Variable Selection for Dynamic Treatment Regimens

Zeyu Bian

Doctor of Philosophy

Department of Epidemiology and Biostatistics

McGill University
Montréal, Québec
May 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

© Copyright Zeyu Bian, 2022

Dedication

This thesis is dedicated to my parents, Mei Gao and Maoyun Bian.

Acknowledgements

I am incredibly grateful for the guidance and enthusiasm that I have received in my research over the past two years from my devoted supervisors, Dr. Sahir Bhatnagar and Dr. Erica Moodie. I want to thank them for leading me to the areas of variable selection and dynamic treatment regimens and guiding me in identifying my own research interests. I really appreciate all the time that Sahir spent teaching me academic writing, coding, and how to create an **R** package. When I first met him, he was a PhD student in the department, and after a year, he became my PhD supervisor, so he was really able to view things from a student's perspective. Sahir's personality is contagious, and I am grateful for his support and encouragement, especially when I was entering the job market. I am also forever indebted to Erica, who introduced me to adaptive treatment strategies and taught me academic integrity, and more importantly, shared her enthusiasm for research with me. When I first entered the program, I never thought that I would one day want to pursue a career in academia. However, with two amazing years of work with Sahir and Erica, I changed my mind—thanks for their passion and inspiration for research. Currently, I have found what I really want to pursue for my career, and I hope to someday teach and inspire my own students like the way they have instructed me.

I would also like to thank my PhD committee member, Dr. Susan Shortreed for providing guidance and input on the three manuscripts in this thesis. I would like to express my deepest gratitude to my uncle Professor Yisheng Li for his influence and guidance during my doctoral studies. I am also grateful to my friend Larry, who helped me revise the French abstract of this thesis; I will always remember all the good old days that we spent together in Burnside Hall and Purvis Hall, sharing our research and dreams with each other. A special thanks to my friends Yikou Yiming Zheng, Xiaohui Zhao, Leo, Gaibian Wei Hu, Xiying Sun, Fan Yang, and Jiaxi, for always being there for me.

Finally, I would like to thank my parents—for everything, I could not have completed all this without them.

Preface

This manuscript-based PhD thesis contains new research in the area of variable selection for dynamic treatment regimens, the elements of the thesis that are considered original scholarship and distinct contributions to knowledge. This document consists of six chapters. Chapter 1 gives a general overview of this thesis, and Chapter 2 provides a comprehensive review of the relevant literature. They were written entirely by Zeyu Bian (ZB) and edited by Sahir Rai Bhatnagar (SRB) and Erica E. M. Moodie (EEMM).

Chapter 3 was conceptualized by EEMM, SRB and ZB. The methodological work, theoretical proofs, writing, programming, and the real data analysis were done by ZB under the guidance of EEMM, SRB, and Susan M Shortreed (SMS). EEMM, SRB and SMS corrected and edited this chapter.

The methodological work in Chapter 4 was conceptualized by EEMM, ZB, SRB and SMS. Sylvie D Lambert (SDL) provided assistance for analyzing the web-based, stress management intervention data. ZB designed and conducted the numerical studies, performed the data analysis, and wrote the draft of the manuscript. The chapter has also been edited by EEMM, SRB, SMS and SDL.

The idea in Chapter 5 was conceptualized by ZB, EEMM, and SRB. The methodology, algorithms, theoretical proof, writing, simulation studies, and the data analysis were carried out by ZB. EEMM, SRB and SMS advised and edited this Chapter; the web-based data were provided by SDL. Chapter 6, the conclusion and summary, was written by ZB and edited by SRB, EEMM, SMS, and SDL.

Abstract

In the precision medicine paradigm, treatment decisions are tailored to each individual, instead of a "one-size-fits-all" approach, which is beneficial in the presence of heterogeneous treatment effects. With the aim of improving individual patients' health outcomes, dynamic treatment regimens (DTRs) recommend effective treatments for individual patients based on their characteristics. However, collected data often contain many irrelevant variables for tailoring treatment. Including all the covariates in an analysis could yield a loss of statistical efficiency and an unnecessarily complicated treatment decision rule, which is difficult for physicians to assess or implement. Thus, variable selection with the objective of optimizing patients' outcome by identifying useful tailoring variables is important. The topic of variable selection in a general context has been well studied, however, it has been less investigated in the area of DTRs. Applying existing variable selection techniques to DTRs estimation methods directly can be challenging: first, the goal of variable selection in DTRs differs from variable selection in a general context. Variable selection for DTRs aims to improve the estimated decision rules instead of predictive performance. Second, in DTRs, we are most interested in selecting variables that may be effect modifiers-a scenario that is rarely considered in the prediction setting. Last, DTRs are often estimated using semi-parametric methods that provide robustness against model misspecification. Many existing methods are complicated and hard to implement (especially for count and binary outcomes), thus it is difficult to extend these to a regularization framework. In such a case, we might want to use a simpler regression-based method.

The overarching goal of this thesis is to develop new variable selection techniques in the DTRs setting. This thesis consists of three manuscripts. In the first manuscript, I extend the estimation approach of dynamic weighted ordinary least squares to a penalized framework, where estimation and variable selection for DTRs can be performed simultaneously. I show that this extension has the double robustness and oracle properties under some con-

ditions. The newly proposed method is applied to data from the Sequenced Treatment Alternatives to Relieve Depression study. The second manuscript considers two practical issues that frequently arise in causal inference and variable selection approaches: confounder selection and tuning parameter selection. The approach from the first paper is combined with a confounder selection method, and this is illustrated on data from a pilot sequential multiple assignment randomized trial of a web-based stress management study. In these first two works, I only considered the case in which the outcome is continuous, while in the third manuscript, I extend the doubly robust penalized weighted regression approach to the discrete outcomes setting.

In this thesis, I show that with a suitable choice of weights, a weighted penalized regression model still enjoys the desired double robustness property, and yet is straightforward to implement. The advantage of the newly proposed approach compared to alternative regularized DTRs estimation methods lies in the fact that it can be viewed from a minimization perspective. Hence, the implementation is simpler, various penalty functions can be used, and the solution can be found using existing computationally efficient tools.

Abrégé

Dans le cadre de la médecine de précision, les décisions quant aux traitements sont adaptées à chaque personne, au lieu d'une approche « taille unique», et celles-ci seront plus bénéfiques en présence de traitement avec des effets hétérogènes. Dans le but d'améliorer les résultats de santé de chaque patient, les régimes de traitement dynamiques (RTD) recommandent des traitements efficaces pour chaque patient en fonction de ses caractéristiques. Toutefois, les données amassées contiennent souvent de nombreuses variables et beaucoup d'entre elles peuvent être impertinentes dans le choix du meilleur traitement. L'inclusion de toutes les covariables dans une analyse pourrait entraîner une perte d'efficacité statistique et une règle de décision de traitement inutilement compliquée qui sera difficile à évaluer ou à mettre en œuvre par les médecins. Ainsi, la sélection de variables dans le but d'optimiser les résultats des patients en identifiant des variables importantes pour le choix du traitement optimal est importante. Le sujet de la sélection de variables dans un contexte général a été bien étudié, mais il a été moins étudié dans le contexte des RTD. L'application directe des techniques de sélection de variables existantes aux méthodes d'estimation des RTD peut être difficile : premièrement, l'objectif de la sélection des variables dans les RTD diffère de la sélection des variables dans un contexte général, c'est-à-dire que la sélection de variables pour les RTD vise à améliorer les règles de décision estimées au lieu des performances prédictives. Deuxièmement, dans les RTD, on s'intéresse plus à la sélection de variables qui peuvent être des modificateurs d'effet - un scénario qui est rarement pris en compte dans le cadre de la prédiction. Enfin, les RTD sont souvent estimés à l'aide de méthodes semi-paramétriques qui offrent une robustesse contre les erreurs de spécification du modèle statistique. De nombreuses méthodes existantes sont compliquées et difficiles à mettre en œuvre (en particulier pour les variables dépendantes discrètes), il est donc difficile de les incorporer dans un cadre de régularisation. Dans ce cas-ci, nous allons utiliser une méthode plus simple et basée sur la régression.

L'objectif principal de cette thèse est de développer de nouvelles techniques de sélection de variables dans le cadre des RTD. Cette thèse se compose de trois manuscrits. Dans le premier manuscrit, je propose une approche d'estimation des moindres carrés ordinaires pondérés dynamiques dans un cadre pénalisé, où l'estimation et la sélection de variables pour les RTD peuvent être effectuées simultanément. Je montre que cette extension possède la double propriété de robustesse et d'oracle sous certaines conditions. La nouvelle méthode proposée est appliquée aux données de l'étude Sequenced Treatment Alternatives to Relieve Depression. Le deuxième manuscrit examine deux problèmes pratiques qui arrivent fréquemment dans les approches d'inférence causale et de sélection de variables : la sélection des facteurs de confusion et la sélection des paramètres de réglage. L'approche du premier article est combinée avec une méthode de sélection des facteurs de confusion, et ceci est illustré à l'aide de données provenant d'un essai randomisé pilote séquentiel d'une étude sur la gestion du stress basée sur le Web. Dans ces deux premiers travaux, je n'ai considéré que le cas où le résultat est continu, tandis que dans le troisième manuscrit, j'étends l'approche de régression pondérée pénalisée doublement robuste au cadre du résultat discret.

Dans cette thèse, je montre qu'avec un choix approprié de poids, un modèle de régression pénalisé pondéré bénéficie toujours de la propriété de double robustesse, tout en étant simple à mettre en œuvre. L'avantage de cette approche-ci par rapport aux autres méthodes d'estimation des RTD régularisés est qu'elle peut être considérée dans une perspective de minimisation. Par conséquent, la mise en œuvre est plus simple, plusieurs fonctions de pénalité peuvent être utilisées et la solution peut être trouvée à l'aide d'outils de calcul efficaces qui existent déjà.

Table of contents

List of Tables

List of Figures

Abbreviations

AIC Akaike Information Criterion

AIPTW Augmented Inverse Probability of Treatment Weighting

ATE Average Treatment Effect

BIC Bayesian Information Criterion

CATE Conditional Average Treatment Effect

CI Confidence Interval

CVD Cardiovascular Disease

DASS Depression Anxiety Stress Scales

DTR Dynamic Treatment Regimen

dWOLS Dynamic Weighted Ordinary Least Squares

IPTW Inverse Probability of Treatment Weighting

ITR Individualized Treatment Rule

LASSO Least Absolute Shrinkage and Selection Operator

MAE Mean Absolute Error

MCS Mental Component Score

MI Motivational Interviewing

MSE Mean Squared Error

PAL Penalized A-Learning

PDR Penalized Doubly Robust

pdWOLS Penalized Dynamic Weighted Ordinary Least Squares

QIDS Quick Inventory of Depressive Symptomatology

REE Regularized Estimating Equation

SCAD Smoothly Clipped Absolute Deviation

SMART Sequential Multiple Assignment Randomized Trial

SMD Standardized Mean Difference

SSRI Selective Serotonin Reuptake Inhibitor

STAR*D Sequenced Treatment Alternatives to Relieve Depression

SUTVA Stable Unit Treatment Value Assumption

VIC Value Information Criterion

Chapter 1

Introduction

This PhD thesis aims to develop new statistical methods to select important tailoring variables for optimal treatment decision-making. In the area of personalized medicine (also known as precision medicine), treatment strategies are tailored to each individual, rather than a "one-size-fits-all" approach, which is believed to be beneficial in the presence of heterogeneous treatment effects. In order to optimize individual patient's health, dynamic treatment regimens (DTRs), or adaptive treatment strategies find optimal treatment decisions for individual patients according to their specific history (Murphy, 2003; Robins, 2004; Chakraborty and Moodie, 2013; Kosorok and Moodie, 2015; Tsiatis et al., 2019). However, with many collected covariates as well as a complex disease process, it is extremely challenging to determine which tailoring variables might be relevant for making treatment decisions. Including all the potential variables at hand in the statistical analysis could result in a loss of statistical efficiency (i.e., the resulting estimator has large variance), a superfluously complicated model, and a needlessly intricate treatment decision rule, which is difficult for medical researchers to assess or use.

Thus, it is important to consider a data-driven approach of selecting relevant tailoring variables in order to optimize patients' outcome, meanwhile, to simplify models to improve

tractability (Lu et al., 2013; Shi et al., 2018; Jeng et al., 2018; Bian et al., 2021). Although the field of variable selection has been well studied in a general statistical context (Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005; Zou, 2006; Zhang, 2010), it has been less explored in DTRs. Directly applying well-established variable selection approaches to DTRs can be challenging: first, the purpose of applying variable selection for DTRs is to improve the estimated decision rules and further improve patients' outcome, which differs from the goal of variable selection in a general context, where the goal is to enhance predictive performance. Second, the types of variables that are of most of interest also differs: in DTRs, we are mainly interested in selecting variables that may be effect modifiers, which is seldomly studied in the general statistical setting. Last but not least, in order to provide robustness against possible model misspecification, DTRs are often estimated using semi-parametric statistical methods (Murphy, 2003; Robins, 2004). Many well-developed methods are complicated and hard to implement (especially for discrete outcomes), hence, it is challenging to extend these to a penalized framework, where important variables can be selected and the model can be estimated simultaneously.

In this thesis, I focus on the use of penalized regression methods for variable selection that possess the desired double robustness property. This thesis consists of six chapters. In Chapter 2, I provide a critical review of the literature on causal inference, DTRs, and variable selection. In Chapter 3, I extend the estimation approach of dynamic weighted ordinary least squares (Wallace and Moodie, 2015) to a penalized framework, where estimation and variable selection for DTRs can be performed simultaneously. In Chapter 4, two practical issues that frequently arise in causal inference and variable selection approaches are considered: confounder selection and tuning parameter selection. The approach developed in Chapter 3 is combined with a confounder selection method. In Chapter 5, I extend the doubly robust penalized weighted regression in Chapter 3 to the discrete outcome setting.

My PhD thesis is in a thesis-by-manuscript format: Chapters 3, 4 and 5 were originally

written as stand-alone papers and therefore, there is some inconsistency in notation and overlap with Chapter 2. Chapter 3 has been published in *Biometrics* (Bian et al., 2021). Chapter 4 is a book chapter accepted as a contribution to *Handbook of Statistical Methods for Precision Medicine*. Chapter 5 will be submitted for publication shortly after the submission of the PhD thesis. In Chapter 6, I conclude with an overview of the three manuscripts, a discussion of limitations, and some directions for future work.

Chapter 2

Literature Review

The literature review consists of seven sections. The first summarizes well-established basic concepts and assumptions in causal inference. The second and third sections describe some estimation methods in observational studies for causal inference and DTRs. We then describe some penalized variable selection methods for main effects and interaction models in the fourth and fifth sections, respectively. This is followed by a brief introduction to variable selection in the context of DTRs.

2.1 Causal Inference

Throughout this thesis, I use upper case letters to denote random variables and lower case letters to denote observed variables. For now, I focus on a point-treatment (cross-sectional) setting; a longitudinal setting will be introduced later in this chapter. Let Y denote the outcome of interest (discrete or continuous), X as the covariates, and A as the binary treatment indicator. We begin this section with an important concept in causal inference: the potential outcome framework (Rubin, 1978). Denote the potential outcome under the binary treatment a as Y^a; the potential outcome Y^a of a subject is the outcome of the patient if

treatment a has been taken.

In causal inference, there are several popular quantities that are used to measure causal effects:

1. Average treatment effect (ATE): $\mathbb{E}(Y^1 - Y^0)$.
2. Average treatment effect on the treated: $\mathbb{E}(Y^1 - Y^0|A = 1)$.
3. Conditional average treatment effect (CATE): $\mathbb{E}(Y^1 - Y^0|X = x)$.

The ATE is the most popular target of inference to assess causality in many fields. The CATE is often the interest when the aim is estimating heterogeneous treatment effects (see more about the heterogeneous treatment effects in Section 2.3).

In this section, we focus on estimation methods of the ATE and CATE. To estimate the ATE, the ideal dataset should be in the form of (a, y^a, y^{1-a}), which is inaccessible, as we can only have the access to the data (a, y), i.e., only one outcome can be observed for a subject. However, with some assumptions, we can still estimate the causal effect $\mathbb{E}(Y^1 - Y^0)$ correctly. For now, we assume the following assumptions hold:

1. Stable unit treatment value assumption (SUTVA) (Rubin, 1980): a patient's potential outcome is not affected by other patients' treatment assignments.
2. Consistency: $Y = AY^1 + (1 - A)Y^0$.
3. Exchangeability (Hernán and Robins, 2020): $Y^a \perp\!\!\!\perp A$.
4. Positivity: $P(A = a) > 0$ almost surely.

The consistency assumption states that, given the observed treatment a, the subject's observed outcome y is equal to the potential outcome y^a. Exchangeability (also known as ignorability) means that the actual treatment allocation A is independent of the counterfactual outcome for Y^0 and Y^1.

In an ideal randomized study, we always have $Y^a \perp\!\!\!\perp A$, hence

$$\begin{aligned}\mathbb{E}(Y^1 - Y^0) &= \mathbb{E}(Y^1) - \mathbb{E}(Y^0) \\ &= \mathbb{E}(Y^1|A=1) - \mathbb{E}(Y^0|A=0) \\ &= \mathbb{E}(Y|A=1) - \mathbb{E}(Y|A=0),\end{aligned}$$

which is identifiable from the observational data.

While randomization is highly valued in causal inference, it has some limitations. First, it is highly costly since the randomized trial has to recruit, allocate subjects and follow them up for the duration of the experiment. The second limitation is ethical consideration, e.g., in a research that studies the relationship between smoking and lung cancer, it is not ethical to randomly allocate a subject to the smoking group. Randomized trials may also have a highly selected population (low generalizability), and could suffer from attrition or non-compliance. As such, it is also important to study statistical methods for causal inference in observational studies.

2.2 Estimation Methods for ATE in Observational Study

In observational studies, the condition $Y^a \perp\!\!\!\perp A$ does not hold in general because of the existence of confounders. An analysis that does not take account the effect of the confounder would induce a bias to the estimator $\widehat{\mathbb{E}}(Y^1 - Y^0)$ (Hernán and Robins, 2020).

In observational studies, the exchangeability assumption needs to be relaxed to conditional exchangeability: $Y^a \perp\!\!\!\perp A \,|\, X$. This assumes that the potential outcome Y^a is independent of the treatment A within levels of the covariates X. Moreover, the positivity condition in previous section also needs some modification: $P(A = a|X) > 0$ almost surely for $a = 0, 1$ and every level of X. Here I review three types of estimation methods: propensity score-based methods, outcome regression, and doubly robust estimation methods (which typically

are a fusion of the former two approaches). Note that there are some other estimation methods such as stratification and matching, which are not covered in this section, as the methodological work in this thesis will not rely on these methods.

2.2.1 Propensity Score-based Methods

A function $b(\cdot)$ of x is called a balancing score if it satisfies $A \perp\!\!\!\perp X \,|\, b(x)$ (Dawid, 1979; Rosenbaum and Rubin, 1983). The propensity score (Rosenbaum and Rubin, 1983) is the coarsest balancing score $b(x)$ such that $b(x) = P(A = 1|X = x)$, which is the probability of treatment received conditional on confounders. Note that the emphasis is on modelling treatment as a function of confounding variables, rather than as a function of all variables that predict treatment. In observational studies, this probability is unknown and hence needs to be modeled. I use $\widehat{\pi}$ to denote the estimated propensity score model $\widehat{P}(A = 1|X = x)$.

Inverse Probability of Treatment Weighting (IPTW) (Hernán and Robins, 2020) is one of the most popular propensity score-based estimation methods. The value of $\mathbb{E}(Y^a)$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i = a)Y_i}{A_i\widehat{\pi}_i + (1 - A_i)(1 - \widehat{\pi}_i)} \tag{2.1}$$

for $i = 1, 2, \ldots, n$, which is a weighted average estimator. That is, if the observed treatment is $A = 1$, then we assign weights $1/\widehat{\pi}_i$ to the outcome of the corresponding subject; as for subjects who do not received the treatment, we give weights $(1 - \widehat{\pi}_i)^{-1}$. The IPTW estimator is a consistent estimator of the value $\mathbb{E}(Y^a)$ given that the propensity score model π is correctly specified. The ATE can be estimated by $\widehat{\mathbb{E}}(Y^1) - \widehat{\mathbb{E}}(Y^0)$, which is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_iY_i}{\widehat{\pi}_i} - \frac{(1 - A_i)Y_i}{1 - \widehat{\pi}_i}\right].$$

2.2.2 Outcome Regression

Another approach to estimate the causal effect is outcome regression. The expected value of the potential outcome under the treatment a can be written as

$$\mathbb{E}(Y^a) = \mathbb{E}_X\, \mathbb{E}(Y^a|X) = \mathbb{E}_X\, \mathbb{E}(Y^a|A=a,X) = \mathbb{E}_X\, \mathbb{E}(Y|A=a,X). \tag{2.2}$$

Expression (2.2) suggests that we could first estimate the value of $\mathbb{E}(Y|A=a,X)$, then marginalize it with respect to X to obtain estimator of $\mathbb{E}(Y^a)$. The quantity $\mathbb{E}(Y|A=a,X)$ could be estimated by regressing Y on A and X. Given that outcome model $\mathbb{E}(Y|A=a,X)$ is correctly specified, the causal effect could be estimated consistently by

$$\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mathbb{E}}(Y_i|A_i=1,X_i) - \widehat{\mathbb{E}}(Y_i|A_i=0,X_i)\right].$$

2.2.3 A Doubly Robust Estimation Method

So far, I have introduced how to use IPTW and outcome regression to estimate causal effect. In this subsection, I briefly discuss an approach that combines the IPTW method and the outcome regression method: augmented inverse probability of treatment weighting (AIPTW).

The advantage of using IPTW is that the propensity score is often easier to model than the outcome model. For example, in a randomized study where treatment A is allocated at random with known probabilities that depend on baseline covariates X, the propensity score is known to the researchers. Nonetheless, the disadvantage of using IPTW is related to efficiency, since the estimator of $\mathbb{E}(Y^a)$ only uses the sample points whose observed treatments are $A=a$, and significant residual variability may remain since there is no attempt to explain variability in the outcome that relates to other covariates. On the other hand, additional efficiency can be gained using the outcome regression approach that utilizes all

the observations in the given dataset and reduces residual errors. However, the proposed outcome model often suffers from model misspecification.

AIPTW could also gain additional efficiency compare to IPTW, in addition, it provides protection to potential model misspecification of the propensity score or the outcome mean model. The value of $\mathbb{E}(Y^a)$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i=a)Y_i}{A_i\widehat{\pi}_i+(1-A_i)(1-\widehat{\pi}_i)}+\left[1-\frac{I(A_i=a)}{A_i\widehat{\pi}_i+(1-A_i)(1-\widehat{\pi}_i)}\right]\widehat{\mathbb{E}}(Y_i|X_i,A_i)\right\}.$$

It can be shown that the AIPTW estimator is consistent (Tsiatis, 2006), given that either one of two nuisance models (π and $\mathbb{E}(Y|A,X)$) is correct (not necessarily both). This property is referred to as the double robust property (Robins et al., 1994; Scharfstein et al., 1999; Bang and Robins, 2005). Double robustness property is a highly desirable property, as it can protect against model misspecification of either the propensity score model or the outcome model. Moreover, in the case that both models are correctly specified, the AIPTW estimator is always more efficient than the IPTW estimator (Tsiatis, 2006).

2.3 Dynamic Treatment Regimens

In the precision medicine paradigm, treatment decisions are tailored to each patient based on their characteristics, instead of a "one-size-fits-all" approach. With the aim of improving individual patients' health outcomes, dynamic treatment regimens (DTRs) (Murphy, 2003; Robins, 2004; Chakraborty and Moodie, 2013; Kosorok and Moodie, 2015; Tsiatis et al., 2019) recommend the most promising treatments for individual patients according to their information. In general, DTRs estimation method can be categorized into value search methods and regression-based methods. Some popular value search methods include dynamic marginal structural models (van der Laan and Petersen, 2007; Orellana et al., 2010), robust learning (Zhang et al., 2012), outcome weighted learning (Zhao et al., 2012), and concordance

learning (Fan et al., 2017). As for regression-based methods, Q-learning (Watkins, 1989) and A-learning (Murphy, 2003; Robins, 2004) are widely used in DTRs. In this Section, I briefly review the background of regression-based methods, as this thesis builds upon such framework.

2.3.1 Single-stage DTR

In a single stage DTR, the tuple $V = (X, A, Y)$ consists of the data for an individual patient, where X is the patient's baseline variables (information or covariates), A is the binary treatment, and Y is the patient's continuous outcome (usually a larger value of Y is more desirable). The objective is to find the optimal treatment decision $\widehat{a}^{opt}$ such that the quantity $\mathbb{E}(Y^a)$ is maximized at $\widehat{a}^{opt}$.

Assume that the observations $V_i, i = 1, \ldots, n$ are independent and identically distributed with probability density $h(V)$ with respect to a measure ν. Moreover, we assume the relationship between Y and (X, A) can be captured by a semi-parametric regression model: $\mathbb{E}(Y|X = x, A = a) = f_0(x; \boldsymbol{\beta}_0) + \gamma(x, a; \boldsymbol{\psi_0})$, where f_0 is an unknown baseline mean function, $\boldsymbol{\beta}_0$ and $\boldsymbol{\psi}_0$ are underlying true parameters, and γ is a known function that satisfies $\gamma(x, 0; \boldsymbol{\psi}) = 0$. This latter function is referred to as the blip function in (Robins, 2004), and taken together, the two functions are referred to as a structural mean model. A blip function can written as: $\gamma(x, a) = \mathbb{E}(Y^a|X = x, A = a) - \mathbb{E}(Y^0|X = x, A = a)$. In this semi-parametric model, f_0 is irrelevant for making decisions (a nuisance model). Hence, our parameter of interest is $\boldsymbol{\psi}$, and the optimal treatment decision is given by $\widehat{a}^{opt} = \arg\max_a \gamma\left(x, a; \widehat{\boldsymbol{\psi}}\right)$. From now on, we use f to denote the posited baseline model, which is not necessarily identical to the true baseline mean function f_0.

Example: In a simple one-stage setting, we could assume that both f and γ are linear in form, e.g., $f(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x$ and $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \psi_1 x)$, and hence the estimated optimal treatment is $\widehat{a}^{opt} = I(\widehat{\psi}_0 + \widehat{\psi}_1 x > 0)$, where $I(\cdot)$ is the indicator function.

Similar to the estimation methods for the ATE that I have described in the previous section, the estimation methods for DTRs can also be categorized into three types of approaches: propensity score-based methods, outcome regression and doubly robust methods.

Propensity Score-based Methods: assume that the blip function is linear: $\gamma(x, a; \boldsymbol{\psi}) = ax^T\boldsymbol{\psi}$, and the propensity score model π is correctly specified, then the blip parameter can be consistently estimated by solving the following estimating equation:

$$\sum_{i=1}^{n}(y_i - a_i\boldsymbol{\psi}^T x_i)(a_i - \widehat{\pi}) = \mathbf{0}.$$

The corresponding estimator is called the E-estimator in Robins et al. (1994) and Robins (2004). The optimal treatment decision is given by $\widehat{a}^{opt} = \arg\max_a \gamma\left(x, a; \widehat{\boldsymbol{\psi}}\right)$.

Outcome Regression: if the posited baseline mean model f is correctly specified, the blip parameters can be obtained using least squares:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\psi}}\{\|\boldsymbol{Y} - f(x; \boldsymbol{\beta}) - \boldsymbol{A} \circ \boldsymbol{X}\boldsymbol{\psi}\|^2\},$$

where $\circ$ is the element-wise product. The optimal treatment decision is given by $\widehat{a}^{opt} = \arg\max_a \gamma\left(x, a; \widehat{\boldsymbol{\psi}}\right)$. Note that this outcome modelling approach, in one interval, corresponds to Q-learning (Watkins, 1989). The modelling can also be accomplished non-parametrically using more flexible regressions, see, for instance, Moodie et al. (2014); Murray et al. (2018); Wager and Athey (2018); Logan et al. (2019) and Hill et al. (2020).

Doubly Robust Methods: A-learning (or G-estimation) (Murphy, 2003; Robins, 2004) can yield consistent estimators of $\boldsymbol{\psi}$ while only requiring one of two nuisance models, either the baseline mean model $f(\boldsymbol{x}; \boldsymbol{\beta})$ or the propensity score model, to be correctly specified. The blip parameter can be obtained by solving the following A-learning estimating equation:

$$\boldsymbol{X}^T \operatorname{diag}(\boldsymbol{A} - \widehat{\pi})(\boldsymbol{Y} - f(x; \widehat{\boldsymbol{\beta}}) - \gamma(x, a; \boldsymbol{\psi})) = \mathbf{0},$$

where the plug-in estimators $\widehat{\pi}$ and $f(x;\widehat{\boldsymbol{\beta}})$ can be estimated using linear regression or other machine learning approaches. The optimal treatment decision is given by $\widehat{a}^{opt} = \arg\max_a \gamma\left(x, a; \widehat{\boldsymbol{\psi}}\right)$. See Moodie et al. (2007) for a detailed comparison between Murphy (2003) and Robins (2004).

Dynamic weighted ordinary least squares (dWOLS) (Wallace and Moodie, 2015) is another approach that can be used to estimate DTRs. It uses a regression approach, achieving double robustness through weighting by a function of the propensity score such that the weights satisfy $\pi(\boldsymbol{x})w(1,\boldsymbol{x}) = (1-\pi(\boldsymbol{x}))w(0,\boldsymbol{x})$, where $w(a,\boldsymbol{x})$ is the weight for a subject with treatment a and covariates $\boldsymbol{x}$. Wallace and Moodie (2015) suggested to use "absolute value" weights of the form $w(a,\boldsymbol{x}) = |a - \mathbb{E}[A|\boldsymbol{X} = \boldsymbol{x}]|$, as these offered better efficiency than other alternatives considered, while yielding consistent estimators of blip parameters if either the treatment or baseline mean model is correctly specified. The main assumption of the dWOLS is that the main effects for all covariates in the blip function are included in the baseline mean model. Violation of this assumption, known as the strong heredity property (Chipman, 1996), can lead to biased estimators of blip parameters. Estimation of the blip parameter using dWOLS is accomplished using the weighted squared-error loss:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\| \sqrt{\boldsymbol{W}} \left(\boldsymbol{Y} - \psi_0 \boldsymbol{A} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j - \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ \boldsymbol{X}_j) \right) \right\|_2^2,$$

where $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \psi_0, \ldots, \psi_p)$, and $\boldsymbol{W} = \text{diag}\left\{w_1(a,\boldsymbol{x}), w_2(a,\boldsymbol{x}), \ldots, w_n(a,\boldsymbol{x})\right\}$ is a $n \times n$ diagonal *balancing weights* matrix. Unlike A-learning, the parameter $\boldsymbol{\beta}$ in the baseline mean model is estimated simultaneously with the blip parameter.

So far we have focused on continuous outcomes such that the relationship between Y and (X, A) is in the form of $\mathbb{E}(Y|X = x, A = a) = f_0(x;\boldsymbol{\beta}_0) + \gamma\left(x, a; \boldsymbol{\psi_0}\right)$. A generalization of this additive model is $g\big(\mathbb{E}(Y|X = x, A = a)\big) = f_0(x;\boldsymbol{\beta}) + \gamma\left(x, a; \boldsymbol{\psi}\right)$, where g is a known link function. Doubly robust A-learning estimating function based on this model can be found in Robins (2004) and Tchetgen Tchetgen et al. (2010), where the link function is the

log and logit link, respectively.

2.3.2 Multi-stage DTRs

Now we move to a more complicated scenario, the multi-stage DTRs, where the treatment decisions are made in each stage according to previous patient history. For a K-stage DTR, $(x_1, a_1, x_2, a_2, \ldots, x_K, a_K, y)$ consists of the observed data for an individual patient, where x_k, a_k are pre-treatment patient covariates and the binary treatment received at stage k, and y is the patient outcome measured at one point in time (usually a larger value of y is more desirable), although y could represent some cumulation or function of previous information or patient outcomes. Denote the patient covariates and treatments up to time k as $\overline{\boldsymbol{x}}_k = (x_1, x_2, \ldots, x_k)$ and $\overline{\boldsymbol{a}}_k = (a_1, a_2, \ldots, a_k)$, respectively, and define the vector of treatment decisions from stage $k+1$ onward as $\underline{\boldsymbol{a}}_{k+1} = (a_{k+1}, a_{k+2}, \ldots, a_K)$. The patient history prior to the kth treatment decision is denoted $\boldsymbol{h}_k = (\overline{\boldsymbol{x}}_k, \overline{\boldsymbol{a}}_{k-1})$. Finally, denote the potential outcome under the treatment $\overline{\boldsymbol{a}}$ as $Y^{\overline{\boldsymbol{a}}}$, where $\overline{\boldsymbol{a}}$ is shorthand for the entire treatment decisions $\overline{\boldsymbol{a}}_K$. The objective is to find the optimal treatment regimens $\overline{\boldsymbol{d}} \equiv \overline{\boldsymbol{d}}_K = (d_1, d_2, \ldots, d_K)$ such that the expected potential outcome $\mathbb{E}(Y^{\overline{\boldsymbol{d}}})$ is maximized, where $d_k \equiv d_k(\boldsymbol{h}_k)$ is the treatment rule (a function) at time point k. To estimate an optimal DTR that maximizes the expected outcome (value function) if it were applied to the whole population, we assume the following assumptions hold:

1. SUTVA (Rubin, 1980).
2. Consistency: $Y = Y^{\overline{\boldsymbol{a}}}$ if the observed treatments are $\overline{\boldsymbol{a}}$.
3. Positivity: $\prod_{k=1}^{K} \mathbb{P}\{d_k(\boldsymbol{h}_k) \,|\, \boldsymbol{h}_k\} > 0$ almost surely.
4. Sequential exchangeability (Robins, 1997): for any possible treatment regimes $\overline{\boldsymbol{d}}_k = (d_1, d_2, \ldots, d_k)$, the stage k treatment is independent of future potential covariates or

outcome conditional on current patient history, i.e.,

$$(\boldsymbol{x}_{k+1}^{\overline{\boldsymbol{d}}_k}, \boldsymbol{x}_{k+2}^{\overline{\boldsymbol{d}}_{k+1}}, \ldots, \boldsymbol{x}_{K}^{\overline{\boldsymbol{d}}_{K-1}}, y^{\overline{\boldsymbol{d}}}) \perp\!\!\!\perp A_k \,|\, \boldsymbol{H}_k = \boldsymbol{h}_k,$$

where $\boldsymbol{d}$ is any treatment regimen that satisfies the positivity condition.

In a K-stage DTR, the aim is to find the optimal treatment regimen $\boldsymbol{d}$ such that the value $\mathbb{E}(Y^{\boldsymbol{d}})$ is maximized, i.e., $\boldsymbol{d}^{opt} = \arg\max_{\boldsymbol{d}} \mathbb{E}(Y^{\boldsymbol{d}})$. The value $\mathbb{E}(Y^{\boldsymbol{d}})$ can be estimated using the following Lemma.

Lemma 2.3.1 (**Robins's G-computation** (Robins, 1986))**.** *For any treatment regimen* $\overline{\boldsymbol{d}} \equiv \overline{\boldsymbol{d}}_K = (d_1, d_2, \ldots, d_K)$ *satisfies Assumptions 1-4, then* $\mathbb{E}(Y^{\overline{\boldsymbol{d}}})$ *can be written as*

$$\mathbb{E}(Y^{\overline{\boldsymbol{d}}}) = \mathbb{E}\{\mathbb{E}\{\ldots\mathbb{E}\{\mathbb{E}\{Y|\boldsymbol{h}_K, d_K\}|\boldsymbol{h}_{K-1}, d_{K-1}\}\ldots|\boldsymbol{x}_1, d_1\}\}.$$

Note, though, that to actually implement g-computation typically requires correct (parametric) specification of the mean model for Y, and all sorts of intermediate expectations.

Knowing how to estimate the value function under a given regimen $\boldsymbol{d}$, now I illustrate how to estimate the optimal treatment regimen. Define the blip function at stage k as the causal effect between those with treatment a_k at stage k and those who received a reference treatment, with the same history $\boldsymbol{h}_k$ and assuming they receive optimal treatment after kth stage:

$$\gamma_k(\boldsymbol{h}_k, a_k) = \mathbb{E}(Y^{\overline{\boldsymbol{a}}_k, \underline{\boldsymbol{a}}_{k+1}^{opt}}|\boldsymbol{H}_k = \boldsymbol{h}_k) - \mathbb{E}(Y^{\overline{\boldsymbol{a}}_{k-1}, a_k=0, \underline{\boldsymbol{a}}_{k+1}^{opt}}|\boldsymbol{H}_k = \boldsymbol{h}_k),$$

where $a_k = 0$ is a reference treatment; in fact, this is referred to as the "optimal blip-to-zero function" (Robins, 2004). The following Theorem states that to maximize $\mathbb{E}(Y^{\boldsymbol{d}})$, it is sufficient to maximize the blip function at each stage.

Theorem 2.3.1 (Murphy (2003))**.** *Assume that* $\mathbb{E}(|Y|\,|\boldsymbol{H}_K, A_K)$ *is bounded almost surely.*

Then

$$\begin{aligned}\mathbb{E}(Y^{\boldsymbol{d}^{opt}}) &= \max_{\boldsymbol{d}} \mathbb{E}(Y^{\boldsymbol{d}}) \\ &= \max_{\boldsymbol{d}} \mathbb{E}\{\mathbb{E}\{\ldots\mathbb{E}\{\mathbb{E}\{Y|\boldsymbol{h}_K, d_K\}|\boldsymbol{h}_{K-1}, d_{K-1}\}\ldots|\boldsymbol{x}_1, d_1\}\} \\ &= \mathbb{E}\{\mathbb{E}\{\ldots\mathbb{E}\{\mathbb{E}\{Y|\boldsymbol{h}_K, d_K^{opt}\}|\boldsymbol{h}_{K-1}, d_{K-1}^{opt}\}\ldots|\boldsymbol{x}_1, d_1^{opt}\}\},\end{aligned}$$

where d_k^{opt} *at each stage is computed as*

$$\begin{aligned}d_k^{opt} &= \arg\max_{a_k} \mathbb{E}\{\ldots\mathbb{E}\{\mathbb{E}\{Y|\boldsymbol{h}_K, d_K^{opt}\}|\boldsymbol{h}_{K-1}, d_{K-1}^{opt}\}\ldots|\boldsymbol{h}_k, a_k\} \\ &= \arg\max_{a_k} \mathbb{E}(Y^{\overline{\boldsymbol{a}}_{k-1}, a_k, \underline{\boldsymbol{d}}_{k+1}^{opt}}|\boldsymbol{h}_k) \qquad (2.3) \\ &= \arg\max_{a_k} \gamma_k(\boldsymbol{h}_k, a_k).\end{aligned}$$

By Theorem 2.3.1, to obtain the optimal treatment regimen, it is sufficient to compute expression (2.3) at each stage (recursively). Moreover, by the definition of the blip function, we have $d_k^{opt} = \arg\max_{a_k} \gamma_k(\boldsymbol{h}_k, a_k)$. That is, it suffices to estimate the blip function in each stage to find the optimal treatment regimen. The following Theorem 2.3.2 can be used to estimate the blip parameter.

Theorem 2.3.2 (Robins (2004))**.** *For treatment regimen* $\overline{\boldsymbol{d}}^{opt}$ *satisfies Assumptions 3 and 4,* $\mathbb{E}(Y^{\overline{\boldsymbol{a}}_k, \underline{\boldsymbol{d}}_{k+1}^{opt}}|\boldsymbol{h}_k, a_k) = \mathbb{E}(\widetilde{Y}_k|\boldsymbol{h}_k, a_k)$*, where* $\widetilde{Y}_k = Y + \sum_{m=k+1}^{K} \mu_m(\boldsymbol{h}_m, a_m)$ *and* $\mu_m(\boldsymbol{h}_m, a_m) = \gamma_m(\boldsymbol{h}_m, d_m^{opt}) - \gamma_m(\boldsymbol{h}_m, a_m)$*.*

By Theorem 2.3.2, $\mathbb{E}(\widetilde{Y}_k|\boldsymbol{h}_k, a_k) = \mathbb{E}(Y^{\overline{\boldsymbol{a}}_{k-1}, 0, \underline{\boldsymbol{d}}_{k+1}^{opt}}|\boldsymbol{h}_k, a_k) + \gamma_k(\boldsymbol{h}_k, a_k) = f_k(\boldsymbol{h}_k) + \gamma_k(\boldsymbol{h}_k, a_k)$, where $f_k(\boldsymbol{h}_k) \equiv \mathbb{E}(Y^{\overline{\boldsymbol{a}}_{k-1}, 0, \underline{\boldsymbol{d}}_{k+1}^{opt}}|\boldsymbol{h}_k, a_k)$ is the baseline mean model at stage k, which is irrelevant for making treatment decisions (nuisance model). Hence our parameter of interest is the blip parameter at each stage. The blip parameter $\boldsymbol{\psi}_k$ at stage k can be obtained using

A-learning estimating equation:

$$\sum_{i=1}^{n} \boldsymbol{h}_{ik}(a_{ik} - \widehat{\pi}_{ik}) \left(\widetilde{y}_{ik} - f_k(\boldsymbol{h}_{ik}; \widehat{\boldsymbol{\beta}}_k) - \gamma_k(\boldsymbol{h}_{ik}, a_{ik}; \boldsymbol{\psi}_k) \right) = \mathbf{0},$$

where $\widetilde{y}_k$ is the "pseudo" outcome defined in Theorem 2.3.2 and can be estimated using

$$\widetilde{y}_k = y + \sum_{m=k+1}^{K} \widehat{\mu}_m(\boldsymbol{h}_m, a_m) = y + \sum_{m=k+1}^{K} \left\{ \gamma_m \left(\boldsymbol{h}_m, \widehat{d}_m^{opt}; \widehat{\boldsymbol{\psi}}_m \right) - \gamma_m \left(\boldsymbol{h}_m, a_m; \widehat{\boldsymbol{\psi}}_m \right) \right\},$$

where $\widehat{d}_m^{opt} = \arg\max_{a_m} \gamma_m(\boldsymbol{h}_m, a_m; \widehat{\boldsymbol{\psi}}_m)$.

2.4 Variable Selection in Regression: Regularization

In this section, I briefly introduce some methods of variable selection in the regression setting. Throughout this section, we assume that the outcome $\boldsymbol{Y} \in \mathbb{R}^n$ and each predictor $X_j \in \mathbb{R}^n$ is centered, for $j = 1, 2, \ldots, p$, so that the intercept can be eliminated; and we consider the following model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the standard Gaussian error term.

In the regression setting, when the number of predictors is less then the number of observations, i.e., $n > p$, the regression coefficients can be estimated using least squares:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

where $\|\cdot\|_2$ is the ℓ_2 norm. However, when $p > n$ or there exists multicollinearity in the design matrix, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is singular, hence the solution $\widehat{\boldsymbol{\beta}}$ is not unique. To overcome this kind of situation, and moreover, to simplify the model and enhance the predictability, as well as to select significant variables to improve the interpretability, statisticians use the idea of regularization (or penalization) by adding a penalty term to the loss function. Now

the resulting estimator can be found by solving the following task:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + n\lambda\rho(|\boldsymbol{\beta}|)\}, \tag{2.4}$$

where λ is the tuning parameter used to control sparsity, and $\rho(|\cdot|)$ is some penalty function. A larger value of λ tends to shrink the coefficients to zero. In what follows we provide several different examples of penalty functions and briefly describe their properties.

2.4.1 Variable Selection Consistency and Oracle Property

Let s be the number of non-zero components of the underlying true parameter $\boldsymbol{\beta}$ and denote S as the set of indices of non-zero components for $\boldsymbol{\beta}$, i.e., $S = \text{supp}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$. In the variable selection literature, the performance of the variable selection procedures are usually measured by the false negative rate (the proportion of times a method wrongly removed a truly important variable) and the false positive rate (the proportion of times a method wrongly included a non-important variable). A variable selection technique is said to have variable selection consistency if it has no false exclusion and no false inclusion, i.e., $\lim_{n\to\infty} P(\text{supp}(\widehat{\boldsymbol{\beta}}) = S) = 1$.

Another property closely related to variable consistency is the oracle property (Fan and Li, 2001). Define the oracle estimator as $\widehat{\boldsymbol{\beta}}_{ora} = \arg\min_{\boldsymbol{\beta}_S} \|\boldsymbol{Y} - \boldsymbol{X}_S\boldsymbol{\beta}_S\|^2$, that is, assume that the true support of the $\boldsymbol{\beta}$ was known in advance. Note that the oracle estimator is just a conceptual idea that is used to help develop theoretical properties for variable selection methods. An estimator is said to have the oracle property if it performs as well as the oracle estimator. By the definition, oracle property implies the variable selection consistency property.

2.4.2 Commonly Used Penalty Functions

I now introduce several commonly used penalty functions $\rho(|\cdot|)$ in Expression (2.4).

Ridge Penalty (Tikhonov regularization) (Hoerl and Kennard, 1970): $\rho(|\boldsymbol{\beta}|) = \|\boldsymbol{\beta}\|_2^2$. In the regression setting, the ridge estimator has an analytical solution: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, where I_p is the identity matrix of size p. Because of the extra term λI_p, the expression $(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p)$ is invertible even in the scenario that n is smaller than p. However, the solution of ridge regression cannot yield sparsity, i.e., estimators cannot be shrunk to 0, thus it cannot perform variable selection.

The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) uses the ℓ_1 penalty: $\rho(|\boldsymbol{\beta}|) = \|\boldsymbol{\beta}\|_1$. Because ℓ_1 norm is non-differentiable at the origin, LASSO can produce sparse solutions (Fan and Li, 2001). The LASSO can be solved by efficient algorithms such as coordinate descent (Hastie et al., 2010), and Nesterov accelerated gradient-based method (Beck and Teboulle, 2009). However, LASSO has some limitations: first, it cannot handle highly correlated variables well (Zou and Hastie, 2005); second, the LASSO estimator has a bias: in the case that the design matrix are orthonormal, the resulting estimator will be shifted by a constant λ (Fan and Li, 2001).

The asymptotic theory for the LASSO has been studied in Fu and Knight (2000) when the dimension p is fixed, where they showed that the LASSO estimator is consistent under some assumptions. The non-asymptotic property of LASSO has been studied in Bickel et al. (2009), where they showed that under the restricted eigenvalue (or restricted strong convexity) condition, the LASSO error bound satisfies $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq 3\lambda\sqrt{ns}/\gamma$ for some $\gamma > 0$, given that the tuning parameter satisfies $\lambda \geq 2\|\boldsymbol{X}^T\boldsymbol{\varepsilon}/n\|_\infty$; the prediction error $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n$ is bounded by the quantity $9s\lambda^2/\gamma$ (Hastie et al., 2019). As for variable selection performance, Zhao and Yu (2006) and Meinshausen and Bühlmann (2006) showed that under the mutual incoherence (or irrepresentable) condition, LASSO achieved variable selection consistency; however, due to the bias induced by the tuning parameter λ, LASSO does not have the desired oracle property.

The Elastic Net (Zou and Hastie, 2005) combines the ℓ_1 and ℓ_2 penalties:

$$\rho(|\boldsymbol{\beta}|) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2,$$

where $\alpha \in [0,1]$ is a tuning parameter used to control the balance between the ℓ_1 and ℓ_2 norm, and thus can produce sparsity while offering good performance even when the features are highly correlated. By the definition, the LASSO is the special case of the elastic net when $\alpha = 1$. When $\alpha = 0$, the elastic net reduces to the ridge penalty.

The Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001) is a non-convex penalty : $\rho'(\beta) = \lambda\{I(\beta \leq \lambda) + \frac{(z\lambda-\beta)_+}{(z-1)\lambda} I(\beta > \lambda)\}$, for some $\beta > 0$ and $z > 2$, which can produce sparse solutions and nearly unbiased estimators. However, the objective function is now non-convex, which makes the optimization problem more challenging than the ℓ_1-based penalty. Fan and Li (2001) showed that the SCAD estimator has the oracle property: as $n \to \infty$, $\widehat{\boldsymbol{\beta}}_{S^c} = 0$ and $\sqrt{n}(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) \to_d \mathcal{N}(0, \boldsymbol{I}_S(\boldsymbol{\beta}))$, where the matrix $\boldsymbol{I}_S \in \mathbb{R}^{s\times s}$ is the fisher information matrix for $\boldsymbol{\beta}_S$.

The Adaptive LASSO (Zou, 2006): $\rho(|\boldsymbol{\beta}|) = \sum_{j=1}^p \widehat{w}_j|\beta_j|$, where $\widehat{w}_j$ is the estimated adaptive weights constructed using some initial estimator, e.g., $\widehat{w}_j = 1/|\widehat{\beta}_j^{ols}|$, where the $\widehat{\beta}_j^{ols}$ is the ordinary least square (OLS) estimators. In this way, the coefficients are not forced to be equally penalized in the ℓ_1 penalty. As n goes to infinity, the weights corresponding to unimportant variables go to infinity, which puts a large penalty on those variables, and the weights corresponding to important variables converge to a finite constant. Thus, small coefficients are removed, and large coefficients are unbiasedly estimated.

Zou (2006) proved that under the fixed p scenario, the adaptive LASSO also has the oracle property. In the setting that the number of predictors are larger than the number of observations, the ordinary least squares estimators are not unique; Zou (2006) proposed to use the ridge penalty to obtain the initial estimator. An alternate approach to obtain the initial estimator is to use the marginal regression, i.e., the outcome Y is regressed separately on

each predictor X_j. Huang et al. (2008) showed that under some conditions, the adaptive LASSO still enjoys the oracle property even if the number of predictors p is much larger than the sample size.

2.4.3 Tuning Parameter Selection

The choice of tuning parameter λ in Expression (2.4) plays a key role in the performance of regularized regression: an inappropriately large or small value of λ will greatly weaken the performance of the resulting estimator with respect to the estimation error, variable selection results and prediction error. In this subsection, we briefly discuss different techniques to select the tuning parameter in regularized regression.

$\mathcal{K}$-fold Cross-validation: Cross-validation is argued to be the most widely used tuning parameter/model selection technique in modern statistics (Hastie et al., 2009). The procedure is described as follows:

1. Split the dataset into $\mathcal{K}$ subsets randomly, such that in each subsample, the sample size is $n/\mathcal{K}$.

2. Choose one from the $\mathcal{K}$ subsample and use it as the test set. The other $(\mathcal{K}-1)$ subsets are used as training data. Compute the test error for each candidate model (equivalently, for each λ).

3. Repeat the last step $(\mathcal{K}-1)$ times such that each subsample has been used once as the test set.

4. Average the test error over all the subsamples for all the models, and choose the model that has the smallest average test error.

In practice, the typical choices of $\mathcal{K}$ is 5 or 10. An interesting special case is the n-fold cross-validation, which is also known as the leave-one-out cross-validation. The average test error is calculated as $n^{-1}\sum_{i=1}^{n}(y_i - \widehat{y}_{-i})^2$, where $\widehat{y}_{-i}$ is the fitted value of observation

i using training data without the ith observation. Golub et al. (1979) showed that the leave-one-out cross-validation error can be approximated using the formula $n^{-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2(1-p/n)^{-2}$, which can greatly reduce the burden of the computation of cross-validation. This approximation formula is called the generalized cross-validation, and the best model is selected in a way such that its corresponding generalized cross-validation error is the lowest among the candidate models. Next I introduce two popular information-based approaches to select the tuning parameter, which do not require the resampling procedure and hence can reduce the computational burden substantially.

Akaike information criterion (AIC) (Akaike, 1974) is defined as AIC $= -2l(\widehat{\boldsymbol{\theta}}) + 2p$, where $l(\widehat{\boldsymbol{\theta}})$ is the maximum log-likelihood of the model. In the linear regression setting, the AIC can be written as $\log(\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2/n)+2p/n$. Stone (1977) showed that the AIC is asymptotically equivalent to the leave-one-out cross-validation.

Similar to AIC, the Bayesian information criterion (BIC) (Schwarz, 1978) is defined as $-2l(\widehat{\boldsymbol{\theta}}) + p\log n$. A comparison of AIC and BIC can be found in Anderson and Burnham (2004) and Yang (2005). In summary, if the true model is contained in the candidate models, then using BIC will select the true model almost surely when the sample size goes to infinity, which is a property that AIC does not have. On the other hand, in the case that the true model is not contained in the candidate models, AIC outperforms BIC with respect to model selection. Other examples of information criterion include Mallows's C_p (Mallows, 2000) and risk inflation criterion (Foster and George, 1994).

2.5 Variable Selection for Interaction Models

As discussed in Section 2.3, DTRs are often estimated from models which include interactions between treatment and a number of tailoring variables. In this section, I briefly discuss penalized regression methods for interaction models.

Consider the following model:

$$\boldsymbol{Y} = \psi_0 \boldsymbol{A} + \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j + \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ \boldsymbol{X}_j) + \boldsymbol{\varepsilon},$$

where "$\circ$" is the element-wise vector multiplication. A straightforward way to perform variable selection in this interaction model is to add one of the penalty functions discussed in Section 2.4 to the loss function. However, this may yield a model in which the estimated interaction term is nonzero while the corresponding main effects are zero, i.e., $\psi_j \neq 0$ while β_j is shrunk to zero. Although there may be situations in which this could happen, Cox (1984) argued that in an interaction model, the main effect term should always be selected into the model before its corresponding interaction term; this is the so-called strong heredity assumption (Chipman, 1996). Later we will see that the strong heredity assumption plays a key role of developed methodological works in Chapters 3-5. As such, variable selection with hierarchical structure is also important to consider. Indeed, this area is quite active and well developed, see, for example, Zhao et al. (2009); Radchenko and James (2010); Bien et al. (2013); Lim and Hastie (2015); Haris et al. (2016); Bhatnagar et al. (2020). Many of the authors listed here induced hierarchical sparsity based on the group LASSO penalty, but there were exceptions. Below I review a technique proposed in Choi et al. (2010), where they use a simple reparametrization to achieve the strong heredity property.

The idea is to first introduce a new set of parameters $\boldsymbol{\tau} = (\tau_1, \tau_2, ...\tau_p)$ and reparametrize the coefficients for the interaction terms ψ_j as a function of τ_j and the main effect parameters β_j and ψ_0 such that $\psi_j = \psi_0 \tau_j \beta_j$. In this way, strong heredity can be met, and the following loss function can be considered:

$$\mathcal{L}(\boldsymbol{Y}; \boldsymbol{\theta}) = \left\| \left(\boldsymbol{Y} - \psi_0 \boldsymbol{A} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j - \sum_{j=1}^{p} \underbrace{\psi_0 \tau_j \beta_j}_{\psi_j} (\boldsymbol{A} \circ \boldsymbol{X}_j) \right) \right\|_2^2,$$

where $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \psi_0, \tau_1, \ldots, \tau_p)$. The penalized estimator can be obtained by solving

the minimization task $\widehat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{Y}; \boldsymbol{\theta}) + \lambda(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\tau}\|_1)$.

2.6 Variable Selection in DTRs

The study of DTRs, often characterized by situations in which there are many covariates and a complex disease process for which competing treatment choices may have heterogeneous effects, it is difficult to know which prognostic factors might be relevant to tailoring treatment. Tailoring medications based on patients' characteristics could improve symptoms. However, tailoring treatment may result in an overly complex treatment rule making it difficult for treating physicians to assess or implement. Therefore, a more data-driven approach of selecting these covariates might simplify and improve the estimated decision rules, and variable selection with the objective of optimizing patients' outcome by identifying useful tailoring variables is important (Lu et al., 2013; Shi et al., 2018; Jeng et al., 2018).

Among the earliest work on selecting tailoring variables in DTRs, Lu et al. (2013) adopted an adaptive LASSO (Zou, 2006) approach within the A-learning framework. The blip parameters can be estimated by minimizing the objective function

$$\|\boldsymbol{Y} - f(\boldsymbol{x}; \boldsymbol{\beta}) - \text{diag}(\boldsymbol{A} - \widehat{\boldsymbol{\pi}})\boldsymbol{X}\boldsymbol{\psi}\|^2 + \lambda \sum_{j=1}^{p} \rho(|\psi_j|).$$

They showed that with a suitable choice of tuning parameter, the blip parameter estimators are approximately unbiased in the penalized framework. Jeng et al. (2018) extended the work of Lu et al. (2013) to the case where p is allowed to grow with the sample size n, and Shi et al. (2016) generalized it to the situation where p is of the non-polynomial order of the n, i.e., $\log p = O(n^q)$ for some $q < 1$.

In singly-robust settings, Song et al. (2015) added the SCAD penalty to the value search method of outcome weighted learning to incorporate sparsity and estimate the optimal treatment decision, where the treatment rule can be viewed as a classification problem. Liang

et al. (2017) added LASSO penalty into their proposed convex surrogate loss function to select the variables. Shi et al. (2018) used the Dantzig selector (Candes and Tao, 2007) to directly penalize the estimating equations of A-learning. Zhu et al. (2019) combined Q-learning with SCAD penalty and conducted the hard-thresholding method proposed in (Moodie and Richardson, 2010) to tackle the challenge of nonregularity. Wu et al. (2021b) proposed a penalized single-index model in a high-dimensional semiparametric framework to select the useful tailoring variables.

All the variable selection methods outlined above are designed for use in DTR estimation. The differences between these approaches and the variable selection methods in prediction setting are twofold. First, the goal differs: variable selection for DTRs aims to improve the estimated decision rules rather than enhance the predictive power of the outcome (of course, both emphasize the interpretability of the model). Second, in DTRs, interest is primarily in effect modification, and hence the selection tends to be focused on the terms in the blip model. In contrast, variable selection approaches in prediction setting usually seek to choose from among all variables with no special status given to effect modifiers.

2.7 Summary

In this literature review, I first discussed some of the key components and assumptions for causal inference. Then I described some estimation methods in observational studies for causal inference and how to generalize them to a DTR setting. Finally, some penalized variable selection methods for main effects, interaction models, and DTRs were reviewed.

Chapter 3

Variable Selection in Regression-based Estimation of Dynamic Treatment Regimes

Preamble to Manuscript 1.

Doubly robust variable selection method for optimal treatment decision-making has been studied in (Shi et al., 2018), where the authors used the Dantzig selector (Candes and Tao, 2007) to penalize the A-learning estimating function. Motivated by this work, we extend the doubly robust DTRs estimation method dynamic weighted ordinary least squares regression (Wallace and Moodie, 2015) to a penalized framework, where estimation and variable selection for DTRs can be conducted simultaneously, and the the double robustness property is inherited from dynamic weighted ordinary least squares regression.

The corresponding manuscript was published in *Biometrics* (Bian et al., 2021), please see https://doi.org/10.1111/biom.13608.

Variable Selection in Regression-based Estimation of Dynamic Treatment Regimes

Zeyu Bian[1], Erica EM Moodie[1], Susan M Shortreed[2,3] and Sahir Bhatnagar[1,4]

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2]*Kaiser Permanente Washington Health Research Institute*

[3]*Department of Biostatistics, University of Washington*

[4]*Department of Diagnostic Radiology, McGill University*

This thesis contains the accepted version of the corresponding paper published in *Biometrics* (Bian et al., 2021).

© Copyright Wiley, 2021 on behalf of International Biometric Society.

Abstract

Dynamic treatment regimes (DTRs) consist of a sequence of decision rules, one per stage of intervention, that aim to recommend effective treatments for individual patients according to patient information history. DTRs can be estimated from models which include interactions between treatment and a (typically small) number of covariates which are often chosen a priori. However, with increasingly large and complex data being collected, it can be difficult to know which prognostic factors might be relevant in the treatment rule. Therefore, a more data-driven approach to select these covariates might improve the estimated decision rules and simplify models to make them easier to interpret. We propose a variable selection method for DTR estimation using penalized dynamic weighted least squares. Our method has the strong heredity property, that is, an interaction term can be included in the model only if the corresponding main terms have also been selected. We show our method has both the double robustness property and the oracle property theoretically; and the newly proposed method compares favorably with other variable selection approaches in numerical studies. We further illustrate the proposed method on data from the Sequenced Treatment Alternatives to Relieve Depression study.

3.1 Introduction

Dynamic treatment regimes (DTRs) (Chakraborty and Moodie, 2013), or adaptive treatment strategies, consist of a sequence of decision rules that aim to improve individual patients' health outcomes by tailoring medical treatment to each patient's information. Statistical methods can be used to identify optimal DTRs, constructing treatment rules tailored over time to individual's information that can optimize the expected patient outcome.

DTRs can be estimated from models that include interactions between treatment and covariates, which are often chosen a priori. However, with many covariates and a complex disease process, for which competing treatment choices have heterogeneous effects, it is difficult to know which prognostic factors might be considered relevant in the treatment rule. A more data-driven approach of selecting these covariates might improve the estimated decision rules and simplify models to improve tractability. We are motivated by the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study (Fava et al., 2003), a randomized multistage trial that aimed to determine optimal treatments for patients with major depressive disorder. With many of covariates such as demographic and clinical characteristics collected throughout the study, it is challenging to select covariates useful for tailoring treatment from among so many based on expert knowledge only. Thus, variable selection with the objective of optimizing individualized treatment decisions becomes important.

Much of the DTR literature focuses on estimation; variable selection with the objective of optimizing treatment decisions has been considered only occasionally. Gunter et al. (2011) proposed a ranking method for variable selection in DTRs. Based on this approach, Fan et al. (2016) developed the sequential advantage selection approach, which considers variables already in the model when deciding whether to include a new variable by the additional improvement provided by this variable. Lu et al. (2013) adopted adaptive LASSO (Zou, 2006) in the context of A-learning (Murphy, 2003), Shi et al. (2018) proposed a method which used the Dantzig selector directly to penalize the estimating equations of A-learning

and has the double robust property, that is, the estimators are consistent if either one of two nuisance models is correct. The topic of variable selection in a general (not DTR) context has seen many innovations (e.g., Tibshirani, 1996; Fan and Li, 2001). Gunter et al. (2011) noted that most variable selection approaches focus on predictive performance, and thus may not perform well in DTRs as these techniques may underestimate the importance of variables that have small predictive ability but that play a significant role in decision making.

In this article, we follow the DTR estimation approach of dynamic ordinary least squares regression (dWOLS) introduced by Wallace and Moodie (2015), an approach which requires only some minor pre-computation and the implementation of standard weighted regression. While having similarities to both Q-learning (Watkins, 1989) and G-estimation (Robins, 2004), it provides simplicity and intuitiveness similar to the former and benefits from the double robustness of the latter although it is suitable only for linear decision rules. By adding two penalty terms in the dWOLS model, we perform estimation and variable selection for DTRs simultaneously. The rest of this article is organized as follows. In Section 2, we introduce the proposed penalized dWOLS (pdWOLS) approach, followed by algorithmic details and theoretical properties. Three simulation studies are given in Section 3. Finally, we apply our method to the STAR*D trial data in Section 4.

3.2 Methodology

3.2.1 Introductory Concepts and Notation

We make assumptions to proceed with estimation of DTRs: (1) Stable unit treatment value assumption (SUTVA) (Rubin, 1980): a patient's potential outcome is not affected by other patients' treatment assignments. (2) Ignorability: ignorability or no unmeasured confounding (Robins, 1997) specifies that for any possible treatment regimes, the stage k treatment is independent of future potential covariates or outcome conditional on current patient history. (3) No interference, no measurement error, and all the individuals have complete

follow-up.

We adopt the setup of Wallace and Moodie (2015). For a K-stages DTR, the following notation is used, with lowercase being used for observed variables and uppercase for their random counterparts: y denotes patient outcome (continuous) which is measured at one point in time. The goal of DTRs is to make treatment decisions that can optimize (typically, maximize) the outcome. The kth binary treatment decision is, e.g., $a_k = 1$ for treatment, $a_k = 0$ for standard care. Patient information available at time k and prior to kth treatment decision is denoted $\boldsymbol{x}_k$. The covariate matrix containing patient history prior to the kth treatment decision is denoted $\boldsymbol{h}_k$; this history can include previous treatments $a_1, \ldots, a_{k-1}$. Finally, $\overline{\boldsymbol{a}}_k = (a_1, a_2, \ldots, a_k)$ is the vector of the first k treatment decisions, and $\underline{\boldsymbol{a}}_k = (a_{k+1}, a_{k+2}, \ldots, a_K)$ is the vector of treatment decisions from stage $k+1$ onward.

The blip (or contrast) function is defined as the difference in expected potential outcome between patients who received treatment a_k at stage k and patients who received a reference treatment denoted, say $a_k = 0$, with the same history and assuming they receive optimal treatment after kth stage:

$$\gamma_k\left(\boldsymbol{h}_k, a_k\right) = \mathbb{E}\left[Y^{\overline{\boldsymbol{a}}_k, \underline{\boldsymbol{a}}_{k+1}^{opt}} - Y^{\overline{\boldsymbol{a}}_{k-1}, a_k=0, \underline{\boldsymbol{a}}_{k+1}^{opt}} | \boldsymbol{H}_k = \boldsymbol{h}_k\right].$$

The regret function (Murphy, 2003) is the expected loss resulting from giving treatment a_k at stage k instead of the optimal treatment a_k^{opt}, assuming optimal treatment is received after k-th stage: $\mu_k\left(\boldsymbol{h}_k, a_k\right) = \mathbb{E}\left[Y^{\overline{\boldsymbol{a}}_{k-1}, \underline{\boldsymbol{a}}_k^{opt}} - Y^{\overline{\boldsymbol{a}}_k, \underline{\boldsymbol{a}}_{k+1}^{opt}} | \boldsymbol{H}_k = \boldsymbol{h}_k\right]$. The blip and regret functions correspond directly: $\mu_k\left(\boldsymbol{h}_k, a_k\right) = \gamma_k\left(\boldsymbol{h}_k, a_k^{opt}\right) - \gamma_k\left(\boldsymbol{h}_k, a_k\right)$. This can be leveraged to simplify some expressions in later sections. Finally, we decompose the expected mean outcome into two components: $\mathbb{E}\left[Y^a | \boldsymbol{H} = \boldsymbol{h}; \boldsymbol{\beta}, \boldsymbol{\psi}\right] = f\left(\boldsymbol{h}_0; \boldsymbol{\beta}\right) + \sum_{k=1}^{K} \gamma_k\left(\boldsymbol{h}_k, a_k; \boldsymbol{\psi}_k\right)$, where $f\left(\boldsymbol{h}_0; \boldsymbol{\beta}\right)$ and $\gamma_k\left(\boldsymbol{h}_k, a_k; \boldsymbol{\psi}_k\right)$ are the so-called treatment-free and blip models, respectively, and $\boldsymbol{h}_0$ are baseline covariates. The function f, being free of any terms relating to the active treatment ($a_k = 1$), is irrelevant for making decisions about optimal treatment selection. For instance,

in a simple one-stage setting, we could assume that both f and γ are linear in form: $f(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x$ and $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \psi_1 x)$, and hence the estimated optimal treatment is $\widehat{a}^{opt} = I(\widehat{\psi}_0 + \widehat{\psi}_1 x > 0)$ where $I(\cdot)$ is the indicator function.

3.2.2 Dynamic weighted ordinary least squares

Dynamic weighted ordinary least squares uses a sequential regression approach, similar to estimate the blip parameter $\boldsymbol{\psi}_k$ in the model for $\mathbb{E}\left[Y^a | \boldsymbol{H} = \boldsymbol{h}; \boldsymbol{\beta}, \boldsymbol{\psi}\right]$, achieving double robustness through weighting by a function of the propensity score (Rosenbaum and Rubin, 1983). The weights must satisfy $\pi(\boldsymbol{x})w(1, \boldsymbol{x}) = (1 - \pi(\boldsymbol{x}))w(0, \boldsymbol{x})$, where $\pi(\boldsymbol{x})$ is the propensity score and $w(a, \boldsymbol{x})$ is the weight for a subject with treatment a and covariates $\boldsymbol{x}$. Wallace and Moodie (2015) suggested to use "absolute value" weights of the form $w(a, \boldsymbol{x}) = |a - \mathbb{E}[A|\boldsymbol{X} = \boldsymbol{x}]|$, as these offered better efficiency than other alternatives considered, while yielding consistent estimators of blip parameters if either the treatment or treatment-free model is correctly specified. Another assumption required by dWOLS is that the treatment-free model must include the main effects for all covariates in the blip model (unlike G-estimation, which can use an intercept-only treatment-free model). Violation of this assumption, known as the strong heredity principle (Chipman, 1996), can lead to biased estimators of blip parameters.

3.2.3 Penalized dWOLS

We first introduce our approach in a one-stage setting with a continuous outcome, letting

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \psi_0 \boldsymbol{A} + \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j + \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ \boldsymbol{X}_j) + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\mathbf{1}$ is the vector of 1's, $\boldsymbol{Y} \in \mathbb{R}^n$ is a continuous response measured on n individuals, $\boldsymbol{X}_j \in \mathbb{R}^n$ are the j-th covariates, $\boldsymbol{X}_i \in \mathbb{R}^p$ are covariates of i-th individual, $\beta_j \in \mathbb{R}$ are the

corresponding parameters for the main effects of covariates, $\psi_j \in \mathbb{R}$ are the blip parameters for $j = 0, 1, \ldots, p$, $\boldsymbol{A}$ is the binary treatment indicator, "$\circ$" is the element wise vector multiplication, and $\boldsymbol{\varepsilon}$ is an error term. This model is a simplification of (Bhatnagar et al., 2020), which considers an additive interaction regression model. In this posited model, the treatment-free model is $\boldsymbol{\beta}_0 + \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j$ and the blip model is $\psi_0 \boldsymbol{A} + \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ \boldsymbol{X}_j)$. To eliminate the intercept β_0, throughout this section, we center the response variable and each input variable in a weighted way, e.g., using $\boldsymbol{Y} - \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$ instead of $\boldsymbol{Y}$ as the outcome.

For a continuous response we use the weighted squared-error loss:

$$\mathcal{L}(\boldsymbol{Y}; \boldsymbol{\theta}) = \frac{1}{2n} \left\| \sqrt{\boldsymbol{W}} \left(\boldsymbol{Y} - \psi_0 \boldsymbol{A} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j - \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ \boldsymbol{X}_j) \right) \right\|_2^2,$$

where $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \psi_0, \ldots, \psi_p)$, and $\boldsymbol{W} = \operatorname{diag}\{w_1(a, \boldsymbol{x}), w_2(a, \boldsymbol{x}), \ldots, w_n(a, \boldsymbol{x})\}$ is a *known* $n \times n$ diagonal matrix with $w_i(a, \boldsymbol{x})$ the "absolute value" weight for the ith individual. Similar to LASSO, we consider the following objective function that includes the ℓ_1 penalty for variable selection:

$$Q(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{Y}; \boldsymbol{\theta}) + \lambda(1 - \alpha)\|\boldsymbol{\beta}\|_1 + \lambda\alpha\|\boldsymbol{\psi}\|_1, \tag{3.2}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ and $\boldsymbol{\psi} = (\psi_1, ..., \psi_p)$, $\lambda > 0$ and $\alpha \in (0, 1)$ are tuning parameters, and the solution is given by $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})$. The parameter α controls the relative penalties for the main effects and the interaction effects. Other choices of the penalty term include the ℓ_2 penalty, the elastic net (Zou and Hastie, 2005) and the SCAD penalty. The ℓ_2 penalty yields ridge regression and hence cannot produce a sparse solution, and the ℓ_1 penalty cannot handle highly correlated variables very well (Zou and Hastie, 2005); the elastic net combines the ℓ_1 and ℓ_2 penalties, and thus can produce sparsity while offering good performance even when the features are highly correlated. The SCAD is a non-convex penalty that can produce sparse solutions and nearly unbiased estimators.

An issue with Equation (3.2) is that since no constraint is placed on the structure of the model, it is possible that an estimated interaction term is nonzero while the corresponding main effects are zero, which violates the strong heredity assumption. To remedy this, our work is built on the strong heredity assumption, a constraint that is often used in practice when estimating interaction effects. Under the strong heredity assumption, an interaction term can be estimated to be non-zero if its corresponding main effects are estimated to be non-zero, whereas a non-zero main effect does not necessarily imply a non-zero interaction term. In DTR analysis, it is most common that there are more confounders than there are potential tailoring variables. Following Choi et al. (2010), we introduce a new set of parameters $\boldsymbol{\tau} = (\tau_1, \tau_2, ...\tau_p)$ and reparametrize the coefficients for the interaction terms ψ_j as a function of τ_j and the main effect parameters β_j and ψ_0: $\psi_j = \psi_0 \tau_j \beta_j$. In this way, strong heredity can be met, and we consider the following model:

$$\mathcal{L}^*(\boldsymbol{Y}; \boldsymbol{\theta}) = \frac{1}{2n} \left\| \sqrt{\boldsymbol{W}} \left(\boldsymbol{Y} - \psi_0 \boldsymbol{A} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j - \sum_{j=1}^{p} \underbrace{\psi_0 \tau_j \beta_j}_{\psi_j} (\boldsymbol{A} \circ \boldsymbol{X}_j) \right) \right\|_2^2,$$

where now $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \psi_0, \tau_1, \ldots, \tau_p)$. This reparametrized model is nonlinear as it involves products of parameters, and the objective function is expressed as:

$$Q(\boldsymbol{\theta}) = \mathcal{L}^*(\boldsymbol{Y}; \boldsymbol{\theta}) + \lambda(1-\alpha)\|\boldsymbol{\beta}\|_1 + \lambda\alpha\|\boldsymbol{\tau}\|_1. \tag{3.3}$$

3.2.4 Algorithm Details

In this section, we describe a blockwise coordinate descent algorithm (Friedman et al., 2007) for fitting the weighted least-squares version of the model in Equation (3.3). "Blockwise" means we breakdown the optimization problem into sub-problems, i.e., we fix the interaction terms $\boldsymbol{\tau}$ and solve for the main effects ψ_0 and $\boldsymbol{\beta}$ and vice versa. Following (Hastie et al., 2010), we fix the value for the tuning parameter α and minimize the objective function over

a decreasing sequence of λ values ($\lambda_{max} > \ldots > \lambda_{min}$).

Denote the n-dimensional residual column vector $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$, where $\widehat{\boldsymbol{Y}}$ is the current fitted value of $\mathbb{E}(\boldsymbol{Y})$ under the posited model. The subgradient equations are given by

$$\frac{\partial Q}{\partial \psi_0} = -\frac{1}{n}(\boldsymbol{A} + \sum_{j=1}^{p} \tau_j \beta_j \boldsymbol{A} \circ \boldsymbol{X}_j)^\top \boldsymbol{W}\boldsymbol{R} = 0 \tag{3.4}$$

$$\frac{\partial Q}{\partial \beta_j} = -\frac{1}{n}\left(\boldsymbol{X}_j + \tau_j \psi_0 \boldsymbol{A} \circ \boldsymbol{X}_j\right)^\top \boldsymbol{W}\boldsymbol{R} + \lambda(1-\alpha)s_1 = \boldsymbol{0} \tag{3.5}$$

$$\frac{\partial Q}{\partial \tau_j} = -\frac{1}{n}\left(\psi_0 \beta_j \boldsymbol{A} \circ \boldsymbol{X}_j\right)^\top \boldsymbol{W}\boldsymbol{R} + \lambda\alpha s_2 = 0 \tag{3.6}$$

where s_1 and s_2 are subgradients of the ℓ_1-norm, i.e., $s_1 \in \text{sign}\,(\beta_j)$ if $\beta_j \neq 0$, $s_1 \in [-1, 1]$ if $\beta_j = 0$; $s_2 \in \text{sign}\,(\tau_j)$ if $\tau_j \neq 0$, $s_2 \in [-1, 1]$ if $\tau_j = 0$.

Define the partial residuals, without the jth predictor for $j = 1, \ldots, p$, as

$$\boldsymbol{R}_{(-j)} = \boldsymbol{Y} - \sum_{\ell \neq j} \boldsymbol{X}_\ell \beta_\ell - \psi_0 \boldsymbol{A} - \sum_{\ell \neq j} \tau_\ell \psi_0 \beta_\ell \left(\boldsymbol{A} \circ \boldsymbol{X}_\ell\right),$$

the partial residual without A as $\boldsymbol{R}_{(-A)} = \boldsymbol{Y} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j$ and the partial residual without the jth interaction for $j = 1, \ldots, p$, as

$$\boldsymbol{R}_{(-jA)} = \boldsymbol{Y} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j - \psi_0 \boldsymbol{A} - \sum_{\ell \neq j} \tau_\ell \psi_0 \beta_\ell \left(\boldsymbol{A} \circ \boldsymbol{X}_\ell\right).$$

From the subgradient Equations (3.4)–(3.6) we see that

$$\widehat{\psi_0} = \frac{\left(\boldsymbol{A} + \sum_{j=1}^{p} \tau_j \beta_j \left(\boldsymbol{A} \circ \boldsymbol{X}_j\right)\right)^\top \boldsymbol{W}\boldsymbol{R}_{(-A)}}{\left(\boldsymbol{A} + \sum_{j=1}^{p} \tau_j \beta_j \left(\boldsymbol{A} \circ \boldsymbol{X}_j\right)\right)^\top \boldsymbol{W} \left(\boldsymbol{A} + \sum_{j=1}^{p} \tau_j \beta_j \left(\boldsymbol{A} \circ \boldsymbol{X}_j\right)\right)}$$

$$\widehat{\beta_j} = \frac{S\left(\left(\boldsymbol{X}_j + \tau_j \psi_0 \left(\boldsymbol{A} \circ \boldsymbol{X}_j\right)\right)^\top \boldsymbol{W}\boldsymbol{R}_{-j}, n \cdot \lambda(1-\alpha)\right)}{\left(\boldsymbol{X}_j + \tau_j \psi_0 \left(\boldsymbol{A} \circ \boldsymbol{X}_j\right)\right)^\top \boldsymbol{W} \left(\boldsymbol{X}_j + \tau_j \psi_0 \left(\boldsymbol{A} \circ \boldsymbol{X}_j\right)\right)}$$

$$\widehat{\tau}_j = \frac{S\left(\left(\psi_0\beta_j\left(\boldsymbol{A}\circ\boldsymbol{X}_j\right)\right)^\top \boldsymbol{W}\boldsymbol{R}_{(-jA)}, n\cdot\lambda\alpha\right)}{\left(\psi_0\beta_j\left(\boldsymbol{A}\circ\boldsymbol{X}_j\right)\right)^\top \boldsymbol{W}\left(\psi_0\beta_j\left(\boldsymbol{A}\circ\boldsymbol{X}_j\right)\right)}$$

where $S(x,u)$ is the soft-thresholding operator defined as $S(x,u) = \text{sign}(x)(|x|-u)_+$ (x_+ is the maximum value of x and 0).

The strong heredity assumption means that finding the λ which shrinks all coefficients to 0, is reduced to finding the smallest λ such that all *main effect* coefficients are shrunk to 0. From the subgradient Equation (3.5), we see that $\beta_j = 0$ is a solution if

$$\left|\frac{1}{n}\left(\boldsymbol{X}_j + \tau_j\psi_0\left(\boldsymbol{A}\circ\boldsymbol{X}_j\right)\right)^\top \boldsymbol{R}_{(-j)}\right| \leq \lambda(1-\alpha).$$

From the subgradient Equation (3.6), we see that $\tau_j = 0$ is a solution if

$$\left|\frac{1}{n}\left(\psi_0\left(\boldsymbol{A}\circ\boldsymbol{X}_j\right)\beta_j\right)^\top \boldsymbol{R}_{(-jA)}\right| \leq \lambda\alpha.$$

Thus the strong heredity assumption implies that the parameter vector $(\beta_1,\ldots,\beta_p,\psi_1,$ $\ldots,\psi_p)$ will be entirely equal to $\mathbf{0}$ if $(\beta_1,\ldots,\beta_p) = \mathbf{0}$. Therefore, the smallest value of λ for which the entire parameter vector reduces to $\lambda_{max} = \frac{1}{n(1-\alpha)}\max_j\left\{\left|\left(\boldsymbol{X}_j\right)^\top \boldsymbol{R}_{(-j)}\right|\right\}$. The computational algorithm to fit all the parameters in a sequence of loops is further detailed in the Supplementary Material (Algorithm 1).

3.2.5 Multiple Intervals Estimation

Knowing how to estimate the blip parameters in a one-stage setting, we now describe how the pdWOLS approach works in a K-stages setting. Starting from the last stage, the estimation procedure is applied to the K-th stage observed outcome $\boldsymbol{y}_K$, treatment $\boldsymbol{a}_K$, and covariates $\boldsymbol{x}_K$. The estimated blip parameters are obtained by maximizing the objective function in Equation (3.3) and the estimated rules $\widehat{a}_K^{opt} = I(\widehat{\psi}_{0K} + \boldsymbol{x}_K\widehat{\boldsymbol{\psi}}_{\boldsymbol{K}} > 0)$, where I is the indicator function. The (K-1)-th stage outcome is based on "optimal responses", that is, the estimation

procedure is applied to the pseudo-outcome $\widetilde{\boldsymbol{y}}_{k-1} = \boldsymbol{y}_K + \mu_K(\boldsymbol{x}_K, a_K; \widehat{\boldsymbol{\psi}}_{\boldsymbol{K}})$, treatment $\boldsymbol{a}_{K-1}$ and covariates $\boldsymbol{x}_{K-1}$, where $\mu_K(\boldsymbol{x}_K, a_K; \widehat{\boldsymbol{\psi}}_{\boldsymbol{K}}) = \gamma_K(\boldsymbol{x}_K, \widehat{a}_K^{opt}; \widehat{\boldsymbol{\psi}}_{\boldsymbol{K}}) - \gamma_K(\boldsymbol{x}_K, a_K; \widehat{\boldsymbol{\psi}}_{\boldsymbol{K}})$ is the regret function at stage K. The pseudo-outcome, $\widetilde{\boldsymbol{y}}_{K-1}$, is optimal since the regret is added to the observed outcome $\boldsymbol{y}_K$. The same procedure continues, recursively working backwards, until stage 1 estimation, such that the blip parameters across all the stages are obtained and all treatment decisions can be made.

3.2.6 Asymptotic Properties of the pdWOLS estimator

We now show that when the number of predictors, p, is fixed and the sample size n approaches infinity, the pdWOLS estimator has both the double robustness and oracle properties (Fan and Li, 2001) under several assumptions. Following the adaptive LASSO (Zou, 2006), we add adaptive weights (or penalty factors) to the objective function (3.3) to obtain

$$\mathcal{L}^*(\boldsymbol{Y}; \boldsymbol{\theta}) + \lambda(1-\alpha)\sum_{j=1}^{p} w_j^{main}|\beta_j| + \lambda\alpha\sum_{j=1}^{p} w_j^{int}|\tau_j|, \tag{3.7}$$

where w_j^{main} and w_j^{int} are adaptive weights of main effect and interaction terms respectively, in this way, the coefficients are not forced to be equally penalized in the ℓ_1 penalty. For instance, we can choose $w_j^{main} = \left|\widehat{\beta}_j^{wls}\right|^{-1}$ and $w_j^{int} = \left|\frac{\widehat{\beta}_j^{wls}\widehat{\psi}_0^{wls}}{\widehat{\psi}_j^{wls}}\right|$ for penalty factors, where $\widehat{\beta}_j^{wls}$ and $\widehat{\psi}_j^{wls}$ are unpenalized weighted least square estimates of the pdWOLS model. As n goes to infinity, the weights corresponding to unimportant variables go to infinity, which puts a large penalty on those variables, and the weights corresponding to important variables converge to a finite constant. Thus, small coefficients are removed, and large coefficients are unbiasedly estimated. Without loss of generality, we can rewrite Equation (3.7) as $\mathcal{L}^*(\boldsymbol{Y}; \boldsymbol{\theta}) + \sum_{j=1}^{p} \lambda_j^{\beta}|\beta_j| + \sum_{j=1}^{p} \lambda_j^{\tau}|\tau_j|$, where $\lambda_j^{\beta} = \lambda(1-\alpha)w_j^{main}$ and $\lambda_j^{\tau} = \lambda\alpha w_j^{int}$.

We assume that the true model follows the strong heredity assumption described above and regularity conditions detailed in the Supplemental Material hold. Note that the regularity conditions of pdWOLS are for quasi-likelihood since the loss function contains data-

dependent weights and the treatment-free model may be misspecified. We describe the asymptotic properties of pdWOLS in the following theorems; proofs are given in the Supplemental Material. Assume that the observations $\boldsymbol{V}_i, i = 1, \ldots, n$ are independent and identically distributed with probability density $g(\boldsymbol{V})$ with respect to a measure ν. Denote the negative quasi-log-likelihood as $L_n^*(\boldsymbol{V}; \boldsymbol{\theta}) = -\sum_{i=1}^n log\, h(\boldsymbol{V}_i, \boldsymbol{\theta})$ (i.e., the dWOLS loss function), where h is the posited family of densities. Let $\boldsymbol{\theta}^*$ be the underlying true parameters, and $\boldsymbol{\theta}_*$ the minimizer of the Kullback–Leibler divergence between h and g (i.e., $\boldsymbol{\theta}_*$ is the closest point to $\boldsymbol{\theta}^*$ in the posited family of densities). Define B_1 as the indices of non-zero components for main effects and B_2 as the indices of non-zero components for interaction terms such that

$$B_1 = \{j : \beta_{*j} \neq 0\}, B_2 = \{j + p + 1 : \tau_{*j} \neq 0\}, B = B_1 \cup B_2,$$

where we define $\boldsymbol{\tau}_*$ in a way such that $\tau_{*j} = \frac{\psi_{*j}}{\psi_{*0}\beta_{*j}}$ if $\beta_{*j} \neq 0$ and 0 otherwise, since we assume the strong heredity property holds. Let na_n be the maximum value of the tuning parameters $(\lambda^\beta, \lambda^\tau)$ such that the corresponding coefficients are non-zero and nb_n be the minimum value of the tuning parameters such that the corresponding coefficients are zero. For λ^τ we only consider the index m such that $\beta_{*m} \neq 0$ and $\psi_{*m} = 0$ (i.e., $m \in B_1$):

$$a_n = \frac{1}{n} max\{\lambda_j^\beta, \lambda_m^\tau : j \in B_1, m + p + 1 \in B_2\}$$

$$b_n = \frac{1}{n} min\{\lambda_j^\beta, \lambda_m^\tau : j \in B_1^c, m + p + 1 \in B_2^c \text{ such that } \beta_{*m} \neq 0\}.$$

Theorem 3.2.1. *Correct Sparsity: Assume that* $\sqrt{n}a_n = O(1)$ *and* $\sqrt{n}b_n \to \infty$*, then there exists a local minimizer* $\widehat{\boldsymbol{\theta}}_n$ *of Equation* (3.7) *such that* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*\| = O_p(n^{-\frac{1}{2}} + a_n)$*. Moreover, we have* $P(\widehat{\boldsymbol{\theta}}_{B^c} = 0) \to 1$.

Theorem 3.2.2. *Asymptotic Normality: Assume that* $\sqrt{n}a_n \to 0$ *and* $\sqrt{n}b_n \to \infty$*, then*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_{*B}) \to_d N\left(0, \boldsymbol{J}^{-1}(\boldsymbol{\theta}_{*B})\boldsymbol{I}(\boldsymbol{\theta}_{*B})\boldsymbol{J}^{-1}(\boldsymbol{\theta}_{*B})\right)$$

where $\boldsymbol{J}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 log\ h(\mathbf{V};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]$ *and* $\boldsymbol{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\left(\frac{\partial log\ h(\mathbf{V};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial log\ h(\mathbf{V};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right]$.

Remark 3.2.1. *Oracle properties of* $\widehat{\boldsymbol{\theta}}_n$ *are established such that the estimator converges to some population parameter instead of the underlying true parameter* $\boldsymbol{\theta}^*$*. Also, the asymptotic covariance matrix no longer equals the inverse of the Fisher's information matrix. If the treatment-free model is correctly specified, then* $\widehat{\boldsymbol{\theta}}_n$ *will converge to* $\boldsymbol{\theta}^*$*. To mimic the oracle, we further assume that all the observational weights are* 1 *(e.g., as in a randomized study).*

Corollary 3.2.1. *Double Robustness: Assume that the blip function is correctly specified and SUTVA and ignorability described in Section 3.2.1 hold, then the resulting blip parameter estimators of pdWOLS are doubly-robust; the estimators are consistent (i.e.,* $\boldsymbol{\psi}_* = \boldsymbol{\psi}^*$*) if either the treatment model or the treatment-free model is correct. Note that correct specification of the blip model permits over-specification - that is, the true blip model may be contained within the analyst-specified model. From Theorems 3.2.1 and 3.2.2, pdWOLS has the same performance as dWOLS, and hence it has the double robustness property.*

Remark 3.2.2. *There are no consistency guarantees for the first-stage estimator if an important confounder is missing in the second-stage model, as this violates an assumption at the second stage such that the estimator of second-stage parameters (subsequently plugged into the first-stage estimating function) may be biased. However, if estimation at the second stage is consistent (no unmeasured confounding, at least one of the nuisance models correct, etc), then double-robustness at the first stage can be assured under key assumptions.*

3.3 Simulation Studies

In this section, we first illustrate the double robustness of pdWOLS and compare its performance to competing approaches through a number of simulations; then we implement the proposed method in a high dimensional setting where $p > n$. Lastly, we present simulation results for a two-stage setting. The tuning parameter α was set to 0.5 for all simulations, and λ was selected using four-fold cross-validation to reduce the computational burden.

In addition to assuming that there are no unmeasured confounders, we assume that the number of confounders is relatively small, so that the propensity score model can be fitted using logistic regression with the entire vector $\boldsymbol{X}$. The propensity score is used to ensure balance between treatment groups. If model misspecification is a concern, one can use data-adaptive techniques, however, care must be taken in using data-adaptive approaches to estimating the propensity score to avoid the risk of selecting instruments, i.e., variables that only predict treatment (Shortreed and Ertefaie, 2017). To consider a general framework, main effects are penalized in Equation (3.3). However, in a low dimensional setting, we may want to retain all available covariates in the outcome model to ensure no weak confounders are erroneously omitted. In such cases, we can choose to not penalize the main effects, setting the corresponding penalty factors in Equation (3.7) to zero.

3.3.1 Competing Methods

We compare the variable selection results, error rate (in terms of the estimated rules as compared to the true optimal treatment), and out-of-sample value (i.e., expected outcome) under the estimated rules of pdWOLS with Q-learning combined with LASSO (Blatt et al., 2004) and penalized A-Learning (PAL) (Shi et al., 2018). Q-learning is a sequential regression approach to DTR estimation; relying only on outcome models; it is not doubly robust. PAL first estimates the treatment-free and propensity score models, then uses the Dantzig selector (Candes and Tao, 2007) to penalize the estimating equations of A-learning: $\widehat{\boldsymbol{\psi}} =$

$argmin_{\boldsymbol{\psi}} \|\boldsymbol{\psi}\|_1$ subject to $\|\boldsymbol{X}^T diag(\boldsymbol{A}-\widehat{\boldsymbol{\pi}})(\boldsymbol{Y}-f(\boldsymbol{x};\widehat{\boldsymbol{\beta}})-\gamma(\boldsymbol{x},a;\boldsymbol{\psi}))\|_\infty \leq n\lambda_{pal}$, where λ_{pal} is the tuning parameter and $\widehat{\boldsymbol{\pi}}$ is the estimated propensity score.

LASSO was implemented using the R package glmnet (Hastie et al., 2010) with λ_{LASSO} selected via four-fold cross-validation. PAL was implemented using the R package ITRSelect (Shi et al., 2018) with the tuning parameter λ_{pal} selected via the Bayesian Information Criteria (BIC) (Schwarz, 1978). The main effect of treatment A is not penalized in any of the three methods. We also present unpenalized estimates of the blip parameters from a two-step approach: that is, after variable selection, the blip parameters are re-calculated by solving the unpenalized weighted least squares via Q-learning, dWOLS, and A-learning with the selected variables, which we term *refitted* procedure.

3.3.2 Experiments Examining Double Robustness Property

We begin with a simple one-stage example with the following data generation procedure:

Step 1: Generate 10 covariates $(\boldsymbol{X}_1-\boldsymbol{X}_{10})$ where $\boldsymbol{X}$ are multivariate normal with zero mean, unit variance, and correlation $Corr(X_j,X_k)=0.25^{|j-k|}$ for $j,k=1,2,\ldots,10$.

Step 2: Generate treatment according to the model:

$$P(A=1|X_1,X_2)=\frac{exp(1+x_1+x_2)}{1+exp(1+x_1+x_2)}.$$

Step 3: Set the blip function, and hence the optimal treatment strategy, to depend only on X_1: $\gamma(x,a;\boldsymbol{\psi})=a(\psi_0+\psi_1 x_1)$ for $\psi_0=1,\psi_1=-1.5$.

Step 4: Set the treatment-free model to $f(\boldsymbol{x};\boldsymbol{\beta})=0.5-0.6e^{x_1}-2x_1-2x_2$.

Step 5: Generate the outcome $Y\sim N(f(\boldsymbol{x};\boldsymbol{\beta})+\gamma(\boldsymbol{x},a;\boldsymbol{\psi}),1)$.

We apply estimation and variable selection approaches with a variety of sample sizes (100, 500, and 2000) in four scenarios, where neither, one, or both of the treatment and treatment-

free models is correctly specified. Specifically, the scenarios are: *Scenario 1 (neither treatment nor treatment-free is correct):* Regress $\boldsymbol{Y}$ on $(\mathbf{1}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{AX})$, and set all observational weights to 1 (similar to assuming a null propensity score model). As this scenario fails to meet the assumptions of correct model specification, consistency is not assured for any approach. *Scenario 2 (treatment correct, treatment-free incorrect):* Regress $\boldsymbol{Y}$ on $(\mathbf{1}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{AX})$, but fit a correctly specified propensity score model whose parameters are estimated via logistic regression. *Scenario 3 (treatment incorrect, treatment-free correct):* Regress $\boldsymbol{Y}$ on $(\mathbf{1}, e^{\boldsymbol{X}_1}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{A}e^{\boldsymbol{X}_1}, \boldsymbol{AX})$, so that the treatment-free model is correctly specified but - as in scenario 1 - set all observational weights to 1. *Scenario 4 (both treatment and treatment-free are correct):* Regress $\boldsymbol{Y}$ on $(\mathbf{1}, e^{\boldsymbol{X}_1}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{A}e^{\boldsymbol{X}_1}, \boldsymbol{AX})$, and estimate the parameters using a correctly specified propensity score.

Since Q-learning does not incorporate any propensity score adjustments, scenarios 1 and 2 yield identical estimates, as do scenarios 3 and scenario 4. All the three methods have the same treatment-free models and the same blip functions to be estimated in the four scenarios. Across all scenarios where at least one nuisance model was correctly specified, refitted estimators performed better than their penalized counterparts in terms of bias (see Figure S1 in the Supplementary Material). When at least one of the treatment or treatment-free models was correctly specified, the blip parameter estimators were consistent for refitted pdWOLS. When the treatment-free model was correct (Scenarios 3 and 4), the refitted Q-learning (LASSO) estimators were consistent, as expected. Surprisingly, PAL failed when the treatment model was incorrect (Scenario 3). This result was not anticipated since PAL is a double robust method, although previous simulations have not considered its performance in terms of parameter estimates (Shi et al., 2018).

The variable selection results for optimal treatment decisions are presented in Table 3.1. In Scenarios 2-4, the important tailoring variable was correctly selected by both pdWOLS and Q-learning (LASSO). PAL failed in scenario 3. However, the false positive rates of

pdWOLS and Q-learning (LASSO) were higher than that of PAL in all scenarios: for example, in Scenario 3, both LASSO and pdWOLS falsely selected the variable Ae^{X_1} 72% of the time.

Table 3.1 also summarizes the error rates (i.e., $\frac{1}{n}\sum_{i=1}^{n} I(a_i^{opt} \neq \widehat{a}_i)$) of the estimated optimal treatment regimes for treatment decision making and value functions. The average value function and the error rates were computed over a testing set of size 10,000 (i.e., a dataset generated according to the process described above in all respects except that treatment was allocated according to the estimated rule). Both the error rate and the value of pdWOLS and Q-learning with LASSO were very close; pdWOLS outperformed other methods in Scenario 2, while Q-learning with LASSO had the best performance in Scenarios 3 and 4. The performance of the refitted versions of pdWOLS and Q-learning were similar; the performance of PAL was uniformly worse than the other methods performed without refitting, however refitting PAL substantially improved its performance.

Table 3.1: Variable selection rate (%) of the blip parameters, error rate (ER, %) and value function over a testing set of size 10,000 under the estimated decision rules using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions ($n = 500$, 400 simulations). The main effect of treatment is not penalized (and hence is always selected).

	Scenario 2			Scenario 3			Scenario 4		
	pdWOLS	QL	PAL	pdWOLS	QL	PAL	pdWOLS	QL	PAL
Ae^{X_1}	-	-	-	72	14	72	42	14	0
AX_1^*	100	100	99	100	100	33	100	100	100
AX_2	53	51	2	73	44	3	52	44	1
AX_3	2	26	2	6	23	0	2	23	1
AX_4	4	28	2	5	24	1	3	24	1
AX_5	4	29	4	4	26	2	2	26	2
AX_6	2	26	2	4	21	1	1	21	1
AX_7	3	25	2	5	22	1	2	22	0
AX_8	3	27	3	6	23	1	2	23	0
AX_9	2	27	3	6	24	1	2	24	1
AX_{10}	2	28	2	5	22	1	1	22	1
ER	3.9	9.8	22.9	5.5	3.4	12.0	4.2	3.4	23.4
ER (Refitted)	4.5	8.5	4.9	3.4	3.6	8.7	3.6	3.6	3.8
Value	0.6	0.6	0.5	0.6	0.7	0.6	0.6	0.7	0.5
Value (Refitted)	0.6	0.6	0.6	0.6	0.7	0.6	0.7	0.7	0.6

* Term with a non-zero coefficient in the data-generating model
Note that Ae^{X_1} was not included in the blip model for scenario 2

3.3.3 Simulations Evaluating Performance in a High-dimensional Setting

Here we present the performance of the new procedure in a high dimensional setting with $p = 400$ and $n = 200$. The data generation procedure is the same as in Section 3.2, except that we now set $P(A = 1)$ to 0.5 for everyone such that no confounding is present. The blip function is $\gamma(\boldsymbol{x}, a; \boldsymbol{\psi}) = a(1 - 1.5x_1)$ where $\psi_0 = 1, \psi_1 = -1.5$ and the treatment-free model is $f(\boldsymbol{x}; \boldsymbol{\beta}) = 0.5 - 0.6e^{x_1} - 2x_1 - 2x_2$. We regress $\boldsymbol{Y}$ on $(\boldsymbol{1}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{AX})$ where the treatment-free model is misspecified.

Figure 3.1 summarizes the blip parameter estimates in the high dimensional setting. Like before, for all the methods, refitted estimators improved the performance of their penalized

counterparts. For ψ_0, Q-learning with LASSO and its refitted estimator had the smallest bias; as for ψ_1, pdWOLS and its refitted version had the smallest bias.

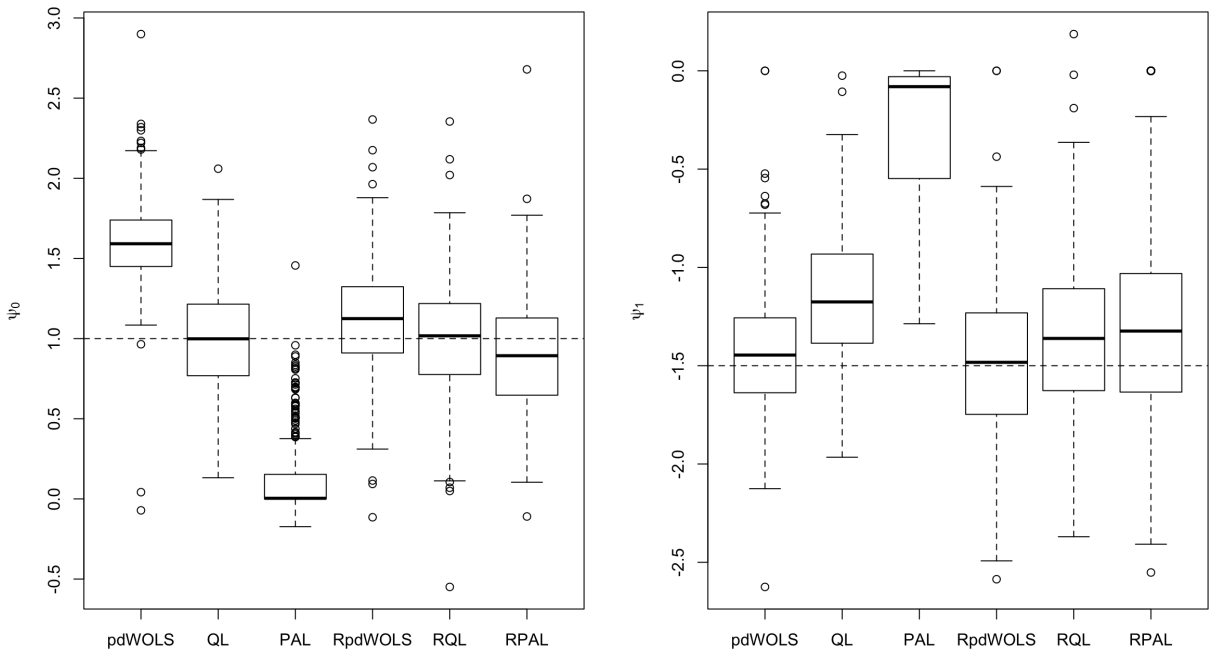

Figure 3.1: Estimates of blip parameters using pdWOLS, Q-learning (LASSO), PAL and their refitted versions with sample size 200 (400 simulations) in a high dimensional ($p = 400$) setting. The true value is represented by the dotted line.

Table 3.2 shows false negative rates (the proportion of times a method wrongly removed a truly important variable), false positive rates (the proportion of times a method wrongly included a non-important variable), error rates, and the value under the estimated rules of the three methods. The average value function and the error rates were computed over a testing set of size 10,000. Q-learning with LASSO achieved a zero false negative rate; pdWOLS and refitted pdWOLS had the lowest false positive rate, error rate, and the highest value, which indicates favorable performance of the newly proposed method. However, unlike before, even the refitted PAL estimator had a smaller bias than the PAL estimator; refitted PAL did not improve the performance of PAL with respect to value and error rate, which shows that

smaller bias in estimation of blip parameters does not necessarily translate into a better performance of the estimated regime.

Table 3.2: False negative (FN, %) rate and false positive (FP, %) rate of variable selection results of the blip parameters, error rate (ER, %) and value using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 200 (400 simulations) in a high dimensional ($p = 400$) setting. The main effect of treatment is not penalized (and hence is always selected).

	FN	FP	ER	Value
pdWOLS	0.3	0.2	12.8	0.7
QL (LASSO)	0.0	1.4	11.3	0.7
PAL	2.6	0.4	24.6	0.6
RpdWOLS	0.3	0.2	9.9	0.7
RQL (LASSO)	0.0	1.4	16.8	0.6
RPAL	2.6	0.4	25.1	0.5

3.3.4 Simulations Evaluating Performance in Multi-stage Setting

In this subsection, we demonstrate the performance of the proposed pdWOLS approach when treatment decisions are made at multiple stages. We consider two different data generation procedures in order to follow previous literature. Setting 1, in which the true treatment-free model does not have an analytical closed-form (misspecified treatment-free model) is presented here. Setting 2, in which the treatment-free models can be computed analytically, is available in the Supplemental Material.

We follow the data generation procedure in (Wallace and Moodie, 2015) with a sample size of 1000:

Step 1: Generate 10 covariates at stage 1: $X_{j1} \sim N(0, 1)$ for $j = 1, 2, ...10$.

Step 2: Generate treatment at stage k according to

$$P(A_k = 1|X_{1k}, X_{2k}) = \frac{exp(x_{1k} - x_{2k})}{1 + exp(x_{1k} - x_{2k})},$$

for $k = 1, 2$.

Step 3: Generate covariates at stage 2, such that $X_{12} \sim N(0.5A_1 + 0.8X_{11}, 1)$ and $X_{j2} \sim N(0.8X_{j1}, 1)$, for $j = 2, 3, ...10$.

Step 4: Set the blip functions to be $\gamma_1(x_1, a_1; \boldsymbol{\psi}_1) = a_1(0.8 - 2x_{11})$ and $\gamma_2(x_2, a_2; \boldsymbol{\psi}_2) = a_2(1 - 1.5x_{12})$, so that $\psi_{01} = 0.8$, $\psi_{11} = -2$, $\psi_{02} = 1$ and $\psi_{12} = -1.5$.

Step 5: Generate the outcome under optimal treatment according to $y^{opt} = 0.5 + 2x_{11} + 2x_{12}$. The observed outcome is generated such that $Y \sim N(y^{opt} - \mu_1 - \mu_2, 1)$ where μ_1 and μ_2 are regret function at stages 1 and 2, defined through the blip functions in step 4.

Recall, that a backward recursive approach can be used to make the treatment decision. Starting from the last stage, the estimation procedure is applied to the observed outcome $\boldsymbol{y}$. The estimated blip parameters and the estimated rules, $\widehat{a}_2^{opt}$, are obtained. Estimation then proceeds to stage 1, where again the estimation procedure is applied to a pseudo-outcome which represents the expected effect of the observed stage 2 treatment with the optimal stage 2 treatment. In pdWOLS, the pseudo-outcome is $\widetilde{y}_1 = y + \gamma_2(\boldsymbol{x}_2, \widehat{a}_2^{opt}; \widehat{\boldsymbol{\psi}}_2) - \gamma_2(\boldsymbol{x}_2, a_2; \widehat{\boldsymbol{\psi}}_2)$, where as for Q-learning with LASSO, the pseudo-outcome is $\widetilde{y}_1^Q = f(\boldsymbol{x}_2; \widehat{\boldsymbol{\beta}}_2) + \gamma_2(\boldsymbol{x}_2, \widehat{a}_2^{opt}; \widehat{\boldsymbol{\psi}}_2)$.

In this setting, the treatment free model in the second stage of estimation aims to represent $y^{opt} - \mu_1 - a_2^{opt}(1 - 1.5x_{12})$ which depends on a_2^{opt}, which in turn is a function of second stage parameters $\boldsymbol{\psi}_2$ and covariate x_2. The treatment free model in this setting cannot be computed analytically. We nevertheless assumed that the treatment-free models were linear in the covariates measured at their respective stages, and thus in these simulations, it is always the case that the treatment-free models were misspecified. For those methods relying on a propensity score, the treatment models were fit using correctly-specified logistic regression models at each stage using all covariates measured at that stage.

Figure 3.2 summarizes the estimates of blip parameters using the three methods in the two-

stage Setting 1. As expected, pdWOLS and PAL work when at least one of the treatment or treatment-free models is correctly specified (in this case, the treatment model is correctly specified), and Q-learning with LASSO failed, since the treatment free model at both stages are misspecified. For pdWOLS and PAL, refitted estimators were nearly unbiased, and they performed better than their penalized counterparts. At stage 1, the bias of PAL estimators decreased to almost zero after refitting. Thus, PAL exhibits excellent performance in variable selection but requires the additional step of refitting for accurate estimation. Unlike PAL, pdWOLS can have small bias even without the refitting procedure.

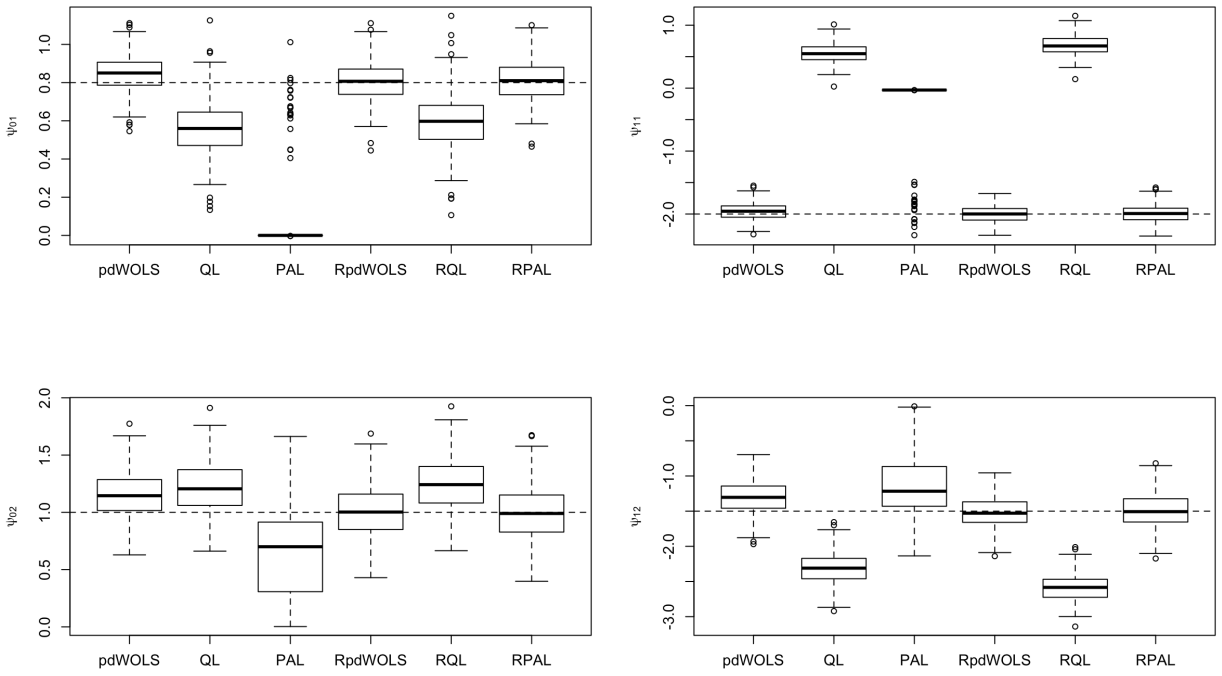

Figure 3.2: Estimates of blip parameters using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 1. The true value is represented by the dotted line.

Table 3.3 presents the variable selection results for optimal treatment decisions. The important tailoring variables were selected by all methods at both stages. At stage 2, the false positive rate of pdWOLS was much smaller than other two methods. For instance, the

selection frequency of $AX_2 - AX_{10}$ were all less than 5%. Note that at stage 1, because the pseudo-outcomes were different for refitted version and their penalized counterparts, the variables selected by the procedures may differ between penalized and unpenalized implementations.

Table 3.3: Variable selection rate (%) of the blip parameters using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 1. The main effect of treatment is not penalized (and hence is always selected).

	Stage 1						Stage 2		
	pdWOLS	QL	PAL	RpdWOLS	RQL	RPAL	pdWOLS	QL	PAL
AX_1 *	100	100	100	100	100	100	100	100	100
AX_2	49	32	1	45	33	2	22	44	33
AX_3	4	28	2	2	34	2	2	41	38
AX_4	3	30	0	2	34	2	3	45	37
AX_5	3	25	1	2	29	2	2	40	40
AX_6	4	25	1	2	29	1	3	40	40
AX_7	4	27	0	2	33	2	3	40	38
AX_8	4	28	0	2	30	1	2	42	38
AX_9	4	26	1	2	32	4	2	41	36
AX_{10}	3	29	2	2	32	2	2	44	38

* Term with a non-zero coefficient in the data-generating model

Table 3.4 summarizes the error rates of the estimated optimal treatment decisions and value functions, computed over a testing set of size 10,000. As before, refitted methods had lower error rate and higher value functions than their penalized counterparts. Penalized dynamic ordinary least squares outperformed other methods at both stages with respect to the error rate and value function; refitting greatly improved the performance of PAL.

Table 3.4: Error Rate (%) and value function using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 1. The total error rate (TER, %) in the estimated optimal treatment across both stages as well as the stage-wise error rates are shown.

	TER	ER (Stage 1)	ER (Stage 2)	Value
pdWOLS	9.2	2.2	7.2	0.4
QL (LASSO)	52.8	50.4	4.6	-0.6
PAL	22.4	14.5	9.8	0.4
RpdWOLS	6.5	2.0	4.6	0.5
RQL (LASSO)	58.0	54.8	6.1	-0.7
RPAL	11.3	2.1	9.5	0.4

Additionally, we compared the choice of tuning parameter α, in order to assess sensitivity of the results to this choice; we considered values of 0.2, 0.5 (as in the analyses above), and 0.8. The results are presented in the Supplemental Material (Figure S4, Tables S3 and S4). To briefly summarize, among all the α's, the bias and the variance of the estimators, the error rate and the estimated value were virtually identical. However, for variable selection, as α increased, the false positive rate decreased notably (See Table S3 in the Supplemental Material), as a larger α will put more penalty on the interaction terms.

3.4 Application to STAR*D Study

In this section, we apply pdWOLS to STAR*D (Fava et al., 2003), a multistage randomized trial that aimed to determine effective treatments for patients with major depressive disorder, where severity was measured using the Quick Inventory of Depressive Symptomatology (QIDS) score (Rush et al., 2003). The study was divided into four levels (one of which had two sub-levels); patients had different treatments at each level and would exit the study upon achieving remission. See the Supplemental Materials for details.

We follow Wallace et al. (2019) and Chakraborty et al. (2013) to perform two-stage analysis based on the use of a selective serotonin reuptake inhibitor (SSRI), with negative QIDS score

as the outcome. Three tailoring variables were considered: (1) the QIDS score measured at the beginning of each level (denoted by q_k at stage k); (2) change in QIDS score divided by the time in the previous level (QIDS slope, denoted by s_k at stage k); and (3) patient preference measured prior to receiving treatment, which is a binary variable (denoted by p_k at stage k). We also generated d iid noise variables at each stage: noise variables at stage 1 were generated using $X_{j1} \sim N(0,1)$ and at stage 2, $X_{j2} \sim N(log|X_{j1}|, 1)$ for $j = 1, 2, \ldots, d$. We consider three scenarios for the analysis where $d = 5, 10, 20$ respectively.

Logistic regression was used to estimate the treatment model adjusting for patient preference only, following the trial design, and weights $w = |A - E(A|X)|$ were used in the analysis. As in Wallace et al. (2019), the treatment-free models were linear in (q_1, s_1, p_1) at stage 1 and (a_1, q_2, s_2, p_2) at stage 2. Linear blip models with covariates (q_1, s_1, p_1) at stage 1 and (a_1, q_2, s_2, p_2) at stage 2 were considered. Note in Wallace et al. (2019), a_1 and p_2 were not included in the blip models to avoid the multicollinearity; this is not necessary in pdWOLS, and hence our model specifications differ.

As in our simulations, the main effect of treatment was not penalized. In all three scenarios and both stages, pdWOLS returned the intercept-only blip model, suggesting that the optimal treatments are treat with SSRI ($A_1 = 1$) and treat with a non-SSRI ($A_2 = 0$) at stage 1 and 2, respectively, for all patients. Penalized A-learning, in contrast, was sensitive to the number of noise variables: when $d = 5$, PAL selected a_j, q_j, s_j, p_j for both stages $j = 1, 2$. When $d = 10$, PAL selected a_2 at stage 2 and a_1, q_1, s_1, p_1 at stage 1, and when $d = 20$, PAL selected a_2 at stage 2 and a_1, p_1 at stage 1. Chakraborty et al. (2013) and Wallace et al. (2019) found that no stage 2 blip covariates were statistically significant (consistent with pdWOLS), while at stage 1, they found only treatment preference was significant.

The false positive rates of PAL at stage 2 and 1 were 100%, 40% ($d = 5$), 10%, 50% ($d = 10$), and 10%, 45% ($d = 20$), respectively; for pdWOLS, the rate was 0% for all d.

3.5 Discussion

In this article, we extended dWOLS to a penalized estimation framework for variable selection and estimating the optimal treatment regimes simultaneously. The proposed method inherits the double robustness property from dWOLS. Our simulations indicated that pdWOLS compares favorably with other variable selection approaches in the context of DTRs.

Our method automatically enforces strong heredity through a simple reparametrization, which guarantees an assumption required by dWOLS. The idea of reparametrization is simple, however, one limitation is that the objective function is non-convex. Hence, it may be of interest, in future work, to investigate approaches that use convex constraints to achieve strong heredity. See, e.g., Bien et al. (2013); Zhao et al. (2009) and Haris et al. (2016).

The standard errors for the estimated blip parameters can be obtained directly; a sandwich formula for computing the covariance of the estimates of the non-zero components can be derived (Fan and Li, 2001). How to derive the standard errors for the estimated blip parameters under the use of refitted pdWOLS requires further investigation. Post selection inference (Lee et al., 2016) should also be addressed.

The proposed method is, fundamentally, based on prediction, selecting any variables that can improve predictive ability. As such, in finite samples, pdWOLS may underestimate the importance of variables that have small predictive ability but that play a significant role in DTRs. Besides, the application of predictive methods directly to causal models may result in inflated variances and self-inflicted bias (Hernán and Robins, 2020). The importance of the distinction between DTRs (causal inference) and prediction must be kept in mind. Variable selection in causal inference is a tough problem: on the one hand, we want to adjust for enough covariates in the analysis to achieve ignorability; on the other hand, adjustment for some other irrelevant variables could induce bias and losses of statistical efficiency (Rotnitzky et al., 2010). Hence, a thoughtful selection of confounders is needed, using expert knowledge to guide variable selection is encouraged. Other discussions about confounder selection can

be found in Shortreed and Ertefaie (2017); Robins and Greenland (1986); Schneeweiss et al. (2009). For pdWOLS, if we are worried about confounding and our focus is on building simple rules, we may want to do minimal selection on main effects but lots of selection on interaction effects, which can be implemented by setting small adaptive weights w_j for the main effects or setting α to a large value. How to choose the tuning parameter λ and α in a DTR framework is an open and intriguing problem worthy of further investigation.

Acknowledgements

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01 MH114873 (co-PIs Shortreed and Moodie). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Moodie is a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine and acknowledges the support of a chercheur de mérite career award from the Fonds de Recherche du Québec, Santé. Bhatnagar acknowledges funding via a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2020-05133.

Conflict of Interest

Dr. Shortreed has been a co-Investigator on Kaiser Permanente Washington Health Research Institute (KPWHRI) projects funded by Syneos Health, who was representing a consortium of pharmaceutical companies carrying out FDA-mandated studies regarding the safety of extended-release opioids.

Data Availability Statement

The data that support the findings of this study are available in the NIMH Data Archive at https://nda.nih.gov/, collection ID number 2148.

Supporting Information

A Web Appendix containing the algorithm referenced in Section 2.4, regularity conditions and proofs of Theorems in Section 2.6, additional simulation results in Sections 3.2, 3.4, STAR*D details in Section 4 and an example of pdWOLS implemented in the R programming language are available with this paper at the Biometrics website on Wiley Online Library.

Chapter 4

Tailoring Variable Selection and Ranking for Optimal Treatment Decisions

Preamble to Manuscript 2. In Manuscript 1, we assumed that the number of confounders is relatively small, so that the propensity score model could be estimated using the entire set of covariates. However, this method would fail when the number of variables is really large. Hence, a data-driven method to select the confounders for the propensity score should be considered. Another topic in Manuscript 1 that requires further investigation is tuning parameter selection. Previously, we used cross-validation and BIC to select the tuning parameter λ. These two approaches are widely used in the penalized likelihood framework, where the goal is prediction. In contrast, DTRs are not usually estimated by likelihood-based methods. How to choose the tuning parameter in a DTRs framework is very important. In Manuscript 2, we incorporate a confounder selection method and an information criterion into the implementation of penalized dynamic weighted ordinary least squares. The corresponding manuscript is a book chapter accepted as a contribution to *Handbook of Statistical Methods for Precision Medicine* edited by Tianxi Cai, Bibhas Chakraborty, Eric B. Laber, Erica E.M. Moodie, and Mark J. van der Laan.

Tailoring Variable Selection and Ranking for Optimal Treatment Decisions

Zeyu Bian[1], Erica EM Moodie[1], Susan M Shortreed[2,3],

Sylvie D Lambert[4,5] and Sahir Bhatnagar[1,6]

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2]*Kaiser Permanente Washington Health Research Institute*

[3]*Department of Biostatistics, University of Washington*

[4]*Ingram School of Nursing, McGill University*

[5]*St. Mary's Research Centre*

[6]*Department of Diagnostic Radiology, McGill University*

Abstract

Dynamic treatment regimes (DTRs) are often estimated from data sources where many covariates are measured, and yet not all may be useful for tailoring treatment. This is particularly the case when DTRs are estimated from observational or non-experimental sources, as the data may have been collected for purposes other than the DTR analysis – however this may also arise in a randomized study where researchers have been thorough and erred on the side of collecting more rather than fewer potentially relevant variables. In such cases, including all the covariates in the model could possibly yield a complex and inappropriate treatment decision. Hence, it may be critical to apply variable selection techniques to a DTR analysis. Currently, most existing literature on DTRs focuses on estimation; variable selection with the objective of optimizing treatment decisions has been less studied. In this Chapter, we review existing tailoring variable selection techniques, with a particular focus on a regression-based variable selection method which can incorporate sparsity and estimate the optimal treatment regimes simultaneously. We also present a value search method that aims to rank the tailoring variables according to the estimated value function, and draw inspiration from this approach to assess the ability of the penalization approach to rank the potential candidate tailoring variables. In simulation studies, we demonstrate how the choice of tuning parameter selection as well as confounder selection affects the performance of the penalization approach. Finally, we apply the penalized approach and the ranking method to a pilot sequential multiple assignment randomized trial of a web-based, stress management intervention using a stepped-care method for cardiovascular diseases patients to determine useful tailoring variables.

4.1 Introduction

Regression-based methods and value search (policy search or classification-based approaches) methods are two popular estimation approaches in the field of dynamic treatment regimes (DTRs). Consider, for instance, the one-interval setting of an individualized treatment rule (ITR). In regression-based methods such as A-learning (Murphy, 2003; Robins, 2004) and Q-learning (Watkins, 1989), the expected mean outcome model is decomposed into baseline mean model and the blip function (throughout, we denote observed variables with lower case and their random counterparts with uppercase):

$$\mathbb{E}\left[Y^a|X=x;\boldsymbol{\beta},\boldsymbol{\psi}\right] = \underbrace{f\left(x;\boldsymbol{\beta}\right)}_{\text{baseline mean model}} + \underbrace{\gamma\left(x,a;\boldsymbol{\psi}\right)}_{\text{blip function}}$$

where Y is the observed outcome, Y^a is the potential outcome under the binary treatment a, x are baseline covariates, f is the baseline mean model with parameter $\boldsymbol{\beta}$, which is irrelevant for making optimal treatment decisions (i.e., it can be considered a nuisance model), and γ is the blip function with parameter $\boldsymbol{\psi}$. The blip function (or contrast function) is defined as the difference in expected potential outcome between a population of patients when treated with a and that same population of patients when treated with a reference treatment, $a = 0$, modelled as a function of some covariates x: $\gamma\left(x,a\right) = \mathbb{E}\left[Y^a - Y^{a=0}|X=x\right]$. Since f does not involve the treatment variable and hence will not affect treatment decisions, the parameter of interest is the blip parameter $\boldsymbol{\psi}$, and the optimal decision rule implied by this model is given by $\widehat{a}^{opt} = \arg\max_a \gamma\left(x,a;\widehat{\boldsymbol{\psi}}\right)$. For example, if the blip function is linear (implying that we assume a linear decision rule), then $\gamma(x,a;\boldsymbol{\psi}) = a(\psi_0 + \sum_{j=1}^p \psi_j x_j)$ and so the estimated optimal treatment is $\widehat{a}^{opt} = I(\widehat{\psi}_0 + \sum_{j=1}^p \widehat{\psi}_j x_j > 0)$ where $I(\cdot)$ is the indicator function.

In contrast, value search methods first specify a finite class of candidate regimes, and then estimate the value function (i.e., expectation) of every possible treatment regime in the class

of candidates. The optimal treatment regime is then chosen to be that which maximizes the estimated value function. Some examples of value search methods include dynamic marginal structural models (van der Laan and Petersen, 2007; Orellana et al., 2010), outcome weighted learning (Zhao et al., 2012), residual weighted learning (Zhou et al., 2017) and the inverse probability weighted estimator proposed in Zhang et al. (2012). The value function $V(\xi)$ can be modelled as a function of ξ non-parametrically, semi-parametrically, e.g., via restricted cubic splines (Cain et al., 2010), or parametrically e.g., via a quadratic function (van der Laan and Petersen, 2007; Shortreed and Moodie, 2012). The optimal treatment regime $\widehat{d}^{opt}(x;\widehat{\xi}^{opt})$ and the estimator $\widehat{\xi}^{opt}$ are selected such that the corresponding value $\widehat{V}(\widehat{\xi}^{opt})$ is maximized.

However, in the current era of high computer memory and computational power, more and more patient information is collected in medical studies or electronic health records. This recording of many covariates brings challenges for the analytic approaches described above. First, the existing methods often suffer from the curse of dimensionality. For example, regression-based methods fail in the case where the number of variables exceeds the sample size ($p > n$). Second, even in the $n > p$ setting, in the presence of many covariates, it is difficult to know which prognostic factors are relevant for treatment decision. Including all covariates in the analytic model can result in a loss of statistical efficiency and needlessly complex treatment decisions. Thus, applying variable selection to precision medicine becomes critical as the number of available covariates increases, since this can improve the treatment decision rules and simplify models to enhance interpretability and ease of implementation in clinical practice.

In this Chapter, we review a penalization approach to variable selection embedded within a regression-based method to ITR estimation–penalized dynamic ordinary least squares (pdWOLS) (Bian et al., 2021) and a variable ranking method which ranks tailoring variables based on the estimated value function associated with an ITR based on each variable

alone. A confounder selection method, outcome adaptive LASSO (Shortreed and Ertefaie, 2017), will be incorporated into the implementation of pdWOLS. The rest of this chapter is organized as follows. In Section 4.2, we provide background on commonly used estimation, variable selection, and ranking methods for DTRs. We present the extension of dynamic ordinary least squares to its penalized counterpart–pdWOLS in Section 4.3. We then provide some numerical results in the form of a simulation study in Section 4.4 and a case study in Section 4.5, applying tailoring variable selection and ranking to a pilot sequential multiple assignment randomized trial of a web-based stress management approach.

4.2 Background

In Section 4.2.1, we first review two regression-based methods and a value search method, then in Sections 4.2.2 and 4.2.3 we discuss how to select and rank the tailoring variables based on these methods.

4.2.1 Some DTRs Estimation Methods

A-learning

We begin this Section with a regression-based method A-learning (or G-estimation) (Murphy, 2003; Robins, 2004), a method which is doubly robust. That is, the approach can yield consistent estimators of $\boldsymbol{\psi}$ while only requiring one of two nuisance models, either the baseline mean model $f(\boldsymbol{x};\boldsymbol{\beta})$ or the propensity score model $P(A=1|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\alpha})=\pi(\boldsymbol{x};\boldsymbol{\alpha})$, to be correctly specified. The blip parameter can be obtained by solving the following A-learning estimating equation:

$$\boldsymbol{X}^T \operatorname{diag}(\boldsymbol{A}-\widehat{\boldsymbol{\pi}})(\boldsymbol{Y}-f(\boldsymbol{x};\widehat{\boldsymbol{\beta}})-\gamma(\boldsymbol{x},a;\boldsymbol{\psi}))=\mathbf{0}$$

where the plug-in estimators $\widehat{\boldsymbol{\pi}}$ and $f(\boldsymbol{x};\widehat{\boldsymbol{\beta}})$ can be estimated using (generalized) linear regression or other machine learning approaches, and α parameterizes the treatment model. When π or f is correctly specified,

$$\begin{aligned}&\mathbb{E}\{(A-\widehat{\pi})\boldsymbol{X}^T(Y-f(\boldsymbol{x};\widehat{\boldsymbol{\beta}})-\gamma(\boldsymbol{x},a;\boldsymbol{\psi}))\}\\&=\mathbb{E}_{\boldsymbol{X}}\mathbb{E}\{(A-\widehat{\pi})\boldsymbol{X}^T(Y-f(\boldsymbol{x};\widehat{\boldsymbol{\beta}})-\gamma(\boldsymbol{x},a;\boldsymbol{\psi}))|\boldsymbol{X}\}=0,\end{aligned}$$

hence $\boldsymbol{\psi}$ is consistently estimated.

While A-learning's robustness against model misspecification is appealing, its implementation can be somewhat complex. In the next subsection, we introduce a method offering easier implementation, while retaining the the double robustness property.

Dynamic Weighted Ordinary Least Squares

Dynamic weighted ordinary least squares (dWOLS) employs a weighted regression approach (Wallace and Moodie, 2015). If we assume that the blip function is correctly specified, stable unit treatment value assumption (Rubin, 1980) and ignorability (Robins, 1997) hold, and the main effects for all covariates in the blip model are included in the baseline mean model, then the resulting blip parameter estimators of dWOLS are consistent if either the treatment model or the baseline mean model is correct and so dWOLS enjoys the attractive double robustness property. The requirement that the baseline mean model include the main effects for covariates in the blip model is essential for dWOLS to consistently estimate ψ, unlike A-learning which imposes no restrictions on the nuisance model and can even simply assume an intercept-only baseline mean model.

Consider the following model with both linear baseline mean model and linear blip function

in a simple one-stage setting:

$$\boldsymbol{Y} = \psi_0 \boldsymbol{A} + \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j + \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ \boldsymbol{X}_j) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$ is a continuous response; A is the binary treatment; $\boldsymbol{X}_j \in \mathbb{R}^{n \times 1}$ are the j-th covariates; $\beta_j \in \mathbb{R}$ are the corresponding parameters for the main effects of covariates; $\psi_j \in \mathbb{R}$ are the blip parameters for $j = 0, 1, \ldots, p$, "$\circ$" is the element wise vector multiplication, and $\boldsymbol{\varepsilon}$ is an error term with standard normal distribution. Estimation for dWOLS is accomplished using the weighted squared-error loss:

$$\mathcal{L}(\boldsymbol{Y}; \boldsymbol{\theta}) = \frac{1}{2n} \| \sqrt{\boldsymbol{W}} \left(\boldsymbol{Y} - \psi_0 \boldsymbol{A} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j - \sum_{j=1}^{p} \psi_j (\boldsymbol{A} \circ \boldsymbol{X}_j) \right) \|_2^2 \tag{4.1}$$

where $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \psi_0, \ldots, \psi_p)$, and $\boldsymbol{W} = \text{diag}\{w_1(a, \boldsymbol{x}), w_2(a, \boldsymbol{x}), \ldots, w_n(a, \boldsymbol{x})\}$ is a $n \times n$ diagonal *balancing weights* matrix used to address confounding, which are a function of the propensity score.

We now discuss the double robustness property of dWOLS. Consider the scenario that the propensity score is correctly specified, thus the weights are all consistently estimated. In this case, estimating the blip parameters from model $f(x; \beta) + \psi_0 a + a\psi^T x$ is equivalent to estimating the causal effect from a conditional structural model $\mathbb{E}(Y^*(a)) = \tilde{\beta}^T x + \psi_0 a + a\psi^T x$ where $\tilde{\beta}$ is the coefficient of the baseline covariates in the structural model (not necessarily identical to the underlying true parameter β^*). Wallace and Moodie (2015) showed that a weighted least square estimator of model $\tilde{\beta}^T x + a\psi^T x$ with weights satisfied $\pi(\boldsymbol{x}) w(1, \boldsymbol{x}) = (1 - \pi(\boldsymbol{x})) w(0, \boldsymbol{x})$ would yield consistent estimates of the blip parameters ψ.

Now assume that the baseline mean model is correct and the estimated weights converge to $\tilde{w}$, dWOLS yields $\widehat{\theta} = (\widehat{\beta}, \widehat{\psi}) = \left(\phi(x, a)^T \tilde{w} \phi(x, a) \right)^{-1} \phi(x, a)^T \tilde{w} y$ where ϕ is the Jacobian matrix of the design matrix $\Phi(x, a)$. Let $\theta^* = (\beta^*, \psi^*)^T$, where θ^* is the true parameter of the model; by some simple algebra, it can be shown that $\mathbb{E}(\widehat{\theta}) = \theta^*$. Moreover, we have

$Var(\widehat{\theta}) = \sigma^2 C \left(\phi(x,a)^T \tilde{w}^{-1} w^* \tilde{w}^{-1} \phi(x,a)\right)^{-1} C$ which is $O(n^{-1})$ under mild conditions, where $C = \left(\phi(x,a)^T \tilde{w}^{-1} \phi(x,a)\right)^{-1}$. Hence, $P(|\widehat{\theta} - \theta^*| > \epsilon) \to 0$ and thus the blip parameters ψ are consistently estimated.

The choice of weights for dWOLS is an open but intriguing problem. Wallace and Moodie (2015) proposed to use the weights, $w(a, \boldsymbol{x}) = |a - \mathbb{E}[A|\boldsymbol{X} = \boldsymbol{x}]|$, as they empirically found it to offer better efficiency than other alternatives such as inverse probability weights in the simulations.

A Simple Value Search Method

We conclude this subsection by introducing a value search method that can be employed in settings where the decision rule is a simple threshold rule: for a tailoring variable X_j, we aim to find the optimal threshold value η_j^{opt} such that the decision rule is of the form $\widehat{a}^{opt} = I(X_j < \eta_j^{opt})$ or $I(X_j > \eta_j^{opt})$. In such value search methods, constructing an augmented data set is often a critical step in the estimation procedure. The augmented data set is created such that for each treatment regime $d(x;\xi)$ in the set of candidate regimes $\mathcal{A}$, and each observation i is repeated. If individual i's observed treatment a_i equals the recommended treatment under $d(x_i;\xi)$, then this subject i's information is copied into the augmented data set, which also contains an extra column which represents the value of ξ. Once the augmented data set is created, it is used to model the value function $V(\xi)$ as a function of ξ. The optimal treatment regime $\widehat{d}^{opt}(x;\widehat{\xi}^{opt})$ and the estimator $\widehat{\xi}^{opt}$ are selected such that the corresponding value $\widehat{V}(\widehat{\xi}^{opt})$ is maximized. In this chapter, we only consider decision rules of the type, $\widehat{a}^{opt} = I(X_j < \eta_j^{opt})$ for convenience. Considering both directions does not involve any extra difficulty, and should be done in real data settings whenever there is any uncertainty about the direction of the threshold rule. In Section 4.2.3, we give a more detailed description of how to rank the tailoring variables based on this framework.

4.2.2 Review of Tailoring Variable Selection Techniques

Among the earliest work on selecting tailoring variables in the context of a regression-based method of DTR estimation, Lu et al. (2013) adopted an adaptive LASSO (Zou, 2006) approach within the A-learning framework. These authors thus adopted a method that allowed the misspecification of the baseline mean model f, but only considered the approach in a randomized design case such that the propensity score is correctly specified and the number of predictors p is fixed. Lu et al. (2013) proposed that the blip parameter vector be estimated by minimizing the objective function

$$\|\boldsymbol{Y} - f(\boldsymbol{x};\boldsymbol{\beta}) - \text{diag}(\boldsymbol{A} - \widehat{\boldsymbol{\pi}})\boldsymbol{X}\boldsymbol{\psi}\|^2 + \lambda_n \sum_{j=1}^{p} \rho(|\psi_j|)$$

where λ_n is the tuning parameter and $\rho(|\cdot|)$ is a penalty function. To see how this approach allows the misspecification of the baseline mean model f, taking the derivative of the loss function (the first term in the above equation) with respect to $\boldsymbol{\psi}$ and set it to zero yields

$$\boldsymbol{X}^T \text{diag}(\boldsymbol{A} - \widehat{\boldsymbol{\pi}})\{\boldsymbol{Y} - f(\boldsymbol{x};\widehat{\boldsymbol{\beta}}) - \text{diag}(\boldsymbol{A} - \widehat{\boldsymbol{\pi}})\boldsymbol{X}\boldsymbol{\psi}\} = \boldsymbol{0},$$

which is an unbiased estimating equation under the assumption that the propensity score is known. Hence, with a suitable choice of tuning parameter, the blip parameter estimators are approximately unbiased in the penalized framework.

Jeng et al. (2018) extended the work of Lu et al. (2013) to the case where p is allowed to grow with the sample size n using debiased LASSO (Zhang and Zhang, 2014), and Shi et al. (2016) generalized it to the situation where p is of the non-polynomial order of the n, i.e., $\log p = O(n^q)$ for some $q < 1$.

Bian et al. (2021) proposed penalized dynamic ordinary least squares, a doubly robustness method which can conduct estimation and selection of the tailoring variables simultaneously; we will review this method in detail in Section 4.3.

In singly-robust settings, Song et al. (2015) added the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) penalty to the value search method of outcome weighted learning to incorporate sparsity and estimate the optimal treatment decision, where the treatment rule can be viewed as a classification problem. Zhu et al. (2019) combined Q-learning with smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) penalty to select the important tailoring variables and conducted the hard-thresholding method proposed in (Moodie and Richardson, 2010) to tackle the challenge of nonregularity.

With the exception of the outcome weighted learning approach, the loss functions of the methods described above were all likelihood-based. In contrast to these methods, Shi et al. (2018) used the Dantzig selector (Candes and Tao, 2007) to directly penalize the estimating equations of A-learning instead of penalizing the likelihood: $\widehat{\boldsymbol{\psi}} = argmin_{\boldsymbol{\psi}} \|\boldsymbol{\psi}\|_1$, subject to $\|\boldsymbol{X}^T \operatorname{diag}(\boldsymbol{A} - \widehat{\boldsymbol{\pi}})\{\boldsymbol{Y} - f(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}) - \gamma(\boldsymbol{x}, a; \boldsymbol{\psi})\}\|_\infty \leq n\lambda_{pal}$ where λ_{pal} is the tuning parameter. In this way, the double-robustness property of A-learning was inherited, and sparsity can also be introduced to the model through the use of the Dantzig selector.

All the variable selection methods outlined above are designed for use in DTR estimation. The differences between these approaches and the variable selection methods in prediction setting are twofold. First, the goal differs: variable selection for DTRs aims to improve the estimated decision rules rather than enhance the predictive power of the outcome (of course, both emphasize the interpretability of the model). Second, in DTRs, interest is primarily in effect modification, and hence the selection tends to be focused on the terms in the blip model. In contrast, variable selection approaches in prediction setting usually seek to choose from among all variables with no special status given to effect modifiers.

4.2.3 Tailoring Variable Ranking

Ranking variables can be viewed as an indirect method of choosing among covariates, or selecting among several potentially useful tailoring variables to focus on a simpler, low-

dimensional treatment rule. The earliest work in this domain is that of Gunter et al. (2011), who proposed a score-based method that can rank the tailoring variables based on the expected increase of the estimated value due to the inclusion of a variable, calling this the "S-score". This approach was primarily used to rank each potential tailoring variable individually, but was also extended to account for multiple tailoring variables in an iterative variable selection process. Fan et al. (2016) extended this work to take into account variables already in the model and determine whether to include a new variable by the additional improvement of the score.

Wu et al. (2021a) considered ranking variables for their tailoring ability in single-variable decision rules of the form $a^{opt}(X_j) = I(X_j < \eta_j^{opt})$ or $I(X_j > \eta_j^{opt})$. They used linear splines to flexibly model the value function and potential tailoring variables were then ranked according to their estimated value. The approach considers the effect of each variable individually, focusing on simple and interpretable rules. While the approach could, in principle, be used repeatedly to consider a "value added" rank (e.g. select the most useful variable, then rank the remaining variables to see which - if any - improve the value in a two-variable rule, and so on), however this has not yet been explored numerically. We now give a brief overview of variable ranking approach in Wu et al. (2021a).

Wu et al. (2021a) consider a decision rule such that $a^{opt} = I(X_j < \eta_j^{opt})$. Instead of modeling an outcome-covariates relationship, the value function $V(\eta_j) = m(\eta_j; \nu)$ can be modeled, where $m(\cdot)$ is some function with parameter ν. We omit the subscript j from here for notational convenience when there is no confusion. Recall from Section 4.1, that value search estimators often rely on the construction of an augmented data set. Denote by N the number of rows in the augmented data, and by r the number of threshold candidate. Wu et al. (2021a) used linear splines to flexibly model $V(\eta)$, with each candidate for η in the augmented data chosen as a potential knot in the linear spline, then employed penalization to shrink some of the coefficients to zero so that only the important thresholds (knots) were

selected.

For a potential tailoring variable with the candidate threshold sequence $(\eta^1, \ldots, \eta^r)$, the coefficients were obtained by minimizing the following function:

$$\sum_{l=1}^{N} w_l \{Y_l - \nu_0 - \nu' \eta^l - \sum_{s=1}^{r} \nu_s (\eta^l - \eta^s)_+\}^2 + \sum_{s=1}^{r} \lambda_s |\nu_s|,$$

where w_l is the corresponding inverse probability of treatment weight for *i-th* observations), ν_0 is the intercept, ν' is the coefficient of η, ν_s is the coefficient of the s-th knot and λ_s is the tuning parameter (not necessarily the same for all s). After variable selection, the coefficients are re-calculated by solving the unpenalized weighted least squares with the selected variables. Then we could use inverse probability weighting to estimate the value function of a particular regime $I(X < \eta^s)$: $\widehat{V}(\eta^s) = \frac{\sum_{l=1}^{N} I(\eta^l = \eta^s) \widehat{Y}_l w_l}{\sum_{l=1}^{N} I(\eta^l = \eta^s) w_l}$, where $\widehat{Y}_l$ is the fitted outcome from the linear spline model. The optimal threshold η^{opt} is chosen such that $\widehat{V}(\eta^{opt}) = max_{1 \leq s \leq r} V(\eta^s)$. This procedure is carried out for each potential tailoring variable, and we rank all the variables by their fitted optimal value function $\widehat{V}(\eta_j^{opt})$ for $j = 1, \ldots p$. A limitation of this ranking method is that it only considered single variable at a time.

As we shall detail in the next section, we posit that the penalized method that we outlined in Section 4.2.2 can also be used to rank the tailoring variables. This could be accomplished in a variety of ways. For example, we could rank them by the magnitude of the absolute value of the coefficient, or by the selected frequency among a sequence of tuning parameters. In the simulation studies, we will show how to rank the tailoring variables using the latter strategy.

4.3 Variable Selection and Ranking in DTRs

In this section, we review the tailoring variable selection method penalized dynamic weighted ordinary least squares, based on the dWOLS approach discussed in Section 4.2.1. We then

discuss how to incorporate *confounder* selection into pdWOLS, and conclude this section with a proposal for using pdWOLS as a means of ranking variables in terms of their utility for tailoring.

4.3.1 Penalized Dynamic Weighted Ordinary Least Squares

We can extend dWOLS to penalized estimation for variable selection and estimating the optimal treatment regimes simultaneously (Bian et al., 2021). One can consider the following objective function that includes the ℓ_1 penalty for variable selection:

$$Q\left(\boldsymbol{\theta}\right) = \mathcal{L}(\boldsymbol{Y};\boldsymbol{\theta}) + \lambda(1-\delta)(\|\boldsymbol{\beta}\|_1 + |\psi_0|) + \lambda\delta\|\boldsymbol{\psi}\|_1 \tag{4.2}$$

where $\mathcal{L}(\boldsymbol{Y};\boldsymbol{\theta})$ is the dWOLS loss function shown in equation (4.1), $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$, $\boldsymbol{\psi} = (\psi_1, ..., \psi_p)$, and $\lambda > 0$, $\delta \in (0, 1)$ are tuning parameters. In a slightly similar manner to elastic net regularization (Zou and Hastie, 2005), δ controls the relative penalties between the main effects and the interaction effects; for example, a δ that is less than 0.5 would tend to penalize the main effects more than the interaction effects. However, there is an issue with (4.2) in its current form: recall that an important assumption required by dWOLS is that the baseline mean model includes the main effects for all covariates in the blip function. In a penalized model such as equation (4.2), this assumption can be violated: it is possible that an estimated interaction term is nonzero while the corresponding main effects are shrunk to zero by the penalization. This can lead to a bias induced by the violation of dWOLS assumptions (rather than bias induced by the penalization itself).

To remedy this, pdWOLS added a constraint to (4.2), requiring the *strong heredity assumption* in the penalized model: an interaction term can be estimated to be non-zero if and only if its corresponding main effects are estimated to be non-zero, whereas a non-zero main effect does not necessarily imply a non-zero interaction term. This can be achieved by introducing a new set of parameters $\boldsymbol{\tau} = (\tau_1, \tau_2, ...\tau_p)$ and reparametrizing the coefficients for

the interaction terms ψ_j as the product of τ_j, β_j and ψ_0 such that $\psi_j = \psi_0\tau_j\beta_j$. In this way, strong heredity is assured and the following model is considered instead:

$$\mathcal{L}^*(\boldsymbol{Y};\boldsymbol{\theta}) = \frac{1}{2n}\|\sqrt{\boldsymbol{W}}\left(\boldsymbol{Y} - \psi_0\boldsymbol{A} - \sum_{j=1}^{p}\boldsymbol{X}_j\beta_j - \sum_{j=1}^{p}\underbrace{\psi_0\tau_j\beta_j}_{\psi_j}(\boldsymbol{A}\circ\boldsymbol{X}_j)\right)\|_2^2$$

where now the objective function is:

$$Q\left(\boldsymbol{\theta}\right) = \mathcal{L}^*(\boldsymbol{Y};\boldsymbol{\theta}) + \lambda(1-\delta)(\|\boldsymbol{\beta}\|_1 + |\psi_0|) + \lambda\delta\|\boldsymbol{\tau}\|_1. \tag{4.3}$$

It has been shown that pdWOLS inherits the double robustness property from dWOLS; the method relies on a model that includes adaptive weights, an efficient algorithm using cyclic coordinate descent to minimize the Equation (4.3), and can be generalized from the one-stage setting to a multi-stage setting (Bian et al., 2021).

4.3.2 Incorporating Confounder Selection into pdWOLS

Here we introduce a variable selection method for confounders in the propensity score model, called outcome-adaptive LASSO (OAL), which was proposed by Shortreed and Ertefaie (2017). This method considers both the treatment-covariate relationship and the outcome-covariates association, and hence avoids prioritizing possibly instrumental variables. The key to the approach is to apply data-adaptive weights to a penalized propensity score model, where the adaptive weights are estimates from a fit of an unpenalized outcome model. For example, we could posit a logit model $\pi(\boldsymbol{x};\boldsymbol{\alpha})$ for the PS, and $\widehat{\boldsymbol{\alpha}}$ is obtained by solving the adaptive penalized logistic regression:

$$\widehat{\boldsymbol{\alpha}}_{oal} = \sum_{i=1}^{n}\{log(1+\boldsymbol{x}_i^T\boldsymbol{\alpha}) - A_i\boldsymbol{x}_i^T\boldsymbol{\alpha}\} + \lambda_{oal}\sum_{j=1}^{p}\tilde{w}_j\alpha_j$$

where λ_{oal} is the tuning parameter, $\tilde{w}_j = |\widehat{\beta}_j|^{-\gamma}$ for some $\gamma \geq 1$, and the $\widehat{\boldsymbol{\beta}}$ is the estimated parameter of the baseline mean model $f(\boldsymbol{x}; \boldsymbol{\beta})$. Shortreed and Ertefaie (2017) proposed to select the tuning parameter λ_{oal} such that the weighted absolute mean difference between the exposure groups

$$\sum_{j=1}^{p} |\widehat{\beta}_j| \left| \frac{\sum_{i=1}^{n} \widehat{\zeta}_i X_{ij} A_i}{\sum_{i=1}^{n} \widehat{\zeta}_i A_i} - \frac{\sum_{i=1}^{n} \widehat{\zeta}_i X_{ij} (1 - A_i)}{\sum_{i=1}^{n} \widehat{\zeta}_i (1 - A_i)} \right|$$

is minimized, where $\widehat{\zeta}_i = A_i \widehat{\pi}(X_i; \widehat{\boldsymbol{\alpha}}_{oal})^{-1} + (1 - A_i)(1 - \widehat{\pi}(X_i; \widehat{\boldsymbol{\alpha}}_{oal}))^{-1}$. This approach to confounder selection for inclusion in the propensity score can be incorporated into pdWOLS in a straightforward manner.

4.3.3 Tailoring Variables Ranking Using pdWOLS

As mentioned in Section 4.2.3, we could rank the tailoring variables by the selected frequency among a sequence of tuning parameters $\lambda_1, \ldots, \lambda_k$, using pdWOLS, as the solution is calculated over this sequence, we have k different solutions in total. For each variable X_j, the selected frequency is calculated as $\sum_{t=1}^{k} I(\widehat{\psi}_{jt} \neq 0)/k$, and we can rank the tailoring variables accordingly.

With these new additions (confounder selection and variable ranking) into the pdWOLS approach, the method can be made even more parsimonious while offering additional insights into the importance of the tailoring variables when used jointly in a multivariate linear decision rule.

4.4 Numerical Studies

We now evaluate the performance of pdWOLS using two different tuning parameter selection approaches for the variable selection rate, and the resulting error rate in the estimated treatment decision as well as the value function of the estimated decision rules. We also

demonstrate its use as a method of ranking tailoring variables. All simulations use sample sizes of 200 and 500 respectively. Further, we consider pdWOLS both with and without the use of confounder selection in the propensity score.

Step 1: Generate 10 covariates ($\boldsymbol{X}_1 - \boldsymbol{X}_{10}$) where X_1 and X_{10} are Bernoulli random variables which take the value 1 with probability 0.5, and $X_2 - X_9$ are multivariate normal with zero mean, unit variance and correlation $Corr(X_j, X_{j'}) = 0.15^{|j-j'|}$ for $j, j' = 2, \ldots, 9$.

Step 2: Generate treatment such that $P(A_i = 1|x_1, x_2, x_3, x_4, x_5) = expit(1 + \sum_{j=1}^{5} x_j)$.

Step 3: Set the blip function as $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \sum_{j=1}^{3} \widehat{\psi}_j x_j)$ for $\psi_0 = 0.5, \psi_1 = -0.8, \psi_2 = -0.5$ and $\psi_3 = -0.5$, and hence the optimal treatment strategy, to depend only on $X_1 - X_3$.

Step 4: Set the baseline mean model to $f(\boldsymbol{x}; \boldsymbol{\beta}) = 0.5 - 0.6e^{x_1} - 2x_1 - 2x_2 + x_3 + 2x_6$.

Step 5: Generate the outcome $Y \sim N(f(\boldsymbol{x}; \boldsymbol{\beta}) + \gamma(\boldsymbol{x}, a; \boldsymbol{\psi}), 1)$.

In the data generating process, $X_1 - X_3$ are confounders, and X_4 and X_5 are instrumental variables. As argued in (De Luna et al., 2011; Greenland, 2008), the ideal propensity score model should include all the confounders and the predictors of outcome while excluding predictors of exposure not associated with outcome as well as all noise variables (covariates that are unrelated to both exposure and outcome). In our case, X_4 and X_5 were only used to generate the treatment, while X_6 is a predictor only of the outcome, hence we hope our propensity score model does not include X_4 and X_5 but does select X_6.

The estimation procedure proceeds as follows:

(a) Use OLS to regress y on $(1, x, ax)$ to obtain $\widehat{\beta}$.

(b) Construct the penalty factors (adaptive weights): $\tilde{w}_j = |\widehat{\beta}_j|^{-\chi}$ for some $\chi > 0$; solve the adaptive penalized logistic regression to select the confounders to include in the propensity score. The choice of χ can be arbitrary, and in this simulation we set it to be 3.

(c) Refit the propensity score model using the selected variables; this will yield the balancing

weights in the pdWOLS. Here, we used weights of the form $w(a, \boldsymbol{x}) = |a - \widehat{\mathbb{E}}[A|\boldsymbol{X} = \boldsymbol{x}]|$ proposed in Wallace and Moodie (2015).

(d) Apply pdWOLS: regress $\boldsymbol{Y}$ on $(\boldsymbol{1}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{AX})$ using the absolute weights obtained in step (c). Even though the baseline mean model is misspecified, the blip parameters can still be consistently estimated due to the double robustness property of pdWOLS.

The choices of tuning parameters in step (b) and step (d) are important. We adopt the ideas of Shortreed and Ertefaie (2017) to select the tuning parameter for outcome-adaptive LASSO such that the weighted absolute mean difference between the exposure groups is minimized. For pdWOLS, we set δ to be 0.5, and evaluate two different approaches, the value information criterion (VIC) (Shi et al., 2021) and the Bayesian information criterion (BIC) (Schwarz, 1978) to choose the λ such that either the VIC or the BIC is maximized. Analogous to the likelihood-based information criteria, $VIC = n\widehat{V}(\psi) - \kappa_n \|\psi\|_0$ where

$$\widehat{V}(\psi) = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{A_i \widehat{a}^{opt}(x_i;\psi) + (1-A_i)(1-\widehat{a}^{opt}(x_i;\psi))}{A_i \widehat{\pi}(x_i) + (1-A_i)(1-\widehat{\pi}(x_i))}$$

and κ_n is a positive sequence. Here we choose $\kappa_n = n^{1/3} log(p)^{2/3} log log(n)$ as this can achieve model selection consistency under certain conditions (Shi et al., 2021).

Figure 4.1 summarizes the estimates of blip parameters using pdWOLS. When $n = 500$, all the four blip parameters were nearly unbiased, regardless of the method to select the tuning parameter and whether we use OAL, which showed that pdWOLS is robust to the misspecification of the baseline mean model. Nevertheless, using OAL can reduce the variance of the estimators. Table 4.3 presents the standard error of the pdWOLS estimators using VIC and BIC with and without applying OAL when sample size is 200, we can see that the standard errors of the estimators obtained with OAL are uniformly smaller than these obtained without applying OAL, regardless using BIC or VIC to select the tuning parameter. When $n = 200$, the median of the estimator of ψ_3 obtained by BIC+OAL is about 0 as the coefficient is often shrunk to 0, while other estimators were similar to the

case when $n = 500$.

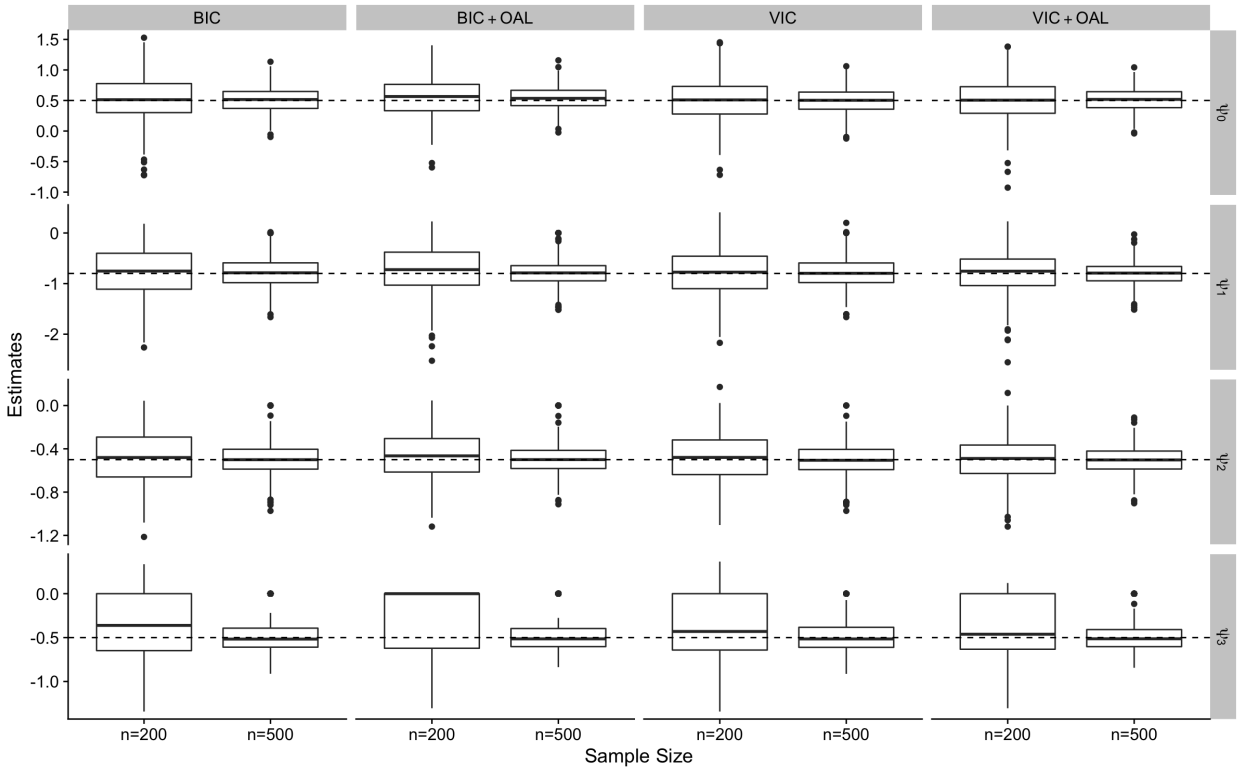

Figure 4.1: Estimates of blip parameters using pdWOLS with and without applying OAL to select the propensity score model with n =200 and 500; 500 simulations. The tuning parameter of pdWOLS was selected by BIC and VIC. The true value is represented by the dotted line.

Table 4.1: Variable selection rate (%) for pdWOLS using VIC and BIC with and without applying OAL (n =200 and 500; 500 simulations). The main effect of treatment is not penalized in the pdWOLS procedure. Variables with $*$ are the truly important variables, the rest are noise variables ($AX_3 - AX_{10}$).

	$n = 200$				$n = 500$			
	VIC+OAL	BIC+OAL	VIC	BIC	VIC+OAL	BIC+OAL	VIC	BIC
AX_1*	97	85	95	86	100	98	99	98
AX_2*	97	86	95	86	100	98	99	99
AX_3*	75	46	71	55	93	84	91	86
AX_4	8	2	10	6	8	1	14	2
AX_5	10	3	8	4	8	1	10	3
AX_6	78	54	72	60	85	64	83	68
AX_7	9	3	11	6	9	2	10	2
AX_8	10	3	10	5	8	1	10	3
AX_9	10	1	8	4	8	1	10	3
AX_{10}	11	2	11	5	7	1	11	2

Table 4.1 shows the variable selection results of pdWOLS with sample sizes of 200 and 500. In general, VIC has a lower false negative rate (i.e. excluding variables that are truly important) and a higher false positive rate (i.e. including variables that should be excluded) compared to BIC; OAL appears to improve the variable selection rate slightly. The irrelevant interaction AX_6 was frequently selected by all the methods, possibly due to the large main effect of X_6. When $n = 200$, the selection rate of the truly important interaction AX_3 was only 46% using BIC+OAL. Table 4.2 presents the error rate and the value function under the estimated rules of pdWOLS computed over a testing set of size 10,000. Despite a small difference for all the four methods, VIC+OAL has the smallest error rate as well as the largest and most accurate (smallest SE) value function.

Table 4.4 shows the variable selection rate for the propensity score using OAL. Note that the OAL selection is the same for both the BIC and VIC-selected penalty in pdWOLS since the confounder selection is a first step in a multi-step estimation procedure. All the truly important variables were correctly picked with a high selection rate, even in the smaller sample size of $n = 200$.

Our simulations suggest that pdWOLS achieves the best performance when combined with VIC and OAL, as it had lowest false negative rate, lowest error rate, and the highest value function. Moreover, applying OAL can improve the efficiency of the estimators.

Table 4.2: Error rate (ER) and value (standard error in parentheses) for pdWOLS using VIC and BIC with and without applying OAL (sample size 500 and 200, 500 simulations and test size 10,000). For comparison, the value function of the true optimal regime, always treated and never treated group are -1.15, -1.45 and -1.56, respectively.

	VIC+OAL	BIC+OAL	VIC	BIC
ER ($n = 200$)	0.18	0.23	0.20	0.23
ER ($n = 500$)	0.10	0.11	0.12	0.12
Value ($n = 200$)	-1.22 (0.07)	-1.25 (0.10)	-1.23 (0.08)	-1.25 (0.09)
Value ($n = 500$)	-1.17 (0.02)	-1.18 (0.05)	-1.18 (0.03)	-1.18 (0.04)

Table 4.3: Standard error of the pdWOLS estimators using BIC and VIC with and without applying OAL (sample size 200, 500 simulations).

	VIC+OAL	BIC+OAL	VIC	BIC
ψ_0	0.33	0.31	0.35	0.34
ψ_1	0.40	0.47	0.47	0.51
ψ_2	0.21	0.25	0.25	0.28
ψ_3	0.30	0.35	0.33	0.36

Table 4.4: Confounder selection rate using OAL. The variables with $*$ are the truly important variables for the propensity score model.

	$n = 200$	$n = 500$
X_1*	100	100
X_2*	100	100
X_3*	99	100
X_4	7	1
X_5	8	1
X_6*	97	98
X_7	3	0
X_8	1	0
X_9	2	0
X_{10}	13	0

The second focus of our simulations was to investigate pdWOLS as a tool for ranking tailoring variables. We do so according to the selected frequency among a sequence of tuning parameters, although as noted above, other approaches to ranking could also be used. Tables 4.5 and 4.6 summarize the proportion of times that tailoring variables are ranked over 500 replications using pdWOLS for n =200 and 500, respectively. From Tables 4.5 and 4.6, we see that pdWOLS ranked the three truly important tailoring variables in the top three with high frequency, however, the variable X_6 – which predicts the outcome but is not a tailoring variable (i.e., does not interact with treatment) was also ranked as third 49% of the time when sample size was equal to 500. This result follows the same pattern as Table 4.1, where X_6 was incorrectly selected into the model with high probability.

Table 4.5: The proportion of times that tailoring variables are ranked over 500 replications using pdWOLS (n =200), which is based on the selected frequency among a sequence of tuning parameters.

Ranking	1	2	3	4	5	6	7	8	9	10
X_1	88.2	9.8	1.2	0.6	0.2	0.0	0.0	0.0	0.0	0.0
X_2	11.2	87.2	0.6	0.8	0.0	0.2	0.0	0.0	0.0	0.0
X_3	0.0	1.0	38.0	59.0	1.2	0.4	0.0	0.0	0.2	0.2
X_4	0.0	0.0	0.0	2.0	19.6	18.8	16.6	12.0	17.4	13.6
X_5	0.0	0.0	0.0	1.8	15.2	16.8	14.6	16.6	17.0	18.0
X_6	0.6	2.0	60.2	32.4	2.4	1.8	0.4	0.0	0.0	0.2
X_7	0.0	0.0	0.0	0.6	17.0	15.8	15.2	17.4	17.4	16.6
X_8	0.0	0.0	0.0	1.2	15.0	17.0	18.0	15.0	18.6	15.2
X_9	0.0	0.0	0.0	0.4	14.8	14.8	19.4	19.0	15.0	16.6
X_{10}	0.0	0.0	0.0	1.2	14.6	14.4	15.8	20.0	14.4	19.6

Table 4.6: The proportion of times that tailoring variables are ranked over 500 replications using pdWOLS (n =500), which is based on the selected frequency among a sequence of tuning parameters.

Ranking	1	2	3	4	5	6	7	8	9	10
X_1	94.4	5.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_2	5.4	94.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_3	0.0	0.0	50.2	49.8	0.0	0.0	0.0	0.0	0.0	0.0
X_4	0.0	0.0	0.0	1.2	21.4	20.8	16.0	13.8	12.8	14.0
X_5	0.0	0.0	0.0	0.2	17.2	15.4	16.8	15.0	17.6	17.8
X_6	0.2	0.2	49.4	48.0	1.8	0.4	0.0	0.0	0.0	0.0
X_7	0.0	0.0	0.0	0.0	15.6	15.8	17.8	19.6	17.8	13.4
X_8	0.0	0.0	0.0	0.2	15.6	17.8	17.2	17.6	16.2	15.4
X_9	0.0	0.0	0.0	0.6	16.6	13.8	15.4	16.4	20.8	16.4
X_{10}	0.0	0.0	0.0	0.0	11.8	16.0	16.8	17.6	14.8	23.0

4.5 Application to the Study of an Adaptive Web-based Stress Management

In this section, we apply pdWOLS to data from a study of an adaptive web-based stress management tool. In this study, a pilot sequential multiple assignment randomized trial (SMART) was undertaken to assess the feasibility and potential effect size of a web-based, stress management intervention adapted over time based on a stepped-care approach for patients with a cardiovascular disease (CVD) patients (Lambert et al., 2021).

A two-stage SMART was piloted: at the first stage, 50 patients with CVD were randomized into two treatment groups with randomization stratified by recruitment source (clinical setting or community organization) and stress level (low or high). The randomization arms at the first stage were a website only group, and a website plus weekly telephone coaching group, each with probability 0.5. Stage two then followed after six weeks: patients who did

not derive the anticipated benefits (non-responders) were re-randomized to either continue with their first stage intervention or to switch to the use of the website plus motivational interviewing (MI) for another six weeks (again, randomization with probability 0.5 to each option); responders maintained their initial intervention from stage one. Responders were those participants whose stress scores improved by at least 50% or who were below the threshold of 16 on the stress subscale of the Depression Anxiety Stress Scales (DASS) (Lovibond and Lovibond, 1996).

We restrict our analysis to the first stage only, and hence a binary treatment is considered: $A = 0$ for the website only group and $A = 1$ for the website plus weekly telephone coaching group. The outcome of interest is the negative of the DASS measured at 6 weeks after stage 1 randomization; we target an individualized treatment rule that minimizes the DASS (conversely, maximizes the negative DASS) since this suggests the presence of fewer symptoms of stress. The aims of our analysis are to determine the potential tailoring variables that can improve the decision rule, and to then estimate the optimal treatment regime of the adaptive web-based stress management study based on these tailoring variables.

Table 4.7: Characteristics of the adaptive web-based stress management study population stratified by stage 1 treatment

	Website-only	Website+coach	SMD
n	25	25	
Age (mean (SD))	61.0 (13.4)	61.3 (10.7)	0.03
DASS Score at baseline (mean (SD))	19.2 (6.7)	18.9 (7.7)	0.03
Physical Component Score (mean (SD))	42.9 (12.2)	45.4 (11.6)	0.20
Mental Component Score (mean (SD))	40.4 (8.6)	40.9 (10.9)	0.04
Sex = Male (%)	13 (52.0)	9 (36.0)	0.33
Marital = Married / Common law (%)	14 (56.0)	15 (60.0)	0.08
Education = University degree (%)	13 (52.0)	16 (64.0)	0.25
Employment = Full time (%)	4 (16.0)	9 (36.0)	0.47
Chronic cardiac condition = Yes (%)	18 (72.0)	16 (64.0)	0.17
Chronic hypertension = Yes (%)	12 (48.0)	10 (40.0)	0.16
Chronic stomach condition = Yes (%)	10 (40.0)	12 (48.0)	0.16
Chronic vision condition = Yes (%)	11 (44.0)	7 (28.0)	0.34
Chronic backpain = Yes (%)	10 (40.0)	14 (56.0)	0.32
Chronic cholesterol condition = Yes (%)	9 (36.0)	11 (44.0)	0.16
Chronic obesity = Yes (%)	7 (28.0)	13 (52.0)	0.50
Chronic osteoarthritis = Yes (%)	9 (36.0)	8 (32.0)	0.09

Table 4.7 summarizes the baseline covariates of the 50 participants, including the standardized mean difference (SMD). Study groups were for the most part comparable with the exceptions of for sex, employment, having a chronic vision condition, chronic back pain and chronic obesity, where SMDs all exceeded 0.25. These differences in baseline characteristics are likely a reflection of the relatively small sample size. For this stratified randomization, a logistic regression model was used to estimate the treatment model (adjusting for recruitment source and stress level), as estimation and use of a parametric propensity score model can improve the efficiency of the estimator compared with using the known propensity score (Henmi and Eguchi, 2004); further, covariate adjustment can remove any potential bias due to chance imbalances between randomization groups. The minimum, median, mean and maximum value of the estimated propensity score are 0.42, 0.48, 0.49 and 0.58 respectively. There are 16 potential tailoring variables: DASS at baseline, age, sex as well as several other measures of physical health (see Table 4.7). We also include three sociodemographic

variables – education, employment, and marital status – which could potentially alter the impact of treatment.

Applying pdWOLS with tuning parameter selected using VIC to this study, we find that eight variables are useful for tailoring treatment (listed in decreasing order according to the pdWOLS variable ranking approach outlined in Section 4.3.3: mental component score (MCS), age, DASS at baseline, sex, marital status, stomach condition, physical component score (PCS), and vision. The estimated treatment rule is $\widehat{a}_1^{opt} = I\{38.3+0.2\text{age}-4.5I(\text{male})-15.4I(\text{unmarried})-1.0\text{DASS}+0.2\text{PCS}-0.7\text{MCS}-9.9I(\text{stomach=yes})+2.5I(\text{vision=yes}) > 0\}$. The estimated value of DASS and the 95% confidence interval (CI) is $13.8(10.6, 17.0)$ where the CI is calculated using 4,000 nonparametric bootstrap (Efron, 1979) resamples. For comparison, the estimated value of DASS for only web-coach and only website group are 15.0 and 18.5, respectively. No minimally important clinical difference has been established for the DASS, however a value of 18.5 is well above our targeted threshold of 16 points, which would suggest a meaningful difference between the tailored strategy and offering website only to all. The difference of 1.2 points on the DASS between the tailored strategy and the website with coaching is less striking, however it is worth noting that the tailored strategy is less resource intensive, offering the coaching to only a subset of the population while still potentially leading to reduced symptoms on average as compared to offering it to all under the untailored, web+coach strategy.

Since the sample size is relatively small, there can still be imbalances even if it is a randomized study. Thus we also apply OAL+pdWOLS to the study, with recruitment source, stress level and all the variables in Table 4.7 contained in the initial PS model. We find that seven variables are useful for tailoring treatment (listed in decreasing order according to the pdWOLS ranking): age, DASS at baseline, PCS, MCS, vision, sex, and employment status. Compare to the pdWOLS without applying OAL, it removes marital status and stomach condition and includes an extra tailoring variable employment status. The estimated

treatment rule is $\widehat{a}_2^{opt} = I\{16.7 + 0.4\text{age} - 0.6\text{DASS} - 0.1\text{PCS} - 0.5\text{MCS} - 4.7I(\text{vision=yes}) - 5.2I(\text{male}) - 7.1I(\text{employed}) > 0\}$. The estimated value of DASS and the 95% confidence interval (CI) is $14.8(11.8, 17.9)$. This yields a larger DASS score than the pdWOLS without applying OAL approach.

The results above suggested that a linear decision rule of eight tailoring variables (pdWOLS) is better than a "one size fits all" approach. Due to the large CI's of these estimated values and the small size of the pilot study, we cannot make any definitive conclusions regarding the benefit of tailoring compare with the only web+coach strategy, nevertheless, these estimated values hint at clinically meaningful benefit and the potential for cost-savings since not all patients need a more costly or intensive treatment.

4.6 Discussion

In this chapter, we reviewed a penalized likelihood-based variable selection method-pdWOLS, which inherits the double robustness property from the regression-based dWOLS. Numerical studies indicated that VIC is a better tuning parameter selection method than the BIC for pdWOLS. Additionally, it was demonstrated that a data-driven confounder selection approach for propensity score model construction can reduce the variance of the pdWOLS estimator. Further, pdWOLS functions well as a means of variable ranking, even with modest sample sizes.

Our analysis of adaptive web-based stress management pilot study data suggests that up to eight tailoring variables may be useful for the optimal treatment decisions, yielding a better clinical outcome than the "only web+coach" and the "only web" approaches (a DASS score of 13.4, 15.0 and 18.5 respectively). While the confidence intervals surrounding these estimated value functions were large and the pilot's size precludes making any definitive conclusions of benefit, these estimated values imply clinically meaningful benefits and provide cost savings, as not all individuals need more expensive treatment options, including non-professional

guidance. These results also underline the importance of considering comorbidities and social support in treating CVD patients' stress, and in collecting detailed information on these and related conditions in any subsequent full-scale SMART.

As shown here, OAL in combination with pdWOLS selects confounders for the propensity score model in a way that is decoupled from the selection of the tailoring variables, and ensures that instruments are unlikely to be included in the propensity score. Lastly, many of the variable selection methods for DTRs are based on penalized likelihood to screen out unneeded tailoring variables. However, DTRs are often estimated using estimating equations, and hence a quasi-likelihood framework is needed. Variable selection techniques in the context of estimating equation should be further studied, see, for example, Fu (2003) and Candes and Tao (2007).

There are several limitations and future directions for the methods discussed in this chapter. The pdWOLS method implemented in this chapter has been implemented only in the continuous outcome case; how to extend it to a more general setting such that the outcome is discrete is an interesting topic requiring further exploration. Post-selection inference should also be addressed; Zhao et al. (2022) proposed a valid tool to study the selective inference for effect modification using LASSO, which may be able to shed some light on how to combine the selection inferential tools with pdWOLS, and indeed with pdWOLS used in combination with OAL confounder selection.

Chapter 5

Variable Selection for Individualized Treatment Rules with Discrete Outcomes

Preamble to Manuscript 3.

In the first two manuscripts, I only considered the case in which the outcome is continuous. However, often the interest of outcome is discrete, for example, the STAR*D data and the adaptive web-based stress management data that we studied in Chapters 3 and 4, respectively. There is a growing literature for selecting variables in individualized treatment rules with continuous outcomes, but the topic has been little studied with discrete outcomes. In Manuscript 3, we propose a doubly robust variable selection method for individualized treatment rules with discrete outcomes. To our knowledge, doubly robust variable selection methods in ITR estimation for discrete outcomes has not been studied in the literature.

Variable Selection for Individualized Treatment Rules with Discrete Outcomes

Zeyu Bian[1], Erica EM Moodie[1], Susan M Shortreed[2,3],
Sylvie D Lambert[4,5] and Sahir Bhatnagar[1,6]

[1]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University*

[2]*Kaiser Permanente Washington Health Research Institute*

[3]*Department of Biostatistics, University of Washington*

[4]*Ingram School of Nursing, McGill University*

[5]*St. Mary's Research Centre*

[6]*Department of Diagnostic Radiology, McGill University*

Abstract

An individualized treatment rule (ITR) is a decision rule that aims to improve individual patients' health outcomes by recommending optimal treatments according to patients' specific information. In observational studies, collected data may contain many variables that are irrelevant for making treatment decisions. Including all available variables in the statistical model for the ITR could yield a loss of efficiency and an unnecessarily complicated treatment rule, which is difficult for physicians to interpret or implement. Thus, a data-driven approach to select important tailoring variables with the aim of improving the estimated decision rules is crucial. While there is a growing body of literature on selecting variables in ITRs with continuous outcomes, relatively few methods exist for discrete outcomes, which pose additional computational challenges even in the absence of variable selection. In this paper, we propose a variable selection method for ITRs with discrete outcomes. We show theoretically and empirically that our approach has the double robustness property, and that it compares favorably with other competing approaches. We illustrate the proposed method on data from a study of an adaptive web-based stress management tool to identify which variables are relevant for tailoring treatment.

Key Words: Double robustness; Individualized treatment rules; Penalization; Precision medicine; Variable selection; Weighted generalized linear model.

5.1 Introduction

In the precision medicine paradigm, treatment decisions are tailored to each individual rather than relying on a "one-size-fits-all" approach; this approach to treatment is beneficial in the presence of heterogeneous treatment effects. With the aim of improving individual patients' health outcomes, individualized treatment rules (ITRs) (Murphy, 2003; Robins, 2004; Chakraborty and Moodie, 2013; Kosorok and Moodie, 2015; Tsiatis et al., 2019) recommend effective treatments for each patient based on their specific characteristics. However, collected data often contain many variables that are irrelevant for tailoring treatment. Including all the variables as the tailoring variables in an analysis could reduce statistical efficiency (Hastie et al., 2009) due to the estimation of redundant variables that are not useful for tailoring treatment and yield an unnecessarily complicated treatment decision rule, which is difficult for physicians to interpret or implement. It is therefore important to develop variable selection methods with the objective of optimizing patients' outcomes by identifying useful tailoring variables. Variable selection for ITRs has been studied in Lu et al. (2013); Jeng et al. (2018); Shi et al. (2018); Bian et al. (2021); all of these works focus on penalized regression-based estimation methods. Lu et al. (2013) and Jeng et al. (2018) only considered a singly robust method in which the propensity score must be correctly specified. Shi et al. (2018) used the Dantzig selector directly to penalize the A-learning (Robins, 2004; Murphy, 2003) estimating equation; Bian et al. (2021) used penalized dynamic weighted ordinary least squares regression to perform variable selection. Both methods considered in Shi et al. (2018) and Bian et al. (2021) are doubly robust, i.e., they can yield consistent estimators while only requiring one of two nuisance models to be correct.

All the methods described above focused solely on cases in which the outcome is continuous. Discrete outcomes introduce additional computational challenges to the estimation of ITR and the variable selection procedure, due to the use of the non-identity link function. Existing literature focusing on discrete outcomes ITR estimation includes Q-learning (Chakraborty

and Moodie, 2013; Linn et al., 2017), Bayesian additive regression trees (Logan et al., 2019), and A-learning (Robins et al., 1992; Tchetgen Tchetgen et al., 2010). However, none of these previous works have been extended to include variable selection. In this article, we focus on developing doubly robust ITR estimation with variable selection for discrete outcomes.

To provide robustness against model misspecification, ITRs are often estimated using estimating equations (Murphy, 2003; Robins, 2004). There are at least two ways to achieve sparsity in the use of estimating equations: via a Dantzig selector (Candes and Tao, 2007) or by a regularized estimating equation (REE). Denote by $\boldsymbol{U}(\boldsymbol{\theta}) \in \mathbb{R}^p$ an estimating equation, where $\boldsymbol{\theta} \in \mathbb{R}^p$. The Dantzig estimator $\widehat{\boldsymbol{\theta}}_{dan}$ can be found by solving the constrained optimization problem: $\widehat{\boldsymbol{\theta}}_{dan} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1$, subject to $\|\boldsymbol{U}(\boldsymbol{\theta})\|_\infty \leq n\lambda$, where λ is a tuning parameter used to control sparsity, and n is the sample size. Another way to induce sparsity is to solve the REE: $\boldsymbol{U}(\boldsymbol{\theta}) = n\lambda q(|\boldsymbol{\theta}|)$, where $q(|\cdot|)$ is the subgradient of a penalty function $\rho(|\cdot|)$, i.e., $q(|\cdot|) = \partial\rho(|\cdot|)$. For example, LASSO (Tibshirani, 1996) regression defined by $\min_{\boldsymbol{\beta}}\{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + n\lambda\|\boldsymbol{\theta}\|_1\}$ is a special case of the REE $\boldsymbol{U}(\boldsymbol{\theta}) = n\lambda\partial\|\boldsymbol{\theta}\|_1$, where $\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix, and $\boldsymbol{Y} \in \mathbb{R}^n$ is the response.

While the Dantzig selector and REE work well for continuous outcomes (Shi et al., 2018), their implementation in ITRs can be difficult for discrete outcomes, which are usually modelled with non-identity link functions. Indeed, the existing doubly robust estimating equations to estimate ITRs for discrete outcomes are non-linear (Robins et al., 1992; Tchetgen Tchetgen et al., 2010, see later in Section 5.2.2), and hence the Dantzig selector cannot be solved using linear programming (James and Radchenko, 2009). As for REE, it has been studied in Johnson et al. (2008) and Wang et al. (2012), where local quadratic approximation (Fan and Li, 2001) was used to solve the REE, which is computationally burdensome since it requires the calculation of the inverse of the Hessian matrix. Finally, even if the solution of the Dantzig selector or the REE can be found, it is challenging to select the tuning parameter in an ITR context since interest lies in inference about treatment effects rather than just

predictive performance. This means that we cannot simply select the tuning parameter that has the lowest prediction error as in the more classical prediction setting.

In this work, we propose two new doubly robust estimating functions to estimate the ITR for count and binary outcomes, respectively. The benefit of our proposed estimating function is that it can be easily generalized to a penalized framework, which permits estimating the optimal treatment regimes and selecting the important variables simultaneously. We show that with a suitable choice of weights, a simple penalized regression model for estimating an ITR enjoys the desired double robustness property, and yet is straightforward to implement. The advantage of the newly proposed approach compared to alternative regularized ITRs estimation methods lies in the fact that it can be viewed from a minimization perspective. Hence, the implementation is simple, various penalty functions can be used, and the solution can be found using existing computationally efficient tools in standard software. A tuning parameter selection procedure is proposed to select the tuning parameter that addresses the fact that the goal of an ITR analysis is estimation of a decision rule rather than prediction. To our knowledge, doubly robust variable selection in ITR estimation for discrete outcomes has not been studied in the existing literature.

The rest of this article is organized as follows. In Section 2, we present some introductory concepts and review some existing doubly robust estimation methods for discrete outcomes. In Section 3, we introduce our proposed estimation methods, and we extend them to a penalized framework in Section 4, followed by statements of theoretical properties. A number of simulation studies are given in Section 5. Finally, we apply our method to data from an adaptive web-based stress management study in Section 6.

5.2 Background

5.2.1 Notations, Assumptions and Introductory Concepts

Throughout, we use upper case letters to denote random variables and lower case letters to denote observed values. We use non-bold letters to denote individual-level data and bold letters to denote all observations in the data, e.g., $X_i \in \mathbb{R}^p$ are the covariates for subject i, while $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ are covariates for all subjects. In a single stage ITR, $V_i = (X_i, A_i, Y_i)$ consists of the data for ith individual patient, where X_i is the patient's baseline covariates, A_i is the binary treatment received, and Y_i is the patient's outcome. In the sequel, we will suppress subscript i where it is clear. We denote the potential outcome (Rubin, 1978) under the treatment a as Y^a. The potential outcome Y^a of a subject is the outcome of the patient if treatment a has been taken, which may be counter to fact. The objective of an ITR analysis is to find the optimal treatment a^{opt} such that the expected potential outcome $\mathbb{E}(Y^a)$ is maximized across the population of individuals. To estimate ITRs, we assume the following standard causal assumptions hold: (1) the stable unit treatment value assumption (SUTVA) (Rubin, 1980): a patient's potential outcome is not affected by other patients' treatment assignments; (2) consistency: $Y = AY^1 + (1-A)Y^0$; (3) conditional exchangeability (Robins, 1997): $Y^a \perp\!\!\!\perp A|X = x$; and (4) positivity: $P(A = a|X = x) > 0$ almost surely for all x and $a = 0, 1$.

Finally, we assume that the observations $V_i,\ i = 1, \ldots, n$ are independent and identically distributed with probability density $h(V)$ with respect to a measure ν. Moreover, we assume the relationship between Y and (X, A) can be captured by a semi-parametric regression model: $g\big(\mathbb{E}(Y^a|X = x)\big) = g\big(\mathbb{E}(Y|X = x, A = a)\big) = f_0(x; \boldsymbol{\beta}) + \gamma\left(x, a; \boldsymbol{\psi}\right)$, where g is a known link function, f_0 is an unknown baseline function, and γ is a known function that satisfies $\gamma\left(x, 0; \boldsymbol{\psi}\right) = 0$, which is referred to as the blip function (Robins, 2004). A blip function can be interpreted as the difference on the linear predictor scale of the transformed

mean potential outcomes

$$\begin{aligned}\gamma(x,a) &= g\big(\mathbb{E}(Y^a|X=x)\big) - g\big(\mathbb{E}(Y^0|X=x)\big) \\ &= g\big(\mathbb{E}(Y^a|X=x,A=a)\big) - g\big(\mathbb{E}(Y^0|X=x,A=a)\big).\end{aligned}$$

In this model, f_0 is irrelevant for making treatment decisions (a nuisance model). Hence, our parameter of interest is $\boldsymbol{\psi}$, and the optimal ITR is given by

$$\begin{aligned}a^{opt} &= \arg\max_a \mathbb{E}(Y^a) = \arg\max_a \mathbb{E}_X\left[g^{-1}(f_0(x;\boldsymbol{\beta}) + \gamma(x,a;\boldsymbol{\psi}))\right] \\ &= \mathbb{E}_X\left[\arg\max_a\ g^{-1}(f_0(x;\boldsymbol{\beta}) + \gamma(x,a;\boldsymbol{\psi}))\right] = \arg\max_a \gamma(x,a;\boldsymbol{\psi}),\end{aligned}$$

given an increasing link function. In the sequel, we only consider a log link for count outcomes and a logit link for binary outcomes.

5.2.2 Existing Estimation Methods for Discrete Outcomes

A-learning for Count Outcomes

Denote by x^ψ the covariates in the blip model and by x^β the covariates in the baseline model; in what follows, the superscript is omitted if they are identical. Suppose that the blip function is known: $\gamma(x^\psi, a; \boldsymbol{\psi}) = a\boldsymbol{\psi}^T x^\psi$, and the link function is the log link. Then the A-learning estimating equation (Robins et al., 1992) for a count outcome is

$$\boldsymbol{U}_1(\boldsymbol{\psi}) = \frac{1}{n}\sum_{i=1}^n x_i^\psi (a_i - \widehat{\pi}_i)\exp\{-\gamma(x_i^\psi, a_i; \boldsymbol{\psi})\}\left(y_i - \exp(f(x_i^\beta; \widehat{\boldsymbol{\beta}}) + \gamma(x_i^\psi, a_i; \boldsymbol{\psi}))\right),$$

where f is the posited baseline model (not necessarily identical to f_0), $\widehat{\boldsymbol{\beta}}$ is a plug-in estimator, and $\widehat{\pi}$ is the estimated propensity score (Rosenbaum and Rubin, 1983). It can be shown that $\boldsymbol{U}_1(\boldsymbol{\psi})$ is an unbiased estimating equation (Robins et al., 1992), provided that at least one nuisance model (propensity score model or the baseline model) is correctly

specified.

A-learning for Binary Outcomes

Estimation is more complicated when the outcome is binary; the blip parameter is estimated by solving the following estimating equation when the link function is the logit link:

$$\boldsymbol{U}_2(\boldsymbol{\psi}) = \frac{1}{n}\sum_{i=1}^{n} x_i^{\psi}(a_i - \widehat{\pi}^*)\left(y_i - \text{expit}(f(x_i^{\beta};\widehat{\boldsymbol{\beta}}) + \gamma(x_i^{\psi}, a_i;\boldsymbol{\psi}))\right),$$

where

$$\widehat{\pi}^* = \left(1 + \frac{(1-\text{expit}(u(x;\widehat{\boldsymbol{\tau}}))\text{expit}(f(x;\widehat{\boldsymbol{\beta}}))}{\text{expit}(u(x;\widehat{\boldsymbol{\tau}}))\text{expit}(f(x;\widehat{\boldsymbol{\beta}}) + \gamma(x,a;\boldsymbol{\psi}))}\right)^{-1},$$

$\text{expit}(t) = \frac{\exp(t)}{1+\exp(t)}$, and $u(x;\boldsymbol{\tau})$ is the nuisance treatment model of $\mathbb{E}(A|Y=0,X)$. Tchetgen Tchetgen et al. (2010) showed that $\boldsymbol{U}_2(\boldsymbol{\psi})$ is an unbiased estimating equation when at least one of $\mathbb{E}(Y|X,A=0)$ or $\mathbb{E}(A|X,Y=0)$ is correctly specified. Note that for logit link, the quantity $\mathbb{E}(A|Y=0,X)$ is modeled instead of the propensity score to assure the double robustness property, because of the symmetry property of the odds ratio:

$$e^{X^\top\boldsymbol{\psi}} = \frac{P(Y=1|A=1,X)P(Y=0|A=0,X)}{P(Y=0|A=1,X)P(Y=1|A=0,X)} = \frac{P(A=1|Y=1,X)P(A=0|Y=0,X)}{P(A=0|Y=1,X)P(A=1|Y=0,X)}.$$

Chen (2007) showed that there are at least two ways to study the association parameter (in our case, the blip parameter): through the density of Y given X and A or through the density of A given X and Y. This provides an intuitive explanation of why $\mathbb{E}(Y|X,A=0)$ and $\mathbb{E}(A|X,Y=0)$ are modeled to assure the double robustness property.

As noted above, the implementation of the Dantzig selector or the REE can be difficult for the A-learning estimating function. In the next section, we propose an alternative estimation method that is also doubly robust and, unlike A-learning, can easily accommodate variable selection.

5.3 Doubly Robust Weighted Generalized Linear Model

In this section, we propose two new estimating equations for count and binary outcomes, respectively, and we show that solving these two estimating equations can be reformulated as an iteratively reweighted generalized linear model (IRGLM). Moreover, the obtained estimators are doubly robust, and the newly proposed estimating equation can be easily generalized to a penalized framework for variable selection.

5.3.1 Count Outcomes

From now on, we posit a linear model for the baseline function, i.e., $f(x;\boldsymbol{\beta}) = x^T\boldsymbol{\beta}$. For count outcomes, we present the following estimating function:

$$\boldsymbol{U}_3(\boldsymbol{\beta},\boldsymbol{\psi}) = \sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\} \left(y_i - \exp(f(x_i^{\beta};\boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi}))\right).$$

This equation takes a similar form to $\boldsymbol{U}_1(\boldsymbol{\psi})$ with the leading term $\exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\}$, and also shared a similar form to Wallace and Moodie (2015) using overlap weights, but is not identical to either.

Assumption 1. *When at least one of the two nuisance models π or f is correctly specified, there exists a unique population parameter $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)$ such that $\mathbb{E}[\mathbf{U}_3(\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)] = \mathbf{0}$.*

Theorem 5.3.1. *Assume that the SUTVA, ignorability, consistency, and positivity conditions described in the previous section and **Assumption 1** hold. If the posited baseline model satisfies $x^{\psi} \subseteq x^{\beta}$, then $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, where $\boldsymbol{\psi}_0$ is the underlying true blip parameter.*

Theorem 5.3.1 states that under standard causal assumptions, the population parameter $\boldsymbol{\psi}^*$ is equivalent to the true data-generating value of the blip (and corresponding ITR) parameter $\boldsymbol{\psi}_0$ if one of two nuisance models π or f is correctly specified. This implies that the blip estimator $\widehat{\boldsymbol{\psi}}$ obtained by solving $\boldsymbol{U}_3(\boldsymbol{\beta},\boldsymbol{\psi})$ is a doubly robust estimator.

Remark 5.3.1. *The condition of the existence of a unique population parameter is similar to the condition of the existence of the quasi-maximum likelihood estimate when the likelihood is misspecified (White, 1982). The assumption that $x^{\psi} \subseteq x^{\beta}$ in the posited model is referred to as the strong heredity assumption (Chipman, 1996): the corresponding main effects of an interaction term must be included in the model.*

Now we demonstrate that $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ can be specified as an IRGLM for which efficient computational solutions exist, and thus a penalized estimator can be constructed from the penalized generalized weighted linear model accordingly. We propose Algorithm 1 to solve $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$. The key is to treat the $|a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\}$ term in $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ as a constant in each iteration t. In this way, Step 7 in Algorithm 1 is equivalent to a weighted generalized linear model with weights $|a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \widetilde{\boldsymbol{\psi}}_t)\}$, where $\widehat{\pi}$ is the estimated propensity score which does not change across iterations and $\widetilde{\boldsymbol{\psi}}$ is the current value of the blip parameter estimate from the most recent iteration update. This can be solved efficiently using, for example, the `glm` function in **R** and specifying the `weights` argument.

Algorithm 1

1: **function** $(x_i, a_i, y_i, \widehat{\pi}_i, \varepsilon)$
2: Set iteration counter $t \leftarrow 0$
3: Initialize $\widetilde{\boldsymbol{\psi}}_0$
4: $w_{i0} \leftarrow |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widetilde{\boldsymbol{\psi}}_0)\}$ for $i = 1, \ldots, n$
5: **repeat**
6: Solve $\boldsymbol{\beta}_t$ and $\boldsymbol{\psi}_t$ such that
7: $\sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\boldsymbol{\psi}} \\ x_i^{\beta} \end{pmatrix} w_{it} \left(y_i - \exp(f(x_i^{\beta}; \boldsymbol{\beta}_t) + \gamma(x_i^{\psi}, a_i; \boldsymbol{\psi}_t)) \right) = 0$
8: $\widetilde{\boldsymbol{\psi}}_{t+1} \leftarrow \boldsymbol{\psi}_t$
9: $w_{i(t+1)} \leftarrow |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widetilde{\boldsymbol{\psi}}_{t+1})\}$
10: $t \leftarrow t + 1$
11: **until** $\|\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}\| < \varepsilon$

5.3.2 Binary Outcomes

A similar framework can be built for binary outcomes using the logit link function. We present estimating equation $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$ for binary outcomes:

$$\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi}) = \sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i^*| \left(y_i - \text{expit}(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi})) \right),$$

where

$$\widehat{\pi}^* = \left(1 + \frac{(1 - \text{expit}(u(x; \widehat{\xi}))\text{expit}(f(x; \widehat{\boldsymbol{\beta}}^*))}{\text{expit}(u(x; \widehat{\xi}))\text{expit}(f(x; \widehat{\boldsymbol{\beta}}^*) + \gamma(x, 1; \boldsymbol{\psi}))} \right)^{-1},$$

and $u(x; \xi)$ is the nuisance treatment model for $\mathbb{E}(A|Y = 0, X)$. Under mild conditions, the solution of $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$ is a doubly robust estimator. Note that all theoretical properties for count outcomes can be applied equally to binary outcomes; for convenience and space, we include the results of binary outcomes in the Appendix. Algorithm 2 can be used to solve $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$, once again treating the term $|a_i - \widehat{\pi}_i^*|$ as a constant in each iteration.

Algorithm 2

function $(x_i, a_i, y_i, \widehat{\pi}_i, \varepsilon)$
  Set iteration counter $t \leftarrow 0$
  Initialize: $\widetilde{\boldsymbol{\psi}}_0$
  $w_{i0} \leftarrow |a_i - \widehat{\pi}_i^*(\widetilde{\boldsymbol{\psi}}_0)|$ for $i = 1, \ldots, n$
  repeat
    Solve $\boldsymbol{\beta}_t$ and $\boldsymbol{\psi}_t$ such that
    $\sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} w_{it} \left(y_i - \text{expit}(f(x_i^{\beta}; \boldsymbol{\beta}_t) + \gamma(x_i^{\psi}, a_i; \boldsymbol{\psi}_t)) \right) = 0$
    $\widetilde{\boldsymbol{\psi}}_{t+1} \leftarrow \boldsymbol{\psi}_t$
    $w_{i(t+1)} \leftarrow |a_i - \widehat{\pi}_i^*(\widetilde{\boldsymbol{\psi}}_{t+1})|$
    $t \leftarrow t + 1$
  until $\|\boldsymbol{\psi}_t - \boldsymbol{\psi}_{t-1}\| < \varepsilon$

5.4 Tailoring Variable Selection

In this section, we introduce sparsity to our proposed estimating function using the formulation of a REE, and show that this REE is asymptotically equivalent to a penalized

weighted generalized linear model given an appropriate initial estimator. Throughout, the main effect of the treatment A is not penalized as our goal is to select the important tailoring variables.

5.4.1 Penalized Doubly Robust Method

Due to the non-linear part (log or logit link) of the estimating equation for discrete outcomes, a Dantzig selector with A-learning estimating equation $\boldsymbol{U}_1(\boldsymbol{\psi})$ or $\boldsymbol{U}_2(\boldsymbol{\psi})$ cannot be solved using linear programming (James and Radchenko, 2009). Hence, we pursue an REE approach to introduce sparsity to the proposed estimating equations $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ and $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$, and once again, reformulate the REE as a penalized weighted GLM. We call this approach the penalized doubly robust (PDR) method, as it will be shown later, the penalized estimator obtained by solving the ITR REE is a doubly robust estimator.

For count outcomes and binary outcomes, respectively, the ITR REE require finding the solution of

$$\sum_{i=1} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a; \boldsymbol{\psi})\} \left(y_i - \exp(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi})) \right) = n\lambda q(|\boldsymbol{\theta}|), \tag{5.1}$$

and

$$\sum_{i=1} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i^*| \left(y_i - \operatorname{expit}(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a; \boldsymbol{\psi})) \right) = n\lambda q(|\boldsymbol{\theta}|). \tag{5.2}$$

To estimate the blip parameters $\boldsymbol{\psi}$ consistently, we require that the penalized model satisfies the following properties: (a) no false exclusion of tailoring variables, and (b) the selected model has the strong heredity property, i.e., $\widehat{\psi}_j \neq 0 \implies \widehat{\beta}_j \neq 0$ (without loss of generality, assume that x^{ψ} has the same "ordering" as x^{β}). Many penalty functions can yield a model

that has variable selection consistency, i.e., no false inclusion and no false exclusion, for example, LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and adaptive LASSO (Zou, 2006). However, these methods all fail to achieve the strong heredity property, thus, further work is required to implement them in this setting. Bian et al. (2021), borrowing on the work in Choi et al. (2010) and Bhatnagar et al. (2020) used the reparametrization to ensure strong heredity when using penalization in the context of ITR. Here, we modify the adaptive LASSO penalty, and show that by using these modified adaptive weights, not only can the blip parameters be unbiasedly estimated (asymptotically), but also the strong heredity constraint can be met.

We omit the subscript for the estimating functions $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ and $\boldsymbol{U}_4(\boldsymbol{\beta}, \boldsymbol{\psi})$ for now, as the properties for both count and binary outcomes can be developed using a general notation $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi})$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\psi}_0)$ denote the underlying true parameters and recall that $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)$ is the unique population parameter such that $\mathbb{E}[\boldsymbol{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)] = 0$. Let s be the number of non-zero components of $\boldsymbol{\psi}_0$ (or equivalently, $\boldsymbol{\psi}^*$) and denote by S the set of indices of non-zero components for $\boldsymbol{\psi}_0$. Denote by S' the set of indices of non-zero components for $\boldsymbol{\beta}^*$; in order to satisfy the strong heredity property, we want the estimated baseline model to satisfy $\widehat{\boldsymbol{\beta}}_{\widetilde{S}} \neq 0$ as n goes to infinity, where $\widetilde{S} = S \cup S'$ (as such, $S \subseteq \widetilde{S}$ and hence strong heredity holds). The goal is to obtain a targeted indices set S^* such that $\widehat{\boldsymbol{\theta}}_{S^*} \neq 0$ and $\widehat{\boldsymbol{\theta}}_{S_c^*} = 0$ with probability tending to 1, where S_c^* is the complement of S^*. We note that this targeted set S^* can be written as $\boldsymbol{\theta}^*_{S^*} = (\boldsymbol{\beta}^*_{\widetilde{S}}, \boldsymbol{\psi}^*_S)$.

Suppose that we have an initial estimator $\widehat{\boldsymbol{\theta}}_{ini} = (\widehat{\boldsymbol{\beta}}_{ini}, \widehat{\boldsymbol{\psi}}_{ini})$ such that $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}^*\| = O_p(1)$ and $\sqrt{n}\|\widehat{\boldsymbol{\psi}}_{ini} - \boldsymbol{\psi}^*\| = O_p(1)$. Following the adaptive LASSO (Zou, 2006) principle, we construct our adaptive weights for the corresponding coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ as follows:

$$\widehat{w}_j^{\beta} = \left\{ \max\left(|\widehat{\beta}_j^{ini}|, |\widehat{\psi}_j^{ini}| \right) \right\}^{-1} \text{ and } \widehat{w}_j^{\psi} = \left| \widehat{\psi}_j^{ini} \right|^{-1}. \tag{5.3}$$

We then use the penalty function $\rho(|\boldsymbol{\theta}|) = \rho(|\boldsymbol{\beta}|) + \rho(|\boldsymbol{\psi}|)$, where

$$\rho(|\boldsymbol{\beta}|) = \sum_{j=1}^{p} \widehat{w}_j^{\beta} |\beta_j| \text{ and } \rho(|\boldsymbol{\psi}|) = \sum_{j=1}^{p} \widehat{w}_j^{\psi} |\psi_j|.$$

In this way, for non-zero coefficients of blip variables, the associated weights and those of their corresponding main effects both converge to finite constants, and thus always remain in the model. We refer to our proposed weights in Expression (5.3) as modified adaptive weights, since these build on the adaptive LASSO framework but differ from it in the choice of $\widehat{w}_j^{\beta}$. Theorem 5.4.1 establishes the existence of a $\sqrt{n}$-consistent solution to the ITR REE (5.1) and (5.2).

Theorem 5.4.1 (***Existence and Selection Consistency***)**.** *Assume that conditions in Theorem 5.3.1 hold, penalty functions are constructed using the modified adaptive weights described in Expression* (5.3)*, and the tuning parameter satisfies* $\sqrt{n}\lambda = o(1)$ *and* $n\lambda \to \infty$*, then there exists a* $\sqrt{n}-$*consistent solution* $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}})$ *of the ITR REE such that* $\widehat{\boldsymbol{\psi}}_S \neq 0$ *and* $\widehat{\boldsymbol{\psi}}_{S_c} = 0$*.*

By Lemma B.1.1 in the Appendix, to establish the existence of the REE solution, it suffices to show that for sufficiently large n, there exists a constant r such that on the boundary of a ball around $\boldsymbol{\theta}^*$ with radius $n^{-1/2}r$, the variational inequality holds for function $\boldsymbol{U}(\boldsymbol{\theta}) - n\lambda q(|\boldsymbol{\theta}|)$ with high probability; that is, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|=n^{-1/2}r} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T [\boldsymbol{U}(\boldsymbol{\theta}) - n\lambda q(|\boldsymbol{\theta}|)] > 0 \right) > 1 - \varepsilon.$$

This technique has been adopted in Portnoy (1984) and Wang (2011) to prove the existence of the M-estimator and GEE estimator when the number of predictors is large. Theorem 5.4.2 establishes the asymptotic normality of the ITR REE estimators under standard regularity conditions (see Appendix for the details).

Theorem 5.4.2 (***Asymptotic Normality***)**.** *For any* $\sqrt{n}-$*consistent solution* $\widehat{\boldsymbol{\theta}}$ *of ITR*

REE,

$$\sqrt{n}\boldsymbol{J}(\boldsymbol{\psi}_S^*)\{\widehat{\boldsymbol{\psi}}_S - \boldsymbol{\psi}_S^* + \boldsymbol{J}(\boldsymbol{\psi}_S^*)^{-1}\lambda\, q(|\boldsymbol{\psi}_S^*|)\} \to_d N\big(0, \mathbf{I}(\boldsymbol{\psi}_S^*)\big),$$

where $\mathbf{I}(\boldsymbol{\theta})$ *is the variance of the estimating equation* $\mathbf{U}(V, \boldsymbol{\theta})$*,* $\boldsymbol{J}(\boldsymbol{\theta})$ *is the quantity* $\mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{\partial \mathbf{U}_3(V,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$*, and* $\mathbf{I}(\boldsymbol{\psi}_S^*)$ *and* $\boldsymbol{J}(\boldsymbol{\psi}_S^*)$ *are the corresponding* $s \times s$ *sub-matrices of* $\mathbf{I}$ *and* $\boldsymbol{J}$ *evaluated at the truth.*

A detailed proof of Theorem 5.4.1 and Theorem 5.4.2 can be found in the Appendix. In order to illustrate the double robustness property of our proposed estimators, we borrow the idea of the oracle estimator (Fan and Li, 2001). Define the oracle estimator $\widehat{\boldsymbol{\psi}}_{ora} \in \mathbb{R}^s$ as the solution of $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi})$ using $f(x_{\widetilde{S}})$ and $\gamma(x_S, a)$ (i.e., assume that the zero and non-zero of the coefficients are known in advance). Since we do not know the truly important variables in the application, the oracle estimator is just a conceptual idea to help us establish the theoretical properties in variable selection. Due to the double robustness of $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\psi})$, $\widehat{\boldsymbol{\psi}}_{ora}$ is a consistent asymptotically normal estimator of $\boldsymbol{\psi}_S^*$ under standard regularity conditions (see Appendix for the details) for M-estimators. The properties of $\widehat{\boldsymbol{\psi}}$ in Theorems 5.4.1 and 5.4.2 are referred to as the oracle property (Fan and Li, 2001), i.e., $\widehat{\boldsymbol{\psi}}$ performs as well as the oracle estimator $\widehat{\boldsymbol{\psi}}_{ora}$.

Corollary 5.4.1 (***Double Robustness***)**.** *It can be seen that the oracle estimator* $\widehat{\boldsymbol{\psi}}_{ora}$ *constructed above is a doubly robust estimator of* $\boldsymbol{\psi}_0$*. Since the resulting estimator* $\widehat{\boldsymbol{\psi}}$ *mimics the oracle estimator* $\widehat{\boldsymbol{\psi}}_{ora}$*,* $\widehat{\boldsymbol{\psi}}$ *is also a doubly robust estimator. That is to say, the resulting estimator* $\widehat{\boldsymbol{\psi}}$ *is a consistent estimator of* $\boldsymbol{\psi}_0$ *if either one of two nuisance models is correct.*

5.4.2 A One-step Estimator

In the setting where the number of variables p is fixed, we present an approximation to solve the ITR REE (5.1) in one-step. Suppose that we can find an initial estimator $\widehat{\boldsymbol{\psi}}_{ini}$ of the blip parameter, such that $\sqrt{n}\|\widehat{\boldsymbol{\psi}}_{ini} - \boldsymbol{\psi}^*\|_2 = O_p(1)$. Then we could plug in $\widehat{\boldsymbol{\psi}}_{ini}$ to the weight term of Expression (5.1) and solve it directly, which is equivalent to maximizing a weighted

penalized likelihood. Taking the count outcomes for example, we could use the solution of the unpenalized estimating equation $\boldsymbol{U}_1(\boldsymbol{\theta})$ or $\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\theta})$ as the initial estimator. Then under mild conditions, using $\widehat{\boldsymbol{\psi}}_{ini}$ as a plug-in estimator will have a negligible effect on the resulting estimator $\widehat{\boldsymbol{\psi}}$. That is, the solution of

$$\sum_{i=1} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widehat{\boldsymbol{\psi}}_{ini})\} \left(y_i - \exp(f(x_i^{\beta}; \boldsymbol{\beta}) + \gamma(x_i^{\psi}, a_i; \boldsymbol{\psi})) \right) = n\lambda \partial \rho(|\boldsymbol{\theta}|)$$

is asymptotically equivalent to the solution of (5.1). In the high dimensional setting where an unpenalized initial estimator cannot easily be computed, the ridge penalty can be used to obtain the initial estimator.

5.4.3 Tuning Parameter Selection

The choice of the tuning parameter λ in Expressions (5.1) and (5.2) plays an important role in the performance of the REE: an inappropriately large or small value of λ will greatly weaken the performance of the resulting estimator with respect to the estimation error and variable selection results. In penalized likelihood, where the goal is prediction, the optimal λ is often chosen in a way such that the corresponding model has the lowest information criterion, usually estimated by a measure of model fit (e.g., negative log-likelihood) with an extra penalty term, eg., the Akaike information criterion (AIC) (Akaike, 1974) or the Bayesian information criterion (BIC) (Schwarz, 1978). However, using AIC or BIC to select the tuning parameter would fail if the likelihood is misspecified (i.e., outcome model is misspecified). An alternative approach is to replace the negative log-likelihood part in the information criterion with the negative estimated value function (Qian and Murphy, 2011; Zhao et al., 2012; Shi et al., 2021):

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \mathbb{I}(A_i = \widehat{a}_i^{opt})}{A_i \widehat{\pi}_i + (1 - A_i)(1 - \widehat{\pi}_i)},$$

where $\mathbb{I}(\cdot)$ is the indicator function; note that the term value function refers to the marginal or population average outcome under a particular treatment strategy–in this case, under the strategy defined by the estimated optimal ITR, $\widehat{a}^{opt}$. However, this requires that the propensity score model is correctly specified. Here we propose an approach to select the tuning parameter while only requiring one of the nuisance models to be correctly specified. We refer to this approach as being a doubly robust criterion.

Recall that our proposed method can be viewed from a minimization perspective, i.e., $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}\{\mathcal{L}_n(\boldsymbol{\theta}; \boldsymbol{y}) + n\lambda\,\rho(|\boldsymbol{\theta}|)\}$. Following the idea used in classical information criteria (Schwarz, 1978; Akaike, 1974; Nishii, 1984), we propose to select the tuning parameter by choosing the model that has the smallest value of $n^{-1}[D_\lambda(\widehat{\boldsymbol{\theta}}, \boldsymbol{y}) + \kappa_n s_\lambda]$, where $D_\lambda(\widehat{\boldsymbol{\theta}}, \boldsymbol{y}) = 2[\mathcal{L}_n^{sat}(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) - \mathcal{L}_n(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})]$ is the quasi-deviance, $\mathcal{L}_n^{sat}$ is the quasi-log-likelihood of the saturated model, κ_n is some positive sequence, and s_λ is the number of non-zero components in the model, for a given λ. We suggest to set κ_n as $\log(\log n)\log p$ following (Fan and Tang, 2013), as this can achieve model selection consistency in a penalized likelihood setting. In practice, we could also use cross-validation to choose the tuning parameter that corresponds to the lowest average loss $\mathcal{L}_n^{cv}(\widehat{\boldsymbol{\theta}}; \boldsymbol{y})$.

5.5 Numerical Studies

In this section, we first illustrate the double robustness of our proposed method as well as how the choice of the initial estimator can impact the resulting estimators, in the presence of a small number of predictors; then we demonstrate the proposed method in a large dimension setting.

5.5.1 Experiments Examining the Double Robustness Property in Low Dimension

Recall that in Section 5.4.2, the initial estimator can be obtained from A-learning or our proposed IRGLM. We now evaluate the performance of our proposed PDR method using two different initial estimators, with respect to the variable selection rate, and the resulting error rate in the estimated treatment decision as well as the value function (expected outcome) of the estimated decision rules. The error rates and the average value function were calculated over a testing set of size 10,000.

The data generation procedure for count outcomes is as follows: Step 1: Generate 15 independent multivariate normal covariates $(X_1, \ldots, X_{15})$ with mean equal to 0.5 and unit variance. Step 2: Generate treatment such that $P(A = 1|x_1, x_2) = \text{expit}(-0.2 + \sum_{j=1}^{2} x_j)$. Step 3: Set the blip function as $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \psi_1 x_1)$ for $\psi_0 = 1$ and $\psi_1 = -2$. Step 4: Set the baseline model to $f(\boldsymbol{x}; \boldsymbol{\beta}) = \exp(-x_1^2 - x_2^2 + x_3 - x_4) + x_1 - 0.2x_2$. Step 5: Generate the outcome $Y \sim \text{Poisson}(\exp(f(\boldsymbol{x}; \boldsymbol{\beta}) + \gamma(\boldsymbol{x}, a; \boldsymbol{\psi})))$. Under this data generation procedure, the optimal treatment is $\mathbb{I}(1 - 2x_1 > 0)$ which corresponds to treatment $A = 1$ for about 50% of subjects; and the marginal mean of the outcome under observed (rather than optimal) treatment is 1.21.

The data generation procedure for binary outcomes is the same for steps 1-3 above, except now we set the nuisance treatment model as $\mathbb{E}(A|Y = 0, X = x) = \exp(-x_1^2 - x_2^2 + x_3 - x_4) + x_1 - 0.2x_2$, and marginalize the conditional expectation over the distribution of Y to obtain the propensity score model $\mathbb{E}(A|X = x)$. Generate the outcome $Y \sim \text{Bernoulli}(\text{expit}(f(\boldsymbol{x}; \boldsymbol{\beta}) + \gamma(\boldsymbol{x}, a; \boldsymbol{\psi})))$. Under this data generation procedure, the optimal treatment corresponds to treatment $A = 1$ for about 50% of subjects; and the marginal mean of the outcome under observed (rather than optimal) treatment is 0.47.

For both count outcomes and binary outcomes, we consider three scenarios with two sample sizes (500 and 1000), where the baseline model is misspecified in scenario 1, and the treatment

model is misspecified in scenario 2. In scenario 3, both models are correctly specified. As the number of predictors is small in this experiment, we compare our proposed PDR with unpenalized doubly robust A-learning. For PDR, we consider two alternative initial estimators: in the first case, referred to as PDR1, it is obtained from A-learning; and in the second, PDR2, from our proposed IRGLM approach. The **R** package `drgee` (Zetterqvist and Sjölander, 2015) is implemented to obtain the A-learning estimates.

Tables 5.1 and 5.2 present the error rate (proportion of times the estimated optimal ITR fails to coincide with the true optimal ITR), value, false negative rate (i.e., setting a tailoring variable's coefficient to 0 when it should be non-zero), false positive rate (i.e., selecting a tailoring variable, when the coefficient should be in fact be zero), mean absolute error (MAE): $\|\boldsymbol{\psi}_0 - \widehat{\boldsymbol{\psi}}\|_1$ and mean squared error (MSE): $\|\boldsymbol{\psi}_0 - \widehat{\boldsymbol{\psi}}\|_2$ of the blip parameter estimates of the three methods for binary and count outcomes respectively. In summary, all three methods have good performance (as they are all doubly robust methods), however, our proposed PDR1 and PDR2 outperform the unpenalized method with respect to the error rate, value, and estimation error for both types of outcomes in all three scenarios regardless of the sample size. No obvious difference in the error rate, value, and variable selection performance were observed between PDR1 and PDR2 in the simulations. Nonetheless, in scenario 2 for binary outcomes, the estimation error (MAE and MSE) is slightly smaller for PDR1 than PDR2, but this difference in estimation error does not translate into a noticeable difference with respect to the error rate or the value function.

Table 5.1: Error rate (ER), value, false negative (FN) rate, and false positive (FP) rate of variable selection results, mean absolute error (MAE) and mean squared error (MSE) using unpenalized estimation (UE) and penalized doubly robust methods (PDR1 and PDR2), with $n = 500$ and 1000, for 400 simulations and a test size 10,000 in three scenarios for a count outcome. For comparison, the value function of the true optimal regime, and the strategies of always treat and never treat are 3.36, 1.82, and 2.08, respectively.

		Scenario 1			Scenario 2			Scenario 3		
		UE	PDR1	PDR2	UE	PDR1	PDR2	UE	PDR1	PDR2
n=500										
	ER	0.13	0.07	0.08	0.09	0.03	0.03	0.12	0.07	0.07
	Value	3.28	3.34	3.33	3.33	3.36	3.36	3.29	3.34	3.34
	FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	FP	1.00	0.16	0.19	1.00	0.04	0.01	1.00	0.19	0.18
	MAE	3.60	1.50	1.50	2.40	0.30	0.30	3.30	1.35	1.35
	MSE	1.80	0.75	0.75	0.75	0.06	0.06	1.35	0.75	0.60
n=1000										
	ER	0.09	0.04	0.04	0.06	0.03	0.03	0.08	0.04	0.04
	Value	3.33	3.36	3.35	3.35	3.36	3.36	3.33	3.35	3.35
	FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	FP	1.00	0.07	0.08	1.00	0.01	0.00	1.00	0.13	0.13
	MAE	2.54	0.82	0.85	1.66	0.18	0.18	2.32	0.91	0.89
	MSE	0.91	0.39	0.39	0.33	0.03	0.03	0.70	0.35	0.33

Table 5.2: Error rate (ER), value, false negative (FN) rate, and false positive (FP) rate of variable selection results, mean absolute error (MAE) and mean squared error (MSE) using unpenalized estimation (UE) and penalized doubly robust methods (PDR1 and PDR2), with $n = 500$ and 1000, for 400 simulations and a test size 10,000 in three scenarios for a binary outcome. For comparison, the value function of the true optimal regime, and the strategies of always treat and never treat are 0.64, 0.48, and 0.48, respectively.

		Scenario 1			Scenario 2			Scenario 3		
		UE	PDR1	PDR2	UE	PDR1	PDR2	UE	PDR1	PDR2
n=500										
	ER	0.18	0.07	0.07	0.18	0.07	0.08	0.18	0.07	0.07
	Value	0.61	0.64	0.64	0.61	0.64	0.64	0.61	0.64	0.64
	FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	FP	1.00	0.05	0.05	1.00	0.04	0.08	1.00	0.06	0.06
	MAE	6.09	1.07	1.10	6.00	0.86	1.31	6.29	1.14	1.15
	MSE	4.20	0.66	0.69	3.97	0.41	0.81	4.43	0.69	0.69
n=1000										
	ER	0.13	0.05	0.05	0.13	0.04	0.05	0.13	0.04	0.04
	Value	0.63	0.64	0.64	0.63	0.64	0.64	0.63	0.64	0.64
	FN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	FP	1.00	0.03	0.02	1.00	0.03	0.05	1.00	0.03	0.03
	MAE	3.81	0.67	0.66	3.85	0.56	0.76	3.89	0.66	0.67
	MSE	1.59	0.28	0.29	1.61	0.19	0.31	1.65	0.27	0.27

5.5.2 Large Dimension Setting

In this section, we decrease the sample size to $n = 300$ and increase the number of covariates p to 30, 60, and 100 (note that the dimension of the model is $2p + 2$, since $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ are estimated simultaneously in the model). Since the number of predictors is now large, we compare our doubly robust method with outcome regression using the LASSO (Tibshirani, 1996) penalty. The LASSO is implemented using the **R** package `glmnet` (Friedman et al., 2007), with the tuning parameter is selected using the information criterion discussed in Section 5.4.3. Finally, we obtain our initial estimator using the ridge penalty; see Algorithm A1 in the Appendix for implementation details.

The data generation procedures for both types of outcomes are similar to the previous low

dimensional setting, except for some changes in the values of the parameters (see Appendix for more details). For both count and binary outcomes, we consider the challenging scenario in which the baseline model is misspecified and the treatment model is correctly specified. Tables 5.3 and 5.4 present the error rate, value, false negative rate, false positive rate, MAE, and MSE of the two methods for the count and binary outcomes, respectively. Our proposed penalized doubly robust method (using the ridge estimator as the initial estimate) outperforms the LASSO with respect to the error rate, value, and MSE for both types of outcomes, regardless of the number of predictors. However, for binary outcomes, the MAE is slightly smaller for LASSO than PDR.

Table 5.3: Error rate (ER), value, false negative (FN) rate, and false positive (FP) rate of variable selection results, mean absolute error (MAE) and mean squared error (MSE) using LASSO and PDR with $n = 300$, for 400 simulations and a test size 10,000 for count outcome. For comparison, the value function of the true optimal regime, and the strategies of always treat and never treat are 2.01, 0.79, and 1.36, respectively.

	$p = 30$		$p = 60$		$p = 100$	
	LASSO	PDR	LASSO	PDR	LASSO	PDR
ER	0.15	0.09	0.17	0.11	0.18	0.11
Value	1.98	2.00	1.97	1.99	1.95	1.99
FN	0.06	0.00	0.07	0.01	0.06	0.00
FP	0.23	0.18	0.11	0.10	0.07	0.07
MAE	3.04	2.09	3.46	2.42	3.67	2.58
MSE	2.62	0.94	3.32	1.39	3.73	1.49

Table 5.4: Error rate (ER), value, false negative (FN) rate, and false positive (FP) rate of variable selection results, mean absolute error (MAE) and mean squared error (MSE) using LASSO and PDR with $n = 300$, for 400 simulations and a test size 10,000 for binary outcome. For comparison, the value function of the true optimal regime, and the strategies of always treat and never treat are 0.57, 0.42, and 0.29, respectively.

	$p = 30$		$p = 60$		$p = 100$	
	LASSO	PDR	LASSO	PDR	LASSO	PDR
ER	0.30	0.21	0.34	0.23	0.37	0.27
Value	0.47	0.52	0.44	0.50	0.43	0.48
FN	0.34	0.07	0.47	0.09	0.51	0.12
FP	0.02	0.21	0.00	0.12	0.00	0.09
MAE	8.33	8.73	8.79	9.70	8.85	12.41
MSE	27.55	16.98	30.82	19.92	31.36	29.42

5.6 Application to an Adaptive Web-based Stress Management Study

We illustrate the newly proposed approach on a dataset from a two-stage pilot sequential multiple assignment randomized trial (Lambert et al., 2021) that aimed to assess a web-based, stress management intervention adapted across time using a stepped-care approach for people with cardiovascular disease (CVD). We focus our analysis on the first stage only, where 50 participants were randomized into two treatment groups each with probability 0.5, stratified by recruitment sources and stress level. The two treatment groups are the website only group ($A = 0$) and the website plus weekly telephone coaching group ($A = 1$).

The primary outcome in this analysis was the stress subscale from the Depression Anxiety Stress Scales (DASS) (Lovibond and Lovibond, 1996), which is a count outcome measured at 6 weeks after stage 1 randomized allocation. As a lower DASS-stress subscale score suggests the presence of fewer symptoms of stress, the optimal treatment decision is made such that one minimizes the DASS-stress subscale score. The aims of our analysis are to determine the tailoring variables related to the decision rule and to obtain the estimated individualized

treatment rule for CVD patients. We restrict our analysis to eight variables: mental component score (MCS), age, DASS-stress subscale score at baseline, sex, marital status, stomach condition, physical component score (PCS), and vision, as they were previously found to be useful for tailoring treatment using Bian et al. (2021).

A logistic regression model is posited to estimate the propensity score adjusted for the recruitment source and stress level. We apply PDR to this study with A-learning as the initial estimator (this approach was referred to as PDR1 in Section 5.5); both the baseline model and the blip model are posited to be linear. We found that five variables are relevant for tailoring treatment: DASS at baseline, sex, marital status, stomach condition, and vision. The estimated treatment rule is

$$\widehat{a}^{opt} = \mathbb{I}\{ -0.78 + 0.09\mathbb{I}(\text{male}) + 0.45\mathbb{I}(\text{unmarried}) + 0.01\text{DASS} + 0.45\mathbb{I}(\text{stomach=yes}) - 0.08\mathbb{I}(\text{vision=yes}) < 0\}.$$

For example, a married woman who does not have either a vision problem nor a stomach ailment, and who has a DASS greater than 13, would be recommended for website plus weekly telephone coaching ($A = 1$). We compared our estimated treatment rule with the results using the approach in Bian et al. (2021), wherein the DASS was treated as a continuous measure, and we found that 74% of the subjects' recommended treatments were the same under the two strategies. Moreover, all five non-zero estimated blip parameters had the same signs as the estimated blip parameters using Bian et al. (2021).

We also considered, for illustrative purposes, an analysis dichotomizes the outcome Y by its median, and used our proposed binary outcome approach. However, due to the small sample size, neither A-learning nor a standard logistic regression yielded a solution due to lack of convergence.

Finally, we illustrate our newly proposed approach on data from the sequenced treatment

alternatives to relieve depression (STAR*D) (Fava et al., 2003). The STAR*D data are considered a benchmark dataset for ITR analyses and have been analyzed in Chakraborty et al. (2013); Shi et al. (2018); Wallace et al. (2019); Bian et al. (2021), among others. While these data are less novel, we considered the comparison relevant and provide results in the Appendix. In summary, the findings in the current analysis using the methods proposed here for both count and binary outcomes align well with the results found in Chakraborty et al. (2013); Wallace et al. (2019); Bian et al. (2021).

5.7 Discussion

In this article, we propose new doubly robust estimating functions to determine an individualized treatment rule when the outcome is discrete and the log or logit link functions are used to model the outcome. The newly proposed approach can be solved using a weighted GLM iteratively given a suitable choice of observational weights. The benefit of our proposed estimating function is that it can be easily generalized to a penalized framework, which permits estimating a parsimonious ITR and selecting the important tailoring variables simultaneously. Based on this, we also present a doubly robust criterion to select the tuning parameter. Numerical studies indicated that the newly proposed penalized doubly robust method compares favorably with other competing approaches in the context of ITRs. To our knowledge, doubly robust variable selection approach for ITRs has not previously been studied.

To obtain a doubly robust estimator, we need a well-behaved initial estimator, which can be found using an unpenalized doubly robust approach. In the setting where the number of predictors is larger than the sample size, we recommend using the ridge estimator to acquire the initial estimate. In future work, we could also build on idea in Huang et al. (2008), who used the marginal regression approach to obtain the initial estimator for the adaptive LASSO, i.e, the outcome is regressed separately on each variable. However, this

technique is more challenging in our setting, as it violates the assumption that the blip model is correctly specified. This is a partial identification problem, and it has been studied in van der Laan et al. (2003), which may be able to shed some light on how to use marginal regression to obtain a valid initial estimator. It also may be of interest, in future work, to investigate the algorithm to directly solve the REE instead of using the approximation. As this alternative does not require an initial estimator, it might perform better in a large p, small n scenario.

The extension of the single stage estimation approach to a multi-stage setting also requires further investigation. In a multi-stage setting, the estimation procedure is conducted recursively using backward induction, and the "outcome" at each stage is set to be a predicted or estimated optimal response. For discrete outcomes, the optimal outcome is usually modeled by multiplicative effects, e.g., the optimal outcome at the $(k-1)$th stage for a count outcome is computed by $\widehat{y}_{k-1}^{opt} = y \times \prod_k^K \exp\{\gamma_k(x_k^{\psi}, \widehat{a}_k^{opt}; \boldsymbol{\psi}_k) - \gamma_k(x_k^{\psi}, a_k; \boldsymbol{\psi}_k)\}$, where K is the total number of stages. A challenge under the multi-stage scenario is that the estimated optimal outcome at any stage for subjects with zero-valued outcome will always remain zero, unless adjustments are made (Wallace et al., 2019), which may lead to a loss of efficiency. Another challenge is that $\widehat{y}_{k-1}^{opt}$ will typically not be discrete anymore; how to apply our proposed method to this setting is an open and intriguing problem.

Chapter 6

Conclusion

6.1 Summary

The three manuscripts (Chapters 3, 4 and 5) presented in this thesis describe a body of work in the context of variable selection methods for optimal treatment decision-making. The first contribution (Chapter 3) generalizes the methodology in (Wallace and Moodie, 2015) to a penalized framework, where variable selection and the estimation of the optimal treatment decision can be conducted simultaneously. Through a simple re-parametrization and the use of adaptive LASSO weights, the newly proposed method possesses the double robustness property. The simulation studies have shown that this extended approach performs well both in single stage and multi-stage settings, with respect to estimation error, variable selection performance (false negative and false positive rate), error rate, and value function estimation. I further illustrated the newly proposed method using the benchmark STAR*D study (Fava et al., 2003). My analysis suggests that the optimal treatments for patients are treat with SSRI at stage 1 and treat with a non-SSRI at stage 2, for all patients. That is, my results suggest that tailoring is not beneficial in the context of choosing an optimal class of antidepressant among patients first treated with and not responding to citalopram.

In the second manuscript (Chapter 4), a confounder selection method and a tuning parameter selection technique are combined with the approach from the first paper (Chapter 3). Simulation studies indicated that the newly proposed method achieves the best performance when combined with the use of the value information criterion to choose the tuning parameter for tailoring variable selection, and outcome adaptive LASSO to choose confounders. Moreover, it was demonstrated that a data-driven confounder selection approach for treatment model construction can improve the efficiency of the resulting estimators. This approach is further illustrated on data from a pilot sequential multiple assignment randomized trial of a web-based stress management study. Our analysis suggests that up to eight variables may be relevant for the individual treatment rules, yielding a better clinical outcome than the "one-size-fits all" approach. This result implies clinically meaningful benefits and cost savings, since only some patients require the more expensive treatment strategy. These results also attach importance to collecting detailed information on selected variables and related information in any subsequent full-scale sequential multiple assignment randomized trial.

The final manuscript (Chapter 5) extended the approach from Chapter 3 to a setting in which the outcomes are discrete. I showed that a simple penalized weighted regression approach can have the desired double robustness property, given a suitable choice of weights. The benefit of the newly proposed approach compared to alternative regularized ITRs estimation methods (in particular, the Dantzig selector or a regularized estimating equation) lies in the fact that it can be viewed from a minimization perspective so that the implementation is simpler, and the solution can be found using existing computationally efficient tools. Simulation studies suggested that my proposed penalized doubly robust method compares favorably with other competing approaches in the context of ITRs. This approach was again illustrated on data from the STAR*D study (the same data as in Chapter 3) and the adaptive web-based stress management tool pilot study (the same data as in Chapter 4). The findings in this work aligned well with the results found in Chapters 3 and 4.

6.2 Limitations and Future Work

While the idea of reparametrization in Chapters 3 and 4 is simple, one limitation is that the objective function is non-convex. Hence, it may be of interest, in future work, to investigate approaches that use convex constraints to achieve strong heredity. See, e.g., Zhao et al. (2009); Bien et al. (2013) and Haris et al. (2016). In Chapter 5, we developed another strategy to achieve the strong heredity using modified adaptive weights, which is also simple and straightforward. Nevertheless, unlike other approaches mentioned above, it can only assure the strong heredity asymptotically, i.e., in finite samples, the resulting estimators may violate the strong heredity.

Post selection inference (Lee et al., 2016) should also be addressed, i.e., inferential methods that can compensate for the fact that the model was picked in a data-dependent way. Zhao et al. (2022) proposed a valid tool to study selective inference for effect modification using LASSO, which may be able to shed some light on how to combine the selection inferential tools with pdWOLS, and pdWOLS used in combination with OAL confounder selection.

In Chapter 5, to obtain a doubly robust estimator, a well-behaved initial estimator is required, which can be estimated using an unpenalized approach. In the case that $p > n$, we could also build on the idea in Huang et al. (2008), who used a marginal regression approach to obtain the initial estimator for the adaptive LASSO, i.e, the outcome is regressed separately on each variable. However, this technique is more challenging in our setting, as it violates the assumption that the blip model is correctly specified. This is a partial identification problem, and it has been studied in van der Laan et al. (2003), which may be able to shed some light on how to use marginal regression to obtain a valid initial estimator. It also may be of interest, in future work, to investigate the algorithm to directly solve the REE problem instead of using the approximation. As this alternative does not require an initial estimator, it might perform better in a large p, small n scenario.

The extension of the single stage ITRs estimation approach proposed in Chapter 5 to a multi-stage scenario also requires further investigation. The estimation procedure is conducted recursively using backward induction in a multi-stage setting, and the "outcome" at each stage is set to be a predicted or estimated optimal response. For discrete outcomes, the optimal outcome is usually modeled by multiplicative effects, e.g., given a K-stage DTR in which the outcome is a count, the optimal outcome at the $(k-1)$th stage is computed by $\widehat{y}_{k-1}^{opt} = y \times \prod_k^K \exp\{\gamma_k(x_k^{\psi}, \widehat{a}_k^{opt}; \boldsymbol{\psi}_k) - \gamma_k(x_k^{\psi}, a_k; \boldsymbol{\psi}_k)\}$. A challenge under this scenario is that the estimated optimal outcome at any stage for subjects with zero-valued outcome will remain zero, unless adjustments are made (Wallace et al., 2019), which may lead to a bias and a loss of efficiency.

6.3 Concluding Remarks

My PhD thesis aims to fill relevant gaps in the precision medicine literature: doubly robust variable selection has been seldom investigated in the area of DTRs for continuous outcomes, and it has not been studied for discrete outcomes at all. Many existing DTRs methods are complicated and hard to implement, thus it is difficult to extend these to a regularization framework. My proposed methods are straightforward to implement while still possessing the desired double robustness property. I hope my thesis can result in further investigation for selecting important tailoring variables for optimal treatment decision-making.

Appendices

APPENDIX A

Appendix to Manuscript 1

In this supplement, we provide the algorithm for estimation, the required regularity conditions and proofs of the main paper, followed by additional simulation results and details of the STAR*D analysis.

Algorithm 3 Blockwise Coordinate Descent with Strong Heredity

function $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{A}, \lambda, \alpha, \epsilon)$
  $\widetilde{\boldsymbol{X}}_j \leftarrow \boldsymbol{A} \circ \boldsymbol{X}_j$ for $j = 1, \ldots, p$
  Initialize: $\psi_0^{(0)} = \beta_j^{(0)} = \tau_j^{(0)} \leftarrow 0$ for $j = 1, \ldots, p$.
  Set iteration counter $t \leftarrow 0$
  $\boldsymbol{R}^* \leftarrow \boldsymbol{Y} - \psi_0^{(t)} A - \sum_j (\boldsymbol{X}_j + \tau_j^{(t)} \psi_0^{(t)} \widetilde{\boldsymbol{X}}_j) \beta_j^{(t)}$
  repeat
    • To update $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$
      $\widetilde{\boldsymbol{X}}_j \leftarrow \beta_j^{(t)} \psi_0^{(t)} \widetilde{\boldsymbol{X}}_j \qquad$ for $j = 1, \ldots, p$
      $\boldsymbol{R} \leftarrow \boldsymbol{R}^* + \sum_{j=1}^p \tau_j^{(t)} \widetilde{\boldsymbol{X}}_j$

$$\boldsymbol{\tau}^{(t)(new)} \leftarrow \arg\min_{\boldsymbol{\tau}} \frac{1}{2n} \| \sqrt{\boldsymbol{W}} (\boldsymbol{R} - \sum_j \tau_j \widetilde{\boldsymbol{X}}_j) \|_2^2 + \lambda \alpha |\tau_j|_1$$

      $\Delta = \sum_j (\tau_j^{(t)} - \tau_j^{(t)(new)}) \widetilde{\boldsymbol{X}}_j$
      $\boldsymbol{R}^* \leftarrow \boldsymbol{R}^* + \Delta$
    • To update $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$
      $\widetilde{\boldsymbol{X}}_j \leftarrow \boldsymbol{X}_j + \tau_j^{(t)} \psi_0^{(t)} \boldsymbol{X}_j$ for $j = 1, \ldots, p$
      $\boldsymbol{R} \leftarrow \boldsymbol{R}^* + \widetilde{\boldsymbol{X}}_j \beta_j^{(t)}$

$$\beta_j^{(t)(new)} \leftarrow \arg\min_{\beta_j} \frac{1}{2n} \| \sqrt{\boldsymbol{W}} (\boldsymbol{R} - \widetilde{\boldsymbol{X}}_j \beta_j) \|_2^2 + \lambda (1 - \alpha) |\beta_j|_1$$

      $\Delta = \widetilde{\boldsymbol{X}}_j (\beta_j^{(t)} - \beta_j^{(t)(new)})$
      $\boldsymbol{R}^* \leftarrow \boldsymbol{R}^* + \Delta$
    • To update ψ_0
      $\widetilde{\boldsymbol{X}}_A \leftarrow \boldsymbol{A} + \sum_j \tau_j^{(t)} \widetilde{\boldsymbol{X}}_j \beta_j^{(t)}$
      $\boldsymbol{R} \leftarrow \boldsymbol{R}^* + \psi_0^{(t)} \widetilde{\boldsymbol{X}}_A$

$$\psi_0^{(t)(new)} \leftarrow \frac{\widetilde{\boldsymbol{X}}_A^\top \boldsymbol{W} \boldsymbol{R}}{\widetilde{\boldsymbol{X}}_A^\top \boldsymbol{W} \widetilde{\boldsymbol{X}}_A}$$

      $\Delta = (\psi_0^{(t)} - \psi_0^{(t)(new)}) \widetilde{\boldsymbol{X}}_A$
      $\boldsymbol{R}^* \leftarrow \boldsymbol{R}^* + \Delta$
    $t \leftarrow t + 1$

  until convergence criterion is satisfied

A.1 Regularity Conditions

A(1): Define $\boldsymbol{V} = (Y, \boldsymbol{X}, A)$. The observations $\boldsymbol{V}_i, i = 1, \ldots, n$ are independent and identically distributed with probability density $g(\boldsymbol{V})$ with respect to a measure ν. Denote the quasi-log-likelihood (i.e., the negative dWOLS loss function) as: $-\sum_{i=1}^{n} log\ h(\boldsymbol{V}_i, \boldsymbol{\theta})$, where h is the posited density. In addition, h is identifiable and satisfies

$$E_g\left[\frac{\partial log\ h(\boldsymbol{V};\boldsymbol{\theta})}{\partial \theta_j}\right] = \int \frac{\partial log\ h(\boldsymbol{V};\boldsymbol{\theta})}{\partial \theta_j} g(\boldsymbol{V}) = 0.$$

A(2): The matrix $\boldsymbol{J}(\boldsymbol{\theta})$ is finite and positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}_*$, where the elements of the matrix $\boldsymbol{J}(\boldsymbol{\theta})$ are

$$\boldsymbol{J}_{jl}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 log\ h(\boldsymbol{V};\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l}\right]$$

and $\boldsymbol{\theta}_*$ is the minimizer of the Kullback–Leibler divergence between h and g.

The following matrix exists at $\boldsymbol{\theta} = \boldsymbol{\theta}_*$:

$$\boldsymbol{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\left(\frac{\partial log\ h(\boldsymbol{V};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial log\ h(\boldsymbol{V};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right].$$

A(3): There exists an integrable function G such that $\left|\frac{\partial^3 log\ h(\boldsymbol{V};\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l \partial \theta_m}\right| \leq G(\boldsymbol{V})$ for all j, l, m and $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_*$.

These assumptions guarantee the asymptotic normality of the quasi-maximum likelihood estimators (White, 1982).

A.2 Proof of Theorem 3.2.1

Theorem 3.2.1. *Correct Sparsity: Assume that* $\sqrt{n}a_n = O(1)$ *and* $\sqrt{n}b_n \to \infty$*, then there exists a local minimizer* $\widehat{\boldsymbol{\theta}}_n$ *of Equation (7) such that* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*\| = O_p(n^{-\frac{1}{2}} + a_n)$*. Moreover, we have* $P(\widehat{\boldsymbol{\theta}}_{B^c} = 0) \to 1$.

PROOF: Denote the objective function as

$$Q(\boldsymbol{\theta}) = \mathcal{L}^*(\boldsymbol{\theta}) + \sum_{j=1}^{p} \lambda_j^{\beta} |\beta_j| + \sum_{j=1}^{p} \lambda_j^{\tau} |\tau_j|.$$

Let $\alpha_n = a_n + n^{-1/2}$. It suffices to show that for any $\epsilon > 0$, there exists a constant C such that $\inf_{\|\boldsymbol{\delta}=C\|} Q(\boldsymbol{\theta}_* + \alpha_n \boldsymbol{\delta}) > Q(\boldsymbol{\theta}_*)$ with probability at least $1 - \epsilon$, where $\boldsymbol{\delta} = (\boldsymbol{u}, \boldsymbol{v})$. Letting $D_n(\boldsymbol{\delta}) = Q(\boldsymbol{\theta}_* + \alpha_n \boldsymbol{\delta}) - Q(\boldsymbol{\theta}_*)$, we have

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= \mathcal{L}^*(\boldsymbol{\theta}_* + \alpha_n \boldsymbol{\delta}) - \mathcal{L}^*(\boldsymbol{\theta}_*) + \sum_{j=1}^{p} \lambda_j^{\beta} (|\beta_{*j} + \alpha_n u_j| - |\beta_{*j}|) + \sum_{j=1}^{p} \lambda_j^{\tau} (|\tau_{*j} + \alpha_n v_j| - |\tau_{*j}|) \\
&\geq \mathcal{L}^*(\boldsymbol{\theta}_* + \alpha_n \boldsymbol{\delta}) - \mathcal{L}^*(\boldsymbol{\theta}_*) + \sum_{j \in B_1} \lambda_j^{\beta} (|\beta_{*j} + \alpha_n u_j| - |\beta_{*j}|) + \sum_{j \in B_2} \lambda_j^{\tau} (|\tau_{*j} + \alpha_n v_j| - |\tau_{*j}|) \\
&\geq \mathcal{L}^*(\boldsymbol{\theta}_* + \alpha_n \boldsymbol{\delta}) - \mathcal{L}^*(\boldsymbol{\theta}_*) - (\sum_{j \in B_1} \lambda_j^{\beta} \alpha_n |u_j| + \sum_{j \in B_2} \lambda_j^{\tau} \alpha_n |v_j|) \\
&\geq \mathcal{L}^*(\boldsymbol{\theta}_* + \alpha_n \boldsymbol{\delta}) - \mathcal{L}^*(\boldsymbol{\theta}_*) - \alpha_n^2 \left(card(B_1) + card(B_2) \right) C \\
&= \alpha_n \nabla \mathcal{L}^*(\boldsymbol{\theta}_*) \boldsymbol{\delta} - 1/2 \, \alpha_n^2 \boldsymbol{\delta}^T \nabla^2 \mathcal{L}^*(\boldsymbol{\theta}_*) \boldsymbol{\delta} (1 + o_p(1)) - C \alpha_n^2 \left(card(B_1) + card(B_2) \right),
\end{aligned}$$

where $card(\cdot)$ is the cardinality of a set. The first term $\alpha_n \nabla \mathcal{L}^*(\boldsymbol{\theta}_*) \boldsymbol{\delta}$ is of order $O_p(n\alpha_n^2)$ using the fact that $\nabla \mathcal{L}^*(\boldsymbol{\theta}_*) = O_p(\sqrt{n})$ and $O_p(\sqrt{n}\alpha_n) = O_p(1) = O_p(n\alpha_n^2)$. The second term $\alpha_n^2 \boldsymbol{\delta}^T \nabla^2 \mathcal{L}^*(\boldsymbol{\theta}_*) \boldsymbol{\delta} (1 + o_p(1)) = n\alpha_n^2 \boldsymbol{\delta}^T \boldsymbol{J}(\boldsymbol{\theta}_*) \boldsymbol{\delta} (1 + o_p(1))$. Since $\boldsymbol{J}(\boldsymbol{\theta}_*)$ is positive definite, clearly the second term can be made to dominate the other two terms by choosing a sufficiently large constant C. Hence, the desired inequality holds:

$$P\left(\inf_{\|\boldsymbol{\delta}=C\|} Q(\boldsymbol{\theta}_* + \alpha_n \boldsymbol{\delta}) > Q(\boldsymbol{\theta}_*) \right) \geq 1 - \epsilon,$$

and there exists a local minimizer $\widehat{\boldsymbol{\theta}}_n$ such that $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*\| = O_p(n^{-\frac{1}{2}} + a_n)$.

Now we prove the correct sparsity property, and we first show that $P(\widehat{\boldsymbol{\beta}}_{B_1^c} = 0) \to 1$. It suffices to show that for any $j \in B_1^c$, $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*\| = O_p(n^{-1/2})$ and $\epsilon_n = Cn^{-1/2}$,

$$\frac{\partial Q(\widehat{\boldsymbol{\theta}}_n)}{\partial \beta_j} > 0 \text{ for } 0 < \widehat{\beta}_j < \epsilon_n \text{ and } \frac{\partial Q(\boldsymbol{\theta}_n)}{\partial \beta_j} > 0 \text{ for } 0 > \widehat{\beta}_j > -\epsilon_n.$$

When $0 < \widehat{\beta}_j < \epsilon_n$, by Taylor's expansion,

$$\begin{aligned}\frac{\partial Q(\widehat{\boldsymbol{\theta}}_n)}{\partial \beta_j} = \frac{\partial \mathcal{L}^*(\boldsymbol{\theta}_*)}{\partial \beta_j} + \sum_{k=1}^{2p+1} \frac{\partial^2 \mathcal{L}^*(\boldsymbol{\theta}_*)}{\partial \beta_j \partial \theta_l}(\widehat{\theta}_l - \theta_{*l}) \\ + \sum_{l=1}^{2p+1}\sum_{m=1}^{2p+1} \frac{\partial^3 \mathcal{L}^*(\tilde{\boldsymbol{\theta}})}{\partial \beta_j \partial \theta_l \partial \theta_m}(\widehat{\theta}_l - \theta_{*l})(\widehat{\theta}_m - \theta_{*m}) + \lambda_j^{\beta}\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_*$. By the regularity conditions, $\frac{\partial Q(\widehat{\boldsymbol{\theta}}_n)}{\partial \beta_j} = \sqrt{n}\left(O_p(1) + n^{-1/2}\lambda_j^{\beta}\right)$, which is greater than 0 by the assumption that $\sqrt{n}b_n \to \infty$. The case that $0 > \widehat{\beta}_j > -\epsilon_n$ can be shown in the same way.

Now we show that $P(\widehat{\boldsymbol{\theta}}_{B_2^c} = 0) \to 1$. In the case that $\beta_{*j} \neq 0$ and $\tau_{*j} = 0$, $\widehat{\tau}_j = 0$ can be proven in a similar manner. On the other hand, when $\widehat{\beta}_j = 0$, $\widehat{\tau}_j$ is also zero by construction. Thus, we have $P(\widehat{\boldsymbol{\theta}}_{B^c} = 0) \to 1$. This completes the proof.

A.3 Proof of Theorem 3.2.2

Theorem 3.2.2. *Asymptotic Normality: Assume that* $\sqrt{n}a_n \to 0$ *and* $\sqrt{n}b_n \to \infty$, *then*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_{*B}) \to_d N\left(0, \boldsymbol{J}^{-1}(\boldsymbol{\theta}_{*B})\boldsymbol{I}(\boldsymbol{\theta}_{*B})\boldsymbol{J}^{-1}(\boldsymbol{\theta}_{*B})\right)$$

where $\boldsymbol{J}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 log\ h(\mathbf{V};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]$ *and* $\boldsymbol{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\left(\frac{\partial log\ h(\mathbf{V};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial log\ h(\mathbf{V};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right]$.

PROOF: By Theorem 3.2.1, we have $\frac{\partial Q(\widehat{\boldsymbol{\theta}}_B)}{\partial\theta_j} = 0$ for any $j \in B$ with probability tending to 1. Denote the penalty term as $\rho_\lambda(\boldsymbol{\theta})$, we have $\frac{\partial Q(\widehat{\boldsymbol{\theta}}_B)}{\partial\theta_j} = \nabla\mathcal{L}^*(\widehat{\boldsymbol{\theta}}_B) + \nabla\rho_\lambda(\widehat{\boldsymbol{\theta}}_B) = 0$. By Taylor's expansion, the first term can be written as

$$\nabla\mathcal{L}^*(\widehat{\boldsymbol{\theta}}_B) = \sqrt{n}\{\frac{1}{\sqrt{n}}\nabla\mathcal{L}^*(\boldsymbol{\theta}_{*B}) + \sqrt{n}\boldsymbol{J}(\boldsymbol{\theta}_{*B})(\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_{*B}) + o_p(1)\} = 0,$$

while the second term $\nabla\rho_\lambda(\widehat{\boldsymbol{\theta}}) = o_p(\sqrt{n})$ using the assumption that $\sqrt{n}a_n \to 0$. By the weak law of large numbers, Slutsky's theorem and the central limit theorem, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}_{*B}) \to_d N\left(0, \boldsymbol{J}^{-1}(\boldsymbol{\theta}_{*B})\boldsymbol{I}(\boldsymbol{\theta}_{*B})\boldsymbol{J}^{-1}(\boldsymbol{\theta}_{*B})\right)$$

which is the desired asymptotic normality property. This completes the proof.

A.4 Simulation Studies

A.4.1 Double Robustness

Figures S1-S2 correspond to results from Section 3.2 in the main paper and demonstrate the double robustness of pdWOLS, comparing its performance to competitor approaches. Figure S1 shows the results under Scenario 2-4 with sample size 100 and 500, while Figure S2 is the full simulations result, which summarizes the estimates of blip parameters using the three methods under four scenarios of model specification. In Scenario 1, all the methods

failed since it fails to meet the assumptions of correct model specification required by them. Specifically, pdWOLS has a larger variance than the other two methods when the sample size is relatively small ($n = 100$). Across all the scenarios and methods, there is very little difference when the sample size was increased from $n = 500$ to $n = 2000$.

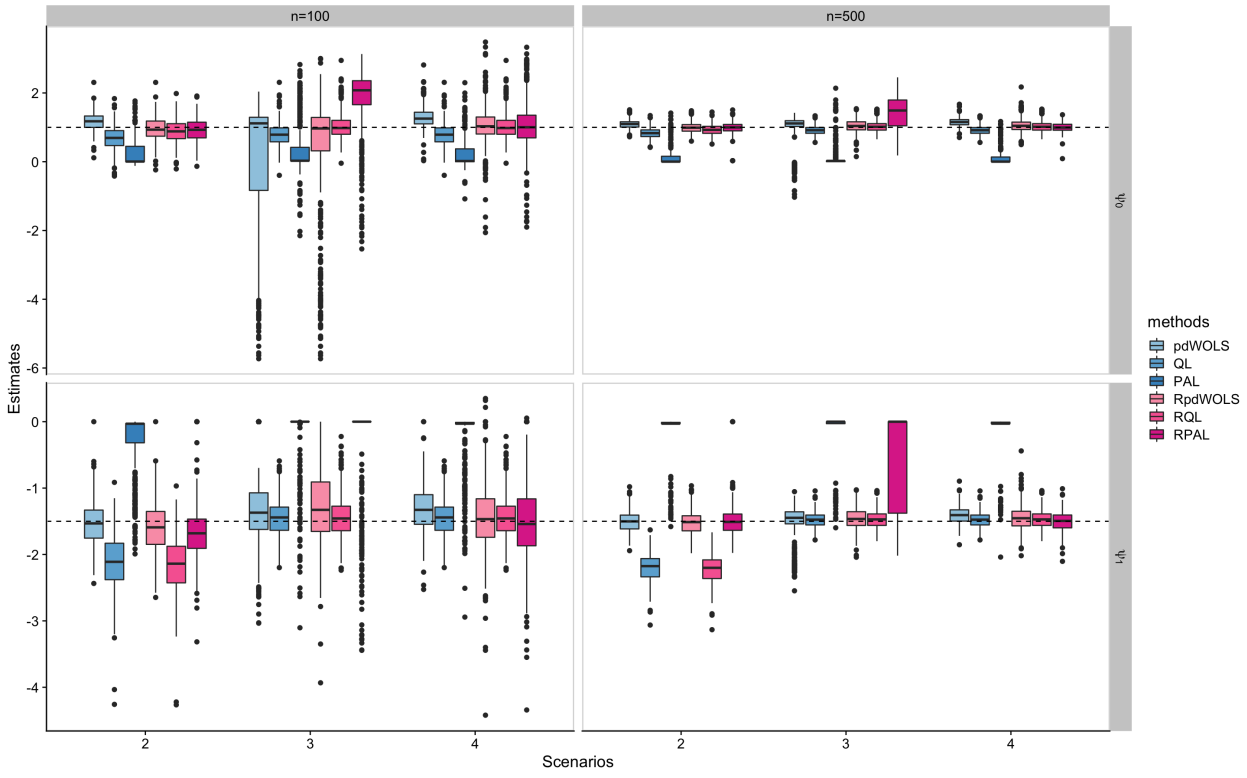

Figure S1: Estimates of blip parameters using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted counterparts (RpdWOLS, RQL and RPAL) when one (Scenarios 2 and 3) or both (Scenario 4) the treatment and treatment-free outcome models are correctly specified with sample size 100 and 500 (400 simulations). The true value is represented by the dotted line.

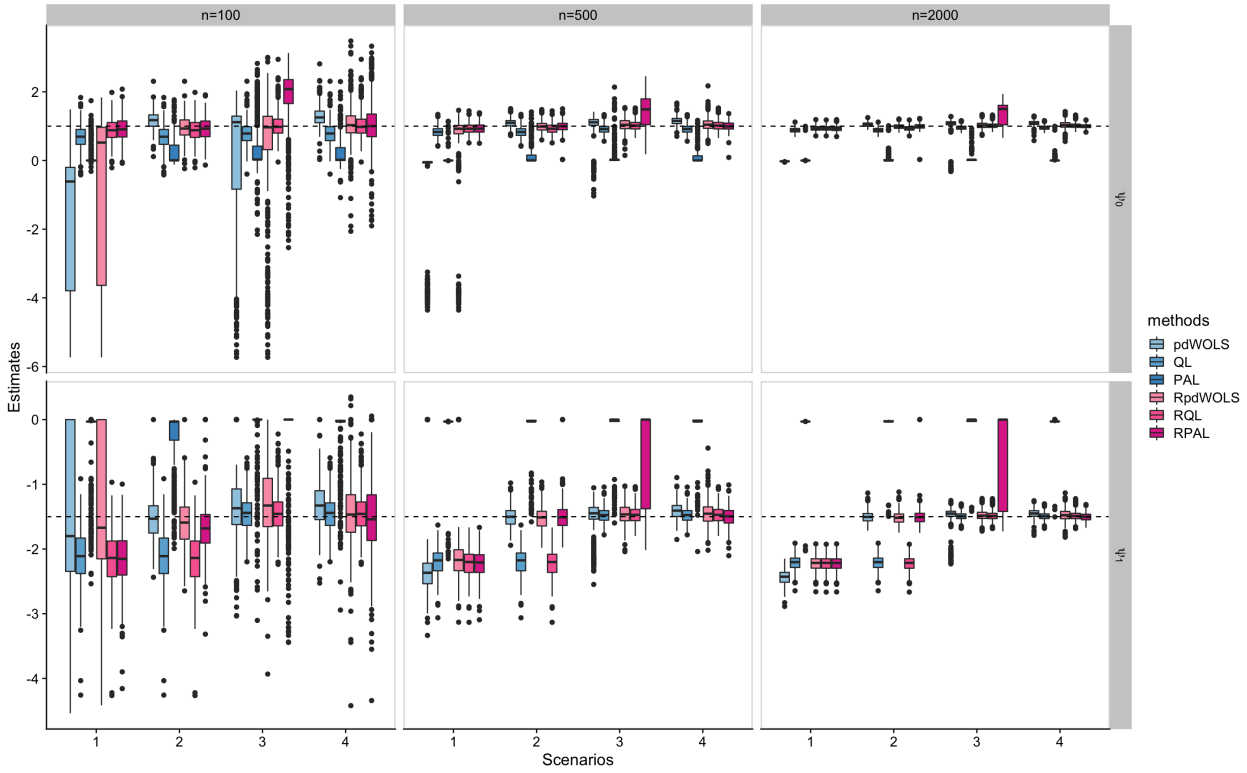

Figure S2: Estimates of blip parameters using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted counterparts (RpdWOLS, RQL and RPAL) when neither (Scenario 1), one (Scenarios 2 and 3) or both (Scenario 4) the treatment and treatment-free outcome models are correctly specified with sample size 100, 500 and 2000 (400 simulations). The true value is represented by the dotted line.

A.4.2 Multi-stage Setting 2

We follow the general approach to data generation procedure in (Simoneau et al., 2018):

Step 1: Generate 10 binary covariates at stage 1 with sample size 1000: $P(X_{j1} = 1) = P(X_{j1} = -1) = 0.5$ for $j = 1, 2, ...10$.

Step 2: Generate randomized treatment at stage k according to $P(A_k = 1) = P(A_k = 0) = 0.5$ for k=1, 2.

Step 3: Generate 10 covariates at stage 2 such that $P(X_{12} = 1|X_{11}, A_1) = expit(\theta_0 x_{11} + \theta_1(2A_1 - 1))$, and the remaining nine covariates such that $P(X_{j2} = 1) = P(X_{j2} = -1) = 0.5$

for $j = 2, 3, ...10$.

Step 4: Generate the outcome according to the model $Y = \beta_0 + \beta_{11}x_{11} + \gamma_0 A_1 + \gamma_1 A_1 x_{11} + \beta_{12}x_{12} + \psi_{02}A_2 + \psi_{12}A_2 x_{12} + \epsilon$ where $\epsilon \sim N(0, 1)$.

Under this setting, it can be shown that the true blip parameters (Simoneau et al., 2018; Chakraborty et al., 2010) in stage 1 are given by:

$$\psi_{01} = \gamma_0 + (q_1 - q_2)(2\beta_{12} + f_1 I(f_1 > 0) - f_2 I(f_2 > 0)),$$

$$\psi_{11} = \gamma_1 + (q_1' - q_2')(2\beta_{12} + f_1 I(f_1 > 0) - f_2 I(f_2 > 0)),$$

where

$$f_1 = \psi_{02} + \psi_{12}, \qquad f_2 = \psi_{02} - \psi_{12},$$

$$q_1 = 0.5(expit(\theta_0 + \theta_1) + expit(-\theta_0 + \theta_1)), \qquad q_2 = 0.5(expit(\theta_0 - \theta_1) + expit(-\theta_0 - \theta_1)),$$

$$q_1' = 0.5(expit(\theta_0 + \theta_1) - expit(-\theta_0 + \theta_1)), \qquad q_2' = 0.5(expit(\theta_0 - \theta_1) - expit(-\theta_0 - \theta_1)).$$

Now set $\theta_0 = \theta_1 = 1, \beta_0 = -1, \beta_{11} = 2, \gamma_0 = -1, \gamma_1 = 1.5,\ \psi_{02} = 1$, and $\psi_{12} = -1.5$. The true blip parameters at stage 2 are $\psi_{02} = 1$ and $\psi_{12} = -1.5$; at stage 1, they are $\psi_{01} = -1.195$ and $\psi_{11} = 1.5$.

At the estimation stage, we assume an outcome model (i.e., treatment-free plus blip model) for y that is linear in terms $(1, x_2, A_2, A_2 x_2)$, and at stage 1, a model for the pseudo-outcome that is linear in terms $(1, x_1, A_1, A_1 x_1)$. In this manner, the treatment-free outcome is correctly specified at the first stage but not the second.

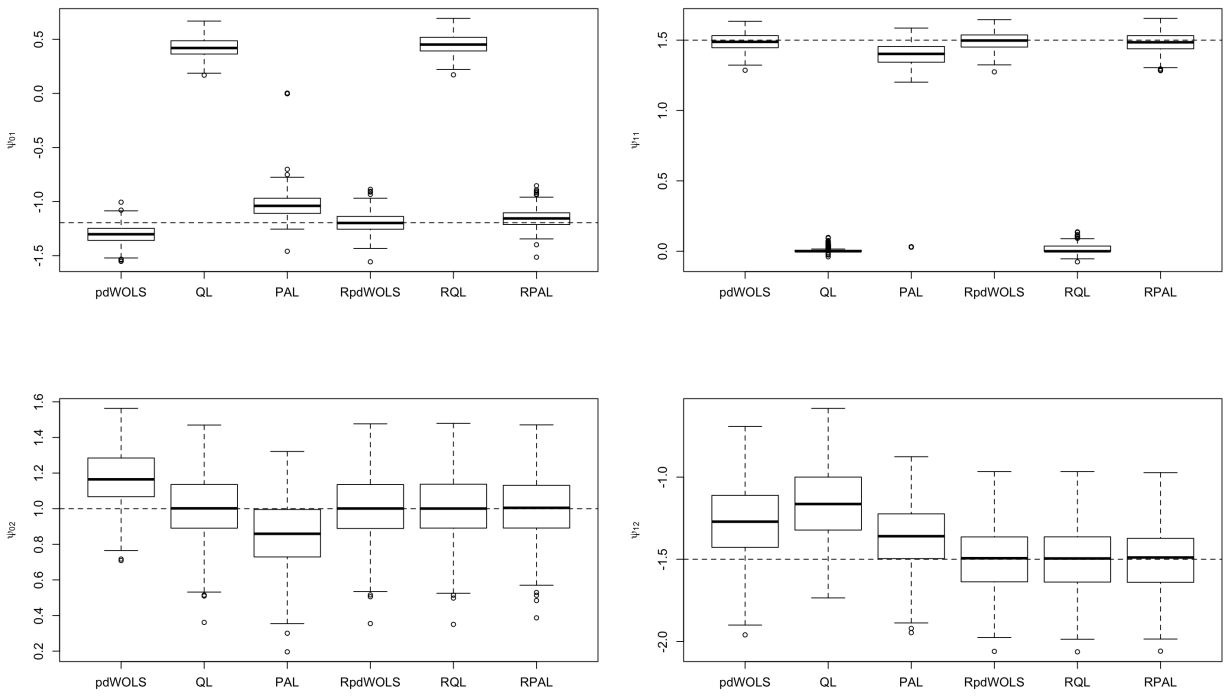

Figure S3: Estimates of blip parameters using pdWOLS, Q-learning (LASSO), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 2. The true value is represented by the dotted line.

Table S1: Variable selection rate (%) of the blip parameters using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 2. The main effect of treatment is not penalized (and hence is always selected).

	Stage 1						Stage 2		
	pdWOLS	QL	PAL	RpdWOLS	RQL	RPAL	pdWOLS	QL	PAL
AX_1 *	100	34	100	100	38	100	100	100	100
AX_2	1	12	27	1	12	28	1	38	46
AX_3	1	12	26	0	11	29	1	37	40
AX_4	1	14	24	1	12	25	2	39	49
AX_5	2	10	25	0	10	25	2	42	50
AX_6	1	16	25	1	14	26	0	33	44
AX_7	1	15	25	0	13	28	2	36	43
AX_8	1	14	30	1	12	32	1	37	45
AX_9	1	13	22	1	13	24	1	36	42
AX_{10}	1	14	24	1	15	28	1	40	42

* Term with a non-zero coefficient in the data-generating model

Table S2: Error Rate (ER, %) and value function using pdWOLS, Q-learning with LASSO (QL), PAL and their refitted versions with sample size 1000 (400 simulations) in two-stage Setting 2. Total error rate (TER, %) and stage-wise error rate in estimated optimal treatment recommendation across both stages are shown.

	TER	ER (Stage 1)	ER (Stage 2)	Value Function
pdWOLS	19.8	1.7	18.6	0.4
QL	62.6	50.8	17.5	-0.9
PAL	5.8	0.2	5.5	0.4
RpdWOLS	2.5	0.4	2.1	0.5
RQL	56.2	50.8	7.9	-0.8
RPAL	12.8	2.9	10.3	0.4

Figure S3 summarizes the estimates of blip parameters using the three methods in the two-stage Setting 2. As expected, pdWOLS and PAL work when at least one of the treatment or treatment-free models is correctly specified (in this case, the treatment model is correctly specified); Q-learning with LASSO failed at stage 1, since the treatment-free model at both stages are misspecified. For pdWOLS and PAL, refitted estimators are nearly unbiased, and they perform better than their penalized counterparts.

Table S1 presents the variable selection results for optimal treatment decisions. The important tailoring variable was selected by pdWOLS and PAL at both stages, while the performance of (refitted) Q-learning with LASSO at stage 1 was poor. At stage 2, the false positive rate of pdWOLS is much smaller than other two methods: for instance, the selection frequency of $AX_2 - AX_{10}$ are all less than 3%, as in stage 1. Note that at stage 1, because the pseudo-outcomes are different for refitted version and their penalized counterparts, the variable selection result varies too.

Table S2 summarizes the error rates of the estimated optimal treatment regimes for treatment decision making and value functions. The average value function and the error rates were computed over a testing set of size 10,000. Refitted methods for pdWOLS and Q-learning have a lower error rate and higher value than their penalized counterparts. However, unlike before where even the refitted PAL estimator has a smaller bias than the PAL estimator, here the refitted PAL did not improve the performance of PAL with respect to the value and error rate.

Sensitivity of Results to Choice of α

The data generation procedure is the same as in the high-dimensional setting, except that p is now 10.

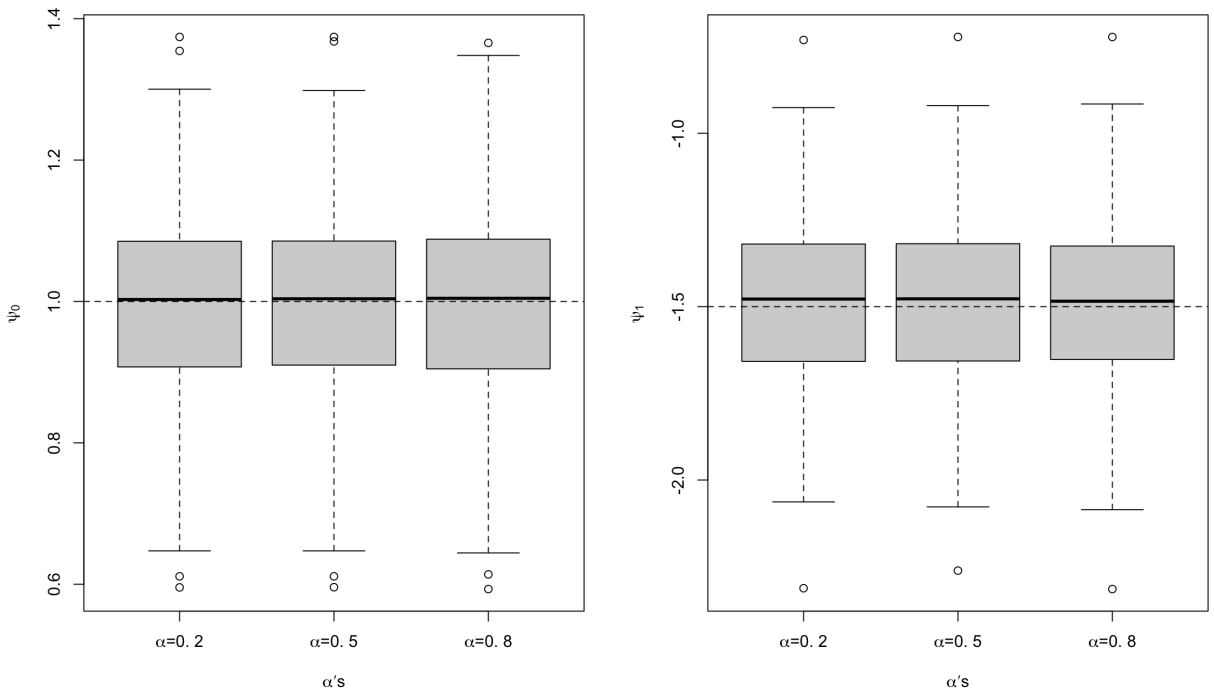

Figure S4: Estimates of blip parameters using pdWOLS with $\alpha = 0.2, 0.5, 0.8$, respectively (sample size 400, 400 simulations). The true value is represented by the dotted line.

Table S3: Variable Selection Rate (%) for pdWOLS with $\alpha = 0.2, 0.5, 0.8$ (sample size 400, 400 simulations). The main effect of treatment using pdWOLS is not penalized. The variables with $*$ are the truly important variables, and others are noise variables ($AX_2 - AX_{10}$).

	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
AX_1 *	100	100	100
AX_2	87	58	20
AX_3	3	3	1
AX_4	3	3	2
AX_5	2	2	2
AX_6	2	2	2
AX_7	2	3	1
AX_8	3	3	2
AX_9	3	2	2
AX_{10}	2	2	1

Table S4: Error rate (ER) and value for pdWOLS with $\alpha = 0.2, 0.5, 0.8$ (sample size 400, 400 simulations). For comparison, the value of the true optimal treatment decision, treat all, and treat none are 0.651, 0.419, and -0.569, respectively.

	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
ER	0.05	0.05	0.04
Value	0.64	0.64	0.64

A.5 Additional STAR*D Details

The STAR*D (Fava et al., 2003) study was divided into four levels. At the beginning of the study (level 1), all patients were prescribed citalopram and followed up regularly. Those who did not enter a state of remission of depression – defined as a Quick Inventory of Depressive Symptomatology (QIDS) score (Rush et al., 2003) less than or equal to 5, could then enter level 2 of the study where seven treatment options were accessible. In level 2, the treatment was shifted from citalopram to one of four different treatments or individuals continued to receive citalopram and received one of three additional treatments. Patients were allowed to enter the sub-level 2A where they received one of the *pharmacological* treatments available at level 2 if they received cognitive therapy at level 2 (either alone or combined with citalopram). Patients who did not experience a remission of depressive symptoms could then proceed to level 3, where their previous treatment was either switched to a new treatment or combined with another treatment. If the depression persisted after treatment in stage 3, they could enter level 4, where again their previous treatment was either switched to a new treatment or combined with another treatment. Following (Wallace et al., 2019) and (Chakraborty et al., 2013), we take level 2 (including 2A) in the trial as stage 1 in our analysis (the first treatment decision) and STAR*D level 3 as stage 2 in the analysis.

Our analytic sample consists of 1027 total participants who entered what we define as the first stage of our analysis (i.e., 273 participants entered level 2/2A of the original trial); of those, 273 moved on to the second stage of analysis. For this analysis, we focus on

treatment with a selective serotonin reuptake inhibitor or a different treatment options, coding treatment without a selective serotonin reuptake inhibitor as A=0 (and 1 otherwise). We define outcome as negative QIDS score at end of treatment. Because participants made a choice to remain in the study after each stage, while all individuals have the outcome Y_1, only those who chose to continue on into the study and enter stage 2 have Y_2 observed. For further details, see (Fava et al., 2003; Chakraborty et al., 2013).

APPENDIX B

Appendix to Manuscript 3

Regularity Conditions for M Estimator

In order to prove the consistency and asymptotic normality of regularized m-estimators, we need the following regularity conditions (Tsiatis, 2006) holds.

A(1): $\mathbb{E}\big(\frac{\partial U(V,\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}\big)$ is non-singular.

A(2): $n^{-1}\sum_{i=1}^{n}\frac{\partial U(V_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \rightarrow \mathbb{E}\big(\frac{\partial U(V,\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}\big)$ uniformly in a neighborhood of $\boldsymbol{\theta}^*$

A(3): The second moment of $U(V,\boldsymbol{\theta}^*)$ exists.

B.1 Fixed Point Problems and Variational Inequality

Lemma B.1.1 (***Fixed Point Problems and Variational Inequality (Dafermos, 1980)***)**.** *Suppose C is a closed convex set, then the variational inequality problem of finding a $x_0 \in C$ such that $F(x_0)^T(x_0 - x) \leq 0$ has a solution if and only if x_0 is a fixed point of the function $\Pi(x - F(x)|C)$, where $\Pi(\cdot\,|C)$ is the projection operator.*

Lemma B.1.1 suggests that solving the fixed point problem $\Pi(x - F(x)|C) = x$ is equivalent to solving a variational inequality problem. This Lemma will be helpful to prove the existence

of the ITR REE solution.

B.2 Proof of Theorem 5.3.1

Theorem 5.3.1 *Assume that the SUTVA, ignorability, consistency, and positivity conditions described in previous section and* ***Assumption 1*** *hold, if the posited baseline model satisfies* $x^{\psi} \subseteq x^{\beta}$*, then* $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$*, where* $\boldsymbol{\psi}_0$ *is the underlying true blip parameter.*

PROOF: We first consider the case where the baseline model is misspecified, and the treatment model is correctly specified. For count outcomes, we can rewrite the A-learning estimating equation

$$\boldsymbol{U}_1(\boldsymbol{\psi}) = \frac{1}{n}\sum_{i=1}^{n} x_i^{\psi}(a_i - \widehat{\pi}) \exp\{-\gamma(x_i^{\psi}, a; \psi)\} \left(y_i - \exp(f(x_i^{\beta}; \widehat{\beta}) + \gamma(x_i^{\psi}, a_i; \psi))\right),$$

as

$$\boldsymbol{U}_1(\boldsymbol{\psi}) = -\mathrm{I}(a=0)x^{\psi}\widehat{\pi}\left(y - \exp(f(x^{\beta}; \widehat{\beta})\right) \tag{B.1}$$

$$+ \mathrm{I}(a=1)x^{\psi}(1-\widehat{\pi}) \exp\{-\gamma(x^{\psi}, 1; \psi)\} \left(y - \exp(f(x^{\beta}; \widehat{\beta}) + \gamma(x^{\psi}, 1; \psi))\right). \tag{B.2}$$

Recall that

$$\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi}) = \begin{pmatrix} ax^{\psi} \\ x^{\beta} \end{pmatrix} |a - \widehat{\pi}| \exp\{-\gamma(x^{\psi}, a; \psi)\} \left(y - \exp(f(x^{\beta}; \beta) + \gamma(x^{\psi}, a; \psi))\right).$$

$\boldsymbol{U}_3(\boldsymbol{\beta}, \boldsymbol{\psi})$ can be split into

$$\mathrm{I}(a=1)x^{\psi}(1-\widehat{\pi}) \exp\{-\gamma(x^{\psi}, 1; \psi)\} \left(y - \exp(f(x^{\beta}; \beta) + \gamma(x^{\psi}, 1; \psi))\right) \tag{B.3}$$

and

$$\begin{aligned}&\mathrm{I}(a=0)x^{\beta}\widehat{\pi}\left(y-\exp(f(x^{\beta};\beta))\right)\\&\quad+\mathrm{I}(a=1)x^{\beta}(1-\widehat{\pi})\exp\{-\gamma(x^{\psi},1;\psi)\}\left(y-\exp(f(x^{\beta};\beta)+\gamma(x^{\psi},1;\psi))\right)\end{aligned}\tag{B.4}$$

Now let $x^{\beta}=x^{\psi}$ [1] and without loss of generality let the plug-in estimator $\widehat{\beta}$ in the A-learning estimating equation equal the baseline estimator obtained from $\boldsymbol{U}_3(\boldsymbol{\beta},\boldsymbol{\psi})$. Subtracting (B.4) from (B.3), we have

$$-\mathrm{I}(a=0)x^{\psi}\widehat{\pi}\left(y-\exp(f(x^{\beta};\beta))\right).\tag{B.5}$$

It is clear that now (B.1)=(B.5) and (B.2)=(B.3), which implies that $\mathbb{E}[\boldsymbol{U}_1(\boldsymbol{\psi}^*)]=0$. Hence the solution of $\boldsymbol{U}_3(\boldsymbol{\beta},\boldsymbol{\psi})$ is a solution of the A-learning estimating function, but not vice versa. Since the A-learning estimator is doubly robust, the solution of $\boldsymbol{U}_3$ is also doubly robust.

Now consider the case where the baseline model is correctly specified, and the treatment model is misspecified. By iterated expectation, $\mathbb{E}[\boldsymbol{U}_3(\boldsymbol{\beta},\boldsymbol{\psi})]=\mathbb{E}[\mathbb{E}(\boldsymbol{U}_3|\,a,x)]$.

$$\begin{aligned}\mathbb{E}(\boldsymbol{U}_3|\,a,x)&=\begin{pmatrix}ax^{\psi}\\x^{\beta}\end{pmatrix}|a-\widehat{\pi}|\exp\{-\gamma(x^{\psi},a;\psi)\}\\&\quad\times\left\{\mathbb{E}(y|a,x)-\exp(f(x^{\beta};\beta)+\gamma(x^{\psi},a;\psi))\right\}\\&=\begin{pmatrix}ax^{\psi}\\x^{\beta}\end{pmatrix}|a-\widehat{\pi}|\exp\{-\gamma(x^{\psi},a;\psi)\}\\&\quad\times\left\{\exp(f(x^{\beta};\beta_0)+\gamma(x^{\psi},a;\psi_0))-\exp(f(x^{\beta};\beta)+\gamma(x^{\psi},a;\psi))\right\}.\end{aligned}$$

[1] To simplify the proof and save the space, here we use a special case of the strong heredity assumption: $x^{\beta}\subseteq x^{\psi}$. The rigorous proof using the strong heredity assumption does not involve any extra difficulty except more tedious sentences.

Clearly, $\boldsymbol{\theta}_0$ is a solution of $\mathbb{E}(\boldsymbol{U}_3) = 0$, by the uniqueness assumption, $\boldsymbol{\psi}_0 = \boldsymbol{\psi}^*$. $\square$

B.3 Proof of Theorem 5.4.1

Theorem 5.4.1 *Assume that conditions in Theorem 5.3.1 hold, penalty functions are constructed using the modified adaptive weights, the tuning parameter satisfies* $\sqrt{n}\lambda = o(1)$ *and* $n\lambda \to \infty$*, then there exists a* $\sqrt{n}$*-consistent solution* $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}})$ *of the ITR REE such that* $\widehat{\boldsymbol{\psi}}_S \neq 0$ *and* $\widehat{\boldsymbol{\psi}}_{S_c} = 0$.

PROOF: Denote S' as the set of indices of non-zero components for $\boldsymbol{\beta}^*$, by the strong heredity constraint, we want our selected variables for baseline model satisfy $\widehat{\boldsymbol{\beta}}_{\widetilde{S}} \neq 0$, where $\widetilde{S} = S \cup S'$. Now let S^* be the set such that $\boldsymbol{\theta}^*_{S^*} = (\boldsymbol{\beta}^*_{\widetilde{S}}, \boldsymbol{\psi}^*_S)$, it suffices to show that there exists $\boldsymbol{\theta} = \boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u}$ for $\|\boldsymbol{u}\| = r$ such that $\boldsymbol{U}(\boldsymbol{\theta}) - n\lambda q(|\boldsymbol{\theta}|) = \mathbf{0}$.

We first show that for $\boldsymbol{\theta} = \boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u}$, $\boldsymbol{\theta}_{S^*_c} = 0$ with probability tending to 1, where S^*_c is the complement of S^*. It suffices to show that

$$\boldsymbol{U}_{S^*_c}(\boldsymbol{\theta}) - n\lambda q(|\boldsymbol{\theta}_{S^*_c}|) = \mathbf{0}, \tag{B.6}$$

which is equivalent to show that $\|\boldsymbol{U}_{S^*_c}(\boldsymbol{\theta})\|_\infty \leq n\lambda \widehat{w}_{S^*_c}$. By Taylor's expansion, $\boldsymbol{U}_{S^*_c}(\boldsymbol{\theta}) = \boldsymbol{U}_{S^*_c}(\boldsymbol{\theta}^*) + \boldsymbol{U}'_{S^*_c}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta}_{S^*_c} - \boldsymbol{\theta}^*_{S^*_c})$, where $\widetilde{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$. Using standard arguments that $\|\boldsymbol{U}_{S^*_c}(\boldsymbol{\theta}^*)\| = O_p(\sqrt{n})$ and $\boldsymbol{U}'_{S^*_c}(\widetilde{\boldsymbol{\theta}}) \to_p -\boldsymbol{J}(\boldsymbol{\theta}^*)$, hence $\|\boldsymbol{U}_{S^*_c}(\boldsymbol{\theta})\| = O_p(\sqrt{n})$ for $\boldsymbol{\theta} = \boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{u}$. Since the penalty function satisfies $\inf_{\|\boldsymbol{\theta}\| \leq Mn^{-1/2}} \sqrt{n}\,\lambda\,\widehat{w}_{S^*_c} = O_p(n\lambda) \to \infty$. Expression (B.6) holds with probability tending to 1.

It now remains to show that $\boldsymbol{U}_{S^*}(\boldsymbol{\theta}_{S^*}, \mathbf{0}) - n\lambda q(|\boldsymbol{\theta}_{S^*}|) = \mathbf{0}$, which is equivalent to the fixed point problem: find $\boldsymbol{\theta}_{S^*}$ such that $\Pi(\boldsymbol{\theta}_{S^*} - F_{S^*}(\boldsymbol{\theta})|C) = \boldsymbol{\theta}_{S^*}$ with $F_{S^*}(\boldsymbol{\theta}) = \boldsymbol{U}_{S^*}(\boldsymbol{\theta}_{S^*}) - n\lambda q(|\boldsymbol{\theta}_{S^*}|)$ and $C = B_{\boldsymbol{\theta}^*_{S^*}}(n^{-1/2}r)$, where $B_{\boldsymbol{\theta}^*_{S^*}}(n^{-1/2}r)$ is the Euclidean ball around $\boldsymbol{\theta}^*_{S^*}$ with radius $n^{-1/2}r$. To see this, note that the projection operator $\Pi(\boldsymbol{x} - F(\boldsymbol{x})|C)$ can be written as $\arg\min_{\boldsymbol{x}_0 \in C}\|\boldsymbol{x}_0 - (\boldsymbol{x} - F(\boldsymbol{x}))\|_2^2$.

By Lemma B.1.1, to establish the existence of the ITR REE solution, it suffices to show that for sufficiently large n, there exists a constant r such that on the boundary of a ball around $\boldsymbol{\theta}^*_{S^*}$ with radius $n^{-1/2}r$, the variational inequality holds for function $\boldsymbol{U}_{S^*}(\boldsymbol{\theta}_{S^*}, \mathbf{0}) - n\lambda q(|\boldsymbol{\theta}_{S^*}|)$ with high probability, i.e., for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*}\| = n^{-1/2}r} (\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*})^T [\boldsymbol{U}_{S^*}(\boldsymbol{\theta}_{S^*}, \mathbf{0}) - n\lambda q(|\boldsymbol{\theta}_{S^*}|)] < 0 \right) > 1 - \varepsilon.$$

Then we have

$$\begin{aligned}
&(\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*})^T [\boldsymbol{U}_{S^*}(\boldsymbol{\theta}_{S^*}, \mathbf{0}) - n\lambda q(|\boldsymbol{\theta}_{S^*}|)] \\
=&(\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*})^T [\boldsymbol{U}_{S^*}(\boldsymbol{\theta}^*) + \boldsymbol{U}'_{S^*}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*}) - n\lambda q(|\boldsymbol{\theta}^*_{S^*}|)] \\
=&(\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*})^T [O_p(\sqrt{n})\boldsymbol{c} - n\boldsymbol{J}(\boldsymbol{\theta}^*_{S^*})(\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*}) - n\lambda q(|\boldsymbol{\theta}^*_{S^*}|)] \\
=&(\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*})^T [O_p(\sqrt{n})\boldsymbol{c} - n\boldsymbol{J}(\boldsymbol{\theta}^*_{S^*})(\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*}) - O_p(n\lambda)\boldsymbol{c}] \\
=&O_p(r) - O_p(r^2) - O_p(n^{1/2}\lambda) \\
=&- O_p(r^2) < 0 \ \text{ (by choosing a sufficiently large } r).
\end{aligned}$$

The first equation uses Taylor's expansion; in the second equation, $\boldsymbol{c}$ is any vector such that $\|\boldsymbol{c}\| = 1$, and we use the fact that $\|\boldsymbol{U}_{S^*}(\boldsymbol{\theta}^*)\|$ is of order $O_p(\sqrt{n})$, and the law of large numbers, i.e., $\boldsymbol{U}'_{S^*}(\widetilde{\boldsymbol{\theta}}) \to_p -n\boldsymbol{J}(\boldsymbol{\theta}^*_{S^*})$; the third equation follows because the penalty function is constructed using the modified adaptive weights, i.e., $n\lambda q(|\boldsymbol{\theta}_{S^*}|) = n\lambda \widehat{w}_{S^*} = O_p(n\lambda)$; the fourth equation holds since $\|\boldsymbol{\theta}_{S^*} - \boldsymbol{\theta}^*_{S^*}\| = n^{-1/2}r$. Thus, we have proved that there exists a $\sqrt{n}$-consistent solution $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}})$ of ITR REE solution such that $\widehat{\boldsymbol{\psi}}_S \neq 0$ and $\widehat{\boldsymbol{\psi}}_{S_c} = 0$. □

B.4 Proof of Theorem 5.4.2

Theorem 5.4.2 *For any $\sqrt{n}-$consistent solution $\widehat{\boldsymbol{\theta}}$ of ITR REE,*

$$\sqrt{n}\boldsymbol{J}(\boldsymbol{\psi}_S^*)\{\widehat{\boldsymbol{\psi}}_S - \boldsymbol{\psi}_S^* + \boldsymbol{J}(\boldsymbol{\psi}_S^*)^{-1}\lambda\, q(|\boldsymbol{\psi}_S^*|)\} \to_d N\big(0, \mathbf{I}(\boldsymbol{\psi}_S^*)\big),$$

where $\mathbf{I}(\boldsymbol{\theta}) \in \mathbb{R}^{2p\times 2p}$ is the variance of the estimating equation $\mathbf{U}(V_i, \boldsymbol{\theta})$, and $\boldsymbol{J}(\boldsymbol{\theta}) \in \mathbb{R}^{2p\times 2p}$ is the quantity $\mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{\partial \mathbf{U}(V_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$, $\mathbf{I}(\boldsymbol{\psi}_S^*)$ and $\boldsymbol{J}(\boldsymbol{\psi}_S^*)$ are the corresponding $s\times s$ sub-matrices of $\mathbf{I}$ and $\boldsymbol{J}$ evaluated at the truth.

Proof: By Theorem 5.4.1, we have $\boldsymbol{U}_{S^*}(\widehat{\boldsymbol{\theta}}_{S^*}, \mathbf{0}) - n\lambda q(|\widehat{\boldsymbol{\theta}}_{S^*}|) = \mathbf{0}$ with probability tending to 1. By Taylor's expansion, the first term can be written as

$$\boldsymbol{U}_{S^*}(\boldsymbol{\theta}^*) + \boldsymbol{U}'_{S^*}(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{S^*} - \boldsymbol{\theta}^*_{S^*}) = n\lambda q(|\boldsymbol{\theta}_{S^*}|),$$

where $\widetilde{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}$. Now we have

$$\sqrt{n}\boldsymbol{J}(\boldsymbol{\theta}^*_{S^*})\{\widehat{\boldsymbol{\theta}}_{S^*} - \boldsymbol{\theta}^*_{S^*} + \boldsymbol{J}(\boldsymbol{\theta}^*_{S^*})^{-1}\lambda q(|\boldsymbol{\theta}_{S^*}|)\} = \boldsymbol{U}_{S^*}(\boldsymbol{\theta}^*) \to_d N\big(0, \boldsymbol{I}(\boldsymbol{\theta}^*_S)\big)$$

by the weak law of large numbers and the central limit theorem, which is the desired asymptotic normality property. It follows that

$$\sqrt{n}\boldsymbol{J}(\boldsymbol{\psi}_S^*)\{\widehat{\boldsymbol{\psi}}_S - \boldsymbol{\psi}_S^* + \boldsymbol{J}(\boldsymbol{\psi}_S^*)^{-1}\lambda\, q(|\boldsymbol{\psi}_S^*|)\} \to_d N\big(0, \boldsymbol{I}(\boldsymbol{\psi}_S^*)\big). \quad \square$$

B.5 Main Results for Binary Outcomes

We posit a linear model for the baseline model, i.e., $f(x;\boldsymbol{\beta}) = x^T\boldsymbol{\beta}$. For binary outcomes, we present estimating function $\boldsymbol{U}_4(\boldsymbol{\beta},\boldsymbol{\psi})$,

$$\boldsymbol{U}_4(\boldsymbol{\beta},\boldsymbol{\psi}) = \sum_{i=1} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} |a_i - \widehat{\pi}_i^*| \left(y_i - \text{expit}(f(x_i^{\beta};\beta) + \gamma(x_i^{\psi}, a; \psi)) \right),$$

where

$$\widehat{\pi}^* = \left(1 + \frac{(1 - \text{expit}(g(x;\widehat{\xi}))\text{expit}(f(x;\widehat{\beta}^*))}{\text{expit}(g(x;\widehat{\xi})\text{expit}(f(x;\widehat{\beta}^*) + \gamma(x,1;\psi))} \right)^{-1},$$

and $g(x;\xi)$ is the nuisance treatment model of $\mathbb{E}(A|Y=0,X)$.

Assumption 2. *When at least one of two nuisance models π or f is correctly specified, there exists a unique population parameter $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)$ such that $\mathbb{E}[\mathbf{U}_4(\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)] = 0$.*

Theorem B.5.1. *Assume that the SUTVA, ignorability, consistency, and positivity conditions described in previous section and **Assumption 2** hold, if the posited baseline model satisfies $x^{\psi} \subseteq x^{\beta}$, then $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, where $\boldsymbol{\psi}_0$ is the underlying true blip parameter.*

Theorem B.5.1 states that $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$ under mild conditions, which implies that the blip estimators $\widehat{\boldsymbol{\psi}}$ obtained by solving $\boldsymbol{U}_4(\boldsymbol{\beta},\boldsymbol{\psi})$ is a doubly robust estimator.

B.6 Algorithm to Obtain the Initial Estimator using Ridge Penalty (Count Outcomes)

Algorithm A1

function $(x_i, a_i, y_i, \varepsilon)$
  Set iteration counter $t \leftarrow 0$
  Initialize $\widetilde{\psi}_0$
  $w_{i0} \leftarrow |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widetilde{\psi}_0)\}$ for $i = 1, \ldots, n$
  repeat
    • Update ψ_t such that
    $\sum_{i=1}^{n} \begin{pmatrix} a_i x_i^{\psi} \\ x_i^{\beta} \end{pmatrix} w_{it} \left(y_i - \exp(f(x_i^{\beta}; \beta) + \gamma(x_i^{\psi}, a_i; \psi_t)) \right) = 2n\lambda\theta_t$
    $\widetilde{\psi}_{t+1} \leftarrow \psi_t$
    $w_{i(t+1)} \leftarrow |a_i - \widehat{\pi}_i| \exp\{-\gamma(x_i^{\psi}, a_i; \widetilde{\psi}_{t+1})\}$
    $t \leftarrow t + 1$
  until $\|\psi_t - \psi_{t-1}\| < \varepsilon$

B.7 Data Generation Procedure in the Large Dimension Setting

The data generation procedure for count outcomes is described as follows: Step 1: Generate p independent multivariate normal covariates $(X_1 - X_p)$ with mean -0.2 and unit variance. Step 2: Generate treatment such that $P(A_i = 1|x_1, x_2, x_3) = \text{expit}(\sum_{j=1}^{3} x_j)$. Step 3: Set the blip function as $\gamma(x, a; \boldsymbol{\psi}) = a(\psi_0 + \psi_1 x_1 + \psi_2 x_2 + \psi_3 x_3)$ for $\psi_0 = -0.8$, $\psi_1 = -2$, $\psi_2 = 2$, and $\psi_3 = 1$. Step 4: Set the baseline model to $f(\boldsymbol{x}; \boldsymbol{\beta}) = -0.3 - 0.8\,|x_1 + x_2 + x_3| + 0.8x_1 - 0.8x_2 + 0.8x_3$. Step 5: Generate the outcome $Y \sim \text{Poisson}(f(\boldsymbol{x}; \boldsymbol{\beta}) + \gamma(\boldsymbol{x}, a; \boldsymbol{\psi}))$. Under this data generation procedure, about 38% of the subjects would receive an optimal treatment with $A = 1$; and the marginal mean of the outcome Y is 1.33.

The data generation procedure for binary outcomes is: Step 1: Generate p multivariate normal covariates $(X_1 - X_p)$ with mean 0.5 and unit variance. Step 2: Set the baseline model to $f(\boldsymbol{x}; \boldsymbol{\beta}) = -\exp(x_1) + 4x_1 - 4x_2$. Step 3: Set the blip function as $\gamma(x, a; \boldsymbol{\psi}) =$

$a(\psi_0 + \psi_1 x_1 + \psi_2 x_2)$ for $\psi_0 = 1.5$, $\psi_1 = -4$, and $\psi_2 = 4$. Step 4: Set the nuisance treatment model as $\mathbb{E}(A|Y = 0, X = x) = -\exp(x_1) + 4x_1 - 4x_2$, and marginalize it over Y to obtain the propensity score model $\mathbb{E}(A|X = x)$; generate the treatment according to $\mathbb{E}(A|X = x)$. Step 5: Generate the outcome $Y \sim \text{Bernoulli}(f(\boldsymbol{x}; \boldsymbol{\beta}) + \gamma(\boldsymbol{x}, a; \boldsymbol{\psi}))$. Under this data generation procedure, about 62% of the subjects would receive an optimal treatment with $A = 1$; and the marginal mean of the outcome Y is 0.24.

B.8 Application to STAR*D Study

In this section, we apply the proposed penalized doubly robust method to the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) data (Fava et al., 2003). STAR*D is a multistage randomized trial whose goal was to find effective treatments for major depressive disorder patients, whose Quick Inventory of Depressive Symptomatology (QIDS) score are greater than 5 (Rush et al., 2003). STAR*D data have been analyzed in Wallace et al. (2019); Chakraborty et al. (2013); Bian et al. (2021), where each of the authors listed conducted a two-stage analysis according to the use of a selective serotonin reuptake inhibitor (SSRI) compared to a different treatment options, with QIDS score as the primary outcome (count) and three tailoring variables: the QIDS score measured at the beginning of each stage, change in QIDS score divided by the time in the previous level (QIDS slope), and patient preference measured prior to receiving treatment. For more information, see Fava et al. (2003); Chakraborty et al. (2013).

In this analysis, we follow the strategy in Wallace et al. (2019); Chakraborty et al. (2013); Bian et al. (2021) but only focus on the first stage. We also follow the strategy in Bian et al. (2021) to generate d noise variables for $d = 5, 10, 20$ respectively to conduct variable selection. The sample size in first stage is 1159, and we denote treatment with a SSRI with $A = 1$, and $A = 0$ for other treatment options. The propensity score was estimated using a logistic regression adjusting for patient preference only. We apply penalized doubly robust

method to this study with A-learning as the initial estimator, both the baseline model and the blip model are posited to be linear.

In all three scenarios ($d = 5, 10, 20$), our method found that only treatment preference is useful for tailoring treatment at stage 1. This result is consistent with the results found in Wallace et al. (2019); Chakraborty et al. (2013), while Bian et al. (2021) found that no blip covariates were useful for tailoring. The false positive rates of our method for the simulated noise variables are 0% ($d = 5$), 10% ($d = 10$), and 5% ($d = 20$), respectively.

Another analysis is performed with the outcome dichotomized by its median. The results are different from the above analysis where the outcome is count: in all three scenarios ($d = 5, 10, 20$), our binary penalized doubly robust method found that no blip variable is useful for tailoring treatment at stage 1, suggesting that the optimal treatments are treat with SSRI ($A = 1$) at stage 1 for all the patients. The false positive rates of our method for the simulated noise variables are 0 for all choices of d. This result is consistent with the result found in Bian et al. (2021).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Anderson, D. and Burnham, K. (2004). Model selection and multi-model inference. *Second Edition. NY: Springer-Verlag*, 63(2020):10.

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

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Bhatnagar, S. R., Lu, T., et al. (2020). A sparse additive model for high-dimensional interactions with an exposure variable. *BioRxiv*, page 445304.

Bian, Z., Moodie, E. E. M., Shortreed, S. M., and Bhatnagar, S. (2021). Variable selection in regression-based estimation of dynamic treatment regimes. *Biometrics*.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A Lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111.

Blatt, D., Murphy, S. A., and Zhu, J. (2004). A-learning for approximate planning. *University of Michigan, Ann Arbor*.

Cain, L. E., Robins, J. M., Lanoy, E., Logan, R., Costagliola, D., and Hernán, M. A. (2010). When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics*, 6(2).

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723.

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

Chakraborty, B., Murphy, S. A., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

Chen, H. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2):413–421.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36.

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

Cox, D. R. (1984). Interaction. *International Statistical Review/Revue Internationale de Statistique*, pages 1–24.

Dafermos, S. (1980). Traffic equilibrium and variational inequalities. *Transportation Science*, 14(1):42–54.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15.

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Fan, A., Lu, W., and Song, R. (2016). Sequential advantage selection for optimal treatment regime. *The Annals of Applied Statistics*, 10(1):32.

Fan, C., Lu, W., Song, R., and Zhou, Y. (2017). Concordance-assisted learning for estimating optimal individualized treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1565–1582.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, pages 531–552.

Fava, M., Rush, A. J., et al. (2003). Background and rationale for the sequenced treatment alternatives to relieve depression (STAR* D) study. *Psychiatric Clinics of North America*, 26(6):457–494.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Fu, W. and Knight, K. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, 59(1):126–132.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Greenland, S. (2008). Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, 167(5):523–529.

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

Haris, A., Witten, D., and Simon, N. (2016). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004.

Hastie, T., Tibshirani, R., and Friedman, J. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, New York.

Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, Boca Raton, FL.

Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91(4):929–941.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

James, G. M. and Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, 96(2):323–337.

Jeng, X. J., Lu, W., and Peng, H. (2018). High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 12(1):2074–2089.

Johnson, B. A., Lin, D., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103(482):672–680.

Kosorok, M. R. and Moodie, E. E. M. (2015). *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, Philadelphia, PA.

Lambert, S. D., Grover, S., Laizner, A. M., McCusker, J., Belzile, E., Moodie, E. E. M., Kayser, J. W., Lowensteyn, I., Vallis, M., Walker, M., et al. (2021). Adaptive web-based stress management programs among adults with a cardiovascular disease: A pilot sequential multiple assignment randomized trial (SMART). *Patient Education and Counseling*.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Liang, S., Lu, W., Song, R., and Wang, L. (2017). Sparse concordance-assisted learning for optimal treatment decision. *J. Mach. Learn. Res.*, 18:202–1.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Linn, K. A., Laber, E. B., and Stefanski, L. A. (2017). Interactive Q-learning for quantiles. *Journal of the American Statistical Association*, 112(518):638–649.

Logan, B. R., Sparapani, R., McCulloch, R. E., and Laud, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Statistical Methods in Medical Research*, 28(4):1079–1093.

Lovibond, S. H. and Lovibond, P. F. (1996). *Manual for the depression anxiety stress scales.* Psychology Foundation of Australia.

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

Mallows, C. L. (2000). Some comments on C_p. *Technometrics*, 42(1):87–94.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

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

Moodie, E. E. M. and Richardson, T. S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*, 37(1):126–146.

Moodie, E. E. M., Richardson, T. S., and Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455.

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

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

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, pages 758–765.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: Main content. *The International Journal of Biostatistics*, 6(2).

Portnoy, S. (1984). Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. I. Consistency. *The Annals of Statistics*, pages 1298–1309.

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

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

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

Robins, J. M. (1997). Causal inference from complex longitudinal data. In Berkane, M., editor, *Latent Variable Modeling and Applications to Causality: Lecture Notes in Statistics*, pages 69–117. Springer.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In Lin, D. Y. and Heagerty, P., editors, *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer.

Robins, J. M. and Greenland, S. (1986). The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123(3):392–402.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495.

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

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

Rotnitzky, A., Li, L., and Li, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika*, 97(4):997–1001.

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

Rubin, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75(371):591–593.

Rush, A. J., Trivedi, M. H., et al. (2003). The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54(5):573–583.

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

Schneeweiss, S., Rassen, J. A., et al. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512.

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

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

Shi, C., Song, R., and Lu, W. (2016). Robust learning for optimal treatment decision with NP-dimensionality. *Electronic Journal of Statistics*, 10:2894–2921.

Shi, C., Song, R., and Lu, W. (2021). Concordance and value information criteria for optimal treatment decision. *The Annals of Statistics*, 49(1):49–75.

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

Shortreed, S. M. and Moodie, E. E. M. (2012). Estimating the optimal dynamic antipsychotic treatment regime: Evidence from the sequential multiple-assignment randomized clinical antipsychotic trials of intervention and effectiveness schizophrenia study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):577–599.

Simoneau, G., Moodie, E. E. M., Platt, R. W., and Chakraborty, B. (2018). Non-regular inference for dynamic weighted ordinary least squares: understanding the impact of solid food intake in infancy on childhood weight. *Biostatistics*, 19(2):233–246.

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

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

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

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019). *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC press, Boca Raton, FL.

van der Laan, M. J., Laan, M., and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, New York.

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1).

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wallace, M. P. and Moodie, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71(3):636–644.

Wallace, M. P., Moodie, E. E. M., and Stephens, D. A. (2019). Model selection for G-estimation of dynamic treatment regimes. *Biometrics*, 75(4):1205–1215.

Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*, 39(1):389–417.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. *King's College*.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 50(1):1–25.

Wu, J., Galanter, N., Shortreed, S. M., and Moodie, E. E. M. (2021a). Ranking tailoring variables for constructing individualized treatment rules: an application to schizophrenia. *Journal of the Royal Statistical Society Series C (in press)*.

Wu, Y., Wang, L., and Fu, H. (2021b). Model-assisted uniformly honest inference for optimal treatment regimes in high dimension. *Journal of the American Statistical Association*, in press.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Zetterqvist, J. and Sjölander, A. (2015). Doubly robust estimation with the R package drgee. *Epidemiologic Methods*, 4(1):69–86.

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 217–242.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zhao, Q., Small, D. S., and Ertefaie, A. (2022). Selective inference for effect modification via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 84(2):382–413.

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

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187.

Zhu, W., Zeng, D., and Song, R. (2019). Proper inference for value function in high-dimensional Q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, 114(527):1404–1417.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67(2):301–320.