Northern Illinois University

Huskie Commons

Graduate Research Theses & Dissertations | Graduate Research & Artistry

2018

Bayesian methods for optimal treatment allocation

Saptarshi Chatterjee

Follow this and additional works at: https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations

Recommended Citation

Chatterjee, Saptarshi, "Bayesian methods for optimal treatment allocation" (2018). *Graduate Research Theses & Dissertations*. 1610.
https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations/1610

This Dissertation/Thesis is brought to you for free and open access by the Graduate Research & Artistry at Huskie Commons. It has been accepted for inclusion in Graduate Research Theses & Dissertations by an authorized administrator of Huskie Commons. For more information, please contact jschumacher@niu.edu.

ABSTRACT

BAYESIAN METHODS FOR OPTIMAL TREATMENT ALLOCATION

Saptarshi Chatterjee, Ph.D.
Department of Mathematical Sciences
Northern Illinois University, 2018
Sanjib Basu, Director

In chronic diseases such as cancer, physicians make multiple treatment decisions over the course of a patient's disease depending on his/her biological characteristics and accrued information. Essentially, the treatment rule at each decision point is a function which takes the patients' biomarker information, treatment and outcome history available up to that point as an input and returns the treatment choice as an output. In the single treatment setting, the optimal treatment decision can be obtained by a regression model on the mean outcome conditional on treatment and covariates, where the optimal treatment is the one that corresponds to the most desirable mean outcome. However, due to its over dependence on the outcome regression model, this method is heavily prone to model misspecification. Also, given data from an observational study, the usual regression method does not control for the confounding bias induced by the covariates affecting both treatment and outcomes.

Causal inference provides a general framework to estimate the treatment causal effect by comparing the potential outcomes under each treatment group. However, for an individual patient, only one potential outcome is observed limiting the direct comparison of potential outcomes at the patient level. A handful number of methods have been proposed in the recent precision medicine literature where they employ semi-parametric estimation methods

such as inverse probability weighting (IPWE) to predict the optimal treatment by maximizing a certain predefined value function. However, the likelihood based methods have received little attention in this area, partly due to making model assumptions. To fill this gap, in this dissertation, we develop two fully Bayesian semiparametric likelihood based methods to predict the optimal treatment for a new patient based on the treatment and covariate information from an observed group of patients. In the first approach (BayesG) we extend the idea of parametric g-formula to include a semiparametric mean function within a marginal structural model framework. In the second approach (PSBayes), we connect the treatment assignment mechanism to a missing data framework and build on the Penalized Spline of Propensity Prediction (PSPP) method in the missing data literature to develop a methodology to predict and compare the potential outcomes of the new patient. The posterior predictive potential outcome distribution is then analyzed to predict the optimal treatment. The performance of the proposed methodologies are illustrated in five different simulation studies covering a wide range of scenarios. Overall, the true specifications of inverse probability methods display comparable performance whereas the misspecified models perform poorly. In the additive mean function scenarios, PSBayes outperform all other methods in having higher accuracy in predicting true optimal treatments, whereas the inverse probability based methods show better performance in nonlinear mean function cases. In the presence of non effect modifiers, the BayesG approach perform better than other methods. We conclude the dissertation by discussing the extension of our proposed methods to a dynamic treatment setting.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

AUGUST 2018

BAYESIAN METHODS FOR OPTIMAL TREATMENT ALLOCATION

BY

SAPTARSHI CHATTERJEE
© 2018 Saptarshi Chatterjee

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICAL SCIENCES

Doctoral Dissertation Director:
Sanjib Basu

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my deepest gratitude to my advisor, Dr. Sanjib Basu, not only for his patient guidance during this work, but also for pushing me to strive for excellence in research. Without his scientific vision, insight, and generous support this work would not bear fruition. I feel fortunate to get the opportunity to work under his supervision and I thank him from the bottom of my heart for everything that he has done for me in the last six years.

I wish to extend my sincere thanks to the other committee members namely Dr. Alan Polansky, Dr. Jeffrey Thunder and Dr. Ryu for kindly agreeing to serve on my committee. My special thanks also goes to Dr. Dulal Bhaumik for kindly agreeing to be my external examiner.

I will forever be indebted to the Division of Statistics for providing me the wonderful opportunity to pursue my PhD here. This is an amazing department and I sincerely thank all the faculty members, instructors and support stuffs for their time and support.

I have been very fortunate to take courses from some of the best professors in my life at NIU. I thank all of them: Dr. Alan Polansky, Dr. Sanjib Basu, Dr. Balakrishna Hosmane, Dr. Nader Ebrahimi, Dr. Duchwan Ryu, Dr. Michelle Xia and Dr. Larry Hua.

I would also like to acknowledge Dr. Rama Lingham and Dr. Sanjib Basu for their initiatives in setting up the account in Gaea. I am particularly thankful to John Winans and

Nicholas Karonis in the Center for Research Computing and Data at NIU, for the support they provided. This work used resources of the Center for Research Computing and Data at Northern Illinois University.

The assistance, cooperation of my fellow friends in DeKalb was also important in the completion of this work. I would like to thank them all.

I wholeheartedly thank my parents, Mr. Subhash Chatterjee and Mrs. Ishani Chatterjee for their unconditional love and support throughout my life and I specially thank my brother, Dr. Shirshendu Chatterjee for introducing me to the fascnating world of Statistics.

Last but not the least, my deepest thanks goes to my wife, Dr. Shrabanti Chowdhury, for encouraging me to come to this part of the world. She motivated me throughout my academic journey and without her support this feat could not have been possible to achieve.

DEDICATION

To the loving memories of my late maternal aunt Mrs. Pratibha Chakrabortty and my late grandmothers, Mrs. Kamala Chatterjee and Mrs. Nirmala Chakrabortty.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Clinical studies are the backbone of modern drug discovery. A clinical trial generally focuses on demonstrating the superiority or non-inferiority of a new intervention over the standard procedure (active control) or over an intervention with no active component (placebo) at the population level. However, in many therapeutic areas such as cancer, patients show heterogeneous response to the same treatment. Thus it is often difficult to discover a treatment which can cure all patients irrespective of their biological characteristics or medical history. To this direction, this dissertation focuses on methodologies for tailoring treatments to match with patients' profiles. Specifically, we propose two Bayesian semi-parametric methods (BayesG and PSBayes) for predicting the best treatment option for a new patient based on an observed group of patients. In the following chapters and sections we will discuss the important aspects related to our methods in detail.

1.1 Observational Study

Clinical research explores whether a new intervention, strategy or medical device is safe and effective for human. Primarily, there are two types of clinical studies - randomized clinical trials or randomized controlled trials (RCT) and observational studies. In a RCT, volunteers are randomly assigned to specific medical product such as a drug, medical device, medical procedures or a specific diet plan according to research design and protocols created and managed by investigators. An important feature of this type of study is that random-

ization is used as a tool to reduce bias among the participants. Study subjects are grouped randomly so that the of treatments are compared on a fair ground as the randomized groups are expected to have similar profile of biological characteristics associated with treatment benefit.

Although RCT is considered the gold standard in clinical research for investigating safety and efficacy of new drug(s) there are a number of factors that limit the practical usage of it. First of all, there is an ethical concern. In a simple RCT patients are randomly allocated to active or control group in order to remove confounding bias and to estimate the difference in outcomes only due to the active component. However, in many scenarios it is very difficult to conduct a blind experiment. For example, in the hypothetical randomized experiment to investigate the causal effect of abortion in increasing the risk of breast cancer, the first step would be to select a large set of pregnant women and then randomly divide them into two groups based on induced abortion and the study would conclude after observing the patients a long period of time. As one can imagine, this type of designed experiment would clearly violate the ethical standard of clinical practice. Besides the ethical concern, a majority of clinical trials fails to reflect the real-life scenarios specifically when the risk of adverse effects needs to be analyzed. Traditional RCTs often do not use large population samples or do not have adequate follow up and thus lack in gaining enough power for detecting rare adverse effects. Besides, patients with high risk of adverse effects are usually excluded from the trial.

Observational studies, on the other hand provide an important tool in providing information on the safety and benefit of the evidence based medications. As a result, these clinical studies are primarily used to detect and analyze the risk factors when a RCT is impossible or unethical to conduct. However, the main challenge in observational studies lies in the fact that the treatment allocation mechanism is not fully or partially randomized and the independent variables are not controlled by the researchers. The participants or the providers of the study determine which intervention or therapy the participants should receive. This

evidence based treatment assignment makes it difficult to separate out the effect of treatment on the outcome from the prognostic effect of covariates or independent variables on the outcome. The usual regression method when applied in this setting, often produces biased estimates of the parameters related to the causal relationship between treatments and observed outcome. The confounding bias in observational studies are well documented in literature and statistical methods have been explored by both econometricians and statisticians to correctly estimate the causal relationship between treatment and patient's response (Robins, 1986; Rubin, 1974; Imbens, 2004; Heckman et al., 2014).

1.2 Personalized Medicine

Traditionally, clinical studies focused on determining treatment effect in the whole population. However, this 'one-size-fits-all' approach to assess treatment efficacy and safety does not reflect patients' heterogeneity in their biological or physiological profiles. For example, a young patient might respond differently to a drug than an older patient do or a treatment might have different effect on a patient with genetic mutation in cytochrome enzyme than others (Berger et al., 2014). The heterogeneity in response to the drugs are evident in the fact that most of the conventional drug development projects are often associated with higher attrition rate and they subsequently succumb to failures. Moreover, a drug which works better in an early phase approved treatments for many therapeutic areas often show less efficacy in follow-up clinical trials. As Temple and Ellenberg (2000) pointed out, this finding can be attributed to patients' diverse genetic profile, clinical markers (e.g. medical history, disease severity), demographics (e.g. sex, age) and social or environmental factors (e.g. smoking habits, education). There are two directions that one can follow to incorporate individual heterogeneity in clinical research. The first one is to define a group of patients

based on their baseline covariates for a predefined intervention such that the treatment effect is maximum for the group of patients. This area of research is formally called subgroup identification which has gained lot of attention in recent years (Tian and Tibshirani, 2010; Lipkovich et al., 2011; Berger et al., 2014; Chen et al., 2015; Loh et al., 2015; Huang et al., 2017). The other perspective to address heterogeneity is based on synthesizing a patient's individual characteristics into making the optimal treatment decision to match with the patient's profile. This area of research, referred to as Personalized Medicine is an emerging field of medical science. We will focus on this perspective.

The primary motivation behind personalized Medicine paradigm comes from the limitation of the classical medicine approach that it cannot synthesize patient's individual heterogeneity in treatment decision making. Personalized medicine has the ability to incorporate all information available on the patient including genetic, physiologic, demographic or clinical information in the process to decide the most effective treatment option for his/her profile. In the simplest form of single treatment decision, it defines a decision rule which takes a patient's information as an input and outputs the treatment option that the patient should receive among the choice of available treatment options. The optimal regime corresponds to the highest gain for a patient if (s)he had followed the regime i.e. if a patient gets a treatment according to the optimal regime it is expected that (s)he would have received greatest clinical benefit in compared to what (s)he would have got otherwise. We will put more light on this aspect in Chapter 2.

1.3 Research Problem

In this dissertation, we focus on developing Bayesian methodologies to predict the optimal treatment allocation for a new patient. Our methods are based on observational study where

we assume to have access to the treatment, response and covariate information on a group of patients from the same population. To identify the optimal treatment for a patient we recast our problem within potential outcome framework which will be detailed in section 3.1.

In observational study, the treatment assignments are usually driven by patients' biological or clinical profiles. Due to the nonrandom treatment assignment mechanism, the individuals in the treated and the untreated groups may have systematic imbalance or heterogeneity in the distribution of the covariates associated with the treatment decision. Additionally, some of the treatment deciding factors may also affect the event of interest. Therefore, the treatment effect on the response gets confounded with the effect of the covariates on the response. Due to the presence of confounding bias, the standard regression analysis on the observed values of the covariate and the treatment leads to biased estimation of the treatment effects and fail to provide the basis of treatment comparison. The main challenge in comparing the treatment options in such settings arises from the fact that we only observe the response corresponding to the treatment option that a patient receives, but the response corresponding to the other treatment(s) stays unobserved. Nevertheless, a full comparison of the treatment options requires information on the potential outcomes that a patient would observe under each possible treatment option.

The main objective of this article is to provide basis for treatment comparison under such setting. Specifically, we follow two different approaches to predict the potential outcomes of the new patient under the available treatment options. In the first method (BayesG) which is based on the notion of Bayesian g-formula, we attempt to predict the potential outcomes by postulating a marginal structural model (MSM) which is a popular class of models in the causal inference literature (Robins, 1998a, 1999a). There are a handful number of articles on estimation of the causal parameters using semi-parametric methods such as inverse probability weighting method (IPWE) (Robins, 1998a, 2000). However, there are not many likelihood based methods primarily due to the fact that parametric models require more

assumptions. Instead of using a full parameterization, we use a likelihood based method within Bayesian semi-parametric setting to model the potential outcomes. In the second approach (PSBayes), we connect the treatment assignment mechanism to the missing data framework and try to predict the missing potential outcomes using a Bayesian semiparametric imputation model. Although, the basic framework of BayesG and PSBayes are different, they both follow the posterior predictive distribution of the new patient's potential outcome leading to obtain the optimal personalized treatment option.

1.4 Outline

In Chapter 2, we review literature on personalized medicine in the context of optimal treatment selection in single stage and in multiple stage settings. In Chapter 3 we discuss the literature on causal inference related to our research problem. To that direction, we introduce the potential outcome framework, discuss different aspects of causal inference in observational study. We also review literature on marginal structural model and the parameter estimation using inverse probability weighting and parametric g-formula. Next we connect the treatment assignment to the missing data framework and conclude this chapter after reviewing a few imputation methods from the missing data literature. In Chapter 4 we discuss the Bayesian parametric g-formula and propose our first method (BayesG) which is built on the extension of the Bayesian parametric g-formula. We detail the model specifications and discuss the optimal treatment selection in the context of it. In Chapter 5, we detail the treatment assignment mechanism and the PSPP method. Next we propose our second method (PSBayes) which is based on the PSPP framework along with a Dependent Dirichlet Process extension of the original specification. We also discuss the model feedback in the context of our proposed methodology. In chapter 6, we discuss five simulation scenarios

and compare the performance of our proposed methods with the inverse probability based methods. Finally in Chapter 7, we conclude with a short summary and discussions of some future research ideas.

CHAPTER 2

PERSONALIZED MEDICINE

2.1 Introduction

Personalized medicine is the tailoring of medical intervention to the patients' needs, characteristics as well as preferences. In traditional clinical trials to test for the drug efficacy, large number of patients are enrolled and are randomized to the drug arm or the placebo double-blindedly, without knowing whether the drug would work best for each individual. The traditional path of drug development which has conventionally influenced the practice of medicine has been based on identifying therapies which target an entire population. However, it has become apparent that most drugs that seem extremely promising in the lab, fail in the clinic. It is also common in medical history that patients with similar symptoms may have different diseases, with varying causes, or the same intervention may not work well for all patients with apparently the same disease. In other words, patients can show significant heterogeneity in response to treatments in many different disease areas. This heterogeneity in the patients' response to the same therapy prompted the researchers to make transition from the traditional 'one-size-fits-all' therapy to the evidence-based and data-driven personalized therapies. Personalized medicine, thus, emerges as a paradigm in medical science that aims to improve the quality of patients' health care by utilizing efficiently all the available information about patient demographics, genetic or genomic features, treatment and outcome history, ability to cope with side-effect burden etc. It also reduces the cost by reducing over-treatment. This allows patients to be treated and monitored more precisely

and effectively and thus providing ways that better meet their individual needs. Developing statistical methodologies in the arena of personalized medicine has brought new challenges that are often beyond the scope of traditional quantitative tools.

Improving clinical intervention involves finding an optimal treatment regime that would maximize the overall benefit leading to the most favorable outcome on average from the patient population. By treatment regime we mean the rule of assigning a treatment, available among the possible treatments in single stage or in multiple stages to the patients based on their observed characteristics, and thus personalizing the treatments to the patients. The two broad perspectives of the precision medicine are to develop and find the right treatment for a targeted subgroup that would very likely be benefited from it and to find the subgroup of patients that share common characteristics and would likely to benefit from the particular treatment. Also inherent is the problem of identifying biomarkers to identify such subgroups. In developing the treatment regime the clinicians make single point or multi-stage/sequence of decisions over the course of the disease. A single treatment decision is the simplest case where the personalized therapy mechanism takes the patients' characteristics as the input and dictates the rule that assigns the best treatment among the available options to the patients. In multi-stage therapies, decisions are made based on the accrued information of a patient and the next treatment action is determined from the possible options and hence making the clinical decisions evidence-based. Updating the treatment in multiple stages based on the accrued information from the previous stage constitutes the dynamic treatment regime (DTR). The data based on which the optimal treatment regime is estimated may come from both clinical trials or observational studies (Murphy, 2003; Robins et al., 2008). In observational setting, Zhang et al. (2012b) proposed a regression framework where semi-parametric methods like inverse probability weighted estimator (IPWE) or doubly robust augmented inverse probability weighted estimator (AIPWE) have been employed to estimate the average potential outcome of the population . In contrast to the well documented inverse-

probability weighting method, the augmented version is robust to misspecification of the outcome regression model or the propensity score model. Zhang et al. (2012a) proposed another direction to address the same problem by estimating the optimal treatment regime by Bayes classifier that minimizes the weighted expected misspecification error. However, the regression setting requires at least one model to be correctly specified which might be a stronger assumption in some situations. In order to avoid this limitation, powerful machine learning approaches have been proposed (Zhao et al., 2012; Moodie et al., 2012; Chakraborty and Moodie, 2013).

Personalized clinical practice is very common and necessary for chronic diseases which require long-term ongoing medical intervention. Examples of chronic diseases could be obesity, diabetes, alcohol and drug abuse, depression, HIV infection etc. For treating patients with chronic diseases, clinicians individualize the treatment type and also makes sequence of decisions following the chronic cure model (CCM) based on the patients' case history, to improve on the patients' outcome. Sequences of treatments are preferred for the long-term chronic diseases because of high inter-patient heterogeneity in response to treatment, likely relapse, presence or emergence of co-morbidities, time-varying side effect severity and reduction of cost and burden needed otherwise for unnecessary treatment. Thus DTR offers a path to implement the series of decisions made by the clinicians involved in the personalized practice.

2.2 Single-stage Decision Problems

As discussed in the introduction both single and multi-stage decision making problems arise in the personalized medicine. Decisions made by clinicians are one of the most important factors controlling the cost and quality of medical care. Medical decisions are the

acts that convert information into action. They help determine what prevention programs are promoted, what diagnoses are made, what tests are ordered, and what treatments are to be assigned to the patients. In the current era of cost containment and competition, it will be the success or failure of medical decision making that will determine the quality of medical practice. Decision theoretic approach has long been contributing in medicine or health care for providing a structure for gathering, organizing and integrating information that are relevant in the decision making. Statistically oriented works in this context include Lindley (1991), French (1986) and Parmigiani (2002). In this section we consider single-stage decision making in the context of personalized medicine.

A single stage regime is a mapping from the available information on the patients; that is, from the space of covariates/prognosticators, to the space of possible treatments/decisions, the best among which is to be assigned to the patients based on the rule. Suppose we consider a simple example where the space of decisions has two treatments $A = \{0, 1\}$ and the covariate space (say, $\mathcal{X}$) has two prognosticators $\mathbf{X} = (X_1, X_2)'$, where $\mathbf{X}$ could come from clinical or pre clinical information (e.g. X_1 could be age and X_2 could be WBC count). A decision rule or a single stage regime is a function $d(\mathbf{X})$, which can be defined as

$$d(\mathbf{X}) = \begin{cases} 1 & \text{if } X_1 < 50 \text{ and } X_2 < 10 \\ 0 & \text{otherwise} \end{cases}$$

If $d(\mathbf{X})$ takes the value 1 then the treatment $A = 1$ is assigned while $d(\mathbf{X}) = 0$ will assign the treatment $A = 0$ to the patients. Thus the regime can be expressed as $d(\mathbf{X}) = I(X_1 < 50 \text{ and } X_2 < 10)$. The rule may also involve a linear combination of the prognostic variables, e.g. $d(\mathbf{X}) = I(X_1 + 8\log(X_2) < 60)$. Chakraborty and Moodie (2013) formulated it using a decision theoretic framework. They conjectured that any decision is statistically evaluated in terms of the utility and the covariate space based on which the decision would be made.

Then for every treatment option a and covariates $\mathbf{X} \in \mathcal{X}$, the decision problem in terms of utility $U(X, a)$ can be expressed in the framework of opportunity loss function $L(\mathbf{X}, a)$ which is defined as

$$L(\mathbf{X}, a) = \sup_a U(\mathbf{X}, a) - U(\mathbf{X}, a), \tag{2.1}$$

For a given $\mathbf{X}$, the optimal treatment option $a^\star$, say, is obtained by minimizing the loss $L(\mathbf{X}, a)$. To have the optimal treatment in the entire population, we need to further minimize $L(\mathbf{X}, a)$ for all $\mathbf{X} \in \mathcal{X}$. The most common way of specifying the utility function is $U(\mathbf{X}, a) = E_a(Y|\mathbf{X})$, where Y is the primary outcome of the patients receiving treatment a with covariates $\mathbf{X}$.

Instead of the primary outcome Zhang et al. (2012b) considered the potential outcome framework. Let Y^a be the potential outcome for the treatment decision a i.e. if a patient receives treatment a his/her outcome is expected to be Y^a. More formally let us consider the average potential outcome of the patients for the treatment rule d or $d(\mathbf{X})$ to reflect that it is a function of $\mathbf{X}$ as $E(Y^d)$ which, under certain assumptions (please refer to section 3.1.4), can be written as

$$E[Y^d] = E[E\{Y^d|\mathbf{X}\}] = E[E\{Y|\mathbf{X}, A = 1\}d(\mathbf{X}) + E\{Y|\mathbf{X}, A = 0\}(1 - d(\mathbf{X}))] \tag{2.2}$$

Here $E(Y^d)$ is called the value function, also denoted by $V(d)$ which has been expressed in terms of the potential outcome. The assumptions of the causal inference that the covariate space contains all the information used for the treatment assignment are implicit. Assuming the higher outcome corresponds to a better treatment decision, the optimal treatment regime is obtained by maximizing the value function $V(d)$. Following Zhang et al. (2012b), it can be shown that the optimal regime, say $d^\star$ can be obtained by comparing the conditional mean response given a value of the covariate for the available treatment

options. Formally, $d^\star = I[E(Y|\mathbf{X} = \mathbf{x}; A = 1) > E(Y|\mathbf{X} = \mathbf{x}; A = 0)]$. Estimation of $d^\star$ for a given data requires the specification of $Q(\mathbf{x}, a) = E(Y|\mathbf{X} = \mathbf{x}; A = a)$. Depending on the outcome, linear or logistic regression model can be posited for $Q(\mathbf{x}, a)$ for estimating $d^\star$ and $V(d)$. However, this estimator is heavily dependent on correct specification of the model $Q(\mathbf{x}, a) = E(Y|\mathbf{X} = \mathbf{x}, A = a)$. In fact, in the setting of single stage regime Qian and Murphy (2011) proposed postulating a regression model and estimating the treatment decision from the model that would maximize the overall average outcome in the population. However, model misspecification may lead to a regime that might not contain the optimal treatment assignment rule that would maximize $V(d)$. To overcome sensitivity to the model misspecification, Zhang et al. (2012b) proposed a doubly robust augmented inverse probability weighted estimator (AIPWE). This approach takes into account possible confounding by including both propensity score model and outcome regression model in the estimator of $V(d)$ and because of the doubly robustness property the estimator remains consistent when either the model for treatment assignment mechanism or the model for the distribution of the response is correctly specified. In addition to the parametric or semi-parametric models, powerful statistical learning methods have also been proposed in the literature for handling data with high complexity in such settings. These are categorized into indirect methods and direct methods. Qian and Murphy (2011), Chakraborty and Moodie (2013), Moodie et al. (2014) and others have worked on the most popular indirect method, called Q-learning which is a two-step regression-based approach. Since the first step of Q-learning involves fitting a regression model (which could be both parametric or non-parametric learning approach) of the conditional mean of outcome, there might be problem of over-fitting leading to non-optimal ITR. Zhao et al. (2012) proposed a direct method in the framework of outcome weighted learning (O-learning). It considers estimation of the optimal treatment regimes from classification perspective. Instead of considering a 0-1 loss they considered a convex surrogate loss as is done in the support vector machine (SVM) via the hinge loss. Because

of formulating the estimation problem in the classification framework, the problem of prediction error associated with the conditional mean approximation using prediction model relating outcome, treatment and prognostic variables can be avoided and thus it provides a better approach to select the targeted therapy. Zhang et al. (2012a) proposed an approach of estimating the mean outcome using regression estimator, IPWE or AIPWE to identify the optimal treatment regimes by Bayes classifier that minimizes the expected weighted misclassification error.

In identifying the individualized treatment rules (ITRs) decision making becomes more accurate as more and more information becomes available on the patients. However, when all of the information do not turn out to be relevant in this context, variable selection becomes inevitably important to impose sparsity in the parameters, that enhance the decision making accuracy. In the existing literature several authors have worked on variable selection using penalized approaches in the context of discovering the ITR; for example, Qian and Murphy (2011) developed a two-stage procedure in the framework of Q-learning to estimate the optimal regimes using L1 penalized least squares where to justify their approach they find a finite sample upper bound of the mean response difference between the estimated ITR and optimal ITR. Other relevant works include penalized quadratic loss in the framework of A-learning proposed by Lu et al. (2013), variable selection methods for qualitative interactions, presenting two variable-ranking quantities by Gunter et al. (2011). However, since these penalization methods are primarily designed for linear regression models, they may not be appropriate for variable selection in treatment decision making under the misspecification of the true model. Motivated by this, Song et al. (2015) developed Penalized Outcome Weighted Learning (POWL) that do not depend on the parametric modeling of the value function and simultaneously estimate the optimal regime and incorporates sparsity in the model parameters retaining the computational advantages. Apart from these, a recent article (Jiang et al., 2017) considers survival time as the primary end-point and proposed two non-

parametric estimators for estimating the survival function of the patients following a given treatment regime, and derived estimation of the optimal treatment regime based on a value-based searching algorithm.

2.3 Dynamic Treatment Regime

Dynamic treatment regime (DTR) is a sequence of decisions that needs to be made at multiple stages through time in response to the patients' needs and demographics that vary over the course of the intervention. DTR is thus also known as treatment strategies (Thall et al., 2000), adaptive treatment strategy (Murphy, 2005; Lavori and Dawson, 2008), where the decision rules are made one per stage of intervention, that takes patients' treatment histories and response up to that stage as input and outputs the recommended treatment plan for the current stage. Thus DTR constitutes series of single stage regimes which are dependent among each other. DTR can be thought of as the key element of chronic care model (CCM) (Chakraborty and Moodie, 2013) where treatments need to be adaptively applied for treating the long term chronic diseases. Consider an example of DTR with two decision points. In the first stage the intervention is induction of chemotherapy which has two states $C = \{c_1, c_2\}$. Based on the response of the patients after stage 1, decision in the second stage is made. Those who responded are given the second stage intervention $M = \{m_1, m_2\}$ while those who did not respond after stage 1 are given the salvation therapy $S = \{s_1, s_2\}$. The decision in the second stage is made based on all accrued information available from stage 1. To put it in the mathematical notations, let the baseline covariates of a patient be denoted by $\mathbf{X}_1$ and hence at the baseline stage the accrued information of the patient is $h_1 = \mathbf{X}_1$. For illustration let us assume that $\mathbf{X}_1$ has the covariates age, WBC, gender, % CD56 expression. The decision rule in this stage is denoted by $d_1(\mathbf{X}_1)$

which outputs the treatment for the patient from the available options in C. Before going to stage 2, all information about the patient up to stage 1 is collected and thus the accrued information h_2 will be $h_2 = \{\mathbf{X}_1,$ chemotherapy at stage 1 $, \mathbf{X}_2\}$, where $\mathbf{X}_2$ involves the additional information after stage 1, for example, hematologic adverse event, post-induction ECOG status, WBC, platelets, and the responder status. Thus in stage 2, 4 treatment options are available, 2 for responder and 2 for non-responder. The decision rule in stage 2 is $d_2(h_2)$ which inputs h_2 and outputs the treatment from available options in M for responder and that from the available choices in S for the non-responder. Thus, the DTR, in this case, is $d = \{d_1(h_1), d_2(h_2)\}$. For a multi-stage decision problem with K stages where decision needs to be made, the sequence of observations/information along with the treatment selection rule is given by $\{O_1, a_1, O_2(a_1), a_2, O_3(\bar{a}_2), \ldots, O_j(\bar{a}_{j-1}), \ldots, O_{K+1}(\bar{a}_K)\}$, where a_j denotes the action space at stage j, $O_j(\bar{a}_{j-1})$ denotes the observations made at stage j corresponding to the treatment sequence $\bar{a}_{j-1}$ and $\bar{a}_j = \{a_1, a_2, \ldots, a_j\}$ denotes the treatment sequence up to stage j. The DTR will be given by $d = \{\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_K\}$, where $\bar{d}_j$ is the rule/mapping from the observation space $\{O_1, a_1, O_2(a_1), a_2, O_3(\bar{a}_2), \ldots, O_j(\bar{a}_{j-1})\}$ to the action a_j at stage j.

The literature review on DTR reveals that Q-learning and A-learning are the two main approaches for estimating optimal DTR. In Q-learning regression model is assumed at each decision point while in A-learning regression model is posited only for the contrasts among treatments and for treatment assignment at each decision point (Zhang et al., 2013). However, the performance of both of these approaches heavily depend on the correctness of the specified models. Thus to circumvent this limitation Zhang et al. (2013) extended the approach of doubly robust estimation of the optimal DTR from single stage decision to two and any finite number of decision stages. This approach not only protects the estimation from model misspecification but also gives comparable or even superior performance compared to the Q- or A- learning methods. Zhao et al. (2015) proposed backward out-

come weighted learning (BOWL) and simultaneous outcome weighted learning (SOWL) for optimizing DTR from classification perspective. Since these approaches are based on non-parametric estimator of the expected long-term outcome, misspecification of the models is not a concern. In BOWL, the estimation of optimal DTR is reformulated as a sequence of weighted classification problems whereas in SOWL reduces the optimal DTR estimation to a single classification problem. As stated in Zhao et al. (2015), a convenient way to formalize the problem in finding optimal DTRs is through potential outcomes (Rubin, 1974, 1978; Robins, 1986). Relevant works include Lavori and Dawson (2004) who used multiple imputation to estimate all potential outcomes, to compare the adaptive regimes using the imputed outcomes, structural model to estimate the mean response that would have been observed if the whole population followed a particular DTR (Murphy et al., 2001). The model considered is for the marginal mean of counter-factual or potential responses in observational data for estimation of the mean response to both nonrandom and random dynamic treatment regimes.

CHAPTER 3

CAUSAL INFERENCE

Causal inference is an important tool for identifying and analyzing the impact of a treatment or an exposure on the consequent outcome. On the other hand, the causal inference is frequently used in missing data analysis which has a close analogy with treatment assignment mechanism in observational studies. In this chapter, we discuss the basic notions of causal inference and missing data and draw their connections to our research problem. Part of the discussion in causal inference is covered in Hernan and Robins (2010), and Imbens and Rubin (2015).

3.1 Potential Outcome Framework

3.1.1 Notation

Consider the research problem we focus in this dissertation. The primary question is that how can we choose the best treatment option for a new patient where the data comes from an observational study? Note that, here we are addressing the problem from a physician's point of view. The way we are thinking is that, suppose, the physician has the ability to guess the possible outcomes under all treatment options. That means, the physician can foresee, if the patient had been assigned to a particular treatment option, then what his/her outcome would be or to put it in other way what would be the patient's potential outcome under some pre-specified treatment option. Then it would be easier for him/her to administer the best

option accordingly. To this direction, we first introduce the potential outcome framework and then mathematically formulate the causal question in our research.

Although there are alternative frameworks suggested in literature (see (Dawid, 2000)) we mainly follow the potential outcome framework as discussed in Rubin (1974, 1978). Suppose there is a treatment variable A, a set of potential confounders $\mathbf{X}$ and a continuous outcome variable Y. The underlying notion in the potential outcome framework presumes that even though a subject (at a particular time) was exposed to a particular treatment, the same subject (at that particular time) can be exposed to another treatment(Imbens and Rubin, 2015). For example, if a person takes aspirin to relieve from headache, the same person could have chosen not to take aspirin or chosen to take another medication to fulfill the same objective. This assumption, known as the positivity assumption will be formally stated in next section. In practice, the treatment or exposure variable can be continuous or binary or categorical. For the simplicity, we are restricting ourselves to consider only the binary exposure; $A = 0$ denoting control and $A = 1$ denoting treatment. Let Y^0 denote the outcome variable that would have been observed under the treatment $A = 0$ and Y^1 denote the outcome variable that would have been observed under the treatment $A = 1$.

The variable Y^0 is referred to as the potential outcome under treatment $A = 0$, whereas the variable Y^1 is termed as the potential outcome under treatment $A = 1$. The terminology "potential outcome" is used because depending on the treatment the subject has received one of the outcomes would potentially be observed. Some authors prefer to use "counterfactual outcome" to emphasize the fact that one outcome represents the situation that has actually occurred (factual) whereas the other outcome does not (counterfactual). For each subject, only one of the potential outcomes (the one for which the treatment was actually received) has been factual. Thus for the subject who has received the treatment $A = a$ the potential outcome Y^a has been observed and hence the observed outcome Y equals the potential

outcome Y^a. This statement is formally conjectured as the consistency assumption in the next section.

3.1.2 Observational Study

Causal inference serves as a central toolbox for many research studies stemming from the health, social and behavioral sciences. Most of the research questions in these fields are motivated by causal questions such as whether a new intervention works or not, whether a new policy succeeds to reduce the crime rate in a city or if an exposure is harmful or not.

The ideal scenario to answer such research questions is to conduct a fully randomized controlled prospective experiment in which treatment or exposures are randomly (blinded) assigned, volunteers strictly follow the guidelines and the relevant data are collected without error. Then assuming the data we use for the analysis comes from the population we want to infer about, the comparison between the treated and untreated group yields a causal interpretation.

However, such ideal experiments might be challenging to conduct in real world applications. Moreover, in many therapeutic areas conducting a randomized trial might not be ethical or practical. These situations motivate us to rely on observational studies to reach the goal. Performing causal inference in observational setting requires formulating the causal question, identifying the exposure, potential confounder and the outcome variable and stating the assumptions to make the connection between the observed data and the pseudo data which is a part of the analysis. Although some assumptions are unverifiable in observational setting, one could enrich the analysis by performing a sensitivity analysis to check the robustness of the method to validation of those assumptions.

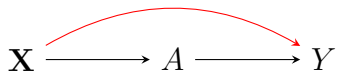

Figure 3.1: Causal diagram

3.1.3 Confounding

One major goal of conducting an observational study is to estimate the treatment causal effect on the outcome. However, in disease area such as cancer, interventions are given based on patient characteristics. Due to the nonrandomized assignment mechanism, there is a high possibility that the treatment groups differ in distribution of patient characteristics associated with the treatment benefit and the observed effect of the intervention is actually the effect of a certain characteristic which makes it difficult to separate out these two effects using standard regression method. This phenomenon is known as confounding and in observational studies it is responsible for inducing bias in the estimation if we employ standard regression approach to estimate treatment effect.

In observational studies confounding can happen in different ways depending on the situations. Let us focus on the problem we are considering in this dissertation. This is an observational study where we assume to have access to the medical information of a group of n patients. The patients' clinical history, biological profile (say $\mathbf{X}$) have influenced doctors to prescribe a certain treatment option (say A). Apart from the patients' covariate and treatment information we also have records on their outcomes Y. Now the outcomes we observe may depend on both treatment and covariate information of the subjects. The important question is how to separate out the effect of the treatment from the effect of the covariates on the outcome.

This situation can be illustrated by figure 3.1. In this scenario, there are two paths through which the treatment A and the outcome Y share information: (1) The direct causal

path $A \rightarrow Y$ and (2) the indirect path $A \leftarrow \mathbf{X} \rightarrow Y$ between A and Y through $\mathbf{X}$. Among these two paths, if the partial path $\mathbf{X} \rightarrow A$ can somehow be removed from path (2), the settings would be similar to a standard regression setting where we would have two independent regressors $\mathbf{X}$ and A affecting the outcome Y and we could separate out the treatment (A) effect from the effect of covariates ($\mathbf{X}$) on the outcome. Under this circumstance, standard regression method would give us an unbiased estimate of the treatment causal effect, because in the absence of confounding the association would be same as causation and the associational measures would be sufficient to provide information on the causal measures. But, the critical question is that under what conditions we actually can ignore the $\mathbf{X} \rightarrow A$ path?

3.1.4 Assumptions

When an extraneous covariate relates to both treatment and response variable, the treatment effect gets confounded by the effect of that covariate on the response. To remove the confounding bias, researchers often use randomized experiments where the confounders are controlled by stratification. One crucial advantage of using randomized experiment is that it produces exchangeability in the sense that the treated and untreated individuals are exchangeable i.e. the potential outcome distribution of treated and untreated individuals are same. Mathematically,

$$Y^a \perp\!\!\!\perp A \quad \text{for all } a.$$

In stratified or conditional randomization one constructs stratified samples and then implement randomization in each group which produces conditional exchangebility within each strata. For example, if we construct strata depending on the values of the confounders $\mathbf{X}$

then randomizing patients within each strata produces same potential outcome distribution for treated and untreated individuals. In formal notations it implies

$$Y^a \perp\!\!\!\perp A|\mathbf{X} \text{ for all } a. \tag{3.1}$$

In conditional randomized experiment one can easily estimate causal effect using Inverse Probability Weighting (IPWE) (detailed in subsection 3.2.3) or parametric g-formula (detailed in subsection 3.2.4) with the help of conditional exchangeability. In observational study we can do the same if we make the following assumptions. Without assuming these conditions the observational data alone is insufficient to identify causal effect (Hernan and Robins, 2010).

Consistency. This assumption states that an individual's observed outcome is actually the potential outcome under the treatment that individual has actually received. That means, if a person receives treatment a then his potential outcome under treatment a would be his observed outcome. Assuming $A \in \{0, 1\}$ this assumption could be formally written as

$$Y = Y^1 A + Y^0(1 - A). \tag{3.2}$$

No Unmeasured Confounder. This assumption is equivalent to conditional exchangeability or sequential randomization (3.1) (Robins, 1999a). Under this assumption within levels of the observed confounder $\mathbf{X}$, all other predictors of outcome are assumed to be equally distributed between the treated and the untreated groups (Hernan and Robins, 2010). Although this assumption usually holds in RCT, it can never be verified in observational studies (analogous to Missing at Random (MAR) assumption) due to the fact that, as the investigator is not in control of the experiment, one can never be assured of whether $\mathbf{X}$ is the only factor that is unequally distributed between the treated and untreated or there is/are

some "unmeasured" confounder(s) (Hernan and Robins, 2010). Although this assumption is not testable, a sensitivity analysis is often performed to investigate the effect of potential unmeasured confounder on the causal parameter estimation (Robins, 1999a,b; Hernán et al., 2000).

Positivity. Sometimes the interest might be in estimating the average causal effect over the population. For this purpose, investigators would need some patients for each treatment group to make the treatment effect contrasts become estimable. This condition, referred to as "positivity" is very important for estimating the average causal effect over a population or a subpopulation. Formally this assumption states that there is a positive probability of receiving each of the treatment level for every combination of treatment and confounder (Cole and Hernán, 2008). In our setup it implies

$$P(A = a|\mathbf{X} = \mathbf{x}) > 0 \quad \text{for all } a \text{ and } \mathbf{x} \tag{3.3}$$

This assumption is also known as experimental treatment assumption because each treatment level is assumed to be received by atleast one subject (Cole and Hernán, 2008). This assumption is violated if some treatments are not assigned to any individual in a study which would yield biased causal estimates (Hernán et al., 2004). Robins et al. (2000) suggested the use of structural nested model in such situation.

3.2 Marginal Structural Model

3.2.1 Introduction

The marginal structural model (MSM) (Robins, 1999a, 2000) was introduced to estimate the average causal effect in a population level. To take an example, consider the following simple linear MSM for the potential outcome under treatment a

$$E\left[Y^a\right] = \alpha_0 + \alpha_1 a. \tag{3.4}$$

This model is different from the usual regression model in the sense that for any subject the quantity in the left hand side is not observed for every value a and hence this model cannot be directly fit to the observed data. Also, note that this model does not have any covariate and is referred as the unconditional or marginal mean model. The most important advantage of using this model is that here the parameters are directly related to the treatment causal effect. For example, $\alpha_1 = E\left[Y^{a=1}\right] - E\left[Y^{a=0}\right]$ and so estimating the parameter α_1 would lead us to obtain the average causal effect in the population.

Suppose, there is a set of covariates or confounders (say $\mathbf{V}$) for which the causal effect of treatment A on the outcome Y varies as the value of $\mathbf{V}$ changes and we are interested in finding how the treatment causal effect varies for different values or levels of $\mathbf{V}$. In such cases, $\mathbf{V}$ is called the effect modifier. When the interest is in estimation of the marginalized causal effect over the whole population, MSMs usually do not include covariates in the model. However, if the interest is in estimating the causal effect for a subgroup of patients (assuming the factor defining the subgroup is an effect modifier) we usually include effect modifier(s) in an interaction model.

The general form of MSM is commensurate with that of generalized linear model (GLM). Like in GLM we can similarly express MSM as

$$E\left[Y^a|\mathbf{V};\boldsymbol{\alpha}\right] = f(A,\mathbf{V},\boldsymbol{\alpha}). \tag{3.5}$$

where, $f(\cdot)$ is a known function of treatment, effect modifier and the parameter related to the structural model. For example, if we consider Y to be binary (say, it can only take values 0 and 1), then the expression above looks like

$$E\left[Y^a|\mathbf{V};\boldsymbol{\alpha}\right] = P\left(Y^a = 1|\mathbf{V};\boldsymbol{\alpha}\right) = \text{logit}^{-1}\left\{f_0(\mathbf{V};\alpha_0) + f_1(A,\mathbf{V};\alpha_1)\right\}.$$

where, $\boldsymbol{\alpha} = (\alpha_0,\alpha_1)'$, α_0 is for the main causal effect of treatment and α_1 measure the change of treatment causal effect as the value of effect modifier changes by one unit. These two parameters often are of interest in causal literature.

3.2.2 Literature Review

Marginal structural models (Robins, 1998a, 2000) represent a class of causal models which can be used to efficiently estimate the treatment causal effect in a time-fixed or time-varying setting. This model was initially proposed as an alternative to the semiparametric g-computation algorithm estimator (Robins, 1986) and to g-estimation of set of parametric counterfactual models, known as structural nested models or SNMs (see Robins (1994, 1997, 1998b,c) for details). In the longitudinal setting with time-dependent treatments and confounders Robins (2000) discussed two major advantages of MSMs over SNMs. First, in case of binary outcome, MSMs can be applied on logistic models to provide semiparametric estimate of the treatment causal effects. Second advantage lies in the application with sur-

vival data in the absence of time dependent covariates. In these situations, time-dependent Cox proportional hazard (PH) model is usually implemented to estimate time dependent treatment effect. MSMs can be easily extended to the time-dependent proportional hazards model.

MSM was originally introduced to provide a valid estimate of the causal effect of a time varying exposure in the presence of time varying confounders (Robins, 1999a). In observational studies, the causal effect can be well estimated assuming all confounders are measured (time dependent or time independent) (Hernán et al., 2000). The standard approach to estimate treatment effect in such setting was to model the current outcome as a function of past treatment and confounder history. However, Robins (1986) proved that the standard approach to control confounding is biased in the presence of complex confounding where the covariates predict subsequent outcome and treatment history and are themselves influenced by past treatment. This finding holds true even though we assume that we observe every confounder which needs to be controlled (Hernán et al., 2000). Under the sequential randomization (also known as exchangeability or no unmeasured confounders) assumption (3.1) the parameters of MSM can be consistently estimated by performing a weighted least square analysis. This method known as inverse probability weighting (Robins, 1998a,b, 2000; Robins et al., 2000) will be detailed in the next subsection.

3.2.3 Inverse Probability Weighting

Let us recall the confounding scenario as illustrated in figure 3.1. The set of characteristics $\mathbf{X}$ of a patient influence the physicians to assign a treatment A. Apart from the covariate $\mathbf{X}$ and the assigned treatment a, the outcome Y is also observed. Suppose, we are interested in estimating the effect of A on Y in this situation. The parameters in the marginal structural

model (3.4) directly correspond to the treatment causal effects. However, fitting this model would require to know the potential outcome under each treatment option for each patient. Clearly, it is not possible since for each patient one of the potential outcomes is counter factual.

A crude analysis in this situation will be to posit the following associational model on the outcome based on treatment

$$E[Y|A] = \alpha_0' + \alpha_1' A \tag{3.6}$$

The associational treatment effect is measured by the estimate of α_1'. The parameters of the associational model (3.6) will be different from that of the causal model (3.4) unless the treatment is not associated with any predictor of the outcome i.e. unless we can ignore the causal path $\mathbf{X} \rightarrow A$ in figure 3.1. If the treatment is unconfounded, the estimate of α_1' would be unbiased for the causal treatment parameter α_1.

In observational study the assumption of unconfounded treatment is not valid and therefore the associational parameter would not be same as the causal parameter. Under the assumption of 'no unmeasured confounders' (3.1) Robins (1998a,b, 2000); Robins et al. (2000) developed a weighted analysis to consistently estimate the causal parameter where the weight of i^{th} subject with treatment option a_i has the following form

$$W_i = \frac{1}{P[A_i = a_i|\mathbf{X}_i]}.$$

The denominator of the weight W_i is the probability that a patient receives his/her own observed treatment, also known as the propensity score. For randomized experiments the propensity score is known to the experimenter, whereas in observational studies the propen-

sity score is generally not known and can be estimated by postulating a logistic regression model of the treatment (A_i) on the observed set of confounders ($\mathbf{X}_i$)

$$\text{logit}\{P(A_i = 1|\mathbf{X}_i, \boldsymbol{\gamma})\} = \mathbf{X}_i'\boldsymbol{\gamma}$$

One can obtain the estimated weight ($\hat{W}_i$) in terms of $\hat{\boldsymbol{\gamma}}$ by fitting this model to the observed data. Thus i^{th} patient with $A_i = 0$ would be assigned the weight of $\hat{W}_i = 1 + \exp(\mathbf{X}_i'\hat{\boldsymbol{\gamma}})$, whereas if i^{th} patient receives $A_i = 1$ the corresponding weight would be $\hat{W}_i = 1 + \exp(-\mathbf{X}_i'\hat{\boldsymbol{\gamma}})$. Once, the estimated weights are obtained we can fit the association model (3.6) on the weighted data. This method is known as inverse probability (IP) weighting and the weighted estimators are referred to as IPWE estimators (Robins et al., 2000).

IP weighting creates a pseudo population consisting of W_i copy of i^{th} patient. Under the assumption that we have measured all confounders, the pseudo population replicates the situation of an experimental study where the treatments are not confounded. Therefore the causal parameters can be consistently estimated by the associational model (3.6) parameter estimates based on the pseudo population data.

IPWE estimation is sensitive to the weights assigned to each patient. If W_i has extreme values (large or small), IPWE leads to inefficient estimation. Robins et al. (2000) recommended to use stabilized weight $SW_i = \frac{P(A_i=a_i)}{P(A_i=a_i|\mathbf{X}_i)}$ in these situations. The mean of stabilized weights is expected to be 1 as the size of the pseudo population equals the study population (Hernan and Robins, 2010).

IPWE estimators was first introduced as a class of semi-parametric estimators for the causal parameters of marginal structural models in time-varying setting (Robins, 1998a,b, 2000; Robins et al., 2000). These articles compared IPWE with previously proposed parametric structural nested models and discussed their advantages and disadvantages. Robins et al. (2000) explained the idea in a simplified point-treatment setting, multilevel treatment

and unsaturated MSMs for binary outcome. Hernán et al. (2000) introduced IPWE estimation for survival outcome and showed the advantage of MSM along with IPWE estimation in the presence of time-dependent risk factor for survival and in the situation when the past treatment history predicts subsequent risk factor level. Although the two stage estimation of IPWE method is well justified in frequentist semi-parametric theory, there are a very few articles concerning IPWE method or propensity score adjustment in the Bayesian paradigm. The primary reason lies in the fact that Bayesian inference is generally based on the full model specification, while IPWE is not derived from likelihood based model. Despite this fact, there have been a few attempts to introduce Bayesian version of IPWE in the causal inference literature (Hoshino, 2008; Kaplan and Chen, 2012; Saarela et al., 2015). Among them Saarela et al. (2015) proposed a Bayesian formulation of IPWE without specifying the MSM in the Bayesian framework. However, as Robins et al. (2015) pointed out, their approach may not be claimed as fully Bayesian as a propensity score model is formulated in the design stage and the model parameters being independent of outcome model parameters, a Bayesian inference cannot be a function of the propensity score. The limitation of casting IPWE to Bayesian framework motivates us to consider an alternative standardization method, known as parametric g-formula as an alternative to IP weighting which will be discussed in next subsection.

3.2.4 Parametric g-formula

Parametric g-formula (Robins, 1986) provides an alternative approach to IP weighting method that can be used to estimate the causal parameter in MSM. Similar to IP weighting method, g-formula can also appropriately adjust for measured confounders in longitudinal setting in presence of time dependent confounders and treatments (Young et al., 2011). This

method was first introduced by Robins (1986) to investigate the causal effect of arsenic on heart disease in an occupationally exposed cohort. Afterwards, it has been applied few times in causal inference literature (Robins et al., 2004; Young et al., 2011; Keil et al., 2014).

The motivation behind this method comes from the fact that the average potential outcome over the whole population can be thought of a weighted average of the stratum specific average of the potential outcomes. More formally, if Y^a is the potential outcome under treatment a then the average potential outcome for the whole population $E[Y^a]$ can be expressed as the weighted average

$$E[Y^a] = \int_{\mathbf{x}} E[Y^a|\mathbf{X}]f(\mathbf{X})d\mathbf{X}$$

where the weight comes from the distribution of the confounders. Moreover, if we assume that we measure all the confounders of the treatment effect ('no unmeasured confounders' or exchangeability (3.1)) and the observed outcome is actually the potential outcome under the treatment a patient receives (consistency (3.2)) the average potential outcome for treatment a can be expressed as

$$\begin{aligned} E[Y^a] &= \int_{\mathbf{X}} E[Y^a|\mathbf{X}]f(\mathbf{X})d\mathbf{X} \\ &= \int_{\mathbf{X}} E[Y^a|\mathbf{X}, A=a]f(\mathbf{X})d\mathbf{X} \qquad \text{[exchangeability]} \\ &= \int_{\mathbf{X}} E[Y|\mathbf{X}, A=a]f(\mathbf{X})d\mathbf{X}. \qquad \text{[consistency]} \end{aligned} \tag{3.7}$$

In ideal settings such as when $\mathbf{X}$ is categorical or low dimensional, one could estimate the conditional mean of the response $E[Y|\mathbf{X}, A=a]$ nonparametrically by dividing the data of the treatment group a into some strata based on different combinations of values of $\mathbf{X}$ and taking average outcome for each strata. However, the nonparametric estimation is not possible even if we have moderately high number of confounders with different levels and

is completely out of the question if we have continuous confounders. In these situations, we have no other option but modeling the conditional response. The parametric g-formula extends the idea of g-formula in a parametric setting where the conditional components are characterized by some parametric models. Specifically, we can reformulate the average potential outcome as

$$E[Y^a|\boldsymbol{\theta}] = \int_{\mathbf{X}} E[Y|\mathbf{X}, A = a, \boldsymbol{\theta}_1] f(\mathbf{X}, \boldsymbol{\theta}_2) d\mathbf{X}$$

where $\boldsymbol{\theta}_1$ is the set of parameters representing the conditional distribution of $Y|\mathbf{X}, A = a$, $\boldsymbol{\theta}_2$ is the set of parameters in the marginal distribution of $\mathbf{X}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$.

Under this formulation, the estimation of causal effect is again a weighted analysis similar to IP weighting method. Here the only difference is the weight, which comes from the marginal distribution $f(\mathbf{X})$ of the confounder, can be directly incorporated in the associational model for the outcome surface and thus it is a one step procedure. The equivalence between parametric g-formula and IP weighting is proved in literature (please refer to Technical Point 2.3 in (Hernan and Robins, 2010) for details). However, estimators based on parametric g-formula are more efficient (smaller variance) than the estimators based on IP weighting possibly due to its dependence on the parametric framework (Young et al., 2011).

Another advantage of parametric g-formula is that, unlike IP weighting, parametric g-formula is based on likelihood based method. Motivated by this fact Keil et al. (2015) introduced a Bayesian framework for parametric g-formula in time-fixed data and discussed its advantage over IP weighting in the context of Bayesian formulation of counterfactual distributions. Bayesian IPWE methods (Saarela et al., 2015) require the prior information from structural outcome models. However, in certain applications, informative prior often comes in the form of conditional, non-structural model parameters (Keil et al., 2015) and so this prior information cannot be fully utilized in Bayesian IPWE setting.

3.3 The Causal Question

Depending on the area of concern, the causal question might be at individual level or at population level. The difference between 'individual' and 'population' requires different tools to answer the questions. For example, when a new drug comes to the market, the healthcare policy regulators are concerned with whether prescribing this drug to a population in certain therapeutic area would increase the frequency of desired outcome keeping the frequency of adverse events within a certain boundary. On the other hand, a physician is concerned with whether recommending this drug to his/her patient would alleviate the cause of the patient's visit. To answer the first scenario one has to estimate the average causal effect of the new drug in the population. In second scenario the interest is in the individual patient level. Typically, in this situation, analysis and inference are done with the observed set of patients under the assumption that the set of patients are sampled from the same population where the new patient is coming from. Then, the result of the analysis is used to predict whether the new patient would respond to the drug or not.

Recently, there have been a few attempts to introduce tailored therapeutics where the authors are interested in finding the best treatment option not for an individual but for a population or a group of patients. In this spirit, Zhang et al. (2012b,a) used potential outcome framework to define the "best" treatment regime (in single stage decision problems the regime is defined by a classification rule) so that if the "best" regime is applied to the whole population that would yield the highest reward (best outcome, assuming larger value is better) on average. In formal notations, if a patient with baseline information $\mathbf{X}$ receives treatment according to the regime d then the potential outcome corresponding to the regime d would be given by

$$Y^d = Y^1 d(\mathbf{X}) + Y^0(1 - d(\mathbf{X})).$$

The optimal regime for the whole population is defined by

$$d^\star = \underset{d}{\text{argmax}}\ E[Y^d],$$

where, $E[Y^d] = E_{\mathbf{X}}[E_{Y^d|\mathbf{X}}[Y^d|\mathbf{X}]]$ is the average value of the regime over the whole population. Keil et al. (2015) considered a Bayesian framework of the parametric g-formula and applied it in the context of obtaining the posterior predictive distribution of potential outcome for a group of patients under some pre-defined regime both in time-fixed and time-varying setting.

In contrast to these methods for a group of patients in a population, we focus in the comparison of different treatment options with respect to a physician or a patient's point of view. The inherent research question lies on the notion of individual causal effect which is defined based on the contrast of potential outcomes of that specific individual. However, as we have discussed earlier, for each patient we only observe the factual outcome corresponding to the treatment (s)he has received and so for estimating individual causal effect for a new patient we must have to rely on potential predictive distribution. Motivated by this line of thinking, our goal is to predict the mean potential outcome for each treatment option. Then assuming higher value of outcome is better, the optimal treatment option would be the one corresponding to the highest average outcome value. In the next two chapters, we will introduce the mathematical formulation of the proposed methods.

3.4 Connection to Missing Data

There is a strong connection between the treatment assignment and the missing data mechanism. Note that, in our setting for a single patient, we do not observe the outcome simultaneously for both treatment options. Rather we only have information on the observed

outcome which corresponds to a treatment that the patient has been administered. Under the consistency assumption (3.2), the observed outcome is the potential outcome under the given treatment. It implies that we only observe one potential outcome for each patient while the other potential outcome corresponding to the treatment that (s)he has not received remains unobserved which raises the analogy between treatment assignment and missing data mechanism. In this section we briefly review some of the missing data methods that we will use in the later chapters.

3.4.1 Missing Data Mechanism

Let $\mathbf{D}_F$ be the full dataset containing all relevant data; observed or unobserved. The unobserved values of a dataset can be characterized by the missing data pattern. Adhering to the notations commonly used in the missing data literature, let the variable R denote the response indicator such that $R = 1$ corresponds to an observed value and $R = 0$ implies the corresponding value is missing. $\mathbf{R}$ denotes the response indicator matrix. In this dissertation, we assume that the missingness can only occur at the outcome variables. Accordingly, the dimension of $\mathbf{R}$ only corresponds to the potential outcomes in $\mathbf{D}_F$. Given the response indicator, we can partition the full dataset $\mathbf{D}_F$ into two parts - one part $\mathbf{D}_F^o$ where the data is fully observed and the other part $\mathbf{D}_F^m$ with missing observations. The missing data mechanism i.e. the relationship between missingness and the variables plays a crucial role in estimating marginal or conditional mean of the variable containing missing values. Rubin (1976) considered R as a stochastic variable and described the missing data mechanism by the conditional distribution of $\mathbf{R}$ given $\mathbf{D}_F$, $f(\mathbf{R}|\mathbf{D}_F, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ represents the unknown relevant parameters. The missing data mechanism can be classified into three categories:

(i) Missing completely at random (MCAR): When missingness does not depend on any data, observed or unobserved

$$f(\mathbf{R}|\mathbf{D}_F, \boldsymbol{\theta}) = f(\mathbf{R}|\boldsymbol{\theta}) \text{ for all } \mathbf{D}_F, \boldsymbol{\theta}$$

(ii) Missing at random (MAR): When missingness only depends on the observed part of the data $\mathbf{D}_F^o$

$$f(\mathbf{R}|\mathbf{D}_F, \boldsymbol{\theta}) = f(\mathbf{R}|\mathbf{D}_F^o, \boldsymbol{\theta}) \text{ for all } \mathbf{D}_F^o, \boldsymbol{\theta}$$

(iii) Missing not at random (NMAR): When missingness depends on both the observed and unobserved part of the data

$$f(\mathbf{R}|\mathbf{D}_F, \boldsymbol{\theta}) = f(\mathbf{R}|\mathbf{D}_F^o, \mathbf{D}_F^m, \boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}$$

The MCAR mechanism is a special case of the MAR mechanism. In practice, the MAR mechanism is often employed by collecting data on covariates related to nonrespondents.

In our setting, if we perceive the full data $\mathbf{D}_F$ with n observations and p predictors as $\mathbf{D}_F = \left(\mathbf{Y}^1_{n\times 1}, \mathbf{Y}^0_{n\times 1}, \mathbf{A}_{n\times 1}, \mathbf{X}_{n\times p}\right)$ then the response indicator matrix $\mathbf{R}$ would be completely specified by the observed treatment indicator $\mathbf{A}$ as it can be assumed as $\mathbf{R} = (\mathbf{A}_{n\times 1}, \mathbf{1}_{n\times 1} - \mathbf{A}_{n\times 1})$. The observed data would be $\mathbf{D}_F^o = \mathbf{D} = (\mathbf{Y}_{n\times 1}, \mathbf{A}_{n\times 1}, \mathbf{X}_{n\times p})$ where $\mathbf{Y}_{n\times 1}$ is the vector of observed outcomes and the missing data would consist of the vector of missing counterfactual outcomes i.e. $\mathbf{D}_F^m = \mathbf{Y}^{\text{miss}}_{n\times 1}$. Also, note that, the response indicators for $\mathbf{Y}^1_{n\times 1}$ and for $\mathbf{Y}^0_{n\times 1}$ sum to $\mathbf{1}_{n\times 1}$ implying that two potential outcomes cannot be simultaneously observed which essentially reiterates the fundamental problem of causal inference.

3.4.2 Imputation Methods

Many methods have been proposed in the missing data literature for imputing the missing observations where the end goal is to reduce bias in estimation of the mean response due to missingness. In an oversimplified approach complete case analysis and weighted complete case analysis (Cochran, 1968; Little, 1986; Horvitz and Thompson, 1952) ignore subjects with missing values and the loss of information from these patients result in less efficient estimator of the mean of the concerned parameter involving missingness. A better alternative could be to assume the following linear regression model for the missing observations

$$Y = \mathbf{X}'\boldsymbol{\beta} + \epsilon$$

where, $\mathbf{X}$, say has p predictors $X_1, X_2, \ldots, X_p$ and $\boldsymbol{\beta}$ are the corresponding regression coefficients. ϵ is the error term with $\epsilon \sim N(0, \sigma^2)$. Suppose, there are r observed cases and $n - r$ missing cases. Then for imputing the i^{th} missing observation for $i = (r+1), \ldots, n$, one can estimate the regression coefficients $\boldsymbol{\beta}$ by MLE based on the complete cases and then predict the missing observation $\hat{Y}_i$ by $\hat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$. After predicting the missing observations the marginal mean of the response μ can be estimated by $\hat{Y} = n^{-1}\left(\sum_{i=1}^{r} Y_i + \sum_{i=r+1}^{n} \hat{Y}_i\right)$.

Gelman et al. (1995) explored Bayesian version of this methodology where they imputed the missing response values from the posterior distribution multiple times and combined the mean values from each imputed set to estimate the marginal mean. Assuming the correct model specification, these parametric methods are effective in correctly estimating μ under MAR mechanism (Zhang and Little, 2009). However, they are prone to produce biased estimates under model misspecification.

There are several ways to achieve robustness in the model specification. In one such attempt (linear in weight prediction (LWP) (Scharfstein et al., 1999; Bang and Robins, 2005)) the inverse of the estimated propensity score is included as a linear weight term in the imputation model as follows,

$$(Y|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \sim N(g(\mathbf{X}, \boldsymbol{\beta}) + \alpha\hat{w}(\mathbf{X}, \boldsymbol{\gamma}), \sigma^2)$$

where, $\hat{w}(\mathbf{X}, \boldsymbol{\gamma}) = 1/\hat{P}(R = 1|\mathbf{X}, \hat{\boldsymbol{\gamma}})$ is the inverse of the estimated propensity score to response for the observed cases. For the cases where Y is missing, $\hat{w}(\mathbf{X}, \boldsymbol{\gamma}) = 1/\left(1 - \hat{P}(R = 1|\mathbf{X}, \hat{\boldsymbol{\gamma}})\right)$. Propensity weighting is a very popular tool in the missing data literature. However, one potential limitation of this approach is that it may cause large variance in estimation as it assigns huge weights to the respondents with very small propensity scores which may result in out-of-range estimates for the means (Little and Rubin, 2014).

Robustness can also be achieved by employing semiparametric and nonparametric methods which relaxes the model assumption by capturing nonlinear relationship between variables through a flexible mean function. To this direction, for a single covariate X, one can build an imputation model based on penalized spline framework (Eilers and Marx, 1996; Berry et al., 2002) $Y|X, \beta \sim N(s(X), \sigma^2)$ with a truncated polynomial basis function

$$s(X) = \sum_{j=0}^{q} \beta_j X^j + \sum_{k=1}^{K} \beta_{qk}(X - \tau_k)_+^q \tag{3.8}$$

where, $1, X, \ldots, X^q, (X - \tau_1)_+^q, \ldots, (X - \tau_K)_+^q$ is the truncated power basis of degree of q, K is the total number of knots, $\tau_1 < \tau_2 < \cdots < \tau_K$ are pre-selected fixed knots. The penalized least square estimator $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_q, \hat{\beta}_{q1}, \ldots, \hat{\beta}_{qK}\right)$ can be obtained by mininimizing the penalized error sum of squares

$$\sum_{i=1}^{n}\left\{Y_i-\sum_{j=0}^{q}\beta_j X_i^j+\sum_{k=1}^{K}\beta_{qk}(X_i-\tau_k)_+^q\right\}^2+\lambda\sum_{k=1}^{K}\zeta\left(\beta_{qk}\right)$$

where, λ is a smoothing parameter which can be estimated by and $\zeta(\cdot)$ is a suitable non-negative penalty function. Although, this imputation model is driven by parameter specifications, it gains flexibility due to the inclusion of a non-linear mean function and imitates a nonparametric model when the number of knot points is very large.

3.4.3 Penalized Spline of Propensity Prediction (PSPP)

The previous approach is very effective for imputing missing observations for a single covariate. In presence of more than one covariate, one might want to extend this method by employing a multivariate spline in the response model. However, due to the curse of high dimensionality, these methods fail to perform effectively and often produce biased estimates. Motivated by this concern, Little and An (2004) proposed an alternative method where instead of including all predictors through a multivariate spline model, they included the spline on a function of a subset of covariates which are more sensitive to model misspecification. Specifically, they estimated the propensity score for each subject and a suitable function of the propensity score is used in the response model in order to adjust for confounding.

Following Zhang and Little (2009) here we describe the method in more detail. Suppose, we observe a group of n subjects whose response vector is denoted by $\mathbf{Y}$ where we use the same response indicator R introduced before such that $R=1$ corresponds to the respondents and $R=0$ implies that the response observation is missing. Let the covariates are denoted by $\mathbf{X}=(X_1,X_2,\ldots,X_p)'$ which are assumed to be fully observed for all patients. We also assume that the missing data pattern follows missing at random (MAR) mechanism. Based on the observed data, the propensity to response can be defined as $P(R=1|X_1,\ldots,X_p)$.

A spline function on the logit of the propensity score (say, $P^\star$) is then included in the imputation model non-parametrically using

$$\begin{aligned}(X_2,\ldots,X_p|P^\star) &\sim N_{p-1}\left(\left(s_2\left(P^\star\right),\ldots,s_p\left(P^\star\right)\right),\boldsymbol{\Sigma}\right)\\ (Y|P^\star,X_2,\ldots,X_p;\boldsymbol{\beta}) &\sim N\left(s\left(P^\star\right)+g\left(P^\star,X_2^\star,\ldots,X_p^\star;\boldsymbol{\beta}\right),\sigma^2\right)\end{aligned} \tag{3.9}$$

where, $s_j(P^\star)=E\left(X_j|P^\star\right), j=2(1)p$ denotes a spline function for the regression of X_j on $P^\star$ with the form (3.8); $X_j^\star=X_j-s_j\left(P^\star\right)$ represents the remainder part of X_j not explained by the propensity score; $s\left(P^\star\right)$ is the spline of Y on $P^\star$ of the same form (3.8) and $g(\cdot)$ is a parametric function characterized by the unknown parameter $\boldsymbol{\beta}$.

The specification of the PSPP model (3.9) requires the covariates $X_2^\star,\ldots,X_p^\star$ to be centered by regressing $X_2,\ldots,X_p$ on splines of $P^\star$. Zhang and Little (2009) simplified this formulation by including the original covariates $X_2,\ldots,X_p$ in the parametric function g without centering as follows:

$$(Y|P^\star,X_2,\ldots,X_p;\boldsymbol{\beta})\sim N\left(s\left(P^\star\right)+g\left(P^\star,X_2,\ldots,X_p;\boldsymbol{\beta}\right),\sigma^2\right). \tag{3.10}$$

Imputing a missing observation by a single value does not reflect the uncertainty involved in the prediction of the unknown missing value. Uncertainty quantification of an estimator is a very critical part of statistical inference which would not be possible here, as the resulting variance of the estimated parameters will be under biased towards 0. To address this concern, An and Little (2008) proposed three approaches to estimation of the standard error calculated from the singly imputed data for incorporating added variability due to imputation, and the properties of the resulting confidence intervals. One of these methods is developed under Bayesian paradigm which involves multiple draws from the joint posterior-predictive distribution of the missing values under the PSPP model and the standard error is computed using multiple-imputation combining methodology proposed in Rubin (2004). We will adopt

a similar approach in chapter 5 to impute the missing potential outcomes for the new patient and then the optimal treatment would be deduced by comparing the potential outcome distributions under both treatment options.

CHAPTER 4

BAYESIAN G-FORMULA BASED OPTIMAL TREATMENT ALLOCATION

4.1 Introduction

In this dissertation, we are focusing on estimating the optimal treatment decision of a future patient where the data in consideration comes from an observational study. The primary questions we are interested here are that how we can define the "best" treatment and how to choose the "best" option for a new patient. The terminology "best" is somewhat subjective and we define it with the help of potential outcome framework. Let $\tilde{\mathbf{X}}$ and $\tilde{Y}$ represent the confounders (biological profile) and the outcome of the new patient. Note that, we are looking at this research problem from a physician's point of view and hence assume that, for the new patient we know the value of $\tilde{\mathbf{X}}$, while $\tilde{Y}$ is unobserved. We define the "best" treatment for the new patient as a function of the observed confounder in the following way

$$a^{\star}(\tilde{\mathbf{X}}) = \underset{a}{\operatorname{argmax}}\ U\left(a, \tilde{\mathbf{X}}\right)$$

where, $U(\cdot)$ is the utility or reward function defined as in a decision theoretic framework and represents the reward of assigning treatment a to a patient with confounder value $\tilde{\mathbf{X}}$. We assume higher utility or reward corresponds to the better treatment. For this dissertation, we are restricting ourselves to a point treatment study with two treatment options (e.g. $A = 0$: Control and $A = 1$: Treatment). We assume that observed patients were assigned one of the two treatment options and the best treatment for the new patient will be one of

these two options. A natural choice of the utility function under the average loss function is $U\left(a, \tilde{\mathbf{X}}\right) = E\left(\tilde{Y}^a|\tilde{\mathbf{X}}\right)$ under which the best treatment decision can be expressed as

$$a^\star(\tilde{\mathbf{X}}) = \underset{a}{\text{argmax}}\ E\left(\tilde{Y}^a|\tilde{\mathbf{X}}\right).$$

where, $\tilde{Y}^a$ is the potential outcome of new patient if (s)he is administered treatment a.

We explore Bayesian methodologies to predict the potential outcome of the new patient which would be observed under a given treatment option. To this end, we obtain the posterior predictive distribution of the potential outcome which eventually leads us to the optimal decision. In the BayesG method, we consider a semi parametric Bayesian approach to estimate the treatment causal parameters using a marginal structural model based on the observed data which facilitates the optimal personalized treatment selection for the new patient. As the data we consider comes from observational study, we adjust for possible confounders for improved estimation of causal parameters. The parameters in the marginal structural model are estimated by g-formula within a semi-parametric Bayesian setting.

Our BayesG method gains precision due to its likelihood based construction. Unlike the common frequentist semi-parametric methods we do not need to specify two separate models for treatment specification and outcome regression based on treatment and confounders. Saarela et al. (2015) proposed a Bayesian approach for IPWE. However, Robins et al.(2015), on a discussion of Saarela et al. (2015) questioned the conceptual validity of propensity score approach in Bayesian paradigm. We circumvented this issue in our approach by standardizing the parameters through a prior on the confounders. The second advantage in our method is that it can be argued to be robust to model misspecification in predicting treatment choice probabilities.

4.2 Method

4.2.1 Framework

We consider the potential framework (Rubin, 1978) as discussed in earlier chapters. Suppose, we observe the data $\mathbf{D} = (\mathbf{Y}, \mathbf{A}, \mathbf{X})$ where $\mathbf{Y}$, $\mathbf{A}$ and $\mathbf{X}$ denote the outcome, treatment and confounder information of n patients in a time fixed setting i.e. the information are collected only once for each patient. In general, the treatment can be continuous (e.g. dose measurement) or categorical (multiple treatment levels). For the sake of simplicity we are restricting ourselves to dichotomous treatment with two options (0 or 1). For now, we only focus on continuous outcome Y where it is assumed that larger values of outcome is better.

Furthermore, we assume that the three assumptions discussed in section 3.1.4 hold true for the observed dataset. That means, by consistency assumption, the observed outcome for a patient is actually the potential outcome under the observed treatment level. Mathematically, $Y^a = Y$ for subjects with treatment $A = a$. By exchangeability or ignorability assumption, we measure all characteristics (confounders) that are involved in treatment decision making so that given a confounder set x, the treatment assignment can be assumed to be randomized. Formally, $p(Y^a|A = a, \mathbf{X} = \mathbf{x}) = p(Y^a|\mathbf{X} = \mathbf{x})$. Positivity assumption implies that for a given level of confounders, there are some patients in each treatment group i.e. $P(A = a|\mathbf{X} = \mathbf{x}) > 0$. Usually, exchangeability assumption does not hold in observational setting as the study design is not in control of the investigator. So one needs to verify it by performing sensitivity analysis.

4.2.2 Marginal Structural Model

The marginal structural model (MSM) provides a formal way to estimate the treatment causal effect on the marginal mean of potential outcomes given a subset of confounders, as discussed in section 3.2. Specifically, let the set of confounders $\mathbf{X} = (\mathbf{V}, \mathbf{W})$ includes two subsets $\mathbf{V}$ and $\mathbf{W}$. $\mathbf{V}$ includes only those confounders for which the treatment causal effect varies with the change of $\mathbf{V}$ in the sense of effect modification and thus are included in the modeling part of MSM. $\mathbf{W}$ is the other part of the confounder set which is not part of effect modification but we still need to adjust it for confounding. Following (3.5), the MSM can take a general form for continuous response. In our case, we assume a linear form as the following

$$E\left[Y^a|\mathbf{V};\boldsymbol{\alpha}\right] = \mathbf{D}^{\star'}\boldsymbol{\alpha} \tag{4.1}$$

where $\mathbf{D}^\star$ is a vector consisting of effect modifier and treatment variables which needs to be specified based on which causal parameters we are interested in estimating. For example, if we are interested in estimating the causal effect in the presence of an effect modifier, we can consider an interaction model in the MSM with $\mathbf{D}^\star = (1, A, V, AV)'$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_{12})'$ where α_1 is the parameter corresponding to the marginal causal effect of the treatment and α_{12} reflects how the treatment causal effect changes w.r.t a unit change in the effect modifier. Several semi-parametric (such as inverse probability weighting) methods have been proposed for estimating the causal parameters (Robins, 1999a,b, 2000; Robins et al., 2000; Saarela et al., 2015). Instead, we consider the g-formula [(Robins, 1986)] in the Bayesian setting utilizing its likelihood based formulation. Although, the standard g-formula requires full parameterization of the outcome model, we consider a semi-parametric approach which we will elaborate in the next subsection. We also note that while most of recent research on Marginal Structural models focused in the estimation perspective, out interest here is on

the prediction perspective of predicting the potential outcome for a patient for which the confounder information is observed.

4.2.3 Bayesian Parametric g-formula

We are interested in the marginal distribution of the potential outcome under treatment a which we denote by Y^a. This prompts us to consider the parametric g-formula (Robins, 1986) for obtaining the likelihood of the potential outcomes. Under the assumptions of (3.1), (3.2) and (3.3) a general form of the parametric g-formula estimate for the potential outcome distribution for treatment a is given by

$$p(Y^a|\boldsymbol{\theta}) = \int_{\mathbf{X}} p(Y|A, \mathbf{X}, \boldsymbol{\theta}_1)p(\mathbf{X}, \boldsymbol{\theta}_2)d\mathbf{X} \tag{4.2}$$

where, the parameter vector $\boldsymbol{\theta}$ includes $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ corresponding to the parameters in the conditional model $p(Y|A, \mathbf{X})$ and the parameters in the probability of confounders $p(\mathbf{X})$.

As Keil et al. (2015) pointed out, there are some advantages of recasting the parametric g-formula within the Bayesian paradigm. The frequentist version of parametric g-formula is dependent on the parametric specification of two unknown quantities - (1) the conditional mean model of the outcome based on the confounder and the treatment and (2) the marginal probability of the confounder. In complex scenarios such as when $\mathbf{X}$ is high dimensional or in presence of high correlation among the observations, the parameters of these two models might be poorly estimated. Under these scenarios, the usual method is to apply a more parsimonious model or to employ a semi or fully Bayesian framework to improve stability of the parameter estimates by incorporating prior information. The Bayesian version of g-formula, on the other hand, gets benefited from the variance reduction properties of Bayesian framework under the setup of parametric g-formula.

Keil et al. (2015) provided a Bayesian parametric formulation of the potential outcome distribution by calculating the posterior predictive distribution of the potential outcome. With the assumptions of conditional exchangeability, consistency and independent parameters, they showed that the potential outcome distribution can be expressed as a function of observed data $\mathbf{D}$, target population distribution of confounders $\mathbf{X}$, the given exposure a under Bayesian Parametric g-formula setup. Mathematically, the potential outcome predictive distribution can be expressed as

$$\begin{aligned} p(\tilde{Y}^a|\mathbf{D}) &= \int_{\boldsymbol{\theta}} p(\tilde{Y}^a, \boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(\tilde{Y}^a|\boldsymbol{\theta}, \mathbf{D}) p(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \end{aligned} \tag{4.3}$$

where, $\tilde{Y}$ denotes the outcome in the context of Bayesian predictive setting, $p(\boldsymbol{\theta}|\mathbf{D})$ is the posterior distribution of the parameters $\boldsymbol{\theta}$ representing the relationship within the observed data and $p(\tilde{Y}^a|\boldsymbol{\theta}, \mathbf{D})$ is the predictive distribution of the potential outcome. Under the assumption of independent parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ for the outcome and confounder model, (4.2) along with (4.3) implies

$$p(\tilde{Y}^a|\mathbf{D}) = \int_{\boldsymbol{\theta}} \int_{\tilde{\mathbf{X}}} p(\tilde{Y}|A, \tilde{\mathbf{X}}, \boldsymbol{\theta}_1, \mathbf{D}) p(\tilde{\mathbf{X}}|\boldsymbol{\theta}_2, \mathbf{D}) p(\boldsymbol{\theta}|\mathbf{D}) d\tilde{\mathbf{X}} d\boldsymbol{\theta}$$

where, $\tilde{\mathbf{X}}$ denotes the confounder in the context of Bayesian predictive setting, $p(\tilde{\mathbf{X}}|\boldsymbol{\theta}_2)$ is the predictive distribution of the confounder $\mathbf{X}$ and $p(\tilde{Y}|A, \tilde{\mathbf{X}}, \boldsymbol{\theta}_1)$ is the predictive distribution corresponding to the observed outcome. The samples drawn from posterior distribution of $\boldsymbol{\theta}$ would contain all information of the observed dataset $\mathbf{D}$ so that $p(\tilde{\mathbf{X}}|\boldsymbol{\theta}_2, \mathbf{D}) = p(\tilde{\mathbf{X}}|\boldsymbol{\theta}_2)$ and $p(\tilde{Y}|A, \tilde{\mathbf{X}}, \boldsymbol{\theta}_1, \mathbf{D}) = p(\tilde{Y}|A, \tilde{\mathbf{X}}, \boldsymbol{\theta}_1)$ making

$$p(\tilde{Y}^a|\mathbf{D}) = \int_{\boldsymbol{\theta}} \int_{\tilde{\mathbf{X}}} p(\tilde{Y}|A, \tilde{\mathbf{X}}, \boldsymbol{\theta}_1) p(\tilde{\mathbf{X}}|\boldsymbol{\theta}_2) p(\boldsymbol{\theta}|\mathbf{D}) d\tilde{\mathbf{X}} d\boldsymbol{\theta} \tag{4.4}$$

Note that, while the parametric g-formula (4.2) directly depends on the parameters $\boldsymbol{\theta}$, the Bayesian version of g-formula marginalizes over the parameters $\boldsymbol{\theta}$ and thus is more robust to model misspecification (Keil et al., 2015).

The Bayesian parametric g-formula does not give us insight on a particular patient's treatment choice as it marginalizes over the possible values of the confounder $\mathbf{X}$, making it suitable for estimating the causal effect at the population level, but not at the individual level. Instead of directly applying Bayesian parametric g-formula, we formulate the potential outcome distribution at the individual level as

$$\begin{aligned} p(\tilde{Y}^a|\tilde{\mathbf{X}}, \mathbf{D}) &= \int_{\boldsymbol{\theta}} p(\tilde{Y}^a, \boldsymbol{\theta}|\tilde{\mathbf{X}}, \mathbf{D}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(\tilde{Y}^a|\boldsymbol{\theta}, \tilde{\mathbf{X}}, \mathbf{D}) p(\boldsymbol{\theta}|\tilde{\mathbf{X}}, \mathbf{D}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(\tilde{Y}^a|\boldsymbol{\theta}, \tilde{\mathbf{X}}) p(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \end{aligned} \tag{4.5}$$

where $\tilde{\mathbf{X}}$ and $\tilde{Y}^a$ denote the new patient's observed confounder and the potential outcome under treatment a. The likelihood $p(\tilde{Y}^a|\boldsymbol{\theta}, \tilde{\mathbf{X}})$ conditions on the new patient's confounder information and hence through this formulation, we can incorporate the future patient's confounder information into the treatment decision making.

Both formulations (4.4) and (4.5) follow the Bayesian setting. In the first one, the common approach is to assume a parametric model for the outcome surface conditional distribution $p\left(\tilde{Y}|\tilde{\mathbf{X}}, \boldsymbol{\theta}\right)$ and put prior on the parameters involved. While this approach may produce more stable estimates, due to its over reliance on parametric assumptions, it may not work well in variety of scenarios. Instead of fully parameterizing the outcome surface model and placing informative or non-informative priors on the parameters, we propose to consider a semi-parametric Bayesian model for the outcome distribution within marginal structural model setup where we directly synthesize the confounders information through a

flexible Gaussian Process (GP) prior (Bernardo et al., 1998) on the confounders distribution. The details of the likelihood model is given in the next subsection.

4.2.4 Bayesian Inference

Likelihood Model. We build a likelihood model within the framework of marginal structural model as in Roy et al. (2016). We assume the outcome distribution to be parametric where the confounders contribute to a part of the mean function of the outcome model through GP prior such that the mean function is commensurate with that of the MSM (4.1). For concreteness, we assume the following expression for the outcome likelihood model:

$$p\left(Y^a|\phi(\mathbf{X});\boldsymbol{\alpha},\sigma_y\right) = N\left(Y; g(A,\mathbf{V};\boldsymbol{\alpha}) + \phi(\mathbf{X}), \sigma_y^2\right) \tag{4.6}$$

where, $\phi(\cdot) : R^p \to R$ is the function of p-variate confounders, $\boldsymbol{\alpha}$ are the parameters in the MSM (4.1) and σ_y is the standard deviation of the outcome model. The likelihood model is formulated under the condition that, the mean potential outcome conditioned on a specific value of the effect modifier $E\left[Y^a|\mathbf{V}\right]$ is same for (4.1) and (4.6). $g(A,\mathbf{V};\boldsymbol{\alpha})$ is a function of the treatment and the effect modifier which works as a buffer in order to maintain the connection between the likelihood model and the MSM.

Gaussian Process Prior. Following the works of Xu et al. (2016); Roy et al. (2016) we assume the functions $\phi(\mathbf{X})$ are independent observations from a Gaussian Process (GP) prior (Bernardo et al., 1998). For a finite (say, k) number of vectors each with p components $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k)'$ the distribution of $\boldsymbol{\phi}(\mathbf{X}) = (\phi(\mathbf{X}_1), \phi(\mathbf{X}_2), \ldots, \phi(\mathbf{X}_k))'$ is given by a k dimensional multivariate normal distribution as follows

$$\boldsymbol{\phi}(\mathbf{X}) \sim N_k(\boldsymbol{\mu}(\mathbf{X}), \mathbf{C}(\mathbf{X}))$$

where, $\boldsymbol{\mu}(\mathbf{X}) = (\mu(\mathbf{X}_1), \mu(\mathbf{X}_2), \ldots, \mu(\mathbf{X}_k))'$ is the $k \times 1$ mean vector and $\mathbf{C}(\mathbf{X}) = \mathbf{C}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k)$ is the $k \times k$ covariance matrix. In short, we denote this by $\boldsymbol{\phi}(\mathbf{X}) \sim \text{GP}(\boldsymbol{\mu}(\mathbf{X}), \mathbf{C}(\mathbf{X}))$. The mean and covariance functions of a GP can assume a very general form.

In our case, we specify GP for the observed dataset with n observations. Let, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)'$ where, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$. We include the regression function of the non effect modifiers $\mathbf{W} = \{W_{ij}, i = 1(1)n, j = 1(1)p\}$ in the prior mean of $\boldsymbol{\phi}(\mathbf{X})$ to model the dependence of $\boldsymbol{\phi}(\mathbf{X})$ as a function of $\mathbf{W}$. Mathematically, we assume,

$$(\boldsymbol{\phi}(\mathbf{X})|\boldsymbol{\beta}, \eta, \rho) \sim \text{GP}\left(\mathbf{W}\boldsymbol{\beta}, \mathbf{C}(\mathbf{X}; \eta, \rho)\right)$$

where, $\boldsymbol{\beta}$ are the parameters related to the mean regression function. $\mathbf{C}(\mathbf{X}; \eta, \rho)$ is the covariance function where the $(i, j)^{\text{th}}$ covariance takes the following form

$$\mathbf{C}(\mathbf{X}; \eta, \rho)|_{i,j} = \eta \exp\left(-\rho \left|\left|\mathbf{X}_i - \mathbf{X}_j\right|\right|^2\right) + \delta_{ij} J^2$$

where, $||\mathbf{X}_i - \mathbf{X}_j||^2$ measures the euclidean distance between the p variate confounder sets of i^{th} and j^{th} patients reflecting the dissimilarity in i^{th} and j^{th} patients' confounder profiles. Assuming each regression parameter $\beta_k, k = 1(1)p$ to have prior mean 0 and variance σ^2_β, the marginal prior covariance between the mean functions of i^{th} and j^{th} subjects can be expressed as

$$\begin{aligned}
\text{cov}(\phi(\mathbf{X}_i), \phi(\mathbf{X}_j)) &= E[\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)] - E[\phi(\mathbf{X}_i)]E[\phi(\mathbf{X}_j)] \\
&= E_{\boldsymbol{\beta}}[E[\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)|\boldsymbol{\beta}]] - E_{\boldsymbol{\beta}}[E[\phi(\mathbf{X}_i)|\boldsymbol{\beta}]]E_{\boldsymbol{\beta}}[E[\phi(\mathbf{X}_j)|\boldsymbol{\beta}]]
\end{aligned}$$

Noting that,

$$E_{\boldsymbol{\beta}}[E[\phi(\mathbf{X}_i)|\boldsymbol{\beta}]] = E_{\boldsymbol{\beta}}[\mathbf{W}_i'\boldsymbol{\beta}] = 0$$

and

$$E[\phi(\mathbf{X}_i)\phi(\mathbf{X}_j)|\boldsymbol{\beta}] = \text{cov}(\phi(\mathbf{X}_i), \phi(\mathbf{X}_j)|\boldsymbol{\beta}) + E[\phi(\mathbf{X}_i)|\boldsymbol{\beta}]E[\phi(\mathbf{X}_j)|\boldsymbol{\beta}]$$
$$= \eta \exp\left(-\rho\,||\mathbf{X}_i - \mathbf{X}_j||^2\right) + \delta_{ij}J^2 + \sum_{k=1}^{p} \mathbf{W}_{ik}\mathbf{W}_{jk}\boldsymbol{\beta}_k^2$$

the marginal prior covariance as in (4.2.4) can be re-expressed as

$$\text{cov}(\phi(\mathbf{X}_i), \phi(\mathbf{X}_j)) = \eta \exp\left(-\rho\,||\mathbf{X}_i - \mathbf{X}_j||^2\right) + \delta_{ij}J^2 + \sum_{k=1}^{p} W_{ik}W_{jk}\sigma_\beta^2$$

In the above formulation, k is the index for the confounders, $\eta > 0$ measures the deviation of the covariance from linearity, $\rho > 0$ scales the contribution of the distance between the confounders. The first term makes sure that the covariance is larger for similar profiles and vice versa. δ_{ij} corresponds to Kronecker's delta taking value 1 if $i = j$ and 0 otherwise. J reflects the amount of jitter, which is similar to noise in regression setting (Bernardo et al., 1998). A small value of J is usually considered (e.g. J=0.1) to make matrix computation more stable. This formulation of covariance function should capture complex scenarios. More general forms of covariance function can certainly be used to capture the variation in confounder values with the addition of more parameters. However, it comes at the cost of intensive computation.

Specification of $g(A, \mathbf{V}; \boldsymbol{\alpha})$. For the full specification of the mean function we need to know the form of $g(A, \mathbf{V}; \boldsymbol{\alpha})$ which can be identified by matching the marginal mean outcome from (4.1) with that of (4.6). In this context, note that,

$$E[Y^a|\mathbf{X}] = E_{\phi(\mathbf{X})|\mathbf{X}}[E[Y^a|\phi(\mathbf{X}), \mathbf{X}]]$$
$$= E_{\phi(\mathbf{X})|\mathbf{X}}[g(A, \mathbf{V}; \boldsymbol{\alpha}) + \phi(\mathbf{X})]$$
$$= g(A, \mathbf{V}; \boldsymbol{\alpha}) + \mathbf{W}'\boldsymbol{\beta}$$

Averaging out $\mathbf{W}$ in the above equation we get

$$\begin{aligned} E[Y^a|\mathbf{V}] &= \int_{\mathbf{W}} (g(A, \mathbf{V}; \boldsymbol{\alpha}) + \mathbf{W}'\boldsymbol{\beta}) dF(\mathbf{W}|\mathbf{V}) \\ &= g(A, \mathbf{V}; \boldsymbol{\alpha}) + \int_{\mathbf{W}} \mathbf{W}'\boldsymbol{\beta} dF(\mathbf{W}|\mathbf{V}) \end{aligned} \tag{4.7}$$

where, $F(\mathbf{W}|\mathbf{V})$ is the distribution of $\mathbf{W}$ conditioned on $\mathbf{V}$. Comparing (4.7) with (4.1) will identify $g(A, \mathbf{V}; \boldsymbol{\alpha})$ as

$$g(A, \mathbf{V}; \boldsymbol{\alpha}) = \mathbf{D}^{\star'}\boldsymbol{\alpha} - \int_{\mathbf{W}} \mathbf{W}'\boldsymbol{\beta} dF(\mathbf{W}|\mathbf{V})$$

The specification of $g(A, \mathbf{V}; \boldsymbol{\alpha})$ would require us to compute the integral $\int_{\mathbf{W}} \mathbf{W}'\boldsymbol{\beta} dF(\mathbf{W}|\mathbf{V})$. Roy et al. (2016) provided a nonparametric solution of this integral under the assumption of a categorical or a low dimensional $\mathbf{V}$. Specifically, they approximated $\int_{\mathbf{W}} \mathbf{W} dF(\mathbf{W}|\mathbf{V})$ by $\frac{1}{n_{\mathbf{v}}} \sum_{i:\mathbf{V}_i=\mathbf{v}} \mathbf{W}_i = \tilde{\mathbf{W}}_{\mathbf{v}}$ where, $n_{\mathbf{v}}$ is the number of patients having the effect modifier value $\mathbf{v}$ which implies

$$g(A, \mathbf{V}; \boldsymbol{\alpha}) \approx \mathbf{D}^{\star'}\boldsymbol{\alpha} - \tilde{\mathbf{W}}'_v\boldsymbol{\beta}.$$

However, for a continuous $\mathbf{V}$ this approximation will not work and we have to resort to a model based framework. We can consider a parametric or non-parametric model for $p(\mathbf{W}|\mathbf{V})$. As the relationship between $\mathbf{W}$ and $\mathbf{V}$ is generally unknown, assuming a parametric model would restrict the model space reducing flexibility of the the final model which is why a non-parametric specification might be a better alternative. For example, for a univariate W and V, we can assume the following cubic regression spline as the specification is flexible enough to capture a wide range of distributions.

$$f(W|V, \boldsymbol{\delta}) = \sum_{j=0}^{3} \delta_j V^j + \sum_{k=1}^{K} \delta_{3k} (V - \tau_{vk})_+^3$$

where, $1, V, \ldots, V^3, (V-\tau_{v1})_+^3, \ldots, (V-\tau_{vK})_+^3$ is cubic p-spline basis, $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_3, \delta_{31}, \ldots, \delta_{3K})$ are corresponding coefficients, $\{\tau_{v1}, \ldots, \tau_{vK}\}$ are K pre-specified knot points over the range of V and $x_+^3 = x^3 I(x > 0)$. For a low dimensional multivariate $\mathbf{W}$ and $\mathbf{V}$ one can extend this specification to multivariate spline model. Accordingly, we can approximate $\int_{\mathbf{W}} \mathbf{W} dF(\mathbf{W}|\mathbf{V})$ by $\boldsymbol{\mu}(\mathbf{V}) = E(\mathbf{W}|\mathbf{V})$ which then determines $g(\cdot)$ as

$$g(A, \mathbf{V}; \boldsymbol{\alpha}) \approx \mathbf{D}^{\star'} \boldsymbol{\alpha} - \boldsymbol{\mu}(\mathbf{V})' \boldsymbol{\beta}.$$

Prior Information. We consider a relatively flat prior for the MSM parameters $\boldsymbol{\alpha}$ by assuming $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{\pi})$ where, $\Sigma_{\boldsymbol{\alpha}}^{\pi}$ is a diagonal matrix with large values in the diagonal. For the regression parameter $\boldsymbol{\beta}$ of the mean function of GP prior, we assume $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\pi}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\pi})$, with known values of $\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\pi}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\pi}$. We have to keep in mind that the scale parameter ρ should not be too small as it may destabilize the inversion of the covariance function of the GP. If the value of ρ is close to zero, it may lead to near singular matrix of covariance function (Roy et al., 2016). Apart from these, we further assume $\sigma_y^{-2} \sim \text{Gamma}(1, 1)$ and $\eta \sim \text{Gamma}(1, 1)$.

4.2.5 Optimal Treatment Decision

The main objective of our method is to allocate the superior treatment choice to the new patients based on their biomarker profiles and the observed data. The "best" treatment is determined by the one that corresponds to the treatment option associated with a higher outcome. Given the pre-treatment variables of the new patient, say $\tilde{\mathbf{X}} = \{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_p\}'$, the optimal treatment, $a^{\star}$ can be expressed as a function of $\tilde{\mathbf{X}}$

$$a^{\star} = \underset{a}{\text{argmax}}\, E(\tilde{Y}^a | \tilde{\mathbf{X}}, \mathbf{D})$$

where, $\tilde{Y}^a$ is the potential outcome of the new patient under treatment a.

With that in mind, we develop a Bayesian methodology to predict the mean potential outcome for the new patient under treatment a for a given $\tilde{\mathbf{X}}$ and the observed data $\mathbf{D}$ which we denote as $\tilde{\mu}(a)$ defined by $\tilde{\mu}(a) = E(\tilde{Y}^a|\tilde{\mathbf{X}}, \mathbf{D}) = g(A, \tilde{\mathbf{V}}; \boldsymbol{\alpha}) + \phi(\tilde{\mathbf{X}})$. Subsequently to estimate $a^\star$ based on the framework we described before. As discussed in section 4.2.5 the mean posterior predictive potential outcome under treatment a for the new patient can be formulated as

$$\tilde{\mu}(a) = E(\tilde{Y}^a|\tilde{\mathbf{X}}, \mathbf{D}) = \int_{\boldsymbol{\theta}} E(\tilde{Y}^a|\tilde{\mathbf{X}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta}$$

where, $\boldsymbol{\theta}$ is the set of unknown parameters $\mu(\cdot)$, $\boldsymbol{\theta} = \{\alpha, \beta, \eta, \rho, \sigma_y\}$.

The posterior predictive computation is performed in two steps. First, for $a = \{0, 1\}$, we draw N Markov chain Monte Carlo (MCMC) samples of $\boldsymbol{\theta}$, (say, $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_N^\star$) from their joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$. Then, for each drawn MCMC sample of $\boldsymbol{\theta}^\star$, we predict the mean potential outcome of the new patient by plugging in $\boldsymbol{\theta}^\star$ as a value of $\boldsymbol{\theta}$ in $E(\tilde{Y}^a|\tilde{\mathbf{X}}, \boldsymbol{\theta})$. Accumulating all predicted mean potential outcomes, $\tilde{\mu}_1(a), \tilde{\mu}_2(a), \ldots, \tilde{\mu}_N(a)$ provide basis to construct empirical predictive distribution of the mean potential outcomes for the new patient.

We can estimate the optimal treatment decision in many different ways. For example, we can choose $a^\star$ following the recommendation of Gelfand and Ghosh (1998) by minimizing a suitable posterior predictive loss function. Considering the average loss function, the optimal treatment can be estimated as

$$\hat{a}^\star = I\left(\bar{\tilde{\mu}}(1) > \bar{\tilde{\mu}}(0)\right)$$

where, $\bar{\tilde{\mu}}(a) = \frac{1}{N}\sum_{i=1}^{N} \tilde{\mu}_i(a)$ with $\tilde{\mu}_i(a)$ being the predicted mean potential outcome of the new patient under treatment a corresponding to i^{th} MCMC draw.

However, instead of averaging the posterior predictive draws, one could follow a fully Bayesian approach by considering the notion of stochastic ordering approach where, the

treatment options are compared for each MCMC draw from the posterior distribution. We can estimate the probability of potential outcome under treatment 1 to be higher than potential outcome under treatment 0 by the following formulation, $P\left(\tilde{\mu}(1) > \tilde{\mu}(0)|\tilde{\mathbf{X}}, \mathbf{D}\right)$ by $\hat{P} = \frac{1}{N}\sum_{i=1}^{N} I(\tilde{\mu}_i(1) > \tilde{\mu}_i(0))$. Then the optimal treatment can be derived by putting a suitable threshold on $\hat{P}$. Using an threshold of 0.5 the optimal treatment can be estimated by

$$\hat{a}^{\star} = I\left(\hat{P} > 0.5\right)$$

The threshold can be determined from a historical or a biological point of view.

4.2.6 A Numerical Study

We conduct a numerical study to illustrate the performance of the BayesG method under a simulation setting where the data generation reflects the causal diagram 3.1. We modified the simulation 1 setting of Zhang et al. (2012b) to include three covariates X_1, X_2, X_3 in the covariate set $\mathbf{X}$. We generate $n = 500$ observations of $\{Y_i, \mathbf{X}_i, A_i\}; i = 1, \ldots, n$ where, $\mathbf{X}_i$ are generated from the following model

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & 0.25 & 0.175 \\ 0 & 0.175 & 0.25 \end{pmatrix}\right]$$

The treatments are generated by the following model

$$A \sim \text{Bern}\{\text{logit}^{-1}(-1 + 0.8X_1^2 + 0.8X_2^2 + 0.8X_3^2)\}$$

and the outcomes are simulated based on the generated treatments and covariates using

$$Y \sim N(\mu(A, \mathbf{X}), 1)$$

where, $\mu(A, \mathbf{X}) = 2 - 1.5X_3^2 + 3X_1X_3 + A(-0.1 - X_1 + X_2)$. For the i^{th} patient the true optimal treatment can be deduced from the outcome model as

$$A_i^\star = I(X_{i1} - X_{i2} < -0.1).$$

Note that, all covariates X_1, X_2, X_3 contribute to generate treatments A and outcomes Y and hence they can be described as confounders. Among them, X_1 and X_2 have interaction with A in the outcome model and so the treatment effect will vary for different levels of these two confounders raising the situation of effect modification which implies $\mathbf{V} = (X_1, X_2)'$. On the other hand, X_3 affects both treatment and outcome, but is not part of effect modification making $W = X_3$.

We apply the BayesG approach in this simulation setting. We assume the true causal model to be $E(Y^a|\mathbf{V}) = \alpha_0 + \alpha_1 A + \boldsymbol{\alpha}_2'\mathbf{V} + \boldsymbol{\alpha}_{12}'A\mathbf{V}$ which we intentionally considered different from the true data generating model in this case. As W is continuous here, we assume a cubic p-spline of the form (4.2.4) for the conditional distribution $f(W|\mathbf{V})$ which is required to compute $E(W|\mathbf{V})$ and consequently the $g(\cdot)$ function. We use the *gam* function from the *mgcv* package in R for this part. We estimate the posterior distributions from the Gibbs sampling procedure. For the posterior computation, we consider Markov Chain Monte Carlo (MCMC) methods. We draw 50,000 MCMC samples from the full conditional distribution of each parameter. Among them 25,000 iterations are considered for burn-in period and the remaining 25,000 iterations are considered for empirically estimating the posterior predictive distribution of the potential outcomes.

Table 4.1: Prediction Measurements of BayesG Method

Accuracy			Sensitivity			Specificity		
Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
0.917	0.93	0.056	0.943	0.965	0.076	0.883	0.92	0.125

We consider three predictive measures (accuracy, sensitivity and specificity) for evaluating the performance of the BayesG method. The sensitivity is measured by the proportion of accurate prediction of treatment 1, while the specificity corresponds to the same for treatment 0. The accuracy measurement gives an overview of both these measures by comparing the predicted optimal treatments to the true ones. Details of these measures are given in the section 6.3.2.

Table 4.1 summarizes the prediction measurements with mean, median and standard deviation which are obtained from 100 holdout samples. Section 6.3.1 provides details of the holdout procedure. From the result table, we see that BayesG has more than 90% overall accuracy for correctly predicting true optimal treatments. Specifically, it predicts true treatment 1 over 95% of the times while, the other treatment options are correctly predicted around 90% of the times. We compare BayesG with other methods in five simulation scenarios including the current settings which we detail in chapter 6.

CHAPTER 5

PROPENSITY PREDICTION BASED OPTIMAL TREATMENT ALLOCATION

In this section, we propose an alternative method (PSBayes) to estimating the optimal treatment decision for a new patient based on his/her biological profile and clinical information in a binary treatment setting. Similar to the BayesG method, the current approach also compares the potential outcomes of a new patient under each treatment option. However, as we already have discussed we do not observe the outcome simultaneously for both treatment options. Instead, we only observe the outcome from the allocated treatment option. Under the consistency assumption (3.2) the observed outcome is the potential outcome under the given treatment and thus only one of the potential outcomes is observed for each patient while the other potential outcome remains counterfactual leading to the fact that each patient has a missing potential outcome corresponding to the treatment option that (s)he has not received. This motivates us to consider methods from missing data literature where the objective is to predict the missing potential outcome from the observed predictor information. We follow an imputation approach in this context for each missing potential outcome so that we can compare the potential outcomes for the new patient which would finally direct us to the optimal treatment allocation.

5.1 Introduction

5.1.1 Framework

As in the previous section, we follow Rubin's (Rubin, 1978) potential outcome framework. For $a \in \{0, 1\}$, we denote Y^a as the potential outcome that would be observed if a patient is administered treatment a. The observed data does not provide information on the potential outcome Y^a under each treatment option a that a patient can have. We only observe a patient's outcome when they are exposed to a certain treatment a, not what the patient's outcome would have been had (s)he been unexposed to a. As a result, the causal effect due to the treatment A is not identifiable at an individual level. To link the potential outcome with the observed outcome, we need to resort to the consistency assumption (3.2) which states that the observed outcome is essentially the potential outcome under the treatment option a patient actually receives. It also implies that, under binary treatment settings when we have two treatment options, say $A = 0$ and $A = 1$, either Y^0 or Y^1 is actually observed while the other one is missing. In general notation, if a patient receives the treatment $A = a$, then the potential outcome Y^{1-a} is missing. Therefore, the treatment variable A can be assumed to be the response indicator of the potential outcome Y^1 or the missing value indicator of the potential outcome Y^0. Therefore, it is evident that estimating individual level treatment comparison can be posed as a missing data problem where the outcome of interest is essentially the potential outcome under each treatment option. With this observation, one approach to make optimal treatment decision, could be by imputing the potential outcomes for the new patient. In section 3.4.1, we have reviewed a few imputation methods from the missing data literature where the objective is to unbiasedly estimate the mean of the outcome having missing observations in MAR setting. In the next section,

we make a connection between the missing data mechanism and the treatment allocation procedure and in the following sections we develop the proposed PSBayes methodology for optimal treatment allocation based on missing data framework.

5.1.2 Treatment Allocation Mechanism

Similar to the missing data mechanism, the treatment assignment can also be classified into three major groups depending on the conditional distribution of the treatment assignment given the covariates. There is a striking similarity between the patterns of the missing data mechanism and the treatment assignment mechanism. For example, in completely randomized experiment the treatment assignment probability is pre-determined which is common for all patients. Hence the conditional distribution of the treatment assignment does not depend on any variable. Denoting the observed covariates by $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, the full conditional distribution of treatment assignment given the potential outcomes and covariates can be expressed as $P(A = a|Y^0, Y^1, \mathbf{X}, \boldsymbol{\theta}) = P(A = a|\boldsymbol{\theta})$, which is similar to the setting of 'missing completely at random' (MCAR) mechanism.

By the no unmeasured confounders assumption (3.1) which is considered throughout the dissertation, given all confounders that are responsible for treatment allocation and/or are associated with outcome, the treatment can be assumed to be randomized among the subjects with the same observed covariates. This assumption is based on the notion of conditional exchangeability which can be considered as a special case of 'missing at random' (MAR) mechanism. The MAR mechanism corresponds to a more general class where the conditioning is on the observed part of the outcome Y^o along with the covariates, but not on the unobserved part of the outcome Y^m. Mathematically, it implies $P(A = a|Y^0, Y^1, \mathbf{X}, \boldsymbol{\theta}) = P(A = a|Y^o, \mathbf{X}, \boldsymbol{\theta})$. On the other hand, in conditional exchangeability assumption, the

treatment assignment only depends on the observed value of the covariates i.e. $P(A = a|Y^0, Y^1, \mathbf{X}, \boldsymbol{\theta}) = P(A = a|\mathbf{X}, \boldsymbol{\theta})$.

In some situations, the treatments are administered not only based on the observed covariates, but the potential outcomes also play role in deciding the best treatment. In such cases, the treatment assignment mechanism depends on both the observed and missing information i.e. $P(A = a|Y^0, Y^1, \mathbf{X}, \boldsymbol{\theta}) = P(A = a|Y^m, Y^o, \mathbf{X}, \boldsymbol{\theta})$, where Y^m is the unobserved part of the outcome. It is very similar to the scenario of 'not missing at random' (NMAR) where the missing indicator not only depends on the observed data but also on the missing data.

Due to the fair similarity between the frameworks of causal inference and missing data analysis, two common approaches are often pursued to handle missing data or treatment allocation mechanism: weighting and imputation. In weighting methods such as inverse probability weighting (IPWE) or augmented inverse probability weighting (AIPWE), a parametric or semi-parametric model is often assumed where the weights are calculated based on the probability of allocation to a certain treatment group. The entire method usually consists of two stages where in the first stage, a propensity score model is employed to fit the probability of treatment allocation $P(A = a|\mathbf{X}, \boldsymbol{\gamma})$ on the observed covariates assuming the treatment decision is driven by the whole set of covariates $\mathbf{X}$. A weighted analysis is then carried out to model the outcome Y based on the covariates $\mathbf{X}$ where the weight is a function of the observed propensity score $\hat{P}(A|\mathbf{X}, \hat{\boldsymbol{\gamma}})$. On the other hand, in the imputation method, a function of the observed covariates is often used to model the outcome and the same model is employed to predict the missing outcome given a set of covariates. Although, both approaches are mathematically interconnected, they often originate from different perspectives and are implemented differently.

5.1.3 Propensity Score

In observational studies, treatment assignment is usually driven by patient characteristic(s) and the response is observed after a treatment is administered. Hence, there is high likelihood that the treatment effect gets confounded with the effect of the characteristic(s) which is associated with treatment selection. In order to remove the confounding bias in treatment effect estimation, it is important to adjust for the association between the treatment and other covariates.

In randomized clinical trial, the treatment assignment mechanism is pre-determined. However, in observational studies, the probability of receiving a certain treatment, also known as propensity score is unknown and is estimated by fitting a model to the treatment based on the relevant covariates. For a binary treatment A, the following model can be used to fit the propensity of receiving $A = 1$.

$$\text{logit}\{P(A = 1|\mathbf{X}, \gamma)\} = \mathbf{X}\gamma \tag{5.1}$$

Instead of logit, one might also use the probit link as

$$\Phi^{-1}\left(P(A = 1|\mathbf{X}, \gamma)\right) = \mathbf{X}\gamma \tag{5.2}$$

Rosenbaum and Rubin (1983) showed that the propensity score has a key balancing property in the sense that given a value of the propensity score, the treatment choice does not depend on the covariates. Thus the treatment assignment mechanism can be assumed to be completely randomized within the group of subjects having similar PS values. Mathe-

matically, the conditional independence between treatments and covariates can be expressed as

$$A \perp\!\!\!\perp \mathbf{X} \mid e(\mathbf{X}).$$

where, $e(\mathbf{X})$ is the conditional probability of assignment to treatment 1 given the covariates i.e. $e(\mathbf{X}) = P(A = 1|\mathbf{X})$.

Propensity score can be used as a matching tool to remove imbalance in the confounder distribution among the treated and untreated. Exploiting this feature, propensity score is commonly used in both weighting and imputation methods to remove or adjust confounding bias in causal inference literature. Robins (1998b,c) introduced a weighted analysis approach to estimate the treatment causal effect where the weight of each subject is obtained by the inverse probability of the treatment, $\hat{w}(\mathbf{X}) = \hat{P}(A = a|\mathbf{X})^{-1}$. For the subjects in treatment 1 group, $\hat{w}(\mathbf{X}) = \hat{e}(\mathbf{X})^{-1}$ and for subjects receiving treatment 0, $\hat{w}(\mathbf{X}) = (1 - \hat{e}(\mathbf{X}))^{-1}$.

The estimated weights create a pseudo population free of treatment confounders, which allows to unbiasedly estimate causal parameters in a regression setting. The weighting methods are very popular in causal inference and survey sampling literature. However, one major limitation of such method is that, due to its inverse formulation, it may assign large weight to a subject with small propensity score yielding estimates with large variance (Little and Rubin, 2014).

Apart from the weighting method, several approaches have been proposed to include weights in the imputation model for predicting nonresponse in the missing data setting. For example, the linear in the weight prediction (LWP) method (Scharfstein et al., 1999; Bang and Robins, 2005; Zhang and Little, 2009) includes the weight as a linear term in the imputation model as follows

$$(Y|\mathbf{X};\beta) \sim N\left(g\left(\mathbf{X};\beta\right) + \alpha\hat{w}(\mathbf{X}), \sigma^2\right).$$

In the proposed methodology, we adopt a similar approach to imputation model specification. But instead of using the inverse function, we include the propensity score in the imputation model through a Bayesian p-spline function resulting in a more robust specification of the mean model. This approach avoids weighting which may result in highly variable estimates for small number of cases. Also, under the Bayesian paradigm PSBayes allows for small sample inferences which can aptly reflect uncertainty in imputing potential outcomes. In the next subsection, we discuss p-spline model for a complete data. In the later sections we describe the proposed likelihood model in the light of Bayesian p-spline regression.

5.1.4 P-Spline Model

Berry et al. (2002) proposed a regression penalized spline (P-spline) model for complete data under both frequentist and Bayesian settings. Consider a simple scenario where there are two sets of vectors - $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ and $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$. Assume that, i^{th} components of $\mathbf{Y}$ and $\mathbf{X}$ are connected by $Y_i = m(X_i) + \epsilon_i$, where ϵ_i is a random error with mean 0 and variance σ^2_ϵ. Let g be a P-spline approximator of m, such that the squared error distance $\sum_{i=1}^{n}\{m(X_i) - g(X_i)\}^2$ is minimized. Under the assumption that m is a smooth function, $m(cdot)$ can be well approximated by $g(\cdot)$. Following the notations from Berry et al. (2002) the regression P-spline model g can be specified by $g(X) = \mathbf{B}(X)'\boldsymbol{\beta}$ where, $\mathbf{B}(X) = (B_1(X), B_2(X), \ldots, B_N(X))'$ is a $N \leq n$ dimensional spline basis with corresponding regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)'$. Under this framework, for a given value of the smoothing parameter α, the penalized least square estimator $\hat{\boldsymbol{\beta}}$ can be obtained by minimizing

$$\sum_{i=1}^{n}\{Y_i - \mathbf{B}(X_i)'\boldsymbol{\beta}\}^2 + \alpha\boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta},$$

and is given by

$$\hat{\boldsymbol{\beta}}(\alpha) = \left(\mathcal{B}'\mathcal{B} + \alpha\mathbf{D}\right)^{-1}\mathcal{B}'\mathbf{Y}$$

where, $\mathcal{B} = \{\mathbf{B}(X_1), \mathbf{B}(X_2), \ldots, \mathbf{B}(X_n)\}'$.

One convenient basis function considers a p degree polynomial of the form $\mathbf{B}(x) = (1, x, x^2, \ldots, x^p, (x-t_1)_+^p, \ldots, (x-t_K)_+^p)'$, where $t_1, t_2, \ldots, t_K$ are K fixed knots, preferably chosen at the quantiles of X. In this specification, the regression coefficients can be divided into two parts - the first part of regression coefficients $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{(p+1)})'$ corresponds to monomial basis functions whereas, the second part $\boldsymbol{\beta}_2 = (\beta_{(p+2)}, \ldots, \beta_N)'$ represents the jumps in the p^{th} derivative of $g(X)$ at the knots, where $N = K + p + 1$. To penalize these jumps, Berry et al. (2002) proposed $\mathbf{D}$ matrix as $N \times N$ diagonal matrix with $p+1$ zeroes followed by K ones along the diagonal so that the roughness penalty term $\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} = \sum_{j=1}^{K}\beta_{1+p+j}^2$ yields the sum of squared jumps.

Under the Bayesian setting, the same specification can be achieved from the following formulation by An and Little (2008)

$$\left(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2_{\boldsymbol{\beta}_2}, \sigma_y^2, \mathbf{X}_1, \mathbf{X}_2\right) \sim N\left(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \sigma_y^2\mathbf{I}_n\right).$$

where,

$$\mathbf{X}_1 = \begin{pmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^p \\ 1 & X_2 & X_2^2 & \cdots & X_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^p \end{pmatrix} \quad \text{and} \quad \mathbf{X}_2 = \begin{pmatrix} (X_1-t_1)_+^p & \cdots & (X_1-t_K)_+^p \\ (X_2-t_1)_+^p & \cdots & (X_2-t_K)_+^p \\ \vdots & \vdots & \vdots \\ (X_n-t_1)_+^p & \cdots & (X_n-t_K)_+^p \end{pmatrix},$$

$\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ where $\boldsymbol{\beta}_1$ is the fixed effect and $\boldsymbol{\beta}_2$ is the random effect with

$$\left(\boldsymbol{\beta}_2 | \sigma^2_{\boldsymbol{\beta}_2}\right) \sim N\left(\mathbf{0}, \frac{\sigma^2_{\beta_2}}{\alpha}\mathbf{I}_K\right).$$

where, α is the tuning parameter as in the frequentist setting. The prior distributions of the unspecified parameters denoted by $\boldsymbol{\theta} = \left(\boldsymbol{\beta}_1, \sigma^2_{\boldsymbol{\beta}_2}, \sigma^2_y\right)$ are given by

$$\pi\left(\boldsymbol{\beta}_1\right) \propto 1,\ \pi\left(\sigma^2_{\boldsymbol{\beta}_2}\right) \sim \text{Inv-Gamma}\left(a_{\beta_2}, b_{\beta_2}\right),\ \pi\left(\sigma^2_y\right) \sim \text{Inv-Gamma}\left(a_\epsilon, b_\epsilon\right)$$

Under this specification, the mean of the posterior distribution of $\boldsymbol{\beta}$ conditional on σ^2_y and α yields the same penalized least square estimator as obtained from the frequentist specification.

5.2 Method

5.2.1 Framework

Suppose we observe a group of n patients with p covariates and the covariate matrix is denoted by $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)'$, where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ is the covariate vector of i^{th} patient. We also observe the treatments the patients have received and the observed treatment vector is denoted by $\mathbf{A}$ where each treatment can only be 0 or 1. Lastly, the response vector of all patients is denoted by $\mathbf{Y}$. For the new patient, we only have information on his/her covariates denoted as $\tilde{\mathbf{X}}$. For simplicity, we assume that there is no missing information within the observed dataset $\mathbf{D} = (\mathbf{X}, \mathbf{A}, \mathbf{Y})$.

The potential outcome under a treatment a is denoted by Y^a. Each patient has two potential outcomes Y^0 and Y^1. Under the consistency assumption, the observed outcome can

Table 5.1: Observed and missing potential outcome under each treatment

$\mathbf{X}$	$\mathbf{A}$	$\mathbf{Y}^0$	$\mathbf{Y}^1$
$\mathbf{X}_1$	0	Y_1	$-$
$\mathbf{X}_2$	0	Y_2	$-$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{X}_r$	0	Y_r	$-$
$\mathbf{X}_{r+1}$	1	$-$	Y_{r+1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{X}_n$	1	$-$	Y_n
$\tilde{\mathbf{X}}$	$-$	$-$	$-$

be expressed in terms of potential outcome and the treatment option (s)he has been administered by $Y = Y^0(1-a)+Y^1(a)$, which implies one of the potential outcome corresponding to the treatment that (s)he has received is observed and the other one is missing which implies the treatment indicator A plays the role of the response indicator for the potential outcome Y^1, as under the consistency assumption, Y^1 is observed when $A = 1$ and is missing when $A = 0$.

Suppose, in the dataset $\mathbf{D}$, first r out of n patients received treatment 0 and the remaining $n - r$ patients received treatment 1. Then from the data point of view, it implies that the potential outcome Y^0 is observed for the first r patients while Y^1 is missing, whereas for the remaining $n-r$ patients, the potential outcome Y^1 is observed and Y^0 is counterfactual. For the new patient, both Y^0 and Y^1 are missing. This observed connection between treatment allocation and missing data mechanism is visually illustrated in table 5.2.1. Although the observed data does not have any missing information, under our assumption, the full data comprised of the potential outcomes denoted by $\mathbf{D}_F = (\mathbf{X}, \mathbf{A}, \mathbf{Y}^0, \mathbf{Y}^1)$ have missing information due to unobserved counterfactual outcomes.

Motivated by observing the close relationship between the treatment assignment and the missing data mechanism, we build the PSBayes methodology within the missing data

framework. Zhang and Little (2009) discussed several regression based imputation methods from the missing data literature where the objective was to estimate the mean of a variable with missing observation. By drawing analogy to the potential outcome settings in causal inference, those methods can also be applied to the framework we are considering here. For example, the imputation model for the potential outcome can be formulated by a parametric model as

$$Y^a = \beta_{0a} + \sum_{j=1}^{p} \beta_{ja} X_j + \epsilon$$

where, $\beta_{0a}, \beta_{1a}, \ldots, \beta_{pa}$ are the regression coefficients characterizing effect of covariates on the potential outcome under treatment a. One can estimate β_{ja} based on the observed outcome by the frequentist or Bayesian approach and predict the missing counterfactual outcome for an observed set of covariates. Then one can estimate the mean potential outcome by combining observed and predicted potential outcomes. This method is very effective under the correct model specification, however performs poorly otherwise.

One way to avoid this limitation is to consider a robust mean function in the model specification similar to the linear in weight prediction (LWP) method (Scharfstein et al., 1999; Bang and Robins, 2005; Zhang and Little, 2009) which includes inverse propensity score as an additive term in the model. Robustness in the imputation model specification can also be achieved by relaxing its parametric specification. Semi-parametric and non-parametric methods have the ability to capture the nonlinear relationship between model variables.

In presence of a single predictor X, we can follow Ruppert et al. (2003); Eilers and Marx (1996) to specify the imputation model based on the penalized spline model $Y^a = s(X, \beta_a) + \epsilon$ with truncated polynomial basis of degree q as follows

$$s(X, \boldsymbol{\beta}_a) = \sum_{j=0}^{q} \beta_{ja} X^j + \sum_{k=1}^{K} \beta_{qka} \left(X - t_k\right)_+^q \tag{5.3}$$

where, $1, X, X^2, \ldots, X^q, (X - t_1)_+^q, \ldots, (X - t_K)_+^q$ is truncated power basis of degree q, similarly defined as in the previous section and $t_1 < \cdots < t_K$ are the fixed knots on the range of X.

The model above is robust and very effective to capture the non-linear relationship between the potential outcome and the predictors. However, there are two potential shortcomings of this method. First, as the observational study entails confounding bias due to the non-randomized treatment allocation, it is very important to adjust for the mismatch in the confounder distribution among the treatment groups, which the method above fails to address. Second, in presence of more than one predictor, which is a very common scenario, one would extend the model by considering multivariate spline in the imputation model. But such approach would be subject to the curse of dimensionality in the high dimensional setting and fitting a nonparametric regression model would be difficult under this circumstance.

The penalized spline of propensity prediction (PSPP) method (Little and An, 2004; Zhang and Little, 2009) overcomes both of these limitations by (1) employing a propensity score model to fit the treatment based on the covariates and by (2) restricting the spline in the penalized spline model (5.3) to the propensity score. PSPP model has three variants - (1) maximum likelihood (PSPP-ML): parameters are estimated using maximum likelihood method and the information matrix or bootstrap approach is employed to obtain the standard error, (2) Bayes (PSPP-B): posterior distribution is used to draw parameters and inference is based on standard Bayesian approach and (3) multiple imputation (PSPP-MI): missing values are imputed multiple times and Rubin's (1987) MI combining method is used for drawing inference. We base our method on PSPP framework and develop a novel approach to predict the optimal treatment of the new patient based on the imputed missing potential

outcomes. In addition, Bayesian posterior predictive inference is employed for quantifying the uncertainty in estimating the missing outcomes.

5.2.2 Bayesian Inference

Likelihood Model. We employ a joint modeling framework consisting of two parts. In the first part, we consider a propensity score (PS) model where we fit the propensity of each treatment assignment based on observed covariates. Specifically, we assume the following PS model for the distribution of A given the observed covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ and the regression parameters $\boldsymbol{\gamma}$

$$\text{logit}\left(P\left(A = 1|\mathbf{X}, \boldsymbol{\gamma}\right)\right) = \mathbf{X}'\boldsymbol{\gamma} = e(\mathbf{X}, \boldsymbol{\gamma}) \text{ (say)} \tag{5.4}$$

The regression method depends on the class of treatment variable. For the scenario of more than one treatment, one could also use a different regression method such as a polytomous regression model.

The second part of the model is developed in the outcome stage where we consider two separate models for the potential outcomes Y^a under both treatment 0 and 1. For each treatment group, we fit the potential outcomes Y^a based on the observed covariates $\mathbf{X}$ and the logit of the propensity of belonging to a particular treatment group a, say $e_a(\mathbf{X}, \boldsymbol{\gamma})$. So, for a patient receiving treatment 1, $e_1(\mathbf{X}, \boldsymbol{\gamma}) = e(\mathbf{X}, \boldsymbol{\gamma})$, but for a patient with treatment 0, $e_0(\mathbf{X}, \boldsymbol{\gamma}) = \text{logit}(1 - \text{logit}^{-1}(e(\mathbf{X}, \boldsymbol{\gamma})))$. Including the propensity score in the outcome model not only facilitates adjustment for the confounding bias, but also addresses the issue arising from high dimensionality, in case there are many predictors. However, the true relationship between the propensity score and the outcome is not known. Following Zhang and Little (2009), we include a non-parametric spline regression function of $e_a(\mathbf{X}, \boldsymbol{\gamma})$ in the

mean function $\mu(\cdot)$ of potential outcome model under treatment a. Formally, the outcome model can be expressed in the following way,

$$(Y^a|\mathbf{X};\boldsymbol{\theta}_a) = \mu(\mathbf{X}, A = a;\boldsymbol{\theta}_a) + \epsilon \tag{5.5}$$

where,

$$\mu(\mathbf{X}, A = a;\boldsymbol{\theta}_a) = s(e_a(\mathbf{X},\boldsymbol{\gamma});\boldsymbol{\alpha}_a) + g\left(e_a(\mathbf{X},\boldsymbol{\gamma}), X_2, \ldots, X_p;\boldsymbol{\beta}_a\right)$$

with $\boldsymbol{\theta}_a$ being all unknown parameters in the model. $s(e_a(\mathbf{X},\boldsymbol{\gamma});\boldsymbol{\alpha}_a)$ is a penalized spline with fixed knots and parameters $\boldsymbol{\alpha}_a$, $g(\cdot)$ is an augmented parametric function of other covariates with parameters $\boldsymbol{\beta}_a$. ϵ is the normal random error associated with the outcome model such that $\epsilon \sim N(0, \sigma_y^2)$. Note that, including the propensity score along with other covariates in $g(\cdot)$ would involve the issue of multicollinearity. Therefore, without loss of generality, X_1 is excluded from $g(\cdot)$. For $a \in \{0, 1\}$ we assume the following q degree polynomial basis for $s(e_a;\alpha_a)$,

$$s(e_a(\mathbf{X},\boldsymbol{\gamma});\boldsymbol{\alpha}_a) = \sum_{j=0}^{q} \alpha_{1ja} e_a(\mathbf{X},\boldsymbol{\gamma})^j + \sum_{k=1}^{K} \alpha_{2ka}(e_a(\mathbf{X},\boldsymbol{\gamma}) - t_{ka})_+^q \tag{5.6}$$

where, $\boldsymbol{\alpha}_a = (\boldsymbol{\alpha}'_{1a}, \boldsymbol{\alpha}'_{2a})'$ with $\boldsymbol{\alpha}_{1a} = (\alpha_{10a}, \alpha_{11a}, \ldots, \alpha_{1qa})'$ representing the fixed effect parameters and $\boldsymbol{\alpha}_{2a} = (\alpha_{21a}, \alpha_{22a}, \ldots, \alpha_{2Ka})'$ denoting random effect parameters with $\boldsymbol{\alpha}_{2a} \sim N_K\left(\mathbf{0}, \sigma^2_{\alpha_2}\mathbf{I}_K\right)$. $t_{1a} < t_{2a} < \cdots < t_{Ka}$ are K knots selected on the range of $e_a(\mathbf{X},\boldsymbol{\gamma})$ and the truncated power function is defined as $f(x) = x_+^q = x^q I(x > 0)$.

For the parametric function $g(\cdot)$ we assume a linear form

$$g(e_a(\mathbf{X},\boldsymbol{\gamma}), X_2, \ldots, X_p;\boldsymbol{\beta}_a) = \beta_{1a} e_a(\mathbf{X},\boldsymbol{\gamma}) + \sum_{j=2}^{p} \beta_{ja} X_j$$

where, $\boldsymbol{\beta}_a = (\beta_{1a}, \ldots, \beta_{pa})'$ are fixed effects. This specification only allows to adjust for the main effects of the covariates. However, if necessary, interaction between $e_a(\mathbf{X}, \boldsymbol{\gamma})$ and the covariates can also be included in the model.

Prior Specification. For the propensity score model parameters, $\boldsymbol{\gamma}$, we assume flat normal priors $\boldsymbol{\gamma} \sim N_{p+1}\left(\boldsymbol{\mu}_\gamma^\pi, \boldsymbol{\Sigma}_\gamma^\pi\right)$ where, $\boldsymbol{\mu}_\gamma^\pi$ and $\boldsymbol{\Sigma}_\gamma^\pi$ are pre-specified hyper-parameters. Specifically, we chose $\boldsymbol{\mu}_\gamma^\pi = \mathbf{0}$ and $\boldsymbol{\Sigma}_\gamma^\pi = 10^3\mathbf{I}_{p+1}$. Following the prior specifications in the Bayesian p-spline regression model (Berry et al., 2002; An and Little, 2008), we assume a flat prior for the coefficients in monomial basis part of the spline function, $\boldsymbol{\alpha}_{1a}$ such that $\boldsymbol{\alpha}_{1a} \propto 1$. For the parametric function $g(\cdot)$ we also assume a normal prior for the coefficients, $\boldsymbol{\beta}$ with high variance values. Specifically, we consider $\boldsymbol{\beta} \sim N_p\left(\mathbf{0}, 10^3\mathbf{I}_p\right)$. We assume flat priors for $\sigma_{\alpha_2}^{-2}$, the precision parameter of the random basis coefficients in the truncated part of the spline function $s(\cdot)$, and also for σ_y^{-2}, the precision parameter of the potential outcome imputation model. Specifically, we chose $\sigma_{\alpha_2}^{-2} \sim \text{Gamma}(0.001, 0.001)$ and $\sigma_y^{-2} \sim \text{Gamma}(0.001, 0.001)$.

5.2.3 Optimal Treatment Allocation

The optimal treatments for PSBayes method can be estimated in the similar fashion as described for the BayesG approach. We construct a Bayesian methodology to predict the mean potential outcome for the new patient under treatment a for a given $\tilde{\mathbf{X}}$ and the observed data $\mathbf{D}$ which is formally expressed as $\tilde{\mu}(a) = E(Y^a|\tilde{\mathbf{X}}, \mathbf{D}) = s(e_a(\tilde{\mathbf{X}}, \boldsymbol{\gamma}); \boldsymbol{\alpha}_a) + g\left(e_a(\tilde{\mathbf{X}}, \boldsymbol{\gamma}), X_2, \ldots, X_p; \boldsymbol{\beta}_a\right)$ and subsequently to estimate $a^\star$ based on the framework we

described before. As discussed in section 4.2.5 the mean posterior predictive potential outcome under treatment a for the new patient can be formulated as

$$\tilde{\mu}(a) = E(\tilde{Y}^a|\tilde{\mathbf{X}}, \mathbf{D}) = \int_{\boldsymbol{\theta}_a} E(\tilde{Y}^a|\tilde{\mathbf{X}}, \boldsymbol{\theta}_a) p(\boldsymbol{\theta}_a|\mathbf{D}) d\boldsymbol{\theta}_a$$

where, $\boldsymbol{\theta}_a$ is the set of unknown parameters in the mean function $\mu(\cdot)$ for treatment group a, $\boldsymbol{\theta}_a = \{\boldsymbol{\gamma}, \boldsymbol{\alpha}_{1a}, \boldsymbol{\alpha}_{2a}, \boldsymbol{\beta}_a, \sigma^{-2}_{\alpha_{2a}}, \sigma^{-2}_y\}$.

The posterior predictive computation can be performed similarly as is done for BayesG approach. We can estimate the optimal treatment decision in many different ways. We follow a fully Bayesian approach where, similar to the notion of stochastic ordering, we compare the treatment options for each MCMC draw from the posterior distribution. We estimate the probability of potential outcome under treatment 1 to be higher than potential outcome under treatment 0 by the following formulation, $P\left(\tilde{\mu}(1) > \tilde{\mu}(0)|\tilde{\mathbf{X}}, \mathbf{D}\right)$ by $\hat{P} = \frac{1}{N}\sum_{i=1}^{N} I(\tilde{\mu}_i(1) > \tilde{\mu}_i(0))$. Then the optimal treatment can be derived by putting a suitable threshold on $\hat{P}$. Using an threshold of 0.5 the optimal treatment can be estimated by

$$\hat{a}^\star = I\left(\hat{P} > 0.5\right).$$

5.3 A Bayesian Nonparametric Specification

The OS model of the PSBayes method estimates regression models for the potential outcome Y^a on (1) a spline of the logit of the propensity score of treatment group a and (2) a linear function of other covariates that are predictive of Y and we assume the error distribution to be normal. Despite specifying a nonparametric function for the mean model, this formulation would be able to capture only a specific class of true models where the mean function is additive in predictors with an unimodal symmetric outcome distribution as we

assume a linear formulation of the parametric $g(\cdot)$ function with gaussian error distribution. However, a flexible non-parametric specification of the outcome distribution is often desired, particularly in presence of multiple modes or skewness. Keeping this in mind, we develop a flexible Bayesian nonparametric model (BNP) based on Dependent Dirichlet Process (DDP) (MacEachern, 2000) for the potential outcome Y^a. Our specification is similar to that of Roy et al. (2016); Xu et al. (2016) in the BNP part, but their objectives were different from ours.

To be specific, let us assume that $Y^a|\mathbf{X} \sim F(\cdot|A, \mathbf{X})$. Unless a parametric setup is specified, F belongs to a family of probability models which sets the basis of nonparametric inference. Under the Bayesian paradigm, we assume a prior distribution on $F(\cdot|A, \mathbf{X})$ in the form of $\pi(F(\cdot|A, \mathbf{X}))$ which is a probability model for the infinite dimensional $F(\cdot|A, \mathbf{X})$. Here we specifically consider a Dirichlet Process (DP) prior for $F(\cdot|A, \mathbf{X})$. Following Xu et al. (2016) we first start the DDP specification with a model for a discrete model distribution $G(\cdot)$. Then we use a Gaussian kernel to specify a DP prior model for the continuous random distribution $F(\cdot)$. Next, we specify the mean of the Gaussian kernels through the mean function to extend it to the prior distribution of $F(\cdot|A, \mathbf{X})$ (5.5). In the following subsections we briefly review the Dirichlet Process Prior models and detail the construction of $G(\cdot)$, $F(\cdot)$ and $F(\cdot|A, \mathbf{X})$ following the frameworks of Müller and Rodriguez (2013); Müller et al. (2016).

5.3.1 Dependent Dirichlet Process

Dirichlet Process. Dirichlet process prior is one of the very popular BNP models which was introduced by Ferguson (1973) in the context of prior distribution on the space of random probability measures (RPM). Formally, let $\mathcal{S}$ be a space and let $\mathcal{B}$ be a σ field of subsets of $\mathcal{S}$. Further let G_0 be a probability measure on $(\mathcal{S}, \mathcal{B})$ and let $M > 0$. Then G, defined on

$(\mathcal{S}, \mathcal{B})$ is a Dirichlet process with parameters (M, G_0) if it assigns probability $G(B)$ to every measurable set B such that for each (measurable) finite partition $\{B_1, B_2, \ldots, B_k\}$ of $\mathcal{S}$, the joint distribution of the random probability vector $\{G(B_1), G(B_2), \ldots, G(B_k)\}$ has a Dirichlet distribution with parameters $\{MG_0(B_1), MG_0(B_2), \ldots, MG_0(B_k)\}$. The parameter M is called the concentration or precision parameter, while G_0 is known as the base measure of the DP. For a large M, G is highly concentrated about G_0 and as $M \to \infty$ the DP process G converges to G_0.

There are many important properties of the DP. First, due to its discrete nature, $G(\cdot)$ can be expressed as a weighted sum of point masses such as $G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot)$ where, $\delta_{\theta_h}(\cdot)$ denote point mass at θ_h and w_h is the corresponding probability weight. As we will discuss shortly, this specification leads to a specific characterization of DP known as the stick breaking construction. Another important property is that the closure of the support of DP includes the space of all distributions with the same support as the base measure G_0 (Müller et al., 2016), which is why the DP prior model is extremely useful for density estimation under iid sampling from an unknown distribution (Ishwaran and James, 2001).

Stick Breaking Construction. Sethuraman (1994) provided an alternative specification of the DP model which is based on its discrete nature such as $G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot)$ where, θ_1, θ_2, ... are iid draws from the base measure G_0 and w_h is defined as $w_h = v_h \prod_{l<h}(1 - v_l)$ with $v_h \overset{iid}{\sim} \text{Beta}(1, M)$. The draw of w_h can be imagined as repeatedly breaking a fraction v_h part of from a stick of originally unit length. Hence this representation was coined as "stick breaking" construction. This specification is useful to generate random samples from a DP prior.

Dirichlet Process Mixture. The DP model G is discrete with probability 1. For continuous density estimation, this limitation can be avoided by specifying continuous kernels in place of point mass $\delta_{\theta_h}(\cdot)$ i.e. by adding a convolution with a continuous kernel to the

discrete distribution G (Ferguson, 1983; Lo, 1984; Escobar, 1990, 1994; Escobar and West, 1995). The Dirichlet Process mixture (DPM) prior on F can be semi-parametrically represented as

$$F(\cdot) = \int p(\cdot|\theta)G(d\theta),\ G \sim \mathrm{DP}(M, G_0) \tag{5.7}$$

where, $p(\cdot|\theta)$ is a parametric distribution characterized by finite dimensional parameter θ and $G(\cdot)$ is a probability distribution on θ. Such mixture representation can constitute a rich family of distributions. For example, Lo (1984) illustrated that a DP location-scale mixture of normals with mean μ and variance σ^2,

$$F(\cdot) = \int N(\cdot|\mu, \sigma^2)G(d\mu, d\sigma^2),\ G \sim \mathrm{DP}(M, G_0)$$

has support on the space of all absolutely continuous distributions. Furthermore, a mixture with respect to location and scale $\theta = (\mu, \sigma)$ in a kernel $k(y|\mu, \sigma) = \sigma^{-1}k((y-\mu)/\sigma)$ for a fixed density k,

$$F(\cdot) = \int \sigma^{-1}k(\cdot|\mu, \sigma)G(d\mu, d\sigma),\ G \sim \mathrm{DP}(M, G_0)$$

approximates any density in the L_1 sense under some mild condition. Similarly, a mixture of uniform distributions with symmetric support around 0,

$$F(\cdot) = \int U(\cdot|-\theta, \theta)G(d\theta),\ G \sim \mathrm{DP}(M, G_0)$$

where, $U(\cdot|-\theta, \theta)$ is a density of an uniform random variable on $[-\theta, \theta]$, has full support on the space of all unimodal symmetric distributions (Müller and Rodriguez, 2013).

The representation in (5.7) can be alternatively characterized with the help of stick breaking construction of the Dirichlet process as

$$F(\cdot) = \sum_{h=1}^{\infty} w_h p(\cdot|\theta_h).$$

This representation essentially presents DPM as a discrete mixture model where, $p(\cdot|\theta_h)$ is the parametric distribution of the h^{th} component of the mixture characterized by θ_h with w_h being the corresponding probability weight, $\theta_h \sim G_0$, $w_h = v_h \prod_{h'<h}\{1 - v_{h'}\}$ and $v_h \sim \text{Beta}(1, M)$. For example, if we use a normal kernel $N(\cdot|\mu, \sigma)$ in place of the parametric distribution $p(\cdot|\theta)$, the DPM with location mixture can be expressed as

$$F(\cdot) = \sum_{h=1}^{\infty} w_h N(\cdot|\mu_h, \sigma).$$

where, $\mu_h \sim G_0$ and other parameters are same as before. This normal mixture characterization is very useful for density estimation under iid sampling from an unknown distribution and it provides a good fit to the data from a wide range of scenarios as virtually any distribution can be closely approximated by the mixture of normal distributions (Ishwaran and James, 2001).

Dependent Dirichlet Process. In the context of non-parametric regression, we are often interested in inference for a family of probability models indexed by the covariate $\mathbf{x}$, $\mathcal{F} = \{F(\cdot|\mathbf{x}), \mathbf{x} \in X\}$. For BNP regression with non-parametric mean function specification, one would consider a mixture of a standard parametric residual distribution and a BNP prior on the mean function, whereas a full non-parametric regression would put the BNP prior on the family of probability models, $\mathcal{F}$ (Müller and Rodriguez, 2013). In this context, MacEachern (2000) extended the DP prior for a single random probability measure to the desired family of random probability measures $p(F(\cdot|\mathbf{x}),\ \mathbf{x} \in X)$ on the $\mathcal{F}$ space which is

named as Dependent Dirichlet Process (DDP). The characterization of the DDP follows the stick breaking representation of the DP. Specifically in the regression setting, a DP prior for $F(\cdot|\mathbf{x})$ can be expressed as

$$F(\cdot|\mathbf{x}) = \sum_{h=1}^{\infty} w_h p(\cdot|\theta_h(\mathbf{x})). \tag{5.8}$$

$\theta_h(\mathbf{x}) \sim G^{\star}_{\mathbf{x}}$, independent across h where $G^{\star}_x$ is the base measure characterized by $\mathbf{x}$ and $w_h = v_h \prod_{h'<h}\{1 - v_{h'}\}$ with $v_h \sim \text{Beta}(1, M)$. Marginalizing over $\theta_h(\mathbf{x})$, this specification ensures that $F(\cdot|\mathbf{x}) \sim \text{DP}(G^{\star}_{\mathbf{x}}, M)$. One key feature of this characterization is that it allows users to introduce dependence of $\theta_h(\mathbf{x})$ across $\mathbf{x}$ and by imposing a dependent prior on $\{\theta_h(\mathbf{x})\}$ one can create dependent random probability measures $F(\cdot|\mathbf{x}, \mathbf{x} \in \mathbf{X})$ (Müller and Rodriguez, 2013). MacEachern (2000) proposed a Gaussian Process prior on $\{\theta_h(\mathbf{x}), \mathbf{x} \in \mathbf{X}\}$ as a default choice. However, other models can also be used. The default specification of DDP model (5.8) is sometimes called as variable location DDP as only locations $\theta_h(\mathbf{x})$ are characterized by $\mathbf{x}$. Generalizing this idea, the weights w_h or both weights and locations could be indexed with $\mathbf{x}$ too, leading to variable weight and location DDP (Müller and Rodriguez, 2013).

5.3.2 Proposed Specification

Likelihood Model. Our objective is to propose a flexible structure for the imputation model so that it can provide a good fit to a wide range of situations. To this direction, we propose a Bayesian semiparametric specification (PSBayes-DDP) for the OS part in our likelihood model. Our proposed model offers flexibility both in terms of the mean function and also in the residual distribution.

Similar to the original PSBayes methodology, here we stick to the joint modeling framework with PS and OS models. For the PS model involving the distribution of A given the

observed covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ and the regression parameters $\boldsymbol{\gamma}$ we assume the same specification (5.4) as before.

On the contrary to the previously proposed OS model, here we assume $(Y^a|\mathbf{X}) \sim F(y|\mathbf{X}, A)$, where $F(\cdot|\mathbf{X}, A)$ belongs to a family of probability models $\mathcal{F} = \{F(\cdot|\mathbf{x}, A);\ \mathbf{x} \in \mathbf{X},\ A \in \{0, 1\}\}$ characterized by $\mathbf{x}$ and a. On the basis of the previous discussion, we can impose a dependent DP mixture prior on $F(y^a|\mathbf{X})$ to include the regression on covariates and treatment. Following the stick breaking construction, we specify

$$F(Y^a|\mathbf{X}) = \sum_{h=1}^{\infty} w_h N(Y; \mu_h(\mathbf{X}, A), \sigma_y^2) \tag{5.9}$$

where, $w_h; h = 1, \ldots, \infty$ are the DP mixture weights, $\mu_h(\mathbf{X}, A)$ is a function of treatment and covariates which endows the h^{th} component of the mixture with the same formulation of the mean outcome model as we specified before. In formal notations $\mu_h(\cdot)$ is defined as follows,

$$\mu_h(\mathbf{X}, A) = s(e_a(\mathbf{X}, \boldsymbol{\gamma}); \boldsymbol{\alpha}_a^h) + g\left(e_a(\mathbf{X}, \boldsymbol{\gamma}), X_2, \ldots, X_p; \boldsymbol{\beta}_a^h\right)$$

where, $s(e_a(\mathbf{X}, \boldsymbol{\gamma}); \boldsymbol{\alpha}_a^h)$ is the similarly defined penalized spline with fixed knots and parameters and $\boldsymbol{\alpha}_a^h$, $g(\cdot)$ is an augmented parametric function of other covariates with parameters $\boldsymbol{\beta}_a^h$ corresponding to h^{th} component of the mixture. Note that, the formulation in (5.9) provides us with an infinite mixture of normals, with weight w_h corresponding to the h^{th} mixture component such that $\sum_{h=1}^{\infty} w_h = 1$ and is very flexible to deal with situations like multiple modes or skewness in the outcome distribution (Roy et al., 2016).

Similar to the original specification, for $a \in \{0, 1\}$ we assume the following q degree polynomial basis for $s(e_a; \alpha_a^h)$,

$$s(e_a(\mathbf{X}, \boldsymbol{\gamma}); \boldsymbol{\alpha}_a^h) = \sum_{j=0}^{q} \alpha_{1ja}^h e_a(\mathbf{X}, \boldsymbol{\gamma})^j + \sum_{k=1}^{K} \alpha_{2ka}^h (e_a(\mathbf{X}, \boldsymbol{\gamma}) - t_{ka})_+^q \tag{5.10}$$

where, $\boldsymbol{\alpha}_a^h = \left(\boldsymbol{\alpha}_{1a}^{h\,\prime}, \boldsymbol{\alpha}_{2a}^{h\,\prime}\right)'$ with $\boldsymbol{\alpha}_{1a}^h = \left(\alpha_{10a}^h, \alpha_{11a}^h, \ldots, \alpha_{1qa}^h\right)'$ representing the fixed effect parameters and $\boldsymbol{\alpha}_{2a}^h = \left(\alpha_{21a}^h, \alpha_{22a}^h, \ldots, \alpha_{2Ka}^h\right)'$ denoting random effect parameters with $\boldsymbol{\alpha}_{2a}^h \sim N_K\left(\mathbf{0}, \sigma_{\alpha_2}^2 \mathbf{I}_K\right)$. $t_{1a} < t_{2a} < \cdots < t_{Ka}$ are K knots selected similarly on the range of $e_a(\mathbf{X}, \boldsymbol{\gamma})$.

For the parametric function $g(\cdot)$ we assume a linear form

$$g(e_a(\mathbf{X}, \boldsymbol{\gamma}), X_2, \ldots, X_p; \boldsymbol{\beta}_a^h) = \beta_{1a}^h e_a(\mathbf{X}, \boldsymbol{\gamma}) + \sum_{j=2}^{p} \beta_{ja}^h X_j$$

where, $\boldsymbol{\beta}_a^h = \left(\beta_{1a}^h, \ldots, \beta_{pa}^h\right)'$ are fixed effects. This specification only allows to adjust for the main effects of the covariates.

Prior Specification. The prior distribution for the weights, $w_h = v_h \prod_{h'<h}(1 - v_{h'})$ is specified as $v_h \sim \text{Beta}(1, M)$, M being the precision parameter. For the prior distribution of M, we assume $M \sim \text{Inv-Gamma}(1, 1)$ following the suggestion by Roy et al. (2016). This prior has longer tail than Gamma$(1, 1)$ and provides stability in posterior computation. For the propensity score model parameters, we assume same priors as before. We also chose same prior for each component of OS parameters. Specifically, we assume $\boldsymbol{\alpha}_{1a}^h \propto 1$. For the parametric function $g(\cdot)$ we also assume $\boldsymbol{\beta}^h \sim N_p\left(\mathbf{0}, 10^3 \mathbf{I}_p\right)$. All the precision parameter priors are remained same.

5.3.3 Optimal Treatment Allocation

In the proposed PSBayes-DDP methodology, the potential outcome is expressed as a discrete mixture of infinitely many normal distributions. We develop a Bayesian methodology to predict $\tilde{Y}^a$ and subsequently to estimate $a^\star$ based on the framework we described before.

As discussed in section 4.2.5 the posterior predictive potential outcome under treatment a for the new patient can be formulated as

$$p\left(\tilde{Y}^a|\tilde{\mathbf{X}},\mathbf{D}\right)=\int_{\boldsymbol{\theta}_a} p\left(\tilde{Y}^a|\boldsymbol{\theta}_a,\tilde{\mathbf{X}}\right)p\left(\boldsymbol{\theta}_a|\mathbf{D}\right)d\boldsymbol{\theta}_a$$

where, $\boldsymbol{\theta}_a$ is the set of unknown parameters for treatment group a, $\boldsymbol{\theta}_a=\{\boldsymbol{\gamma},\boldsymbol{\alpha}_{1a},\boldsymbol{\alpha}_{2a},\boldsymbol{\beta},\sigma^{-2}_{\alpha_{2a}},\sigma^{-2}_{y}\}$.

For the posterior predictive computation, first we draw N MCMC samples of $\boldsymbol{\theta}_a$, (say, $\boldsymbol{\theta}^{\star}_{a1},\ldots,\boldsymbol{\theta}^{\star}_{aN}$) for $a=\{0,1\}$, from their joint posterior distribution $p\left(\boldsymbol{\theta}_a|\mathbf{D}\right)$. Then, for each drawn MCMC sample of $\boldsymbol{\theta}^{\star}_{a}$, we impute the potential outcome of the new patient say, $\tilde{Y}^{a^*}$ by drawing a sample from $p\left(\tilde{Y}^a|\boldsymbol{\theta}_a,\tilde{\mathbf{X}}\right)$. Given all imputed potential outcomes, $\tilde{Y}^{a^*}_1,\tilde{Y}^{a^*}_2,\ldots,\tilde{Y}^{a^*}_N$ we can construct empirical predictive distribution of the potential outcomes for the new patient.

Based on the empirical posterior predictive distribution of the potential outcomes, we can estimate the optimal treatment allocation in many different ways. For example, we can choose $a^{\star}$ following the recommendation of Gelfand and Ghosh (1998) by minimizing a suitable posterior predictive loss function. Considering the average loss function, it can be formulated as

$$a^{\star}=I\left(E(\tilde{Y}^1|\tilde{\mathbf{X}},\mathbf{D})>E(\tilde{Y}^0|\tilde{\mathbf{X}},\mathbf{D})\right)$$

which can be estimated by

$$\hat{a}^{\star}=I\left(\bar{\tilde{Y}}^{1^*}>\bar{\tilde{Y}}^{0^*}\right)$$

where, $\bar{\tilde{Y}}^a=\frac{1}{N}\sum_{i=1}^{N}\tilde{Y}^{a^*}_i$ with $\tilde{Y}^{a^*}_i$ being the imputed potential outcome under treatment a of the new patient corresponding to i^{th} MCMC draw.

Based on the empirical posterior predictive distribution of the potential outcomes, we adopt a fully Bayesian stochastic ordering appraoch which allows us to compare the treatment options in each draw and estimate the probability of potential outcome under treat-

ment 1 to be higher than potential outcome under treatment 0, $P\left(\tilde{Y}^1 > \tilde{Y}^0 | \tilde{\mathbf{X}}, \mathbf{D}\right)$ by $\hat{P} = \frac{1}{N}\sum_{i=1}^{N} I(\tilde{Y}_i^{1^*} > \tilde{Y}_i^{0^*})$. Then the optimal treatment can be derived by a suitable threshold on $\hat{P}$. Using an threshold of 0.5 the optimal treatment can be estimated by

$$\hat{a}^\star = I\left(\hat{P} > 0.5\right).$$

5.4 Model Feedback

5.4.1 Introduction

In observational studies treatments are selected based on the patient characteristics. Because of the non-randomized allocation, methods proposed in observational study literature often consider treatment as stochastic variable and postulate the propensity score model on the treatment surface to analyze the dependence of treatment allocation on observed covariates. In point treatment settings, the propensity score (PS) model can be employed to estimate the probability of allocation of certain treatment option as a function of the observed covariates. In randomized study, treatment comparison is direct as the distribution of patient covariates in the treatment groups are fairly similar. However, in non randomized design such as observational studies direct comparison between treatment groups may not be feasible since the patient characteristics differ across different treatment groups. Originating from survey sampling methods, there are two major streams of approaches to treatment comparison in observational studies. Among them, one group of methods rely on a two stage approach where the analysis is carried out in two stages. In the first stage, a propensity score model is fitted to predict the probability of receiving certain treatment option for each patient. Typically a logistic regression model of the treatment variable based on the

measured confounders is employed for this purpose. In the second stage of the method various approaches are proposed in the literature depending on the objective of the study. To adjust for confounding by the measured covariates Rosenbaum and Rubin (1983) advocated stratification of patients based on quantiles of the PS values. In stead of directly comparing the outcomes between different treatment groups, the outcomes from treatment groups are matched within same PS quantile values. In the regression setting, a regression model is employed on the response surface to compare the treatment options by matching the outcomes from subjects with similar PS values. In the context of treatment causal effect estimation, a weighted regression analysis is often performed on the response surface in the second stage where the weights are obtained from the fitted PS values. A second group of methods such as the linear weight in prediction (LWP) (Scharfstein et al., 1999; Bang and Robins, 2005) similarly use PS model to estimate the propensity score, but instead of performing a weighted analysis, they treat the weights as covariate and thus adjusts for covariate imbalance. No matter what weighting method is used, in the traditional frequentist approach the first stage and second stage models are fitted sequentially where the estimated PS values are treated as fixed quantities in the second stage.

On the other hand, the Bayesian methods have a natural appeal in this setting as combining two separate models into a joint likelihood model would facilitate simultaneous estimation of PS and OS models. A major advantage of the Bayesian methods is that by jointly modeling the treatment-covariate and outcome-covariate-treatment associations the uncertainty in propensity score estimation can be incorporated in the OS stage, whereas in frequentist counterpart treatment effect estimation does not acknowledge uncertainty in propensity score which might make falsely precise confidence interval for treatment effect. McCandless et al. (2009) proposed one such method where the propensity score is treated as a latent variable. The striking feature of their method is that the posterior estimation

of marginal treatment effect incorporates treatment selection uncertainty which is averaged out over the latent variable in the outcome stage of the model.

One implicit characteristic of the joint modeling framework in the Bayesian PS estimation is that the posterior inference allows the PS parameters to get updated by the OS model information. This notion, known as "model feedback" (Zigler et al., 2013) from OS to PS model has raised some concerns against the use of Bayesian method in the propensity score setting. Rubin (2007, 2008) argued that PS models are implemented to approximate the design of randomized study and should be completely independent of the outcome. Zigler et al. (2013) demonstrated that the flow of information from OS model to PS model weakens the property of PS model to adjust for confounding which leads to biased estimate of the treatment effect. However, the issue is partially addressed if we include individual covariates in the outcome model in addition to PS adjustment.

5.4.2 An Example

Zigler et al. (2013) illustrated the effect of model feedback on treatment causal effect estimation using a simple example with binary treatment and binary outcome in the regression setting similar to that of (McCandless et al., 2009). For A and Y, we follow the same notations as before, but for X, we only use bold font when it includes patients' information. Here also, we consider $A \in \{0, 1\}$. Unlike Zigler et al. (2013), we are assuming the outcome variable to have continuous observations. $\mathbf{X}$ denotes a $n \times p$ covariate matrix containing information on n patients with p covariates. $\mathbf{A}$ denotes the treatment vector of n patients and $\mathbf{Y}$ represents the outcome information of all patients.

The first stage of sequential PS methods involves a model for the probability of $A = 1$ based on a predetermined function of observed covariates X. For the sake of illustration we

consider a probit regression model with parameters $\boldsymbol{\gamma}$ which gives the following likelihood of the PS model

$$l\left(\boldsymbol{\gamma}|\mathbf{A},\mathbf{X}\right)=\prod_{i=1}^{n}\left\{\Phi^{-1}(\mathbf{X}_i'\boldsymbol{\gamma})\right\}^{A_i}\left\{1-\Phi^{-1}(\mathbf{X}_i'\boldsymbol{\gamma})\right\}^{1-A_i} \tag{5.11}$$

where, $\Phi(\cdot)$ is the normal CDF, $i=1,\ldots,n$ denote indices of n patients and $\{Y_i,\mathbf{X}_i,A_i\}$ denote information on i^{th} patient with $\mathbf{X}_i=(X_{i1},\ldots,X_{ip})'$.

For the OS part, propensity score plays a key role by embedding treatment-covariate relationship information into the outcome model. Keeping that in mind, we assume the following normal regression model for the outcome stage where a function of the PS, say $e(X,\gamma)$ is included as a covariate along with the treatment variable A which is included to estimate treatment causal effect.

$$Y|A,X,\boldsymbol{\gamma},\boldsymbol{\xi},\delta=\xi_0+\xi_1 e(X,\boldsymbol{\gamma})+\delta A+\epsilon \tag{5.12}$$

where, ξ_1 controls the adjustment of PS in the outcome model, δ is the treatment effect and ϵ is the random error assumed to follow $N(0,1)$. The functional form of $e(X,\boldsymbol{\gamma})$ determines how PS is adjusted in the outcome model. For example, if we specify $e(X,\boldsymbol{\gamma})=X'\boldsymbol{\gamma}$ the PS would be linearly adjusted in the outcome model (McCandless et al., 2009). Other than directly including PS into the outcome model, we can also specify a membership such as by its quantiles as discussed in Rosenbaum and Rubin (1983) in the context of matching methods. The OS model corresponds to the following likelihood function

$$l\left(\boldsymbol{\xi},\delta,\boldsymbol{\gamma}|\mathbf{Y},\mathbf{A},\mathbf{X}\right)\propto\exp\left\{-0.5\sum_{i=1}^{n}\left(Y_i-\xi_0-\xi_1 e(\mathbf{X}_i,\boldsymbol{\gamma})-\delta A_i\right)^2\right\} \tag{5.13}$$

In the traditional sequential PS methods, the estimation of δ is carried out in two separate stages. The estimates of $\boldsymbol{\gamma}$, say $\hat{\boldsymbol{\gamma}}$ are obtained from the PS likelihood (5.11) which is then plugged in the OS likelihood function (5.13) and δ is estimated from $l\left(\boldsymbol{\xi},\delta,\hat{\boldsymbol{\gamma}}|\mathbf{Y},\mathbf{A},\mathbf{X}\right)$.

Unlike the sequential approach of the frequentist method, the Bayesian counterpart follows from the joint likelihood of the parameters $\boldsymbol{\xi}, \delta, \boldsymbol{\gamma}$ combining both PS (5.11) and OS model (5.13) likelihood functions.

$$l\left(\boldsymbol{\xi}, \delta, \boldsymbol{\gamma} | \mathbf{Y}, \mathbf{A}, \mathbf{X}\right) \propto \prod_{i=1}^{n} \left\{ \Phi^{-1}(\mathbf{X}_i'\boldsymbol{\gamma}) \right\}^{A_i} \left\{ 1 - \Phi^{-1}(\mathbf{X}_i'\boldsymbol{\gamma}) \right\}^{1-A_i} \tag{5.14}$$

$$\times \exp\left\{ -0.5\left(Y_i - \xi_0 - \xi_1 e(\mathbf{X}_i, \boldsymbol{\gamma}) - \delta A_i\right)^2 \right\} \tag{5.15}$$

Note that, in this joint formulation the PS parameter $\boldsymbol{\gamma}$ is considered stochastic along with other parameters and by inclusion of the PS function $e(X, \boldsymbol{\gamma})$ in the outcome model the uncertainty in PS gets channelized to the estimation of δ. However, under this setting, the posterior inference of $\boldsymbol{\gamma}$ involves both PS and OS model parameters leading the situation of model feedback. Under a similar framework, Zigler et al. (2013) provided a mathematical reasoning of this phenomenon following a latent variable formulation of probit link regression proposed by Albert and Chib (1993). According to their proposal, sampling the binary response A from the PS likelihood (5.11) is equivalent to drawing a latent truncated normal variable $A^\star$ with mean $X'\boldsymbol{\gamma}$ and variance 1. With this notion, assuming a flat prior distribution on the real line for all parameters, the posterior distribution of the model parameters in (5.14) and (5.15) is given by

$$p\left(\boldsymbol{\xi}, \delta, \boldsymbol{\gamma} | \mathbf{Y}, \mathbf{A}, \mathbf{A}^\star, \mathbf{X}\right) \propto \exp\left[-0.5\{\left(\mathbf{A}^\star - \mathbf{X}\boldsymbol{\gamma}\right)'\left(\mathbf{A}^\star - \mathbf{X}\boldsymbol{\gamma}\right) + \mathbf{Y}^{\star\prime}\mathbf{Y}^\star\}\right]$$

where, $\mathbf{Y}^\star = (\mathbf{Y} - \xi_0 \mathbf{1}_n - \xi_1 e(\mathbf{X}, \boldsymbol{\gamma}) - \delta\mathbf{A})$, 1_n is the n dimensional vector with all entries 1. From the above joint formulation, the full conditional distribution of $\boldsymbol{\gamma}$ is given by

$$p\left(\boldsymbol{\gamma}|\cdot\right) \propto \exp\left[-0.5\left\{\boldsymbol{\gamma}'\left(\mathbf{X}'\mathbf{X}\left(1+\xi_1^2\right)\right)\boldsymbol{\gamma} - 2\boldsymbol{\gamma}'\mathbf{X}'\{\mathbf{A}^\star + \xi_1\left(\mathbf{Y} - \xi_0\mathbf{1}_n - \delta\mathbf{A}\right)\}\right\}\right].$$

which corresponds to the kernel of a multivariate Normal distribution with mean $\{\mathbf{X}'\mathbf{X}\,(1+\xi_1^2)\}^{-1}\,[\mathbf{X}'\{\mathbf{A}^\star+\xi_1\,(\mathbf{Y}-\xi_0\mathbf{1}_n-\delta\mathbf{A})\}]$ and variance covariance matrix $\{\mathbf{X}'\mathbf{X}\,(1+\xi_1^2)\}^{-1}$. We can clearly see that the mean part is dependent on the outcome model and if $\xi_1 \neq 0$, the outcome information will influence the posterior update of $\boldsymbol{\gamma}$ thus causing model feedback.

Zigler et al. (2013) provided a more detailed account of the effect of model feedback. Assuming $e(X,\boldsymbol{\gamma}) = X'\boldsymbol{\gamma}$, the OS mean part of the joint likelihood function (5.15) can be expressed as a re-parameterization of the response surface model

$$\xi_0+\xi_1 e(X,\boldsymbol{\gamma})+\delta A = (\xi_0+\xi_1\gamma_0)+\xi_1\gamma_1 X_1+\xi_2\gamma_2 X_2+\cdots+\xi_p\gamma_p X_p+\delta A \tag{5.16}$$

This re-parameterization illustrates the fact that in this formulation the outcome model fails to reflect the association between the outcome and the individual covariates on the response surface, rather it conditions on an one dimensional summary of the covariates in the form of PS. Because of this dimension reduction, the OS stage parameters of the true data generating model cannot be fully identified by the analysis model. To elaborate on this fact, Zigler et al. (2013) considered the following true models on the treatment-covariate and outcome-covariate surfaces

$$\text{PS: } \Phi^{-1}\{P(A=1|X,\boldsymbol{\gamma})\} = \gamma_0+\gamma_1 X_1+\cdots+\gamma_p X_p \tag{5.17}$$

$$\text{OS: } Y|X,A,\boldsymbol{\alpha},\delta = \alpha_0+\alpha_1 X_1+\cdots+\alpha_p X_p+\delta A+\epsilon \tag{5.18}$$

where, α_k reflects true association between outcome and k^{th} covariate. Note that, the joint likelihood functions (5.14)-(5.15) can only recover the true treatment-covariate surface in (5.17), but fail to estimate individual covariate-outcome association in (5.18) as it only adjusts for the PS. Combining the above data generating model for the OS part with (5.16) we get $\alpha_0 = \xi_0+\xi_1\gamma_0,\ \alpha_1 = \xi_1\gamma_1,\cdots,\ \alpha_p = \xi_p\gamma_p$. Thus the analysis outcome model

would correspond to the OS part of the data generating model only if $\alpha_k = \xi_1\gamma_k$ for $k \neq 0$. As Zigler et al. (2013) pointed out, if this relationship does not hold true then feedback from the outcome model would affect the posterior estimation of $\boldsymbol{\gamma}$ in such a way that the true association between treatment and covariates (5.17) cannot be recovered leading to the violation of balancing score property of PS. On the other hand, in frequentist approach this issue does not arise as the outcome model parameters do not update the PS model parameters reflecting the true assignment mechanism to ensure balancing score property.

5.4.3 Model Feedback in PSBayes

Here we concentrate on the model feedback property in the context of original PSBayes specification. Similar to the previous example, we are assuming two models - one at the PS stage and another for the outcome stage. In the PS model we include all covariates so that we can recover the entire covariate-treatment surface through the propensity score. However, for the sake of illustration, we assume the following probit model in the PS model specification

$$\Phi^{-1}\left(P\left(A = 1|\mathbf{X}, \gamma\right)\right) = \mathbf{X}'\boldsymbol{\gamma}$$

which corresponds to the same PS likelihood model as in (5.11). Following Albert and Chib (1993), the same likelihood corresponds to the specification of a latent vector $\mathbf{A}^\star$ with mean vector $\mathbf{X}\boldsymbol{\gamma}$ and variance covariance matrix $\mathbf{I}_n$ which leads to the following likelihood model

$$p\left(\boldsymbol{\gamma}|\mathbf{A}, \mathbf{A}^\star, \mathbf{X}\right) \propto \exp\left[-0.5\{\left(\mathbf{A}^\star - \mathbf{X}\boldsymbol{\gamma}\right)'\left(\mathbf{A}^\star - \mathbf{X}\boldsymbol{\gamma}\right)\}\right] \tag{5.19}$$

For the OS model, we consider a robust specification of the mean part by including the spline function of the logit of the propensity score. In this process, both PS and OS models share PS parameters causing an update of PS values in posterior calculation which results in model feedback. For the sake of mathematical illustration, we focus on the spline function in the mean function of the OS model (5.5) and ignore the parametric part $g(\cdot)$. Hence, the OS model is specified as

$$Y^a|\mathbf{X};\boldsymbol{\gamma},\boldsymbol{\alpha}_a = \sum_{j=0}^{q}\alpha_{1ja}e_a(\mathbf{X},\boldsymbol{\gamma})^j + \sum_{k=1}^{K}\alpha_{2ka}(e_a(\mathbf{X},\boldsymbol{\gamma})-t_{ka})_+^q + \epsilon$$

which corresponds to the following formulation of the OS model incorporating all patients' information

$$\mathbf{Y}^a|\mathbf{X};\boldsymbol{\gamma},\boldsymbol{\alpha}_a = \mathbf{C}_1(\boldsymbol{\gamma})\alpha_{1a} + \mathbf{C}_2(\boldsymbol{\gamma})\alpha_{2a} + \boldsymbol{\epsilon}$$

. where,

$$\mathbf{C}_1(\boldsymbol{\gamma}) = \begin{pmatrix} 1 & e_a(\mathbf{X}_1,\boldsymbol{\gamma}) & e_a^2(\mathbf{X}_1,\boldsymbol{\gamma}) & \cdots & e_a^q(\mathbf{X}_1,\boldsymbol{\gamma}) \\ 1 & e_a(\mathbf{X}_2,\boldsymbol{\gamma}) & e_a^2(\mathbf{X}_2,\boldsymbol{\gamma}) & \cdots & e_a^q(\mathbf{X}_2,\boldsymbol{\gamma}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & e_a(\mathbf{X}_{n_a},\boldsymbol{\gamma}) & e_a^2(\mathbf{X}_n,\boldsymbol{\gamma}) & \cdots & e_a^q(\mathbf{X}_n,\boldsymbol{\gamma}) \end{pmatrix},$$

$$\mathbf{C}_2(\boldsymbol{\gamma}) = \begin{pmatrix} (e_a(\mathbf{X}_1,\boldsymbol{\gamma})-t_1)_+^q & \cdots & (e_a(\mathbf{X}_1,\boldsymbol{\gamma})-t_{Ka})_+^q \\ (e_a(\mathbf{X}_2,\boldsymbol{\gamma})-t_{1a})_+^q & \cdots & (e_a(\mathbf{X}_2,\boldsymbol{\gamma})-t_{Ka})_+^q \\ \vdots & \vdots & \vdots \\ (e_a(\mathbf{X}_n,\boldsymbol{\gamma})-t_{1a})_+^q & \cdots & (e_a(\mathbf{X}_n,\boldsymbol{\gamma})-t_{Ka})_+^q \end{pmatrix},$$

$\boldsymbol{\alpha}_{2a} \sim N_K\left(\mathbf{0},\sigma_{\alpha_2}^2\mathbf{I}_K\right)$ and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0},\sigma_y^2\mathbf{I}_n)$.

The above specification gives the potential outcome distribution under treatment a as

$$\mathbf{Y}^a|\mathbf{X};\boldsymbol{\gamma},\boldsymbol{\alpha}_a \sim N_n\left(\mathbf{C}_1(\boldsymbol{\gamma})\boldsymbol{\alpha}_{1a},\boldsymbol{\Sigma}_y(\boldsymbol{\gamma})\right)$$

where, $\boldsymbol{\Sigma}_y(\boldsymbol{\gamma}) = \left(\sigma^2_{\alpha_2}\mathbf{C}_2(\boldsymbol{\gamma})\mathbf{C}_2(\boldsymbol{\gamma})' + \sigma^2_y\mathbf{I}_n\right)$. The OS likelihood is given by

$$p\left(\boldsymbol{\gamma}, \boldsymbol{\alpha}_a|\mathbf{Y}^a, \mathbf{X}, \mathbf{A}\right) \propto \exp\left[-0.5\{\left(\mathbf{Y}^a - \mathbf{C}_1(\boldsymbol{\gamma})\right)' \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\gamma})\left(\mathbf{Y}^a - \mathbf{C}_1(\boldsymbol{\gamma})\right)\}\right] \tag{5.20}$$

Combining the PS and OS likelihoods (5.19) and (5.20) the full conditional distribution of PS parameters $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} p(\boldsymbol{\gamma}|\cdot) \quad \propto \quad & \exp\left[-0.5\left\{(\boldsymbol{\gamma}'\mathbf{X}'\mathbf{X}\boldsymbol{\gamma} - 2\boldsymbol{\gamma}'\mathbf{X}'\mathbf{A}^\star)\right.\right. && (5.21) \\ & \left.\left. + \left(\mathbf{Y}^a - \mathbf{C}_1(\boldsymbol{\gamma})\right)' \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\gamma})\left(\mathbf{Y}^a - \mathbf{C}_1(\boldsymbol{\gamma})\right)\right\}\right] && (5.22) \end{aligned}$$

Clearly, we see that although (5.21) only involves treatment and covariates, (5.22) involves quantities from outcome and PS model parameters which implies that the full posterior distribution of $\boldsymbol{\gamma}$ depends on the response surface model quantities in the posterior calculation. Hence, the propensity score is also updated by this process causing model feedback.

In a similar framework An and Little (2008) avoided this situation by approximating the posterior distribution of $\boldsymbol{\gamma}$ by $N_p\left(\hat{\boldsymbol{\gamma}}, I\left(\hat{\boldsymbol{\gamma}}\right)^{-1}\right)$ where $\hat{\boldsymbol{\gamma}}$ is the maximum likelihood estimator of $\boldsymbol{\gamma}$ in (5.19) and $I\left(\hat{\boldsymbol{\gamma}}\right)$ is the observed Fisher information matrix. Due to this approximation, the PS parameters $\boldsymbol{\gamma}$ are not impacted by the outcome model quantities as they are not drawn from their exact posterior distributions and hence can avoid model feedback. However, by treating the estimated PS parameters as a known quantity in the outcome stage, this approach faces same limitations as the frequentist counterpart in misstating the uncertainty in causal estimates.

5.4.4 Augmentation

As noted by Zigler et al. (2013), the identifiability issue due to the implied parameterization in Bayesian PS formulation as discussed in the example we considered before is not because of model feedback, but rather due to the dimension reduction from the entire covariate-treatment surface to one dimensional summary of covariates in the form of PS. Because of the constrained parameterization, the Bayesian approach fails to recover the entire covariate-outcome response surface. Instead of considering the OS model (5.12) with a function of PS values only, Zigler et al. (2013) suggested an augmentation of individual covariates in the outcome model (5.12) and the final model can be expressed as

$$Y|A, X, \boldsymbol{\gamma}, \boldsymbol{\xi}, \delta, \boldsymbol{\beta} = \xi_0 + \xi_1 e(X, \boldsymbol{\gamma}) + \delta A + g(X^+, \boldsymbol{\beta}) + \epsilon \tag{5.23}$$

where, $g(X^+, \boldsymbol{\beta})$ is a function of individual covariates except at least one covariate which is/are excluded from X in order to avoid multicollinearity and the corresponding parameters. Specifically, if we assume $g(X^+, \boldsymbol{\beta}) = X^+\boldsymbol{\beta}$ where $X^+ = (X_2, X_3, \ldots, X_p)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p-1})'$, then the RHS of (5.16) yields

$$(\xi_0 + \xi_1\gamma_0) + \xi_1\gamma_1 X_1 + (\xi_2\gamma_2 + \beta_1)X_2 + \cdots + (\xi_p\gamma_p + \beta_{p-1})X_p + \delta A \tag{5.24}$$

which by comparing with the data generating outcome model (5.18) implies $\alpha_1 = \xi_1\gamma_1$ and for $k \geq 2$, $\alpha_k = \xi_k\gamma_k + \beta_{k-1}$. By including the individual covariates in the response surface this model ensures identifiability of the data generation model parameters at the outcome stage.

The PSPP framework (Little and An, 2004; Zhang and Little, 2009) also builds the methodology on the same notion of augmentation and includes a parametric function of

individual covariates in the outcome model. They further showed that the PSPP method possesses key double robustness property in the sense that the exact specification of the augmentation function $g(\cdot)$ can be misstated without biasing estimates of marginal parameters of interest as long as the relationship between Y and $e(X, \boldsymbol{\gamma})$ is modeled correctly. In the proposed method we follow the same direction and include a similar $g(X_2, \ldots, X_p, \boldsymbol{\beta})$ function in the OS model in addition to $e(X, \boldsymbol{\gamma})$ which results in the specification of the final OS model (5.5).

CHAPTER 6

SIMULATION STUDY

6.1 Overview

We conduct extensive simulation studies to evaluate and compare the performance of BayesG and PSBayes methods in predicting the "best" treatment for a new patient across a wide range of scenarios. The simulation setups are inspired by the scenario 1 from Zhang et al. (2012b) although the metrics we consider are different from theirs. The simulation dataset consists of 500 subjects having a binary treatment A, continuous covariates $\mathbf{X}$ and a continuous outcome Y reflecting the causal diagram 3.1 in the sense that the treatment of each subject is generated following a PS model solely based on the covariates and the outcome is generated following a OS model based on both the treatments and the covariates. As the covariates are present in both PS and OS data generating models they can be specified as confounders as well. To effectively evaluate the optimal treatment prediction performance, we consider five simulation scenarios where we consider linear and nonlinear mean functions with single to multi modes in the outcome model mimicking wide range of real life situations. The treatments are generated by the same PS model across all scenarios. To inspect and validate the performance, we generate 100 synthetic datasets following the data generation models and randomly partition each dataset allocating 80% of the observations to the training data and 20% to the test data. We fit the models on the training data comprising of 400 observations and the optimal treatments are predicted based on the covariate information of the subjects in the test data consisting of 100 observations. The predicted optimal treatments

are matched with the true optimal treatments of the test data and prediction accuracy measures are calculated.

6.2 Method Specification

We compare our proposed methods based on Bayesian g-formula approach (BayesG) and PSPP Bayesian approach (PSBayes) with two versions of inverse probability weighting (IPWE) and augmented inverse probability weighting methods (AIPWE) (Zhang et al., 2012b), one corresponding to the correct specifications of PS and OS models, while in the other one, the models are misspecified.

6.2.1 Proposed Methods

In the BayesG method, we consider the true causal model to be $E\left[Y^a|\mathbf{V};\boldsymbol{\alpha}\right] = \alpha_0+\alpha_1 A+ \alpha_2 V+\alpha_{12} AV$ and we estimate $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_{12})'$ in order to predict the optimal treatment. The hyper-parameters of the prior distributions are specified following the prior distribution suggestions in Roy et al. (2016). Specifically, we chose the variance covariance matrix of the prior distributions of MSM parameters $\boldsymbol{\alpha}$ to be $\Sigma^{\pi}_{\boldsymbol{\alpha}} = 10^3\mathbf{I}_4$. The same specification is used for the variance covariance matrix of the GP mean regression function parameters $\boldsymbol{\beta}$ while its mean vector is assumed to be $\mathbf{0}$. For the Gibbs sampler, we specify the initial values as $\boldsymbol{\alpha} = \mathbf{1}$, $\boldsymbol{\beta} = \mathbf{1}$, $\sigma^2_y = 1$, $\eta = 0.5$, $\rho = 50$. For this approach, we directly draw from the full posterior distributions of the parameters. We use 25,000 draws for the burn-in period and an additional 25,000 draws in the iteration period provide sufficiently small MC error in all scenarios.

For the PSBayes approach, we consider a p-spline model of order 3 in the mean function of the outcome model as this specification is sufficient for desired robustness in most cases. The prior distribution hyper-parameters are determined following the prior specification of our method. Because of the significantly higher number of parameters and complex model specifications, we choose 100,000 number of draws for the burn-in and the same number of draws for posterior computation. The whole MCMC procedure is performed using *rjags* package in R which only requires the model specifications of the stochastic nodes along with the specifications of the fixed nodes and initial values of the parameters. We treat missing potential outcomes as parameters and we update them by repetitive draws from their posterior predictive distribution. In this way, we account for the uncertainty in missing data prediction.

6.2.2 Comparator Methods

We want to examine the performance of our methods compared to other methods proposed in the literature of optimal regime selection. Specifically, we considered two comparator methods - inverse probability treatment weighting (IPWE) and the doubly robust augmented inverse probability treatment weighting (AIPWE) (Zhang et al., 2012a), which are described in section 2.2. The IPWE method requires one to correctly specify the propensity score model while for AIPWE we need to additionally specify the outcome surface model. However, due to the double robustness property, AIPWE method needs only one correct specification. It is unlikely to correctly specify the regression models in real life situation. Keeping this in mind, we consider two variants of IPWE and AIPWE specifications as our comparator methods viz. IPWE-True, IPWE-Mis, AIPWE-True and AIPWE-Mis. Among them, the true versions are based on the true data generating models. More specifically, in

IPWE-True method the analysis PS model has same specification as the data generating PS model, whereas the IPWE-Mis method has misspecified PS model. Similarly, for AIPWE-True method, both PS and OS models are correctly specified, but for AIPWE-Mis method, both of the models are misspecified. The specifications of all the comparator methods are inspired from Zhang et al. (2012b) simulation 1.

Irrespective of the simulation scenarios, the outcome regression models specify a class of treatment regimes based on the observed covariates $\mathcal{G}_\eta = I(\mathbf{X}'\boldsymbol{\eta} > 0)$, so that the optimal treatment $A^\star \in \mathcal{G}_{\boldsymbol{\eta}}$. Both IPWE and AIPWE provide estimation of the average potential outcome indexed by $\boldsymbol{\eta}$ and maximize these estimators in order to estimate $\boldsymbol{\eta}$ and thus define the optimal treatment selection rule. However, both IPWE and AIPWE estimators are non-smooth functions of $\boldsymbol{\eta}$ and hence the traditional optimization routines may fail. As an alternative, following Zhang et al. (2012b) we used a genetic algorithm which is implemented in the *rgenoud* package in R. We used the default argument values except we took *max*=T, *optim.method*=Neider-Mead and *pop.size*=3000 as suggested in the documentation of the package and also in the Zhang et al. (2012b). We took the *starting.values*=c(0,0,...,0) and set $(-2, -2, \ldots, -2)'$ and $(2, 2, \ldots, 2)'$ as lower and upper bounds of $\boldsymbol{\eta}$.

6.3 Performance Comparison

6.3.1 Validation

The primary objective of this simulation study is to evaluate how well the methods can predict the treatment assignment for a new patient. Evaluation of prediction performance requires the consideration of two separate datasets - one for model building and estimation (training) and another for prediction and validation (test). One way to carry out such

validation technique is to follow hold-out strategy where the whole dataset is randomly divided into two parts for training and testing. The model is fit using the training set and the optimal treatments are predicted for the test data using the test data confounders. The predicted treatment assignments are matched with the observed treatment assignments in the test data which gives a summary of the prediction accuracy. Finally, this whole process is repeated multiple times to give an aggregated prediction accuracy measurement. Due to heavy computational cost, we parallelized the whole process into 100 array jobs which was run in 10 parallel nodes of Gaea, a high performance 60 node CPU/GPU hybrid cluster at Northern Illinois Universitys Center for Research Computing and Data. Each process was assigned computation for 1 holdout data. For BayesG approach, each holdout computation took 5 hours on average for 50,000 iterations while the PSBayes approach took 1 hour on average with 200,000 iterations.

6.3.2 Prediction Measure

For each holdout sample, we can obtain a 2×2 confusion matrix by matching the predicted optimal treatments with the true optimal treatments from the simulation settings which would provide us with the sensitivity, specificity and prediction accuracy. Let Z be a dichotomous variable which can take values 0 or 1. Let $Z^\star$ be the true value of Z and $\hat{Z}$ be the predicted value of Z. Then in the classification setting, the sensitivity and specificity are given by $P(\hat{Z} = 1|Z^\star = 1)$ and $P(\hat{Z} = 0|Z^\star = 0)$ respectively and the prediction accuracy gives the overall strength of matching and is defined as $P(\hat{Z} = Z^\star)$.

In our case, let $A^{\star}_{ik}$ be the true optimal treatment choice for i^{th} patient in k^{th} holdout test data and let $\hat{A}_{ik}$ be the corresponding predicted optimal treatment. Then the sensitivity and specificity for k^{th} holdout test data can be estimated as

$$\text{sens}_k = \frac{\sum_{i=1}^{n_{\text{test}}} I\left(\hat{A}_{ik} = A^{\star}_{ik} = 1\right)}{\sum_{i=1}^{n_{\text{test}}} I\left(A^{\star}_{ik} = 1\right)},\ \text{spec}_k = \frac{\sum_{i=1}^{n_{\text{test}}} I\left(\hat{A}_{ik} = A^{\star}_{ik} = 0\right)}{\sum_{i=1}^{n_{\text{test}}} I\left(A^{\star}_{ik} = 0\right)}$$

where, n_{test} is the number of patients in the test data. The prediction accuracy is estimated by matching the predicted treatment with the observed

$$\text{acc}_k = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} I\left(A^{\star}_{ik} = \hat{A}_{ik}\right)$$

6.4 Simulation 1

In this simulation, we follow the simulation 1 setting of Zhang et al. (2012b) to generate 500 observations $(Y_i, A_i, \mathbf{X}_i)$, $i = 1(1)500$ where, $\mathbf{X}_i = (X_{i1}, X_{i2})'$. X_{i1} and X_{i2} are independently generated from $U(-1.5, 1.5)$. A_i is generated from a Bernoulli distribution with success probability $\pi(\mathbf{X}_i)$ where the following logit model is used to generate $\pi(\mathbf{X}_i)$ given the values of $\mathbf{X}_i$

$$\text{logit}\{\pi(\mathbf{X})\} = -1 + 0.8X_1^2 + 0.8X_2^2$$

and Y_i is generated using the following outcome model given A_i and $\mathbf{X}_i$

$$(Y|A, \mathbf{X}) = \mu(A, \mathbf{X}) + \epsilon$$

Table 6.1: Simulation 1 prediction measures.

Model	Accuracy			Sensitivity			Specificity		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
BayesG	0.768	0.78	0.099	0.999	1	0.005	0.511	0.536	0.209
PSBayes	0.85	0.925	0.136	0.954	0.979	0.053	0.732	0.903	0.288
IPWE-True	0.876	0.91	0.11	0.887	0.918	0.125	0.863	0.923	0.188
IPWE-Mis	0.674	0.61	0.138	0.906	0.912	0.054	0.417	0.302	0.283
AIPWE-True	0.914	0.92	0.048	0.917	0.931	0.085	0.911	0.953	0.106
AIPWE-Mis	0.525	0.54	0.044	1	1	0	0	0	0

where, the mean function $\mu(A, \mathbf{X}) = 2 - 1.5X_1^2 - 1.5X_2^2 + 3X_1X_2 + A(-0.1 - X1 + X2)$ and the error $\epsilon \sim N(0, 1)$. From the outcome model, the true optimal treatment of i^{th} patient can be deduced as

$$A_i^\star = I(X_{i1} - X_{i2} < -0.1)$$

To estimate the optimal regime, we assume the same specification for the proposed methods as discussed before. For the comparator methods, we assume a correct and an incorrect specification of PS and OS models. In particular, for the correctly specified PS model we posit $\text{logit}\{\pi_t(\mathbf{X}; \boldsymbol{\gamma})\} = \gamma_0 + \gamma_1 X_1^2 + \gamma_2 X_2^2$ which belong to the same class of the data generating PS model and for the misspecified version of the PS model, we assume $\text{logit}\{\pi_m(\mathbf{X}; \boldsymbol{\gamma})\} = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2$. As for the outcome regression model, we consider $\mu_t(A, \mathbf{X}, \boldsymbol{\beta}) = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_1 X_2 + A(\beta_4 + \beta_5 X_1 + \beta_6 X_2)$ corresponding to the correct specification and $\mu_m(A, \mathbf{X}, \boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + A(\beta_3 + \beta_4 X_1 + \beta_5 X_2)$ for the misspecified model. Both PS and OS model parameters are estimated by standard GLM procedure.

Table 6.1 shows the comparison of accuracy, sensitivity and specificity measures among all methods we discussed here. The mean, median and standard deviation of each prediction measure over 100 holdout samples are displayed. The empirical distribution of the predic-

tion measurements may not be symmetric and hence we consider median as the centrality measure.

As we see from the result, the BayesG approach predict the true optimal treatments 76.8% of the times on average. Interestingly, the median sensitivity is 1 i.e. this approach correctly predicts all true treatment choices that are 1. However, the median specificity is only 53.6% resulting moderate overall accuracy performance. On the other hand, we see that PSBayes approach performs substantially well in this setting as it correctly predicts true treatment 1 almost 98% and true treatment 0 90% of the times. Overall this approach correctly predicts the treatments 92.5% on average. Among the inverse probability methods the true variants of AIPWE and IPWE perform considerably as both models are correctly specified, however, both of them have marginally lower accuracy measure than PSBayes on average. But in real life situations, we rarely anticipate the true data generating models and correctly specify them as the analysis models. Therefore, the misspecified models would give us a glimpse of real life scenarios. Here, also we see that while the true specifications of IP based methods perform comparably, their misspecified counterparts perform poorly. For instance, while IPWE with correct PS model could predict the true optimal treatment around 90% of the times, the misspecified version could only show 61% prediction accuracy. Under misspecified models, IPWE methods show better performance than AIPWE as in IPWE only the PS model is misspecified, but in AIPWE both PS and OS models are misspecified. Taken together, these key findings lead into the conclusion that incorrectly specified models heavily affect the prediction performance and the extent of under performance depends on the extent of misspecification.

Table 6.2: Simulation 2 prediction measures.

Model	Accuracy			Sensitivity			Specificity		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
BayesG	0.756	0.77	0.125	0.943	1	0.099	0.549	0.636	0.293
PSBayes	0.733	0.76	0.137	0.825	0.854	0.138	0.624	0.8	0.348
IPWE-True	0.923	0.93	0.046	0.918	0.926	0.075	0.927	0.955	0.088
IPWE-Mis	0.82	0.84	0.122	0.943	0.96	0.049	0.681	0.713	0.246
AIPWE-True	0.959	0.96	0.027	0.959	0.978	0.046	0.957	0.978	0.055
AIPWE-Mis	0.889	0.94	0.139	0.949	0.963	0.053	0.822	0.924	0.276

6.5 Simulation 2

Data Generation. This setting is similar to the simulation 1 except the mean function of the outcome data generation model is not an additive model. Similar to Simulation 1, we independently generate the covariates X_{i1} and X_{i2} from $U(-1.5, 1.5)$ and A_i from the logit PS model

$$\text{logit}\{P(A = 1)|\mathbf{X}\} = -1 + 0.8X_1^2 + 0.8X_2^2$$

. Y_i is generated using the following outcome model given A_i and $\mathbf{X}_i$

$$(Y|A, \mathbf{X}) = \mu(A, \mathbf{X}) + \epsilon$$

where, the mean function $\mu(A, \mathbf{X}) = \exp\left(2 - 1.5X_1^2 - 1.5X_2^2 + 3X_1X_2 + A(-0.1 - X_1 + X_2)\right)$ and the error $\epsilon \sim N(0, 1)$. Here also, the true optimal treatment is defined as

$$A^\star = I(X_1 - X_2 < -0.1)$$

Result. The simulation 2 results are displayed in table 6.2. Note that, in this setting, we purposely did not include an additive mean function in the outcome data generating model

to check the performance of the comparator methods in such scenarios. In summary, both correct and incorrect specifications of IPWE and AIPWE display better performance than our proposed approaches probably due to the fact that the model space of the proposed outcome models does not include non-linear models.

In particular, both BayesG and PSBayes methods correctly predict the true optimal treatments around 77% of the times, whereas the true variants of IPWE and AIPWE are 93% and 96% accurate. AIPWE-Mis achieves 94% overall accuracy, which approximates the true version, while IPWE with an incorrect PS model is 84% accurate. Among our methods, PSBayes perform marginally better than BayesG and predicts the true optimal treatment 77% of the times. The other measures such as sensitivity and specificity follows the same pattern.

6.6 Simulation 3

Data Generation. In this setting, we consider three confounders $\mathbf{X} = (X_1, X_2, X_3)'$ and generate $\mathbf{X}_i, i = 1(1)500$ as follows.

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & 0.25 & 0.175 \\ 0 & 0.175 & 0.25 \end{pmatrix} \right]$$

All three covariates contribute to generate treatments using the following model

$$\text{logit}(P(\mathbf{X})) = -1 + 0.8X_1^2 + 0.8X_2^2 + 0.8X_3^2$$

Table 6.3: Simulation 3 prediction measures

Model	Accuracy			Sensitivity			Specificity		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
BayesG	0.917	0.93	0.056	0.943	0.965	0.076	0.883	0.92	0.125
PSBayes	0.887	0.9	0.068	0.891	0.931	0.1	0.881	0.919	0.122
IPWE (True)	0.811	0.815	0.102	0.887	0.925	0.122	0.711	0.796	0.257
IPWE (Mis)	0.797	0.815	0.091	0.95	0.981	0.072	0.594	0.629	0.216
AIPWE(True)	0.872	0.88	0.078	0.907	0.958	0.11	0.825	0.873	0.182
AIPWE (Mis)	0.84	0.86	0.089	0.888	0.925	0.116	0.775	0.814	0.196

Based on the treatment A_i and covariates $\mathbf{X}_i$, the outcomes Y_i are simulated as

$$(Y|A, \mathbf{X}) \sim N(\mu(A, \mathbf{X}), 1)$$

where, $\mu(A, \mathbf{X}) = \exp(2 - 1.5X_3^2 + 3X_1X_3 + A(-0.1 - X_1 + X_2))$. Note that, X_3 is included in the mean function but not included in the interaction term with treatment and can be considered as non-effect modifier W as in the causal inference literature. As the interaction term does not change the optimal treatment is defined by the same decision rule as in simulation 1 and simulation 2.

Result. From the simulation result in Table 6.3, we see that the BayesG method outperforms all other methods in terms of higher sensitivity, specificity and overall accuracy in the prediction of true optimal treatments. The reason behind this finding directs to the presence of the non-effect modifier X_3 in the data generating model. The PSBayes has shown marginally inferior prediction profiles. Among the IP methods, AIPWE show a better prediction measures than IPWE.

Specifically, we see that BayesG is around 93% accurate in predicting the true best treatment. PSBayes also performs comparably and shows 90% accuracy. Among the comparator methods, AIPWE performs closely since the true specification is 88% accurate and the misspecification is 86% accurate. The IPWE only relies on PS models and does not have

enough power to detect true signals and performs suboptimally by predicting the true treatment around 80% of the times. If we consider other measures, all methods except PSBayes have higher sensitivity than specificity which means most of the methods correctly predict treatment 1 most of the times, while it is not true for treatment 0. PSBayes has similar sensitivity and specificity profile which might be due to the fact that it fits the outcome model for both treatment groups separately.

6.7 Simulation 4

Data Generation. In this simulation setting we consider a bimodal mean function. Here we simulate the covariates $\mathbf{X}$ and the treatment A in the similar fashion as in Simulation 1 or 2. Additionally, we simulate a binary variable B_i, $i = 1(1)500$ from $B \sim \text{Bernoulli}(0.5)$ as a cluster indicator. Next based on the treatments, the covariates and cluster membership variable we simulate the outcomes as

$$(Y|A, \mathbf{X}) \sim N(\mu(A, \mathbf{X}), 1)$$

where, $\mu(A, \mathbf{X}) = 2 - 1.5X_1^2 - 1.5X_2^2 + 3X_1X_2 + A(-0.1 - X_1 + X_2) + 10(B - \bar{B})$.

Here, also the optimal treatment would be decided by the same decision rule as in the other simulation settings as the interaction term is same across all settings.

Result. In this setting, the outcome distribution is not unimodal and somewhat reflects a non-ideal scenario. Here as we see from the table 6.4, the DDP specification of the PSBayes method outperforms all other methods in having higher prediction accuracy, sensitivity and specificity by a substantial margin. The PSBayes method has overall 94% accuracy while BayesG and the IP methods with true specifications are less than 80% accurate in prediction

Table 6.4: Simulation 4 prediction measures.

Model	Accuracy			Sensitivity			Specificity		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
BayesG	0.742	0.77	0.133	0.952	1	0.101	0.494	0.526	0.286
PSBayes-DDP	0.838	0.94	0.183	0.9	0.981	0.184	0.763	0.932	0.345
IPWE-True	0.708	0.73	0.177	0.696	0.734	0.26	0.723	0.786	0.279
IPWE-Mis	0.673	0.65	0.154	0.836	0.904	0.173	0.482	0.415	0.31
AIPWE-True	0.768	0.78	0.144	0.768	0.863	0.237	0.763	0.833	0.241
AIPWE-Mis	0.665	0.63	0.153	0.854	0.9	0.14	0.442	0.372	0.295

of the true best treatment option for the new patients. The misspecified versions of the IP method perform poorly as they only show around 65% accuracy. The BayesG method has a very high sensitivity and predicts the true treatment 1 almost all the times, but has a lower accuracy due to having only 52.6% specificity. Here also, we see a clear disparity between the sensitivity and specificity measures for all methods except PSBayes.

6.8 Simulation 5

Data Generation. Here we consider a skewed outcome distribution. To do that, we simulate the covariates $\mathbf{X}$ and the treatment A in the similar fashion as in Simulation 1, 2 or 4. For the outcome generation, we follow $(Y|A, \mathbf{X}) = \mu(A, \mathbf{X}) + \epsilon$ where, we assume $\mu(A, \mathbf{X}) = 2 - 1.5X_1^2 - 1.5X_2^2 + 3X_1X_2 + A(-0.1 - X_1 + X_2)$ and ϵ is generated from a skewed mixture distribution. Specifically, we consider a mixture setup to generate ϵ as $\epsilon|\epsilon^\star \sim N(\epsilon^\star, 3)$ with $\epsilon^\star \sim \text{LogNormal}(2, 1)$.

Here also the optimal treatment would be decided by the same decision rule as in the other simulation settings as the interaction term is same across all settings.

Result. This situation reflects a skewed outcome case which occurs frequently in real life scenarios. Here also we see, the DDP specification of the PSBayes method outperforms all

Table 6.5: Simulation 5 prediction measures

Model	Accuracy			Sensitivity			Specificity		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
BayesG	0.622	0.595	0.191	0.744	0.838	0.283	0.483	0.517	0.352
PSBayes-DDP	0.749	0.8	0.164	0.79	0.945	0.254	0.708	0.814	0.314
IPWE (True)	0.572	0.6	0.225	0.57	0.604	0.302	0.576	0.629	0.332
IPWE (Mis)	0.609	0.64	0.164	0.832	0.869	0.159	0.353	0.367	0.232
AIPWE(True)	0.59	0.625	0.224	0.597	0.673	0.294	0.59	0.62	0.342
AIPWE (Mis)	0.595	0.585	0.205	0.682	0.733	0.262	0.497	0.392	0.335

other methods in having higher prediction accuracy, sensitivity and specificity by a significant margin. Only the PSBayes method has more than 80% overall accuracy. Close to 95% of the times, it predicted correct true treatment 1 cases, whereas for true treatment 0 cases the accuracy is around 82%. In contrast, all other methods show overall accuracy around 60%. The sensitivity of BayesG, misspecified IP models are significantly higher than the specificity, while the true versions have similar sensitivity and specificity profiles. Interestingly, the incorrect specification of the IP methods sometimes outperform the true version of these methods highlighting the fact that the performance difference is more clear when the outcome distribution is symmetric. But in presence of skewed distribution, the difference is more nuanced.

CHAPTER 7

DISCUSSION

In this dissertation, we propose two Bayesian semi-parametric methods; BayesG and PSBayes to estimate the optimal treatment decision for a new patient based on his/her biological profile in a binary treatment setting. Our methods are based on an observational study where we have access to the information of a group of patients about their biological profile, assigned treatments and their response to the treatments. The BayesG method builds on the notion of Bayesian parametric g-formula and extends it to specify GP prior within a marginal structural model framework which allows for nonparametric estimation of the mean function. Advantage of using this method includes higher efficiency in estimation of the causal parameters due to adherence to the likelihood based parametric g-formula and incorporating prior information through Bayesian setting to reduce variability in the estimation. However, one limitation of the BayesG approach is that the treatment decision is driven by the marginal structural model which only involves the effect modifier among the confounders. Hence, one should be very cautious in identifying the true effect modifiers and in specifying a correct MSM such that it belongs to the class of true MSMs. In the simulation studies, the MSMs of the analysis model are specified based on the outcome model. Observing the close analogy between treatment assignment and missing data mechanisms our second approach is developed on the PSPP framework from the missing data literature. We train our model with the information of the observed group of patients and impute missing potential outcomes of the new patient under each treatment group. We then compare the empirical distributions of the treatment groups and estimate the optimal treatment employing a stochastic ordering approach.

Finally, we consider five different simulation scenarios reflecting wide range of situations where we compare the prediction performance of our methods with comparator semi-parametric estimation approaches such as inverse probability weighting (IPWE) and the double robust augmented inverse probability weighting (AIPWE). In the presence of additive mean functions, we found that our methods show better performance than the IP based methods and vice-versa. Specifically, in the presence of non effect modifier W, the BayesG method performs better than all other method, whereas in other additive mean function scenarios PSBayes shows higher median accuracy in predicting true optimal treatments. Furthermore, the BNP specification of PSBayes method outperform other methods in the last two scenarios, where the outcomes are generated from a multimodal or a positively skewed distribution.

Throughout this dissertation, we discussed the methods and literature related to time-fixed setting. However, as we discussed in chapter 2, in lot of practical applications involving chronic diseases the data often originate from a longitudinal setting raising a natural question - how to determine the "best" treatment decision at a certain stage of the multistage therapy aka dynamic treatment regime (DTR) based on the accrued information on a patient. We seek to answer this question by extending our proposed method in presence of time varying confounders and treatments as one of our future works.

For concreteness, let $\{Y(t), A(t), \mathbf{X}(t)\}$ be the outcome, treatment and confounder information of a patient collected at a discrete time t. In addition, let the history of exposure upto and including time t be denoted as $\bar{A}(t) = \{A(u) : 0 \leq u \leq t\}$. Similarly, the time varying covariates and outcomes are denoted as $\bar{\mathbf{X}}(t) = \{\mathbf{X}(u) : 0 \leq u \leq t\}$ and $\bar{Y}(t) = \{Y(u) : 0 \leq u \leq t\}$. Under the temporal assumption that at every time point $\mathbf{X}(t)$ occurs before $A(t)$ we can specify two models to facilitate Bayesian parametric g-formula - (1) for the confounder at time t based on accrued information available upto time $t-1$ i.e. $p(\mathbf{X}(t)|\bar{a}(t), \bar{\mathbf{x}}(t-1), \bar{y}(t-1), \bar{a}(t-1))$ and (2) for the outcome $Y(t)$ at time t, based

on the accrued information available on treatment and confounder history upto time t and the outcome history upto time $t-1$ i.e. $p(Y(t)|\bar{a}(t), \bar{\mathbf{x}}(t), \bar{y}(t-1))$. Then similar to the BayesG method, we can integrate the nuisance parameter and covariate to match with the MSM specification. Several works have been done in this area in the context of MSM and DTR (Robins, 1998a, 1999a, 2000; Robins et al., 2000; Hernán et al., 2000; Murphy et al., 2001; Keil et al., 2015). Keil et al. (2015) considered a similar method for estimating the causal parameters in the setting of Bayesian parametric g-formula. However, their method relies on parametric assumption which we can relax by considering flexible non-parametric likelihood model.

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

An, H. and Little, R. J. (2008). Robust model-based inference for incomplete data via penalized spline propensity prediction. *Communications in StatisticsSimulation and Computation®*, 37(9):1718–1731.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Berger, J. O., Wang, X., and Shen, L. (2014). A bayesian approach to subgroup identification. *Journal of biopharmaceutical statistics*, 24(1):110–129.

Bernardo, J., Berger, J., Dawid, A., Smith, A., et al. (1998). Regression and classification using gaussian process priors. *Bayesian statistics*, 6:475.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97(457):160–169.

Chakraborty, B. and Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Chen, G., Zhong, H., Belousov, A., and Devanarayan, V. (2015). A prim approach to predictive-signature development for patient stratification. *Statistics in medicine*, 34(2):317–342.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

Escobar, M. D. (1990). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means.

Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pages 287–302. Elsevier.

French, S. (1986). *Decision theory: an introduction to the mathematics of rationality*. Halsted Press.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gunter, L., Zhu, J., and Murphy, S. (2011). Variable selection for qualitative interactions. *Statistical methodology*, 8(1):42–55.

Heckman, J. J., Lopes, H. F., and Piatek, R. (2014). Treatment effects: A bayesian perspective. *Econometric reviews*, 33(1-4):36–67.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Hernan, M. A. and Robins, J. M. (2010). *Causal inference*. CRC Boca Raton, FL:.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Hoshino, T. (2008). A bayesian propensity score adjustment for latent variable modeling and mcmc algorithm. *Computational Statistics & Data Analysis*, 52(3):1413–1429.

Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L., and Devanarayan, V. (2017). Patient subgroup identification for clinical drug development. *Statistics in Medicine*, 36(9):1414–1428.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jiang, R., Lu, W., Song, R., and Davidian, M. (2017). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1165–1185.

Kaplan, D. and Chen, J. (2012). A two-step bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, 77(3):581–609.

Keil, A. P., Daza, E. J., Engel, S. M., Buckley, J. P., and Edwards, J. K. (2015). A bayesian approach to the g-formula. *Statistical Methods in Medical Research*, page 0962280217694665.

Keil, A. P., Edwards, J. K., Richardson, D. R., Naimi, A. I., and Cole, S. R. (2014). The parametric g-formula for time-to-event data: towards intuition with a worked example. *Epidemiology (Cambridge, Mass.)*, 25(6):889.

Lavori, P. W. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clinical trials*, 1(1):9–20.

Lavori, P. W. and Dawson, R. (2008). Adaptive treatment strategies in chronic disease. *Annu. Rev. Med.*, 59:443–453.

Lindley, D. V. (1991). Making decisions.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21):2601–2621.

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, pages 949–968.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique*, pages 139–157.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357.

Loh, W.-Y., He, X., and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine*, 34(11):1818–1833.

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical methods in medical research*, 22(5):493–504.

MacEachern, S. N. (2000). Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, pages 1–40.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in medicine*, 28(1):94–112.

Moodie, E. E., Chakraborty, B., and Kramer, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics*, 40(4):629–645.

Moodie, E. E., Dean, N., and Sun, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6(2):223–243.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2016). *Bayesian nonparametric data analysis*. Springer.

Müller, P. and Rodriguez, A. (2013). *Nonparametric bayesian inference*. Institute of Mathematical Statistics; American Statistical Association.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481.

Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423.

Parmigiani, G. (2002). *Modeling in medical decision making: a Bayesian approach*. Wiley.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180.

Robins, J., Orellana, L., and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in medicine*, 27(23):4678–4721.

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

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412.

Robins, J. M. (1997). Causal inference from complex longitudinal data. *Latent variable modeling and applications to causality*, pages 69–117.

Robins, J. M. (1998a). 1997 proceedings of the american statistical association, section on bayesian statistical science.

Robins, J. M. (1998b). Correction for non-compliance in equivalence trials. *Statistics in medicine*, 17(3):269–302.

Robins, J. M. (1998c). Structural nested failure time models. *Encyclopedia of Biostatistics*.

Robins, J. M. (1999a). Association, causation, and marginal structural models. *Synthese*, 121(1):151–179.

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

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M., Hernán, M. A., and SiEBERT, U. (2004). Effects of multiple interventions. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*, 1:2191–2230.

Robins, J. M., Hernán, M. A., and Wasserman, L. (2015). On bayesian estimation of marginal structural models. *Biometrics*, 71(2):296.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Roy, J., Lum, K. J., and Daniels, M. J. (2016). A bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics*, 18(1):32–47.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

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

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, pages 808–840.

Ruppert, D., Wand, M., and Carroll, R. (2003). Semiparametric regression.,(cambridge university press: Cambridge, uk.).

Saarela, O., Stephens, D. A., Moodie, E. E., and Klein, M. B. (2015). On bayesian estimation of marginal structural models. *Biometrics*, 71(2):279–288.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Song, R., Kosorok, M., Zeng, D., Zhao, Y., Laber, E., and Yuan, M. (2015). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat*, 4(1):59–68.

Temple, R. and Ellenberg, S. S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: ethical and scientific issues. *Annals of internal medicine*, 133(6):455–463.

Thall, P. F., Millikan, R. E., Sung, H.-G., et al. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in medicine*, 19(8):1011–1028.

Tian, L. and Tibshirani, R. (2010). Adaptive index models for marker-based risk stratification. *Biostatistics*, 12(1):68–86.

Xu, Y., Müller, P., Wahed, A. S., and Thall, P. F. (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association*, 111(515):921–950.

Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., and Hernán, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in biosciences*, 3(1):119.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, G. and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3):911–918.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1):263–273.