_1^0(\Psi_0)])^T\} \\ &= 0 \end{aligned}$$

The last equality is because Λ_1 and Λ_2 are orthogonal to each other and $E[g|X, Y_1^0(\Psi_0)] \in \Lambda_1$.

For $\prod(g|\Lambda_3) = E[g|A, X, Y_1^0(\Psi_0)] - E[g|X, Y_1^0(\Psi_0)]$, given any $a_3 \in \Lambda_3$,

$$\begin{aligned}
& E[\{g - E[g|A, X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)]\}a_3^T] \\
&= E\{E[\{g - E[g|A, X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)]\}a_3^T|A, X, Y_1^0(\Psi_0)]\} \\
&= E\{a_3(E[\{g - E[g|A, X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)]\}|A, X, Y_1^0(\Psi_0)]^T)\} \\
&= E\{a_3(E[g|X, Y_1^0(\Psi_0)]^T)\} \\
&= 0
\end{aligned}$$

The last equality is because Λ_1 and Λ_3 are orthogonal to each other. \square

Therefore,

$$\prod(g|\Lambda) = E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] + E[g|A, X, Y_1^0(\Psi_0)] - E[g|X, Y_1^0(\Psi_0)],$$

and

$$\prod(g|\Lambda^\perp) = g - E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] - E[g|A, X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)].$$

It can be concluded that

$$\Lambda^\perp = \{g - E[g|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] - E[g|A, X, Y_1^0(\Psi_0)] + E[g|X, Y_1^0(\Psi_0)], g \in \mathcal{H}\}. \quad (4.1.3)$$

4.1.4 The Efficient Score

Consider the score function of our interest

$$S_{\Psi_0} = \frac{\partial \log(f_{Y_1, Y_2, A, X})}{\partial \Psi} \quad (4.1.4)$$

$$\begin{aligned}
&= \frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} + \frac{\partial \log \frac{\partial h_2(\Psi_0)}{\partial y_2}}{\partial \Psi} + \frac{\partial \log f_{Y_1^0, X}(y_1^0(\Psi_0), x)}{\partial \Psi} \\
&\quad + \frac{\partial \log f_{Y_2^0 | X, Y_1^0}(y_2^0(\Psi_0) | x, y_1^0(\Psi_0))}{\partial \Psi} + \frac{\partial \log f_{A | X, Y_1^0}(a | x, y_1^0(\Psi_0))}{\partial \Psi}.
\end{aligned}$$

Note that S_{Ψ_0} is a q dimensional vector.

The efficient score is then

$$S_{eff} = S_{\Psi_0} - E[S_{\Psi_0} | Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] + E[S_{\Psi_0} | X, Y_1^0(\Psi_0)] - E[S_{\Psi_0} | A, X, Y_1^0(\Psi_0)].$$

We do a few more steps of calculation to see what the efficient score looks like.

First, re-write S_{Ψ_0} as

$$\begin{aligned}
S_{\Psi_0} &= \frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} + \frac{\partial \log \frac{\partial h_2(\Psi_0)}{\partial y_2}}{\partial \Psi} + \frac{\partial \log f_{Y_1^0, X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \\
&\quad + \frac{\partial \log f_{Y_2^0 | X, Y_1^0}(y_2^0(\Psi_0) | x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \\
&\quad + \frac{\partial \log f_{Y_2^0 | X, Y_1^0}(y_2^0(\Psi_0) | x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} \frac{\partial h_2(\Psi_0)}{\partial \Psi} \\
&\quad + \frac{\partial \log f_{A | X, Y_1^0}(a | x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi}
\end{aligned}$$

Then, if we factorize the likelihood as

$$\begin{aligned}
&f_{Y_1, Y_2, A, X}(y_1, y_2, a, x) \\
&= \frac{\partial h_1}{\partial y_1} \frac{\partial h_2}{\partial y_2} f_{X, Y_1^0}(x, y_1^0(\Psi_0)) \\
&\quad \times f_{Y_2^0 | X, Y_1^0}(y_2^0(\Psi_0) | x, y_1^0(\Psi_0)) f_{A | X, Y_1^0, Y_2^0}(a | x, y_1^0(\Psi_0), y_2^0(\Psi_0)),
\end{aligned}$$

obviously,

$$E\left[\frac{\partial \log f_{A | X, Y_1^0, Y_2^0}(a | x, y_1^0(\Psi_0), y_2^0(\Psi_0); \eta_{30})}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] = 0_{q \times 1}.$$

It is easy to see that

$$\begin{aligned}
& E[S_{\Psi_0} | Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] \\
&= E\left[\frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} + \frac{\partial \log \frac{\partial h_2(\Psi_0)}{\partial y_2}}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \\
&\quad + \frac{\partial \log f_{Y_1^0|X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} E\left[\frac{\partial h_1(\Psi_0)}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \\
&\quad + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0) | x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} E\left[\frac{\partial h_1(\Psi_0)}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \\
&\quad + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0) | x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} E\left[\frac{\partial h_2(\Psi_0)}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right].
\end{aligned}$$

If we factorize the likelihood as

$$\begin{aligned}
& f_{Y_1, Y_2, A, X}(y_1, y_2, a, x) \\
&= \frac{\partial h_1}{\partial y_1} f_{X, Y_1^0}(x, y_1^0(\Psi_0)) f_{A, Y_2|X, Y_1^0}(a, y_2 | x, y_1^0(\Psi_0)),
\end{aligned}$$

obviously

$$E\left[\frac{\partial \log f_{A, Y_2|X, Y_1^0}(a, y_2 | x, y_1^0(\Psi_0))}{\partial \Psi} \middle| X, Y_1^0(\Psi_0)\right] = 0_{q \times 1}$$

It is easy to verify that

$$\begin{aligned}
& E[S_{\Psi_0} | X, Y_1^0(\Psi_0)] \\
&= E\left[\frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} \middle| X, Y_1^0(\Psi_0)\right] + \frac{\partial \log f_{Y_1^0|X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} E\left[\frac{\partial h_1(\Psi_0)}{\partial \Psi} \middle| X, Y_1^0(\Psi_0)\right].
\end{aligned}$$

Similarly, if we factorize the likelihood as

$$\begin{aligned}
& f_{Y_1, Y_2, A, X}(y_1, y_2, a, x) \\
&= \frac{\partial h_1}{\partial y_1} f_{X, Y_1^0}(x, y_1^0(\Psi_0)) f_{A|X, Y_1^0}(a | x, y_1^0(\Psi_0)) f_{Y_2|A, X, Y_1^0}(y_2 | a, x, y_1^0(\Psi_0)),
\end{aligned}$$

obviously

$$E\left[\frac{\partial \log f_{Y_2|A,X,Y_1^0}(y_2|a, x, y_1^0(\Psi_0))}{\partial \Psi} \middle| A, X, Y_1^0(\Psi_0)\right] = 0_{q \times 1}$$

It is easy to see that

$$\begin{aligned} & E[S_{\Psi_0} | A, X, Y_1^0(\Psi_0)] \\ &= \frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} + \frac{\partial \log f_{Y_1^0, X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \\ & \quad + \frac{\partial \log f_{A|X, Y_1^0}(a|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi}. \end{aligned}$$

Then we consider

$$\begin{aligned} & S_{eff} \\ &= S_{\Psi_0} - E[S_{\Psi_0} | Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] + E[S_{\Psi_0} | X, Y_1^0(\Psi_0)] - E[S_{\Psi_0} | A, X, Y_1^0(\Psi_0)] \\ &= \frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} + \frac{\partial \log \frac{\partial h_2(\Psi_0)}{\partial y_2}}{\partial \Psi} + \frac{\partial \log f_{Y_1^0, X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \\ & \quad + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \\ & \quad + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} \frac{\partial h_2(\Psi_0)}{\partial \Psi} \\ & \quad + \frac{\partial \log f_{A|X, Y_1^0}(a|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \\ & \quad - \left\{ E\left[\frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} + \frac{\partial \log \frac{\partial h_2(\Psi_0)}{\partial y_2}}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \right. \\ & \quad + \frac{\partial \log f_{Y_1^0, X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} E\left[\frac{\partial h_1(\Psi_0)}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \\ & \quad + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} E\left[\frac{\partial h_1(\Psi_0)}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \\ & \quad \left. + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} E\left[\frac{\partial h_2(\Psi_0)}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \right\} \\ & \quad + \left\{ E\left[\frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} \middle| X, Y_1^0(\Psi_0)\right] \right\} \end{aligned}$$

$$\begin{aligned}
& + \frac{\partial \log f_{Y_1^0, X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} E\left[\frac{\partial h_1(\Psi_0)}{\partial \Psi} \middle| X, Y_1^0(\Psi_0)\right] \Big\} \\
& - \left\{ \frac{\partial \log \frac{\partial h_1(\Psi_0)}{\partial y_1}}{\partial \Psi} + \frac{\partial \log f_{Y_1^0, X}(y_1^0(\Psi_0), x)}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \right. \\
& \left. + \frac{\partial \log f_{A|X, Y_1^0}(a|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \frac{\partial h_1(\Psi_0)}{\partial \Psi} \right\}
\end{aligned}$$

After a lot of cancelations among the terms, the final expression for S_{eff} is that

$$\begin{aligned}
& S_{eff} \tag{4.1.5} \\
& = \left\{ \frac{\partial \log \frac{\partial h_2(\Psi_0)}{\partial y_2}}{\partial \Psi} - E\left[\frac{\partial \log \frac{\partial h_2(\Psi_0)}{\partial y_2}}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \right\} \\
& \quad + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \left(\frac{\partial h_1(\Psi_0)}{\partial \Psi} - E\left[\frac{\partial h_1(\Psi_0)}{\partial \Psi} \middle| X, Y_1^0(\Psi_0)\right] \right) \\
& \quad + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} \left(\frac{\partial h_2(\Psi_0)}{\partial \Psi} - E\left[\frac{\partial h_2(\Psi_0)}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right] \right)
\end{aligned}$$

For the convenience of later discussion, we also consider writing S_{eff} as

$$\begin{aligned}
& S_{eff} \tag{4.1.6} \\
& = \frac{\partial \log f_{Y_2|X, Y_1^0}(y_2|x, y_1^0(\Psi_0))}{\partial \Psi} - E\left(\frac{\partial \log f_{Y_2|X, Y_1^0}(y_2|x, y_1^0(\Psi_0))}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)\right)
\end{aligned}$$

4.2 Locally Efficient Doubly Robust RAL Estimator

4.2.1 The Estimator that Achieves Semi-parametric Efficiency Bound

Given the efficient score function (4.1.5) we derived in the previous section, we can define

$$\begin{aligned}
& S_{eff,i}(\Psi) \\
& = \left\{ \frac{\partial \log \frac{\partial h_2(\Psi)}{\partial y_2}}{\partial \Psi} - E\left[\frac{\partial \log \frac{\partial h_2(\Psi)}{\partial y_2}}{\partial \Psi} \middle| Y_2^0(\Psi), X, Y_1^0(\Psi)\right] \right\}
\end{aligned}$$

$$\begin{aligned}
& + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi)|x, y_1^0(\Psi))}{\partial y_1^0(\Psi)} \left(\frac{\partial h_1(\Psi)}{\partial \Psi} - E\left[\frac{\partial h_1(\Psi)}{\partial \Psi} | X, Y_1^0(\Psi)\right] \right) \\
& + \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi)|x, y_1^0(\Psi))}{\partial y_2^0(\Psi)} \left(\frac{\partial h_2(\Psi)}{\partial \Psi} - E\left[\frac{\partial h_2(\Psi)}{\partial \Psi} | Y_2^0(\Psi), X, Y_1^0(\Psi)\right] \right)
\end{aligned}$$

where i indicates each subject.

We solve for $\hat{\Psi}$ such that $\sum_i S_{eff,i}(\hat{\Psi}) = 0$ and $\hat{\Psi}$ will be the RAL estimator that achieves the semi-parametric efficiency bound, under the sole condition that (4.1.1) is true. The semi-parametric efficiency bound is given by $E[S_{eff}(S_{eff})^T]^{-1}$. However, this estimating equation requires that we know the true density function $f_{Y_2^0|X, Y_1^0}$ and that we can evaluate the conditional expectation involved in the estimating equation under the true distribution, which are usually unpractical.

4.2.2 Construction of a Locally Efficient Doubly Robust RAL Estimator

A more practical approach would be that we posit certain parametric model for $f_{Y_2^0|X, Y_1^0}$ and $f_{A|X, Y_1^0}$. Let $p_1(y_2^0|x, y_1^0; \alpha_1)$ be the working model for $f_{Y_2^0|X, Y_1^0}$, and let $p_2(a|x, y_1^0; \alpha_2)$ be the working model for $f_{A|X, Y_1^0}$. We assume that α_1 is of dimension m_1 and α_2 is of dimension m_2 . We can estimate α_1 and α_2 by maximizing the working conditional likelihood functions:

$$\begin{aligned}
\hat{\alpha}_1 &= \operatorname{argmin} \prod_i p_1(y_2^0(\Psi^I)|x, y_1^0(\Psi^I); \alpha_1) \\
\hat{\alpha}_2 &= \operatorname{argmin} \prod_i p_2(a|x, y_1^0(\Psi^I); \alpha_2)
\end{aligned}$$

where i indicates different individuals and Ψ^I is some initial estimate of Ψ^I that is root- n consistent.

We then define

$$\begin{aligned}
U(\Psi) = & \left\{ \frac{\partial \log \frac{\partial h_2(\Psi)}{\partial y_2}}{\partial \Psi} - \int \frac{\partial \log \frac{\partial h_2(\Psi)}{\partial y_2}}{\partial \Psi} p_2(a|x, y_1^0(\Psi); \hat{\alpha}_2) da \right\} \\
& + \frac{\partial \log p_1(y_2^0(\Psi)|x, y_1^0(\Psi); \hat{\alpha}_1)}{\partial y_1^0(\Psi)} \left(\frac{\partial h_1(\Psi)}{\partial \Psi} - \int \frac{\partial h_1(\Psi)}{\partial \Psi} p_2(a|x, y_1^0(\Psi); \hat{\alpha}_2) da \right) \\
& + \frac{\partial \log p_1(y_2^0(\Psi)|x, y_1^0(\Psi); \hat{\alpha}_1)}{\partial y_2^0(\Psi)} \left(\frac{\partial h_2(\Psi)}{\partial \Psi} - \int \frac{\partial h_2(\Psi)}{\partial \Psi} p_2(a|x, y_1^0(\Psi); \hat{\alpha}_2) da \right)
\end{aligned} \tag{4.2.1}$$

We then consider estimator for Ψ obtained by solving

$$\sum_i U_i(\Psi) = 0 \tag{4.2.2}$$

where i indicates different subjects. We will show that this estimator is locally efficient and doubly robust.

4.2.3 Locally Efficiency

The estimator obtained from (4.2.2) is locally efficient in the sense that if the working models $p_1(y_2^0|x, y_1^0; \alpha_1)$ and $p_2(a|x, y_1^0; \alpha_2)$ are the correct models for $f_{Y_2^0|X, Y_1^0}$ and $f_{A|X, Y_1^0}$, the estimator will be an RAL estimator that has the influence function proportional to the efficient score (4.1.5). No other estimators will have a smaller asymptotic variance under the sole restriction that (4.1.1) is true.

Before proving the result, we show the following useful lemma.

Lemma 4.2.1. *Assume that Z_1, Z_2, \dots, Z_n are i.i.d. with finite mean and variance and that $g(z, \theta)$ is smooth with respect to θ and that $\frac{\partial g(z, \theta)}{\partial \theta}$ is bounded. If $E[g(Z_i, \theta)] = 0$ for θ in a neighborhood of θ_0 and θ_n converges to θ_0 with \sqrt{n} rate in probability,*

where θ_n could be a function of Z_1, \dots, Z_n , then

$$\sum_{i=1}^n [g(Z_i, \theta_n) - g(Z_i, \theta_0)] = O_p(1).$$

Proof. By the assumption, with high probability, $|\theta_n - \theta_0| \leq \frac{C}{\sqrt{n}}$. Then with high probability,

$$\left| \sum_{i=1}^n [g(Z_i, \theta_n) - g(Z_i, \theta_0)] \right| \leq \sup_{|\theta' - \theta_0| \leq \frac{C}{\sqrt{n}}} \left| \sum_{i=1}^n [g(Z_i, \theta') - g(Z_i, \theta_0)] \right|$$

Consider the variance of $\sum_{i=1}^n [g(Z_i, \theta') - g(Z_i, \theta_0)]$. Since $E[g(Z_i, \theta')] = 0$ and $E[g(Z_i, \theta_0)] = 0$,

$$\begin{aligned} & \text{Var} \left\{ \sum_{i=1}^n [g(Z_i, \theta') - g(Z_i, \theta_0)] \right\} \\ &= \sum_{i=1}^n E \{ [g(Z_i, \theta') - g(Z_i, \theta_0)]^2 \} \\ &= n(\theta' - \theta_0)^2 E[g'(Z_i, \theta^*)] \\ &\leq C^* \end{aligned}$$

for $|\theta' - \theta_0| \leq \frac{C}{\sqrt{n}}$, where C^* is some positive constant.

This shows that with high probability the variance of $\sum_{i=1}^n [g(Z_i, \theta_n) - g(Z_i, \theta_0)]$ is also finite. Therefore

$$\sum_{i=1}^n [g(Z_i, \theta_n) - g(Z_i, \theta_0)] = O_p(1).$$

□

Lemma 4.2.2. *If $p_1(y_2^0|x, y_1^0; \alpha_{10})$ and $p_2(a|x, y_1^0; \alpha_{20})$ are the correct models for $f_{Y_2^0|X, Y_1^0}$ and $f_{A|X, Y_1^0}$, where α_{10} and α_{20} are the true parameter values, under certain regularity conditions on the smoothness of p_1 and p_2 with respect to α_1 and α_2*

respectively, the estimator $\hat{\Psi}$ by solving (4.2.2) have the influence function that is proportional to the efficient score (4.1.5).

Proof. If p_1 and p_2 are the true models, under mild regularity conditions, our maximum likelihood estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are \sqrt{n} consistent estimates for α_{10} and α_{20} .

Denote

$$p(y_2|x, y_1^0(\Psi); \alpha_1) = \frac{\partial h_2(\Psi)}{\partial y_2} p_1(y_2^0(\Psi)|x, y_1^0(\Psi); \alpha_1).$$

Each individual $U(\Psi_0)$ in (4.2.2) can be written as

$$U(\Psi_0) = \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} p_2(a|x, y_1^0; \hat{\alpha}_2) da \quad (4.2.3)$$

We add and subtract a few terms

$U(\Psi_0)$

$$= \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|x, y_1^0; \alpha_{20}) da \right\} \quad (4.2.4)$$

$$+ \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right. \quad (4.2.5)$$

$$\left. - \int \left(\frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) p_2(a|x, y_1^0; \alpha_{20}) da \right\} \\ - \left\{ \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} [p_2(a|x, y_1^0(\Psi_0); \hat{\alpha}_2) - p_2(a|x, y_1^0(\Psi_0); \alpha_{20})] da \right\} \quad (4.2.6)$$

$$- \left\{ \int \left(\frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) \right. \quad (4.2.7) \\ \left. \times [p_2(a|x, y_1^0(\Psi_0); \hat{\alpha}_2) - p_2(a|x, y_1^0(\Psi_0); \alpha_{20})] da \right\}$$

If p_1 and p_2 are the true models, The first term (4.2.4) is the efficient score function.

Consider the second term (4.2.5).

$$\begin{aligned}
& \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \\
& - \int \left(\frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) p_2(a|x, y_1^0; \alpha_{20}) da \\
& = \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} p_2(a|x, y_1^0(\Psi_0); \alpha_{20}) da \right\} \\
& - \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|x, y_1^0(\Psi_0); \alpha_{20}) da \right\}
\end{aligned}$$

Define the following g_1 function

$$\begin{aligned}
& g_1(Z, \alpha_1) \\
& = \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_1)}{\partial \Psi} - \int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_1)}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_{20}) da
\end{aligned}$$

where $Z = (A, X, Y_1^0, Y_2^0)$. If p_1 and p_2 are smooth enough, g will satisfy the smoothness condition in Lemma 4.2.1. We calculate

$$\begin{aligned}
E[g_1(Z, \alpha_1)] & = E \left\{ \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_1)}{\partial \Psi} \right. \\
& \quad \left. - E \left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_1)}{\partial \Psi} \mid X, Y_1^0(\Psi_0), Y_2^0(\Psi_0) \right] \right\} \\
& = 0
\end{aligned}$$

Note that the expectations are calculated under the true distribution. Referring to Lemma 4.2.1, we can show that

$$\begin{aligned}
O_p(1) & = \sum_i \left\{ \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right. \\
& \quad \left. - \int \left(\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) \right.
\end{aligned}$$

$$\times p_2(a|X, Y_1^0; \alpha_{20})da \}$$

Similarly, for (4.2.6), we define

$$g_2(Z, \alpha_2) = \int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_2) da.$$

Under regularity conditions, g_2 is smooth enough in α_2 . We compute

$$\begin{aligned} E[g_2(Z, \alpha_2)] &= E\left[\int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_2) da \right] \\ &= E\left\{ E\left[\int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_2) da \mid X, Y_1^0(\Psi_0) \right] \right\} \\ &= E\left\{ \int E\left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \mid X, Y_1^0(\Psi_0), A = a \right] \right. \\ &\quad \left. \times p_2(a|X, Y_1^0(\Psi_0); \alpha_2) da \right\} \\ &= 0 \end{aligned}$$

By Lemma 4.2.1,

$$\begin{aligned} &O_p(1) \\ &= \sum_i \left\{ \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} [p_2(a|x, y_1^0(\Psi_0); \hat{\alpha}_2) - p_2(a|x, y_1^0(\Psi_0); \alpha_{20})] da \right\} \end{aligned}$$

For (4.2.7),

$$\begin{aligned} &\left| \int \left(\frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) \right. \\ &\quad \left. \times [p_2(a|x, y_1^0(\Psi_0); \hat{\alpha}_2) - p_2(a|x, y_1^0(\Psi_0); \alpha_{20})] da \right| \\ &= |(\hat{\alpha}_1 - \alpha_{10})(\hat{\alpha}_2 - \alpha_{20}) \frac{\partial}{\partial \alpha_2} \int \frac{\partial^2 \log p(y_2|x, y_1^0(\Psi_0); \alpha_1^*)}{\partial \Psi \partial \alpha_1} p_2(a|x, y_1^0(\Psi_0); \alpha_2^*) da| \end{aligned}$$

Under certain smoothness regularity condition,

$$\left| \frac{\partial}{\partial \alpha_2} \int \frac{\partial^2 \log p(y_2|x, y_1^0(\Psi_0); \alpha_1^*)}{\partial \Psi \partial \alpha_1} p_2(a|x, y_1^0(\Psi_0); \alpha_2^*) da \right|$$

could be bounded by a constant C' . Then (4.2.7) could be bounded by $C'|(\hat{\alpha}_1 - \alpha_{10})(\hat{\alpha}_2 - \alpha_{20})|$. Considering that $|(\hat{\alpha}_1 - \alpha_{10})|$ is $O_p(\frac{1}{\sqrt{n}})$ and $|(\hat{\alpha}_2 - \alpha_{20})|$ is also $O_p(\frac{1}{\sqrt{n}})$, we get,

$$\begin{aligned} & \sum_i \left\{ \int \left(\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) \right. \\ & \quad \left. \times [p_2(a|X, Y_1^0(\Psi_0); \hat{\alpha}_2) - p_2(a|X, Y_1^0(\Psi_0); \alpha_{20})] da \right\} \\ & = n [O_p(\frac{1}{\sqrt{n}})]^2 \\ & = O_p(1) \end{aligned}$$

Combining all the results above, we have

$$\sum_i U_i(\Psi_0) = \sum_i S_{eff,i}(\Psi_0) + O_p(1)$$

Notice that $\sum_i S_{eff,i}$ is of order $O_p(\sqrt{n})$. Therefore, $\hat{\Psi}$ that solve $\sum_i U_i(\Psi) = 0$ will have the same influence function as the estimator that solves $\sum_i S_{eff,i}(\Psi) = 0$. $\hat{\Psi}$ achieves the semi-parametric bound under the sole condition of (4.1.1). \square

Hence, Lemma 4.2.2 proves that the estimator from (4.2.2) is locally efficient, in the sense that if both p_1 and p_2 are the correct model, the estimator achieves the semi-parametric efficiency bound. However, if p_1 and p_2 are both wrong the estimator may not be consistent. In the following section, we will show that the second best thing for the estimator is true, i.e., as long as one of p_1 and p_2 is the correct model, the estimator will be consistent.

4.2.4 Double Robustness

In this section, p_1 and p_2 may not be the true distribution, but we still assume that our estimated $\hat{\alpha}_1$ and $\hat{\alpha}_2$ converge to some α_{10} and α_{20} respectively with \sqrt{n} rate.

We do Taylor expansion for Ψ around Ψ_0 with $\sum_i U_i(\Psi) = 0$.

$$0 = \frac{1}{n} \sum_i \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} p_2(a|x, y_1^0; \hat{\alpha}_2) da \right\} \quad (4.2.8)$$

$$+ \left\{ \frac{1}{n} \sum_i \frac{\partial}{\partial \Psi} \left(\frac{\partial \log p(y_2|x, y_1^0(\Psi^*); \hat{\alpha}_1)}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi^*); \hat{\alpha}_1)}{\partial \Psi} p_2(a|x, y_1^0; \hat{\alpha}_2) da \right)^T \right\} \quad (4.2.9)$$

$$\times (\hat{\Psi} - \Psi_0)$$

where Ψ^* is between $\hat{\Psi}$ and Ψ_0 . Denote (4.2.9) to be B_n . Under certain regularity conditions, B_n will be nonsingular and $|B_n|^{-1}$ will be bounded.

Then,

$$\begin{aligned} & \hat{\Psi} - \Psi_0 \\ &= B_n^{-1} \frac{1}{n} \sum_i \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} p_2(a|x, y_1^0; \hat{\alpha}_2) da \right\} \end{aligned} \quad (4.2.10)$$

$$\begin{aligned} &= B_n^{-1} \times \\ & \left\{ \frac{1}{n} \sum_i \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|x, y_1^0; \alpha_{20}) da \right\} \right\} \end{aligned} \quad (4.2.11)$$

$$\begin{aligned}
& + \frac{1}{n} \sum_i \left\{ \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right. \\
& \quad \left. - \int \left(\frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) p_2(a|x, y_1^0; \alpha_{20}) da \right\} \\
\end{aligned} \tag{4.2.12}$$

$$\begin{aligned}
& - \frac{1}{n} \sum_i \left\{ \int \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right. \\
& \quad \left. \times [p_2(a|x, y_1^0(\Psi_0); \hat{\alpha}_2) - p_2(a|x, y_1^0(\Psi_0); \alpha_{20})] da \right\} \\
\end{aligned} \tag{4.2.13}$$

$$\begin{aligned}
& - \frac{1}{n} \sum_i \left\{ \int \left(\frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \hat{\alpha}_1)}{\partial \Psi} - \frac{\partial \log p(y_2|x, y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \right) \right. \\
& \quad \left. \times [p_2(a|x, y_1^0(\Psi_0); \hat{\alpha}_2) - p_2(a|x, y_1^0(\Psi_0); \alpha_{20})] da \right\} \\
\end{aligned} \tag{4.2.14}$$

Since $\hat{\alpha}_1 \rightarrow \alpha_{10}$ with \sqrt{n} rate, each term in (4.2.12) will be of order $\frac{1}{\sqrt{n}}$. The summation will be of order \sqrt{n} , and (4.2.12) itself is of order $\frac{1}{\sqrt{n}}$. Similarly, we can argue that (4.2.13) and (4.2.14) are of order $\frac{1}{\sqrt{n}}$. It is worth noting that if both p_1 and p_2 are true models, by the proof of Lemma 4.2.2, (4.2.12), (4.2.13) and (4.2.14) are of order $\frac{1}{n}$. When p_1 and p_2 are not the true models, this is not true. However, we only need them to be of order $\frac{1}{\sqrt{n}}$ for the purpose of this section.

(4.2.11) is the average of n i.i.d. random variables. By the law of large numbers, it will converge to its expectation with \sqrt{n} rate,

$$E \left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - \int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_{20}) da \right]. \tag{4.2.15}$$

If this expectation is 0, the right hand side of (4.2.10) is converging to 0 with \sqrt{n} rate. Thus, $\hat{\Psi}$ will be \sqrt{n} consistent.

It is easy to show that (4.2.15) is 0 if either p_1 or p_2 is the true model.

If p_1 is the correct model,

$$\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} = \frac{\partial \log p(Y_2|A, X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi}$$

because of ignorability (4.1.1). It is the true conditional score function conditional on $A, X, Y_1^0(\Psi_0)$. Thus,

$$\begin{aligned} & E\left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - \int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_{20}) da\right] \\ &= E\left\{E\left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - \int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_{20}) da \mid X, Y_1^0(\Psi_0)\right]\right\} \\ &= E\left\{E\left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \mid X, Y_1^0(\Psi_0)\right] - \int E\left[\frac{\partial \log p(Y_2|A = a, X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \mid A = a, X, Y_1^0(\Psi_0)\right] \right. \\ &\quad \left. \times p_2(a|X, Y_1^0(\Psi_0); \alpha_{20}) da\right\} \\ &= 0 \end{aligned}$$

If p_2 is the correct model

$$\begin{aligned} & E\left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - \int \frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} p_2(a|X, Y_1^0(\Psi_0); \alpha_{20}) da\right] \\ &= E\left[\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} - E\left(\frac{\partial \log p(Y_2|X, Y_1^0(\Psi_0); \alpha_{10})}{\partial \Psi} \mid X, Y_1^0(\Psi_0)\right)\right] \\ &= 0 \end{aligned}$$

Therefore, as long as one of p_1 and p_2 is the correct model, $\hat{\Psi}$ from solving (4.2.2) will be \sqrt{n} consistent. $\hat{\Psi}$ is doubly robust.

To summarize, we have constructed an RAL estimator that is locally efficient and doubly robust. It is worth noting that our locally efficiency is “weaker” than the locally efficiency in standard g-estimation. Our estimator achieves the semi-parametric efficiency bound under the sole condition that ignorability (4.1.1) is true. In standard g-estimation, the semi-parametric efficiency bound under the sole restriction of standard ignorability is the same as the semi-parametric efficiency bound under the restriction that both standard ignorability and the propensity score model are true (see Robins et al., 1992). However, this is not the case with our relaxed ignorability (4.1.1), as the propensity score model in our setting also involves Ψ .

4.3 Important Special Cases

4.3.1 When the Treatment is Binary

If the treatment is binary, the efficient score can be simplified.

$$\begin{aligned}
& S_{eff} \\
&= S_{\Psi_0} - E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0)] + E[S_{\Psi_0}|X, Y_1^0(\Psi_0)] - E[S_{\Psi_0}|A, X, Y_1^0(\Psi_0)] \\
&= A \times E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 1] \\
&\quad + (1 - A) \times E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 0] \\
&\quad - P(A = 1|X, Y_1^0(\Psi_0)) \times E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 1]
\end{aligned}$$

$$\begin{aligned}
& - (1 - P(A = 1|X, Y_1^0(\Psi_0))) \times E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 0] \\
& + P(A = 1|X, Y_1^0(\Psi_0)) \times E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 1] \\
& + (1 - P(A = 1|X, Y_1^0(\Psi_0))) \times E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 0] \\
& - A \times E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 1] - (1 - A) \times E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 0] \\
= & [A - P(A = 1|X, Y_1^0(\Psi_0))] \\
& \times \{E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 1] - E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 0]\} \\
& - [A - P(A = 1|X, Y_1^0(\Psi_0))] \\
& \times \{E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 1] - E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 0]\} \\
= & [A - P(A = 1|X, Y_1^0(\Psi_0))] \\
& \times \{E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 1] - E[S_{\Psi_0}|Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 0] \\
& \quad - E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 1] + E[S_{\Psi_0}|X, Y_1^0(\Psi_0), A = 0]\}
\end{aligned}$$

If we plug in the functional form of S_{Ψ_0} , the efficient score can be further simplified to

$$\begin{aligned}
S_{eff} = & [A - P(A = 1|X, Y_1^0(\Psi_0))] \\
& \times \left\{ E\left[\frac{\partial \log f_{Y_2|X, Y_1^0}(y_2|x, y_1^0(\Psi_0))}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 1 \right] \right. \\
& \left. - E\left[\frac{\partial \log f_{Y_2|X, Y_1^0}(y_2|x, y_1^0(\Psi_0))}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 0 \right] \right\}
\end{aligned}$$

Using the general theory in Section 4.2, we can consider the following estimating equation

$$\sum_i (A - P(A = 1|X, Y_1^0(\Psi); \hat{\alpha}))g(Y_2^0(\Psi), Y_1^0(\Psi), X) = 0$$

where $\hat{\alpha}$ can be estimated by maximizing conditional likelihood as in Section 4.2.2, and g is any function. The estimator will be consistent, as long as our propensity score model is correct. It is also locally efficient in the sense that if function g happens to be

$$E\left[\frac{\partial \log f_{Y_2|X, Y_1^0}(y_2|x, y_1^0(\Psi_0))}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 1\right] \\ - E\left[\frac{\partial \log f_{Y_2|X, Y_1^0}(y_2|x, y_1^0(\Psi_0))}{\partial \Psi} \middle| Y_2^0(\Psi_0), X, Y_1^0(\Psi_0), A = 0\right],$$

the influence function for the estimator will be proportional to the efficient score function.

4.3.2 Special Blip-down Functions

We now assume that $\frac{\partial h_2(\Psi_0)}{\partial y_2} = 1$, $\frac{\partial h_1(\Psi_0)}{\partial \Psi} = -A$, and $\frac{\partial h_2(\Psi_0)}{\partial \Psi} = -cA$, where c is a constant. Note that this implicitly assumes that Ψ is of one dimension. The efficient score becomes

$$S_{eff} = - (A - E[A|X, Y_1^0(\Psi_0)]) \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} \\ - c(A - E[A|X, Y_1^0(\Psi_0)]) \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} \\ = - (A - E[A|X, Y_1^0(\Psi_0)]) \\ \times \left[\frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} + c \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} \right]$$

Assuming we can correctly model $E[A|X, Y_1^0(\Psi_0)]$ as $E[A|X, Y_1^0(\Psi_0); \alpha_0]$, the following estimating equation can also be used for any g function,

$$\sum_i (A - E[A|X, Y_1^0(\Psi); \hat{\alpha}]) g(Y_2^0(\Psi), Y_1^0(\Psi), X) = 0.$$

where $\hat{\alpha}$ can be estimated using least square principle. Under mild regularity conditions, $\hat{\alpha}$ is \sqrt{n} consistent. The estimator for Ψ by solving the estimating equation will be \sqrt{n} consistent. It is also locally efficient in the sense that if function g happens to be

$$\frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} + c \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)},$$

the influence function for the estimator is the efficient score.

4.3.3 Identification Issue

The semi-parametric efficiency bound is $E[S_{eff}(S_{eff})^T]^{-1}$. Therefore, for Ψ to be identifiable, S_{eff} must not be a constant, i.e., 0.

This implies that

$$A - E[A|X, Y_1^0(\Psi_0)] \neq 0$$

and that

$$\frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} + c \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} \neq 0.$$

The former equation is a usual assumption. The latter equation might be violated if e.g.,

$$\log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0)) = C - \frac{(y_2^0(\Psi_0) - cy_1^0(\Psi_0) - \beta x)^2}{2\sigma}$$

i.e., assuming a normal regression model for $Y_2^0|X, Y_1^0$ and the regression coefficient for Y_1^0 is exactly c . This will induce

$$\frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_1^0(\Psi_0)} + c \frac{\partial \log f_{Y_2^0|X, Y_1^0}(y_2^0(\Psi_0)|x, y_1^0(\Psi_0))}{\partial y_2^0(\Psi_0)} = 0.$$

And hence Ψ will not be identifiable.

One example when the above case is true is that, assuming A is binary, $Y_2^0 = cY_1^0 + \beta X + \epsilon$, where ϵ is normal with zero mean, $Y_2^1 = Y_2^0 + c\Psi$, and $Y_1^1 = Y_1^0 + \Psi$. It is easy to see that $Y_2^1 = cY_1^1 + \beta X + \epsilon$. Therefore, knowing (Y_1^0, Y_1^1) is equivalent to knowing (Y_2^0, Y_2^1) . Y_2 does not contain more information about the causal effect Ψ , which makes Ψ un-identifiable.

However, if $Y_2^0 = cY_1^0 + \beta X + \epsilon$, $Y_2^1 = Y_2^0 + d\Psi$, $d \neq c$, and $Y_1^1 = Y_1^0 + \Psi$, Ψ will be identifiable. Now $Y_2^1 = cY_1^1 + (d - c)\Psi + \beta X + \epsilon$, and (Y_2^0, Y_2^1) contain additional information about the causal effect Ψ .

4.4 Extension to Multi-period Case

We consider a K -period study, assuming that

- A_0, A_1, \dots, A_{K-1} are the treatment levels at time $0, 1, \dots, K - 1$.
- Y_0, Y_1, \dots, Y_K are the actual outcomes at time $0, 1, \dots, K$.
- L_0, L_1, \dots, L_{K-1} are the covariates at time $0, 1, \dots, K - 1$. L_t does not include Y_t .
- $Y_k^{\bar{a}}$ is the counterfactual outcome at time k under treatment regime $\bar{a} = (a_0, a_1, \dots, a_{K-1})$.
- $Y_t^{k,0}$ is the counterfactual outcome at time $t \geq k + 1$, under treatment regime $\bar{a} = (A_0, A_1, \dots, A_k, 0, \dots, 0)$.

Note that consistency assumption requires that $Y_{k+1}^{k,0} = Y_{k+1}$.

We assume a rank preserving model:

$$Y_{k+2}^{k,0} = h_{k,k+2}(Y_{k+2}^{k+1,0}, Y_{k+1}^{k,0}, \bar{L}_k, \bar{Y}_k, \bar{A}_k; \Psi_0)$$

...

$$Y_m^{k,0} = h_{k,m}(Y_m^{k+1,0}, \bar{Y}_{(k+1):(m-1)}^{k,0}, \bar{L}_k, \bar{Y}_k, \bar{A}_k; \Psi_0), m > k + 1$$

where $\bar{L}_k = (L_0, \dots, L_k)$, $\bar{A}_k = (A_0, \dots, A_k)$, $\bar{Y}_{(k+1):(m-1)}^{k,0} = (Y_{k+1}^{k,0}, Y_{k+2}^{k,0}, \dots, Y_{m-1}^{k,0})$, and $h_{k,m}$'s are known rank preserving functions. We assume Ψ is the parameter of our interest and is a q dimensional vector. Ψ achieves its true value at Ψ_0 .

We next show that with this model, given $Y_m, \bar{L}_{m-1}, \bar{Y}_{m-1}, \bar{A}_{m-1}$ and $h_{k,m}$'s, we are able to blip Y_m down to $Y_m^{\bar{0}}$.

- With function $h_{k-2,k}$, $2 \leq k \leq m$ and Y_k , $1 \leq k \leq m$, we can get $Y_k^{k-2,0}$, $2 \leq k \leq m$.
- With function $h_{k-3,k}$, $3 \leq k \leq m$ and $Y_k^{k-2,0}$, $2 \leq k \leq m$, we can get $Y_k^{k-3,0}$, $3 \leq k \leq m$.
- ...
- With function $h_{0,m-1}$ and $h_{1,m}$ and the counterfactuals above, we can get $Y_{m-1}^{\bar{0}}$ and $Y_m^{1,0}$.
- With function $h_{0,m}$ and the counterfactuals we obtained above, we can get $Y_m^{\bar{0}}$.

We denote

$$Y_m^{\bar{0}} = h_m(Y_m, \bar{L}_{m-1}, \bar{Y}_{m-1}, \bar{A}_{m-1}; \Psi)$$

We also assume the following ignorability assumption

$$Y_m^{k,0} \perp\!\!\!\perp A_k | \bar{L}_k, \bar{Y}_k, \bar{A}_{k-1}, Y_{k+1}^{k-1,0}, m > k + 1 \quad (4.4.1)$$

With the rank preserving model described above, the ignorability assumption is equivalent to

$$Y_m^{\bar{0}} \perp\!\!\!\perp A_k | \bar{L}_k, \bar{Y}_k, \bar{A}_{k-1}, Y_{k+1}^{\bar{0}}, m > k + 1 \quad (4.4.2)$$

4.4.1 Likelihood Function

We consider the likelihood function of \bar{A}_{K-1} , \bar{L}_{K-1} and \bar{Y}_K , for one subject. We note the convention that $\bar{A}_{-1} = \phi$ and $\bar{L}_{-1} = \phi$.

$$\begin{aligned} & f_{\bar{A}_{K-1}, \bar{L}_{K-1}, \bar{Y}_K}(\bar{a}_{K-1}, \bar{l}_{K-1}, \bar{y}_K) \\ &= f_{\bar{A}_{K-1}, \bar{L}_{K-1}, \bar{Y}_K^{\bar{0}}}(\bar{a}_{K-1}, \bar{l}_{K-1}, \bar{y}_K^{\bar{0}}(\Psi)) \prod_{1 \leq m \leq K} \frac{\partial h_m}{\partial y_m} \\ &= \left(\prod_{1 \leq m \leq K} \frac{\partial h_m}{\partial y_m} \right) \times f_{\bar{Y}_K^{\bar{0}}}(\bar{y}_K^{\bar{0}}(\Psi)) \\ & \quad \times \prod_{k=0}^{K-1} f_{L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}}(l_k | \bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_K^{\bar{0}}(\Psi)) f_{A_k | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}}(a_k | \bar{l}_k, \bar{a}_{k-1}, \bar{y}_K^{\bar{0}}(\Psi)) \\ &= \left(\prod_{1 \leq m \leq K} \frac{\partial h_m}{\partial y_m} \right) \times f_{\bar{Y}_K^{\bar{0}}}(\bar{y}_K^{\bar{0}}(\Psi)) \\ & \quad \times \prod_{k=0}^{K-1} f_{L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}}(l_k | \bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_K^{\bar{0}}(\Psi)) f_{A_k | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(a_k | \bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi)) \\ &= \left(\prod_{1 \leq m \leq K} \frac{\partial h_m}{\partial y_m}(\Psi) \right) \times f_{\bar{Y}_K^{\bar{0}}}(\bar{y}_K^{\bar{0}}(\Psi); \eta_y) \end{aligned}$$

$$\begin{aligned}
& \times \prod_{k=0}^{K-1} f_{L_k|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}}(\bar{l}_k|\bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_K^{\bar{0}}(\Psi); \eta_{lk}) \\
& \times f_{A_k|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(\bar{a}_k|\bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi); \eta_{ak})
\end{aligned}$$

The third equality is because of ignorability assumption and the rank preserving model. Note that in practice, under rank preserving model, we usually do the following parametrization

$$f_{A_0|L_0, \bar{Y}_1^{\bar{0}}}(a_0|l_0, \bar{y}_1^{\bar{0}}(\Psi); \eta_{a0}) = f_{A_0|L_0, Y_0, Y_1^{\bar{0}}}(a_0|l_0, y_0, y_1^{\bar{0}}(\Psi_0); \eta_{a0}^*)$$

and

$$f_{A_k|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(a_k|\bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi); \eta_{ak}) = f_{A_k|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_k, Y_{k+1}^{\bar{0}}}(a_k|\bar{l}_k, \bar{a}_{k-1}, \bar{y}_k, y_{k+1}^{\bar{0}}; \eta_{ak}^*).$$

Denote the true values of the parameters as $\Psi_0, \eta_{y0}, \eta_{lk0}$ and η_{ak0} .

4.4.2 Nuisance Tangent Space

Consider any parametric sub-model

$$\begin{aligned}
& f_{\bar{A}_{K-1}, \bar{L}_{K-1}, \bar{Y}_K}(\bar{a}_{K-1}, \bar{l}_{K-1}, \bar{y}_K) \\
& = \left(\prod_{1 \leq m \leq K} \frac{\partial h_m}{\partial y_m}(\Psi) \right) \times f_{\bar{Y}_K^{\bar{0}}}(\bar{y}_K^{\bar{0}}(\Psi); \gamma_y) \\
& \quad \times \prod_{k=1}^{K-1} f_{L_k|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}}(\bar{l}_k|\bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_K^{\bar{0}}(\Psi); \gamma_{lk}) \\
& \quad \times f_{A_k|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(\bar{a}_k|\bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi); \gamma_{ak})
\end{aligned}$$

where γ_y is of dimension n_y , γ_{lk} is of dimension n_{lk} , and γ_{ak} is of dimension n_{ak} .

The nuisance score of this parametric submodel is

$$S_\gamma = S(\bar{a}_{K-1}, \bar{l}_{K-1}, \bar{y}_K; \Psi_0, \gamma_0) = \frac{\partial \log f_{\bar{A}_{K-1}, \bar{L}_{K-1}, \bar{Y}_K}(\bar{a}_{K-1}, \bar{l}_{K-1}, \bar{y}_K; \Psi_0, \gamma_0)}{\partial \gamma} \Big|_{\gamma=\gamma_0}$$

Denote

$$\begin{aligned} S_{\gamma_y} &= \frac{\partial \log f_{\bar{Y}_K}(\bar{y}_K(\Psi); \gamma_y)}{\partial \gamma_y} \\ S_{\gamma_{lk}} &= \frac{\partial \log f_{L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K}(l_k | \bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_K(\Psi); \gamma_{lk})}{\partial \gamma_{lk}} \\ S_{\gamma_{ak}} &= \frac{\partial \log f_{A_k | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}}(a_k | \bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}(\Psi); \gamma_{ak})}{\partial \gamma_{ak}} \end{aligned}$$

Define

$$\begin{aligned} \Lambda_\gamma &= \{B_{q \times (n_y + \sum n_{ak} + \sum n_{lk})} S_\gamma\} \\ \Lambda_{\gamma_y} &= \{B_{q \times n_y} S_{\gamma_y}\} \\ \Lambda_{\gamma_{lk}} &= \{B_{q \times n_{lk}} S_{\gamma_{lk}}\} \\ \Lambda_{\gamma_{ak}} &= \{B_{q \times n_{ak}} S_{\gamma_{ak}}\} \end{aligned}$$

By definition, it is easy to see that

$$\Lambda_\gamma = \Lambda_{\gamma_y} \oplus \left\{ \bigoplus_{k=0}^{K-1} \Lambda_{\gamma_{lk}} \oplus \Lambda_{\gamma_{ak}} \right\}.$$

We define $\Lambda, \Lambda_y, \Lambda_{lk}, \Lambda_{ak}$ to be the mean-square closure of all $\Lambda_\gamma, \Lambda_{\gamma_y}, \Lambda_{\gamma_{lk}}, \Lambda_{\gamma_{ak}}$, respectively. It can also be proved that

$$\Lambda = \Lambda_y \oplus \left\{ \bigoplus_{k=0}^{K-1} \Lambda_{lk} \oplus \Lambda_{ak} \right\}.$$

Here Λ is the nuisance tangent space.

We also notice that $\Lambda_y, \Lambda_{lk}, \Lambda_{ak}$ can be characterized as follows:

Lemma 4.4.1.

$$\Lambda_y = \{a_y(\bar{Y}_K^{\bar{0}}(\Psi)) : E[a_y(\bar{Y}_K^{\bar{0}}(\Psi))] = 0\}$$

$$\Lambda_{lk} = \{a_{lk}(\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi)) : E[a_{lk}(\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi)) | \bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi)] = 0\}$$

$$\Lambda_{ak} = \{a_{ak}(\bar{A}_k, \bar{L}_k, \bar{Y}_{k+1}^{\bar{0}}(\Psi)) : E[a_{ak}(\bar{A}_k, \bar{L}_k, \bar{Y}_{k+1}^{\bar{0}}(\Psi)) | \bar{A}_{k-1}, \bar{L}_k, \bar{Y}_{k+1}^{\bar{0}}(\Psi)] = 0\}$$

The proof of this lemma is exactly the same as the proofs from Tsiatis (2006). It can also be proved that these subspaces are orthogonal to each other, using properties of conditional expectation and the ignorability assumption.

For example, to prove that $\Lambda_{lk_1} \perp \Lambda_{ak_2}$, $k_1 \leq k_2$, pick any $h_1 \in \Lambda_{lk_1}$ and $h_2 \in \Lambda_{ak_2}$.

$$\begin{aligned} E[h_1 h_2^T] &= E\{E[h_1 h_2^T | \bar{L}_{k_1}, \bar{A}_{k_1-1}, \bar{Y}_K^{\bar{0}}(\Psi)]\} \\ &= E\{h_1 E[h_2^T | \bar{L}_{k_1}, \bar{A}_{k_1-1}, \bar{Y}_K^{\bar{0}}(\Psi)]\} \\ &= E\{h_1 E[E\{h_2^T | \bar{L}_{k_2}, \bar{A}_{k_2-1}, \bar{Y}_K^{\bar{0}}(\Psi)\} | \bar{L}_{k_1}, \bar{A}_{k_1-1}, \bar{Y}_K^{\bar{0}}(\Psi)]\} \\ &= E\{h_1 E[E\{h_2^T | \bar{L}_{k_2}, \bar{A}_{k_2-1}, \bar{Y}_{k_2+1}^{\bar{0}}(\Psi)\} | \bar{L}_{k_1}, \bar{A}_{k_1-1}, \bar{Y}_K^{\bar{0}}(\Psi)]\} \\ &= E\{h_1 E[0_{1 \times q} | \bar{L}_{k_1}, \bar{A}_{k_1-1}, \bar{Y}_K^{\bar{0}}(\Psi)]\} \\ &= 0 \end{aligned}$$

The third equality is because $k_1 \leq k_2$. The fourth equality uses the ignorability assumption. Proofs for the orthogonality of other pairs follow the same fashion.

4.4.3 The Efficient Score

Consider the score function of interest

$$S_{\Psi_0} = \frac{\partial \log(f_{\bar{A}_{K-1}, \bar{L}_{K-1}, \bar{Y}_K})}{\partial \Psi}$$

The efficient score is

$$S_{eff} = S_{\Psi_0} - \prod(S_{\Psi_0} | \Lambda_y) - \sum_{k=0}^{K-1} [\prod(S_{\Psi_0} | \Lambda_{lk}) + \prod(S_{\Psi_0} | \Lambda_{ak})].$$

We then show the following facts.

Lemma 4.4.2.

$$\begin{aligned} \prod(S_{\Psi_0} | \Lambda_y) &= E[S_{\Psi_0} | \bar{Y}_K^0] \\ \prod(S_{\Psi_0} | \Lambda_{lk}) &= E[S_{\Psi_0} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^0] - E[S_{\Psi_0} | \bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^0] \\ \prod(S_{\Psi_0} | \Lambda_{ak}) &= E[S_{\Psi_0} | \bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^0] - E[S_{\Psi_0} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^0] \end{aligned}$$

Proof. For $\prod(S_{\Psi_0} | \Lambda_y) = E[S_{\Psi_0} | \bar{Y}_K^0]$, we can show that for any $h \in \Lambda_y$

$$\begin{aligned} &E[\{S_{\Psi_0} - E[S_{\Psi_0} | \bar{Y}_K^0]\}h^T] \\ &= E\{E[(S_{\Psi_0} - E[S_{\Psi_0} | \bar{Y}_K^0])h^T | \bar{Y}_K^0]\} \\ &= E\{E[(S_{\Psi_0} - E[S_{\Psi_0} | \bar{Y}_K^0]) | \bar{Y}_K^0]h^T\} \\ &= 0 \end{aligned}$$

For $\prod(S_{\Psi_0}|\Lambda_{lk}) = E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}] - E[S_{\Psi_0}|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]$, we can show that for any $h \in \Lambda_{lk}$,

$$\begin{aligned}
& E[\{S_{\Psi_0} - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}] + E[S_{\Psi_0}|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]\}h^T] \\
&= E\{E[(S_{\Psi_0} - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}] + E[S_{\Psi_0}|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}])h^T|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]\} \\
&= E\{E[(S_{\Psi_0} - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}] + E[S_{\Psi_0}|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}])|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]h^T\} \\
&= E\{E[E[S_{\Psi_0}|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]h^T\} \\
&= E\{E[S_{\Psi_0}|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]h^T\} \\
&= 0
\end{aligned}$$

For $\prod(S_{\Psi_0}|\Lambda_{ak}) = E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}] - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}]$, we can show that for any $h \in \Lambda_{ak}$

$$\begin{aligned}
& E[\{S_{\Psi_0} - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}] + E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}]\}h^T] \\
&= E\{E[(S_{\Psi_0} - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}] + E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}])h^T|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}]\} \\
&= E\{E[(S_{\Psi_0} - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}] + E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}])|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}]h^T\} \\
&= E\{E[E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}]|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}]h^T\} \\
&= E\{E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}]h^T\} \\
&= 0
\end{aligned}$$

□

By the previous lemma, we get

$$S_{eff} = S_{\Psi_0} - E[S_{\Psi_0}|\bar{Y}_K^{\bar{0}}]$$

$$\begin{aligned}
& - \sum_{k=1}^{K-1} \{E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}] - E[S_{\Psi_0}|\bar{L}_{k-1}, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}]\} \\
& - \sum_{k=1}^{K-1} \{E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}] - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}]\} \\
& = \sum_{k=0}^{K-1} \{E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}] - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}] \\
& \quad - E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}] + E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}]\}.
\end{aligned}$$

Let

$$\begin{aligned}
S_{eff}(\Psi) & = \sum_{k=0}^{K-1} \{E[S_{\Psi}|\bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi)] - E[S_{\Psi}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi)] \\
& \quad - E[S_{\Psi}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}(\Psi)] + E[S_{\Psi}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\Psi)]\}
\end{aligned}$$

The efficient estimate of Ψ is thus given by the solution from $\sum_i S_{eff}(\Psi) = 0$, where i indicates different individuals. This is the optimal estimating equation for the controlling-the-future method.

We would like to simplify the formula for S_{eff} . The following formulas are obtained using the similar trick of factorization of the likelihood as before.

$$\begin{aligned}
& E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0)] \\
& = E\left[\sum_{1 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \\
& \quad + E\left[\frac{\partial \log[f_{\bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}}(\bar{l}_k, \bar{a}_k, \bar{y}_K^{\bar{0}}(\Psi_0))]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \\
& = E\left[\sum_{1 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] + \frac{\partial \log[f_{\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}}(\bar{l}_k, \bar{a}_k, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi}
\end{aligned}$$

$$+ E\left[\frac{\partial \log[f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\{\bar{y}_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K|\bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))}}{\partial \Psi}]\right]_{\bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0)}$$

Similarly

$$\begin{aligned} & E[S_{\Psi_0}|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)] \\ &= E\left[\sum_{1 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi}\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)} \\ & \quad + E\left[\frac{\partial \log[f_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}}(\bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_K^{\bar{0}}(\Psi_0))]}{\partial \Psi}\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)} \\ &= E\left[\sum_{1 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi}\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)} \\ & \quad + E\left[\frac{\partial \log[f_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(\bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi}\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)} \\ & \quad + E\left[\frac{\partial \log[f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\{\bar{y}_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K|\bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))}}{\partial \Psi}]\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)} \\ &= E\left[\sum_{1 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi}\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)} \\ & \quad + E\left[\frac{\partial \log[f_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(\bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi}\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)} \\ & \quad + E\left[\frac{\partial \log[f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K|\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\{\bar{y}_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K|\bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))}}{\partial \Psi}]\right]_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)} \end{aligned}$$

$$\begin{aligned} & E[S_{\Psi_0}|\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)] \\ &= \sum_{1 \leq m \leq k+1} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} + \frac{\partial \log[f_{\bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}}(\bar{l}_k, \bar{a}_k, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi} \end{aligned}$$

and that

$$\begin{aligned}
& E[S_{\Psi_0} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)] \\
&= \sum_{1 \leq m \leq k+1} E\left[\frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)\right] \\
&\quad + E\left[\frac{\partial \log[f_{\bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(\bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)\right]
\end{aligned}$$

Given the above and with a lot of cancelation, we get

$$\begin{aligned}
& E[S_{\Psi_0} | \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0)] - E[S_{\Psi_0} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)] \\
&\quad - E[S_{\Psi_0} | \bar{L}_k, \bar{A}_k, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)] + E[S_{\Psi_0} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)] \\
&= E\left[\sum_{1 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0)\right] \\
&\quad - E\left[\sum_{1 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)\right] \\
&\quad - \sum_{1 \leq m \leq k+1} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} + \sum_{1 \leq m \leq k+1} E\left[\frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}(\Psi_0)\right] \\
&\quad + \left\{ E\left[\frac{\partial \log[f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(\{y_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K | \bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi} \right. \right. \\
&\quad \quad \left. \left. \middle| \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0)\right] \right. \\
&\quad \left. - E\left[\frac{\partial \log[f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}}(\{y_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K | \bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi} \right. \right. \\
&\quad \quad \left. \left. \middle| \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)\right] \right\} \\
&= E\left[\sum_{k+2 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0)\right] \\
&\quad - E\left[\sum_{k+2 \leq m \leq K} \frac{\partial \log[\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} \middle| \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0)\right]
\end{aligned}$$

$$\begin{aligned}
& + \left\{ E \left[\frac{\partial \log [f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}} (\{y_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K | \bar{l}_k, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi} \right. \right. \\
& \quad \left. \left. | \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \right. \\
& \quad \left. - E \left[\frac{\partial \log [f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_{k+1}^{\bar{0}}} (\{y_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K | \bar{l}_{k-1}, \bar{a}_{k-1}, \bar{y}_{k+1}^{\bar{0}}(\Psi_0))]}{\partial \Psi} \right. \right. \\
& \quad \left. \left. | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \right\}
\end{aligned}$$

$$\equiv S_{eff,k}$$

Thus

$$\begin{aligned}
& S_{eff,k} \\
& = E \left[\sum_{k+2 \leq m \leq K} \frac{\partial \log [\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} | \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \\
& \quad - E \left[\sum_{k+2 \leq m \leq K} \frac{\partial \log [\frac{\partial h_m}{\partial y_m}(\Psi_0)]}{\partial \Psi} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \\
& \quad + \sum_{t=k+2}^K \left(\frac{\partial \log [f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K} | \bar{A}_{k-1}, \bar{L}_k, \bar{Y}_{k+1}^{\bar{0}}} (\{y_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K | \bar{a}_{k-1}, \bar{l}_k, \bar{y}_{k+1})]}{\partial Y_t^{\bar{0}}} \right. \\
& \quad \times \left. \left\{ E \left[\frac{\partial Y_t^{\bar{0}}(\Psi_0)}{\partial \Psi} | \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] - E \left[\frac{\partial Y_t^{\bar{0}}(\Psi_0)}{\partial \Psi} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \right\} \right) \\
& \quad + \left(\frac{\partial \log [f_{\{Y_t^{\bar{0}}\}_{t=k+2}^K} | \bar{A}_{k-1}, \bar{L}_k, \bar{Y}_{k+1}^{\bar{0}}} (\{y_t^{\bar{0}}(\Psi_0)\}_{t=k+2}^K | \bar{a}_{k-1}, \bar{l}_k, \bar{y}_{k+1})]}{\partial Y_{k+1}^{\bar{0}}} \right. \\
& \quad \times \left. \left\{ E \left[\frac{\partial Y_{k+1}^{\bar{0}}(\Psi_0)}{\partial \Psi} | \bar{L}_k, \bar{A}_k, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] - E \left[\frac{\partial Y_{k+1}^{\bar{0}}(\Psi_0)}{\partial \Psi} | \bar{L}_k, \bar{A}_{k-1}, \bar{Y}_K^{\bar{0}}(\Psi_0) \right] \right\} \right)
\end{aligned}$$

We get an efficient score that is a natural extension of the efficient score in the single period case. Locally efficient estimators that are doubly robust can be constructed in a similar way as in Section 4.2, which we omit here.

4.5 Conclusion

In this chapter, we have developed the semi-parametric theory for the relaxed ignorability assumption and the controlling-the-future method we used in Chapter 2. In particular, we have characterized the nuisance tangent space under the sole restriction that the relaxed ignorability (4.1.1) is true, and calculated the efficient score and the semi-parametric efficiency bound. Motivated by the form of the efficient score function, we propose a locally efficient and doubly robust estimator.

The multi-period generalization of the theory is straightforward. It is worth noting that even though we have only considered ignorability assumption that only allows treatment assignment depend on the next period potential outcome given the historical treatment and covariates, the formulas and the derivation are almost the same for extended assumptions that allows treatment assignment to depend on more than one period of future potential outcomes (see extended formulations in Joffe and Robins (2009)). Similar locally efficient and doubly robust estimator can be constructed in the same fashion as in the single period model.

We admit that the discussion in this chapter is incomplete. In this chapter, we have required that there are more than one outcomes associated with the treatment, and our discussion of the multi-period case has focused on studies with repeated measurements of the outcomes. In the extended formulation of Joffe and Robins (2009), it is possible to extend the ideas of the basic controlling-the-future method to cases when we only have a single measurement of outcome and the treatment could

depend on the potential outcome in known functional form, possibly parametrized by a finite dimensional parameter. This extension does not always lead to identification. Work is in progress studying when identification can be achieved and how to construct useful estimator with good properties.

Chapter 5

Appendices

5.1 Estimating Covariance Matrix of Estimated Parameters

The formulas in this section can be used to estimate the covariance matrix of the estimated parameters from naive g-estimation of Section 2.2.1, modified g-estimation of Section 2.2.3, and the controlling-the-future estimation of Section 2.4.1.

We denote $\theta = (\Psi, \beta)$. In Section 2.2.1 and Section 2.2.3, β is the parameter in the propensity score model. In Section 2.4.1, $\beta = (\beta_X, \beta_h)$ is the parameter in the propensity score model. Let $U(\theta)$ be the vector of the left hand side of the estimating equations (Equation (2.2.5) in Section 2.2.1, Equation (2.2.7) in Section 2.2.3, and Equation (2.4.4) in Section 2.4.1, respectively). We also denote

$$U_{i,k,m}(\theta) \equiv (A_{i,k}^* - p_{i,k}(\beta))[g(Y_{i,m}^{0*}(\Psi), X_{i,k}^*), X_{i,k}^*]^T,$$

for the naive g-estimation and the modified g-estimation, and denote

$$U_{i,k,m}(\Psi; \beta_X, \beta_h) \equiv (A_{i,k}^* - p_{i,k}(\Psi; \beta_X, \beta_h))[g(Y_{i,m}^{0*}(\Psi), X_i^*, h_i), X_{i,k}^*, h_{i,k}]^T,$$

for the controlling-the-future estimation. Then $U(\theta) = \sum U_{i,k,m}$.

Let $B(\theta) = E[\frac{\partial U(\theta)}{\partial \theta}]$, which can be estimated as

$$\hat{B}(\theta) = - \sum_{i,k,m} \left\{ \frac{\partial U_{i,k,m}}{\partial \theta} \right\} \Big|_{\theta=\hat{\theta}}$$

where $\hat{\theta}$ is the solution from the corresponding estimating equations, and $k < m$ in both g-estimations and $k < m - 1$ in controlling-the-future estimation. Then the covariance matrix of the estimator $\hat{\theta}$ can be estimated as

$$Cov(\hat{\theta}) = \hat{B}^{-1}(\theta) Cov[\hat{U}(\theta)] \hat{B}^{-1}(\theta)'$$

by the Delta-method, where $Cov[U(\theta)]$ is estimated by

$$Cov[\hat{U}(\theta)] = \sum_i U_i(\hat{\theta}) U_i(\hat{\theta})'$$

where $U_i = \sum_{k,m} U_{i,k,m}(\hat{\theta})$, $k < m$ in both g-estimations and $k < m - 1$ in controlling-the-future estimation.

5.2 Definition of N_t and Explicit Formula of λ_t

This section gives the definition of N_t , the counting process that counts the number of changes in the treatment process, and an explicit formula of λ_t , the intensity process of N_t with respect to the filtration of $\sigma(\bar{Z}_t)$. The definitions will be used in the proof of Theorem 2.3.3.

Following the standard definition of a counting process defined from a *càdlàg* process, we first define a sequence of stopping times:

$$\tau_1 = \inf_{t>0} \{t : A_t \neq A_0\}$$

$$\tau_2 = \inf_{t > \tau_1} \{t : A_t \neq A_{\tau_1}\}$$

$$\tau_3 = \inf_{t > \tau_2} \{t : A_t \neq A_{\tau_2}\}$$

...

Then, N_t can be defined as

$$\{N_t = k\} = \{\tau_k \leq t < \tau_{k+1}\} \quad (5.2.1)$$

Assume A_t is a binary *càdlàg* process such that N_t defined in (5.2.1) is a counting process on $[0, K]$, satisfying

- N_t is a non-negative integer.
- $N_s \leq N_t$ for $s \leq t$.
- $dN_t = N_t - N_{t-}$ is either 0 or 1.
- $E[N_t] < \infty$.

Given N_t , λ_t is the intensity process with respect to $\sigma(\bar{Z}_t)$. We will give a formula for λ_t in terms of A_t , which can then be explicitly related to the propensity score of A_t . First, we denote

$$r_t(\delta) = (1 - A_{t-})A_{t+\delta} + A_{t-}(1 - A_{t+\delta}).$$

Then define

$$\lambda_t \equiv \lim_{\delta \downarrow 0} \frac{E[r_t(\delta) | \sigma(\bar{Z}_{t-})]}{\delta} \quad (5.2.2)$$

As a regularity condition, we assume that the limit on the right hand side of (5.2.2) always exists and is finite. The following lemma shows that the so-defined λ_t is the intensity process for N_t .

Lemma 5.2.1. λ_t defined in Equation (5.2.2) is the intensity process for counting process N_t , w.r.t. $\sigma(\bar{Z}_t)$. In other words,

$$\lim_{\delta \downarrow 0} \frac{Pr(N_{t+\delta} - N_{t-} = 1 | \sigma(\bar{Z}_{t-}))}{\delta} = \lambda_t$$

Proof. The proof is simple.

$$\begin{aligned} & Pr(N_{t+\delta} - N_{t-} = 1 | \sigma(\bar{Z}_{t-})) \\ &= Pr(\tau_{N_{t-}+1} \leq t + \delta < \tau_{N_{t-}+2} | \sigma(\bar{Z}_{t-})) \\ &= Pr(A_{t+\delta} \neq A_{t-} | \sigma(\bar{Z}_{t-})) - Pr(A_{t+\delta} \neq A_{t-}, N_{t+\delta} - N_{t-} \geq 2 | \sigma(\bar{Z}_{t-})) \\ &= Pr(A_{t+\delta} \neq A_{t-} | \sigma(\bar{Z}_{t-})) - O(\delta^2) \\ &= E[r_t(\delta) | \sigma(\bar{Z}_{t-})] - O(\delta^2) \end{aligned}$$

Divide both sides by δ and take the limit with $\delta \downarrow 0$, we can get the desired result. □

5.3 Proof of FTSR Implying CTSR

We assume that Z_t is a *càdlàg* process, and everything we discuss is in an *a.s.* sense.

By the definition of continuous time sequential randomization in Definition 2.2.1, we only need to prove that λ_t defined in the Appendix B is also the intensity process for N_t with respect to the filtration of $\sigma(\bar{Z}_t, \underline{Y}_{t+}^0)$.

First, we denote

$$\mathcal{H}_{t-} \equiv \sigma(\bar{Z}_{t-})$$

$$\mathcal{F}_{t-,t+} \equiv \sigma(\bar{Z}_{t-}, \underline{Y}_{t+}^0)$$

We also define

$$\eta_t \equiv \lim_{\delta \downarrow 0} \frac{E[r_t(\delta) | \mathcal{F}_{t-,t+}]}{\delta}.$$

As a regularity condition, we assume that η_t exists and is finite. By a similar proof as in Lemma 5.2.1, η_t is the intensity process of N_t with respect to the filtration $\mathcal{F}_{t-,t+}$. Therefore, proving Theorem 2.3.3 is equivalent to proving that $\lambda_t = \eta_t$.

To bridge our intuition in the discrete time case into the continuous time case, we assume the following regularity conditions:

1. We assume that η_t defined above always exists and is positive random functions.

We also assume that η_t is bounded by some constant that is independent of t .

(Note that by the Dominated Convergence Theorem of conditional expectation, λ_t also exists and is positive, and can be bounded by the same constant.)

2. We assume that for any finite sequence of time points, $t_1 \leq t_2 \leq t_3 \leq \dots \leq t_n$, the density $pr(Z_{t_1} = z_1, Z_{t_2} = z_2, \dots, Z_{t_n} = z_n)$ is well-defined, and is locally uniformly bounded, i.e. there exists a constant D and a rectangle $B \equiv [t_1 - \delta_1, t_1 + \delta_1] \times [t_2 - \delta_2, t_2 + \delta_2] \times \dots \times [t_n - \delta_n, t_n + \delta_n]$, for any $(t'_1, t'_2, \dots, t'_n)^T \in B$ and for any possible value of $(z_1, z_2, \dots, z_n)^T$,

$$pr(Z_{t'_1} = z_1, Z_{t'_2} = z_2, \dots, Z_{t'_n} = z_n) \leq D$$

For any conditional expectation involving finite sequence of time points, we choose the version that is defined by the joint density.

3. Given any finite sequence of time points, $t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_n$ and any possible value of $(z_1, z_2, \cdots, z_n)^T$, we assume that the following convergence is uniform in a closed neighborhood of $\tilde{t} \equiv (t_1, t_2, t_3, \cdots, t_n)$

$$\begin{aligned} & pr(Z_{t'_1} = z_1, Z_{t'_2} = z_2, \cdots, Z_{t'_n} = z_n) \\ &= \lim_{\Delta \downarrow 0} \frac{Pr(Z_{t'_1} \in [z_1, z_1 + \Delta_1], Z_{t'_2} \in [z_2, z_2 + \Delta_2], \cdots, Z_{t'_n} \in [z_n, z_n + \Delta_n])}{\Delta_1 \times \Delta_2 \times \cdots \times \Delta_n} \end{aligned}$$

where $(t'_1, t'_2, \cdots, t'_n)^T$ is in a neighborhood of \tilde{t} .

4. Given any finite sequence of time points, $t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_i \leq \cdots \leq t_n$ and any possible value of $(z_1, z_2, \cdots, z_n)^T$, we define

$$f(\delta) = \frac{pr(A_{t_i+\delta} \neq A_{t_i} | Z_{t_1} = z_1, Z_{t_2} = z_2, \cdots, Z_{t_n} = z_n)}{\delta}.$$

We assume that $\lim_{\delta \downarrow 0} f(\delta)$ exists and is positive and finite. We also assume that $f(\delta)$ is finite and is right-continuous in δ , and the continuity is uniform with respect (δ, t_i) in $[0, \delta_0] \times B(t_i)$, where $B(t_i)$ is a closed neighborhood of t_i . Further, we assume that the above assumption is true if any of the Z in f is in its left-limit value rather than the concurrent value.

Remark 5.3.1. The third regularity condition is needed when we want to prove convergence in density. For example, consider that when $\delta \downarrow 0$, we have $Z_{t_2+\delta} \rightarrow Z_{t_2}$.

Then, we can see that

$$\begin{aligned}
& \lim_{\delta \downarrow 0} pr(Z_{t_1} = z_1, Z_{t_2+\delta} = z_2, Z_{t_3} = z_3) \\
&= \lim_{\delta \downarrow 0} \lim_{\substack{\Delta_1 \downarrow 0 \\ \Delta_2 \downarrow 0 \\ \Delta_3 \downarrow 0}} \frac{Pr(Z_{t_1} \in [z_1, z_1 + \Delta_1], Z_{t_2+\delta} \in [z_2, z_2 + \Delta_2], Z_{t_3} \in [z_3, z_3 + \Delta_3])}{\Delta_1 \Delta_2 \Delta_3} \\
&= \lim_{\substack{\Delta_1 \downarrow 0 \\ \Delta_2 \downarrow 0 \\ \Delta_3 \downarrow 0}} \lim_{\delta \downarrow 0} \frac{Pr(Z_{t_1} \in [z_1, z_1 + \Delta_1], Z_{t_2+\delta} \in [z_2, z_2 + \Delta_2], Z_{t_3} \in [z_3, z_3 + \Delta_3])}{\Delta_1 \Delta_2 \Delta_3} \\
&= \lim_{\substack{\Delta_1 \downarrow 0 \\ \Delta_2 \downarrow 0 \\ \Delta_3 \downarrow 0}} \frac{Pr(Z_{t_1} \in [z_1, z_1 + \Delta_1], Z_{t_2} \in [z_2, z_2 + \Delta_2], Z_{t_3} \in [z_3, z_3 + \Delta_3])}{\Delta_1 \Delta_2 \Delta_3} \\
&= pr(Z_{t_1} = z_1, Z_{t_2} = z_2, Z_{t_3} = z_3)
\end{aligned}$$

The validity of interchanging the limits at second equality is because of the third regularity condition. The third equality comes from the fact that probabilities are expectations of indicator functions and that dominated convergence theorem applies.

Next, we notice the following lemma:

Lemma 5.3.2.

$$E[\eta_t | \mathcal{H}_{t-}] = \lambda_t$$

Proof.

$$\begin{aligned}
E[\eta_t | \mathcal{H}_{t-}] &= E\left[\lim_{\delta \downarrow 0} \frac{E[r_t(\delta) | \mathcal{F}_{t-, t+}]}{\delta} \middle| \mathcal{H}_{t-}\right] \\
&= \lim_{\delta \downarrow 0} \frac{E[r_t(\delta) | \mathcal{H}_{t-}]}{\delta} \\
&= \lambda_t
\end{aligned}$$

The second equality is because of Dominant Convergence Theorem and Tower Property of conditional expectation (see Rogers and Williams 1994, p139-140), since $\mathcal{H}_{t-} \subset \mathcal{F}_{t-,t+}$. \square

Remark 5.3.3. If η_t is also \mathcal{H}_{t-} -measurable, we will have

$$\lambda_t = E[\eta_t | \mathcal{H}_{t-}] = \eta_t$$

Therefore, the main step to prove Theorem 2.3.3 is to prove that η_t is \mathcal{H}_{t-} -measurable, when finite time sequential randomization is true.

Before proving η_t is \mathcal{H}_{t-} -measurable, we need two more lemmas.

Lemma 5.3.4. *If the càdlàg process Z_t follows the **finite time sequential randomization** as defined in Definition 2.3.2, then the following version of **FTSR** is also true,*

$$\begin{aligned} & pr(A_{t_n} | \bar{L}_{t_{n-1}}, L_{t_n-}, \bar{A}_{t_{n-1}}, \bar{Y}_{t_{n-1}}^0, Y_{t_n-}^0, \underline{Y}_{t_n+}^0) \\ &= pr(A_{t_n} | \bar{L}_{t_{n-1}}, L_{t_n-}, \bar{A}_{t_{n-1}}, \bar{Y}_{t_{n-1}}^0, Y_{t_n-}^0) \end{aligned} \quad (5.3.1)$$

where $\bar{L}_{t_{n-1}} = (L_{t_1}, L_{t_2}, \dots, L_{t_{n-1}})$, $\bar{A}_{t_{n-1}} = (A_{t_1}, A_{t_2}, \dots, A_{t_{n-1}})$, $\bar{Y}_{t_{n-1}}^0 = (Y_{t_1}^0, Y_{t_2}^0, \dots, Y_{t_{n-1}}^0)$, and $\underline{Y}_{t_n+}^0 = (Y_{t_{n+1}}^0, Y_{t_{n+2}}^0, \dots, Y_{t_{n+l}}^0)$.

Remark 5.3.5. The difference between (5.3.1) and the original definition of FTSR is that in (5.3.1) most L 's and Y^0 's are stated in their concurrent values, while in Definition 2.3.2, they are all stated in their left limits. Lemma 5.3.4 is only for technical convenience.

Proof. Without loss of generality, we only need to prove that if we have

$$pr(A_{t_2}|L_{t_1-}, L_{t_2-}, A_{t_1}, Y_{t_1-}^0, Y_{t_2-}^0, Y_{t_3-}^0) = pr(A_{t_2}|L_{t_1-}, L_{t_2-}, A_{t_1}, Y_{t_1-}^0, Y_{t_2-}^0)$$

for any $t_1 < t_2 < t_3$, we will have

$$pr(A_{t_2}|L_{t_1}, L_{t_2-}, A_{t_1}, Y_{t_1}^0, Y_{t_2-}^0, Y_{t_3}^0) = pr(A_{t_2}|L_{t_1}, L_{t_2-}, A_{t_1}, Y_{t_1}^0, Y_{t_2-}^0).$$

Or equivalently, if we have

$$\begin{aligned} & pr(A_{t_2}, L_{t_1-}, L_{t_2-}, A_{t_1}, Y_{t_1-}^0, Y_{t_2-}^0, Y_{t_3-}^0)pr(L_{t_1-}, L_{t_2-}, A_{t_1}, Y_{t_1-}^0, Y_{t_2-}^0) \quad (5.3.2) \\ & = pr(A_{t_2}, L_{t_1-}, L_{t_2-}, A_{t_1}, Y_{t_1-}^0, Y_{t_2-}^0)pr(L_{t_1-}, L_{t_2-}, A_{t_1}, Y_{t_1-}^0, Y_{t_2-}^0, Y_{t_3-}^0) \end{aligned}$$

for any $t_1 < t_2 < t_3$, we need to prove that

$$\begin{aligned} & pr(A_{t_2}, L_{t_1}, L_{t_2-}, A_{t_1}, Y_{t_1}^0, Y_{t_2-}^0, Y_{t_3}^0)pr(L_{t_1}, L_{t_2-}, A_{t_1}, Y_{t_1}^0, Y_{t_2-}^0) \quad (5.3.3) \\ & = pr(A_{t_2}, L_{t_1}, L_{t_2-}, A_{t_1}, Y_{t_1}^0, Y_{t_2-}^0)pr(L_{t_1}, L_{t_2-}, A_{t_1}, Y_{t_1}^0, Y_{t_2-}^0, Y_{t_3}^0) \end{aligned}$$

Since (5.3.2) is true for any triple of $t_1 < t_2 < t_3$, we hope to find a sequence of $t_{1,k} \rightarrow t_1$ and $t_{3,k} \rightarrow t_3$, such that $L_{t_{1,k}-} \rightarrow L_{t_1}$, $Y_{t_{1,k}-}^0 \rightarrow Y_{t_1}^0$ and $Y_{t_{3,k}-}^0 \rightarrow Y_{t_3}^0$.

Considering L_{t_1} for example, since L_t is a *càdlàg* process, we choose any $t_{1,k} \downarrow t_1$.

We then choose $s_{1,k} \in (t_1, t_{1,k})$, such that $|L_{s_{1,k}} - L_{t_{1,k}-}| < \frac{1}{k}$. Notice that

$$|L_{t_1} - L_{t_{1,k}-}| \leq |L_{t_1} - L_{s_{1,k}}| + |L_{s_{1,k}} - L_{t_{1,k}-}|$$

Let $k \rightarrow \infty$, since L is right continuous, the first term on the right hand side converges to zero, and the second term is controlled by $\frac{1}{k}$. Therefore, $L_{t_{1,k}-}$ converges to L_{t_1} , a.s.. (Note that the proof is for a point-wise convergence. $s_{1,k}$ may be a random

function of ω , but $t_{1,k}$ is a deterministic sequence.) Similar proofs holds for $Y_{t_1}^0$ and $Y_{t_3}^0$.

By definition, (5.3.2) holds for every set of $(t_{1,k}, t_2, t_{3,k})$. Therefore, we can take the limit to both sides of (5.3.2) when $k \rightarrow \infty$. Using a similar argument as in Remark 5.3.1, we can take limit inside the density function and thus prove that (5.3.3) is true. \square

Lemma 5.3.6. *Suppose FTSSR is true. If we define*

$$\begin{aligned}\mathcal{F} &= \sigma(Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t-}, Y_{t_{n+1}}^0, \dots, Y_{t_{n+l}}^0) \\ \mathcal{H} &= \sigma(Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t-})\end{aligned}$$

we have

$$\lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{F}]}{\delta} = \lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{H}]}{\delta} \quad (5.3.4)$$

Proof. First, we notice that the limits on both sides of equation (5.3.4) exist and finite. This fact follows from the regularity condition 1 that η_t exist and is finite.

Take $\lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{F}]}{\delta}$ for example.

$$\begin{aligned}& \lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{F}]}{\delta} \\ &= \lim_{\delta \downarrow 0} \frac{E[E[r_t(\delta)|\sigma(\bar{Z}_{t-}, \underline{Y}_t^0)]|\mathcal{F}]}{\delta} \\ &= E[\lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\sigma(\bar{Z}_{t-}, \underline{Y}_t^0)]}{\delta} | \mathcal{F}] \\ &= E[\eta_t | \mathcal{F}]\end{aligned}$$

The existence is guaranteed by the dominated convergence theorem, and $E[\eta_t | \mathcal{F}]$ is obviously finite.

Given equation (2.3.3) and Lemma 5.3.4, we always have

$$E[I_{A_t \neq A_{t_n}} | \bar{L}_{t-}, \bar{A}_{t_n}, \bar{Y}_{t-}^0, \underline{Y}_{t+}^0] = E[I_{A_t \neq A_{t_n}} | \bar{L}_{t-}, \bar{A}_{t_n}, \bar{Y}_{t-}^0] \quad (5.3.5)$$

where $\bar{L}_{t-} = (L_{t_1}, L_{t_2}, \dots, L_{t_{n-1}}, L_{t_n}, L_{t-})^T$, $\bar{A}_{t_n} = (A_{t_1}, A_{t_2}, \dots, A_{t_n})^T$, $\bar{Y}_{t-}^0 = (Y_{t_1}^0, Y_{t_2}^0, \dots, Y_{t_n}^0, Y_{t-}^0)^T$, and $\underline{Y}_{t+}^0 = (Y_{t_{n+1}}^0, Y_{t_{n+2}}^0, \dots, Y_{t_{n+l}}^0)^T$.

In the regularity conditions, since we assumed existence of joint density, the usual definition of conditional probability is a version of the conditional expectation defined using σ -fields. In our case, we have

$$\begin{aligned} & \lim_{\delta \downarrow 0} \frac{E[r_t(\delta) | \mathcal{F}]}{\delta} \\ &= \lim_{\delta \downarrow 0} \frac{\text{pr}(A_{t+\delta} \neq A_t | Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t-}, Y_{t_{n+1}}^0, \dots, Y_{t_{n+l}}^0)}{\delta} \\ &= \lim_{\delta \downarrow 0} \lim_{t_n \uparrow t-} \frac{\text{pr}(A_{t+\delta} \neq A_{t_n} | Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t_n}, L_{t-}, Y_{t-}^0, Y_{t_{n+1}}^0, \dots, Y_{t_{n+l}}^0)}{\delta + (t - t_n)} \\ &= \lim_{t_n \uparrow t-} \lim_{\delta \downarrow 0} \frac{\text{pr}(A_{t+\delta} \neq A_{t_n} | Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t_n}, L_{t-}, Y_{t-}^0, Y_{t_{n+1}}^0, \dots, Y_{t_{n+l}}^0)}{\delta + (t - t_n)} \\ &= \lim_{t_n \uparrow t-} \frac{\text{pr}(A_t \neq A_{t_n} | Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t_n}, L_{t-}, Y_{t-}^0, Y_{t_{n+1}}^0, \dots, Y_{t_{n+l}}^0)}{t - t_n} \\ &= \lim_{t_n \uparrow t-} \frac{E[I_{A_t \neq A_{t_n}} | \bar{L}_{t-}, \bar{A}_{t_n}, \bar{Y}_{t-}^0, \underline{Y}_{t+}^0]}{t - t_n} \end{aligned}$$

The second equality is guaranteed by the third regularity condition. The interchangeability of limits are guaranteed by the fourth regularity condition, since we have

$$\begin{aligned} & \lim_{\delta \downarrow 0} \frac{\text{pr}(A_{t+\delta} \neq A_{t_n} | Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t_n}, L_{t-}, Y_{t-}^0, Y_{t_{n+1}}^0, \dots, Y_{t_{n+l}}^0)}{\delta + (t - t_n)} \\ &= \frac{\text{pr}(A_t \neq A_{t_n} | Z_{t_1}, \dots, Z_{t_{n-1}}, Z_{t_n}, L_{t-}, Y_{t-}^0, Y_{t_{n+1}}^0, \dots, Y_{t_{n+l}}^0)}{t - t_n} \end{aligned}$$

being uniform in t_n .

Similarly, we can prove that

$$\lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{H}]}{\delta} = \lim_{t_n \uparrow t^-} \frac{E[I_{A_t \neq A_{t_n}} | \bar{L}_{t^-}, \bar{A}_{t_n}, \bar{Y}_{t^-}^0]}{t - t_n}$$

Therefore, we have

$$\begin{aligned} \lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{F}]}{\delta} &= \lim_{t_n \uparrow t^-} \frac{E[I_{A_t \neq A_{t_n}} | \bar{L}_{t^-}, \bar{A}_{t_n}, \bar{Y}_{t^-}^0, \underline{Y}_{t^+}^0]}{t - t_n} \\ &= \lim_{t_n \uparrow t^-} \frac{E[I_{A_t \neq A_{t_n}} | \bar{L}_{t^-}, \bar{A}_{t_n}, \bar{Y}_{t^-}^0]}{t - t_n} \\ &= \lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{H}]}{\delta} \end{aligned}$$

□

Now we prove the final key lemma

Lemma 5.3.7. *Given FTSR, η_t is \mathcal{H}_{t^-} -measurable.*

Proof. We prove the result by the definition of a measurable function with respect to a σ -field.

For any $a \in \mathcal{R}$, consider the following set

$$B \equiv \{\omega : \eta_t \leq a\}$$

Since η_t is measurable w.r.t. \mathcal{F}_{t^-, t^+} , $B \in \mathcal{F}_{t^-, t^+}$.

By Lemma (25.9) (Rogers and Williams, 1994), B is a σ -cylinder, and it can be decided by variables from countably many time points. Suppose the collection of these countably many time points is S . $S = S_1 \cup S_2$, where $t_{1,i} < t$ for $t_{1,i} \in S_1$, and $t_{2,j} > t$ for $t_{2,j} \in S_2$.

Let \mathcal{F}_S denote the σ -field generated by $(Z_{t_{1,i}}, i \in \mathcal{N}; Z_{t-}; Y_{t_{2,j}}^0, j \in \mathcal{N})$. We have augmented the σ -field generated by variables from S with Z_{t-} .

Next define the following series of σ -fields:

$$\mathcal{F}_1 = \sigma(Z_{t_{1,1}}, Z_{t-}, Y_{t_{2,1}}^0)$$

$$\mathcal{F}_2 = \sigma(\mathcal{F}_1, Z_{t_{1,2}}, Y_{t_{2,1}}^0)$$

...

$$\mathcal{F}_\infty = \mathcal{F}_S$$

Considering the following sets:

$$B_1 = \{\omega : E[\eta_t | \mathcal{F}_1] \leq a\}$$

$$B_2 = \{\omega : E[\eta_t | \mathcal{F}_2] \leq a\}$$

...

$$B_S = B_\infty = \{\omega : E[\eta_t | \mathcal{F}_S] \leq a\}$$

We have $B_k \in \mathcal{F}_k$.

It's easy to see that

$$B_1 \supset B_2 \supset \dots \supset B_S$$

because

$$E[E[\eta_t | \mathcal{F}_k] | \mathcal{F}_{k-1}] = E[\eta_t | \mathcal{F}_{k-1}]$$

and taking conditional expectation preserves the direction of inequality.

Also, with the above definitions, $\mathcal{F}_k \uparrow \mathcal{F}_S$. Therefore, by Theorem (5.7) from Durrett 2005, Chapter 4, we know that

$$E[\eta_t | \mathcal{F}_k] \rightarrow E[\eta_t | \mathcal{F}_S] \text{ a.s.}$$

Then, it is easy to see that $I_{B_1} \rightarrow I_{B_S}$ a.s., and that

$$B_S = \bigcap_{i=1}^{\infty} B_i$$

with difference up to a null set.

We now claim that

$$B_S = B \tag{5.3.6}$$

with difference up to a null set.

Obviously $B \subset B_S$. Suppose $Pr(B_S - B) > 0$. Since $B_S - B \in \mathcal{F}_S$, we have

$$\int_{B_S - B} \eta_t Pr(d\omega) = \int_{B_S - B} E[\eta_t | \mathcal{F}_S] Pr(d\omega)$$

Then

$$LHS > aPr(B_S - B)$$

and

$$RHS \leq aPr(B_S - B)$$

This is a contradiction.

Therefore, $B = \bigcap_{i=1}^{\infty} B_i$ with difference up to a null set.

Next, we define

$$\mathcal{H}_1 = \sigma(Z_{t_{1,1}}, Z_{t-})$$

$$\mathcal{H}_2 = \sigma(\mathcal{H}_1, Z_{t_1,2})$$

...

Given FTSR, by Lemma 5.3.6, we have

$$E[\eta_t | \mathcal{F}_k] = E[\eta_t | \mathcal{H}_k]$$

Therefore, every $B_k \in \mathcal{H}_k$, and thus $B_k \in \mathcal{H}_{t-}$.

Since $B = \bigcap_{i=1}^{\infty} B_i$, $B \in \mathcal{H}_{t-}$ as well. By the definition of a measurable function, η_t is measurable with respect to \mathcal{H}_{t-} . □

Combining all the results in this Appendix, we have proved Theorem 2.3.3.

5.4 Proof of Theorem 2.3.4

Proof. Denote $\mathcal{G}_t = \sigma(Y_{t-}^0, L_{t-}, A_{t-})$ and $\mathcal{G}_{t_0,t} = \sigma(\{Y_{l-}^0 : t_0 \leq l \leq t\}, \{L_{l-} : t_0 \leq l \leq t\}, \{A_l : t_0 \leq l < t\})$. Recall the definition of $r_t(\delta) = (1 - A_{t-})A_{t+\delta} + A_{t-}(1 - A_{t+\delta})$, and $Z_t = (Y_t^0, L_t, A_t)^T$. By the Markovian property, we have

$$E[r_t(\delta) | \sigma(\bar{Z}_{t-})] = E[r_t(\delta) | \mathcal{G}_{t_0,t}],$$

for any $t_0 < t$. Since $\mathcal{G}_{t_0,t} \downarrow \mathcal{G}_t$, by Durrett 2005 (Chapter 4, Theorem 6.3), $E[r_t(\delta) | \mathcal{G}_{t_0,t}] \rightarrow E[r_t(\delta) | \mathcal{G}_t]$. Therefore,

$$E[r_t(\delta) | \sigma(\bar{Z}_{t-})] = E[r_t(\delta) | \mathcal{G}_t],$$

Similarly, we can show that

$$E[r_t(\delta) | \sigma(\bar{Z}_{t-}, Y_{t+s}^0)] = E[r_t(\delta) | \sigma(\mathcal{G}_t, Y_{t+s}^0)].$$

Therefore, we have a reduced form of *continuous time sequential randomization*

$$\begin{aligned}
\lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\sigma(\mathcal{G}_t, Y_{t+s}^0)]}{\delta} &= \lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\sigma(\bar{Z}_{t-}, Y_{t+s}^0)]}{\delta} \\
&= \lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\sigma(\bar{Z}_{t-})]}{\delta} \\
&= \lim_{\delta \downarrow 0} \frac{E[r_t(\delta)|\mathcal{G}_t]}{\delta}
\end{aligned}$$

First, we notice that if we can prove

$$pr(Y_{t+s}^0, A_{t-}|Y_{t-}^0, L_{t-})pr(A_t|Y_{t-}^0, L_{t-}) = pr(Y_{t+s}^0, A_t|Y_{t-}^0, L_{t-})pr(A_{t-}|Y_{t-}^0, L_{t-}), \tag{5.4.1}$$

we can conclude (2.3.4). The reason is as follows: assuming that we have (5.4.1) to be true, we integrate A_{t-} out on both sides of the equation. We will get

$$pr(Y_{t+s}^0|Y_{t-}^0, L_{t-})pr(A_t|Y_{t-}^0, L_{t-}) = pr(Y_{t+s}^0, A_t|Y_{t-}^0, L_{t-}).$$

Divide the above equation by $pr(Y_{t+s}^0|Y_{t-}^0, L_{t-})$, we obtain (5.4.1).

Consider

$$g(\delta_1, \delta_2) \equiv pr(Y_{t+s}^0|A_{t+\delta_1} = a_1, A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})$$

where $\delta_1 > 0$ and $\delta_2 > 0$.

We observe that

$$\begin{aligned}
&\lim_{\delta_1 \downarrow 0} \lim_{\delta_2 \downarrow 0} g(\delta_1, \delta_2) \\
&= \lim_{\delta_1 \downarrow 0} pr(Y_{t+s}^0|A_{t+\delta_1} = a_1, A_{t-} = a_2, Y_{t-}^0, L_{t-}) \\
&= \lim_{\delta_1 \downarrow 0} \frac{pr(Y_{t+s}^0, A_{t+\delta_1} = a_1|A_{t-} = a_2, Y_{t-}^0, L_{t-})}{pr(A_{t+\delta_1}|A_{t-} = a_2, Y_{t-}^0, L_{t-})}
\end{aligned}$$

$$\begin{aligned}
&= pr(Y_{t+s}^0 | A_{t-} = a_2, Y_{t-}^0, L_{t-}) \lim_{\delta_1 \downarrow 0} \frac{pr(A_{t+\delta_1} = a_1 | Y_{t+s}^0, A_{t-} = a_2, Y_{t-}^0, L_{t-})}{pr(A_{t+\delta_1} = a_1 | A_{t-} = a_2, Y_{t-}^0, L_{t-})} \\
&= \begin{cases} pr(Y_{t+s}^0 | A_{t-} = a_2, Y_{t-}^0, L_{t-}) \times \lim_{\delta_1 \downarrow 0} \frac{1 - pr(A_{t+\delta_1} \neq a_1 | Y_{t+s}^0, A_{t-} = a_2, Y_{t-}^0, L_{t-})}{1 - pr(A_{t+\delta_1} \neq a_1 | A_{t-} = a_2, Y_{t-}^0, L_{t-})} & \text{if } a_1 = a_2 \\ \\ pr(Y_{t+s}^0 | A_{t-} = a_2, Y_{t-}^0, L_{t-}) \times \lim_{\delta_1 \downarrow 0} \frac{\frac{pr(A_{t+\delta_1} \neq a_1 | Y_{t+s}^0, A_{t-} = a_2, Y_{t-}^0, L_{t-})}{\delta}}{\frac{pr(A_{t+\delta_1} \neq a_1 | A_{t-} = a_2, Y_{t-}^0, L_{t-})}{\delta}} & \text{if } a_1 \neq a_2 \end{cases} \\
&= pr(Y_{t+s}^0 | A_{t-} = a_2, Y_{t-}^0, L_{t-})
\end{aligned}$$

Here taking limit inside density is guaranteed by the third regularity condition, and the last equality is because of continuous time sequential randomization assumption.

We also observe that

$$\begin{aligned}
&\lim_{\delta_2 \downarrow 0} \lim_{\delta_1 \downarrow 0} g(\delta_1, \delta_2) \\
&= \lim_{\delta_2 \downarrow 0} pr(Y_{t+s}^0 | A_{t-\delta_2}, A_t, Y_{t-}^0, L_{t-}) \\
&= \lim_{\delta_2 \downarrow 0} pr(Y_{t+s}^0 | A_t, Y_{t-}^0, L_{t-}) \\
&= pr(Y_{t+s}^0 | A_t, Y_{t-}^0, L_{t-})
\end{aligned}$$

The second equality used the Markov property.

If we can interchange the limits, then

$$pr(Y_{t+s}^0 | A_{t-}, Y_{t-}^0, L_{t-}) = pr(Y_{t+s}^0 | A_t, Y_{t-}^0, L_{t-}).$$

Equation (5.4.1) follows from the definition of conditional density.

We now establish the fact that

$$\lim_{\delta_2 \downarrow 0} \lim_{\delta_1 \downarrow 0} g(\delta_1, \delta_2) = \lim_{\delta_1 \downarrow 0} \lim_{\delta_2 \downarrow 0} g(\delta_1, \delta_2)$$

by showing that $\lim_{\delta_1 \downarrow 0} g(\delta_1, \delta_2)$ is uniform in δ_2 .

Define $g_1(\delta_2) = \lim_{\delta_1 \downarrow 0} g(\delta_1, \delta_2)$, then

$$\begin{aligned} & |g(\delta_1, \delta_2) - g_1(\delta_2)| \\ &= \left| \frac{\text{pr}(Y_{t+s}^0, A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})}{\text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})} \right. \\ & \quad \left. - \frac{\text{pr}(Y_{t+s}^0, A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})}{\text{pr}(A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})} \right| \\ &= \text{pr}(Y_{t+s}^0 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}) \\ & \quad \times \left| \frac{\text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0)}{\text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})} \right. \\ & \quad \left. - \frac{\text{pr}(A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0)}{\text{pr}(A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})} \right| \end{aligned}$$

Consider the ratio $\frac{\text{pr}(A_{t+\delta_1}=a_1 | A_{t-\delta_2}=a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0)}{\text{pr}(A_{t+\delta_1}=a_1 | A_{t-\delta_2}=a_2, Y_{t-}^0, L_{t-})}$. We claim that it converges to $\frac{\text{pr}(A_t=a_1 | A_{t-\delta_2}=a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0)}{\text{pr}(A_t=a_1 | A_{t-\delta_2}=a_2, Y_{t-}^0, L_{t-})}$ uniformly in δ_2 .

If $a_1 = a_2$, density $\text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})$ is bounded from below by a positive number. By the third regularity condition,

$$\begin{aligned} & \text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0) \\ & \rightarrow \text{pr}(A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0) \end{aligned}$$

and

$$\text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}) \rightarrow \text{pr}(A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})$$

uniformly in δ_2 , as $\delta_1 \downarrow 0$. When the denominators are bounded from below by a positive number, the ratio also converges uniformly.

If $a_1 \neq a_2$, by the fourth regularity condition,

$$\begin{aligned} & \frac{\text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0)}{\delta_1 + \delta_2} \\ & \rightarrow \frac{\text{pr}(A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-}, Y_{t+s}^0)}{\delta_2} \end{aligned}$$

and

$$\frac{\text{pr}(A_{t+\delta_1} = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})}{\delta_1 + \delta_2} \rightarrow \frac{\text{pr}(A_t = a_1 | A_{t-\delta_2} = a_2, Y_{t-}^0, L_{t-})}{\delta_2}$$

uniformly in δ_2 , as $\delta_1 \downarrow 0$. Also the denominator $\frac{\text{pr}(A_{t+\delta_1}=a_1|A_{t-\delta_2}=a_2, Y_{t-}^0, L_{t-})}{\delta_1+\delta_2}$ is bounded from below by a positive number. Hence we establish the uniformly convergence of the ratio.

Combining the two cases above, $|g(\delta_1, \delta_2) - g_1(\delta_2)|$ is bounded by $O(\delta_1)$ that does not depend on δ_2 , so $g(\delta_1, \delta_2) \rightarrow g_1(\delta_2)$ uniformly in δ_2 . Therefore,

$$\lim_{\delta_2 \downarrow 0} \lim_{\delta_1 \downarrow 0} g(\delta_1, \delta_2) = \lim_{\delta_1 \downarrow 0} \lim_{\delta_2 \downarrow 0} g(\delta_1, \delta_2)$$

By the argument at the beginning of the proof, we have proved the first part of the theorem.

To show that (2.3.4) implies FTSR, without of loss of generality, we consider

$$\begin{aligned} & \frac{\text{pr}(A_t | L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0, Y_{t+s}^0)}{\sum_{i=0,1} \text{pr}(A_t = i, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0, Y_{t+s}^0)} \\ & = \frac{\text{pr}(A_t, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0, Y_{t+s}^0)}{\sum_{i=0,1} \text{pr}(A_t = i, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0, Y_{t+s}^0)} \end{aligned}$$

$$\begin{aligned}
&= \frac{pr(Y_{t+s}^0 | A_t, L_{t-}, Y_{t-}^0) pr(A_t, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0)}{\sum_{i=0,1} pr(Y_{t+s}^0 | A_t = i, L_{t-}, Y_{t-}^0) pr(A_t = i, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0)} \\
&= \frac{pr(Y_{t+s}^0 | L_{t-}, Y_{t-}^0) pr(A_t, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0)}{\sum_{i=0,1} pr(Y_{t+s}^0 | L_{t-}, Y_{t-}^0) pr(A_t = i, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0)} \\
&= \frac{pr(A_t, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0)}{\sum_{i=0,1} pr(A_t = i, L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0)} \\
&= pr(A_t | L_{t-}, Y_{t-}^0, A_{t-m}, L_{(t-m)-}, Y_{(t-m)-}^0)
\end{aligned}$$

The second equality is because of Markov property. The third equality used equation (2.3.4). We have proved the second half of the theorem. \square

5.5 Simulation Parameters

In all simulation models from M1 to M4, we specify parameters as follows:

- Let $g(V, t) = C$, a constant. Let $C = 100$.
- For M1 (also in M3 and M4), let $\theta = 0.2$ and $\sigma = 1$.
- For M2, let $m = 2$, $\theta_1 = 0.2$, $\sigma_1 = 1$ and $\theta_2 = 1$, $\sigma_2 = 0.5$. The transition

probability of J_t would be $P(t) = e^{At}$, where $A = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$.

- For initial value, e_0 is generated from $N(0, \frac{\sigma}{\sqrt{2\theta}})$.
- The causal parameter $\Psi = 1$.
- In M1, M2 and M3, $s(A_t, Y_t) = e^{\alpha_0 + \alpha_1 A_t + \alpha_2 Y_t + \alpha_3 A_t Y_t}$. Let $\alpha_1 = -0.3$, $\alpha_2 = -0.005$, $\alpha_3 = 0.007$ and $\alpha_0 = -0.2$.

- In M4, A_t is generated as follows: if $Y_{t-0.5} > 101$ and $Y_t > 101$, A_t jumps to 0 with probability 0.7, if A_t has not been 0; if $Y_{t0.5} < 99$ and $Y_t < 99$, A_t jumps to 1 with probability 0.7, if it has not been 1; otherwise, A_t is generated following the same model as similar to that in M1, except that $s(A_t, L_t^*) = e^{\alpha_0 + \alpha_1 A_t + \alpha_2 L_t^* + \alpha_3 A_t L_t^*}$. The values of the α 's are the same as before.
- In M4, η_t follows an Ornstein-Uhlenbeck process with parameters $\theta = 0.2$ and $\sigma = 1$.
- For initial value, A_0 is generated from $Bernoulli(\text{expit}(\alpha_0 + \alpha_2 Y_0))$.
- $K = 5$ is the number of periods.
- Number of subjects $n = 5000$.

5.6 Continuous Time Ignorability

In this section, we give a technical definition of *continuous time ignorability assumption*, and prove a sufficient condition for a Markov process to satisfy the continuous time ignorability assumption.

This definition of continuous time ignorability follows and extends the formulations by Lok (Lok, 2008), whose formulation is only for a single outcome. It also generalizes the formulation in Chapter 2, whose definition is for a rank preserving model.

Definition 5.6.1 (Continuous Time Ignorability). Assume that $X_t = (A_t^*, L_t, Y_t, Y_t^{\bar{A}_t^*-0})^T$ is a continuous time *càdlàg* process and that A_t^* is a discrete jumping process.

Let

$$\mathcal{F}_{t,h} = \sigma(\bar{L}_{t-}, \bar{Y}_{t-}, \bar{A}_{t-}^*, Y_{t+h}^{\bar{A}_{t-}^*}, 0),$$

and

$$\mathcal{F}_t = \sigma(\bar{L}_{t-}, \bar{Y}_{t-}, \bar{A}_{t-}^*).$$

Assume that

$$E[I_{A_s^* \text{ jumps more than once within } [t, t+h]} | \mathcal{F}_{t,h}] = o_1(h).$$

where $o_1(h)/h \rightarrow 0$ a.s. when $h \rightarrow 0$.

We say that the process satisfies **continuous time ignorability** assumption, if

$$E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}_{t,h}] = hs(\bar{L}_{t-}, \bar{Y}_{t-}, \bar{A}_{t-}^*) + o_2(h)$$

where $s(\cdot)$ is a nonnegative functional, bounded and measurable with respect to \mathcal{F}_t , and $o_2(h)/h \rightarrow 0$ a.s. when $h \rightarrow 0$.

The definition basically states that treatment at time t only depends on observable historical covariates, outcomes and treatments prior to time t , and does not depend on the future potential outcome.

The following lemma can be used to prove that our model follows the continuous time ignorability.

Lemma 5.6.2. *Assume that $X_t = (A_t^*, L_t, Y_t, Y_t^{\bar{A}_{t-}^*, 0})^T$ is a continuous time Markov process and that it satisfies the regularity conditions in Definition 5.6.1. Let*

$$\mathcal{G}_{t,h} = \sigma(L_{t-}, Y_{t-}, A_{t-}^*, Y_{t-}^{\bar{A}_{t-}^*, 0}, Y_{t+h}^{\bar{A}_{(t+h)-}^*, 0}),$$

and

$$\mathcal{H}_t = \sigma(L_{t-}, Y_{t-}, A_{t-}^*).$$

If

$$E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{G}_{t,h}] = h s^*(L_{t-}, Y_{t-}, A_{t-}^*) + o_3(h), \quad (5.6.1)$$

where $s^*(\cdot)$ is a nonnegative functional, bounded and measurable with respect to \mathcal{H}_t , and $o_3(h)/h \rightarrow 0$ a.s. when $h \rightarrow 0$, then the Markov process satisfies the continuous time ignorability.

Proof. Let $\mathcal{G}_t = \sigma(L_{t-}, Y_{t-}, A_{t-}^*, Y_{t-}^{\bar{A}_{t-}^*, 0})$, $\mathcal{F}'_t = \sigma(\bar{L}_{t-}, \bar{Y}_{t-}, \bar{A}_{t-}^*, \bar{Y}_{t-}^{\bar{A}_{t-}^*, 0})$, and $\mathcal{F}'_{t,h} = \sigma(\bar{L}_{t-}, \bar{Y}_{t-}, \bar{A}_{t-}^*, \bar{Y}_{t-}^{\bar{A}_{t-}^*, 0}, Y_{t+h}^{\bar{A}_{(t+h)-}^*, 0})$.

We claim that

$$E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}'_{t,h}] = E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{G}_{t,h}]$$

For any finite set of $t_1 < t_2 < \dots < t_n < t$, let

$$\begin{aligned} \mathcal{F}'_{n,t,h} = & \sigma(L_{t_1}, \dots, L_{t_n}, L_{t-}, Y_{t_1}, \dots, Y_{t_n}, Y_{t-}, A_{t_1}^*, \dots, A_{t_n}^*, A_{t-}^*, Y_{t_1}^{\bar{A}_{t_1-}^*, 0}, \dots, \\ & Y_{t_n}^{\bar{A}_{t_n-}^*, 0}, Y_{t-}^{\bar{A}_{t-}^*, 0}, \bar{Y}_{(t+h)-}^{\bar{A}_{(t+h)-}^*, 0}). \end{aligned}$$

It is easy to see that

$$E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}'_{n,t,h}] = E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{G}_{t,h}], \text{ a.s.},$$

because of the Markov property of X_t and the equivalence of conditional expectation.

By the definition of conditional expectation,

$$\int_B I_{A_{t+h}^* \neq A_{t-}^*} dP = \int_B E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{G}_{t,h}] dP, a.s.$$

for any $B \in \mathcal{F}'_{n,t,h}$. Since $\mathcal{F}'_{t,h}$ is generated by all such $\mathcal{F}'_{n,t,h}$'s, by dominated convergence theorem, the above equation is true for any $B \in \mathcal{F}'_{t,h}$. Therefore, we have proved that

$$E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}'_{t,h}] = E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{G}_{t,h}]$$

Therefore,

$$E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}'_{t,h}] = hs^*(L_{t-}, Y_{t-}, A_{t-}^*) + o_3(h)$$

and

$$\begin{aligned} E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}_{t,h}] &= E\{E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}'_{t,h}] | \mathcal{F}_{t,h}\} \\ &= hs^*(L_{t-}, Y_{t-}, A_{t-}^*) + E[o_3(h) | \mathcal{F}_{t,h}] \end{aligned}$$

As we have assumed that $o_3(h)/h \rightarrow 0$ a.s. as $h \rightarrow 0$, $E[o_3(h)/h | \mathcal{F}_{t,h}] \rightarrow 0$ a.s. as $h \rightarrow 0$ by dominated convergence theorem. Denote $o_4(h) = E[o_3(h) | \mathcal{F}_{t,h}]$

We have proven that

$$E[I_{A_{t+h}^* \neq A_{t-}^*} | \mathcal{F}_{t,h}] = hs^*(L_{t-}, Y_{t-}, A_{t-}^*) + o_4(h)$$

which satisfies the definition of the continuous time ignorability. \square

Using Lemma 5.6.2 to prove that our model satisfies Definition 5.6.1 is straightforward. It is easy to see that the model we defined in Section 3.2 guarantees equation (5.6.1).

5.7 Simulation of Endpoint-Conditioned Bounded Simple Random Walk

This section describes how we simulate the path of A^* given the number of switches in the path and that the path matches certain starting point and ending point. We reformulate the problems as the follows. We need to simulate a series of binary x'_1, x'_2, \dots, x'_M being either $+1$ or -1 , such that $s_k = \sum_{j=1}^k x'_j$ is always between L and U inclusively, $L < U$, and that $s_M = N$, assuming all qualifying paths have the same probability.

We start by considering all the paths that only matches the end points, and assume all these paths have the same probability. We denote the proportion of these paths that are bounded between L and U to be $p(M, N, L, U)$. It is then very easy to get a recursive equation

$$\begin{aligned} p(M, N, L, U) &= \frac{M - N}{2M} p(M - 1, N + 1, L + 1, U + 1) \\ &\quad + \frac{M + N}{2M} p(M - 1, N - 1, L - 1, U - 1) \end{aligned} \quad (5.7.1)$$

with boundary conditions properly defined.

If the function $q(M, N, L, U)$ can be calculated easily, we can calculate

$$\begin{aligned} &P(x'_1 = 1 | \{x'_j\}_{j=1}^M \text{ is a qualified path.}) \\ &= \frac{P(x'_1 = 1, \{x'_j\}_{j=1}^M \text{ is a qualified path.})}{P(\{x'_j\}_{j=1}^M \text{ is a qualified path.})} \\ &= \frac{P(x'_1 = 1)p(M - 1, N - 1, L - 1, U - 1)}{p(M, N, L, U)} \end{aligned}$$

$$= \frac{\frac{M+N}{2M} p(M-1, N-1, L-1, U-1)}{p(M, N, L, U)},$$

and

$$\begin{aligned} & P(x'_2 = 1 | x'_1 = 1, \{x'_j\}_{j=1}^M \text{ is a qualified path.}) \\ &= \frac{P(x'_2 = 1 | x'_1 = 1) P(\{x'_j\}_{j=1}^M \text{ is a qualified path.} | x'_1 = 1, x'_2 = 1)}{P(\{x'_j\}_{j=1}^M \text{ is a qualified path.} | x'_1 = 1)} \\ &= \frac{\frac{M+N-1}{2M-2} p(M-2, N-2, L-2, U-2)}{p(M-1, N-1, L-1, U-1)}. \end{aligned}$$

Similarly, we can calculate any

$$P(x'_k = 1 | x'_1 = i_1, \dots, x'_{k-1} = i_{k-1}, \{x'_j\}_{j=1}^M \text{ is a qualified path.})$$

in a similar fashion, and thus simulate x'_j sequentially.

The key to the computation is how to evaluate $p(M, N, L, U)$ efficiently. The recursive equation (5.7.1) with proper boundary conditions can be used to evaluate the function, but it is very inefficient. Instead, we give a closed form formula for $p(M, N, L, U)$, which can be evaluated much faster. Define n_q as the number of qualifying paths. Using reflection principle repeatedly, it can be shown that

$$\begin{aligned} n_q = \sum_{k=0}^M & \left[\binom{M}{k(U+1) - k(L-1) + (M+N)/2} \right. \\ & - \binom{M}{(k+1)(U+1) - k(L-1) + (M+N)/2} \\ & \left. - \binom{M}{k(U+1) - (k+1)(L-1) + (M-N)/2} \right] \end{aligned}$$

$$+ \left(\begin{array}{c} M \\ (k+1)(U+1) - (k+1)(L-1) + (M-N)/2 \end{array} \right) \Bigg]$$

Then

$$p(M, N, L, U) = \frac{n_q}{\binom{M}{(M+N)/2}}$$

Here n_q can be used to calculate the proposal distribution in Section 3.3.3.

Bibliography

- Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica* *70*(1), 223–262.
- Blackwell, P. (2003). Bayesian inference for markov processes with diffusion and discrete components. *Biometrika* *90*, 613–627.
- Bladt, M. and Sørensen, M. (2005). Statistical inference for discretely observed markov jump processes. *Journal of Royal Statistical Society, Series B* *68*, 767–784.
- Brumback, B. A., Hernán, M. A., Haneuse, S. J., and Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, **23**(5), 749–767.
- Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, **128**, 134–155.
- Daulaire, N. M., Starbuck, E. S., Houston, R. M., Church, M. S., Stukel, T. A., and

- Pandey, M. R. (1992). Childhood mortality after a high dose of vitamin a in a high risk population. *British Medical Journal* *304*(6821), 207–210.
- del Ninno, C., Dorosh, P., Smith, L. and Roy, D. (2001). The 1998 floods in Bangladesh: disaster impacts, household coping strategies and response. *International Food Policy Research Institute Research Report No. 122*. IFPRI: Washington, D.C.
- del Ninno, C. and Lundberg, M. (2005), Treading water: the long-term impact of the 1998 flood on nutrition in Bangladesh. *Economics and Human Biology*, **3**, 67–96.
- Durrett, R. (2005), Probability: Theory and Examples. *Duxbury Advanced Series*, Third Edition, Brooks/Cole-Thomson Learning: Belmont, CA.
- Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* *69*(4), 959–993.
- Gasser, T., Muller, H., Kohler, W., Molinari, L., and Prader, A. (1984), Nonparametric Regression Analysis of Growth Curves. *The Annals of Statistics*, **12:1**, 210–229.
- Guerrant, R. L., Kosek, M., Lima, A. A. M., Lorntz, B., and Guyatt, H. L. (2002), Updating the Dalys for diarrhoeal disease. *Trends in Parasitology*, **18**, 191–193.
- Hernán, M. A., Brumback, B., and Robins, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, **21**, 1689–1709.

- Hobolth, A. (2008). A markov chain monte carlo expectation maximization algorithm for statistical analysis of dna sequence evolution with neighbor-dependent substitution rates. *Journal of Computational and Graphical Statistics* 17, 138–162.
- Humphrey, J. H., West, Jr, K. P., and Sommer, A. (1992). Vitamin a deficiency and attributable mortality among under-5-year-olds. *Bull World Health Organ* 70(2), 225–232.
- Joffe, M. M. and Robins, J. M. (2009). Controlling the future: revised assumptions and methods for causal inference with repeated measures outcomes. Working Paper.
- Johannes, M. S., Polson, N. G., and Stroud, J. R. (2009). Optimal filtering of jump diffusions: Extracting latent states from asset prices. *Review of Financial Studies* 22, 2559–2599.
- Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association* 80(392), 863–871.
- Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. Jr. (1987), The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, **126**, 310–318.

- Kosek, M., Bern, C. and Guerrant, R. L. (2003), The global burden of diarrhoeal disease as estimated from studies published between 1992 and 2000. *Bulletin of the World Health Organization*, **81**, 197–204.
- Lok, J. J. (2004), Mimicking counterfactual outcomes for the estimation of causal effects. <http://www.citebase.org/abstract?id=oai:arXiv.org:math/0409045>.
- Lok, J. J. (2008), Statistical modeling of causal effects in continuous time. *The Annals of Statistics*. **36**, 1464–1507.
- Mark, S. D. and Robins, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in Medicine*, **12**, 1605–1628.
- Moore, S. R., Lima, A. A. M., Conaway, M. R., Schorling, J. B., Soares, A. M. and Guerrant, R. L. (2001), Early childhood diarrhea and helminthiases associate with long-term linear growth faltering. *International Journal of Epidemiology*, **30**, 1457–1464.
- Nielsen, R. (2002). Mapping mutations on phylogenies. *Systematic Biology* *51*, 729–739.
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–1512 (errata 1987, **14**, 917–921).

- Robins, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, **79**, 2, 321–334.
- Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, **23**, 2379–2412.
- Robins, J. M. (1997). Causal inference from complex longitudinal data. *Latent Variable Modelling and Applications to Causality. Lecture Notes in Statistics*, **120**, M. Berkane, Editor. NY: Springer Verlag, 69–117.
- Robins, J. M. (1998). Marginal structural models. *1997 Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*, 1–10. Reproduced courtesy of the American Statistical Association.
- Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. *Statistical Models in Epidemiology: The Environment and Clinical Trials*, M.E. Halloran and D. Berry, Editors, IMA Volume 116, NY: Springer-Verlag, 95–134.
- Robins, J. M. (2008) Causal Models for Estimating the Effects of Weight Gain on Mortality. *International Journal of Obesity*, **32**, S15–S41.
- Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, **3**, 319–336.

- Robins, J. M., Greenland, S., and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* *94*, 687–700.
- Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.
- Robins, J. M., and Tsiatis, A. (1991). Correcting for non-compliance in randomized trials using rank-preserving structural failure time models. *Communications in Statistics*, **20**, 2609–2631.
- Rogers, L. C. G. and D. Williams (1994). Diffusions, Markov processes, and martingales, Volume 1 of *Wiley series in probability and mathematical statistics*. Chichester, New York: John Wiley and Sons.
- Rosenbaum, P. R. (1984). The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment, *Journal of the Royal Statistical Society. Series A*, **147:5**, 656–666.
- Rosenbaum, P. R. (2002). *Observational Studies*, 2nd edition, Springer.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* *66*, 688–701.
- Singer, B. (1981). Estimation of nonstationary Markov chains from panel data. *Sociological Methodology*, **12**, 319–337.

- Sommer, A., Katz, J., and Tarwotjo, I. (1983). Increased mortality in children with mild vitamin a deficiency. *American Journal of Clinical Nutrition* 40, 1090–1095.
- Tsiatis, A. A. (2006). *Semiparametric theory and missing data*, Springer.
- van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* 10(1), 1–50.
- Vijayaraghavan, K., Radhaiah, G., Prakasam, B. S., Sarma, K. V., and Reddy, V. (1990). Effect of massive dose vitamin a on morbidity and mortality in indian children. *Lancet* 336(8727), 1342–1345.
- Wei, J. and Norman, E. (1963). Lie algebraic solution of linear differential equations. *Journal of Mathematical Physics* 4, 575–581.
- Wei, J. and Norman, E. (1964). On global representation of the solutions of linear differential equations as a product of the exponentials. *Proceedings of American Mathematical Society* 15, 327–334.
- West, K. P. J., Pokhrel, R. P., Katz, J., LeClerq, S. C., Khattry, S. K., Shrestha, S. R., Pradhan, E. K., Tielsch, J. M., Pandey, M. R., and Sommer, A. (1991). Efficacy of vitamin a in reducing preschool child mortality in nepal. *Lancet* 338(8759), 67–71.
- Zeger, S. L. and Karim, M. R. (1991). Generalized linear model with randomized effects: a gibbs sampling approach. *Journal of the American Statistical Association* 84(413), 79–86.

Zeger, S. L. and Liang, K.-Y. (1991). Feedback models for discrete and continuous time series. *Statistica Sinica* 1, 51–64.

Zhang, M., Joffe, M. M., and Small, D. S. (2009). Causal inference for continuous time processes when covariates are observed only at discrete times. *Working Paper*.

Zhang, M. and Small, D. S. (2009). Effect of Vitamin A deficiency on respiratory infection: causal inference for a discretely observed continuous time non-stationary Markov process. *Working Paper*.