Novel Flexible Statistical Methods for Missing Data Problems and Personalized Health Care

by

Yilun Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Lu Wang, Chair
Assistant Professor Peisong Han
Research Associate Professor Matthew J. Schipper
Associate Professor Emily Somers

Yilun Sun

yilunsun@umich.edu

ORCID iD: 0000-0002-1341-382X

ACKNOWLEDGEMENTS

I want to extend thanks to my instructors, peer students, friends, and family, who so generously contributed to the work presented in this thesis.

First and foremost, I am indebted to my advisor, Lu Wang. Lu is always generous with her time, she is always a good listener, and she always provides insightful advice. Her valuable suggestions, comments, and guidance encourage me to stay persistent towards my goals. Lu is an inspiring advisor in many ways, and her guidance has been significantly helping me gain knowledge, develop research interests, grow as a research scholar, as well as balance work and family. This dissertation would not have been possible without the consistent guidance, encouragement, and support from Lu.

I want to thank Matthew Schipper, for supervising my GSRA research, supporting my Ph.D. study, and serving as my dissertation committee member. Matt has been very understanding and supportive since my first day working with him. He has been giving me both independence and guidance, teaching me to conduct collaborative research effectively and helping me gain valuable experience in developing statistical tools for exciting and meaningful radiation oncology applications.

I am also grateful to the rest of my dissertation committee members, Peisong Han, and Emily Somers. I collaborated with Peisong in my first thesis project and benefited greatly from his tremendous passion for research, broad knowledge, and insightful comments. I am thankful to Emily for agreeing to serve on my dissertation committee

on such short notice, and her invaluable insights and suggestions helped me improve my thesis.

I want to express my gratitude to Jeremy Taylor. Jeremy supervised my research in the summer even before I officially started my Ph.D. study, and has always been very supportive. His excellent guidance and mentorship helped me get on the right track and realize my research interest. Moreover, his rigorous attitude toward scholarship has dramatically influenced my work ever since.

Thanks are also due to the faculty, staff, and peer students in the Department of Biostatistics, University of Michigan. Also, I am grateful to my collaborators from Michigan Medicine.

Special thanks to my beloved family members. I want to thank my mom, Xi Wang, without her selfless support, I would not be able to make this accomplishment. I also want to thank my wife, Yang Jiao, and my two lovely kids, Joanna and Lucas. I would like to thank them for all the support, joy, and love. I am fortunate to have them in my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Coarsened data, including a broad class of incomplete data structures, are ubiquitous in biomedical research. Various methods have been proposed when data are coarsened at random, where it is often required to specify working models for the coarsening mechanism, or the conditional outcome regression, or both. However, some practical issues emerge: for example, it is not unusual to have multiple working models that all fit the observed data reasonably well, or working model specification can be challenging in the presence of a large number of variables. In this dissertation, we propose new flexible statistical methods to tackle the challenges mentioned above with two types of coarsened data: missing data and counterfactual data when optimizing the dynamic treatment regime (DTR), which is a sequence of decision rules that adapt treatment to the time-varying medical history of each individual.

In the first project, we develop a multiply robust kernel estimating equations (MR-KEEs) method for nonparametric regression that can accommodate multiple working models for either the missing data mechanism or the outcome regression or both. The resulting estimator is consistent if any one of those models is correctly specified. When including correctly specified models for both, the proposed estimator achieves the optimal efficiency within the class of augmented inverse probability weighted (AIPW) kernel estimators.

In the second project, we develop a stochastic tree-based reinforcement learning method, ST-RL, for estimating the optimal DTR in a multi-stage multi-treatment setting with data from either randomized trials or observational studies. At each stage, ST-RL constructs a decision tree by first modeling the mean counterfactual outcomes via Bayesian nonparametric regression models, and then stochastically searching for

the optimal tree-structured regime using a Markov Chain Monte Carlo algorithm. We implement the proposed method in a backward inductive fashion through multiple decision stages. Compared to existing methods, ST-RL delivers the optimal DTR with better interpretability and does not require explicit working model specifications. Besides, ST-RL demonstrates stable and outstanding performance with moderately high dimensional data.

In the third project, we propose a variable selection technique to screen potential tailoring variables for estimating the optimal DTR with large observational data. The proposed method is based on nonparametric sparse additive models and therefore enjoys superior flexibility. Also, it allows treatments with more than two levels, as well as continuous doses. In particular, the selection procedure enforces strong heredity constraint, i.e., the interactions can only be included when both of its main effects have already been selected.

CHAPTER I

Introduction

Coarsened data (*Heitjan*, 1993; *Gill et al.*, 1997) are ubiquitous in biomedical research. The term coarsened data represents a broad class of incomplete data structures, where the observed data can be viewed as many-to-one functions of complete data. For example, investigators aimed to study the relationship between adjuvant therapy and sexual dysfunction in gynecologic oncology patients. Data were collected from the clinical records (Y) in addition to a cross-sectional survey of outpatients in the gynecologic oncology clinic, where the survey contained both patients information X_1 and behavior questions X_2. However, some patients chose not to complete the behavior part possibly because of privacy concern. As a result, only the coarsened data (X_1, Y) were observed for these patients, while the complete data (X_1, X_2, Y) were obtained for other patients. Other coarsened data examples include censored data in survival analysis, missing data problems, and potential outcomes in causal inference.

There are three coarsened-data mechanisms: coarsening completely at random (CCAR), where coarsening mechanism does not depend on data; coarsening at random (CAR), that is, the probability of coarsening depends on observed data only; and non-coarsening at random (NCAR), where coarsening depends on unobserved data. When data are coarsened at random, existing methods often require specification of work-

ing models for the coarsening mechanism, or the conditional outcome regression, or both. However, some issues emerge in practice: for example, it is not unusual to have multiple working models that all fit the observed data reasonably well, or working model specification can struggle in the presence of a large number of variables.

In this dissertation, we consider two kinds of coarsened data: missing data and counterfactual data in estimating optimal dynamic treatment regimes (DTRs), which is a sequence of decision rules that adapt treatment to the time-varying medical history of each individual. The overarching goal is to propose new flexible statistical methods to tackle the aforementioned challenges. Missing data are commonly seen in almost all scientific research areas, and various methods have been proposed in the statistical literature. Popular approaches include multiple imputation or likelihood-based methods (*Little and Rubin*, 2002; *Little*, 1992). Another stream of research is weighting-based methods, including inverse probability weighting (IPW, *Horvitz and Thompson*, 1952) and augmented inverse probability weighting (AIPW, *Robins et al.*, 1995). Existing literature often involves working model specification for missing mechanisms or conditional outcome regression, which often requires a considerable amount of guesswork; thus it remains a substantial challenge to obtain consistent estimates when working models are subject to misspecification.

On the other hand, dynamic treatment regimes (DTRs) are sequences of treatment decision rules, one per stage of intervention. Each decision rule maps up-to-date patient information to a recommended treatment (*Robins*, 2004; *Murphy*, 2003). DTRs build the basis of common medical practice, and it is essential to identify and evaluate optimal DTRs for personalized therapy and tailored management of chronic health conditions such as cancer, cardiovascular disease, behavioral disorders, and infections. Recently DTR has become a rapidly growing field, and various statistical methods have been developed to identify optimal DTRs. However, there are

still many unsolved questions in this area. For example, most existing methodologies require specifying parametric or semiparametric models when modeling the counterfactual outcome, which directly impacts the quality of estimated DTR. Usually, the consistency of the estimation can only be guaranteed using data from randomized trials; however, in practice, observational data are more often encountered because they are much cheaper and easier to obtain. As a consequence, model misspecification when using large observational data is a big concern, especially given the inherent difficulty of modeling high-dimensional information in a time-varying setting. Moreover, it is crucial to deliver interpretable treatment regimes for human experts to understand and gain insights, and however, previous literature has primarily emphasized on estimation accuracy instead of interpretability.

In Chapter II, we propose a multiply robust estimation method for nonparametric regression models in the presence of missing data, which provide not only great flexibility to allow data-driven dependence of the response on covariates, but also successfully correct for selection bias due to missing data. When data are missing at random (*Little and Rubin*, 2002), i.e., the missingness can be fully accounted for by variables that are completely observed, there has only been limited work addressing estimation in nonparametric regression problems. We propose new multiply robust kernel estimating equations (MRKEEs) that can simultaneously accommodate multiple working models for either the missingness mechanism or the outcome regression, or both. In particular, we consider a new weighted kernel estimating equation approach using counterpart ideas as empirical likelihood, which incorporates potentially many working models for outcome regression and missingness mechanism into the weights. We derive the asymptotic properties using empirical likelihood theory and show that as long as one of the postulated models is correctly specified, the proposed estimator is consistent, and therefore is multiply robust. It means that compared to the kernel methods in the literature, MRKEEs provide extra protection against working

model misspecification. Furthermore, we demonstrate that when including correctly specified models for both the missingness mechanism and the outcome regression, our proposed estimator achieves the optimal efficiency. This work significantly improves model robustness against misspecification in the presence of missing data and is the pioneering work in studying the theories of multiply robust estimators in the context of nonparametric regression.

Chapter III and IV address the problem of estimating the optimal DTR using observational data. To achieve additional flexibility, we explore nonparametric approaches that require a minimal amount of working model specification. However, optimal DTRs estimated from nonparametric models often lack interpretability. To reconcile the tension between interpretability and flexibility, in Chapter III we develop a stochastic tree-based reinforcement learning method, ST-RL, for estimating the optimal DTR in a multi-stage multi-treatment setting with data from either randomized trials or observational studies. At each stage, ST-RL constructs a decision tree by first modeling the mean of counterfactual outcomes via Bayesian nonparametric regression models, and then stochastically searching for the optimal tree-structured regime using a Markov Chain Monte Carlo algorithm. We implement the proposed method in a backward inductive fashion through multiple decision stages. Compared to greedy tree-growing algorithms such as CART, the stochastic tree search can search the decision tree space more efficiently. Besides, we show in theory that the stochastic tree search will not overfit and that the resulting tree-structured optimal regime is consistent by deriving finite sample bound. Compared to existing methods, ST-RL delivers optimal DTRs with better interpretability and does not require explicit working model specifications. Also, ST-RL demonstrates stable and outstanding performance with moderately high dimensional data.

In large observational data, especially Omics data, the dimensionality further in-

creases and eventually necessitates variable selection before estimating optimal DTRs. In Chapter IV, we propose a variable selection technique, Sparse Additive Selection (SpAS), for identifying potential prescriptive and predictive variables. Existing methods have some limitations - first, they require modeling the contrast functions and thus only allow binary treatments, and second, existing methods apply to randomized trial data only. SpAS allows treatments with more than two levels or even continuous doses. Also, SpAS takes advantage of nonparametric sparse additive models and therefore has superior flexibility. In particular, SpAS enforces strong heredity constraint, i.e., the interactions can only be included when both of its main effects have already been selected.

CHAPTER II

Multiply Robust Estimation in Nonparametric Regression with Missing Data

2.1 Introduction

For many biomedical studies, nonparametric regression becomes more and more attractive as it allows for more flexibility on how the outcome depends on a covariate. For example, how the hormone changes over time (*Zhang et al.*, 2000), how disease risk changes over a certain biomarker (*Kennedy et al.*, 2013), and how cardiovascular function changes over air pollution exposure (*Donnelly et al.*, 2011). In this chapter, we consider nonparametric regression in the presence of missing data. To be specific, suppose the outcome is subject to missingness. The data we intend to collect consist of a random sample of n subjects: $(Y_i, Z_i, \boldsymbol{U}_i)$, $i = 1, \cdots, n$, where Y is the outcome, Z is a scalar covariate and $\boldsymbol{U}$ is a vector of auxiliary variables. We are interested in making inference about $\mathbb{E}(Y|Z = z)$ through a generalized nonparametric model

$$\mathbb{E}(Y|Z) = \mu\{\theta(Z)\}, \tag{2.1}$$

where μ is a known monotonic link function (*McCullagh and Nelder*, 1989, Chap. 2) with a continuous first derivative, z is an arbitrary value in the support of Z and $\theta(z)$

is an unknown smooth function of z that needs to be estimated. The collection of auxiliary variables $\boldsymbol{U}$ is commonly seen in practice and some examples can be found in *Pepe* (1992), *Pepe et al.* (1994), and *Wang et al.* (2010). Such variables can help explain the missingness mechanism and improve the estimation efficiency. Let R be a missingness indicator so that $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise. Given the covariate Z and a rich collection of auxiliary variables $\boldsymbol{U}$, we assume that the missingness of Y is independent of Y. That is

$$\Pr\left(R = 1|Z, \boldsymbol{U}, Y\right) = \Pr\left(R = 1|Z, \boldsymbol{U}\right), \tag{2.2}$$

which is known as the missing at random (MAR) mechanism (e.g., *Little and Rubin*, 2002, Chap. 1).

To our best knowledge, there has only been limited work addressing the above nonparametric regression problem with MAR data. *Wang et al.* (2010) showed that the inverse probability weighted (IPW) generalized kernel estimating equations (KEEs) result in consistent estimation when the missingness probability is known or correctly modeled. They also made the extension to augmented IPW (AIPW) (*Robins et al.*, 1994) KEEs by modeling $\mathbb{E}\left(Y|Z, \boldsymbol{U}\right)$ in addition, and their estimator is doubly robust in the sense that it is consistent if either the missingness probability or $\mathbb{E}\left(Y|Z, \boldsymbol{U}\right)$ is correctly modeled. Double robustness allows only one working model for each quantity, yet in practice, it is not uncommon to have multiple working models that all have a reasonable fit to the observed data and none rules out the possibility of others, especially when the dimension of $\boldsymbol{U}$ is large. Through simulations, *Kang et al.* (2007) demonstrated that the AIPW estimators could be severely biased when both working models are only mildly misspecified, and this observation makes it desirable to have a more robust estimation procedure that can simultaneously accommodate multiple working models so that consistency is achieved if any working model is correctly

specified.

Existing multiply robust estimation methods are for marginal mean estimation with missing data (e.g., *Han and Wang*, 2013; *Chan et al.*, 2014; *Han*, 2014a; *Chen and Haziza*, 2017), parametric regression with missing data (e.g., *Han*, 2014b, 2016a) and causal inference (e.g., *Naik et al.*, 2016; *Wang and Tchetgen*, 2016). In this chapter, we make an extension to nonparametric regression and propose a class of multiply robust KEEs (MRKEEs) that offer more protection against working model misspecification than AIPW KEEs. The MRKEE estimators are consistent if any one of the multiple working models for either the missingness probability or $\mathbb{E}(Y|Z, \boldsymbol{U})$ is correctly specified. When correct working models for both are used, the MRKEE estimators achieve the maximum efficiency based on the observed data. The multiple robustness considered in this chapter is different from that in *Tchetgen* (2009), *Molina et al.* (2017) and *Rotnitzky et al.* (2017), where the likelihood function can be factorized into multiple components, for each of which two working models are postulated, and estimation consistency is achieved if for each component there is one model correctly specified. Please refer to *Molina et al.* (2017) for more discussion.

The rest of this chapter is organized as follows. Section 2.2 describes the proposed MRKEEs and their numerical implementation. Section 2.3 investigates the large sample properties of the proposed estimators. Sections 2.4 and 2.5 present the numerical studies and an application example, respectively. Some relevant discussions are provided in Section 2.6, and Appendix A contains some technical details.

2.2 The Proposed Estimator

Without loss of generality, we focus on local linear kernel estimators throughout this chapter. The proposed estimators and their corresponding properties can be generalized to higher order kernel estimators based on similar developments. Let

$K_h(s) = h^{-1}K(s/h)$, where $K(\cdot)$ is a mean-zero density function and h is the bandwidth parameter. Define $\boldsymbol{G}(z) = (1, z)^{\mathrm{T}}$, where z is an arbitrary scalar, and denote $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$. For any target point z, the local linear kernel estimator approximates $\theta(Z)$ in the neighborhood of z by a linear function $\boldsymbol{G}(Z - z)^T\boldsymbol{\alpha}$. Hereafter in our notation we occasionally suppress the dependence on data when it does not cause confusion. To proceed, let $m = \sum_{i=1}^{n} R_i$ be the number of complete cases, and index these subjects by $i = 1, \cdots, m$ without loss of generality. We consider the following multiply robust kernel estimating equations (MRKEEs)

$$\sum_{i=1}^{m} \hat{w}_i K_h(Z_i - z)\mu_i^{(1)} V_i^{-1} \boldsymbol{G}(Z_i - z) \left[Y_i - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}\right] = \mathbf{0}, \tag{2.3}$$

where $\mu_i^{(1)}$ is the first derivative of $\mu(\cdot)$ evaluated at $\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}$, V_i is a working variance model for $\mathrm{Var}\,(Y_i|Z_i)$ indexed by parameter $\boldsymbol{\zeta}$, and $\hat{w}_i$ are certain weights assigned to the complete cases, which we propose to derive based on empirical likelihood techniques and multiple working models for $\Pr(R = 1|Z, \boldsymbol{U})$ and $\mathbb{E}(Y|Z, \boldsymbol{U})$ to achieve multiple robustness. More details will be given below. Our proposed multiply robust estimator of $\theta\,(z)$ is $\widehat{\theta}_{\mathrm{MR}}\,(z) = \widehat{\alpha}_0$ where $\widehat{\boldsymbol{\alpha}}_{MR} = (\widehat{\alpha}_0, \widehat{\alpha}_1)$ is the solution to (2.3). Usually $\boldsymbol{\zeta}$ is unknown and can be estimated via weighted moment equations $\sum_{j=1}^{m} \hat{w}_j V_j^{(1)} \left[\{Y_j - \hat{\alpha}_{0,j}(\boldsymbol{\zeta})\}^2 - V\left\{\hat{\alpha}_{0,j}(\boldsymbol{\zeta}), \boldsymbol{\zeta}\right\}\right] = \mathbf{0}$, where $V_j^{(1)} = \partial V\left\{\hat{\alpha}_{0,j}(\boldsymbol{\zeta}); \boldsymbol{\zeta}\right\}/\partial\boldsymbol{\zeta}$ and $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{\zeta}) = \{\hat{\alpha}_{0,j}(\boldsymbol{\zeta}), \hat{\alpha}_{1,j}(\boldsymbol{\zeta})\}^T$ solves (2.3) with $z = Z_j, j = 1, ..., n$. As noted in *Wang et al.* (2010) and shown in our derivation of Theorem 3 in Appendix, unlike parametric regression where an incorrect working variance model leads to compromised efficiency, the efficiency in estimating $\theta(z)$ will not benefit from correctly estimating $\boldsymbol{\zeta}$. The asymptotic results in Section 2.3 also show that the working variance model plays no role in improving efficiency of $\widehat{\theta}_{\mathrm{MR}}\,(z)$. Thus one can simply set it to be the identity matrix for continuous independent data structure.

If the weights $\widehat{w}_i$ in (2.3) are set to be all equal to one, then (2.3) becomes the naive

complete-case KEEs, which result in biased estimation if the missingness probability is not completely at random, i.e. $\Pr(R=1|Z,\boldsymbol{U},Y)=\Pr(R=1)$. If instead $\hat{w}_i=\Pr(R_i=1|Z_i,\boldsymbol{U}_i)^{-1}$, then (2.3) becomes the IPW KEEs. Consistency of the IPW KEE estimator requires the missingness probability $\Pr(R=1|Z,\boldsymbol{U})$ to be known or correctly modeled. The AIPW KEEs in *Wang et al.* (2010) added an augmentation term involving an outcome regression model for $\mathbb{E}(Y|Z,\boldsymbol{U})$ to the IPW KEEs so that estimation consistency is achieved if either the model for $\Pr(R=1|Z,\boldsymbol{U})$ or the model for $\mathbb{E}(Y|Z,\boldsymbol{U})$ is correctly specified. Our aim is to further improve the robustness by allowing multiple working models for both $\Pr(R=1|Z,\boldsymbol{U})$ and $\mathbb{E}(Y|Z,\boldsymbol{U})$ so that estimation consistency is achieved if any one of these working models is correctly specified, which is the so-called multiple robustness property.

Consider two sets of working models $\mathcal{P}=\{\pi^j(\boldsymbol{\nu}^j):\ j=1,\cdots,J\}$ for $\Pr(R=1|Z,\boldsymbol{U})$ and $\mathcal{A}=\{a^k(\boldsymbol{\gamma}^k):\ k=1,\cdots,K\}$ for $\mathbb{E}(Y|Z,\boldsymbol{U})$, where $\boldsymbol{\nu}^j$ and $\boldsymbol{\gamma}^k$ are the corresponding parameters. We use $\hat{\boldsymbol{\nu}}^j$ and $\hat{\boldsymbol{\gamma}}^k$ to denote the estimators of $\boldsymbol{\nu}^j$ and $\boldsymbol{\gamma}^k$, respectively. Usually, $\hat{\boldsymbol{\nu}}^j$ is taken to be the maximizer of the binomial likelihood

$$\prod_{i=1}^{n}\{\pi_i^j(\boldsymbol{\nu}^j)\}^{R_i}\{1-\pi_i^j(\boldsymbol{\nu}^j)\}^{1-R_i}. \tag{2.4}$$

On the other hand, from (2.2) we have $Y\perp R|(Z,\boldsymbol{U})$ and thus $\mathbb{E}(Y|Z,\boldsymbol{U})=\mathbb{E}(Y|R=1,Z,\boldsymbol{U})$. Therefore, $\hat{\boldsymbol{\gamma}}^k$ can be derived by fitting the model $a^k(\boldsymbol{\gamma}^k)$ based on the complete cases. We consider w_i, $i=1,\cdots,m$, that satisfy the following constraints

$$\begin{aligned}
&\sum_{i=1}^{m} w_i\pi_i^j(\hat{\boldsymbol{\nu}}^j)=n^{-1}\sum_{i=1}^{n}\pi_i^j(\hat{\boldsymbol{\nu}}^j)\quad(j=1,\cdots,J),\\
&\sum_{i=1}^{m} w_i\boldsymbol{\psi}_i^k(\hat{\boldsymbol{\alpha}}^k,\hat{\boldsymbol{\gamma}}^k)=n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}_i^k(\hat{\boldsymbol{\alpha}}^k,\hat{\boldsymbol{\gamma}}^k)\quad(k=1,\cdots,K),
\end{aligned} \tag{2.5}$$

where $\boldsymbol{\psi}^k(\boldsymbol{\alpha},\boldsymbol{\gamma}^k)=K_h(Z-z)\mu^{(1)}V^{-1}\boldsymbol{G}(Z-z)\left[a^k(\boldsymbol{\gamma}^k)-\mu\{\boldsymbol{G}(Z-z)^T\boldsymbol{\alpha}\}\right]$ depends

on the location z, and $\hat{\boldsymbol{\alpha}}^k$ can be obtained by solving the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} K_h(Z_i - z)\mu_i^{(1)} V_i^{-1} \boldsymbol{G}(Z_i - z)\left[R_i Y_i + (1 - R_i)a_i^k(\hat{\boldsymbol{\gamma}}^k) - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}\right] = \mathbf{0}. \tag{2.6}$$

Here (2.6) is actually KEEs for $\boldsymbol{\alpha}$ with the missing outcomes substituted by the fitted values based on the k-th outcome regression model $a^k(\boldsymbol{\gamma}^k)$, and thus $\hat{\boldsymbol{\alpha}}^k$ can be regarded as an imputation-type estimator of $\boldsymbol{\alpha}$ using model $a^k(\boldsymbol{\gamma}^k)$.

The constraints in (2.5) match the weighted averages of certain functions of the observed data based on complete cases to the unweighted averages of those functions based on the whole sample. This construction of constraints is similar to the calibration idea in survey sampling (*Deville and Särndal*, 1992) and helps achieve multiple robustness.

In addition to (2.5), we further impose the positivity constraint $w_i > 0$ and a normalization that $\sum_{i=1}^m w_i = 1$. The compatibility between the positivity and normalization constraints with those in (2.5), or equivalently the existence of a set of weights satisfying all these constraints, can be shown by following the same arguments as those in *Han* (2014b). The $\hat{w}_i$ we propose to use in (2.3) are then derived by maximizing $\prod_{i=1}^m w_i$ subject to $w_i > 0$, $\sum_{i=1}^m w_i = 1$ and the constraints in (2.5).

For ease of presentation, let $\Pi^j(\boldsymbol{\nu}^j) = \frac{1}{n}\sum_{i=1}^n \pi_i^j(\boldsymbol{\nu}^j)$ and $\boldsymbol{\Psi}^k(\boldsymbol{\alpha}, \boldsymbol{\gamma}^k) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}_i^k(\boldsymbol{\alpha}, \boldsymbol{\gamma}^k)$, and write $\hat{\boldsymbol{\nu}}^{\mathrm{T}} = \{(\hat{\boldsymbol{\nu}}^1)^{\mathrm{T}}, \cdots, (\hat{\boldsymbol{\nu}}^J)^{\mathrm{T}}\}$, $\hat{\boldsymbol{\alpha}}^{\mathrm{T}} = \{(\hat{\boldsymbol{\alpha}}^1)^{\mathrm{T}}, \cdots, (\hat{\boldsymbol{\alpha}}^K)^{\mathrm{T}}\}$, $\hat{\boldsymbol{\gamma}}^{\mathrm{T}} = \{(\hat{\boldsymbol{\gamma}}^1)^{\mathrm{T}}, \cdots, (\hat{\boldsymbol{\gamma}}^K)^{\mathrm{T}}\}$ and

$$\begin{aligned}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^{\mathrm{T}} \;=\; & [\pi_i^1(\hat{\boldsymbol{\nu}}^1) - \Pi^1(\hat{\boldsymbol{\nu}}^1), \cdots, \pi_i^J(\hat{\boldsymbol{\nu}}^J) - \Pi^J(\hat{\boldsymbol{\nu}}^J), \\ & \{\boldsymbol{\psi}_i^1(\hat{\boldsymbol{\alpha}}^1, \hat{\boldsymbol{\gamma}}^1) - \boldsymbol{\Psi}^1(\hat{\boldsymbol{\alpha}}^1, \hat{\boldsymbol{\gamma}}^1)\}^{\mathrm{T}}, \cdots, \{\boldsymbol{\psi}_i^K(\hat{\boldsymbol{\alpha}}^K, \hat{\boldsymbol{\gamma}}^K) - \boldsymbol{\Psi}^K(\hat{\boldsymbol{\alpha}}^K, \hat{\boldsymbol{\gamma}}^K)\}^{\mathrm{T}}].\end{aligned}$$

Using the empirical likelihood theory (e.g., *Qin and Lawless*, 1994), we have

$$\hat{w}_i = \frac{1}{m}\frac{1}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})} \quad (i=1,\cdots,m),$$

where $\hat{\boldsymbol{\rho}}^{\mathrm{T}} = (\hat{\rho}_1,\cdots,\hat{\rho}_{J+2K})$ is a $(J+2K)$-dimensional Lagrange multiplier solving

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{1+\boldsymbol{\rho}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})} = \mathbf{0}. \tag{2.7}$$

Because of the positivity of $\hat{w}_i$, $\hat{\boldsymbol{\rho}}$ must satisfy

$$1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) > 0 \quad (i=1,\cdots,m). \tag{2.8}$$

For numerical implementation, calculating $\hat{\boldsymbol{\rho}}$ by directly solving (2.7) is not recommended because (2.7) may have multiple roots and the $\hat{\boldsymbol{\rho}}$ we need is the one that satisfies (2.8). Instead, $\hat{\boldsymbol{\rho}}$ can be derived by minimizing $F_n(\boldsymbol{\rho}) = -n^{-1}\sum_{i=1}^{n} R_i \log\{1+\boldsymbol{\rho}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\}$. Following the same arguments as in *Han* (2014b), it can be shown that this minimization is a convex minimization where a unique minimizer always exists, at least when the sample size is not too small. This minimizer naturally satisfies (2.8) and solves the equation $\partial F_n(\boldsymbol{\rho})/\partial\boldsymbol{\rho} = \mathbf{0}$, which turns out to be (2.7). Refer to *Chen et al.* (2002) and *Han* (2014b) for more details, and a Newton-Raphson-type algorithm for implementation.

Selecting the bandwidth parameter h and estimating the conditional variance $Var(Y|Z = z)$ are crucial for our proposed method. We defer the presentation of these topics to Section 3.2 after we introduce the notation and derive the asymptotics.

2.3 Large Sample Properties

2.3.1 Multiple Robustness

To derive the asymptotic properties, we assume that z is an interior point of the support of Z and that $h = h_n$ is a sequence of bandwidths selected while the sample size n changes such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. In addition, we assume the following regularity conditions hold:

(i) z is in the interior of the support of f_Z, that is, $z \in \text{supp}(f_Z)$, where f_Z is the density of Z.

(ii) For each $z \in \text{supp}(f_Z)$, $(\mu^{-1})^{(1)}\{\mathbb{E}(Y|Z = z)\}$, $V = \text{Var}(Y|Z = z)$ and $[(\mu^{-1})^{(1)}\{\mathbb{E}(Y|Z = z)\}^2 V]^{-1}$ are nonzero, where $(\mu^{-1})^{(1)}$ is the first derivative of the inverse function of μ;

(iii) For each boundary point z_b of $\text{supp}(f_Z)$, there exists an interval $\mathcal{Z}_b$ containing z_b with non-null interior such that $\inf_{z \in \mathcal{Z}_b} f_Z(z) > 0$;

(iv) The functions f'_Z, $\theta^{(2)}(z)$, V, V'' and $(\mu^{-1})^{(3)}$ are continuous;

(v) The function $\frac{\partial^2}{\partial \boldsymbol{\alpha}^2}\boldsymbol{\phi}$ is continuous and uniformly bounded at any $z \in \text{supp}(f_Z)$, where $\boldsymbol{\phi}$ is naive kernel estimating equation.

Condition (iv) requires that the underlying conditional mean outcome, $\theta(z)$, to be twice continuously differentiable. This implies that local linear regression is very flexible in terms that it accommodates a broad class of underlying conditional mean functions, including all smooth functions and less smooth ones as long as they have continuous second derivatives. Moreover, in such cases, using restrictive parametric models such as linear or polynomial regression is unlikely to fit the data well. In this case, MRKEE is an ideal choice for modeling $\theta(z)$ when missing data are present

because of its flexibility. As a trade-off, we sacrifice estimation efficiency, because essentially the inference is also made locally.

We first establish the consistency of $\hat{\theta}_{\mathrm{MR}}(z)$ when $\mathcal{P}$ contains a correctly specified model for $\Pr(R=1|Z,\boldsymbol{U})$. Without loss of generality, let $\pi^1(\boldsymbol{\nu}^1)$ be this correct model and let $\boldsymbol{\nu}_0^1$ denote the true value of $\boldsymbol{\nu}^1$ so that $\pi^1(\boldsymbol{\nu}_0^1)=\Pr(R=1|Z,\boldsymbol{U})$.

Define $\hat{\boldsymbol{\lambda}}^{\mathrm{T}}=(\hat{\lambda}_1,\cdots,\hat{\lambda}_{J+2K})$ in such a way that $\hat{\rho}_1=(\hat{\lambda}_1+1)/\Pi^1(\hat{\boldsymbol{\nu}}^1)$ and $\hat{\rho}_l=\hat{\lambda}_l/\Pi^1(\hat{\boldsymbol{\nu}}^1)$, $l=2,\cdots,J+2K$. Then (2.7) becomes

$$\begin{aligned}
\mathbf{0} &= \frac{1}{\Pi^1(\hat{\boldsymbol{\nu}}^1)}\frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{1+\left\{\frac{\hat{\lambda}_1+1}{\Pi^1(\hat{\boldsymbol{\nu}}^1)},\frac{\hat{\lambda}_2}{\Pi^1(\hat{\boldsymbol{\nu}}^1)},\cdots,\frac{\hat{\lambda}_{J+2K}}{\Pi^1(\hat{\boldsymbol{\nu}}^1)}\right\}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})} \\
&= \frac{1}{\Pi^1(\hat{\boldsymbol{\nu}}^1)}\frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{1+\frac{\pi_i^1(\hat{\boldsymbol{\nu}}^1)-\Pi^1(\hat{\boldsymbol{\nu}}^1)}{\Pi^1(\hat{\boldsymbol{\nu}}^1)}+\left\{\frac{\hat{\boldsymbol{\lambda}}}{\Pi^1(\hat{\boldsymbol{\nu}}^1)}\right\}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})} \\
&= \frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})/\pi_i^1(\hat{\boldsymbol{\nu}}^1)}{1+\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})/\pi_i^1(\hat{\boldsymbol{\nu}}^1)} \qquad (2.9)
\end{aligned}$$

and

$$\hat{w}_i=\frac{1}{m}\frac{\Pi^1(\hat{\boldsymbol{\nu}}^1)/\pi_i^1(\hat{\boldsymbol{\nu}}^1)}{1+\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})/\pi_i^1(\hat{\boldsymbol{\nu}}^1)}. \qquad (2.10)$$

Since $\hat{\boldsymbol{\lambda}}$ solves (2.9), we have $\hat{\boldsymbol{\lambda}}\xrightarrow{p}\mathbf{0}$ from the Z-estimator theory (e.g., *van der Vaart*, 1998, Chap. 5). Let $\boldsymbol{\phi}_i(\boldsymbol{\alpha})=K_h(Z_i-z)\mu_i^{(1)}V_i^{-1}\boldsymbol{G}(Z_i-z)\left[Y_i-\mu\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}\}\right]$ and denote the true parameter value as $\boldsymbol{\alpha}_0$, whose first component is $\theta(z)$. The

MRKEEs (2.3) evaluated at $\boldsymbol{\alpha}_0$ become

$$\begin{aligned}
\sum_{i=1}^{m} \hat{w}_i \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) &= \frac{\Pi^1(\hat{\boldsymbol{\nu}}^1)}{m} \sum_{i=1}^{n} \frac{R_i/\pi_i^1(\hat{\boldsymbol{\nu}}^1)}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\pi_i^1(\hat{\boldsymbol{\nu}}^1)} \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) + o_p(1) \\
&\simeq \mathbb{E}\left\{ \frac{R}{\pi^1(\boldsymbol{\nu}_0^1)} K_h(Z - z) \mu^{(1)} V^{-1} \boldsymbol{G}(Z - z) \left[Y - \mu\{\boldsymbol{G}(Z - z)^T \boldsymbol{\alpha_0}\} \right] \right\} \\
&= f_Z(z) \mu^{(1)} \left\{ \theta(z) \right\} V^{-1} \left\{ \theta(z) \right\} (1, 0)^{\mathrm{T}} \left[\mathbb{E}(Y | Z = z) - \mu(\theta(z)) \right] \\
&\xrightarrow{p} \mathbf{0},
\end{aligned}$$

where $f_Z(\cdot)$ denotes the density of Z and $\simeq$ denotes asymptotic equality because of an omitted $o_p(1)$ term. This result implies the consistency of $\hat{\theta}_{\mathrm{MR}}(z)$ for $\theta(z)$.

We now establish the consistency of $\hat{\theta}_{\mathrm{MR}}(z)$ when $\mathcal{A}$ contains a correctly specified model for $\mathbb{E}(Y|Z, \boldsymbol{U})$. Without loss of generality, let $a^1(\boldsymbol{\gamma}^1)$ be this correct model and $\boldsymbol{\gamma}_0^1$ the true value of $\boldsymbol{\gamma}^1$ such that $a^1(\boldsymbol{\gamma}_0^1) = \mathbb{E}(Y|Z, \boldsymbol{U})$. Let $\boldsymbol{\nu}_*^j$, $\boldsymbol{\alpha}_*^k$, $\boldsymbol{\gamma}_*^k$ and $\boldsymbol{\rho}_*$ denote the probability limits of $\hat{\boldsymbol{\nu}}^j$, $\hat{\boldsymbol{\alpha}}^k$, $\hat{\boldsymbol{\gamma}}^k$ and $\hat{\boldsymbol{\rho}}$, respectively. We then have $\Pi^j(\hat{\boldsymbol{\nu}}^j) \xrightarrow{p} \Pi_*^j$ and $\boldsymbol{\Psi}^k(\hat{\boldsymbol{\alpha}}^k, \hat{\boldsymbol{\gamma}}^k) \xrightarrow{p} \boldsymbol{\Psi}_*^k$ where $\Pi_*^j = \mathbb{E}\{\pi^j(\boldsymbol{\nu}_*^j)\}$ and $\boldsymbol{\Psi}_*^k = \mathbb{E}\{\boldsymbol{\psi}^k(\boldsymbol{\alpha}_*^k, \boldsymbol{\gamma}_*^k)\}$. Write $\boldsymbol{\nu}_*^{\mathrm{T}} = \{(\boldsymbol{\nu}_*^1)^{\mathrm{T}}, \cdots, (\boldsymbol{\nu}_*^J)^{\mathrm{T}}\}$, $\boldsymbol{\alpha}_*^{\mathrm{T}} = \{(\boldsymbol{\alpha}_*^1)^{\mathrm{T}}, \cdots, (\boldsymbol{\alpha}_*^K)^{\mathrm{T}}\}$, $\boldsymbol{\gamma}_*^{\mathrm{T}} = \{(\boldsymbol{\gamma}_*^1)^{\mathrm{T}}, \cdots, (\boldsymbol{\gamma}_*^K)^{\mathrm{T}}\}$, and

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)^{\mathrm{T}} &= [\pi^1(\boldsymbol{\nu}_*^1) - \Pi_*^1, \cdots, \pi^J(\boldsymbol{\nu}_*^J) - \Pi_*^J, \\
&\quad \left\{\boldsymbol{\psi}^1(\boldsymbol{\alpha}_*^1, \boldsymbol{\gamma}_*^1) - \boldsymbol{\Psi}_*^1\right\}^{\mathrm{T}}, \cdots, \left\{\boldsymbol{\psi}^K(\boldsymbol{\alpha}_*^K, \boldsymbol{\gamma}_*^K) - \boldsymbol{\Psi}_*^K\right\}^{\mathrm{T}}]. \quad (2.11)
\end{aligned}$$

Since $a^1(\boldsymbol{\gamma}^1)$ is correctly specified, we must have $\boldsymbol{\gamma}_*^1 = \boldsymbol{\gamma}_0^1$ and $\boldsymbol{\alpha}_*^1 = \boldsymbol{\alpha}_0$. After some algebra, we can write

$$\sum_{i=1}^{m} \hat{w}_i \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) = \sum_{i=1}^{m} \hat{w}_i \{\boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) - \boldsymbol{\psi}_i^1(\hat{\boldsymbol{\alpha}}^1, \hat{\boldsymbol{\gamma}}^1)\} + \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}_i^1(\hat{\boldsymbol{\alpha}}^1, \hat{\boldsymbol{\gamma}}^1).$$

Further calculation shows that this quantity is equal to

$$\begin{aligned}
& \frac{1}{m}\sum_{i=1}^{n}\frac{R_i\{\boldsymbol{\phi}_i(\boldsymbol{\alpha}_0)-\boldsymbol{\psi}_i^1(\hat{\boldsymbol{\alpha}}^1,\hat{\boldsymbol{\gamma}}^1)\}}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}+\mathbb{E}\{\boldsymbol{\psi}^1(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0^1)\}+o_p(1) \\
\xrightarrow{p} \quad & \frac{1}{P(R=1)}\mathbb{E}\left[\frac{R\{\boldsymbol{\phi}(\boldsymbol{\alpha}_0)-\boldsymbol{\psi}^1(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0^1)\}}{1+\boldsymbol{\rho}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)}\right] \\
= \quad & \frac{1}{P(R=1)}\mathbb{E}\left(\mathbb{E}\left[\frac{R\{\boldsymbol{\phi}(\boldsymbol{\alpha}_0)-\boldsymbol{\psi}^1(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0^1)\}}{1+\boldsymbol{\rho}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)}\,\middle|\, Z,\boldsymbol{U}\right]\right)=\mathbf{0},
\end{aligned}$$

which implies the consistency of $\hat{\theta}_{\mathrm{MR}}(z)$ for $\theta(z)$.

Summarizing the above results, we have the following theorem on the multiple robustness of $\hat{\theta}_{\mathrm{MR}}(z)$.

Theorem 2.1. *Under Conditions (i) - (iv), when $\mathcal{P}$ contains a correctly specified model for $Pr(R=1|Z,\boldsymbol{U})$ or $\mathcal{A}$ contains a correctly specified model for $\mathbb{E}(Y|Z,\boldsymbol{U})$, we have $\hat{\theta}_{MR}(z) \xrightarrow{p} \theta(z)$ as $n\to\infty$, $h\to 0$ and $nh\to\infty$.*

2.3.2 Asymptotic Distribution and Efficiency

We derive the asymptotic distribution of the proposed estimator when $\Pr(R=1|Z,\boldsymbol{U})$ is correctly modeled, as is typical in the missing data literature (e.g., *Robins et al.*, 1994, 1995; *Tsiatis*, 2006, Chap. 7). In this case, the previously shown result $\hat{\boldsymbol{\lambda}} \xrightarrow{p} \mathbf{0}$ and the asymptotic expansion of $\sqrt{nh}\hat{\boldsymbol{\lambda}}$ given by the lemma in the Appendix guarantee a closed-form asymptotic variance, which facilitates explicit assessment and possible improvement of the efficiency of MRKEEs. In addition, this case is also of practical importance, because in many two-stage design studies (e.g., *Pepe*, 1992; *Pepe et al.*, 1994), the missingness is determined by the investigator and thus $\Pr(R=1|Z,\boldsymbol{U})$ is known or can be correctly modeled. On the other hand, when no models for $\Pr(R=1|Z,\boldsymbol{U})$ is correctly specified, the derivation of a Taylor series-based asymptotic variance estimator of MRKEEs requires asymptotic expansion of $\sqrt{nh}(\hat{\boldsymbol{\rho}}-\boldsymbol{\rho}^*)$.

Since the true value $\boldsymbol{\rho}^*$ is unknown, the resulting estimator provides little insight into the efficiency of MRKEEs in this case. Moreover, in the case with one correct model for $\mathbb{E}(Y|Z,\boldsymbol{U})$, the asymptotic distribution of θ_{MR} depends on which model is correctly specified. Furthermore, with that information, one would directly estimate $\mathbb{E}(Y|Z,\boldsymbol{U})$ by substituting the missing outcome with the fitted value using the correct model. Therefore, there is little practical interest in deriving the asymptotic distribution of MRKEEs estimator $\hat{\theta}_{MR}(Z)$ when no model for $\Pr(R=1|Z,\boldsymbol{U})$ is correctly specified.

Denote $\pi(Z,\boldsymbol{U}) = \Pr(R=1|Z,\boldsymbol{U})$,

$$\boldsymbol{L} = \mathbb{E}\left\{\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0)\frac{\boldsymbol{g}(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi(Z,\boldsymbol{U})}\right\}, \quad \boldsymbol{M} = \mathbb{E}\left\{\frac{\boldsymbol{g}(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)^{\otimes 2}}{\pi(Z,\boldsymbol{U})}\right\}, \tag{2.12}$$

$$\boldsymbol{Q}(z) = \frac{R}{\pi(Z,\boldsymbol{U})}\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0) - \frac{R-\pi(Z,\boldsymbol{U})}{\pi(Z,\boldsymbol{U})}\boldsymbol{L}\boldsymbol{M}^{-1}\boldsymbol{g}(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*), \tag{2.13}$$

and $c_2(K) = \int s^2K(s)ds$, where $\boldsymbol{B}^{\otimes 2} = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ for any matrix $\boldsymbol{B}$. The asymptotic distribution of $\hat{\theta}_{\mathrm{MR}}(z)$ is given by the following theorem with the proof given in the Appendix.

Theorem 2.2. *Under Conditions (i) - (iv), suppose that $\mathcal{P}$ contains a correctly specified model for $Pr(R=1|Z,\boldsymbol{U})$, then*

$$\sqrt{nh}\left\{\hat{\theta}_{MR}(z) - \theta(z) - \frac{1}{2}h^2\theta''(z)c_2(K) + o(h^2)\right\} \xrightarrow{d} N\left\{0, W_{MR,\pi}(z)\right\}$$

as $n\to\infty$, $h\to 0$ and $nh\to\infty$, where $W_{MR,\pi}(z)$ is the $(1,1)$-element of matrix

$$\left[\mathbb{E}\left\{\frac{\partial\boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial\boldsymbol{\alpha}^T}\right\}\right]^{-1}\mathbb{E}\left\{\boldsymbol{Q}(z)^{\otimes 2}\right\}\left[\mathbb{E}\left\{\frac{\partial\boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial\boldsymbol{\alpha}}\right\}\right]^{-1}.$$

The leading bias term of $\hat{\theta}_{\mathrm{MR}}(z)$ is the same as that of the IPW and AIPW estimators

(*Wang et al.*, 2010). In general, there is no clear comparison of the asymptotic efficiency among these estimators. However, when $\mathcal{A}$ also contains a correctly specified model for $\mathbb{E}(Y|Z,\boldsymbol{U})$, $\hat{\theta}_{\mathrm{MR}}$ becomes more efficient than the IPW estimator as a result of the following theorem, the proof of which is provided in the Appendix as well.

Theorem 2.3. *Under Conditions (i) - (iv), when $\mathcal{P}$ contains a correctly specified model for $Pr(R=1|Z,\boldsymbol{U})$ and $\mathcal{A}$ contains a correctly specified model for $\mathbb{E}(Y|Z,\boldsymbol{U})$, we have*

$$\sqrt{nh}\left\{\hat{\theta}_{MR}(z)-\theta(z)-\frac{1}{2}h^2\theta''(z)c_2(K)+o(h^2)\right\}\xrightarrow{d} N\left\{0, W_{MR,\ opt}(z)\right\}$$

as $n\to\infty$, $h\to 0$ and $nh\to\infty$, where

$$W_{MR,\ opt}(z)=b_K(z)\,\mathbb{E}\left[\frac{Var(Y|Z,\boldsymbol{U})}{\pi(Z,\boldsymbol{U})}+\left[\mathbb{E}(Y|Z,\boldsymbol{U})-\mu\{\theta(Z)\}\right]^2\middle| Z=z\right],$$

$b_K(z)=\int K^2(s)ds/\left\{[\mu^{(1)}\{\theta(z)\}]^2 f_Z(z)\right\}$.

Note that $W_{\mathrm{MR,\ opt}}(z)$ is also the asymptotic variance of the AIPW estimator with both $\Pr(R=1|Z,\boldsymbol{U})$ and $\mathbb{E}(Y|Z,\boldsymbol{U})$ correctly modeled. But $\hat{\theta}_{\mathrm{MR}}(z)$ achieves this efficiency in the presence of multiple models without the knowledge of exactly which ones are correctly specified in $\mathcal{P}$ and $\mathcal{A}$. Some simple algebra shows that, the difference of the asymptotic variances between the IPW estimator and $\hat{\theta}_{\mathrm{MR}}$ is given by

$$b_K(z)\,\mathbb{E}\left[\left(\pi(Z,\boldsymbol{U})^{-1}-1\right)\times\left[\mathbb{E}(Y|Z,\boldsymbol{U})-\mu\{\theta(Z)\}\right]^2|Z=z\right],$$

and thus the efficiency improvement of $\hat{\theta}_{\mathrm{MR}}(z)$ over the IPW estimator is 0 only when $\mathbb{E}(Y|Z,\boldsymbol{U})=\mu\{\theta(Z)\}$. When $\boldsymbol{U}$ and Y are not independent given Z, which is typically the case in practice, $\hat{\theta}_{\mathrm{MR}}(z)$ with one of the $\mathbb{E}(Y|Z,\boldsymbol{U})$ models correctly specified is always more efficient than the IPW estimator.

Remark 1: Bandwidth Selection. The asymptotically optimal bandwidth minimizes the weighted mean integrated squared errors (MISE). The rate of convergence of the optimal bandwidth, is $h_{MR} = [\{\int W_{MR}(z)dz\}/\{c_2^2(K)\int \theta''^2(z)f(z)dz\}]^{1/5}n^{-1/5}$ when $\approx$ contains a correctly specified model for $\Pr(R = 1|Z, \boldsymbol{U})$. In addition, when $\mathcal{P}$ contains a correctly specified model for $\Pr(R = 1|Z, \boldsymbol{U})$ and $\mathcal{A}$ contains a correctly specified model for $\mathbb{E}(Y|Z, \boldsymbol{U})$, the MISE optimal bandwidth $h_{MR,opt} = [\{\int W_{MR,opt}(z)dz\}/\{c_2^2(K)\int \theta''^2(z)f(z)dz\}]^{1/5}n^{-1/5}$. When neither $\approx$ nor $\mathcal{P}$ contains correctly specified model, the MISE optimal rate $h \propto n^{-1/5}$. In practice, we can use a generalized data-driven bandwidth selection approach for nonparametric regression following the empirical bias bandwidth selection (EBBS) method of *Ruppert* (1997). The goal is to select the optimal bandwidth $h_{\text{MR, opt}}$ that minimizes the empirical mean squared error $EMSE\{z; h(z)\}$ of $\widehat{\theta}_{\text{MR}}(z)$, where $EMSE\{z; h(z)\} = \widehat{bias}\left\{\widehat{\theta}_{\text{MR}}(z)\right\}^2 + \widehat{\text{Var}}\left\{\widehat{\theta}_{\text{MR}}(z)\right\}$. We calculate $EMSE\{z; h(z)\}$ at a series of z and $h(z)$, and choose the $h(z)$ that minimizes $EMSE\{z; h(z)\}$. Instead of using plug-in estimators, $\widehat{bias}\left\{\widehat{\theta}_{\text{MR}}(z)\right\}$ is calculated empirically: we fit a univariate polynomial regression model on $(\boldsymbol{h}, \hat{\theta}(z, \boldsymbol{h}))$, where the independent variable $\boldsymbol{h}$ is a series of bandwidths in a neighborhood of the target bandwidth, and outcome $\hat{\theta}(z, \boldsymbol{h})$ is the kernel estimators evaluated at these bandwidths with fixed z. The desired empirical bias is estimated using sum of the nonzero-order terms from this model evaluated at target bandwidth.

Remark 2: Estimation of $Var(Y|Z = z)$. The variance of $\hat{\theta}(Z)$ can be estimated using the sandwich estimator in Theorem 2.2 in certain scenarios, such as two-stage design, where the selection probability is known or can be correctly modeled. The

quantities $\boldsymbol{L}$, $\boldsymbol{M}$, $\boldsymbol{Q}$ and $\mathbb{E}\left\{\frac{\partial \boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\right\}$ can be estimated empirically, where

$$\frac{\partial \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} = K_h(Z_i - z)\mu^{(2)}V^{-1}\boldsymbol{G}(Z_i - z)\boldsymbol{G}^{\mathrm{T}}(Z_i - z)\left[Y_i - \mu\{\boldsymbol{G}(Z_i - z)^{\mathrm{T}}\boldsymbol{\alpha_0}\}\right] - K_h(Z_i - z)[\mu^{(1)}]^2V^{-1}\boldsymbol{G}(Z_i - z)\boldsymbol{G}^{\mathrm{T}}(Z_i - z).$$

To obtain $\hat{\phi}_i$ and $\widehat{\phi^{(1)}}_i$, we estimate the conditional variance, $V_i = Var[\mu\{\boldsymbol{G}(Z_i - z)^{\mathrm{T}}\boldsymbol{\alpha_0}\}]$, using a parametric model and its correct specification is not required to make valid inference. Specifically, we assume a parametric model $Var[\mu\{\boldsymbol{G}(Z_i - z)^{\mathrm{T}}\boldsymbol{\alpha_0}\}, \xi]$, indexed by parameter ξ. The parameters ξ and $\boldsymbol{\alpha}$ can be estimated by iteratively solving proposed MRKEE and the weighted estimating equations $\sum_{i=1}^{n} \hat{w}_i V_i^{(1)}[\{Y_i - \hat{\alpha}_0\}^2 - V\{\hat{\alpha}_0, \xi\}] = 0$, where $\hat{w}_i$ is calculated as in (2.10), $V^{(1)}$ is the first derivative of proposed parametric model, and $\hat{\alpha}_0$ is obtained by plugging $\hat{V}(\hat{\xi})$ from previous iteration into the MRKEE. Furthermore, when $\mathcal{A}$ also contains a correctly specified model for $\mathbb{E}(Y|Z, \boldsymbol{U})$, we can use the same procedure to estimate the optimal variance $W_{\mathrm{MR,\ opt}}(z)$. In general cases where the missingness probability is unknown, we recommend to calculate $\widehat{\mathrm{Var}}\left\{\widehat{\theta}_{\mathrm{MR}}(z)\right\}$ using the bootstrap for constructing point-wise confidence intervals (*McMurry and Politis*, 2008). Specifically, we fit a pilot kernel regression and obtain centered residual estimates $\widetilde{\epsilon}_i = \hat{\epsilon}_i - \frac{1}{n}\sum_{i=1}^{n_s} \hat{\epsilon}_i,\ i = 1, \cdots, n_s$ where $\hat{\epsilon} = Y - \hat{\theta}_h(Z)$ is the empirical residual and n_s is obtained by discarding the residual estimates near the boundary. Bootstrap samples can then be constructed from $Y^* = \hat{\theta}(Z) + \hat{\epsilon}^*$, where $\hat{\epsilon}^*$ are sampled randomly with replacement from $\widetilde{\epsilon}$.

2.4 Simulation Studies

In this section, we conduct numerical studies to investigate the finite-sample performance of the proposed MRKEEs. We consider the local linear regression with a continuous outcome. A random sample of size n is generated as (Z, Y, U, R). Re-

gressor Z is generated from Uniform(0,1) and the auxiliary variable U is generated independently from Uniform(0,6). The outcome Y is normally distributed with variance 2 and mean

$$\mathbb{E}(Y|Z,U) = 4 \cdot m(Z) + 1.3 \cdot U$$

where $m(\cdot) = F_{8,8}(.)$, a unimodal function $F_{p,q}(x) = \Gamma(p+q)\{\Gamma(p)\Gamma(q)\}^{-1}x^{p-1}(1-x)^{q-1}$. The selection indicator R follows a binomial distribution

$$\Pr(R = 1|Z,U) = \text{expit}\{-1.5 + \exp(U-3)\},$$

which makes Y missing at random (MAR) with missingness percentage about 50% on average. The correctly specified models for $\Pr(R=1|Z,U)$ and $\mathbb{E}(Y|Z,U)$ are then $\text{logit}\{\pi^1(\boldsymbol{\nu}^1)\} = \nu_0^1 + \nu_1^1 \cdot \exp(U-3)$ and $a^1(\boldsymbol{\gamma}^1) = \gamma_1^1 \cdot m(Z) + \gamma_2^1 \cdot U$, respectively. We use the following incorrect models in our simulation study for illustration: $\text{logit}\{\pi^2(\boldsymbol{\nu}^2)\} = \nu_0^2 + \nu_1^2 \cdot \exp(U)$ and $a^2(\boldsymbol{\gamma}^2) = \gamma_1^2 \cdot \sin(2\pi \cdot Z)\text{I}(Z \geq 0.8) + \gamma_2^2 \cdot U$. In this simulation study, we use the generalized EBBS bandwidth selection described in Section 3.1. The number of replications for the simulation study is 500 with sample sizes 1000 and 2000. For each simulated data set, we compute $\hat{\theta}_{\text{nomiss}}$ (assuming no missing data), $\hat{\theta}_{\text{naive}}$ (naive complete-case estimator) and $\hat{\theta}_{\text{AIPW}}$ and their variances using the sandwich estimators. We also compute the multiply robust estimator $\hat{\theta}_{\text{MR}}$ with at least one model in each class and estimate its variance using both formula-based and bootstrap estimator over 500 replications. Each estimator is indexed by a four-digit number, and each digit, from left to right, indicates whether $\pi^1(\boldsymbol{\nu}^1)$, $\pi^2(\boldsymbol{\nu}^2)$, $a^1(\boldsymbol{\gamma}^1)$ or $a^2(\boldsymbol{\gamma}^2)$ is used, respectively. We will suppress the dependence of all estimators on z for brevity. For example, $\hat{\theta}_{\text{MR},\,1011}$ denotes the proposed multiply robust estimator based on the correctly specified model $\pi^1(\boldsymbol{\nu}^1)$ and the two outcome regression models $a^1(\boldsymbol{\gamma}^1)$ and $a^2(\boldsymbol{\gamma}^2)$.

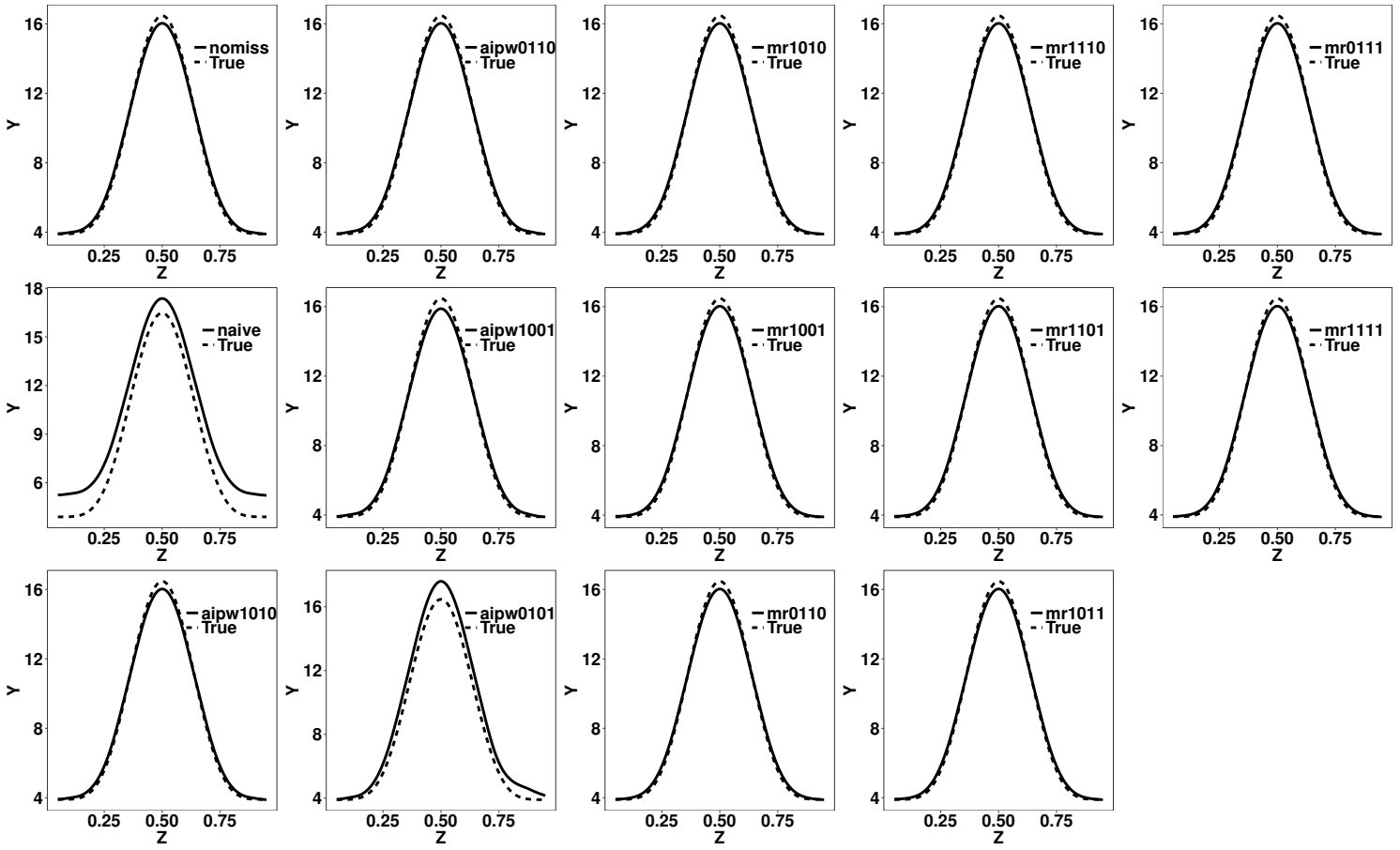

Figure 2.1: Simulation results of the estimated nonparametric functions using naive, AIPW and MR kernel methods based on 500 replications with sample size $n = 2000$.

Figure 1 depicts the empirical mean of estimated curves based on 500 replications with sample size 2000. The naive complete-case estimate is severely biased. The AIPW kernel estimate is unbiased when either the selection probability model or the outcome regression model is correct, but is biased when both are incorrect. The proposed multiply robust estimate is close to the true $\theta(z)$ whenever one of the working models is correctly specified. Figure 3 for sample size 1000 also shows the same pattern. Table 1 summarizes the performance of each estimator using metrics integrated over the support of Z. Consistent with Figure 1 and and Figure 3, as well as the theory in Section 2.3, similar trends on bias are observed for both $n = 1000$ and $n = 2000$ when comparing different methods. Additional information is shown in Table 1 with respect to estimation efficiency of each estimator. Although $\hat{\theta}_{\mathrm{AIPW},1001}$ based on a correct model for $\Pr(R = 1|Z, U)$ and an incorrect model for $\mathbb{E}(Y|Z, U)$ has small relative bias, it has a significant loss of efficiency compared to $\hat{\theta}_{\mathrm{AIPW},1010}$: $\hat{\theta}_{\mathrm{AIPW},1001}$ is only half as efficient in terms of variance and loses 70% of efficiency in terms of MISE. This observation agrees with existing findings for doubly robust estimators. For our proposed method, the relative bias is small whenever a correctly specified model, either for $\Pr(R = 1|Z, U)$ or for $\mathbb{E}(Y|Z, U)$, is used. For example, $\hat{\theta}_{\mathrm{MR},1010}$, $\hat{\theta}_{\mathrm{MR},1001}$, $\hat{\theta}_{\mathrm{MR},0110}$, $\hat{\theta}_{\mathrm{MR},0111}$, $\hat{\theta}_{\mathrm{MR},1011}$, $\hat{\theta}_{\mathrm{MR},1101}$, $\hat{\theta}_{\mathrm{MR},1110}$ and $\hat{\theta}_{\mathrm{MR},1111}$, all have small relative bias ranging from 0.034 to 0.036 for n=2000, and from 0.045 to 0.047 for n=1000. Moreover, when the model of $\Pr(R = 1|Z, U)$ is correctly specified while the regression model is incorrect, in contrast to $\hat{\theta}_{\mathrm{AIPW},1001}$, our multiply robust estimators $\hat{\theta}_{\mathrm{MR},1001}$ and $\hat{\theta}_{\mathrm{MR},1101}$ not only still has little bias, but also are three times as efficient as $\hat{\theta}_{\mathrm{AIPW},1001}$ in terms of the empirical MISE. In addition, $\hat{\theta}_{\mathrm{MR},0101}$, represents the scenario in which both propensity models are misspecified, still has small relative bias and relatively small variance increase compared to AIPW estimator ($\hat{\theta}_{\mathrm{AIPW},0101}$).

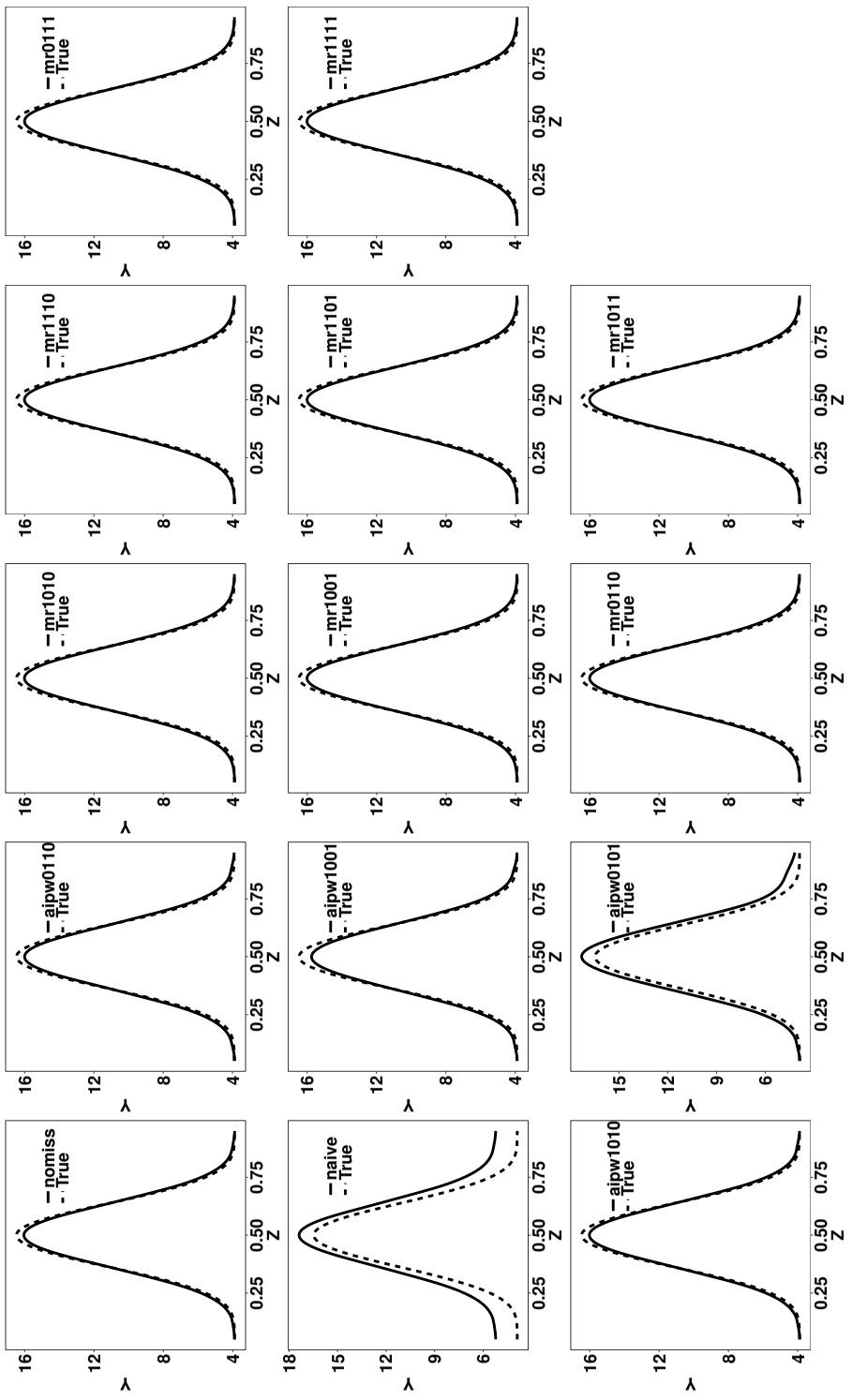

Figure 2.2: Simulation results of the estimated nonparametric functions using naive, AIPW and MR kernel methods based on 500 replications with sample size $n = 1000$.

As an explanation, first note that for an arbitrary function of Z and $\boldsymbol{U}$, $b(Z, \boldsymbol{U})$,

$$\mathbb{E}(w(Z, \boldsymbol{U})[b(Z, \boldsymbol{U}) - \mathbb{E}\{b(Z, \boldsymbol{U})\}]|R = 1) = 0, \tag{2.14}$$

where $w(Z, U) = \Pr(R = 1|Z, \boldsymbol{U})^{-1}$ is the true propensity weight. The constraints we use to construct the weights w in (2.5) can be viewed as an empirical version of (2.14). Thus as we include more working models in weight construction, the resulting w_i's will get closer to normalized true propensity weights. That being said, including more working models – as long as the number is not too large to trigger numerical issue – facilitates the achievement of consistency regardless of the specification of $b(Z, \boldsymbol{U})$. This phenomenon has also been noted in literature (e.g., *Han*, 2016b; *Chen and Haziza*, 2017).

We also evaluate the proposed variance estimators by comparing empirical and estimated standard errors, denoted as EMPSE and ESTSE, respectively. The EMPSE measures the variability over simulation replications, which can be viewed as an alternative to the true underlying variance; the ESTSE measures the average variability estimated using either formula- or bootstrap- based estimators. In particular, we demonstrate the performance of the aforementioned bootstrap variance estimator and compare it with formula-based estimator in Table 1 below. In general, bootstrap performs better than formula-based estimator in terms that estimated and empirical standard errors are closer. When the sample size increases, the differences are getting smaller. In general, with finite sample size, the class of estimators $\hat{\theta}_{\text{MR}}$ have more stable behaviors in terms of bias and efficiency. Comparing to the other methods listed in Table 1, and Figures 1 & 3, MRKEEs provide reliable protection against both bias and severe loss of efficiency.

Table 2.1: Simulation results of relative biases, S.E.s and MISEs of the naive, AIPW and MR estimates of $\theta(z)$ based on 500 replications.

	n = 2000				n = 1000			
	rbias[1]	EMPSE[2]	ESTSE[3]	EMISE[4]	rbias[1]	EMPSE[2]	ESTSE[3]	EMISE[4]
no missing	0.030	0.160	0.155	0.073	0.038	0.232	0.215	0.103
naive	0.226	0.199	0.191	1.901	0.226	0.288	0.263	1.949
aipw1010	0.034	0.196	0.189	0.088	0.044	0.283	0.260	0.136
aipw0110	0.039	0.239	0.229	0.110	0.051	0.342	0.311	0.180
aipw1001	0.049	0.409	0.399	0.294	0.064	0.563	0.540	0.501
aipw0101	0.101	0.615	0.586	1.125	0.112	0.818	0.776	1.396
mr0110	0.035	0.204	0.202	0.091	0.045	0.299	0.287	0.146
mr0101	0.036	0.215	0.201	0.099	0.048	0.316	0.284	0.159
mr0111	0.035	0.205	0.203	0.092	0.046	0.302	0.290	0.148
mr1010	0.034	0.198	0.168	0.088	0.045	0.292	0.227	0.139
mr1010^{b}[5]	0.034	0.199	0.194	0.089	0.045	0.292	0.275	0.141
mr1001	0.036	0.208	0.180	0.094	0.047	0.306	0.244	0.151
mr1001^{b}	0.036	0.209	0.193	0.094	0.047	0.307	0.272	0.153
mr1110	0.034	0.198	0.168	0.088	0.045	0.293	0.227	0.139
mr1110^{b}	0.034	0.200	0.194	0.089	0.045	0.292	0.275	0.141
mr1101	0.036	0.208	0.180	0.094	0.047	0.306	0.244	0.151
mr1101^{b}	0.036	0.209	0.193	0.094	0.047	0.307	0.273	0.153
mr1011	0.034	0.199	0.167	0.089	0.045	0.295	0.225	0.140
mr1011^{b}	0.034	0.200	0.195	0.089	0.045	0.294	0.277	0.143
mr1111	0.034	0.199	0.167	0.089	0.045	0.295	0.225	0.141
mr1111^{b}	0.034	0.200	0.195	0.089	0.045	0.294	0.278	0.143

[1] Relative bias, calculated as $\int |\widehat{bias}\{\hat{\theta}(z)\}/\theta(z)| dF(z)$.

[2] Empirical S.E., defined as $\int \widehat{SE}_{EMP}\{\hat{\theta}(z)\} dF(z)$, where $\widehat{SE}_{EMP}\{\hat{\theta}(z)\}$ is the sampling S.E. of the replicated $\hat{\theta}(z)$.

[3] Estimated S.E., defined as $\int \widehat{SE}_{EST}\{\hat{\theta}(z)\} dF(z)$, where $\widehat{SE}_{EST}\{\hat{\theta}(z)\}$ is the sampling average of the replicated sandwich estimates $\widehat{SE}\{\hat{\theta}(z)\}$.

[4] empirical MISE, defined as $\int \{\hat{\theta}(z) - \theta(z)\}^2 dF(z)$.

[5] b: bootstrap-based variance estimation

2.5 Data Applications

We consider data collected among 2078 high-risk cardiac patients enrolled in a cardiac rehabilitation program at the University of Michigan from 2003 to 2011 (*Giorgini et al.*, 2015a,b). All participants have at least one clinical indication for cardiac rehabilitation, including coronary artery disease (CAD), heart failure, heart valve

repair or replacement, as well as heart transplantation. Our main interest in this application is to evaluate the effect of ozone exposure two days prior to the resting seated systolic blood pressure (SBP). We consider a subset of 704 subjects with measured ozone exposure, of which 308 (44%) have missing SBP because of failure to appear at the evaluation exam. For this study cohort, subjects are between 20 and 86 years old, 73% are male and 31% are current or former smokers. The ozone exposure data are collected two days prior to the exam date from an air pollutants monitoring site maintained by the Michigan Department of Environmental Quality to allow for investigation on the delayed effect on the BP outcomes after ozone exposure. The ozone exposure among our study cohort ranges from 13.1 to 75.5 ppb, with a median 36.5 ppb and interquartile-range (31.5, 41.8). Although the ozone exposure has been reported to be significantly associated with increased blood pressure (*Day et al.*, 2017), the functional form of such an association remains unclear, and is known as non-linear scientifically.

We apply the MRKEEs method to investigate such a potentially non-linear relationship. Since this is an observational study with missing data and thus we are not sure of the missingness mechanism, we fit two generalized linear regression models, using logit link and probit link separately, both with ozone level, age, gender, BMI and smoking status as the potential predictors. We also include quadratic terms of BMI and age, as suggested by previous literature (*Hirano et al.*, 2003; *Wooldridge*, 2007), where they found that overparameterization of missingness probability leads to a more precisely determined point estimate. We use linear regression to fit two conditional mean models with different specifications: one with quadratic age and BMI terms and one without, along with other covariates. Model diagnosis detects no significant deficiencies in fitting all these four models. Thus we applied MRKEEs with these four working models together. We also fit naive KEEs to the data for comparison and use EBBS for bandwidth selection for both methods. The variances

are estimated using bootstrap for MRKEE and sandwich estimator for naive KEEs, respectively. The estimated curves and their 95% CIs are shown in Figure 2. Here we focus our discussion mainly on subjects with ozone exposure from 20 to 50 ppb, because there are very few patients and kernel estimates are unstable outside this range. From Figure 2, the multiply robust kernel regression curve shows that increasing ozone exposure is associated with higher SBP, consistent with findings in the literature (*Day et al.*, 2017). Such an interesting relationship tends to reach a flato as the ozone level reaches high, which makes sense because the SBP cannot explode after reaching high enough. In contrast, the naive complete-case analysis suggests that SBP is relatively stable (around 115 mmHg) and may even slightly decrease as the ozone exposure increases. Notice that among this study cohort, older participants have a higher chance to be absent from the exam, but in the meanwhile, they could be more susceptible to ozone in terms of blood pressure. Therefore, the naive KEEs estimator using only complete cases fails to capture the increasing trend and thus leads to a biased estimate. In contrast, our MRKEEs analysis accounts for the bias due to missing data and suggests that SBP increases relatively fast when subjects start getting exposed to ozone, with the SBP increase slows down as ozone exposure keeps increasing after a certain level.

2.6 Discussion

In this chapter, we proposed a novel multiply robust kernel estimating equations (MRKEEs) method for local polynomial regression when the outcome is missing at random. This method incorporates the auxiliary information by utilizing weights computed through the empirical likelihood method based on multiple working models for the selection probability and/or the outcome regression. Compared to the doubly robust AIPW kernel estimation, the proposed MRKEEs method provides more protection against model misspecification, and the resulting estimator is consistent when

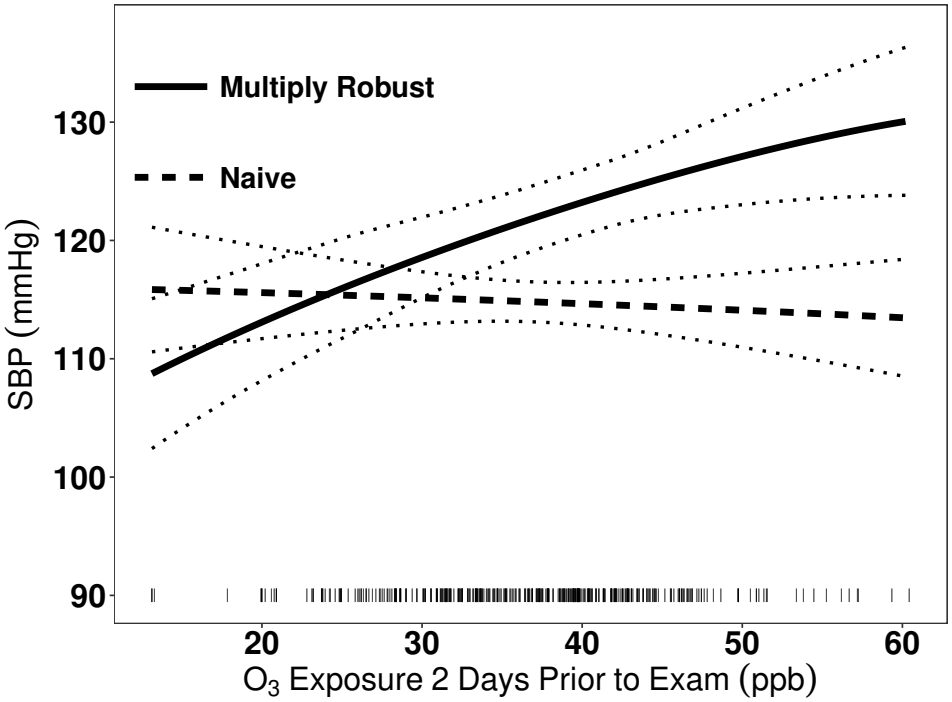

Figure 2.3: The naive and multiply robust KEE estimates of ozone exposure (in ppb) on systolic blood pressure. Each vertical tick mark along the x-axis stands for one observation.

any one of those working models is correctly specified. Moreover, when correct models are used for both quantities, this MRKEE estimator achieves the optimal efficiency that the optimal AIPW estimator can achieve. Simulation studies indicate that the proposed estimator generally has better finite sample performance in terms of both bias and efficiency. Although there is no theory regarding the bias when neither class contains a correct model, under MAR we are able to investigate the working models for $\Pr(R = 1|Z, \boldsymbol{U})$ and $\mathbb{E}(Y|Z, \boldsymbol{U})$ using observed data. Regular model checking and diagnostics are useful in practice, and we will gain efficiency when any one of the working models for $\mathbb{E}(Y|Z, \boldsymbol{U})$ is approximately correct.

As pointed out in *Han and Wang* (2013) and *Han* (2014b), the weight calculation may encounter numerical issues when the sample size is small, or the number of constraints is large. This might happen more often for kernel regression due to the

locality. Therefore some caution is needed when applying this method to datasets with extremely small sample sizes.

The proposed method can be generalized to cases with multiple covariates. One possible extension would be modifying the proposed methods to generalized partially linear model

$$\mathbb{E}(Y|\boldsymbol{X}, Z) = \mu\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \theta\left(Z\right)\right)$$

where $\boldsymbol{X}$ denotes a covariate vector, $\boldsymbol{\beta}$ is the parameter vector and θ is an unknown smooth function. When Y is missing at random, $\boldsymbol{\beta}$ and $\theta(z)$ can be estimated in an iterative fashion based on MRKEEs and a multiply robust version of the profile estimating equations for $\boldsymbol{\beta}$.

Another possible extension would be on single index models assuming that the conditional mean depends on $\boldsymbol{X}$ and Z through a linear combination $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_X + Z\beta_Z$; i.e.

$$\mathbb{E}(Y|\boldsymbol{X}, Z) = \mu\left\{\theta\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_X + Z\beta_Z\right)\right\},$$

where $\theta\left(\cdot\right)$ is an unknown smooth function that we wish to estimate. Parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_X, \beta_Z)$ and $\theta(\cdot)$ can be estimated using a similar iterative method. These extensions will be reported elsewhere.

CHAPTER III

Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes

3.1 Introduction

The emerging field of precision medicine has gained prominence in the scientific community. It aims to improve healthcare quality through tailoring treatment by considering patient heterogeneity. One way to formalize precision health care is dynamic treatment regimes (DTRs, e.g. *Murphy*, 2003; *Robins*, 2004), which are sequential decision rules, one per stage, mapping patient-specific information to a recommended treatment. Consequently, DTRs provide health care that is individualized and also adapted over time to changes in patient status. This is especially valuable in chronic health management (e.g. *Zhao et al.*, 2015). Typically, we define optimal DTRs as the ones that maximize each individual's long term clinical outcome when applied to a population of interest, and thus identification of optimal DTRs becomes the key to precision health care.

Various methods for estimating optimal DTRs have been proposed; some examples include marginal structural models (e.g. *Murphy et al.*, 2001; *Wang et al.*, 2012), G-estimation of structural nested mean models (e.g. *Robins*, 2004), likelihood-based

approaches (e.g. *Thall et al.*, 2007), Q-learning (e.g. *Nahum-Shani et al.*, 2012), A-learning (e.g. *Murphy*, 2003; *Schulte et al.*, 2014), outcome weighted learning (OWL, *Zhao et al.*, 2012, 2015), and other classification or supervised learning methods (e.g. *Zhang et al.*, 2012, 2013). Most of these methods require specifying working models for the treatment assignment mechanism, or the conditional outcome, or both, which can be difficult when limited knowledge of observed data is available. Besides, the working model specification can be especially challenging in the presence of a moderate-to-large number of covariates. *Qian and Murphy* (2011), *Zhao et al.* (2011) and *Moodie et al.* (2014) developed various nonparametric Q-learning methods, and *Murray et al.* (2018) developed a Bayesian machine learning approach based on nonparametric Bayesian regression models. Such data-driven methods are flexible and mitigate the risk of model misspecification; however, the resulting DTRs are difficult to interpret and thus obstruct human experts from understanding the regimes. Therefore, it is often desirable to have interpretable and parsimonious DTRs, as they bridge the gap between clinician's domain expertise and data-driven treatment strategies. The tension between model interpretability and prediction accuracy occurs because of competing objectives: interpretability favors simple and generalizable models, while accuracy often means specialization and sophistication.

In response, a recent research stream has focused on rule-based learning methods for estimating interpretable optimal treatment regimes. *Zhang et al.* (2015, 2018b) proposed a list-based approach to estimate optimal DTRs as a sequence of if-then clauses. This method is flexible since nonparametric kernel ridge regression is used for regime value modeling, but one drawback is that such list-based methods are computationally demanding and therefore, each clause only thresholds two covariates. As an alternative, *Laber and Zhao* (2015) and *Tao et al.* (2018) proposed tree-based methods through sequentially maximizing the improvement in purity measures for the quality of the regimes. However, these semiparametric methods will perform poorly

when working models do not have a good approximation of the underlying truth, which is likely the case with fairly complex data.

To overcome these limitations, we propose a stochastic tree-based reinforcement learning method, ST-RL, which combines the powerful predictive model with an interpretable decision tree structure, for estimation of optimal DTRs in a multi-treatment, multi-stage setting. At each stage, ST-RL evaluates the regime quality using nonparametric Bayesian additive regression trees, and then stochastically constructs an optimal regime using a Markov Chain Monte Carlo (MCMC) tree search algorithm. Our proposed ST-RL has stable performance even when the outcome of interest is complicated by nonlinearity and low-order interactions. ST-RL significantly reduces the guesswork to specify working models, which makes it desirable, especially when data come from complex observational studies. Moreover, compared to existing nonparametric Q-learning methods, ST-RL constructs tree-structured DTRs that are easy to interpret and visualize, which allows clinicians to verify, validate and refine the treatment recommendations using their domain expertise. The greedy splitting in all existing tree-based methods is likely to result in locally optimal or overly complicated trees (*Murthy and Salzberg*, 1995; *Duda et al.*, 2012). In contrast, ST-RL improves the optimality and model parsimony by taking advantage of stochastic tree search.

The rest of this chapter is organized as follows. Section 3.2 formalizes the problem of estimating optimal DTRs and describes the proposed method along with the computational algorithm. Theoretical results are presented in Section 3.3. Sections 3.4 and 3.5 present the numerical studies and an application example, respectively, followed by Section 3.6 that concludes with some discussion.

3.2 Stochastic Tree-based Reinforcement Learning

3.2.1 Dynamic Treatment Regimes

Suppose our data consist of n i.i.d. trajectories $\{(\mathbf{X}_t, A_t, R_t)_{t=1}^T\}$ that come from either a randomized trial or an observational study, where $t \in \{1, 2, \cdots, T\}$ denotes the t^{th} stage, $\mathbf{X}_t$ denotes the vector of patient characteristics accumulated during the treatment period t, A_t denotes a multi-categorical or ordinal treatment indicator with observed value $a_t \in \mathcal{A}_t = \{1, \ldots, K_t\}$, K_t ($K_t \geq 2$) is the number of treatment options at the t^{th} stage, and R_t denotes the immediate reward following A_t. We further let $\mathbf{H}_t$ denote patient history before A_t, i.e. $\mathbf{H}_t = \{(\mathbf{X}_i, A_i, R_i)_{i=1}^{t-1}, \mathbf{X}_t\}$. We consider the overall outcome of interest as $Y = f(R_1, \ldots, R_T)$, where $f(\cdot)$ is a prespecified function (e.g., sum, last value, etc.), and we assume that Y is bounded, with higher values of Y preferable.

We denote a DTR, a sequence of individualized treatment rules, as $\mathbf{g} = (g_1, \ldots, g_T)$, where g_t maps from the domain of patient history $\mathbf{H}_t$ to the domain of treatment assignment A_t. To define the optimal DTRs, we use the counterfactual outcome framework of causal inference and start from the last stage in a reverse sequential order. At the final stage T, let $Y^*(A_1, \ldots, A_{T-1}, a_T)$, or $Y^*(a_T)$ for brevity, denote the counterfactual outcome had a patient been treated with a_T conditional on previous treatments $(A_1, \ldots, A_{T-1})$, and define $Y^*(g_T)$ as the counterfactual outcome under regime g_T, i.e.,

$$Y^*(g_T) = \sum_{a_T=1}^{K_T} Y^*(a_T) I\{g_T(\mathbf{H}_T) = a_T\}.$$

The performance of g_T is measured by its value function $V(g_T)$ (*Qian and Murphy*, 2011), which is defined as the mean counterfactual outcome had all patients followed g_T, i.e. $V(g_T) \equiv \mathbb{E}\{Y^*(g_T)\}$. Therefore, the optimal regime g_T^{opt} satisfies $V(g_T^{opt}) \geq V(g_T)$ for all $g_T \in \mathcal{G}_T$, where $\mathcal{G}_T$ denotes the set of regimes of interest. In

order to identify optimal DTRs using observed data, we make the standard assumptions to link the distribution law of counterfactual data with that of observational data (*Murphy et al.*, 2001; *Robins and Hernán*, 2009). First we assume consistency, i.e. the observed outcome is the same as the counterfactual outcome under the treatment actually assigned, i.e. $Y = \sum_{a_T=1}^{K_T} Y^*(a_T) I(A_T = a_T)$, which also implies that there is no interference between subjects. We also assume $\{Y^*(1), \ldots, Y^*(K_T)\} \perp\!\!\!\perp A_T \mid \mathbf{H}_T$, where $\perp\!\!\!\perp$ denotes statistical independence, i.e. no unmeasured confounding assumption (NUCA). Finally, we assume $\Pr(A_T = a_T|\mathbf{H}_T) \in (c_0, c_1)$ is bounded by c_0 and c_1, where $0 < c_0 < c_1 < 1$. Under these assumptions, the optimal regime at stage T can be written as

$$g_T^{opt} = \arg\max_{g_T \in \mathcal{G}_T} \mathbb{E}\left[\sum_{a_T=1}^{K_T} \mathbb{E}(Y|A_T = a_T, \mathbf{H}_T) I\{g_T(\mathbf{H}_T) = a_T\}\right],$$

where the outer expectation is taken with respect to the joint distribution of the observed data $\mathbf{H}_T$.

At an intermediate stage t ($1 \leq t \leq T-1$), we consider $Y^*(A_1, \ldots, A_{t-1}, g_t, g_{t+1}^{opt}, \ldots, g_T^{opt})$, a conditional counterfactual outcome under optimal regimes for all future stages, had a patient following g_t at stage t (*Murphy*, 2005; *Moodie et al.*, 2012). Similarly under the three aforementioned assumptions, the optimal regime g_t^{opt} at stage t can be defined as

$$\begin{aligned} g_t^{opt} &= \arg\max_{g_t \in \mathcal{G}_t} \mathbb{E}\left\{Y^*(A_1, \ldots, A_{t-1}, g_t, g_{t+1}^{opt}, \ldots, g_T^{opt})\right\} \\ &= \arg\max_{g_t \in \mathcal{G}_t} \mathbb{E}\left[\sum_{a_t=1}^{K_t} \mathbb{E}(\widetilde{Y}_t|A_t = a_t, \mathbf{H}_t) I\{g_t(\mathbf{H}_t) = a_t\}\right], \end{aligned}$$

where $\mathcal{G}_t$ is the set of all potential regimes at stage t, $\widetilde{Y}_T = Y$ at stage T, and at any

earlier stage t, $\widetilde{Y}_t$ can be defined recursively using Bellman's optimality:

$$\widetilde{Y}_t = \mathbb{E}\left\{\widetilde{Y}_{t+1}|A_{t+1} = g_{t+1}^{opt}(\mathbf{H}_{t+1}), \mathbf{H}_{t+1}\right\}, \, t = 1, \cdots, T-1,$$

i.e. the expected outcome assuming optimal regimes are followed at all future stages.

3.2.2 Bayesian Additive Regression Trees

Accurate estimation of value function is the key to estimating optimal DTRs. To mitigate the risk of model misspecification, multiple nonparametric regression methods have been introduced in estimating optimal DTRs, and some recent literature advocates the use of Bayesian nonparametric approaches for causal inference (*Hill*, 2011; *Murray et al.*, 2018). Specifically, Bayesian additive regression trees (BART) has become popular because it requires minimal effort for model specification and can well approximate complex functions involving nonlinearity and interactions. Recent developments in BART have provided theoretical support for its superior performance (*Rockova and van der Pas*, 2017; *Rockova and Saha*, 2018). Moreover, modifications of BART have been proposed to adapt to sparsity and various level of smoothness (*Linero*, 2018; *Linero and Yang*, 2018), which further increase its potential in a wide range of scenarios.

We use BART to fit the conditional outcome regression model $\mathbb{E}(\widetilde{Y}_t|A_t = a_t, \mathbf{H}_t)$ and then predict the counterfactual outcomes for each subject. Other nonparametric regression models can also be used. Specifically, we model the pseudo-outcome at an arbitrary stage t using an ensemble of regression trees:

$$\widetilde{Y}_t = \sum_{i=1}^{m} f_i(\mathbf{H}_t, A_t; \mathcal{M}_i) + \epsilon,$$

where $\epsilon \sim N(0, 1)$, each of f_i's is a binary tree and $\mathcal{M}_i$ is the parameter set charac-

terizing the i-th tree. The number of trees, m, is a tuning parameter which can be selected using cross-validation. BART combines a large number of weak decision tree learners into a strong one to approximate complex functions, and uses a regularization prior for each tree to penalize overly complex trees and hence prevent overfitting. We use the average of after burn-in samples, which approximates the posterior mean, to obtain a prediction $\widehat{\mathbb{E}}(\widetilde{Y}_t|A_t, \mathbf{H}_t)$. For technical details readers are refered to *Chipman et al.* (2010). As a result, the optimal regime at stage t, g_t^{opt} can be estimated as

$$\hat{g}_t^{opt} = \arg\max_{g_t \in \mathcal{G}_t} \mathbb{P}_n \left[\sum_{a_t=1}^{K_t} \widehat{\mathbb{E}}(\widetilde{Y}_t|A_t = a_t, \mathbf{H}_t) I\{g_t(\mathbf{H}_t) = a_t\} \right]. \tag{3.1}$$

3.2.3 Stochastic Tree Search Algorithm

Constructing an optimal binary decision tree is known to be NP-complete (*Laurent and Rivest*, 1976); an exhaustive search is infeasible with even a fairly small number of nodes. Existing tree-based methods iteratively maximize the improvement in a purity function at each splitting. Although such greedy algorithms often work reasonably well, they may be affected by local optimality. As a result, the estimated sub-optimal, unnecessarily complex tree-structured DTRs may have compromised quality and interpretability. Consider a toy example with two baseline variables, X_1 and X_2, both uniformly distributed on the interval [0,1], and a randomly assigned binary treatment $A = 0$ or 1. The underlying optimal regime $g^{opt}(X_1, X_2) = 1$ when ($X_1 > 0.5$, $X_2 > 0.5$), or ($X_1 < 0.5$, $X_2 < 0.5$), and otherwise $g^{opt}(X_1, X_2) = 0$. We assume a linear reward function $Y = X_1 + X_2 + \beta I\{A = g^{opt}(X_1, X_2)\} + \epsilon$, where $\epsilon \sim N(0, 1)$. It is obvious that no matter how we partition the feature space using a tree-based method to assign treatments at the first step, in principle each treatment group always has half of patients ending up with suboptimal treatments. Therefore, the target purity remains the same regardless of any partition in this case, and thus the existing tree-based methods (e.g. *Laber and Zhao*, 2015) will fail to identify the

true optimal regime.

To balance exploration and exploitation, we propose to stochastically search the tree space for the optimal regime using a Markov chain Monte Carlo (MCMC) algorithm (*Chipman et al.*, 1998; *Denison et al.*, 1998; *Wu et al.*, 2007). To make the presentation clear in this section, we suppress the stage subscript t, but the following procedure works for any stage t. Denote a tree-based regime as $g = (\mathcal{P}, \mathcal{R})$, where $\mathcal{P}$ is the parameter set which characterizes the tree topology – i.e., splitting variables, splitting threshold, and the topological arrangement of nodes and edges – and $\mathcal{R}$ is the treatment assignment rule for each leaf node. Note for any given $\mathcal{P}$, $\mathcal{R}$ can be determined by $\mathcal{R} = \arg\max_{\mathcal{R} \in \boldsymbol{R}} V(g)$, i.e., the treatment assignment at each leaf which maximizes the expected regime value. Given the observed data, we sample $\mathcal{P}$ from

$$\pi(\mathcal{P}|A, \mathbf{H}) \propto \pi(\mathcal{P}) f(A|\mathcal{P}, \mathbf{H}), \tag{3.2}$$

where A denotes the observed treatment at the current stage and $\boldsymbol{H}$ denotes patient medical history up to the current stage. However, this sampling procedure is inefficient especially when a large proportion of patients did not receive their optimal treatments, in which case the regular stochastic search might fail to converge to the optimal regime. Alternatively, we use $\pi(\mathcal{P}|\hat{A}^{opt}, \mathbf{H})$ to approximate (3.2), where $\hat{A}^{opt} = \arg\max_{a \in (1,\cdots,K)} \widehat{\mathbb{E}}(\widetilde{Y}|A = a, \mathbf{H})$ is the estimated optimal treatment and can be viewed as a warm start.

We specify $\pi(\mathcal{P})$ using a tree growing process. Let $\mathcal{P} = (\mathcal{T}, \boldsymbol{\rho}, \boldsymbol{\eta})$, where $\mathcal{T}$ is the topological arrangement of nodes and edges (i.e. a tree skeleton), $\boldsymbol{\rho}$ denotes the splitting variables, and $\boldsymbol{\eta}$ denotes the corresponding splitting thresholds. Then the prior distribution over $\mathcal{P}$ can be decomposed as

$$\pi(\mathcal{P}) = \pi(\mathcal{T})\pi(\boldsymbol{\rho}|\mathcal{T})\pi(\boldsymbol{\eta}|\mathcal{T}, \boldsymbol{\rho}). \tag{3.3}$$

Throughout this process, we set $\pi(\boldsymbol{\rho}|\mathcal{T})$ and $\pi(\boldsymbol{\eta}|\mathcal{T},\boldsymbol{\rho})$ to be uniform, i.e. given the tree structure, each variable is equally likely to be selected for splitting, and splitting can take place uniformly on the domain of selected splitting variable. The distribution $\pi(\mathcal{T})$ is specified using a tree growing process. We first draw the tree size $s(\mathcal{T})$ from a shifted Poisson distribution $Pois(\lambda)+1$, with the shift used to avoid an empty tree. Growing a tree with a given size $s(\mathcal{T})$ can be viewed as cascading down $s(\mathcal{T})$ leaf nodes from the root. Assuming that at each internal node the available leaf nodes independently pick the left or right subtree with equal probability, $\pi(\mathcal{T})$ can be written as $\pi_\lambda^{\text{Pois}}\{s(\mathcal{T})-1\}\prod_{u\in\mathcal{T}}\beta(s_{ul}|s_u)$, where $\pi_\lambda^{\text{Pois}}$ denotes the probability mass function of a Poisson random variable with mean λ, u denotes any internal node, s_u is the number of available leaf nodes at u, and s_{ul} is the number of leaf nodes picking the left subtree of u. We set $\beta(s_{ul}|s_u)$ to be uniform $\beta(s_{ul}|s_u)=1/(s_u-1)$, for $s_{ul}=1,\cdots,s_u-1$. Such specification of $\pi(\mathcal{T})$ favors balanced trees, meaning that for a given size, the algorithm penalizes deep trees by assigning them smaller probabilities. For more details regarding the prior specification, the readers are referred to *Wu et al.* (2007). Note that the use of uniform distributions assumes no domain knowledge available a priori; however, other distributions could be used to incorporate human expertise. In practice, the choice of Poisson parameter should reflect preferences over the regime complexity, especially when data are limited.

To obtain $f(\hat{A}^{opt}|\mathcal{P},\mathbf{H})$, we assume that multi-level $\hat{A}^{opt}$ within each leaf node follows i.i.d. multinomial distributions (binomial in case of two-level treatment), leaf node parameters follow Dirichlet distributions, and leaf nodes are independent. Consequently,

$$f(\hat{A}^{opt}|\mathcal{P},\mathbf{H})\propto\int_{\boldsymbol{\Theta}}f(\hat{A}^{opt}|\mathcal{P},\mathbf{H},\boldsymbol{\Theta})\pi(\boldsymbol{\Theta}|\mathcal{P})d\boldsymbol{\Theta}$$

can be calculated analytically because of Dirichlet-Multinomial conjugacy. Thus leaf parameters $\boldsymbol{\Theta}$ are not involved in the sampling scheme.

Algorithm 1: MCMC Algorithm

Result: $\hat{\boldsymbol{g}}^{opt} = (\hat{g}_t^{opt})_{t=1}^T$

Initialize $\mathcal{P}^{(0)} \sim \pi(\mathcal{P})$;

for $i \leftarrow 1, 2, \cdots$ **do**

1. Propose $\mathcal{P}^*$ using one of the proposals $q(\mathcal{P}^*|\mathcal{P}^{(i-1)})$: Grow/Prune, Change, Swap, Restructure;
2. Acceptance probability: $\alpha(\mathcal{P}^*|\mathcal{P}^{(i-1)}) = \min\{1, \frac{q(\mathcal{P}^{(i-1)}|\mathcal{P}^*)\pi(\mathcal{P}^*|\hat{A}^{opt},\mathbf{H})}{q(\mathcal{P}^*|\mathcal{P}^{(i-1)})\pi(\mathcal{P}^{(i-1)}|\hat{A}^{opt},\mathbf{H})}\}$;
3. Sample $u \sim$ Uniform(0,1):

if $u < \alpha$ **then**

Set $\mathcal{P}^{(i)} = \mathcal{P}^*$;

else

Set $\mathcal{P}^{(i)} = \mathcal{P}^{(i-1)}$;

end

end

To implement the MCMC algorithm, we use multiple proposals, including grow/prune, change and swap (*Chipman et al.*, 1998; *Denison et al.*, 1998). More specifically, the grow/prune proposal either splits a randomly chosen leaf node or merges two sibling leaves into one leaf node, with equal probabilities. The change proposal randomly selects an internal node and resamples the splitting rule using $\pi(\boldsymbol{\rho}|\mathcal{T})$ and $\pi(\boldsymbol{\eta}|\mathcal{T}, \boldsymbol{\rho})$. In addition to these two local movements, swap proposal is more aggressive: it randomly picks a node N_p and one of its children N_c, both of which are internal nodes, and swaps their splitting rules, i.e., swaps both splitting variables and the corresponding thresholds. As a result, the whole subtree originally rooted from N_p will be restructured. To facilitate a broader stochastic search, we adopt restructure proposal as in *Wu et al.* (2007): for any given tree, the restructure proposal attempts to construct a different tree while maintaining the current partitioning, i.e., searching for a new tree with the same leaf nodes. This radical proposal drastically modifies the structure

of a tree while preserving the information accumulated so far, which allows escaping from local optimum and leads to a more efficient search on the tree space. Each proposal has an associated probability; in our MCMC implementation, we assign the grow/prune proposal with a probability of 0.5, the change proposal with a probability of 0.5, the swap proposal with a probability of 0.2 and the restructure proposal with a probability of 0.05. The algorithm is summarized as in Algorithm 1.

As mentioned above, for each sampled tree, the treatment assignment rule in each leaf node $\mathcal{R}$ can be uniquely determined and the tree can be evaluated using $V(\hat{g})$. Therefore, the MCMC algorithm provides a powerful tool for stochastic regime construction.

3.2.4 Implementation of ST-RL

ST-RL is implemented in a reverse sequential order. At the final stage T, the pseudo outcome $\widetilde{Y}_T = Y$, and therefore it can be directly used in regime estimation. At each intermediate stage $t < T$, the pseudo outcome $\widetilde{Y}_t$ relies on the optimal treatment regimes in all future stages and needs to be estimated. To prevent bias accumulation, we estimate $\widetilde{Y}_t$ using the actual observed outcomes at stage t plus the predicted future loss due to sub-optimal treatments (*Huang et al.*, 2015); that is

$$\hat{\widetilde{Y}}_t = Y + \sum_{j=t+1}^{T} \{\widehat{\mathbb{E}}[\hat{\widetilde{Y}}_j | g_j^{opt}(\mathbf{H}_j), \mathbf{H}_j] - \widehat{\mathbb{E}}[\hat{\widetilde{Y}}_j | A_j = a_j, \mathbf{H}_j]\},$$

where $\hat{\widetilde{Y}}_T = Y$ and the conditional means are estimated using BART. The performance of BART can be tuned by some hyperparameters, such as number of trees, prior hyperparameters for node parameters and noise. These tuning parameters can be selected using cross-validation; however, in simulation studies, we find that the default parameters from *Chipman et al.* (2010) work well in practice. The algorithm of ST-RL is outlined in Algorithm 2, where the value is estimated using

$\widehat{V}(g_t) = \mathbb{P}_n[\sum_{a_t=1}^{K_t} \widehat{\mathbb{E}}(\hat{\widetilde{Y}}_t|A_t = a_t, \mathbf{H}_t) I\{g_t(\mathbf{H}_t) = a_t\}]$, and the conditional mean is also estimated using BART.

Algorithm 2: ST-RL Algorithm

Result: $\hat{\boldsymbol{g}}^{opt} = (\hat{g}_t^{opt})_{t=1}^T$

Initialize $\widetilde{Y}_T = Y$;

for $t \leftarrow T$ **to** *1* **do**

1. Estimate $\widehat{\mathbb{E}}(\hat{\widetilde{Y}}_t|A_t = a_t, \mathbf{H}_t) = \sum_{i=1}^m \hat{f}_i(\mathbf{H}_t, A_t; \widehat{\mathcal{M}}_i)$ using BART for each subject, where $a_t = 1, \cdots, K_t$;
2. For each subject, obtain $\hat{A}_t^{opt} = \arg\max_{a_t \in (1,\cdots,K_t)} \hat{\mathbb{E}}(\hat{\widetilde{Y}}_t|A_t = a_t, \mathbf{H}_t)$;
3. Perform a stochastic tree search using MCMC Algorithm 1;
4. Estimate the optimal regime using $\hat{g}_t^{opt} = \arg\max_{g_t \in \mathcal{G}_t} \hat{V}(g_t)$;
5. **if** $t > 1$ **then**
   Set $\hat{\widetilde{Y}}_{t-1} = \hat{\widetilde{Y}}_t + \widehat{\mathbb{E}}[\hat{\widetilde{Y}}_t|\hat{g}_t^{opt}(\mathbf{H}_t), \mathbf{H}_t] - \widehat{\mathbb{E}}[\hat{\widetilde{Y}}_t|A_t = a_t, \mathbf{H}_t]$;

else
   Stop;

end

end

3.3 Theoretical Results

In this section we show that the estimated DTRs using ST-RL are well generalizable on new data. We use $\lesssim$ and $\gtrsim$ to denote inequality up to a constant, $||\cdot||_\infty$ and $||\cdot||_n$ to denote the sup norm and empirical norm, respectively. At stage $t \in (1, \cdots, T)$, there are K_t treatment options $a_1, \cdots, a_{K_t}$, and we assume K_t is finite. An arbitrary tree-structured treatment regime g_t (with s leaf nodes) can be denoted as $g_t = (\boldsymbol{p}_t, \boldsymbol{A}_t)$, where $\boldsymbol{p}_t$ denotes any partition of $\mathbf{H}_t$ with K_t subsets and $\boldsymbol{A}_t$ is a vector of the corresponding treatment assignments. More specifically, $\boldsymbol{p}_t = \{p_{it} : p_{it} \subset (L_1, \cdots, L_s)\}_{i=1}^{K_t}$ is coarser than $\boldsymbol{L}$, and $\boldsymbol{L} = \{L_1, \cdots, L_s\}$ is the partition formed by leaf nodes of g_t,

where the i-th leaf $L_i = \{a_{mi} < H_{tm} < b_{mi},\, m = 1, \cdots, d_{\mathbf{H}_t}\}$ and $d_{\mathbf{H}_t}$ is the number of variables in $\mathbf{H}_t$. The partitioning subsets $p_{it} \cap p_{jt} = \emptyset$ for any $i \neq j$, and $\bigcup_{i=1}^{K_t} p_{it} = \mathbf{H}_t$. However, there is no natural ordering in a partition; to avoid ambiguity, we explicitly order partitioning subsets by defining variables and thresholds: suppose subset p_{it} contains two hyper-rectangular cells L_i and $L_{i'}$, then it can be indexed by a vector of lower bound in each dimension: $\boldsymbol{I}_i = \{\min(a_{mi}, a_{mi'}) : m = 1, \cdots, d_{\mathbf{H}_t}\}$. Two index vectors can be compared element-wise: there must exist some integer $m' \in (1, \cdots, d_{\mathbf{H}_t})$ such that $\boldsymbol{I}_{im'} < \boldsymbol{I}_{jm'}$ and $\boldsymbol{I}_{im''} = \boldsymbol{I}_{jm''}$ for $\forall m'' < m'$ if $i < j$. On the other hand, treatment assignment $\boldsymbol{A}_t$ represents the recommended treatment sequence corresponding to the ordered subsets in any partition $\boldsymbol{p}_t$. Each subset has a unique treatment recommendation, and $A_{it} \neq A_{jt}$ for any $i \neq j$. We use $\sigma(\mathcal{T})$ to denote all tree-induced partitions with size K_t as described above. Similarly, we denote the set of all possible treatment assignments as $\sigma(\mathcal{A})$.

Remark. In reality, it might be the case that some treatment options do not associate to any partitions, i.e., such treatments are sub-optimal for all patients. When this is the case, we let the corresponding $p_i = \emptyset$.

To facilitate the derivation, we need a distance metric between two partitions, $d(\boldsymbol{p}, \boldsymbol{p}')$. Various metrics have been proposed: examples include Hamming distance, symmetric difference and rank-based partition distance (*Rossi*, 2011). We use the minimum sum of pair-wise symmetric difference of two partitions: $d(\boldsymbol{p}_t, \boldsymbol{p}'_t) = \sum_{i=1}^{K_t} \rho(p_{it}, p_{f(i)t})$, where $f(\cdot)$ is a bijection from index set $\{1, \cdots, K_t\}$ to itself, and $\rho(p_i, p_j) = \Pr(\mathbf{H}_t \in p_i \triangle p_j) = \Pr(\mathbf{H}_t \in p_i \cup p_j \setminus p_i \cap p_j)$. It is easy to see $d(\boldsymbol{p}_t, \boldsymbol{p}'_t) \geq 0$, $d(\boldsymbol{p}_t, \boldsymbol{p}'_t) = d(\boldsymbol{p}'_t, \boldsymbol{p}_t)$ and it follows triangle inequality; moreover $d(\boldsymbol{p}_t, \boldsymbol{p}'_t) = 0$ does not imply $\boldsymbol{p}_t = \boldsymbol{p}'_t$. Therefore $d(\boldsymbol{p}_t, \boldsymbol{p}'_t)$ is a semi-metric.

Denote the conditional outcome regression model $\mathbb{E}(\widetilde{Y}_t | A_t, \mathbf{H}_t)$ as $Q_t(A_t, \mathbf{H}_t)$. For

each t, our proposed estimator can then be rewritten as

$$
\begin{aligned}
(\hat{\boldsymbol{p}}_t, \hat{\boldsymbol{A}}_t) &= \underset{\boldsymbol{p}_t \in \sigma(\mathcal{T}_t),\, \boldsymbol{A}_t \in \sigma(\mathcal{A}_t)}{\arg\max} \mathbb{P}_n \left[\sum_{a=1}^{K_t} \hat{Q}_t(a, \mathbf{H}_t) \sum_{i=1}^{K_t} I(\mathbf{H}_t \in p_{it}) I(A_{it} = a) \right] \\
&= \underset{\boldsymbol{p}_t \in \sigma(\mathcal{T}_t),\, \boldsymbol{A}_t \in \sigma(\mathcal{A}_t)}{\arg\max} \mathbb{P}_n \hat{F}_t(\boldsymbol{p}_t, \boldsymbol{A}_t),
\end{aligned}
$$

where $\hat{F}_t(\boldsymbol{p}_t, \boldsymbol{A}_t)$ denotes $\sum_{a=1}^{K_t} \hat{Q}_t(a, \mathbf{H}_t) \sum_{i=1}^{K_t} I(\mathbf{H}_t \in p_{it}) I(A_{it} = a)$, and $\hat{Q}_t(a, \mathbf{H}_t) = \hat{\mathbb{E}}(\hat{\widetilde{Y}}_t | A_t = a, \mathbf{H}_t)$. Similarly, the optimal underlying tree-structured regime $g^{*\text{opt}}$ is defined as

$$
(\boldsymbol{p}_t^*, \boldsymbol{A}_t^*) = \underset{\boldsymbol{p}_t \in \sigma(\mathcal{T}_t),\, \boldsymbol{A}_t \in \sigma(\mathcal{A}_t)}{\arg\max} \mathbb{E} F_t(\boldsymbol{p}_t, \boldsymbol{A}_t). \tag{3.4}
$$

In addition to consistency, NUCA, and positivity assumptions, we also make the following assumptions. First, we assume the variables are bounded. Note when the dimension of $\mathbf{H}_t$ is fixed, the results presented later in this section are still valid, but in that case, the following assumption of $d_{\mathbf{H}_t}$ can be dropped.

Assumption 1. *For each t, $||\mathbf{H}_t||_\infty \lesssim \log^{1/2} n$, $|\widetilde{Y}_t| \lesssim \log^{1/2} n$ and $d_{\mathbf{H}_t} \lesssim \log^{1/2} n$.*

Next we assume certain smoothness of $Q_t(a_t, \cdot) = \mathbb{E}(\widetilde{Y}_t | A_t = a_t, \mathbf{H}_t)$.

Assumption 2. *For each t, $Q_t(a_t, \cdot)$ is α-Hölder continuous, and $0 < \alpha \leq 1$. That is,*

$$
Q_t(a_t, \cdot) : [0,1]^p \to \mathbb{R}; \sup_{\mathbf{H}_{t1}, \mathbf{H}_{t2} \in [0,1]^p} \frac{|Q_t(a_t, \mathbf{H}_{t1}) - Q_t(a_t, \mathbf{H}_{t2})|}{||\mathbf{H}_{t1} - \mathbf{H}_{t2}||_2^\alpha} < \infty,
$$

where $0 < \alpha \leq 1$. Moreover, $||Q_t(a_t, \cdot)||_\infty \lesssim \log^{1/2} n$ for all t.

To avoid overfitting, the class of trees under consideration cannot be too complex. *Lugosi et al.* (1996) proved the consistency of a binary decision tree by restricting the number of leaf nodes s_t to $o(n/\log n)$. Recently, *Rockova and van der Pas* (2017) proved in regression case that the sampled tree size is bounded in probability when the

prior tree size is drawn from a Poisson distribution. These conditions are equivalent; we have similar results for our proposed method, as shown in the following theorem.

Theorem 3.1 (Tree Complexity)**.** *For each t, if the stochastic tree search is as described previously, then the sampled tree size, i.e., the number of leaf nodes s_t satisfies*

$$Pr(s_t \gtrsim n^{d_{\mathbf{H}_t}/2\alpha + d_{\mathbf{H}_t}} | \mathbf{H}_t, \hat{A}_t^{opt}) \to 0$$

in P_0^n probability.

Theorem 3.1 guarantees that the tree size of our estimated optimal DTR using ST-RL grows sub-linearly. In other words, ST-RL prevents from overfitting. Compared to other existing tree-based methods, ST-RL does not require stopping rules and pruning procedures, which are typically ad hoc. In addition, we make the following assumption, which guarantees that the optimal tree-structured DTR exists and is unique, i.e. any other optimal tree-structured DTRs $(\boldsymbol{p}^{*'}, \boldsymbol{A}^{*'})$ will satisfy $d(\boldsymbol{p}^{*'}, \boldsymbol{p}^*) = 0$ and $\boldsymbol{A}^{*'} = \boldsymbol{A}^*$.

Assumption 3. *For each t, the following inequalities hold:*

- *For all $\epsilon > 0$, $\sup_{d(\boldsymbol{p},\boldsymbol{p}^*)\geq\epsilon} \mathbb{E}F(\boldsymbol{p}, \boldsymbol{A}^*) < \mathbb{E}F(\boldsymbol{p}^*, \boldsymbol{A}^*)$; moreover, there exists a constant $\kappa > 0$ such that $\mathbb{E}F(\boldsymbol{p}, \boldsymbol{A}^*) - \mathbb{E}F(\boldsymbol{p}^*, \boldsymbol{A}^*) \leq -\kappa d^2(\boldsymbol{p}, \boldsymbol{p}^*)$ as $d(\boldsymbol{p}, \boldsymbol{p}^*) \to 0$.*

- *There exists a constant $\epsilon > 0$ such that $\sup_{\boldsymbol{A}\neq\boldsymbol{A}^*} \mathbb{E}F(\boldsymbol{p}, \boldsymbol{A}) < \mathbb{E}F(\boldsymbol{p}^*, \boldsymbol{A}^*) - \epsilon$.*

Given the above assumptions, the theorem below establishes the stage-specific finite sample bound to evaluate the performance of $\hat{g}_t^{opt}$ compared to g_t^{opt} in terms of classification error and difference in value. For each t, we use $V_t(g_t) = \mathbb{E}[\sum_{a_t=1}^{K_t} \mathbb{E}(\widetilde{Y}_t | A_t = a_t, \mathbf{H}_t) I\{g_t(\mathbf{H}_t) = a_t\}]$ to denote the stage-specific value function for an arbitrary

regime g_t.

Theorem 3.2 (Finite Sample Bound)**.** *Under regularity conditions, for each t and any $\epsilon > 0$, we have*

$$Pr(\hat{g}_t^{opt} \neq g_t^{opt}) \lesssim n^{-r_t+\epsilon},$$
$$Pr\left\{V_t(g_t^{opt}) - V_t(\hat{g}_t^{opt}) \gtrsim \tau n^{-r_t+\epsilon})\right\} \leq e^{-\tau},$$

where $r_T = \frac{2}{3}\frac{\alpha_T}{2\alpha_T + d_{\mathbf{H}_T}}$, *and* $r_t = \frac{2}{3}\min(\frac{\alpha_t}{2\alpha_t + d_{\mathbf{H}_t}}, r_{t+1})$ *for* $t < T$.

Yang (1999) showed that the minimax convergence rate for nonparametric classification in our case is $O(n^{\alpha/(2\alpha+d)})$. The scaling factor $2/3$ arises as a result of non-regular arg max continuous mapping theorem (*Kim et al.*, 1990; *Kosorok*, 2008) and implies that interpretability is achieved at the cost of efficiency. At the final stage T, the rate of convergence is determined by the convergence rate of the conditional outcome regression estimator, which is shown to be $O(n^{2\alpha/(2\alpha+d)} \log n)$ (*Rockova and Saha*, 2018). It becomes more complex at an earlier stage t since the rate of convergence at stage t is also affected by convergence rates at later stages.

3.4 Simulation Study

We demonstrated the performance of ST-RL through a series of simulation studies, including four single-stage scenarios and four two-stage scenarios. In each scenario we generated training data with a sample size $n = 500$ and the dimension of baseline variables $p = 50$, 100 and 200. The baseline covariates X_i, $i = 1, \cdots, p$ were generated from a uniform distribution $U(0, 1)$, with an AR1 correlation structure and a correlation coefficient of 0.4. Following the same mechanism, we generate independent test sets with a sample size of 1000.

We estimated the optimal regime using training data and then predicted the optimal treatments for subjects in the test sets using our estimated optimal DTRs. The following metrics were then calculated using test sets: *opt%* refers to the percentage of subjects correctly classified to their optimal treatments, $\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$ refers to the estimated counterfactual mean outcome under the estimated optimal regime $\hat{g}^{opt}$. *opt%* measures the performance of the estimated optimal regime in assigning future subjects to their optimal treatments, and $\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$ measures the population desirable outcome if following $\hat{g}^{opt}$. ST-RL was implemented using modified BART with Dirichlet splitting rule prior to induce sparsity (*Linero*, 2018). Four methods were compared with ST-RL: Q-learning with linear model (Q-Lin), nonparametric Q-learning with random forest (Q-RF), outcome weighted learning implemented using CART (OWL-CART; *Zhao et al.*, 2015) and tree-based reinforcement learning (T-RL; *Tao et al.*, 2018). In each scenario, the two metrics were averaged over 500 simulation replications.

3.4.1 Single-stage Scenarios

In single-stage setting, we considered four scenarios, all with three-level treatment A taking values in $\{0, 1, 2\}$. In scenarios I and II, the treatment A was generated from a multinomial distribution with probabilities $\{\pi_0, \pi_1, \pi_2\}$, where $\pi_0 = 1/\{1 + \exp(\boldsymbol{X}\boldsymbol{\beta}_1) + \exp(\boldsymbol{X}\boldsymbol{\beta}_2)\}$, $\pi_1 = \exp(\boldsymbol{X}\boldsymbol{\beta}_1)/\{1 + \exp(\boldsymbol{X}\boldsymbol{\beta}_1) + \exp(\boldsymbol{X}\boldsymbol{\beta}_2)\}$ and $\pi_2 = 1 - \pi_0 - \pi_1$, where $\boldsymbol{\beta}_1 = (-1, -1, 0, 2, 0, -1, 1, -1, -1, 2, 0, \cdots, 0)$ and $\boldsymbol{\beta}_2 = (-1, -1, -1, -1, 2, 1, 1, 1, 1, 0, 0, \cdots, 0)$. The underlying optimal treatment regime $g^{opt}(\mathbf{H})$ was specified as:

$$g^{opt}(\mathbf{H}) = \begin{cases} 0 & X_3 \leq 0.35 \\ 1 & X_3 > 0.35,\ 2.5X_2^2 - 0.5X_1 \leq 0.5 \\ 2 & X_3 > 0.35,\ 2.5X_2^2 - 0.5X_1 > 0.5 \end{cases}$$

In scenario I, a simple reward function Y was used with equal loss on sub-optimal treatments $Y = 1 - X_4 + X_5 - X_1 + X_2 + 2 * I(g^{opt} = A) + \epsilon$, and hereafter $\epsilon \sim N(0, 1)$. In contrast, scenario II used a simple reward but with different losses; i.e., the incurred loss depended on the sub-optimal treatment received: $Y = 1 + X_4 + X_5 + 2I(A = 0)\{2 * I(g^{opt} = 0) - 1\} + 1.5I(A = 2)\{2 * I(g^{opt} = 2) - 1\} + \epsilon$.

Scenarios III and IV generated treatment A from a rule-based treatment assignment mechanism: $A = I(X_1^2 + X_2^2 < 0.5) + I(X_4^2 + X_5^2 < 0.5)$, and a relatively complex underlying optimal treatment regime: $g^{opt}(\mathbf{H}) = I(2X_4^2 + X_4^2 > 0.6)[I\{X_2 + 0.5\exp(X_1)^2 + X_7 < 2.25\} + I\{X_3^2 - \log(X_5) - X_6 < 3.5\}]$. Moreover, scenarios III and IV studied complicated reward functions. Scenario III used a reward with equal loss due to sub-optimal treatments $Y = 1 + 2X_5 + 2\exp(X_1) + 5\Phi(X_1)X_2^2 + 5X_3\log(X_5 + 0.1) + \exp\{2X_2 + 1.5I(g^{opt} = A)\} + \epsilon$, while scenario IV used a reward with different losses $Y = 1 + 2X_5 + 2\exp(X_1) + 5\Phi(X_1)X_2^2 + 5X_3\log(X_5 + 0.1) + \exp[2X_2 + 2.5I(A = 0)\{2 * I(g^{opt} = 0) - 1\} + 2I(A = 1)\{2 * I(g^{opt} = 1) - 1\} + 1.5I(A = 2)\{2 * I(g^{opt} = 2) - 1\}] + \epsilon$, where $\Phi(\cdot)$ stands for the cumulative distribution function of the standard normal distribution.

Table 3.1 summarizes the performance of the compared methods in all 4 scenarios. As both T-RL and OWL-CART require estimation of the treatment assignment probabilities, a multinomial logistic regression was fitted, with the observed treatment as the dependent variable and all baseline covariates as explanatory variables. In addition, T-RL requires specification of an outcome regression model for $\mathbb{E}(Y|\boldsymbol{X})$, which we assumed to be a linear regression model. In all four scenarios, ST-RL had outstanding performance compared to the other methods, even when there is a moderately large number of variables (p=200) and a relatively small sample size (n=500). The list-based method had competitive performance in Scenarios I and II; however, its performance was compromised when the underlying optimal regime

Table 3.1: Simulation results for single-stage scenarios I-IV, with 50, 100, 200 baseline covariates and sample size 500. The results are averaged over 500 replications. *opt%* shows the median and IQR of the percentage of test subjects correctly classified to their optimal treatments. $\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and the empirical standard deviance of the expected counterfactual outcome under the estimated optimal regime.

Number of		Scenario I		Scenario II		Scenario III		Scenario IV	
Baseline Covariates		$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*	$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*	$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*	$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*
50	Q-Lin	3.3 (0.1)	65.0 (4.5)	2.5 (0.1)	64.8 (5.2)	16.8 (0.4)	73.5 (3.9)	26.7 (0.9)	73.4 (4.1)
	Q-RF	3.8 (0.1)	89.9 (7.0)	2.5 (0.2)	60.4 (17.2)	16.9 (0.5)	70.6 (5.5)	27.8 (1.0)	67.5 (4.1)
	OWL	2.7 (0.1)	36.2 (5.8)	1.8 (0.2)	38.2 (6.9)	15.7 (0.4)	60.7 (4.3)	24.0 (1.2)	65.0 (5.7)
	List	3.9 (0.1)	95.0 (3.0)	3.1 (0.1)	94.6 (3.3)	17.6 (0.5)	77.0 (5.2)	27.1 (1.1)	63.8 (6.5)
	T-RL	3.7 (0.2)	88.0 (9.0)	2.9 (0.2)	83.2 (16.3)	17.7 (1.0)	85.6 (10.8)	28.0 (1.7)	70.6 (5.8)
	ST-RL	3.9 (0.1)	95.4 (2.0)	3.2 (0.0)	95.8 (1.5)	18.2 (0.6)	90.4 (3.1)	29.9 (0.9)	89.4 (3.8)
100	Q-Lin	2.9 (0.1)	44.1 (6.3)	2.0 (0.2)	44.6 (6.8)	14.6 (0.9)	56.5 (8.5)	21.8 (2.2)	57.5 (9.2)
	Q-RF	3.7 (0.2)	89.2 (9.2)	2.4 (0.2)	54.0 (19.2)	16.7 (0.5)	68.8 (5.4)	27.6 (1.0)	66.5 (4.4)
	OWL	2.7 (0.1)	34.4 (3.1)	1.7 (0.1)	35.2 (4.4)	15.9 (0.4)	63.1 (4.4)	24.6 (1.1)	67.1 (5.1)
	List	3.9 (0.1)	95.3 (2.7)	3.1 (0.1)	94.8 (3.3)	17.5 (0.5)	76.2 (5.6)	26.9 (1.1)	62.4 (5.9)
	T-RL	3.0 (0.4)	52.4 (31.5)	2.2 (0.4)	50.9 (31.1)	16.4 (2.0)	76.0 (23.0)	25.2 (4.5)	65.0 (12.0)
	ST-RL	3.9 (0.1)	95.2 (2.5)	3.1 (0.0)	95.8 (1.5)	18.0 (0.8)	90.0 (3.3)	29.7 (1.2)	89.2 (3.7)
200	Q-Lin	2.7 (0.1)	34.6 (4.3)	1.7 (0.1)	35.1 (4.4)	12.0 (0.7)	33.5 (6.9)	14.7 (1.9)	33.9 (6.8)
	Q-RF	3.7 (0.2)	88.6 (9.5)	2.3 (0.2)	50.4 (20.7)	16.5 (0.5)	67.2 (5.4)	27.4 (1.2)	65.7 (4.5)
	OWL	2.7 (0.1)	34.0 (2.5)	1.7 (0.1)	34.4 (2.7)	15.9 (0.4)	63.1 (3.7)	24.7 (1.1)	67.5 (5.0)
	List	3.9 (0.1)	95.1 (3.4)	3.1 (0.1)	94.8 (3.3)	17.5 (0.5)	75.8 (5.7)	26.9 (1.1)	61.8 (5.3)
	T-RL	2.7 (0.1)	33.5 (3.0)	1.7 (0.2)	33.4 (2.9)	12.7 (1.5)	35.8 (11.4)	15.9 (4.1)	35.6 (10.9)
	ST-RL	3.8 (0.2)	94.8 (3.8)	3.1 (0.1)	95.7 (1.7)	17.7 (1.2)	89.3 (4.8)	29.1 (2.3)	88.6 (4.9)

and reward became complex (scenarios III and IV). T-RL performed well when the data dimension is relatively low, even when both propensity and outcome regression models were incorrectly specified. When the number of variables increased, T-RL performed poorly. Q-Learning methods also significantly varied by case. Both Q-Lin and Q-RF performed better with fewer noise variables, with Q-RF being more stable against increased dimensionality. Consistent with the simulation findings in *Tao et al.* (2018), OWL-CART had an unsatisfactory performance, which is probably due to both misspecification of propensity score models and a low percentage of subjects receiving optimal treatments in the simulated trial or observational data.

3.4.2 Two-stage Scenarios

We considered four two-stage scenarios with a three-level treatment at each stage, and the outcome was defined as the sum of immediate rewards from each stage, i.e. $Y = R_1 + R_2$. In scenarios V and VI, stage 1 treatment A_1 was generated from a multinomial distribution similar to Section 3.4.1, with parameters $\boldsymbol{\beta}_{11} = (-1, -1, 0, 2, 0, -1, 1, 1, -1, 0, 0, \cdots, 0)$ and $\boldsymbol{\beta}_{12} = (-1, -1, -1, -1, 0, 1, 0, 0, 1, 2, 0, \cdots, 0)$. In scenario V, the underlying optimal DTR had a tree structure $g^{opt}_{1,tree} = I(X_1 < 0.7)\{I(X_5 < 0.7) + I(X_{10} > 0.3)\}$, while in scenario VI the optimal DTR did not have a tree structure $g^{opt}_{1,nontree} = I(X_3 > 0.35)\{I(2.5X_2^2 - 0.5X_1 > 0.5) + 1\}$. In both scenarios V and VI, the immediate reward R_1 was generated from $Y = 1 + X_2 + X_4 + X_6 + 2I(A = 0)\{2 * I(g^{opt} = 0) - 1\} + 1I(A = 1)\{2I(g^{opt} = 1) - 1\} + 3I(A = 2)\{2I(g^{opt} = 2) - 1\} + \epsilon$.

At stage 2, in both scenarios V and VI the treatment $A_2 = \{0, 1, 2\}$ was generated from a multinomial distribution with probabilities $\{\pi_{20}, \pi_{21}, \pi_{22}\}$, where $\pi_{20} = 1/\{1 + \exp(\boldsymbol{X}\boldsymbol{\beta}_{21} + 0.2R_1) + \exp(\boldsymbol{X}\boldsymbol{\beta}_{22} + 0.2R_1)\}$, $\pi_{21} = \exp(\boldsymbol{X}\boldsymbol{\beta}_{21} + 0.2R_1)/\{1 + \exp(\boldsymbol{X}\boldsymbol{\beta}_{21} + 0.2R_1) + \exp(\boldsymbol{X}\boldsymbol{\beta}_{22} + 0.2R_1)\}$ and $\pi_{22} = 1 - \pi_{20} - \pi_{21}$, with $\boldsymbol{\beta}_{21} = (-1, -1, 1, 2, 0, -1, 1, 0, 1, -2, 0, \cdots, 0)$ and $\boldsymbol{\beta}_{22} = (1, 1, -1, -1, -2, 1, 1, 1, 0, -1, 0, \cdots,$

0). Similarly to stage 1, scenario V used a tree-type optimal treatment regime $g^{opt}_{2,tree} = I(R_1 > 1)\{I(X_{10} < 0.6) + 1\}$ and scenario VI used a non-tree-type regime $g^{opt}_{2,non-tree} = I(1.5X_2 + X_4 + 0.5R_1) + I(0.5R_1 + X_1 + X_9 > 2)$. The stage 2 reward R_2 was then generated from $R_2 = 1 + X_1 + 2X_5 + 2I(A_2 = 0)\left\{2I(g^{opt}_2 = 0) - 1\right\} + 1.5I(A_2 = 2)\left\{2I(g^{opt}_2 = 2) - 1\right\} + \epsilon$.

We also evaluated the performance of our proposed ST-RL in a randomized trial setting through scenarios VII and VIII. In both scenarios, 3-level treatments A_1 and A_2 were generated from a multinomial distribution with probabilities $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Scenario VII used simple tree-type regimes at both stages: $g^{opt}_{2,tree} = I(X_1 > 0.15)\{I(X_2 > 0.25) + I(X_2 > 0.6)\}$, and $g^{opt}_{2,tree} = I(X_3 > 0.2)\{I(R_1 > 0) + I(X_5 < 0.8)\}$. Scenario VIII used non-tree-type underlying optimal regimes: $g^{opt}_{1,nontree} = I(2X_4^2 + X_4 > 0.6)\{I(X_2 + 0.5\exp(X_1)^2 + X_7 < 2.25) + I(X_3^2 - \log(X_5) - X_6 < 3.5)\}$ and $g^{opt}_{2,nontree} = I(1.5X_2 + X_4 + 0.5R_1 > 2) + I(0.3R_1 + X_1 + X_9 > 2.5)$. In both scenarios, a complex and non-linear reward function was used for stages 1 and 2: $R_1 = \exp\{2 + 0.5X_4 - |4.5X_1 - 1|(A_1 - g^{opt}_1)^2\} + \epsilon$ and $R_2 = \exp\{1 + X_2 - |1.5X_3 + X_4 + 1|(A_2 - g^{opt}_2)^2\} + \epsilon$.

Simulation results of two-stage treatment regimes are summarized in Table 3.2. Essentially the comparison with other methods showed similar trends as observed in one-stage scenarios. For OWL, we applied the backward OWL (BOWL) method in *Zhao et al.* (2015). ST-RL had a reliable performance in both confounded (scenarios V and VI) and randomized (scenarios VII and VIII) settings, even with a large amount of noise interference. The performance of list-based method varied significantly by simulation scenarios: when the underlying optimal regime was fairly complex (VI and VIII), the constraint of at most two variables in each clause might be too restrictive and thus list-based method did not work well. In addition, the list-based method did not always perform consistently even in a simple setting (V). T-RL performed well in randomized settings (VII and VIII), but was severely affected when working models

Table 3.2: Simulation results for two-stage scenarios V-VIII, with 50, 100, 200 baseline covariates and sample size 500. The results are averaged over 500 replications. *opt%* shows the median and IQR of the percentage of test subjects correctly classified to their optimal treatments. $\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and the empirical standard deviance of the expected counterfactual outcome under the estimated optimal regime.

Number of		Scenario V		Scenario VI		Scenario VII		Scenario VIII	
Baseline Covariates		$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*	$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*	$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*	$\hat{\mathbb{E}}\{Y^*(\hat{g}^{opt})\}$	*opt%*
50	Q-Lin	6.0 (0.1)	40.8 (4.3)	6.8 (0.2)	55.8 (5.8)	10.7 (0.2)	43.2 (2.9)	12.2 (0.2)	69.3 (3.8)
	Q-RF	6.6 (0.3)	56.6 (11.3)	7.5 (0.4)	72.4 (16.0)	13.2 (0.4)	81.3 (6.8)	12.6 (0.5)	68.3 (10.1)
	OWL	4.3 (0.2)	13.3 (2.7)	4.2 (0.2)	11.2 (2.8)	8.7 (0.6)	24.8 (6.3)	8.2 (0.6)	20.2 (5.6)
	List	6.3 (0.2)	34.1 (3.2)	6.7 (0.2)	32.4 (10.0)	13.2 (0.7)	83.8 (8.1)	10.6 (0.3)	21.6 (3.0)
	T-RL	6.7 (0.6)	51.9 (26.0)	7.7 (0.4)	75.1 (16.8)	13.4 (0.6)	86.2 (14.1)	13.4 (0.2)	85.7 (7.6)
	ST-RL	7.6 (0.1)	89.4 (3.9)	8.2 (0.1)	88.5 (1.9)	14.1 (0.1)	96.8 (2.1)	13.6 (0.1)	89.0 (3.3)
100	Q-Lin	4.9 (0.4)	21.0 (8.9)	5.6 (0.4)	33.0 (7.4)	9.6 (0.2)	31.5 (3.2)	10.8 (0.3)	51.7 (4.5)
	Q-RF	6.2 (0.4)	47.4 (13.5)	7.3 (0.5)	67.6 (20.2)	12.9 (0.4)	76.2 (8.5)	12.2 (0.5)	62.6 (11.4)
	OWL	4.3 (0.2)	13.5 (2.7)	4.2 (0.1)	11.3 (2.5)	8.0 (0.6)	19.1 (6.4)	7.6 (0.4)	16.7 (4.0)
	List	6.3 (0.2)	33.6 (2.5)	6.5 (0.3)	34.6 (4.2)	13.4 (0.5)	85.1 (5.2)	10.6 (0.6)	21.6 (4.1)
	T-RL	5.8 (0.6)	29.9 (22.5)	7.2 (0.5)	65.1 (22.8)	12.5 (0.6)	67.0 (16.1)	13.1 (0.5)	80.3 (8.5)
	ST-RL	7.6 (0.2)	89.2 (4.0)	8.1 (0.1)	88.6 (1.8)	14.0 (0.2)	96.3 (2.5)	13.6 (0.2)	88.5 (4.1)
200	Q-Lin	4.1 (0.1)	9.2 (2.5)	4.0 (0.2)	9.1 (2.8)	6.9 (0.3)	11.5 (2.4)	6.9 (0.3)	11.8 (3.5)
	Q-RF	5.9 (0.4)	38.0 (12.5)	7.0 (0.6)	61.3 (25.8)	12.4 (0.6)	69.2 (9.2)	12.0 (0.6)	58.2 (13.4)
	OWL	4.8 (0.2)	18.9 (3.8)	4.3 (0.1)	13.2 (2.4)	7.6 (0.6)	15.7 (5.4)	7.5 (0.4)	15.9 (3.8)
	List	6.3 (0.2)	34.0 (2.9)	6.5 (0.3)	34.9 (4.0)	13.4 (0.6)	85.4 (5.0)	10.4 (0.4)	20.6 (5.0)
	T-RL	4.1 (0.2)	10.9 (5.7)	4.1 (0.2)	10.6 (13.4)	6.9 (0.6)	13.0 (6.9)	6.9 (0.7)	17.2 (16.2)
	ST-RL	7.4 (0.4)	86.1 (10.9)	8.1 (0.3)	88.2 (2.1)	13.8 (0.4)	93.7 (5.7)	13.5 (0.2)	87.8 (4.8)

are wrong. Moreover, T-RL was more susceptible to noise interference.

3.5 Data Applications

We applied the proposed ST-RL to an esophageal cancer dataset of 1170 patients collected at MD Anderson Cancer Center from 1998 to 2012. At baseline, patients were between 20 to 91 years old, 85% male, 56% with advanced overall cancer stage (stage III/IV), 19% with squamous cell carcinoma, 81% with Adenocarcinoma, and 44% with well/moderate tumor differentiation. Patients had an average tumor length of 33mm with an interquartile range from 30mm to 37mm. A general treatment strategy for optimizing long-term survival is neoadjuvant chemoradiation followed by esophagectomy (surgery) (*Nieman and Peters*, 2013).

In this application, we primarily focused on the two-stage neoadjuvant chemoradiation disease management before surgery. At baseline, patient characteristics were recorded and denoted as $\boldsymbol{X}$, which include patient information like age, gender, BMI, and disease statuses such as ECOG performance, PET SUV, hypertension status, lesion location, histology, differentiation, tumor length, and overall stage. At the initial treatment stage, 41% patients received induction chemotherapy (ICT). We denote this initial treatment as A_1, with the value YES if treated with ICT and NO otherwise. An intermediate measure of tumor response, R_1, was recorded after this initial treatment to evaluate the effect of A_1, and R_1 ranged from 0 to 5, with 0 being progression and 5 being the complete response.

At treatment stage 2, there were three radiation modalities: 39% of patients received 3D conformal radiotherapy (3DCRT), 45% patients received intensity-modulated radiation therapy (IMRT) and 16% received proton therapy (PT). We denote the stage 2 treatment as A_2, following which, the tumor response R_2 (same scale as R_1) and any development of new lesion N_2 (0 if developed new lesion and 1 if not) were evaluated.

Furthermore, side effects were also recorded, including nausea $S_{2,1}$ (0 if experienced nausea and 1 if not) and anorexia $S_{2,2}$ (0 if experienced anorexia and 1 if not). We defined a composite reward of interest $Y = R_1 + R_2 + 2N_2 + 2S_{2,1} + 2S_{2,2}$ to measure the effectiveness of this two-stage treatment regime. This composite reward balanced multiple competing clinical priorities by incorporating tumor responses measured at the end of two stages and accounting for new lesion development and side effects along the treatment process. Missing data are imputed using IVEware (*Raghunathan et al.*, 2002).

We then applied ST-RL algorithm to the data described above. We used BART to fit the response surface of $\mathbb{E}(\widetilde{Y}_2|A_2 = a_2, \mathbf{H}_2)$, where $\widetilde{Y}_2 = Y$, and obtain the optimal tree-structured regime $\hat{g}_2^{opt}$. The same procedure was repeated for treatment stage 1 to obtain $\hat{g}_1^{opt}$, with $\boldsymbol{H}_1 = \boldsymbol{X}$ and $\widetilde{Y}_1 = Y + \widehat{\mathbb{E}}(\widetilde{Y}_2|\hat{g}_2^{opt}, \mathbf{H}_2) - \widehat{\mathbb{E}}(\widetilde{Y}_2|A_2, \mathbf{H}_2)$. The estimated optimal DTR was $\hat{\boldsymbol{g}}^{opt} = (\hat{g}_1^{opt}, \hat{g}_2^{opt})$, where

$$\hat{g}_1^{opt} = \begin{cases} \text{YES} & \text{if tumor differentiation = poor and cancer stage = I/II and} \\ & \text{tumor length <41mm} \\ \text{NO} & \text{otherwise} \end{cases} \tag{3.5}$$

and

$$\hat{g}_2^{opt} = \begin{cases} \text{PT} & \text{if } A_1 = \text{YES and tumor length} > 29\text{mm} \\ \text{IMRT} & \text{if } A_1 = \text{YES and tumor length} < 29\text{mm} \\ \text{3DCRT} & \text{otherwise} \end{cases} \tag{3.6}$$

The estimated optimal DTR suggests that at the initial treatment stage, ICT is recommended for early-stage esophageal cancer patients with reasonably small-sized tumors that are poorly differentiated. Although a randomized clinical trial (*Ajani et al.*, 2013) suggested that ICT before preoperative chemoradiation would not significantly benefit the esophageal patient population in terms of complete pathologic

response and overall survival, our result can serve as a secondary analysis to help identify subgroups that might benefit from ICT in terms of immediate tumor response and new lesion development. At treatment stage 2, our result showed that patients who received ICT and had larger tumors should use PT while patients who received ICT and had smaller tumors should use IMRT, all other patients should receive 3DCRT. Although it is difficult to draw definitive conclusions regarding clinical outcomes from the existing literature due to lack of published large trials comparing the three modalities, clinical findings demonstrate that IMRT and PT allow the radiation beam to be shaped more precisely to smaller tumors compared to 3DCRT (*Xu and Lin*, 2016). This is consistent with our results that $\hat{g}_2^{opt}$ suggests using PT/IMRT as an alternative treatment for patients if they already used ICT as their initial treatment, which indicates they have a smaller, early-stage tumor according to $\hat{g}_1^{opt}$.

3.6 Discussion

The rapid developments in statistical machine learning have significantly enhanced our ability to flexibly and accurately estimate the most appropriate treatment to patients. However, explicit and interpretable decision rules are preferred by clinicians rather than a machine learning black-box. How to balance the accuracy gain and compromise of interpretability has drawn attention. Efficient communication is the key to any data-driven dynamic treatment regimes to make a practical impact. Therefore, developing methodologies that are both theoretically sound and computationally efficient is critical for further promoting the broader impact of dynamic treatment regimes (DTRs).

We have proposed a new method ST-RL to estimate optimal DTRs in multi-stage multi-treatment settings using potentially large and complex observational data. ST-RL combines flexible nonparametric BART estimation and stochastic policy search

to improve interpretability and quality of data-driven DTRs. Many existing methods focus on randomized trial data; however, in practice, observational data is much more common. Especially in large observational studies, often little is known about the intervention assignment mechanism or the relationship between outcome and covariates. In this case, our newly proposed data-driven and interpretable method ST-RL can mitigate the model misspecification risk, and perform stably across different scenarios. The estimated results are helpful to generate scientific hypotheses and to provide insights on future biomedical research.

The proposed methods can be extended to other types of outcomes, such as binary and right-censored survival data. One remaining question is how to improve the scalability of the stochastic tree search when the data are sparse. With high-dimensional sparse data, one can perform explicit variable selection before regime estimation; however, how to ensure the selection consistency over multiple stages remains a challenge.

CHAPTER IV

A Flexible Tailoring Variable Screening Approach for Estimating Optimal Dynamic Treatment Regimes in Large Observational Studies

4.1 Introduction

Although the traditional "one-size-fits-all" medicine remains a common practice, it is inadequate to provide optimal healthcare to each patient. In contrast, the emerging personalized medicine considers population heterogeneity and is now widely recognized as a promising approach for disease management and treatment for complex health conditions. The idea of using precisely tailored therapies to accommodate patient treatment response heterogeneity over time has been formalized as dynamic treatment regimes (DTRs) in statistics (*Murphy*, 2003). DTRs are sequential decision rules that individualize treatments while adapting to disease progression over time. Optimal DTRs are mappings from up-to-date patient information to treatment recommendations at each stage, such that an average desirable long-term clinical reward is maximized. Various methods have been proposed to estimate optimal DTRs, including marginal structural models (*Murphy et al.*, 2001), G-estimation of structural nested mean models (*Robins*, 2004), likelihood-based approaches (*Thall et al.*, 2007),

robust approaches (*Zhang et al.*, 2012, 2013), machine learning methods (*Zhao et al.*, 2012, 2015; *Laber and Zhao*, 2015).

Nowadays, it is often the case that a large number of variables are collected in observational studies, which has brought new challenges in estimating optimal treatment regimes. One popular example is the Omics data. More specifically, most existing methods require working model specification, such as the model of conditional outcome regression, or the model of treatment assignment probability, or both. High-dimensional data would increase the risk of model misspecification. Although non-parametric techniques can potentially eradicate the need for working models, they are known to be impacted by the curse of dimensionality. Furthermore, many existing methods cannot even work without non-trivial modifications when $p \gg n$. Therefore, it is of great interest to perform explicit variable selection before estimating optimal DTRs when there is a large number of variables. Even though variable selection has been extensively studied in the literature, it is understudied in the area of dynamic treatment regimes. It was not until recently that this topic has drawn some attention. In the pioneering work by *Gunter et al.* (2011), the authors proposed S-Score, a framework for variable selection in treatment individualization. The authors reintroduced the concept of prescriptive variables, i.e., variables having qualitative interactions with treatment. *Fan et al.* (2016) extended this method to a multi-stage scenario and proposed sequential advantage selection (SAS). SAS is closely related to Q-learning and identifies prescriptive variables one at a time while taking into consideration the variables that are already selected in previous steps. Another stream of research is based on the framework of A-leaning. *Lu et al.* (2013) developed a penalized regression approach in single-stage setting; *Shi et al.* (2018) further proposed penalized A-learning (PAL) in multi-stage scenario. In parallel, *Zhang et al.* (2018a) proposed forward minimal misclassification error rate (ForMMER) selection from classification perspective. Another line of research focuses on testing for qualitative interactions;

examples include *Gail and Simon* (1985); *Chang et al.* (2015); *Hsu* (2017); *Shi et al.* (2019).

However, existing methods have some limitations. First of all, most existing methods, if not all, involve modeling the treatment effect contrast functions. As a result, these methods are naturally restricted to treatments with two levels. Second, existing methods require working model specification. S-Score and SAS sequentially build conditional outcome regression models; C-learning and PAL involve modeling the contrasts using doubly robust estimator and therefore require modeling propensity scores. For these methods, the quality of working models will determine the variable selection performance; however, it is challenging to construct working models in high-dimensional data. For example, SAS selects variables sequentially, therefore will likely be biased when some confounding variables have not been included in the model. Besides, ForMMER needs pre-screening procedure when constructing augmented inverse probability weighted estimator (AIPWE) for contrasts when $p > n$. Partially due to this reason, all existing methods are limited to randomized trial data.

In this chapter, we propose a variable selection procedure, sparse additive selection (SpAS), to identify predictive and potential prescriptive variables in estimating optimal DTRs. Our proposed method is closely related to Q-learning and is implemented using backward induction. When modeling the Q-function, i.e., $\mathbb{E}(Y|X, A)$, we use the nonparametric sparse additive model with strong heredity constraint: interactions can only enter the model when both of its main effects are selected. This will help improve model interpretability and plausibility. Such hierarchical variable selection methods have been studied in the case of linear regression (*Yuan et al.*, 2007; *Choi et al.*, 2010; *Bien et al.*, 2013), and also in additive model (*Radchenko and James*, 2010). Moreover, existing methods do not provide the feature of main effect selection; however, it is of importance to identify these variables: literature has shown

they should be included in the working models for propensity scores to improve their quality (*Shortreed and Ertefaie*, 2017). As an improvement, SpAS allows treatments with more than two levels and continuous doses. Moreover, when little is known about the data-generating process, the superior flexibility of SpAS makes it an ideal choice to reduce the number of variables under consideration or being collected.

The rest of this chapter is organized as follows. In Section 4.2 we present the SpAS method and also provide a back-fitting algorithm for model fitting. To demonstrate our proposed method, Section 4.3 compares SpAS with other existing methods in a series of simulation studies. We further illustrate our method in a real data application in Section 4.4, and provide some discussions in Section 4.5.

4.2 Method

4.2.1 Optimal Treatment Regime and Additive Models

We start by presenting the proposed method in single-stage setting and later extend to multi-stage setting in Section 4.2.3. Assume independent data $(\boldsymbol{X}_1, A_1, Y_1), \cdots (\boldsymbol{X}_n, A_n, Y_n)$ are collected, where Y_i denotes real-valued outcome, and $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})$ denotes baseline covariate vector with dimension p, that is potentially large relative to the number of observations, n. Here we assume treatment A_i to be categorical with d_A levels, however our method can generalize to continuous treatments. A treatment regime, $g(\cdot)$, maps from the $\boldsymbol{X}$ to the domain of recommended treatment options. Throughout this chapter, we assume larger values of Y are preferred.

To define optimal treatment regime g^{opt}, we adopt the counterfactual framework in causal inference. Let $Y^*(a)$ denote the counterfactual outcome for a patient that would have been observed had this patient received treatment a, similarly let $Y^*(g)$ denote the counterfactual outcome under regime g. Thus the optimal treatment

regime is the one that maximizes mean counterfactual outcome had all patients followed this regime, i.e., $g^{\text{opt}} = \arg\max_{g\in\mathcal{G}} \mathbb{E}\{Y^*(g)\}$, where $\mathcal{G}$ is the set of all regimes under consideration. To identify the counterfactual quantities from observed data, we make the following assumptions. First we assume consistency, i.e., the observed outcome is the same as the counterfactual outcome under the treatment a patient is actually given, i.e., $Y = \sum_{a=1}^{d_A} Y^*(a)I(A = a)$, where $I(\cdot)$ is the indicator function that takes the value 1 if $\cdot$ is true and 0 otherwise. This assumption also implies that there is no interference between subjects. Moreover, we make no unmeasured confounding assumption (NUCA) that $\{Y^*(1), \ldots, Y^*(d_A)\} \perp\!\!\!\perp A \mid \boldsymbol{X}$. Last, we assume positivity, that is, there exists constants $0 < c_0 < c_1 < 1$ such that with probability 1, the probability $c_0 < Pr(A = a|\boldsymbol{X}) < c_1$.

Under aforementioned assumptions, the optimal regime g^{opt} can be written as $= \arg\max_{g\in\mathcal{G}} \mathbb{E}[\sum_{a=1}^{d_A} Q(A = a, \boldsymbol{X})I\{g(\boldsymbol{X}) = a\}]$, where the Q-function is defined as $Q(A, \boldsymbol{X}) = \mathbb{E}(Y|A, \boldsymbol{X})$. The Q-function is crucial in estimating optimal DTRs. Moreover, in the Q-function, the predictive variables correspond to main effects, while covariates that have interactions with treatment are potentially prescriptive variables. To estimate the Q-function, we consider the following additive model:

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + f_A(A_i) + \sum_{j=1}^{p} f_{Aj}(X_{ij}, A_i) + \epsilon_i, \tag{4.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and is independent of X_i; f_j, f_A and f_{Aj} are functions of covariates that are bounded. For simplicity purpose, we assume Y and all the f's are centered so that intercept is omitted. To avoid trivial solutions, we convert these functions to finite dimensions using basis functions with truncation parameter d_m and d_{in} for main effects and interactions, respectively. Specifically, let $(\psi_1(\cdot), \cdots, \psi_{d_m}(\cdot))$ be a family of uniformly bounded, orthonormal basis with dimension d_m for main effects, and $(\phi_1(\cdot, \cdot), \cdots, \phi_{d_{in}}(\cdot, \cdot))$ be a family of uniformly bounded, orthonormal basis with

dimension d_{in} for interactions. Then we use n by d_m matrix $\boldsymbol{\Psi}_j$ to denote the evaluation of $(\psi_1(\cdot), \cdots, \psi_{d_m}(\cdot))$ on X_{ij}, $i = 1, \cdots, n$, with the (i,k)-th entry being $\psi_k(X_{ij})$. Similarly, we use matrix $\boldsymbol{\Phi}_{Aj}$ to denote the n by d_{in} matrix with (i,k)-th entry being $\phi_k(X_{ij}, A_i)$. On the other hand, $\boldsymbol{\Psi}_A$ denotes the n by d_A treatment indicator matrix with (i,k)-th entry being $I(A_i = k)$. Therefore the Q-function becomes

$$Q(A, \boldsymbol{X}) = \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\beta}_j + \boldsymbol{\Psi}_A \boldsymbol{\beta}_A + \sum_{j=1}^{p} \boldsymbol{\Phi}_{Aj} \boldsymbol{\beta}_{Aj}, \tag{4.2}$$

where $\boldsymbol{\beta}_j$, $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_{Aj}$ are vectors of basis coefficients with lengths d_m, d_A, d_{in} for j-th main effect, treatment and j-th interactions, respectively. To perform variable selection, we take regularization-based approach and consider minimization of the following loss function:

$$\frac{1}{2n} \left\| \boldsymbol{Y} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\beta}_j - \boldsymbol{\Psi}_A \boldsymbol{\beta}_A - \sum_{j=1}^{p} \boldsymbol{\Phi}_{Aj} \boldsymbol{\beta}_{Aj} \right\|_2^2 + \boldsymbol{P}_{\boldsymbol{\lambda}}, \tag{4.3}$$

where $\boldsymbol{P}_{\boldsymbol{\lambda}}$ is the penalty term indexed by regularization parameters λ. There are different ways to enforce strong heredity constraint through formulation of $\boldsymbol{P}_{\boldsymbol{\lambda}}$: one strategy is directly imposing penalty on basis coefficients $\boldsymbol{\beta}_j$, $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_{Aj}$ (*Bien et al.*, 2013). As an alternative, the penalty can be put on each term of $\boldsymbol{\Psi}_j \boldsymbol{\beta}_j$, $\boldsymbol{\Psi}_A \boldsymbol{\beta}_A$ and $\boldsymbol{\Phi}_{Aj} \boldsymbol{\beta}_{Aj}$ (*Radchenko and James*, 2010). In this chapter we take the latter approach because it is reported to have better empirical performance (*She et al.*, 2018). Specifically, we let

$$\begin{aligned}
\boldsymbol{P}_{\boldsymbol{\lambda}} =& P_{\lambda_1} + P_{\lambda_2} \\
=& \lambda_1 \left[\sum_{j=1}^{p} \left(\left\| \boldsymbol{\Psi}_j \boldsymbol{\beta}_j \right\|_2^2 + \left\| \boldsymbol{\Phi}_{Aj} \boldsymbol{\beta}_{Aj} \right\|_2^2 \right)^{1/2} + \left(\left\| \boldsymbol{\Psi}_A \boldsymbol{\beta}_A \right\|_2^2 + \sum_{j=1}^{p} \left\| \boldsymbol{\Phi}_{Aj} \boldsymbol{\beta}_{Aj} \right\|_2^2 \right)^{1/2} \right] \\
& + \lambda_2 \sum_{j=1}^{p} \left\| \boldsymbol{\Phi}_{Aj} \boldsymbol{\beta}_{Aj} \right\|_2 .
\end{aligned}$$

Note P_{λ_1} penalizes total effect of each variable including main effects and interactions, while P_{λ_2} puts an extra penalty on interaction terms. They together enforce strong hierarchy, details are shown in Section 4.2.2.

4.2.2 Back-fitting Algorithm for Sparse Additive Selection

The optimization in 4.3 can be solved via block-wise coordinate descent (*Yuan and Lin*, 2006). In every iteration, the algorithm updates main effects of baseline variables, treatment and their interactions, one at a time, while holding all other terms fixed. Correspondingly, we denote the projection matrices for these three types of terms as $\mathcal{P}_j = \boldsymbol{\Psi}_j(\boldsymbol{\Psi}_j\boldsymbol{\Psi}_j^\top)^{-1}\boldsymbol{\Psi}_j^\top$, $\mathcal{P}_A = \boldsymbol{\Psi}_A(\boldsymbol{\Psi}_A\boldsymbol{\Psi}_A^\top)^{-1}\boldsymbol{\Psi}_A^\top$ and $\mathcal{P}_{Aj} = \boldsymbol{\Phi}_{Aj}(\boldsymbol{\Phi}_{Aj}\boldsymbol{\Phi}_{Aj}^\top)^{-1}\boldsymbol{\Phi}_{Aj}^\top$. The algorithm is outlined in Algorithm 3.

The soft thresholding operators can be obtained for baseline variables, treatment and interactions separately using a solver for linear equations. For j-th main effect $\boldsymbol{\Psi}_j\hat{\boldsymbol{\beta}}_j$, when its interaction with treatment $\left\|\boldsymbol{\Phi}_{Aj}\hat{\boldsymbol{\beta}}_{Aj}\right\|_2 = 0$, the soft thresholding operator has closed-form $\mathcal{S}_j = 1/(1-\lambda_1/\left\|\hat{\boldsymbol{P}}_j\right\|_2)_+$ (*Ravikumar et al.*, 2009), where $(\cdot)_+$ denotes the positive part. On the other hand, when j-th interaction $\left\|\boldsymbol{\Phi}_{Aj}\hat{\boldsymbol{\beta}}_{Aj}\right\|_2 \neq 0$, i.e., it has already been selected, the operator $\mathcal{S}_j$ needs to be solved numerically through:

$$\mathcal{S}_j\left(1+\frac{\lambda_1}{\sqrt{\mathcal{S}_j^2\left\|\hat{\boldsymbol{P}}_j\right\|_2^2+\left\|\boldsymbol{\Phi}_{Aj}\hat{\boldsymbol{\beta}}_{Aj}\right\|_2^2}}\right)=1.$$

For treatment main effect, the thresholding operator $\mathcal{S}_A$ has closed-form $\mathcal{S}_A = 1/(1-\lambda_1/\left\|\hat{\boldsymbol{P}}_A\right\|_2)_+$ when $\sum_{k=1}^p\left\|\boldsymbol{\Phi}_{Ak}\hat{\boldsymbol{\beta}}_{Ak}\right\|_2^2$ is zero, i.e. none of the interactions are selected. In this case our model 4.1 degenerates to a regular sparse additive model. When any

of the interactions are selected, $\mathcal{S}_A$ can be obtained by solving

$$\mathcal{S}_A\left(1+\frac{\lambda_1}{\sqrt{\mathcal{S}_A^2\left\|\hat{\boldsymbol{P}}_A\right\|_2^2+\sum_{k=1}^p\left\|\boldsymbol{\Phi}_{Ak}\hat{\boldsymbol{\beta}}_{Ak}\right\|_2^2}^2}\right)=1.$$

The above results show that a term will only be selected when the corresponding projected residual, $\left\|\hat{\boldsymbol{P}}\right\|_2$ is above certain threshold. In summary, when j-th interaction term $f_{Aj}(X_j, A)$ is not selected, i.e. $\left\|\boldsymbol{\Phi}_{Aj}\hat{\boldsymbol{\beta}}_{Aj}\right\|_2=0$, both corresponding main effects, $\hat{f}_j(X_j)$ and $\hat{f}_A(A)$ needs to be greater than λ_1 in order to be selected. Otherwise when $f_{Aj}(X_j, A)$ is selected, the entering threshold for both main effects drops to 0. In other words, this mechanism guarantees strong heredity principle: an interaction can only be selected when both its main effects are selected. Similarly for the treatment main effect, the threshold is λ_1 when no interactions are selected; however as soon as one interaction enters the model, treatment will be selected automatically.

The shrinkage operators for interaction terms are slightly more complicated. For j-th interaction term $\boldsymbol{\Phi}_{Aj}\hat{\boldsymbol{\beta}}_{Aj}$, the shrinkage operator $\mathcal{S}_{Aj}=1/(1-(2\lambda_1+\lambda_2)/\left\|\hat{\boldsymbol{P}}_{Aj}\right\|_2)_+$ when both j-th main effect $\left\|\boldsymbol{\Psi}_j\hat{\boldsymbol{\beta}}_j\right\|_2^2$ and the residual treatment effect, i.e. treatment effect excluding j-th interaction $\left\|\boldsymbol{\Psi}_A\hat{\boldsymbol{\beta}}_A\right\|_2^2+\sum_{k\neq j}\left\|\boldsymbol{\Phi}_{Ak}\hat{\boldsymbol{\beta}}_{Ak}\right\|_2^2$ are 0. When only one of quantities $\left\|\boldsymbol{\Psi}_j\hat{\boldsymbol{\beta}}_j\right\|_2^2$ and $\left\|\boldsymbol{\Psi}_A\hat{\boldsymbol{\beta}}_A\right\|_2^2+\sum_{k\neq j}\left\|\boldsymbol{\Phi}_{Ak}\hat{\boldsymbol{\beta}}_{Ak}\right\|_2^2$ is zero, let's say $\left\|\boldsymbol{\Psi}_j\hat{\boldsymbol{\beta}}_j\right\|_2^2=0$ without loss of generality, $\mathcal{S}_{Aj}$ can be obtained by solving

$$\mathcal{S}_{Aj}\left\|\hat{\boldsymbol{P}}_{Aj}\right\|_2\left(1+\frac{\lambda_1}{\left\|\boldsymbol{\Psi}_A\hat{\boldsymbol{\beta}}_A\right\|_2^2+\sum_{k=1}^p\left\|\boldsymbol{\Phi}_{Ak}\hat{\boldsymbol{\beta}}_{Ak}\right\|_2^2}\right)=\left(\left\|\hat{\boldsymbol{P}}_{Aj}\right\|_2-\lambda_1-\lambda_2\right)_+, \quad (4.4)$$

Algorithm 3: Sparse Backfitting Algorithm

Result: $\hat{\boldsymbol{\beta}}_j$, $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_{Aj}$

Initialize $\hat{\boldsymbol{\beta}}_j$, $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_{Aj}$ for $j = 1, \cdots, p$;

while *Convergence is False* **do**

  for $j \leftarrow 1$ **to** p **do**

    Compute the residual, $\hat{\boldsymbol{R}}_j = \boldsymbol{Y} - \sum_{k \neq j} \boldsymbol{\Psi}_k \hat{\boldsymbol{\beta}}_k - \boldsymbol{\Psi}_A \hat{\boldsymbol{\beta}}_A - \sum_{l=1}^{p} \boldsymbol{\Phi}_{Al} \hat{\boldsymbol{\beta}}_{Al}$

    Projection $\hat{\boldsymbol{R}}_j$ onto $\boldsymbol{X}_j$, $\hat{\boldsymbol{P}}_j = \mathcal{P}_j \hat{\boldsymbol{R}}_j$

    Soft thresholding: $\hat{\boldsymbol{f}}_j = \mathcal{S}_j \hat{\boldsymbol{P}}_j$, $\hat{\boldsymbol{\beta}}_j = \mathcal{S}_j (\boldsymbol{\Psi}_j^\top \boldsymbol{\Psi}_j)^{-1} \boldsymbol{\Psi}_j^\top \hat{\boldsymbol{R}}_j$

  end

  Compute the residual, $\hat{\boldsymbol{R}}_A = \boldsymbol{Y} - \sum_{l=1}^{p} \boldsymbol{\Psi}_l \hat{\boldsymbol{\beta}}_l - \sum_{l=1}^{p} \boldsymbol{\Phi}_{Al} \hat{\boldsymbol{\beta}}_{Al}$

  Project residuals onto A, $\hat{\boldsymbol{P}}_A = \mathcal{P}_A \hat{\boldsymbol{R}}_A$

  Soft thresholding: $\hat{\boldsymbol{f}}_A = \mathcal{S}_A \hat{\boldsymbol{P}}_A$, $\hat{\boldsymbol{\beta}}_A = \mathcal{S}_A (\boldsymbol{\Psi}_A^\top \boldsymbol{\Psi}_A)^{-1} \boldsymbol{\Psi}_A^\top \hat{\boldsymbol{R}}_A$

  for $j \leftarrow 1$ **to** p **do**

    Compute the residual, $\hat{\boldsymbol{R}}_{Aj} = \boldsymbol{Y} - \sum_{l=1}^{p} \boldsymbol{\Psi}_l \hat{\boldsymbol{\beta}}_l - \boldsymbol{\Psi}_A \hat{\boldsymbol{\beta}}_A - \sum_{k \neq j} \boldsymbol{\Phi}_{Ak} \hat{\boldsymbol{\beta}}_{Ak}$

    Project residuals onto $(\boldsymbol{X}_j, A)$, $\hat{\boldsymbol{P}}_{Aj} = \mathcal{P}_{Aj} \hat{\boldsymbol{R}}_{Aj}$

    Soft thresholding: $\hat{\boldsymbol{f}}_{Aj} = \mathcal{S}_{Aj} \hat{\boldsymbol{P}}_{Aj}$, $\hat{\boldsymbol{\beta}}_{Aj} = \mathcal{S}_{Aj} (\boldsymbol{\Phi}_{Aj}^\top \boldsymbol{\Phi}_{Aj})^{-1} \boldsymbol{\Phi}_{Aj}^\top \hat{\boldsymbol{R}}_{Aj}$

  end

end

and in the symmetric case when only $\left\| \boldsymbol{\Psi}_A \hat{\boldsymbol{\beta}}_A \right\|_2^2 + \sum_{k \neq j} \left\| \boldsymbol{\Phi}_{Ak} \hat{\boldsymbol{\beta}}_{Ak} \right\|_2^2$ equals zero, $\mathcal{S}_{Aj}$ can be solved in a similar way. When both quantities $\left\| \boldsymbol{\Psi}_j \hat{\boldsymbol{\beta}}_j \right\|_2^2$ and $\left\| \boldsymbol{\Psi}_A \hat{\boldsymbol{\beta}}_A \right\|_2^2 + \sum_{k \neq j} \left\| \boldsymbol{\Phi}_{Ak} \hat{\boldsymbol{\beta}}_{Ak} \right\|_2^2$ are nonzero, $\mathcal{S}_{Aj}$ can be solved via

$$\mathcal{S}_{Aj} \left\| \hat{\boldsymbol{P}}_{Aj} \right\|_2 \left(1 + \frac{\lambda_1}{\sqrt{\left\| \boldsymbol{\Psi}_j \hat{\boldsymbol{\beta}}_j \right\|_2^2 + \left\| \boldsymbol{\Phi}_{Aj} \hat{\boldsymbol{\beta}}_{Aj} \right\|_2^2}} + \frac{\lambda_1}{\sqrt{\left\| \boldsymbol{\Psi}_A \hat{\boldsymbol{\beta}}_A \right\|_2^2 + \sum_{k=1}^{p} \left\| \boldsymbol{\Phi}_{Ak} \hat{\boldsymbol{\beta}}_{Ak} \right\|_2^2}} \right) = \left(\left\| \hat{\boldsymbol{P}}_{Aj} \right\|_2 - \lambda_2 \right)_+ .$$

Therefore, the threshold for an interaction term changes according to whether its main effects are already in the model. For an arbitrary j-th interaction, if neither $\boldsymbol{f}_j$ nor A is in the model, the threshold is $2\lambda_1 + \lambda_2$, which is likely to happen at early stage of selection. If one of $\boldsymbol{f}_j$ and A has been selected, this threshold decreases to $\lambda_1 + \lambda_2$ if either main effect is selected, as selecting this interaction only introduces one additional variable instead of two in the previous scenario. When both main effects are already in the model, the threshold further drops to λ_2. This is also intuitive because in this case including such an interaction adds no new predictors to the model, therefore it should be prioritized compared to interactions with one or no main effects selected.

It is crucial to select the tuning parameters. In practice, we search the tuning parameter (λ_1, λ_2) on a 2-dimensional grid. When searching the grid, the solution from the previous grid point is used as a warm start. Moreover, the active set strategy is used to speed up the algorithm, which is based on the fact that the set of nonzero functions (i.e., active set) are often the same or differ by one element when moving to an adjacent point on the fine grid. Thus when moving to the next point on the grid, we first iterate on the current active set until convergence and then sweep all the variables to check for any new selected variables. To determine optimal tuning parameters, multiple criterions such as GCV, BIC have been tested. When the sample size is sufficiently large compared to the number of variables, all popular metrics have similar performance. In cases where $p > n$, we find none of the metrics can dominate others in general. In our simulation study, we used a BIC type criterion (*Gao and Carroll*, 2017):

$$\mathrm{BIC}(\lambda) = n \log\{\mathrm{RSS}(\lambda)/n\} + 6(1+\gamma)\log(p)d(\lambda),$$

where $d(\lambda)$ is the number of variables selected and γ is an arbitrary positive number.

In this chapter, we set γ to be 0.1 in simulation studies and data application.

4.2.3 Extension to Multi-stage Setting

The proposed SpAS can be extended to the scenario with multiple decision points, where the data can come from either sequential multiple assignment randomized trials (SMART) or observational studies. At each decision point, the treatment can have multiple levels or be continuous. Assume the three causal assumptions, consistency, NUCA and positivity hold. Let $\{(\mathbf{X}_t, A_t)_{t=1}^T, Y\}$ denote the data, where $t \in \{1, 2, \cdots, T\}$ denotes t^{th} stage, $\mathbf{X}_t$ denotes the vector of patient characteristics accumulated during treatment period t, A_t denotes the treatment variable with observed value $a_t \in \mathcal{A}_t = \{1, \ldots, K_t\}$, and K_t ($K_t \geq 2$) is the number of treatment options at the t^{th} stage. Therefore patient history prior to A_t can be written as $\mathbf{H}_t = \{(\mathbf{X}_i, A_i)_{i=1}^{t-1}, \mathbf{X}_t\}$. We assume that Y is bounded, and higher values of Y are preferable. A dynamic treatment regime (DTR) is then denoted as $\mathbf{g} = (g_1, \ldots, g_T)$, where at stage t, g_t maps from the domain of patient history $\mathbf{H}_t$ to the domain of treatment assignment A_t.

To perform variable selection across multiple stages, we start from the final stage T and proceed in reverse sequential order. At each stage t, $(1 < t < T)$, we define pseudo-outcome $\widetilde{Y}_t$ recursively using Bellman's optimality:

$$\widetilde{Y}_t = \mathbb{E}\left\{\widetilde{Y}_{t+1} | A_{t+1} = g_{t+1}^{opt}(\mathbf{H}_{t+1}), \mathbf{H}_{t+1}\right\}, t = 1, \cdots, T-1,$$

I.e. the expected outcome assuming optimal regimes are followed at all future stages (*Murphy*, 2005; *Moodie et al.*, 2012). Q-function at stage t can be defined using $\widetilde{Y}_t$: $Q_t(A_t, \mathbf{H}_t) = \mathbb{E}(\widetilde{Y}_t | A_t, \mathbf{H}_t)$. At final stage T, because $\widetilde{Y}_T = Y$, we directly fit additive model for $Q_t(A_t, \mathbf{H}_t)$ and perform variable selection using the proposed SpAS approach. At intermediate stage $t < T$, $\widetilde{Y}_t$ is estimated using the actual observed

outcome, Y, plus the predicted future loss due to sub-optimal future treatments to prevent bias accumulation: $\hat{\widetilde{Y}}_t = Y + \sum_{j=t+1}^{T}[\hat{Q}_j\{\hat{g}_j^{opt}(\mathbf{H}_j), \mathbf{H}_j\} - \hat{Q}_j(A_j, \mathbf{H}_j)]$. Here $\hat{g}_j^{opt}(\mathbf{H}_j)$ refers to the estimated optimal treatment regime at stage j. $\widetilde{Y}_t$ can be estimated either directly using fitted sparse additive model, or using other methods by only including variables selected from SpAS. In our simulations, we estimate $\widetilde{Y}_t$ by using random forests-based conditional mean estimates (*Breiman*, 2001).

4.3 Simulation Study

In this section, we conducted simulation studies to evaluate the performance of the proposed method. In particular, we considered data-generating mechanisms with confounding variables, including four single-stage scenarios and two two-stage scenarios.

4.3.1 Single-stage Scenarios

In the single-stage scenarios, the baseline variables $\boldsymbol{X} = (X_1, \cdots, X_p)$ follow a multivariate normal distribution: for each entry the marginal distribution is $N(0, 1)$, and the correlation structure is AR1 with a correlation coefficient of 0.2 or 0.8. In Scenarios I and II, we considered a binary treatment $A = (0, 1)$, and $A = 1$ when $X_1 + X_2 < 0$ and $X_9 + X_{10} > 0$. In Scenarios III, we considered a three-level treatment $A = (0, 1, 2)$ generated from a multinomial distribution with probabilities $\{\pi_0,\ \pi_1,\ \pi_2\}$, where $\pi_0 = 1/\{1 + \exp(\boldsymbol{X}\boldsymbol{\beta}_1^{\pi}) + \exp(\boldsymbol{X}\boldsymbol{\beta}_2^{\pi})\}$, $\pi_1 = \exp(\boldsymbol{X}\boldsymbol{\beta}_1^{\pi})/\{1 + \exp(\boldsymbol{X}\boldsymbol{\beta}_1^{\pi}) + \exp(\boldsymbol{X}\boldsymbol{\beta}_2^{\pi})\}$ and $\pi_2 = 1 - \pi_0 - \pi_1$, where $\boldsymbol{\beta}_1^{\pi} = (-1, -1, 0, 2, 0, -1, 1, 1, -1, 0, 0, \cdots, 0)$ and $\boldsymbol{\beta}_2^{\pi} = (-1, -1, -1, -1, 0, 1, 0, 0, 1, 2, 0, \cdots, 0)$. In addition, Scenario IV represents a randomized trial setting, and the three treatment options are equally likely to be assigned.

We used the following models to generate outcomes:

- Scenario I: $Y = 1 + \boldsymbol{X}\gamma_1 + \beta_{A1}A + A\boldsymbol{X}\beta_1 + \epsilon$, with $\gamma_1 = (1, 1, \mathbf{0}_6, -1, -1, \mathbf{0}_{p-10})$,

$\beta_{A1} = 0.2$, $\beta_1 = (0, 0, 1, 1, 1, \mathbf{0}_{p-5})$.

- Scenario II: $Y = 1 + \sin(\boldsymbol{X})\gamma_1 + \beta_{A1}A + A\sin(\boldsymbol{X})\beta_1 + \epsilon$.

- Scenario III & IV: $Y = 1 + 0.25(1 + \boldsymbol{X}\gamma_3)^2 + \beta_{A3}I(A = 1) + \beta_{A4}I(A = 2) + (\boldsymbol{X} - 1)^2\beta_{31} \cdot I(A = 1) + \sin(\boldsymbol{X})\beta_{32} \cdot I(A = 2) + \epsilon$, with $\gamma_3 = (1, 1, 0, -1, \mathbf{0}_5, 1, \mathbf{0}_{p-10})$, $\beta_{31} = (1, \mathbf{0}_7, -1, 2, \mathbf{0}_{p-10})$, $\beta_{A3} = \beta_{A4} = 0.1$ and $\beta_{32} = (2, 1, \mathbf{0}_6, 1, -2, \mathbf{0}_{p-10})$.

In all four scenarios, the noise ϵ is normally distributed with mean 0 and variance 0.25. Scenarios I and II have three interaction variables X_3, X_4, X_5, and Scenarios III and IV have four interaction variables X_1, X_2, X_9 and X_{10}. Besides, strong heredity constraint does not fully hold in any scenarios considered above. We replicated the simulation 500 times with sample size $n = 400$ and $p = 1000$.

We compared our proposed method with four competing methods: the S-Score method (*Gunter et al.*, 2011), forward minimal misclassification error rate (ForMMER, *Zhang et al.*, 2018a), sequential advantage selection (SAS, *Fan et al.*, 2016) and penalized A-learning (PAL, *Shi et al.*, 2018). The contrast function in ForMMER was implemented using augmented inverse probability weighted estimator (AIPWE), and the tuning parameter, α was set to be 0.05. Moreover, in order to model the conditional outcome regression for AIPWE, we adopted the same AIC-based forward selection approach as in *Zhang et al.* (2018a). These four methods were proposed for binary treatment, and it is not trivial to generalize to treatments with more than two levels. Thus for illustrative purpose, in Scenarios III and IV, we took a naive one-against-all approach as a workaround: in each case, $A = 1$ vs. (2,3), $A = 2$ vs. (1,3) and $A = 3$ vs. (1,2), a separate selection procedure was executed to identify prescriptive variables. Moreover, we set the final selected variables as the union of prescriptive variables identified in the three regimes.

We evaluated the variable selection performance using two metrics: size and TP (true

positive). Size is the number of selected interaction variables, and TP is the number of selected correct interaction variables. The results of the Scenarios I and II are summarized in Table 4.1 below. In both scenarios, our proposed SpAS has stable and outstanding performance in terms of reasonable sizes and high TPs. S-Score tends to be conservative in terms that both methods select a small number of interactions. In Scenario I where the correct underlying model is linear, ForMMER performs well in terms of massive TPs. In particular, when baseline variables are weakly correlated, ForMMER can correctly identify all the interactions in every replication. The reason is that the AIPW estimator in ForMMER is correctly specified and therefore, the estimated contrasts approximate the truth well. Similar to SpAS, SAS also involves fitting conditional outcome regression models; however, SAS has much lower TPs compared to SpAS. The reason is that SAS fits a sequence of regression models by including one covariate at a time. Compared to the simultaneous model fitting and variable selection in SpAS, this approach is more vulnerable to confounding effects that are yet to be included. Also, highly correlated baseline covariates do not affect the performance of SpAS. In Scenario II, the data generating model is non-linear, and the working models in all competing methods are misspecified. This has a detrimental effect on their performance. On the other hand, SpAS is still able to deliver reliable selection results.

Table 4.2 summarizes results from Scenarios III and IV, where the outcome models are non-linear. Note the competing methods are not directly applicable and are included for illustrative purpose. Therefore we only demonstrated the comparison when baseline covariates are weakly correlated. In both scenarios, SAS, ForMMER, and S-Score all tend to over-select the interaction variables. The sizes selected by PAL differ a lot between the two scenarios. It is worth noting that in Scenario IV, where the treatments are randomly assigned, the competing methods have undesirable results in terms of excessively large sizes and small TPs using the one-against-all workaround.

Table 4.1: Simulation results for single-stage Scenarios I and II based on 500 replications. Size: number of interactions selected; TP: number of true interactions selected.

	Weak Correlation ($\rho = 0.2$)		Strong Correlation ($\rho = 0.8$)	
Methods	Size	TP	Size	TP
		Scenario I		
S-Score	1.03 (0.45)	0.84 (0.49)	1.28 (0.52)	0.69 (0.53)
ForMMER	8.72 (1.81)	3.00 (0.00)	8.67 (2.02)	2.82 (0.44)
SAS	5.09 (3.29)	2.13 (1.36)	4.03 (6.88)	1.55 (0.55)
PAL	1.44 (0.57)	1.43 (0.58)	1.10 (0.30)	1.09 (0.30)
SpAS	4.35 (0.55)	2.95 (0.33)	4.36 (0.72)	2.88 (0.34)
		Scenario II		
S-Score	0.87 (0.67)	0.56 (0.56)	1.17 (0.40)	0.81 (0.42)
ForMMER	2.30 (2.79)	1.21 (0.90)	4.88 (2.76)	1.24 (0.56)
SAS	0.41 (3.26)	0.13 (0.58)	3.50 (7.59)	1.39 (0.52)
PAL	1.36 (0.84)	1.11 (0.49)	1.35 (0.82)	0.97 (0.45)
SpAS	4.12 (0.72)	2.86 (0.64)	4.26 (0.65)	2.81 (0.62)

Even though in this scenario, the propensity model is guaranteed to be consistent, the compromised performance of PAL and ForMMER could be due to the failure to well approximate propensity scores with a large number of covariates. The proposed SpAS results in reasonable sizes and decent TPs in both scenarios, without fitting the model multiple times.

4.3.2 Two-stage Scenarios

We illustrated the multi-stage decision making performance of the proposed method with two-stage Scenarios V and VI. The two scenarios have confounding variables at both stages. The baseline variables $\boldsymbol{X}_1 = (X_{1,1}, \cdots, X_{1,p_1})$ were generated from the same multivariate normal distribution as in Section 4.3.1. To get meaningful comparison results, we considered binary treatments at both stages in Scenarios V and VI. The first stage treatment $A_1 = 1$ if $X_{1,1} + X_{1,2} > 1$ or $X_{1,9} + X_{1,10} < 0$. The intermediate covariates collected at stage two is denoted as $\boldsymbol{X}_2 = (X_{2,1}, \cdots, X_{2,p_2})$.

Table 4.2: Simulation results for single-stage Scenarios III and IV based on 500 replications with $\rho = 0.2$. The methods S-Score, ForMMER, SAS and PAL are implemented using one-versus-all approach. Size: number of interactions selected; TP: number of true interactions selected.

	size	TP
	Scenario III	
S-Score	51.2 (38.63)	2.24 (0.53)
ForMMER	13.74 (2.91)	3.33 (0.70)
SAS	21.22 (5.94)	3.96 (0.20)
PAL	4.72 (2.50)	2.39 (0.57)
SpAS	5.15 (1.57)	3.75 (0.41)
	Scenario IV	
S-Score	19.12 (42.99)	2.00 (0.54)
ForMMER	11.85 (2.40)	2.70 (0.80)
SAS	45.15 (6.15)	3.35 (0.54)
PAL	15.00 (5.43)	2.95 (0.62)
SpAS	8.37 (2.67)	3.70 (0.80)

We let $p_2 = 5$ and $\boldsymbol{X}_2 = 0.5 \cdot (X_{1,11}, \cdots, X_{1,15}) + \epsilon_2$, where $\epsilon_2 \sim N(0, 0.25)$. The stage two treatment $A_2 = 1$ if $X_{2,3} + X_{2,4} < 0$ or $X_{2,1} + X_{2,2} < -1$.

We generated the outcomes observed at the end of stage two from the following mechanisms:

- Scenario V: $Y = \boldsymbol{X}_1\gamma_{5,1} + \boldsymbol{X}_2\gamma_{5,2} + \beta_{A_1A_2}A_1A_2 + A_2(a_2 + \boldsymbol{X}_2\beta_{5,2}) + A_1(a_1 + \boldsymbol{X}_1\beta_{5,1}) + \epsilon$, with $\gamma_{5,1} = (1, \mathbf{0}_{p_1-1})$, $\gamma_{5,2} = (1, \mathbf{0}_{p_2-1})$, $\beta_{A_1A_2} = 4$, $a_1 = 1$, $a_2 = 0.5$, $\beta_{5,1} = (\mathbf{1.5}_5, \mathbf{0}_{p_1-5})$, $\beta_{5,2} = \mathbf{-2}_{p_2}$.

- Scenario VI: $Y = \sin(\boldsymbol{X}_1)\gamma_{5,1} + \sin(\boldsymbol{X}_2)\gamma_{5,2} + \beta_{A_1A_2}A_1A_2 + A_2\{a_2 + \cos(\boldsymbol{X}_2)\beta_{5,2}\} + A_1\{a_1 + \sin(\boldsymbol{X}_1)\beta_{5,1}\} + \epsilon$.

The noise ϵ is normally distributed with mean 0 and variance 0.25. In Scenario V, it is easy to see the optimal regime at the second stage is $g_2^{opt}(\boldsymbol{X}_1, A_1, \boldsymbol{X}_2) = I(\beta_{A_1A_2}A_1 + a_2 + \boldsymbol{X}_2\beta_{5,2} > 0)$ and therefore is determined by six variables $(A_1, \boldsymbol{X}_2)$.

At stage one, the Q-function is

$$\begin{aligned} Q_1(\boldsymbol{X}_1, A_1) =& \mathbb{E}\{\boldsymbol{X}_2\gamma_{5,2} + A_2^{opt}(\beta_{A_1A_2}A_1 + a_2 + \boldsymbol{X}_2\beta_{5,2})|\boldsymbol{X}_1, A_1\} + \boldsymbol{X}_1\gamma_{5,1} \\ &+ A_1(a_1 + \boldsymbol{X}_1\beta_{5,1}) \\ =& \mathbb{E}\{(\beta_{A_1A_2}A_1 + a_2 + \boldsymbol{X}_2\beta_{5,2})_+|\boldsymbol{X}_1, A_1\} + \Omega(\boldsymbol{X}_1, A_1) \\ =& \sigma_2\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\mu_1^2}{2\sigma_2^2}\right\} + \mu_1[1 - \Phi(-\frac{\mu_1}{\sigma_2})] + \Omega(\boldsymbol{X}_1, A_1) \end{aligned} \tag{4.5}$$

where $\Omega(\boldsymbol{X}_1, A_1) = \mathbb{E}(\boldsymbol{X}_2\gamma_{5,2}|\boldsymbol{X}_1, A_1)+\boldsymbol{X}_1\gamma_{5,1}+A_1(a_1+\boldsymbol{X}_1\beta_{5,1})$, and $\mu_1 = \beta_{A_1A_2}A_1+ a_2 + 0.5 \cdot (X_{1,11}, \cdots, X_{1,15})$. The optimal regime at stage 1 is then $g_1^{opt}(\boldsymbol{X}_1) = I\{Q_1(\boldsymbol{X}_1, 1) > Q_1(\boldsymbol{X}_1, 0)\}$, and involves ten important variables $(X_1, \cdots, X_5, X_{11}, \cdots, X_{15})$. Note the interactions between A_1 and $(X_{1,11}, \cdots, X_{1,15})$ is induced by μ_1, i.e., by assuming to follow optimal regime at stage 2. Hereafter we refer to these variables as Q-interactions, which stand for interactions that arise from backward induction. Following similar argument, Scenario VI has the same important interactions at both stages.

Each Scenario was simulated 500 times with sample size $n = 400$ and $p_1 = 500$. Because S-Score was only proposed in the single-stage scenario, therefore we compared the proposed method with ForMMER, SAS, and PAL. The results are summarized in Table 4.3. In Scenario V, ForMMER has decent performance at stage 2 when baseline covariates are weakly correlated; increasing correlation coefficient will result in a drop of TP. Similar to the single-stage scenario, PAL tends to be conservative and under-selects variables. At stage 1, all competing methods perform poorly, and none of them (except for ForMMER in Scenario V, $\rho = 0.2$) have TPs fairly close to the truth. This is due to the small magnitude of Q-interactions in the stage 1 Q-function (4.5), i.e., interactions between A_1 and $(X_{1,11}, \cdots, X_{1,15})$. As an empirical verification, the column TP(μ_1) in Table 4.3 suggests the competing methods are highly unlikely to identify any Q-interactions. SpAS has an outstanding performance

Table 4.3: Simulation Results in Two Stage Scenarios V and VI. Size: number of interactions selected; TP: number of true interactions selected; TP (μ_1): number of true interactions between A_1 and $(X_{1,11}, \cdots, X_{1,15})$ selected.

	Stage 2		Stage 1		
	Size	TP	Size	TP	TP (μ_1)
Scenario V, $\rho = 0.2$					
ForMMER	9.26 (2.08)	4.11 (1.16)	5.16 (3.70)	2.69 (1.59)	0.12 (0.35)
SAS	5.14 (1.67)	2.96 (1.08)	14.51 (3.37)	5.05 (0.67)	0.28 (0.48)
PAL	5.48 (3.18)	1.68 (0.92)	6.20 (2.49)	3.58 (0.98)	0.06 (0.23)
SpAS	8.48 (1.28)	5.65 (1.01)	10.80 (1.55)	7.75 (1.76)	2.85 (0.83)
Scenario VI, $\rho = 0.2$					
ForMMER	2.86 (0.91)	1.19 (0.64)	9.12 (1.65)	4.84 (0.58)	0.08 (0.27)
SAS	0.00 (0.00)	0.00 (0.00)	9.73 (11.88)	0.23 (0.63)	0.09 (0.32)
PAL	0.00 (0.00)	0.00 (0.00)	8.87 (3.44)	2.22 (0.98)	0.07 (0.26)
SpAS	7.95 (1.28)	5.79 (0.96)	11.21 (2.44)	8.01 (1.53)	2.87 (1.06)
Scenario V, $\rho = 0.8$					
ForMMER	7.71 (1.73)	3.06 (0.62)	6.79 (1.73)	2.53 (0.87)	0.38 (0.51)
SAS	4.32 (1.07)	1.90 (0.46)	4.73 (1.46)	3.62 (0.74)	0.91 (0.43)
PAL	3.93 (2.85)	0.73 (0.70)	2.74 (0.75)	2.68 (0.70)	0.00 (0.00)
SpAS	8.68 (1.66)	5.70 (1.34)	12.37 (3.44)	7.66 (1.28)	2.69 (1.16)
Scenario VI, $\rho = 0.8$					
ForMMER	4.27 (1.38)	1.63 (0.57)	7.65 (1.79)	2.90 (0.80)	0.11 (0.33)
SAS	0.09 (0.34)	0.07 (0.25)	8.19 (11.78)	0.45 (0.95)	0.14 (0.38)
PAL	0.00 (0.00)	0.00 (0.00)	3.90 (1.94)	2.20 (0.79)	0.02 (0.19)
SpAS	8.27 (1.32)	5.86 (1.09)	12.87 (3.44)	7.82 (1.31)	2.63 (1.21)

in stage 2. It also has much better TP and TP(μ_1) compared to other methods. Also, SpAS tends to result in slightly larger sizes.

In Scenario VI, similar trends are observed. SpAS has stable performance when the data generating a model is non-linear, while all competing methods perform poorly even in stage 2. In stage 1, the three competing methods perform better than they do in stage 2. This could be due to the larger magnitude of interactions between A_1 and $(X_1, \cdots, X_5)$. TP(μ_1) indicates the competing methods still fail to select the Q-interactions.

4.4 Application

We applied the proposed method to a hepatocellular carcinoma (HCC) dataset of 227 patients collected at the VA Ann Arbor Healthcare System between January 2006 and December 2013. The median follows up for this cohort is 353 days. Patient pre-treatment objective clinical/tumor information was collected and summarized in Table 4.4 below. Besides, 2540 body factor biomarkers were calculated using analytic morphomics technique from pretreatment CT studies to assess patient body composition, such as body dimensions, visceral fat, and muscle mass. Patients were excluded if they lacked CT imaging before HCC-directed treatment, or had technical issues with CT imaging precluding analytic morphomics.

Table 4.4: Patient Demographics in HCC Study

Age: Median (IQR)	61 (57, 66)
Race (% Caucasian)	95 (41.9%)
Etiology	
Hepatitis C	167 (73.6%)
Alcohol-induced	16 (7.0%)
NASH/cryptogenic	12 (5.3%)
Multifocal HCC	110 (48.5%)
Child pugh class	
A	130 (57.3)
B	70 (30.8)
C	27 (11.9)
MELD Score	9.0 (8.0, 12.0)
ECOG performance status	1.0 (0.0, 2.0)
TNM stage (I/II/III/IV)	97/49/59/22
Treatment	
Resection	25 (11.1%)
TACE	83 (36.7%)
Other	118 (52.2%)

In this application, we considered three types of intervention: resection, transarterial chemoembolization (TACE), and others. Besides, there are one-third censored observations in this dataset. For illustration purpose, we imputed these censored

observations using a one-step approach as part of recursively imputed survival tree algorithm (RIST, *Zhu and Kosorok*, 2012). Missing covariates were imputed using hot deck imputation (*Little and Rubin*, 2019).

It is of interest to learn an optimal treatment rule which maximizes expected survival time by combining routine clinical information with the analytic morphomics biomarkers; however, the limited sample size, high-dimensionality, and three-level treatment impose challenges, and none of existing methods mentioned before can be applied. Thus it is more realistic to conduct early-stage exploratory analyses; one option is to identify variables that might determine optimal treatment strategy. Therefore we analyzed this data using the proposed SpAS method to identify the predictive and potentially prescriptive variables. We searched for tuning parameters over a two-dimensional grid. We identified 18 main effects and 17 interaction variables. The selected variables are summarized in Table 4.5 below. The presence of a multifocal tumor is the only prognostic variable that was not identified as a treatment effect modifier. Besides, all clinical factors and one morphomics biomarker representing muscle measurement have been reported to be prognostic (*Singal et al.*, 2016; *Parikh et al.*, 2018). Therefore, 15 new markers were identified as potential tailoring variables, and many of them are measurements for body dimension. In addition, we further fixed the ratio between the two tuning parameters, λ_1 and λ_2 to 1.5 and calculated the regularization path, as shown in Figure 4.4. Regularization path plots can provide insight into the relative variable importance, and have important applications in medical science. For example, sometimes, it is preferred to reduce the number of variables being considered or collected to a certain level due to budget concern. In such cases, the regularization path can help the investigators determine how many and what variables to include in further research.

Table 4.5: The variable selection results for HCC data.

Variable name	Type of measurement	Interaction	Previously reported	Description
SPLEENMINBBOXZ BY VISCERALFATAREA	Organ	Yes	No	The spleen size divided by the area of fat-intensity pixels in the visceral cavity
VISCERALFATHU BY MEANPORTALVEINHU	Fat	Yes	No	Median fat pixel intensity inside the visceral cavity divided by average HU value on the portal vein
FASCIACIRCUMFERENCE BY VBSLABHEIGHT	Body dimension	Yes	No	Circumference of fascia perimeter divided by Height of the body slab for this vertebra
PSPVOLOFVB BY VB2FASCIA	Muscle	Yes	Yes	Volume of the dorsal muscle group divided by the distance between the vertebra to the facial envelope
PSPVOLOFVB BY TOTALBODYAREA	Muscle	Yes	No	Volume of the dorsal muscle group divided by total body area
VISCERALFATHU BY SUBCUTFATHU	Fat	Yes	No	The ratio of median fat pixel intensities between visceral cavity and subcutaneous region
VISCERALFATHU NORMALIZED	Fat	Yes	No	Normalized median fat pixel intensity inside the visceral cavity
DIST INFANTPT2SUPANTPT	Body dimension	Yes	No	Height of the vertebral body at anterior aspect
VBSLABHEIGHT	Body dimension	Yes	No	Height of the body slab for this vertebra
VB2FRONTSKIN BY FASCIACIRCUMFERENCE	Body dimension	Yes	No	Distance from the vertebral body to the front skin divided by circumference of fascia perimeter
SPLEENMINBBOXZ BY BODYWIDTH	Organ	Yes	No	The spleen size divided by the body width
VB2FRONTSKIN BY FASCIAAREA	Body dimension	Yes	No	Distance from the vertebral body to the front skin divided by area of the visceral cavity
FASCIAAREA BY TOTALBODYCIRCUMFERENCE	Body dimension	Yes	No	Area of the visceral cavity divided by total body circumference
SPLEENMINBBOXZ BY VBSLABHEIGHT	Organ	Yes	No	The spleen size divided by the height of the body slab for this vertebra
Multifocal	Clinical factor	No	Yes	Presence of multifocal tumor
Albumin	Clinical factor	Yes	Yes	Albumin
Child pugh class	Clinical factor	Yes	Yes	Child pugh class
TNM Stage	Clinical factor	Yes	Yes	TNM stage

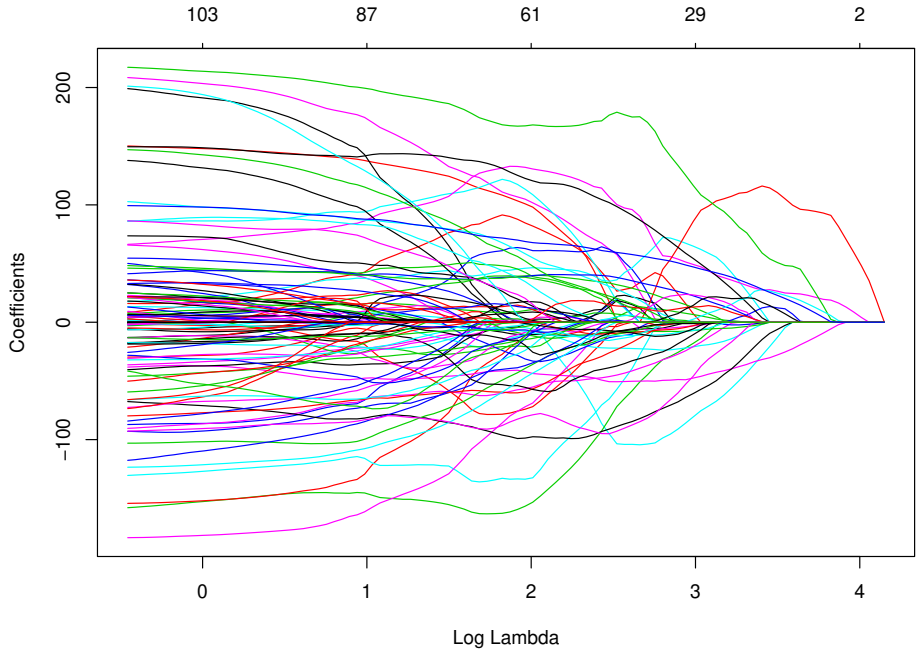

Figure 4.1: The regularization path calculated for HCC data. The ratio between the two tuning parameters is fixed at $\lambda_1/\lambda_2 = 1.5$.

4.5 Discussion

Estimating optimal dynamic treatment regimes with large observational data has recently started to draw attention in the statistics community. However, most existing variable selection methods in estimating optimal DTRs, if not all, can only allow randomized trial data and binary treatments. In this chapter, we proposed SpAS for identifying predictive and potentially prescriptive variables in multi-stage, multi-treatment settings using observational data. At each stage, we fit a sparse additive model which requires little effort for model specification. Furthermore, the proposed method improves the interpretability and plausibility of fitted models by enforcing strong heredity constraint, i.e., an interaction can only be included if both corresponding main effects have already entered the model. Besides, the proposed method explicitly identifies predictive variables. As pointed out by *Shortreed and Ertefaie* (2017), these variables can be used to refine propensity score models to account for

confounding bias while maintaining statistical efficiency. Thus the selected predictive variables can improve the quality of estimated dynamic treatment regime.

Qian and Murphy (2011) studied the mismatch between minimizing squared error loss when fitting outcome regression models and maximizing the mean counterfactual outcome when learning the optimal treatment regimes. In response, recently, there has been a surge of direct methods for estimating optimal DTRs. Likewise, in this area, all existing variable selection methods directly target at optimizing the mean counterfactual outcome based on models for contrast in outcome regression between two treatment levels. Such a strategy does not solve the challenge of how to construct working models, especially when $p \gg n$. Nevertheless, these methods are restricted to binary treatment by their nature. On the contrary, our proposed method is appealing in terms that it allows multi-level treatments and continuous doses, as well as observational data.

It is still worth noting that SpAS selects treatment effect modifiers, which are not necessarily tailoring variables. Therefore the proposed method would result in a larger set of variables. However, once the number of selected variables becomes manageable, domain experts can help further narrow down the variable list to improve the quality of estimated DTRs and/or to achieve cost-effectiveness. In addition, choosing the optimal tuning parameter is crucial. The proposed method uses the high-dimensional BIC for model tuning; however, this criterion was not specifically proposed for our purpose. It is of both theoretical and practical interest to further study the performance of available metrics or to develop new metrics for parameter tuning in variable selection for estimating optimal DTRs.

CHAPTER V

Summary and Future Work

In this dissertation, we have explored some flexible modeling strategies for coarsened data problems including nonparametric kernel regression when the outcome is missing at random (MAR) and estimating interpretable optimal dynamic treatment regimes using observational data. The overarching goal is to develop flexible, easy-to-use statistical methods while improving model robustness.

The MRKEE method proposed in Chapter II is an essential addition to the literature of multiple robustness. It achieves consistency when any one of missing mechanism or conditional outcome regression models is known or can be correctly specified, thus provides more protection against working model misspecification. MRKEE also has great potential in applications such as flexible dose-response modeling when data are subject to missingness, which is often the case in radiation oncology. Chapter III and IV studied the role of nonparametric machine learning methods in estimating optimal dynamic treatment regimes using observational data. The use of such powerful tools further relaxes model assumptions and requires minimal guesswork in model specification. The ST-RL method proposed in Chapter III estimates optimal DTRs as a sequence of decision trees, one per stage, and thus are interpretable to clinicians and human experts. It also scales well to a moderately large number of covariates. Chapter III contributes to theory development regarding the finite sample bounds

when using non-greedy tree search algorithm for estimating optimal DTRs. The variable selection method proposed in Chapter IV can identify potential predictive and prescriptive variables when more than two treatment options are available using observational data.

Some extensions can further enhance the versatility of nonparametric regression in the presence of missing data. First of all, it is of interest to extend MRKEE to allow multiple predictors as this is often the case in practice. It is also vital to extend MRKEE to accommodate other common types of outcomes, such as binary, time-to-event, and longitudinal data.

One important future research direction of ST-RL is to allow continuous treatment, such as radiation doses. Although there has been some exploration of optimal dosing strategy (e.g., *Laber and Zhao*, 2015; *Chen et al.*, 2016), existing methods are computationally demanding and lacking satisfactory empirical performance. Therefore it is still of great interest to develop dynamic treatment regimes for multi-stage dose optimization. Furthermore, the time-to-event outcome is often of interest in clinical studies. The statistics community only recently started to focus on optimal DTR estimation using survival data (e.g., *Jiang et al.*, 2017; *Hager et al.*, 2018; *Simoneau et al.*, 2019). Extending our tree-based method to survival outcome can greatly improve the interpretability of estimated DTRs.

In high-dimensional setting, especially $p \gg n$, we explored the variable selection method in Chapter V to facilitate the construction of optimal DTRs. Similar to ST-RL, the extension to accommodate survival outcome is also of great importance. Another important research direction is to further relax the additivity assumption by considering a more flexible regression framework.

APPENDICES

APPENDIX A

Proofs for Chapter II

Lemma 1.1. *When $\pi^1(\boldsymbol{\nu}^1)$ is the correctly specified model for $Pr(R=1|Z,\boldsymbol{U})$ and $\boldsymbol{\nu}_0^1$ is the true parameter value, we have*

$$\sqrt{nh}\hat{\boldsymbol{\lambda}} \;=\; (nh)^{1/2}\boldsymbol{M}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i-\pi_i^1(\boldsymbol{\nu}_0^1)}{\pi_i^1(\boldsymbol{\nu}_0^1)}\boldsymbol{g}_i(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\right\}+o_p(1).$$

where $\boldsymbol{M}$ is given by (2.12) and $\boldsymbol{g}(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)$ is given by (2.11).

Proof of Lemma. We first show that the asymptotic distribution of $\hat{\boldsymbol{\lambda}}$ stays the same whether π_0^1 is known or can be estimated consistently at $\sqrt{n}$-rate, i.e. $\sqrt{n}(\hat{\boldsymbol{\nu}}^1-\boldsymbol{\nu}_0^1)=O_p(1)$. Suppose, under some regularity conditions, $\partial\hat{\theta}_{MR}\{z;\pi(\boldsymbol{\nu}^1)\}/\partial\boldsymbol{\nu}^{1\mathrm{T}}$ is bounded in a neighborhood of $\boldsymbol{\nu}_0^1$, *i.e.*,

$$\partial\hat{\theta}_{MR}\{z;\pi(\boldsymbol{\nu}^1)\}/\partial\boldsymbol{\nu}^{1\mathrm{T}}|_{\boldsymbol{\nu}^1\in\mathcal{N}(\boldsymbol{\nu}_0^1)}=O_p(1),$$

where $\mathcal{N}(\boldsymbol{\nu}_0^1) \supset \{\boldsymbol{\nu}^{\mathbf{1}} : ||\boldsymbol{\nu}^1 - \boldsymbol{\nu}_0^1|| < ||\hat{\boldsymbol{\nu}}^1 - \boldsymbol{\nu}_0^1||\}$. We have

$$
\begin{aligned}
&\sqrt{nh}[\hat{\theta}_{MR}\{z; \pi(\hat{\boldsymbol{\nu}}^1)\} - \theta(z)] \\
&= \sqrt{nh}[\hat{\theta}_{MR}\{z; \pi(\hat{\boldsymbol{\nu}}^1)\} - \hat{\theta}_{MR}\{z; \pi(\boldsymbol{\nu}_0^1)\}] + \sqrt{nh}[\hat{\theta}_{MR}\{z; \pi(\boldsymbol{\nu}_0^1)\} - \theta(z)] \\
&= \sqrt{h}\left[\frac{\partial \hat{\theta}_{MR}\{z; \pi(\boldsymbol{\nu}^1)\}}{\partial \boldsymbol{\nu}^{1\mathrm{T}}}|_{\boldsymbol{\nu}^{1*}}\right] \sqrt{n}(\hat{\boldsymbol{\nu}}^1 - \boldsymbol{\nu}_0^1) + \sqrt{nh}[\hat{\theta}_{MR}\{z; \pi(\boldsymbol{\nu}_0^1)\} - \theta(z)] \qquad \text{(A.1)}
\end{aligned}
$$

for some $\boldsymbol{\nu}^{1*} \in \{\boldsymbol{\nu}^1 : ||\boldsymbol{\nu}^1 - \boldsymbol{\nu}_0^1|| < ||\hat{\boldsymbol{\nu}}^1 - \boldsymbol{\nu}_0^1||\}$. Note $\sqrt{n}(\hat{\boldsymbol{\nu}}^1 - \boldsymbol{\nu}_0^1) = O_p(1)$, $\partial \hat{\theta}_{MR}\{z; \pi(\boldsymbol{\nu}^1)\} / \partial \boldsymbol{\nu}^{1\mathrm{T}}|_{\boldsymbol{\nu}^{1*}} = O_p(1)$, and $h \to 0$ as $n \to \infty$, the first term in (A.1) is $o_p(1)$. Therefore, the asymptotic distribution of $\hat{\theta}_{MR}\{z; \pi(\hat{\boldsymbol{\nu}}^1)\}$ when $\boldsymbol{\nu}^1$ is estimated consistently at $\sqrt{n}$-rate is the same as that of $\hat{\theta}_{MR}(z; \pi_0)$ when π_0 is known.

Similar argument shows that the asymptotic results remain the same if $(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\gamma}})$ is replaced with its probability limit $(\boldsymbol{\nu}_*, \boldsymbol{\gamma}_*)$. Taking Taylor expansion of the left-hand side of (2.9) around $(\mathbf{0}^{\mathrm{T}}, \boldsymbol{\alpha}_*^{\mathrm{T}})$ leads to

$$
\begin{aligned}
\mathbf{0} &= (nh)^{1/2} \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \hat{\boldsymbol{g}}_i(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) - \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)^{\otimes 2}}{\pi_i^1(\boldsymbol{\nu}_0^1)} \right\} (nh)^{1/2} \hat{\boldsymbol{\lambda}} \\
&\quad + \sum_{k=1}^{K} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \begin{bmatrix} \mathbf{0}_{\{J+2(k-1)\}\times 2} \\ \frac{\partial \boldsymbol{\psi}_i^k(\boldsymbol{\alpha}_*^k, \boldsymbol{\gamma}_*^k)}{\partial \boldsymbol{\alpha}^k} - \frac{1}{n} \sum_{h=1}^{n} \frac{\partial \boldsymbol{\psi}_h^k(\boldsymbol{\alpha}_*^k, \boldsymbol{\gamma}_*^k)}{\partial \boldsymbol{\alpha}^k} \\ \mathbf{0}_{\{2(K-k)\}\times 2} \end{bmatrix} \right) (nh)^{1/2} (\hat{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}_*^k) \\
&= (nh)^{1/2} \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \hat{\boldsymbol{g}}_i(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) - (nh)^{1/2} \boldsymbol{M} \hat{\boldsymbol{\lambda}} + o_p(1).
\end{aligned}
$$

On the other hand, it is easy to check that

$$
(nh)^{1/2} \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \hat{\boldsymbol{g}}_i(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) = (nh)^{1/2} \frac{1}{n} \sum_{i=1}^{n} \frac{R_i - \pi_i^1(\boldsymbol{\nu}_0^1)}{\pi_i^1(\boldsymbol{\nu}_0^1)} \boldsymbol{g}_i(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1).
$$

Then, solving for $\hat{\boldsymbol{\lambda}}$ from the above Taylor expansion gives the result. □

Proof of Theorem 2. From the previous proof, we can replace $(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\gamma}})$ estimated at $\sqrt{n}$-rate by its probability limit $(\boldsymbol{\nu}_*, \boldsymbol{\gamma}_*)$ without affecting the asymptotic distribution.

Since $\hat{\boldsymbol{\alpha}}_{\mathrm{MR}}$ satisfies

$$\mathbf{0} = \sum_{i=1}^{m} \hat{w}_i \boldsymbol{\phi}_i(\hat{\boldsymbol{\alpha}}_{\mathrm{MR}}) = \frac{1}{m} \sum_{i=1}^{n} \frac{R_i \Pi^1(\hat{\boldsymbol{\nu}}^1)/\pi_i^1(\hat{\boldsymbol{\nu}}^1)}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\pi_i^1(\hat{\boldsymbol{\nu}}^1)} \boldsymbol{\phi}_i(\hat{\boldsymbol{\alpha}}_{\mathrm{MR}}),$$

we have

$$\begin{aligned}
\mathbf{0} &= \frac{1}{m} \left\{ \frac{1}{n} \sum_{h=1}^{n} \pi_h^1(\boldsymbol{\nu}_0^1) \right\} (nh)^{1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) \\
&\quad - \frac{1}{m} \left\{ \frac{1}{n} \sum_{h=1}^{n} \pi_h^1(\boldsymbol{\nu}_0^1) \right\} \left\{ \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi_i^1(\boldsymbol{\nu}_0^1)} \right\} (nh)^{1/2} \hat{\boldsymbol{\lambda}} \\
&\quad + \frac{1}{m} \left\{ \frac{1}{n} \sum_{h=1}^{n} \pi_h^1(\boldsymbol{\nu}_0^1) \right\} \left\{ \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \frac{\partial \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \right\} (nh)^{1/2} (\hat{\boldsymbol{\alpha}}_{\mathrm{MR}} - \boldsymbol{\alpha}_0) + o_p(1) \\
&= \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)} \sqrt{h} \boldsymbol{\phi}_i(\boldsymbol{\alpha}_0) - \sqrt{n} \boldsymbol{L} \hat{\boldsymbol{\lambda}} + \mathbb{E} \left\{ \frac{\partial \boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \right\} (nh)^{1/2} (\hat{\boldsymbol{\alpha}}_{\mathrm{MR}} - \boldsymbol{\alpha}_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Q}_i(z) + \mathbb{E} \left\{ \frac{\partial \boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \right\} (nh)^{1/2} (\hat{\boldsymbol{\alpha}}_{\mathrm{MR}} - \boldsymbol{\alpha}_0) + o_p(1).
\end{aligned}$$

We suppress the denpendence of $\boldsymbol{Q}_i(z)$ on z and denote it as $\boldsymbol{Q}_i$. We then have

$$\mathbf{0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Q}_i + \mathbb{E} \left\{ \frac{\partial \boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \right\} (nh)^{1/2} (\hat{\boldsymbol{\alpha}}_{\mathrm{MR}} - \boldsymbol{\alpha}_0) + o_p(1).$$

Solving for $\sqrt{nh}(\hat{\boldsymbol{\alpha}}_{\mathrm{MR}} - \boldsymbol{\alpha}_0)$ leads to

$$\sqrt{nh}\{\hat{\boldsymbol{\alpha}}_{\mathrm{MR}} - \boldsymbol{\alpha}_0\} = - \left[\mathbb{E} \left\{ \frac{\partial \boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \right\} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Q}_i,$$

which gives $W_{\mathrm{MR},\pi}(z)$.

The bias term can be derived as follows. We expand the derivative term:

$$\begin{aligned}\mathbb{E}\left\{\frac{\partial \boldsymbol{\phi}(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}}\right\} &= \mathbb{E}\left[K_h(Z-z)\left\{\mu^{(1)}(z,\boldsymbol{\alpha}_0)\right\}^2 V^{-1}(z,\boldsymbol{\alpha}_0)\boldsymbol{G}(Z-z)\boldsymbol{G}(Z-z)^T\right]+o_p(1)\\ &= -f_Z(z)\left(\mu^{(1)}\{\theta(z)\}\right)^2 V^{-1}\{\theta(z)\}\boldsymbol{D}(K)+o_p(1)\end{aligned}$$

where $\boldsymbol{D}(K)$ is a 2×2 matrix with the (j,k)th element $c_{j+k-2}(K)\times h^{(j+k-2)}$, and $c_r(K)=\int s^r K(s)ds$.

Moreover, we rewrite $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{Q}_i = \sqrt{nh}\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{Q}_i/\sqrt{h} = \sqrt{nh}\cdot(\boldsymbol{Q}_{1n}+\boldsymbol{Q}_{2n}-\boldsymbol{Q}_{3n})$, where

$$\begin{aligned}\boldsymbol{Q}_{1n} &= n^{-1}\sum_{i=1}^n \frac{R_i}{\pi^1_{i0}}K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha}_0)\boldsymbol{G}(Z_i-z)\left[Y_i-\mu\{\theta(Z_i)\}\right]\\ \boldsymbol{Q}_{2n} &= n^{-1}\sum_{i=1}^n \frac{R_i}{\pi^1_{i0}}K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha}_0)\boldsymbol{G}(Z_i-z)\left[\mu\{\theta(Z_i)\}-\mu\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}_0\}\right]\\ \boldsymbol{Q}_{3n} &= n^{-1}\sum_{i=1}^n \frac{R_i-\pi^1_{i0}}{\pi^1_{i0}}\boldsymbol{L}\boldsymbol{M}^{-1}\boldsymbol{g}_i(\boldsymbol{\nu}_*,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*).\end{aligned}$$

It is easily seen that when $\pi^1(\boldsymbol{\nu}^1)$ is correctly specified, $\boldsymbol{Q}_{1n}$ and $\boldsymbol{Q}_{3n}$ are asymptotically normal with mean zero. Therefore, $\boldsymbol{Q}_{2n}$ is the leading bias term, and under MAR we have $bias\{\boldsymbol{Q}_{2n}\}=$

$$\begin{aligned}&\mathbb{E}\left\{K_h(Z-z)\mu^{(1)}(z,\boldsymbol{\alpha}_0)V^{-1}(z,\boldsymbol{\alpha}_0)\left[\mu\left\{\theta(Z)\right\}-\mu\left\{\boldsymbol{G}(Z-z)^T\boldsymbol{\alpha}_0\right\}\right]\boldsymbol{G}(Z-z)\right\}+o_p(1)\\ &\qquad=\frac{1}{2}\theta''(z)\left[\mu^{(1)}\{\theta(z)\}\right]^2 V^{-1}\{\theta(z)\}f_Z(z)\boldsymbol{H}(K)+o(h^2),\end{aligned}$$

where $\boldsymbol{H}(K)$ is a 2×1 vector with the kth element $c_{k+1}(K)\times h^{(k+1)}$. Note that the variance of $\boldsymbol{Q}_{2n}$ is of order $o(1/nh)$, and hence can be ignored asymptotically. □

Proof of Theorem 3. Write

$$\boldsymbol{H} = \frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0), \quad \boldsymbol{A} = \frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}\boldsymbol{g}(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*).$$

It is easy to check that $\boldsymbol{L} = \mathbb{E}(\boldsymbol{H}\boldsymbol{A}^{\mathrm{T}})$ and $\boldsymbol{M} = \mathbb{E}(\boldsymbol{A}^{\otimes 2})$. Now $\mathcal{P}$ contains a correctly specified model for $P(R=1|Z,\boldsymbol{U})$ (denoted as $\pi^1(\boldsymbol{\nu}_0^1)$). When $\mathcal{A}$ contains a correctly specified model for $\mathbb{E}(Y|Z,\boldsymbol{U})$ (denoted as $a^1(\boldsymbol{\gamma}_0^1)$), $\mathbb{E}\{\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0)|Z,\boldsymbol{U}\}$ is a component of $\boldsymbol{g}(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)$, and thus $\mathbb{E}\{\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0)|Z,\boldsymbol{U}\}R/\pi^1(\boldsymbol{\nu}_0^1)$ is in the linear space spanned by $\boldsymbol{A}$. Since

$$\mathbb{E}\left(\left[\boldsymbol{H} - \frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}\mathbb{E}\{\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0)|Z,\boldsymbol{U}\}\right]\left\{\frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}f(Z,\boldsymbol{U})\right\}\right) = \boldsymbol{0}$$

for any function $f(Z,\boldsymbol{U})$ and all components of $\boldsymbol{g}(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)$ are functions of Z and $\boldsymbol{U}$ only, we have that

$$\boldsymbol{L}\boldsymbol{M}^{-1}\boldsymbol{A} = \mathbb{E}(\boldsymbol{H}\boldsymbol{A}^{\mathrm{T}})\{\mathbb{E}(\boldsymbol{A}^{\otimes 2})\}^{-1}\boldsymbol{A} = \frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}\mathbb{E}\{\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0)|Z,\boldsymbol{U}\}.$$

This fact yields that $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{g}(\boldsymbol{\nu}_*, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) = \mathbb{E}\{\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0)|Z,\boldsymbol{U}\}$, and thus

$$\boldsymbol{Q} = \frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0) - \frac{R-\pi^1(\boldsymbol{\nu}_0^1)}{\pi^1(\boldsymbol{\nu}_0^1)}\mathbb{E}\{\sqrt{h}\boldsymbol{\phi}(\boldsymbol{\alpha}_0)|Z,\boldsymbol{U}\}.$$

Similar to previous proof, we write $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{Q}_i = \sqrt{nh}\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{Q}_i/\sqrt{h} = \sqrt{nh}\cdot(\boldsymbol{Q}_{1n} + \boldsymbol{Q}_{2n} - \boldsymbol{Q}_{3n})$, where $\boldsymbol{Q}_{1n}$ is the same as in the previous proof, and

$$\boldsymbol{Q}_{2n} = n^{-1}\sum_{i=1}^n\{\frac{R_i}{\pi_i^1(\boldsymbol{\nu}_0^1)}-1\}K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha}_0)\left[a_i^1(\boldsymbol{\gamma}_0^1) - \mu\{\theta(Z_i)\}\right]\boldsymbol{G}(Z_i-z),$$

and

$$\boldsymbol{Q}_{3n} = n^{-1}\sum_{i=1}^n K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha}_0)\left[\mu\{\theta(Z_i)\} - \mu\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}_0\}\right]\boldsymbol{G}(Z_i-z).$$

It is easily seen that $\boldsymbol{Q}_{1n}$ and $\boldsymbol{Q}_{2n}$ have mean 0. The third term $\boldsymbol{Q}_{3n}$ is the leading bias term, and simple calculations show that $\mathbb{E}\left[\boldsymbol{Q}_{3n}\right]$ is equal to (A.2). It follows that

$$bias\{\hat{\boldsymbol{\alpha}}_{MR}(z)\} = \frac{1}{2}h^2\theta''(z)c_2(K) + o(h^2).$$

Note that the variance of $\boldsymbol{Q}_{3n}$ is of order $o(1/nh)$, and hence can be ignored asymptotically. When both $\mathcal{P}$ and $\mathcal{A}$ contain a correctly specified model, $\boldsymbol{Q}_{1n} + \boldsymbol{Q}_{2n}$ is asymptotically normal with mean 0 and variance

$$\mathrm{Var}\left\{\boldsymbol{Q}_{1n} + \boldsymbol{Q}_{2n}\right\} = \frac{1}{n}\mathrm{Var}\left\{\boldsymbol{Q}_{1,2}\right\},$$

where

$$\begin{aligned}\boldsymbol{Q}_{1,2} &= K_h(Z-z)\mu^{(1)}(z,\boldsymbol{\alpha}_0)V^{-1}(z,\boldsymbol{\alpha}_0)\boldsymbol{G}(Z-z) \\ &\quad\times\left(\frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}\left[Y-\mu\{\theta(Z)\}\right] - \left\{\frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}-1\right\}\left[a^1(\boldsymbol{\gamma}_0^1)-\mu\{\theta(Z)\}\right]\right)\end{aligned}$$

Further calculation shows that $n^{-1}\mathrm{Var}\left\{\boldsymbol{Q}_{1,2}\right\}$ equals

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\Bigg[K_h^2(Z-z)\left\{\mu^{(1)}(z,\boldsymbol{\alpha}_0)\right\}^2 V^{-2}(z,\boldsymbol{\alpha}_0)\boldsymbol{G}(Z-z)\boldsymbol{G}(Z-z)^T \\ &\qquad\times\left(\frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}\left[Y-\mu\{\theta(Z)\}\right] - \left\{\frac{R}{\pi^1(\boldsymbol{\nu}_0^1)}-1\right\}\left[a^1(\boldsymbol{\gamma}_0^1)-\mu\{\theta(Z)\}\right]\right)^2\Bigg],\end{aligned}$$

which can be simplified as

$$\begin{aligned}&\frac{1}{nh}f_Z(z)\left[\mu^{(1)}\{\theta(z)\}\right]^2 V^{-2}\{\theta(z)\}\mathbb{E}\Bigg[\frac{\mathrm{Var}(Y|Z,\boldsymbol{U})}{\pi_0(Z,\boldsymbol{U})} \\ &\qquad\qquad + \left[\mathbb{E}(Y|Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right]^2\Big|\, Z=z\Bigg]\boldsymbol{D}(K^2) + o(\frac{1}{nh})\end{aligned}$$

Summarizing all these results gives Theorem 3. □

APPENDIX B

Proofs for Chapter III

Proof of Theorem 3.1. Let us consider one-stage case for now. Our proof is a direct extension of *Rockova and van der Pas* (2017) (refer to as RP17 hereafter). Since we stochastically search optimal regime using Bayesian CART algorithm, it is sufficient to prove that $\Pr(s \gtrsim n^{d_{\mathbf{H}}^0/2\alpha+d_{\mathbf{H}}^0}|\mathbf{H}, \hat{A}^{opt}) \to 0$, where $d_{\mathbf{H}}^0$ is the number of signal variables in the true tree-structured regime that assigns patients to $\hat{A}^{opt}$. The sieve $\mathcal{F}_n$, consisting of step functions over small trees that split only on a few variables, can be constructed for given $n \in \mathbb{N}$ consisting of step functions over small trees (s_n) that split only on a few (q_n) variables using the same way as in RP17:

$$\mathcal{F}_n = \bigcup_{q=0}^{q_n} \bigcup_{s=1}^{s_n} \bigcup_{\mathbf{H}:d_{\mathbf{H}}=q} \mathcal{F}(\mathcal{V}_{\mathbf{H}}^s),$$

where $|d_{\mathbf{H}}|$ is the number of variables in $\mathbf{H}$ that actually determine the decision rules, and $\mathcal{V}_{\mathbf{H}}^s$ denotes a family of valid tree partitions with s leaves based on $\mathbf{H}$ variables. The set of step functions supported by $\mathcal{V}_{\mathbf{H}}^s$ is then defined as

$$\mathcal{F}(\mathcal{V}_{\mathbf{H}}^s) = \left\{ f_{\mathcal{T},\boldsymbol{\beta}} : [0,1]^p \to (1,\cdots,K); f_{\mathcal{T},\boldsymbol{\beta}}(\mathbf{H}) = \sum_{l=1}^{s} \beta_l I_{\Omega_l}(\mathbf{H}); \mathcal{T} \in \mathcal{V}_{\mathbf{H}}^s, \boldsymbol{\beta} \in (1,\cdots,K)^s \right\}.$$

Denote the conditional probabilities $(\eta_{1,\cdots,K}) = \Pr(\hat{A}^{opt} = 1, \cdots, K|\mathbf{H})$, and $(\eta^*_{1,\cdots,K})$ denotes the conditional probabilities under true data generating process. Define metric $d(f, f^*) = \sqrt{\sum_{k=1}^K ||\eta_k - \eta_k^*||_n^2}$. Therefore for sets $\mathcal{F}_n \subset \mathcal{F}$, we want to show $\Pi(\mathcal{F} \setminus \mathcal{F}_n) = o(e^{-(\delta+2)n\epsilon_n^2})$, where sequence $\epsilon_n^2 \to 0$ and $n\epsilon_n^2$ is bounded away from 0, and $\delta > 0$ satisfies

$$\sup_{\epsilon_n > \epsilon} \log N(\epsilon/36, \{f \in \mathcal{F}_n : d(f, f^*) < \epsilon_n\}, d(\cdot, \cdot)) \leq n\epsilon_n^2, \tag{B.1}$$

$$\Pi(f \in \mathcal{F}_n : d(f, f^*) \leq \epsilon_n) \geq e^{-\delta n \epsilon_n^2}. \tag{B.2}$$

The ϵ-covering numbers of the set $\{f : d(f, f^*) \leq \epsilon\}$ for $d(\cdot, \cdot)$ is bounded by the $\epsilon\sqrt{n}/\bar{C}$-covering numbers of a Euclidean ball $\{\boldsymbol{\eta} \in [0,1]^{s \times K} = (\boldsymbol{\eta}_{1,\cdots,K}) \in [0,1]^s : \sqrt{\sum_{k=1}^K ||\boldsymbol{\eta}_k - \boldsymbol{\eta}_k^*||_2^2} \leq \epsilon\sqrt{n}/\bar{C}\}$. The tree size and variable dimensions that define the sieve can be selected as $q_n = \lceil C \min\{d_{\mathbf{H}}, n^{q_0/(2\alpha+q_0)} \log^{2\beta} n / \log(p \vee n)\} \rceil$ and $s_n = \lfloor C' n\epsilon_n^2 / \log n \rfloor \asymp n^{q_0/(2\alpha+q_0)} \log^{2\beta-1} n$.

It can be seen that $\Pi(\mathcal{F} \setminus \mathcal{F}_n) < \Pi(s > s_n) + \Pi(q > q_n)$. Using the same arguments as in Section 8.3 of RP17 completes the proof by showing $\Pi(s > s_n) \asymp o(e^{-(\delta+2)n\epsilon_n^2})$ and $\Pi(s > s_n) \asymp o(e^{-(\delta+2)n\epsilon_n^2})$. □

The proof of Theorem 3.2 requires several ancillary results shown below.

The following lemma is a modified version of Theorem 7.1 in *Rockova and van der Pas* (2017), which provides the finite sample bound of BART. Note the convergence is evaluated for $\hat{\mathbb{E}}(\widetilde{Y}|A, \mathbf{H})$, not $\hat{Q}$, and they are not equivalent except at the last stage.

Proof. Since $||\hat{f}||_\infty$ and $||f_0||_\infty$ are bounded, we can verify $||\hat{f} - f_0||_2^4 < \infty$. Then by

Bernstein's inequality (*Christmann and Steinwart*, 2008), we have

$$\Pr\left(\left|\mathbb{P}_n||\hat{f}-f_0||_2^2-\mathbb{E}||\hat{f}-f_0||_2^2\right| \gtrsim \tau/n+\sqrt{\tau/n}\right) \leq e^{-\tau}.$$

Moreover, under the assumptions, *Rockova and van der Pas* (2017) showed that with probability approaching 1, $||\hat{f}-f_0||_n \lesssim n^{-\frac{\alpha}{2\alpha+d_{\mathbf{H}}}} \log^{1/2} n$. Therefore we plug in this result, and

$$\Pr\left(\mathbb{E}||\hat{f}-f_0||_2^2 \gtrsim n^{-\frac{2\alpha}{2\alpha+d_{\mathbf{H}}}} \log n+\tau/n\right) \leq e^{-\tau}.$$

□

The following lemma establishes the convergence rate of $\hat{Q}_t$, $t=1,\cdots,T$.

Theorem 2.1. *Suppose the assumptions in Section 3 hold,* $Pr\left\{\mathbb{P}_n(\widetilde{Y}_t-\hat{\widetilde{Y}}_t)^2 \gtrsim \tau n^{-\zeta}\right\} \leq e^{-\tau}$*, where* $\xi, \zeta > 0$*, then it follows*

$$Pr\left(\mathbb{E}||\hat{Q}_t-Q_t||_2^2 \gtrsim n^{-\frac{2\alpha_t}{2\alpha_t+d_{\mathbf{H}_t}}} \log n+\tau n^{-\min(1,\zeta)}\right) \leq e^{-\tau}.$$

Proof. Recall $\hat{Q}_t=\hat{\mathbb{E}}(\hat{\widetilde{Y}}_t|A_t,\mathbf{H}_t)$, and $Q_t=\mathbb{E}(\widetilde{Y}_t|A_t,\mathbf{H}_t)$. To facilitate the proof, we denote $Q_t^n=\mathbb{E}(\hat{\widetilde{Y}}_t|A_t,\mathbf{H}_t)$. Then we have

$$\mathbb{E}||\hat{Q}_t-Q_t||_2^2 \leq \mathbb{E}||\hat{Q}_t-Q_t^n||_2^2+\mathbb{E}||Q_t^n-Q_t||_2^2.$$

To bound the second term, define $Z_t \equiv \hat{\widetilde{Y}}_t-\widetilde{Y}_t$, thus $Q_t^n-Q_t=\mathbb{E}(Z_t|A_t,\mathbf{H}_t)$. Consider the regression model $Z_t=\mathbb{E}(Z_t|A_t,\mathbf{H}_t)+\epsilon$. It can be seen $||Z_t||_2^2 \geq ||\mathbb{E}(Z_t|A_t,\mathbf{H}_t)||_2^2-||\epsilon||_2^2$ and $||\epsilon||_2^2=O_p(1)$, then $||Q_t^n-Q_t||_2^2 \lesssim ||Z_t||_2^2$. As a result, and again by Bernstein's inequality, $\mathbb{E}||\hat{Q}_t-Q_t||_2^2 \lesssim \mathbb{E}||Z_t||_2^2+\mathbb{E}||\hat{f}-f_0||_2^2$, and

$$\Pr\left(\mathbb{E}||\hat{Q}_t - Q_t||_2^2 \gtrsim n^{-\frac{2\alpha_t}{2\alpha_t + d_{\mathbf{H}_t}}} \log n + \tau n^{-\min(1,\zeta)}\right) \leq e^{-\tau}.$$

□

With these results, now we prove Theorem 3.2.

Proof of **Theorem 3.2**. For finite sample bound derivation, we assume the data are bounded, i.e. there exist some $B \in \mathbb{R}^+$ such that $||Y||_\infty < B$. We start from the final stage T and omit the subscript notation for stage now. Using the triangular inequality, we have

$$\Pr(\hat{g}^{opt} \neq g^{*\text{opt}}) \leq \sum_{i=1}^{K} \Pr(\hat{A}_i \neq A_i^*) + d(\hat{\boldsymbol{p}}, \boldsymbol{p}^*). \tag{B.3}$$

One can easily notice that

$$\begin{aligned}
&\sum_{i=1}^{K} \Pr(\hat{A}_{p_i} \neq A_{p_i}^*) \\
\leq &\Pr\left\{ \sup_{\boldsymbol{p}\in\sigma(\mathcal{T}),\, \boldsymbol{A}_p \neq \boldsymbol{A}_p^*} \mathbb{P}_n \hat{F}(\boldsymbol{p}, \boldsymbol{A}_p) \geq \mathbb{P}_n \hat{F}(\boldsymbol{p}^*, \boldsymbol{A}_p^*) \right\} \\
\leq &\Pr\left\{ \sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} \left| \mathbb{P}_n \hat{F}(\boldsymbol{p}, \boldsymbol{A}_p) - \mathbb{E}F(\boldsymbol{p}, \boldsymbol{A}_p) \right| \geq \tau/2 \right\}.
\end{aligned} \tag{B.4}$$

The second inequality holds because in assumptions we assume for arbitrarily small τ, $\mathbb{E}F(\boldsymbol{p}^*, \boldsymbol{A}_p^*) \geq \sup_{\boldsymbol{p}\in\sigma(\mathcal{T}),\, \boldsymbol{A}_p \neq \boldsymbol{A}_p^*} \mathbb{E}F(\boldsymbol{p}, \boldsymbol{A}_p) + \tau$.

In order to bound this probability, we write

$$\begin{aligned}
&\sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} |\mathbb{P}_n\{\hat{F}(\boldsymbol{p},\boldsymbol{A}_p)\} - \mathbb{E}\{F(\boldsymbol{p},\boldsymbol{A}_p)\}| \\
&\leq \sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} |\mathbb{P}_n\{\hat{F}(\boldsymbol{p},\boldsymbol{A}_p)\} - \mathbb{P}_n\{F(\boldsymbol{p},\boldsymbol{A}_p)\}| + \sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} |\mathbb{P}_n\{F(\boldsymbol{p},\boldsymbol{A}_p)\} - \mathbb{E}\{F(\boldsymbol{p},\boldsymbol{A}_p)\}|.
\end{aligned}$$

The first term can then be bounded:

$$\begin{aligned}
&\sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} |\mathbb{P}_n\{\hat{F}(\boldsymbol{p},\boldsymbol{A}_p)\} - \mathbb{P}_n\{F(\boldsymbol{p},\boldsymbol{A}_p)\}| \\
\leq &\sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} |\mathbb{P}_n \sum_{a=1}^{K_t} \left[\hat{Q}_t(a,\mathbf{H}_t) - Q_t(a,\mathbf{H}_t)\right] \sum_{i=1}^{K_t} I(\mathbf{H}_t \in p_{it}) I(A_{it}=a)| \\
\leq &|\mathbb{P}_n \sum_{i=1}^{K} \left[\hat{Q}(i,\mathbf{H}) - Q(i,\mathbf{H})\right]| \\
\leq &\sum_{i=1}^{K} \left\{\mathbb{P}_n \left[\hat{Q}(i,\mathbf{H}) - Q(i,\mathbf{H})\right]^2\right\}^{1/2} \\
\lesssim &n^{-\frac{\alpha}{2\alpha+p}} \log^{1/2} n.
\end{aligned}$$

The second term can be bounded by the property of VC-class. Following the arguments in *Zhang et al.* (2018b), we have

$$\Pr\left\{\sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} |\mathbb{P}_n\{F(\boldsymbol{p},\boldsymbol{A}_p)\} - \mathbb{E}\{F(\boldsymbol{p},\boldsymbol{A}_p)\}| \gtrsim 1/\sqrt{n} + \sqrt{\tau/n}\right\} \leq e^{-\tau}$$

up to a constant determined by the VC index. Therefore, it follows that

$$\Pr\left\{\sup_{\boldsymbol{p}\in\sigma(\mathcal{T})} \left|\mathbb{P}_n \hat{F}(\boldsymbol{p},\boldsymbol{A}_p) - \mathbb{E}F(\boldsymbol{p},\boldsymbol{A}_p)\right| \gtrsim n^{-\frac{\alpha}{2\alpha+p}} \log^{1/2} n\right\} \leq e^{-\tau}, \tag{B.5}$$

and consequently (B.4) becomes $\sum_{i=1}^{K} \Pr(\hat{A}_{p_i} \neq A^*_{p_i}) \lesssim \exp(n^{\frac{\alpha}{2\alpha+p}}/\log^{1/2} n)$.

In order to bound $d(\hat{\boldsymbol{p}}, \boldsymbol{p}^*)$, we first look at

$$\sup_{\boldsymbol{p}\in\sigma(\mathcal{T}),d(\boldsymbol{p},\boldsymbol{p}^*)\leq d} \left|\mathbb{P}_n\hat{F}(\boldsymbol{p},\boldsymbol{A}^*) - \mathbb{E}F(\boldsymbol{p},\boldsymbol{A}^*) - \left\{\mathbb{P}_n\hat{F}(\boldsymbol{p}^*,\boldsymbol{A}^*) - \mathbb{E}F(\boldsymbol{p}^*,\boldsymbol{A}^*)\right\}\right| \quad \text{(B.6)}$$

$$\leq \sup_{\boldsymbol{p}\in\sigma(\mathcal{T}),d(\boldsymbol{p},\boldsymbol{p}^*)\leq d} |\mathbb{P}_nF(\boldsymbol{p},\boldsymbol{A}^*) - \mathbb{E}F(\boldsymbol{p},\boldsymbol{A}^*) - \{\mathbb{P}_nF(\boldsymbol{p}^*,\boldsymbol{A}^*) - \mathbb{E}F(\boldsymbol{p}^*,\boldsymbol{A}^*)\}| \quad \text{(B.7)}$$

$$+ \sup_{\boldsymbol{p}\in\sigma(\mathcal{T}),d(\boldsymbol{p},\boldsymbol{p}^*)\leq d} \left|\mathbb{P}_n\hat{F}(\boldsymbol{p},\boldsymbol{A}) - \mathbb{P}_nF(\boldsymbol{p},\boldsymbol{A}) - \left\{\mathbb{P}_n\hat{F}(\boldsymbol{p}^*,\boldsymbol{A}) - \mathbb{P}_nF(\boldsymbol{p}^*,\boldsymbol{A})\right\}\right|. \quad \text{(B.8)}$$

The first term (B.7) (denoted as T_6) can be bounded using VC class property. Again by VC preservation theorem,

$$\mathcal{F}_d = \left\{\sum_{i=1}^{K} Q(A_{p_i},\mathbf{H}) - Q(A_{p_i^*},\mathbf{H})\}I(\mathbf{H}\in p_i \triangle p_i^*) : \boldsymbol{p}\in\sigma(\mathcal{T}), d(\boldsymbol{p},\boldsymbol{p}^*)\leq d\right\}$$

is also VC class. Also because $||f||_\infty \leq 2B$ for $\forall f \in \mathcal{F}_d$ and $Var(f) \leq \mathbb{E}f^2 \leq dB^2$, by concentration inequality we have

$$\Pr\left\{||\mathbb{P}_nf - \mathbb{E}f||_\infty \gtrsim \left[\frac{d^{1/2}\log^{1/2}(1/d) + d^{1/2}\tau^{1/2}}{\sqrt{n}} + \frac{\log(1/d)+\tau}{n}\right]\right\} \leq e^{-\tau},$$

and by applying $\log(1/d) \lesssim d^{-\delta}$ for $\forall\delta\in\mathbb{R}^+$,

$$\Pr\{T_6 \gtrsim d^{1/2-\delta}(n^{-1/2} + n^{-1/2}\tau^{1/2}) + d^{-\delta}(n^{-1} + n^{-1}\tau)\} \leq e^{-\tau}. \quad \text{(B.9)}$$

To bound the second term (B.8) (denoted as T_7), we have

$$\begin{aligned}
T_7 \leq & \left| \mathbb{P}_n \left\{ \sum_{i=1}^{K} \{\hat{Q}(A_{p_i}, \mathbf{H}) - \hat{Q}(A_{p_i^*}, \mathbf{H})\} I(\mathbf{H} \in p_i \triangle p_i^*) \right. \right. \\
& \left. \left. - \sum_{i=1}^{K} \{Q(A_{p_i}, \mathbf{H}) - Q(A_{p_i^*}, \mathbf{H})\} I(\mathbf{H} \in p_i \triangle p_i^*) \right\} \right| \\
\leq & \left| \mathbb{P}_n \left\{ \sum_{i=1}^{K} \{\hat{Q}(A_{p_i}, \mathbf{H}) - Q(A_{p_i}, \mathbf{H})\} I(\mathbf{H} \in p_i \triangle p_i^*) \right. \right. \\
& \left. \left. - \sum_{i=1}^{K} \{\hat{Q}(A_{p_i^*}, \mathbf{H}) - Q(A_{p_i^*}, \mathbf{H})\} I(\mathbf{H} \in p_i \triangle p_i^*) \right\} \right|,
\end{aligned}$$

where $p_i \triangle p_i^*$ denotes the symmetric difference, equivalent to $p_i \cup p_i^* / p_i \cap p_i^*$. We further let $\epsilon = \sum_{i=1}^{K} [\mathbb{E}\{\hat{Q}(A_{p_i^*}, \mathbf{H}) - Q(A_{p_i^*}, \mathbf{H})\}^2]^{1/2}$. Now it follows

$$\begin{aligned}
& \mathbb{E} \sum_{i=1}^{K} \{\hat{Q}(A_{p_i^*}, \mathbf{H}) - Q(A_{p_i^*}, \mathbf{H})\} I(\mathbf{H} \in p_i \triangle p_i^*) \\
\leq & \sum_{i=1}^{K} \mathbb{E}[\{\hat{Q}(A_{p_i^*}, \mathbf{H}) - Q(A_{p_i^*}, \mathbf{H})\} I(\mathbf{H} \in p_i \triangle p_i^*)] \\
\leq & \sum_{i=1}^{K} [\mathbb{E}\{\hat{Q}(A_{p_i^*}, \mathbf{H}) - Q(A_{p_i^*}, \mathbf{H})\}^2]^{1/2} [\mathbb{E} I(\mathbf{H} \in p_i \triangle p_i^*)]^{1/2} \\
\leq & d^{1/2} \epsilon.
\end{aligned}$$

As a result $\mathbb{E}T_7 \lesssim d^{1/2}\epsilon$. We further notice that its variance can be bounded by $c'd$. Therefore, by concentration inequalities,

$$\Pr\{T_7 \gtrsim d^{1/2-\delta}(\epsilon + n^{-1/2} + n^{-1/2}\tau^{1/2}) + d^{-\delta}(n^{-1} + n^{-1}\tau)\} \leq e^{-\tau}.$$

Now we take a look at quantity (B.6), hereby denoted as T5:

$$\Pr\{T_5 \gtrsim n^{-\frac{\alpha}{2\alpha+p}} \log^{1/2} n\tau\} \leq e^{-\tau}.$$

By Lemma 18 in *Zhang et al.* (2018b), $\Pr\{d(\hat{\boldsymbol{p}}, \boldsymbol{p}^*) \gtrsim \tau n^{-2/3\frac{\alpha}{2\alpha+p}} \log^{1/2} n\} \leq e^{-\tau}$. Thus at stage T, we have $\Pr(\hat{g}_T^{opt} \neq g_T^{opt}) \lesssim n^{-2/3\frac{\alpha_T}{2\alpha_T+d_{\mathbf{H}_T}}} \log^{1/2} n \lesssim n^{-r_T+\epsilon}$, and $\Pr(V(g_T^{opt}) - V(\hat{g}_T^{opt}) \gtrsim \tau n^{-r_T+\epsilon}) \leq e^{-\tau}$.

For stage $t = T-1$, we have $\Pr(||\widetilde{Y}_t - \hat{\widetilde{Y}}_t||_n \gtrsim \tau n^{-r_T+\epsilon}) \leq e^{-\tau}$, then by Lemma 2.1, one can easily get

$$\Pr\left(\mathbb{E}||\hat{Q}_t - Q_t||_2^2 \gtrsim n^{-\frac{2\alpha_t}{2\alpha_t+d_{\mathbf{H}_t}}} \log n + \tau n^{-2r_T+\epsilon}\right) \leq e^{-\tau},$$

i.e. $\Pr\left(\mathbb{E}||\hat{Q}_t - Q_t||_2^2 \gtrsim \tau n^{-2\min(\frac{\alpha_t}{2\alpha_t+d_{\mathbf{H}_t}}, r_T)+\epsilon}\right) \leq e^{-\tau}$, i.e. the rate depends on the convergence rate of BART assuming $\widetilde{Y}_t$ is observed, and also depends on the convergence rate carried over from previous stage estimations. Using similar arguments, we have $\Pr(\hat{g}_t^{opt} \neq g_t^{opt}) \lesssim n^{-2/3\min(\frac{\alpha_t}{2\alpha_t+d_{\mathbf{H}_t}}, r_T)} \log^{1/2} n \lesssim n^{-r_t+\epsilon}$, and $\Pr(V(g_T^{opt}) - V(\hat{g}_T^{opt}) \gtrsim \tau n^{-r_t s+\epsilon}) \leq e^{-\tau}$.

Similarly, results can be obtained for all t.

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Ajani, J., et al. (2013), A phase ii randomized trial of induction chemotherapy versus no induction chemotherapy followed by preoperative chemoradiation in patients with esophageal cancer, *Annals of oncology*, *24*(11), 2844–2849.

Bien, J., J. Taylor, and R. Tibshirani (2013), A lasso for hierarchical interactions, *Annals of statistics*, *41*(3), 1111.

Breiman, L. (2001), Random forests, *Machine learning*, *45*(1), 5–32.

Chan, K. C. G., S. C. P. Yam, et al. (2014), Oracle, multiple robust and multipurpose calibration in a missing response problem, *Statistical Science*, *29*(3), 380–396.

Chang, M., S. Lee, and Y.-J. Whang (2015), Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements, *The Econometrics Journal*, *18*(3), 307–346.

Chen, G., D. Zeng, and M. R. Kosorok (2016), Personalized dose finding using outcome weighted learning, *Journal of the American Statistical Association*, *111*(516), 1509–1521.

Chen, J., R. Sitter, and C. Wu (2002), Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika*, *89*(1), 230–237.

Chen, S., and D. Haziza (2017), Multiply robust imputation procedures for the treatment of item nonresponse in surveys, *Biometrika*, *104*(2), 439–453.

Chipman, H. A., E. I. George, and R. E. McCulloch (1998), Bayesian cart model search, *Journal of the American Statistical Association*, *93*(443), 935–948.

Chipman, H. A., E. I. George, R. E. McCulloch, et al. (2010), Bart: Bayesian additive regression trees, *The Annals of Applied Statistics*, *4*(1), 266–298.

Choi, N. H., W. Li, and J. Zhu (2010), Variable selection with the strong heredity constraint and its oracle property, *Journal of the American Statistical Association*, *105*(489), 354–364.

Christmann, A., and I. Steinwart (2008), Support vector machines.

Day, D. B., et al. (2017), Association of ozone exposure with cardiorespiratory pathophysiologic mechanisms in healthy adults, *Jama Internal Medicine*, *177*(9), 1344–1353.

Denison, D. G., B. K. Mallick, and A. F. Smith (1998), A bayesian cart algorithm, *Biometrika*, *85*(2), 363–377.

Deville, J.-C., and C.-E. Särndal (1992), Calibration estimators in survey sampling, *Journal of the American statistical Association*, *87*(418), 376–382.

Donnelly, A., B. Misstear, and B. Broderick (2011), Application of nonparametric regression methods to study the relationship between no 2 concentrations and local wind direction and speed at background sites, *Science of the Total Environment*, *409*(6), 1134–1144.

Duda, R. O., P. E. Hart, and D. G. Stork (2012), *Pattern classification*, John Wiley & Sons.

Fan, A., W. Lu, and R. Song (2016), Sequential advantage selection for optimal treatment regime, *The annals of applied statistics*, *10*(1), 32.

Gail, M., and R. Simon (1985), Testing for qualitative interactions between treatment effects and patient subsets., *Biometrics*, *41*(2), 361–372.

Gao, X., and R. J. Carroll (2017), Data integration with high dimensionality, *Biometrika*, *104*(2), 251–272.

Gill, R. D., M. J. Van Der Laan, and J. M. Robins (1997), Coarsening at random: Characterizations, conjectures, counter-examples, in *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294, Springer.

Giorgini, P., et al. (2015a), Higher fine particulate matter and temperature levels impair exercise capacity in cardiac patients, *Heart*, pp. heartjnl–2014.

Giorgini, P., et al. (2015b), Particulate matter air pollution and ambient temperature: opposing effects on blood pressure in high-risk cardiac patients, *Journal of hypertension*, *33*(10), 2032–2038.

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

Hager, R., A. A. Tsiatis, and M. Davidian (2018), Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data, *Biometrics*, *74*(4), 1180–1192.

Han, P. (2014a), A further study of the multiply robust estimator in missing data analysis, *Journal of Statistical Planning and Inference*, *148*, 101–110.

Han, P. (2014b), Multiply robust estimation in regression analysis with missing data, *Journal of the American Statistical Association*, *109*(507), 1159–1173.

Han, P. (2016a), Combining inverse probability weighting and multiple imputation to improve robustness of estimation, *Scandinavian Journal of Statistics*, *43*, 246–260.

Han, P. (2016b), Combining inverse probability weighting and multiple imputation to improve robustness of estimation, *Scandinavian Journal of Statistics*, *43*(1), 246–260.

Han, P., and L. Wang (2013), Estimation with missing data: beyond double robustness, *Biometrika*, *100*(2), 417–430.

Heitjan, D. F. (1993), Ignorability and coarse data: Some biomedical examples, *Biometrics*, pp. 1099–1109.

Hill, J. L. (2011), Bayesian nonparametric modeling for causal inference, *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Hirano, K., G. W. Imbens, and G. Ridder (2003), Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, *71*(4), 1161–1189.

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

Hsu, Y.-C. (2017), Consistent tests for conditional treatment effects, *The econometrics journal*, *20*(1), 1–22.

Huang, X., S. Choi, L. Wang, and P. F. Thall (2015), Optimization of multi-stage dynamic treatment regimes utilizing accumulated data, *Statistics in medicine*, *34*(26), 3424–3443.

Jiang, R., W. Lu, R. Song, and M. Davidian (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.

Kang, J. D., J. L. Schafer, et al. (2007), Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statistical science*, *22*(4), 523–539.

Kennedy, E. H., W. L. Wiitala, R. A. Hayward, and J. B. Sussman (2013), Improved cardiovascular risk prediction using nonparametric regression and electronic health record data, *Medical care*, *51*(3), 251.

Kim, J., D. Pollard, et al. (1990), Cube root asymptotics, *The Annals of Statistics*, *18*(1), 191–219.

Kosorok, M. R. (2008), *Introduction to empirical processes and semiparametric inference.*, Springer.

Laber, E., and Y. Zhao (2015), Tree-based methods for individualized treatment regimes, *Biometrika*, *102*(3), 501–514.

Laurent, H., and R. L. Rivest (1976), Constructing optimal binary decision trees is np-complete, *Information processing letters*, *5*(1), 15–17.

Linero, A. R. (2018), Bayesian regression trees for high-dimensional prediction and variable selection, *Journal of the American Statistical Association*, *113*(522), 626–636.

Linero, A. R., and Y. Yang (2018), Bayesian regression tree ensembles that adapt to smoothness and sparsity, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 1087–1110.

Little, R., and D. Rubin (2002), *Statistical analysis with missing data*, Wiley, Hoboken.

Little, R. J. (1992), Regression with missing x's: a review, *Journal of the American Statistical Association*, *87*(420), 1227–1237.

Little, R. J., and D. B. Rubin (2019), *Statistical analysis with missing data*, vol. 793, Wiley.

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

Lugosi, G., A. Nobel, et al. (1996), Consistency of data-driven histogram methods for density estimation and classification, *The Annals of Statistics*, *24*(2), 687–706.

McCullagh, P., and J. Nelder (1989), *Generalized linear models*, Chapman and Hall, London New York.

McMurry, T. L., and D. N. Politis (2008), Bootstrap confidence intervals in nonparametric regression with built-in bias correction, *Statistics & Probability Letters*, *78*(15), 2463–2469.

Molina, J., A. Rotnitzky, M. Sued, and J. Robins (2017), Multiple robustness in factorized likelihood models, *Biometrika*.

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

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

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

Murray, T. A., Y. Yuan, and P. F. Thall (2018), A bayesian machine learning approach for optimizing dynamic treatment regimes, *Journal of the American Statistical Association*, *113*(523), 1255–1267.

Murthy, S. K., and S. Salzberg (1995), Decision tree induction: How effective is the greedy heuristic?, in *KDD*, pp. 222–227.

Nahum-Shani, I., M. Qian, D. Almirall, W. E. Pelham, B. Gnagy, G. A. Fabiano, J. G. Waxmonsky, J. Yu, and S. A. Murphy (2012), Q-learning: A data analysis method for constructing adaptive interventions., *Psychological methods*, *17*(4), 478.

Naik, C., E. J. McCoy, and D. J. Graham (2016), Multiply robust estimation for causal inference problems, *arXiv preprint arXiv:1611.02433*.

Nieman, D. R., and J. H. Peters (2013), Treatment strategies for esophageal cancer, *Gastroenterology Clinics*, *42*(1), 187–197.

Parikh, N. D., P. Zhang, A. G. Singal, B. A. Derstine, V. Krishnamurthy, P. Barman, A. K. Waljee, and G. L. Su (2018), Body composition predicts survival in patients with hepatocellular carcinoma treated with transarterial chemoembolization, *Cancer research and treatment: official journal of Korean Cancer Association*, *50*(2), 530.

Pepe, M. S. (1992), Inference using surrogate outcome data and a validation sample, *Biometrika*, *79*(2), 355–365.

Pepe, M. S., M. Reilly, and T. R. Fleming (1994), Auxiliary outcome data and the mean score method, *Journal of Statistical Planning and Inference*, *42*(1-2), 137–160.

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

Qin, J., and J. Lawless (1994), Empirical likelihood and general estimating equations, *The Annals of Statistics*, pp. 300–325.

Radchenko, P., and G. M. James (2010), Variable selection using adaptive nonlinear interaction structures in high dimensions, *Journal of the American Statistical Association*, *105*(492), 1541–1553.

Raghunathan, T. E., P. W. Solenberger, and J. Van Hoewyk (2002), Iveware: Imputation and variance estimation software.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009), Sparse additive models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(5), 1009–1030.

Robins, J. M. (2004), Optimal structural nested models for optimal sequential decisions, in *Proceedings of the second seattle Symposium in Biostatistics*, pp. 189–326, Springer.

Robins, J. M., and M. A. Hernán (2009), Estimation of the causal effects of time-varying exposures, in *Advances in Longitudinal Data Analysis*, edited by G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, pp. 553–599, Chapman and Hall/CRC Press, Boca Raton.

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

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995), Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the american statistical association*, *90*(429), 106–121.

Rockova, V., and E. Saha (2018), On theory for bart, *arXiv preprint arXiv:1810.00787*.

Rockova, V., and S. van der Pas (2017), Posterior concentration for bayesian regression trees and their ensembles, *arXiv preprint arXiv:1708.08734*.

Rossi, G. (2011), Partition distances, *arXiv preprint arXiv:1106.4579*.

Rotnitzky, A., J. Robins, and L. Babino (2017), On the multiply robust estimation of the mean of the g-functional, *arXiv preprint arXiv:1705.08582*.

Ruppert, D. (1997), Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation, *Journal of the American Statistical Association*, *92*(439), 1049–1062.

Schulte, P. J., A. A. Tsiatis, E. B. Laber, and M. Davidian (2014), Q-and a-learning methods for estimating optimal dynamic treatment regimes, *Statistical science: a review journal of the Institute of Mathematical Statistics*, *29*(4), 640.

She, Y., Z. Wang, and H. Jiang (2018), Group regularized estimation under structural hierarchy, *Journal of the American Statistical Association*, *113*(521), 445–454.

Shi, C., A. Fan, R. Song, W. Lu, et al. (2018), High-dimensional *a*-learning for optimal dynamic treatment regimes, *The Annals of Statistics*, *46*(3), 925–957.

Shi, C., R. Song, W. Lu, et al. (2019), On testing conditional qualitative treatment effects, *The Annals of Statistics*, *47*(4), 2348–2377.

Shortreed, S. M., and A. Ertefaie (2017), Outcome-adaptive lasso: Variable selection for causal inference, *Biometrics*, *73*(4), 1111–1122.

Simoneau, G., E. E. Moodie, J. S. Nijjar, R. W. Platt, and S. E. R. A. I. C. Investigators (2019), Estimating optimal dynamic treatment regimes with survival outcomes, *Journal of the American Statistical Association*, (just-accepted), 1–24.

Singal, A. G., et al. (2016), Body composition features predict overall survival in patients with hepatocellular carcinoma, *Clinical and translational gastroenterology*, *7*(5), e172.

Tao, Y., L. Wang, and D. Almirall (2018), Tree-based reinforcement learning for estimating optimal dynamic treatment regimes, *The Annals of Applied Statistics*.

Tchetgen, E. J. T. (2009), A commentary on g. molenberghs's review of missing data methods, *Drug Information Journal*, *43*(4), 433–435.

Thall, P. F., L. H. Wooten, C. J. Logothetis, R. E. Millikan, and N. M. Tannir (2007), Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring, *Statistics in medicine*, *26*(26), 4687–4702.

Tsiatis, A. (2006), *Semiparametric theory and missing data*, Springer, New York.

van der Vaart, A. W. (1998), *Asymptotic statistics*, Cambridge University Press, Cambridge, UK New York, NY, USA.

Wang, L., and E. T. Tchetgen (2016), Bounded, efficient and triply robust estimation of average treatment effects using instrumental variables, *arXiv preprint arXiv:1611.09925*.

Wang, L., A. Rotnitzky, and X. Lin (2010), Nonparametric regression with missing outcomes using weighted kernel estimating equations, *Journal of the American Statistical Association*, *105*(491), 1135–1146.

Wang, L., A. Rotnitzky, X. Lin, R. E. Millikan, and P. F. Thall (2012), Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer, *Journal of the American Statistical Association*, *107*(498), 493–508.

Wooldridge, J. M. (2007), Inverse probability weighted estimation for general missing data problems, *Journal of Econometrics*, *141*(2), 1281–1301.

Wu, Y., H. Tjelmeland, and M. West (2007), Bayesian cart: Prior specification and posterior simulation, *Journal of Computational and Graphical Statistics*, *16*(1), 44–66.

Xu, C., and S. H. Lin (2016), Esophageal cancer: comparative effectiveness of treatment options, *Comparative Effectiveness Research*, *6*, 1–12.

Yang, Y. (1999), Minimax nonparametric classification. i. rates of convergence, *IEEE Transactions on Information Theory*, *45*(7), 2271–2284.

Yuan, M., and Y. Lin (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Yuan, M., V. R. Joseph, and Y. Lin (2007), An efficient variable selection approach for analyzing designed experiments, *Technometrics*, *49*(4), 430–439.

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

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

Zhang, B., M. Zhang, et al. (2018a), Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariates, *The Annals of Applied Statistics*, *12*(4), 2335–2358.

Zhang, D., X. Lin, and M. Sowers (2000), Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles, *Biometrics*, *56*(1), 31–39.

Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian (2015), Using decision lists to construct interpretable and parsimonious treatment regimes, *Biometrics*, *71*(4), 895–904.

Zhang, Y., E. B. Laber, M. Davidian, and A. A. Tsiatis (2018b), Estimation of optimal treatment regimes using lists, *Journal of the American Statistical Association*, pp. 1–9.

Zhao, Y., D. Zeng, M. A. Socinski, and M. R. Kosorok (2011), Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer, *Biometrics*, *67*(4), 1422–1433.

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

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

Zhu, R., and M. R. Kosorok (2012), Recursively imputed survival trees, *Journal of the American Statistical Association*, *107*(497), 331–340.