Approaches for Developing Treatment Rules

Jeremy Roth

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Noah Simon, Chair

Holly Janes

Susanne May

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Approaches for Developing Treatment Rules

Jeremy Roth

Chair of the Supervisory Committee:
Dr. Noah Simon
Biostatistics

The availability of scientific knowledge and the strength of supporting statistical evidence for efficacy of available treatments varies considerably across clinical settings. Nonetheless, the goal of clinicians remains largely unchanged: to recommend patients the most beneficial course of treatment available. In this dissertation, we examine the problem of estimating treatment efficacy for heterogeneous individuals in a target population with varied levels of abstraction and with an eye toward varied study designs.

In Chapter 2, we restrict ourselves to RCT data and frame the problem of identifying individual characteristics that affect treatment response as a global hypothesis test for qualitative interaction using a convex optimization problem that is solved either with or without the constraint of qualitative interaction under limited modeling assumptions. We also present a permutation-based testing procedure that yields a p-value or false discovery rate.

In Chapter 3, we move away from the RCT setting and instead focus on observational study designs that introduce the significant complication of treatment not being assigned independently of patient characteristics. At the core of Chapter 3 is a principled framework and user-friendly `R` implementation in the `DevTreatRules` package that allow practitioners to develop a function (known as a *treatment rule*) to recommend treatment based on individual characteristics, while also obtaining a trustworthy estimate of the treatment rule's benefit in the target population. We also introduces a four-category classification of characteristics

collected in a given observational study based on whether each variable might influence treatment assignment and whether it is expected to be observed in independent clinical settings. Our framework and `R` implementation emphasize the distinct roles these variable types should play in a principled analysis to ensure that an estimated treatment rule is applicable in clinical settings and that the estimate of the rule benefit is reliable.

We begin Chapter 4 by exploring the popular *outcome-weighted learning* (OWL) method that takes a "direct" approach to estimating a treatment rule rather than the "indirect" approach taken by the split-regression procedure in Chapter 3. We present a simple Bayesian interpretation of OWL that offers a clear equivalence with split-regression when the outcome is binary and a more nuanced connection when the outcome is continuous. We show how OWL fits into the principled framework of Chapter 3 and we accordingly expand the `R` package `DevTreatRules` to accommodate OWL. We then conduct a simulation study that uses `DevTreatRules` to develop and compare the performance of treatment rules from OWL and split-regression under a range of scenarios. We also implement another promising direct approach to estimating treatment rules, referred to as *direct-interactions*, in `DevTreatRules` and include it in the simulation study. We share our proposed remedies for a few subtle but critical computational issues we encountered during our simulation study that have a substantial impact on the performance of OWL and direct-interactions in practice.

For Anu,

my sun and my stars

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

The availability of scientific knowledge and the strength of supporting statistical evidence for efficacy of available treatments varies considerably across clinical settings. Nonetheless, the goal of clinicians remains largely unchanged: to recommend patients the most beneficial course of treatment available. Perhaps the disease pathway is well-understood biologically and a particular treatment option has been shown through extensive randomized controlled trials (RCTs) to disrupt that pathway with minimal side effects; perhaps the disease pathway is hypothesized to vary across individuals and the perceived benefits of potential treatment are similarly heterogeneous and supported only by suggestive results from less-than-ideal observational studies that fall short of yielding conclusive scientific knowledge. Whatever the situation, what clinicians seek from statisticians is a reliable, data-driven recommendation of which treatment will best serve each individual.

In this dissertation, we examine the problem of estimating treatment efficacy for heterogeneous individuals in a target population. We explore the problem with varied levels of abstraction and with an eye toward varied study designs. In Chapter 2, we restrict ourselves to RCT data and frame the problem of identifying individual characteristics that affect treatment response as a global hypothesis test for qualitative interaction. We formulate a convex optimization problem whose solution estimates outcomes either with or without the constraint of qualitative interaction under limited *a priori* modeling assumptions. We provide a general statement of our optimization problem that accommodates a mix of continuous and categorical characteristics in low-dimensional or high-dimensional settings via an additive model, and we also offer flexibility in choosing the loss and penalty functions to cater to specific applications. Armed with the solution to the convex problem, we also present a

permutation-based testing procedure that formally conducts the hypothesis test and yields a p-value or false discovery rate. We provide a `Python` implementation that is restricted to the fused lasso penalty function and to use with a handful of characteristics due to its reliance on a general convex solver. We illustrate the implementation's potential utility by applying it to gene expression data and estimating clinical outcome as a function of either one or two potentially informative genes. We share the code needed to implement our method and reproduce our results on GitHub.

In Chapter 3, we move away from the RCT setting and instead focus on observational study designs that introduce the significant complication of treatment not being assigned independently of patient characteristics. At the core of Chapter 3 is a principled framework and user-friendly `R` implementation in our `DevTreatRules` package that allow practitioners to develop a function (known as a *treatment rule*) that recommends treatment based on individual characteristics, while also obtaining a trustworthy estimate of the treatment rule's benefit in the target population based on careful data-splitting.

Chapter 3 also introduces a four-category classification of characteristics collected in a given observational study based on whether each variable might influence treatment and whether it is expected to be observed in independent clinical settings. Our framework and `R` implementation emphasize the distinct roles these variable types should play in a principled analysis to ensure that an estimated treatment rule is applicable in future clinical settings (i.e. only based on patient characteristics that will actually be available) and that the estimate of the rule's benefit is reliable (i.e. confounding is handled appropriately when variables and their roles may differ across studies).

We motivate our principled framework in Chapter 3 by writing out foundational quantities from the causal inference literature based on our variable classification and noting the most natural form of the targeted treatment rule would require complicated estimation/averaging of conditional densities; however, this onerous procedure can be replaced with a minimization problem that leads to straightforward and interpretable estimation. We refer to the simplified estimation procedure as the *split-regression* approach to estimating a treatment rule.

We illustrate the principled split-regression framework and software implementation in Chapter 3 with an application to the Women's Health Initiative Observational Study (WHI-OS) dataset to develop and evaluate a treatment rule that assigns baseline hormone therapy to postmenopausal women if it is expected to reduce 10-year incidence of heart disease or 10-year incidence of breast cancer. Although the WHI-OS it not publicly available, we share the code used to conduct our data example based on the raw WHI-OS data files so that readers with access to the data can reproduce our results.

We begin Chapter 4 by exploring the popular outcome-weighted learning (OWL) method developed by Zhao et al. (2012) that takes the so-called "direct" approach to estimating a treatment rule rather than the "indirect" approach taken by split-regression in Chapter 3. We believe the distinction between OWL (which counter-intuitively predicts *treatment assignment* as a function of individual characteristics, weighted by outcome) and split-regression (which predicts outcome as a function of individual characteristics under each treatment option) is representative of the general gulf between direct and indirect methods in the literature: We feel direct methods tend to have a very convincing justification for use as treatment rules but often imply somewhat hazy estimation procedures, while indirect methods offer an intuitive estimation procedure but a hazier justification for their use in developing treatment rules.

We take several routes in Chapter 4 to try to bridge some of the gap separating OWL and split-regression. We present a simple Bayesian interpretation of OWL that offers a clear equivalence with split-regression when the outcome is binary and a more nuanced connection to split-regression with a continuous outcome. We show how OWL easily fits into the principled framework of Chapter 3 and we accordingly expand the `R` package `DevTreatRules` presented in Chapter 3 to accommodate OWL. We then conduct a simulation study that uses `DevTreatRules` to develop and compare the performance of treatment rules from OWL and split-regression under a range of scenarios. We also implement another promising direct approach to estimating treatment rules (Tian et al., 2014; Chen et al., 2017), referred to as *direct-interactions*, in `DevTreatRules` and include it in the simulation study. We share

our proposed remedies for a few subtle but critical computational issues we encountered during our simulation study that have a huge impact on the performance of OWL and direct-interactions in practice. Last, we present a data example using the WHI-OS dataset that compares the performances of treatment rules estimated by the direct and indirect approaches.

Throughout Chapter 4 we emphasize that `DevTreatRules` allows a practitioner to form a development/validation/evaluation partitioning of the data, develop treatment rules from all three alternative approaches on the development dataset, compare estimated performances on the validation set, and, based on that exploration, lock in the one or two most promising rules to carry forward to the evaluation set, where a reliable estimate of performance can be calculated. We took that approach in our data example and share the code needed to reproduce those results (along with the results of the simulation study). We hope that by sharing our entire implementation in an `R` package and by providing the code to reproduce our simulations and data examples, we will be able to support future work in both research and clinical settings alike.

Chapter 2

A FRAMEWORK FOR ESTIMATING AND TESTING QUALITATIVE INTERACTIONS WITH APPLICATIONS TO PREDICTIVE BIOMARKERS

This is a pre-copyedited, author-produced version of an article accepted for publication in *Biostatistics* following peer review. The version of record [Jeremy Roth & Noah Simon. A framework for estimating and testing qualitative interactions with applications to predictive biomarkers. *Biostatistics* (2018) 19 (3): 263-280] is available online at: `https://academic.oup.com/biostatistics/article/19/3/263/4093306` with DOI 10.1093.

Abstract

An effective treatment may only benefit a subset of patients enrolled in a clinical trial. We translate the search for patient characteristics that predict treatment benefit to a search for qualitative interactions, which occur when the estimated response-curve under treatment crosses the estimated response-curve under control. We propose a regression-based framework that tests for qualitative interactions without assuming linearity or requiring pre-specified risk strata; this flexibility is useful in settings where there is limited *a priori* scientific knowledge about the relationship between features and the response. Simulations suggest that our method controls Type I error while offering an improvement in power over a procedure based on linear regression or a procedure that pre-specifies evenly spaced risk strata. We apply our method to a publicly available dataset to search for a subset of HER2+ breast cancer patients who benefit from adjuvant chemotherapy. We implement our method in Python and share the code/data used to produce our results on GitHub (`https://github.com/jhroth/data-example`).

2.1 Introduction

If the results of a clinical trial indicate that a proposed treatment is not beneficial overall, the treatment may still benefit a smaller subset of patients. Existing methods to predict which patients benefit from treatment generally use one of two general strategies.

The first approach, sometimes called classification-based (in Zhang et al. (2015), for example), assumes the structure of the treatment rule (a function that maps a patient's characteristics to a treatment option) and then looks for the optimal rule over that class of possible treatment rules. For example, Zhang et al. (2015) assume the optimal treatment rule is a sequence of "if-then" statements with less than three variables involved in each condition, then searches for the rule of this form that maximizes an estimate of treatment rule value. Outcome-weighted learning (Zhao et al., 2012, 2015) and marginal structural mean models (Robins et al., 2008; Orellana et al., 2010) are other classification-based approaches to predicting treatment benefit.

The second approach, sometimes called regression-based, usually assumes a conditional mean model of the response variable given a treatment assignment and patient characteristics. For an observation with a particular set of characteristics, these methods generally predict the conditional mean response under different treatment assignments to infer the best treatment option. Examples include methods leveraging the Q-learning framework (Zhao et al., 2009, 2011; Moodie et al., 2014), other statistical learning tools (e.g. boosting in Kang et al. (2014)), or a combination of parametric, semiparametric, and nonparametric estimators (such as the two-stage approach of Cai et al. (2011)).

In the econometric literature, Chang et al. (2015) and Crump et al. (2008) develop nonparametric theory to test the null hypothesis that the direction of the treatment effect is the same across all values of pre-specified covariates. This null hypothesis is sometimes called testing for treatment effect heterogeneity.

In the regression setting, the search for predictive biomarkers can be thought of as a test for a clinically meaningful interaction between treatment and patient characteristics. A *qual-*

itative interaction occurs when treatment is neither superior nor inferior over all subsets of patients (Gail and Simon, 1985; Pan and Wolfe, 1997; Peto, 1982). Qualitative interactions stand in contrast to *non-qualitative* interactions, where the magnitude of treatment effect varies across patient subsets but the direction of the treatment effect is the same in every subset (i.e. either all patients benefit or all patients do not benefit). Qualitative interactions are clinically meaningful because they identify a treatment strategy with the ability to improve patient outcomes. Figure 2.1 shows a hypothetical example of a qualitative interaction that can inform a treatment decision (left panel) and a non-qualitative interaction that cannot inform a treatment decision (right panel). If the response variable Y is desirable (e.g. survival time), the left panel suggests that treatment is only beneficial for patients with values of candidate biomarker X above about 1.7.

We propose a regression-based framework that tests for qualitative interaction without assuming a linear mean model or requiring a candidate biomarker to be divided into risk strata beforehand. To do this, we translate the general settings of qualitative and non-qualitative interaction into convex optimization problems that data-adaptively estimate the control and treatment groups' underlying trends in treatment response over the range of a candidate biomarker. We form a simple likelihood-based test statistic and evaluate its significance using a permutation test.

We emphasize that our framework does not directly identify a subpopulation of patients who should receive treatment; rather, our method provides a global test of the null hypothesis that treatment is either uniformly beneficial or uniformly not beneficial across the ordered values of a candidate biomarker. In the convex problem that allows for qualitative interaction, the values of the candidate biomarker at which the estimated response-curves for treatment groups cross each other *suggest* a subset of patients who benefit from treatment. However, there is no formal test run evaluating average treatment effect in that subset.

Section 2.2 outlines our general framework for estimation and testing. Section 2.3 describes our testing procedure in more detail. Section 2.4 briefly discusses the Python implementation of our method. We present results from a data example in Section 2.5 and

simulation results in Section 2.6. Section 2.7 gives a brief conclusion.

2.2 Estimation and Testing Framework

We begin with the formal definition of a qualitative interaction in the multivariate setting.

Suppose we observe patients randomized to either new treatment ($T = 1$) or standard of care ($T = 0$). On each patient we observe p covariate values $\mathbf{x} \equiv (x_1, \ldots, x_p)$ and a numeric response y. Suppose further that $(y|x_1, \ldots, x_p, T) \sim L(\cdot, \theta(\mathbf{x}, T))$ where L is a known distribution parametrized by a single parameter, $\theta(\mathbf{x}, T)$, that is an unknown function of the features and treatment assignment. Further suppose that, for any y, $L(y, \theta)$ is log-concave in θ. In what follows we will rewrite $\theta(\mathbf{x}, T)$ as $\theta_T(\mathbf{x})$. In addition let G denote the joint distribution of $\mathbf{x}$ and let $\text{supp}(G) = \{\mathbf{x}|G(\mathbf{x}) \neq 0\}$ denote the support of G. We are interested in testing the null hypothesis:

$$
\begin{aligned}
&H_0(\text{no qualitative interaction}): \\
&\quad \text{Either } \forall \mathbf{x} \in \text{supp}(G),\ \theta_1(\mathbf{x}) \leq \theta_0(\mathbf{x}) \ \text{ OR } \forall \mathbf{x} \in \text{supp}(G),\ \theta_0(\mathbf{x}) \leq \theta_1(\mathbf{x}) \\
&\qquad \text{versus} \\
&H_A(\text{qualitative interaction}): \\
&\quad \exists \mathbf{x}_1, \mathbf{x}_0 \in \text{supp}(G) \text{ with } \theta_0(\mathbf{x}_0) > \theta_1(\mathbf{x}_0) \text{ and } \theta_1(\mathbf{x}_1) > \theta_0(\mathbf{x}_1).
\end{aligned}
$$

Often rather than a bi-directional null hypothesis, we will instead be interested in testing whether new treatment is more effective than control for any subset of patients. If higher values of y are more beneficial, then in this case H_0 is reduced to: $\forall \mathbf{x}\ \theta_1(\mathbf{x}) \leq \theta_0(\mathbf{x})$; and the alternative hypothesis is correspondingly changed.

We note that this tests a global null hypothesis (that one response curve lies everywhere above the other). Rejecting this null does not indicate *where* the curves cross. The testing procedure we will propose does give an estimate of the subset of patients for whom treatment is more effective than control (and vice versa). However, we do not run a formal statistical

test of average treatment effect in those subsets.

For testing this hypothesis suppose we have data on $n_0 + n_1 = n$ independently drawn individuals, where n_0 individuals are randomized to the control group and n_1 to the treatment group. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be a matrix of p features observed for the n individuals and $Y \in \mathbb{R}^n$ be our vector of responses. We will let $\mathbf{x}_{i,\cdot}$ denote the i-th row of $\boldsymbol{X}$, $\mathbf{x}_{\cdot,j}$ denote the j-th column of $\boldsymbol{X}$, and x_{ij} the (i,j) entry of $\boldsymbol{X}$. In what follows we will give a proposal for a general likelihood ratio based test statistic — the null distribution of this statistic will be calculated via permutation.

2.2.1 Forming the test statistic

To form the standardized likelihood ratio statistic we must estimate $\theta_T(\mathbf{x})$ for $T = 0, 1$ under both the null and alternative hypotheses. Let $\mathcal{L}(\cdot, \theta)$ denote the negative log-likelihood of $L(\cdot, \theta)$ and recall the definition of deviance is

$$\mathcal{D}\left(y, \hat{\theta}\right) = 2\left[\mathcal{L}\left(y, \hat{\theta}\right) - \min_\theta \mathcal{L}(y, \theta)\right],$$

This is just a standardized version of the negative-log-likelihood (forced to be positive). Our test statistic is

$$Z = \frac{\sum_{i=1}^n \mathcal{D}\left(y_i, \hat{\theta}_{T_i}^{null}(\mathbf{x}_i)\right) - \sum_{i=1}^n \mathcal{D}\left(y_i, \hat{\theta}_{T_i}^{alt}(\mathbf{x}_{i,\cdot})\right)}{\sum_{i=1}^n \mathcal{D}\left(y_i, \hat{\theta}_{T_i}^{alt}(\mathbf{x}_{i,\cdot})\right)}, \tag{2.1}$$

where for $T = 0, 1$, $\hat{\theta}_T^{null}$ and $\hat{\theta}_T^{alt}$ are estimates under the null and alternative hypotheses respectively. In the case of gaussian data with fixed variance, we have $Z = \frac{\sum_i \left(y_i - \hat{\theta}_{T_i}^{null}(\mathbf{x}_i)\right)^2 - \sum_i \left(y_i - \hat{\theta}_{T_i}^{alt}(\mathbf{x}_i)\right)^2}{\sum_i \left(y_i - \hat{\theta}_{T_i}^{alt}(\mathbf{x}_i)\right)^2}$. This is just a scaled F-statistic. For logistic loss, $\min_\theta \mathcal{L}(y, \theta) = 0$, so we have $Z = \frac{\sum_{i=1}^n \mathcal{L}\left(y_i, \hat{\theta}_{T_i}^{null}(\mathbf{x}_i)\right) - \sum_{i=1}^n \mathcal{L}\left(y_i, \hat{\theta}_{T_i}^{alt}(\mathbf{x}_{i,\cdot})\right)}{\sum_{i=1}^n \mathcal{L}\left(y_i, \hat{\theta}_{T_i}^{alt}(\mathbf{x}_{i,\cdot})\right)}$. In order to control Type I error in our permutation-based testing procedure, it is important for the denominator of Z to remain positive; the deviance based test statistic in (2.1) achieves this.

In estimating $\hat{\theta}_T^{null}$ and $\hat{\theta}_T^{alt}$, we often do not want to make parametric assumptions. In what follows we will discuss non-parametric estimation of $\hat{\theta}_T^{null}$ and $\hat{\theta}_T^{alt}$ under general shape constraints.

2.2.2 A Convex Formulation for Estimating $\hat{\theta}^{null}$ and $\hat{\theta}^{alt}$

Our approach leverages penalized likelihood methods (Hastie et al., 2008) to estimate our parameters. In the more traditional, single class setting where $y_i \sim L(\cdot, \theta(\mathbf{x}))$, the penalized likelihood approach solves the problem:

$$\hat{\theta} \leftarrow \text{argmin}_\theta \sum_i \mathcal{L}(y_i, \theta(\mathbf{x}_{i,\cdot})) + \lambda P(\theta), \tag{2.2}$$

where P is a structure-inducing penalty and $\lambda > 0$ is a tuning parameter that trades off between enforcing goodness of fit and structure. P is chosen based on the type of smoothness or structure one expects: Sobolev-type penalties, $P(\theta) = \int \left\|\theta^{(k)}\right\|^2$, can be used to induce general smoothness; total-variation/trend-filtering penalties, in one dimension, can be used to obtain piecewise polynomial fits. Other penalties, based on the *convex indicator* (defined as $I(z \in S) \equiv \infty * \delta\{z \notin S\}$ where $\delta(\cdot)$ is the usual $0, 1$ indicator function) can enforce other structure such as monotonicity $I(\theta \text{ non-decreasing})$, a parametric form $I(\theta \in \text{span}\{\psi_1, \ldots, \psi_K\})$ for prespecified functions $\psi_1, \ldots, \psi_K$, or additivity $I\left(\theta(\mathbf{x}) = \sum_{j=1}^p \phi_j(x_j) \text{ for some } \phi_1, \ldots \phi_p\right)$. In the additive case, we can combine this convex indicator with smoothness-type penalties on each individual component. In all of these cases because $\mathcal{L}$ is convex, the problem in (2.2) is convex (Boyd and Vandenberghe, 2004).

We now extend this to our two class framework.

2.2.3 Estimation of $\hat{\theta}^{alt}$

For the unrestricted model under H_A, we can use the penalized likelihood framework to jointly solve for $\hat{\theta}_0^{alt}$ and $\hat{\theta}_1^{alt}$. We optimize

$$\hat{\theta}_1^{alt}, \hat{\theta}_0^{alt} \leftarrow \underset{\theta_1,\theta_0}{\operatorname{argmin}} \sum_{T_i=1} \mathcal{L}\left(y_i, \theta_1(\mathbf{x}_{i,\cdot})\right) + \lambda_1 P\left(\theta_1\right) + \sum_{T_i=0} \mathcal{L}\left(y_i, \theta_0(\mathbf{x}_{i,\cdot})\right) + \lambda_0 P\left(\theta_0\right) \tag{2.3}$$

We note that this can be minimized separately in θ_1 and θ_0, requiring us merely to solve two problems of the form (2.2). The tuning parameters λ_0 and λ_1 determine the amount of regularization; we use split-sample/cross validation (as described in Section 2.3) to choose these tuning parameters.

2.2.4 Estimation of $\hat{\theta}^{null}$

Estimating the null-restricted parameters is more difficult; here we aim to constrain our estimates not to cross on the support of G. Because G is generally unknown, we use instead the empirical distribution of $\mathbf{x}$, imposing the constraint that either $\hat{\theta}_0\left(\mathbf{x}_{i,\cdot}\right) \leq \hat{\theta}_1\left(\mathbf{x}_{i,\cdot}\right)$ for all $i = 1, \ldots, n$ or the mirror $\hat{\theta}_0\left(\mathbf{x}_{i,\cdot}\right) \geq \hat{\theta}_1\left(\mathbf{x}_{i,\cdot}\right)$ for all $i = 1, \ldots, n$ is true. Thus we solve the following penalized, constrained, maximum likelihood problem:

$$\hat{\theta}_1^{null}, \hat{\theta}_0^{null} \leftarrow \underset{\theta_1,\theta_0}{\operatorname{argmin}} \sum_{T_i=1} \mathcal{L}\left(y_i, \theta_1(\mathbf{x}_{i,\cdot})\right) + \lambda_1 P\left(\theta_1\right) + \sum_{T_i=0} \mathcal{L}\left(y_i, \theta_0(\mathbf{x}_{i,\cdot})\right) + \lambda_0 P\left(\theta_0\right) \tag{2.4}$$

$$s.t. \quad \textbf{Either } \theta_0\left(\mathbf{x}_{i,\cdot}\right) \leq \theta_1\left(\mathbf{x}_{i,\cdot}\right) \text{ for all } i = 1, \ldots, n \tag{C1}$$

$$\textbf{Or } \theta_0\left(\mathbf{x}_{i,\cdot}\right) \geq \theta_1\left(\mathbf{x}_{i,\cdot}\right) \text{ for all } i = 1, \ldots, n \tag{C2}$$

This is not a convex problem because the constraint is not convex. However, (C1) or (C2) alone is a convex constraint. Because this `Either/Or` condition is actually optimizing over the union of the sets defined by (C1) and (C2) we can still find the global optimum quite simply. We merely solve (2.4) with *only* constraint (C1) and again with *only* constraint (C2) and then choose, from those 2 candidate solutions, the solution that has the lower objective

value. For each single constraint (2.4) is a convex problem and can be solved efficiently (Boyd and Vandenberghe, 2004).

It should be noted that the null hypothesis is only constrained to have no qualitative interaction *at the observed data points*, rather than to have no qualitative interaction over the entire design space. In cases where minimal smoothness is assumed (particularly with a single feature), fitted values at unobserved data points will be close to a simple average of the fitted values at nearby observed data points. Here, without qualitative interactions at the observed data points, we would not expect qualitative interactions to occur for unobserved points within the design space. When more structure is assumed, such as in the additive model framework discussed in Section 2.2.6, fits under the null may actually allow qualitative interactions for unobserved combinations of covariates in the design space. That is, the null model permits more flexibility than it should under the assumption of no qualitative anywhere in the design space, and consequently our testing procedure will be conservative.

If the covariate values from our sample are representative of the population, then any qualitative interaction permitted under the null (at unobserved covariate combinations) would necessarily apply to a small subpopulation: Otherwise that subpopulation would have been represented in our sample. If observed features are not representative of the larger population, however, then the conservative nature of the method could be impactful — it may be that a covariate combination which is unobserved, and at which we allow a qualitative interaction under the null is actually a significant segment of the population. Our test would be unable to detect that sub-population. An alternative way to form constraints under the null would be to discretize the entire design space, then apply the (C1) and (C2) constraints to each element of the grid; this is an interesting approach, but one we choose not to pursue primarily for computational reasons.

2.2.5 Structure Induced on Fitted Values

For some choices of penalty, (2.2) does not immediately estimate an entire function θ; instead the function is only estimated at a finite list of specified $\mathbf{x}$-values (often taken to be the

observed values $\mathbf{x}_1, \ldots \mathbf{x}_n$). This is the case for fused lasso (Tibshirani et al., 2005) and higher order ℓ_1-trend filtering (Tibshirani, 2013), among others. There, the penalty generally only involves those specified $\mathbf{x}$-values; for example, 0th order trend filtering with a single covariate x, ordered such that $x_1 < x_2 < \cdots < x_n$, yields $P(\theta) = \sum_{i=2}^{n} |\theta(x_i) - \theta(x_{i-1})|$. However to estimate our fits under the null hypothesis we must solve (2.4) — this problem has constraints that require estimation at all x-values observed under either treatment for each of θ_1 and θ_0. In this case we adjust the usual fused lasso estimator to include all x-values in the penalty. More specifically, if we reorder our data such that $x_1 < x_2 < \ldots < x_n$ then our optimization problem is:

$$\begin{aligned}
\hat{\theta}_1^{null}, \hat{\theta}_0^{null} \leftarrow \underset{\theta_1,\theta_0}{\operatorname{argmin}} & \sum_{T_i=1} \mathcal{L}\left(y_i, \theta_1(\mathbf{x}_{i,\cdot})\right) + \lambda_1 \sum_{i=2}^{n} |\theta_1(x_i) - \theta_1(x_{i-1})| + \\
& \sum_{T_i=0} \mathcal{L}\left(y_i, \theta_0(\mathbf{x}_{i,\cdot})\right) + \lambda_0 \sum_{i=2}^{n} |\theta_0(x_i) - \theta_0(x_{i-1})| \\
s.t. \quad & \textbf{Either } \theta_0(\mathbf{x}_{i,\cdot}) \leq \theta_1(\mathbf{x}_{i,\cdot}) \text{ for all } i = 1, \ldots, n \qquad \text{(C1)} \\
& \textbf{Or } \theta_0(\mathbf{x}_{i,\cdot}) \geq \theta_1(\mathbf{x}_{i,\cdot}) \text{ for all } i = 1, \ldots, n \quad . \qquad \text{(C2)}
\end{aligned}$$

Note that in fitting the alternative (2.3), for $T = 0$ we need only include those x_i with $T_i = 0$ in the penalty (and similarly for $T = 1$).

2.2.6 Estimation with Multiple Features

When $\mathbf{x}_{i,\cdot} \in \mathbb{R}^p$ for $p > 1$, we can implement our framework using an additive model (Hastie and Tibshirani, 1990; Hastie et al., 2008). Suppose $\mathbf{x}_{i,\cdot}$ is partitioned into continuous and categorical variables as $\mathbf{x}_{i,\cdot} = \{\mathbf{x}_{i,\cdot}^{\text{cont.}}, \mathbf{x}_{i,\cdot}^{\text{cat.}}\}$ with $\mathbf{x}_{i,\cdot}^{\text{cont.}} \in \mathbb{R}^q$ and $\mathbf{x}_{i,\cdot}^{\text{cat.}} \in \mathbb{R}^s$. Also write $\lambda_j = \{\lambda_j^{\text{cont.}}, \lambda_j^{\text{cat.}}\}$ for $j = 0, 1$. Define

$$\Theta_1 = \left[\theta_{1,1}^{\text{cont.}} \quad \ldots \quad \theta_{1,q}^{\text{cont.}} \quad \theta_{1,1}^{\text{cat.}} \quad \ldots \quad \theta_{1,s}^{\text{cat.}}\right],$$

and similarly for Θ_0. Under the alternative hypothesis, we would then solve

$$\left(\widehat{\Theta}_1, \hat{\gamma}_1\right) \leftarrow \underset{\Theta_1,\gamma_1}{\operatorname{argmin}} \sum_{T_i=1} \mathcal{L}\left(y_i, \left[\gamma_1 + \sum_{j=1}^{q} \theta_{1,j}^{\text{cont.}}(x_{ij}^{\text{cont.}}) + \sum_{k=1}^{s} \theta_{1,k}^{\text{cat.}}(x_{ik}^{\text{cat.}})\right]\right) + \lambda_1^{\text{cont.}} \sum_{j=1}^{q} P_j^{\text{cont.}}\left(\theta_{1,j}^{\text{cont.}}\right) + \lambda_1^{\text{cat.}} \sum_{k=1}^{s} P_k^{\text{cat.}}\left(\theta_{1,k}^{\text{cat.}}\right),$$

$$\left(\widehat{\Theta}_0, \hat{\gamma}_0\right) \leftarrow \underset{\Theta_0,\gamma_0}{\operatorname{argmin}} \sum_{T_i=1} \mathcal{L}\left(y_i, \left[\gamma_0 + \sum_{j=1}^{q} \theta_{0,j}^{\text{cont.}}(x_{ij}^{\text{cont.}}) + \sum_{k=1}^{s} \theta_{0,k}^{\text{cat.}}(x_{ik}^{\text{cat.}})\right]\right) + \lambda_0^{\text{cont.}} \sum_{j=1}^{q} P_j^{\text{cont.}}\left(\theta_{0,j}^{\text{cont.}}\right) + \lambda_0^{\text{cat.}} \sum_{k=1}^{s} P_k^{\text{cat.}}\left(\theta_{0,k}^{\text{cat.}}\right),$$

where $\gamma_0, \gamma_1 \in \mathbb{R}$. To solve under the null hypothesis, we add the appropriate `Either/Or` conditions and (C1) and (C2) as additional constraints. The fitted value under the alternative for the ith individual would then be

$$\sum_{t=0}^{1} I(T_i = t)\left[\gamma_t + \sum_{j=1}^{q} \theta_{t,j}^{\text{cont.}}(x_{ij}^{\text{cont.}}) + \sum_{k=1}^{s} \theta_{t,k}^{\text{cat.}}(x_{ik}^{\text{cat.}})\right]. \tag{2.5}$$

There are many potential choices for the continuous and categorical penalties. One choice that seems to perform well, which we use in our `Python` implementation, is $P^{\text{cat.}}$ as the 1-dimensional fused lasso penalty and $P_k^{\text{cat.}}$ as a ridge penalty:

$$P_k^{\text{cat.}}\left(\theta_{t,j}^{\text{cat.}}\right) = \sum_{T_i=t} \left[\theta_{t,j}^{\text{cat.}}(x_{ij})\right]^2,$$

$$P_j^{\text{cont.}}\left(\theta_{t,j}^{\text{cont.}}\right) = \left\| D_{j,t}^{(0)}\left[\theta_{t,j}^{\text{cont.}}\left(x_{\cdot,j}^{\text{cont.}}\right)\right]\right\|_1,$$

where $D_{j,t}^{(0)}(\mathbf{x}_{\cdot,j})$ returns a vector of first-differences between the fitted values at consecutive ordered values of the feature $\mathbf{x}_{\cdot,j}$, among observations in the $T = t$ group.

2.2.7 Minimum Clinically Relevant Effect

Rather than testing a null hypothesis of no crossing, we may instead be interested in testing a null of the form *no clinically relevant crossing*, where clinical relevance will be application-specific. For ease of exposition we will work with only the one sided null/alternative wherein under the null we expect that outcomes under the new treatment ($T = 1$) will fall largely or entirely below under the control ($T = 0$). However, the opposite direction or the bidirectional null/alternative are also simple to test. In this case we can formalize our null hypothesis as:

$$H_0 : \int \left[\theta_1 (\mathbf{x}) - \theta_0 (\mathbf{x}) - \delta\right]_+ dG (\mathbf{x}) \leq C \tag{2.6}$$

for some prespecified δ and $C > 0$, where $[z]_+ = z * I(z \geq 0)$. For $\delta = 0$ this null hypothesis states that, if the population is treated under the optimal rule, the population average benefit from new treatment over control is at most C. For $C = 0$, and $\delta \neq 0$ this null hypothesis states that no patients in the population have expected benefit of greater than δ from new treatment over control. For $\delta = C = 0$ we return to the original null hypothesis.

To test this null we use the same formulation as above; now however we need estimates $\hat{\theta}_1^{null}$, $\hat{\theta}_0^{null}$ for a slightly different null. As before we use the empirical distribution of the covariates as a surrogate for G. We solve

$$\hat{\theta}_1^{null}, \hat{\theta}_0^{null} \leftarrow \underset{\theta_1, \theta_0}{\operatorname{argmin}} \sum_{T_i=1} \mathcal{L} \left(y_i, \theta_1(\mathbf{x}_{i,\cdot})\right) + \lambda_1 P \left(\theta_1\right) + \sum_{T_i=0} \mathcal{L} \left(y_i, \theta_0(\mathbf{x}_{i,\cdot})\right) + \lambda_0 P \left(\theta_0\right) \tag{2.7}$$

$$s.t. \quad \frac{1}{n} \sum_{i=1}^{n} \left[\theta_1 (\mathbf{x}_i) - \theta_0 (\mathbf{x}_{i,\cdot}) - \delta\right]_+ \leq C \tag{C1*}$$

This is a convex problem that can be solved efficiently. As before one could replace (C1*) with an equivalent (C2*), or conduct a bidirectional test on an `Either/Or` constraint.

2.2.8 Selecting a Smoothing Method

The choice of penalty function P determines the type of smoothness our solutions will have. We illustrate various candidates in Table 2.1.

Our `Python` implementation uses the 1-dimensional fused lasso penalty. We make this choice because we have found that the 1-dimensional fused lasso penalty performs well empirically, and solutions to (2.4) and (2.3) using this penalty can be found efficiently and accurately in our implementation. Figure 2.2 shows example solutions to (2.4) and (2.3) for the fused lasso with a continuous response where we have simulated noisy data from two crossing sinusoidal functions. We see that, in this simple example, the fused lasso fits effectively detect and localize the crossings.

2.3 Testing Procedure

We are interested in testing our null hypothesis of no qualitative interaction based on the ingredients (test statistic and estimated functions) outlined in Section 2.2. To develop a formal test we combine those ingredients with permutations to evaluate the null distribution of our statistic. The exact algorithm is the following:

1. Calculate $\hat{\theta}_T^{alt}$, for $T = 0, 1$ by solving (2.3) across sequences of λ_0 and λ_1 values. Using 10-fold cross-validation, choose the tuning-parameter values λ_0^* and λ_1^* that maximize the cross-validated log-likelihood.

2. Calculate $\hat{\theta}_T^{null}$ for $T = 0, 1$ by solving (2.4) with $\lambda_0 = \lambda_0^*, \lambda_1 = \lambda_1^*$.

3. Calculate the test-statistic, Z, by plugging in $\hat{\theta}_T^{alt}$ and $\hat{\theta}_T^{null}$ for $T = 0, 1$ to (2.1):

$$Z = \frac{\sum_{i=1}^{n} \mathcal{D}\left(y_i, \hat{\theta}_{T_i}^{null}(\mathbf{x}_i)\right) - \sum_{i=1}^{n} \mathcal{D}\left(y_i, \hat{\theta}_{T_i}^{alt}(\mathbf{x}_{i,\cdot})\right)}{\sum_{i=1}^{n} \mathcal{D}\left(y_i, \hat{\theta}_{T_i}^{alt}(\mathbf{x}_{i,\cdot})\right)},$$

4. Calculate the permutation null of Z by running the following for $b = 1, \ldots, B$

(a) Permute the treatment labels T_i for $i = 1, \ldots, n$

(b) With the new treatment labels rerun steps $1 - 3$. This will produce a permuted statistic Z^{*b}.

5. Compare Z to $\left\{Z^{*b}\right\}_{b=1}^{B}$ using either a p-value

$$p^* = \frac{\#\left\{Z^{*b} \geq Z\right\}}{B} \tag{2.8}$$

or an estimate of the false discovery rate (Benjamini and Hochberg, 1995; Tusher et al., 2001), which is further discussed in Appendix A.2.

We use the optimal tuning parameters under the alternative hypothesis to solve under the null hypothesis. This decision was primarily made for computational reasons. The extra constraints (C1) and (C2) required under the null make the problem more computationally costly to solve. In particular, the specialized algorithm given in Section 2.4 for solving under the alternative no longer works under the null, and a general-purpose convex solver is used instead. In our experience, penalty parameters selected via CV under the alternative, perform well under the null in practice. Nevertheless, there may be problems where this is not appropriate, and performing CV on the null is required for good performance.

2.3.1 Conservativeness of Permutation Tests

In the particular case that $\theta_0(x) = \theta_1(x)$ for all x under the null, then our permutation test evaluates an *exact* p-value based on the conditional distribution of our statistic.

However, our [one sided] null only requires that $\theta_0(x) \leq \theta_1(x)$ for all x. In this case, permutations appear to result in a conservative p-value. To see this, consider the situation wherein the null is true, and the entire θ_1 curve falls well above (or well below) the θ_0 curve. In this case the true null distribution of Z will have significant mass at exactly 0. However, by permuting we treat both θ_0 and θ_1 as the average of the two curves, making it much more

likely for our empirical estimates to cross, and for our permuted statistic to be positive (i.e. for the fit under the alternative to be a superior fit to the data).

We provide some theoretical evidence of conservativeness in Appendix A.3 for the simplified setting wherein we estimate our conditional means using piecewise constant functions with prespecified knots and assume our outcomes are gaussian distributed with equal variance in each treatment arm.

2.4 Computing

We implement our method in `Python` using a 1-dimensional fused lasso penalty. The GitHub repository `https://github.com/jhroth/data-example` contains a working example of applying our method to data from a randomized controlled trial investigating a breast cancer treatment (Prat and Bianchini, 2014). The data are publicly available on the Gene Expression Omnibus with accession number GSE50948 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50948`). We describe GSE50948 dataset in more detail in Section 2.5.

2.4.1 Solving Under H_0

We solve the optimization problem under H_0 in (2.4) for either a binary response (using a logistic loss function) or continuous response (using a squared-error loss function) with `CVXPY`, a general solver for convex optimization problems implemented in `Python` (Diamond et al., 2014). Our method calls the the splitting conic solver (SCS) (O'Donoghue et al., 2016a,b) when using the 1-dimensional fused lasso penalty.

2.4.2 Solving Under H_A

If the response variable is continuous, we solve under H_A using a `C` implementation of a very efficient fused lasso solver that uses a dynamic programming approach (Johnson, 2013) (when we have one feature) or using `CVXPY` with the SCS solver (in an additive model when

we have more than one feature). For a binary response variable, we use logistic loss and apply a generalized gradient descent algorithm (Parikh et al., 2014) that leans on the solver from (Johnson, 2013). We give detail on the derivation of this algorithm in Appendix A.1. For a binary response variable and more than one feature, we again fit an additive model and call the the splitting conic solver (SCS) (O'Donoghue et al., 2016a,b).

For multivariate problems, our use of a general purpose solver can result in long runtimes. In the worst case, for the examples in this manuscript, we found that a single model fit can take several minutes. This is not prohibitive, as fits for various features/permutations can be run in parallel. In particular, simulation and data examples in this manuscript were run on Amazon Web Services (`aws.amazon.com`). In addition, we plan to develop a custom first solver (leveraging sparsity) to further alleviate this issue.

2.5 Data Example: GSE50948

The code used to produce and plot the results for the data example is available in the GitHub repository `https://github.com/jhroth/data-example`. We use our methodology to evaluate expression markers from the Prediction Analysis of Microarray 50 (PAM50) diagnostic test (Parker et al., 2009) for use in characterizing those clinically-defined HER2+ breast cancer patients (defined by pathology/copy number) who are likely to benefit from trastuzumab.

Between 20% and 30% of breast cancer are characterized by over-expression of human epidermal growth factor receptor type 2 (HER2); this is known as *HER2+* breast cancer. Trastuzumab, an intravenous antibody therapy that inhibits production of HER2, has been shown to lengthen disease-free survival for women with HER2+ breast cancer (Joensuu et al., 2006) and is now part of the standard-of-care chemotherapy regimen for HER2+ breast cancer patients (Hudis, 2007).

PAM50 is a diagnostic test that uses the expression levels of 50 genes to characterize cancer in 5 molecular subtypes. One of these subtypes is "HER2-enriched" (HER2-E). While clinical and PAM50-based molecular classification of HER2+/- often agree, they do

not always.

Prat and Bianchini (2014) find evidence that clinically HER2+ breast cancer patients classified to the HER2-E subtype by PAM50 have a higher rate of 3-year pathologic complete response (pCR) after receiving trastuzumab and chemotherapy, compared to chemotherapy without trastuzumab. No statistically significant differences in pCR between treatment arms were found for those clinically HER2+ patients classified (in contradiction) by PAM50 as having any of the other four cancer subtypes. Here, pCR is defined as the absence of residual invasive breast cancer at the end of the 3-year period.

We reanalyze the publicly available subset of data from the NeoAdjuvant Herceptin (NOAH) randomized clinical trial (Gianni et al., 2010), which can be accessed in the Gene Expression Omnibus under the accession number GSE50948 (Edgar et al., 2002). We explore, which, if any, of the PAM50 genes can be used to characterize benefit from trastuzumab in clinically HER2 patients. The NOAH study population consisted of 334 women, 235 of whom had clinically HER2+ breast cancer. Clinically HER2+ patients in the NOAH trial were randomized to either receive treatment with trastuzumab in addition to neoadjuvant chemotherapy ($T = 1$) or to receive neoadjuvant chemotherapy alone ($T = 0$). Gene expression levels that passed quality control are available for the baseline tumor biopsies of 111 HER2+ patients.

Our response variable of interest is a binary indicator of 3-year pCR. In the $T = 0$ group, 25.5% of participants experienced pCR after three years; in the treatment group, 46.7% experienced pCR after three years. This is consistent with the finding in Gianni et al. (2010) that trastuzumab was beneficial overall for women with clinically HER2+ breast cancer. There is no mention of censoring in (Prat and Bianchini, 2014) and we use logistic loss for this analysis.

The supplemental materials in Prat and Bianchini (2014) list 9 unique probes used to represent the PAM50 genes classifying patients to the HER2+ subtype; we use combinations of these 9 expression levels as the candidate biomarkers for our data example. For one, we use our method to fit a univariate model with each of the 9 probes as a single feature.

We also fit our method with 36 additive models, as described in Section 2.2.6, where each possible combination of two probes is used as the set of features.

We compute one observed and 100 permuted test statistics as described in Section 2.3 for each of 9 univariate models and 36 2-dimensional additive models. We estimate a false discovery rate (FDR) as described in Appendix A.2.

Figure 2.3 shows the fitted values for the single probe (ACTR3B) in the univariate model with the lowest FDR estimate (0.30). Figure 2.3 suggests that trastuzumab only improves pCR for HER2+ patients with low levels of ACTR3B, but the associated FDR estimate (0.30) is relatively high. For comparison, Figure 2.4 shows the predicted values for both treatment groups from a logistic regression where the response variable is pCR and the features are ACTR3B, T, and (ACTR3B $\times$ T). The estimated response-curves from linear-logistic regression do not cross each other for any value of ACTR3B in the sample range.

Figure 2.5 shows a surface plot for the 2-probe combination from the additive framework model (ACTR3B and BAG1), which had an estimated FDR of 0.15. Figure 2.5 suggests that HER2+ breast cancer patients with relatively low expression levels of ACTR3B *and* low levels of BAG1 may not benefit from trastuzumab, while all other patients do benefit. If this finding is validated in independent datasets, an implication is that clinically HER2+ patients who have low levels of ACTR3B and low levels of BAG1 can avoid treatment with trastuzumab.

2.6 Simulations

2.6.1 Simulation Scenarios

We simulate data from 12 different scenarios across a range of signal-to-noise ratios (SNRs), using $n = 100$ observations and 100 permutations for testing. Figure 2.6 shows the simulated response variable as a function of the candidate biomarker for an SNR of 1.5.

In every scenario we simulate candidate biomarker $x_i \sim \text{Uniform}(0, 1)$ and treatment indicator $t_i \sim \text{Bernoulli}(0.5)$ for $i = 1, \ldots, n$.

In the four mean-shift simulation scenarios (first row of Figure 2.6), the response-curves for the treatment groups are mean shifts of each other and we simulate the response variable as

$$y_i \mid t_i = f(x_i) - \delta t_i + \epsilon_i,$$

for $i = 1, \cdots, n$, where we and we chose $\delta = 0.6$ and used noise parameter $\nu^2 = \frac{\widehat{\text{Var}}(f(X) - \delta T)}{\text{SNR}}$ to generate $\epsilon_i \sim N(0, \nu^2)$. The function f is either linear, piecewise linear, piecewise constant, or sinusoidal.

In the second row of Figure 2.6 (non-qualitative interaction), we simulate the response variable with

$$y_i \mid t_i = f_0(x_i)t_i + f_1(x_i)(1 - t_i) + \epsilon_i t_i + \gamma_i(1 - t_i)$$

for $i = 1, \cdots, n$, where we specify noise parameters $\nu_0^2 = \frac{\widehat{\text{Var}}(f_0(x))}{\text{SNR}}$ and $\nu_1^2 = \frac{\widehat{\text{Var}}(f_0(x))}{\text{SNR}}$ to generate $\gamma_i \sim N(0, \nu_0^2)$ and $\epsilon_i \sim N(0, \nu_1^2)$. The functions f_0 and f_1 yield linear, piecewise linear, piecewise constant, or sinusoidal response-curves for the treatment groups that do not cross each other.

In the last row of Figure 2.6 (qualitative interaction), the response-curves population curves for the treatment groups cross each other and we generate

$$y_i \mid t_i = f_1(x_i)t_i + f_2(x_i)(1 - t_i) + \epsilon_i$$

for $i = 1, \cdots, n$, where we specify noise parameter $\nu^2 = \frac{\sum_{i=1}^{n} \left((f_0(x_i) - f_1(x_i))^+ \right)^2}{\text{SNR}}$ to generate $\epsilon_i \sim N(0, \nu^2)$. The functions f_0 and f_1 yield linear, piecewise linear, piecewise constant, or sinusoidal response-curves for the treatment groups that cross each other.

2.6.2 Simulation Results

Comparison to Linear Regression

Let $\mathcal{X} \subset \mathbb{R}$ be values of candidate biomarker X that have clinical relevance. Then a simple alternative to our testing procedure is fitting the linear regression

$$E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i + \beta_2 t_i + \beta_3 (x_i \times t_i)$$

Under this model, qualitative interaction occurs if there exist $x, x', x'' \in \mathcal{X}$ such that

$$\beta_2 + \beta_3 x = 0, \quad \beta_2 + \beta_3 x' < 0, \quad \beta_2 + \beta_3 x'' > 0$$

so we are testing

$$H_0 : \text{there do } \textit{not} \text{ exist } x, x', x'' \in \mathcal{X} \text{ such that } x = -\frac{\beta_2}{\beta_3}, x' < -\frac{\beta_2}{\beta_3}, x'' > -\frac{\beta_2}{\beta_3}$$

We derive our procedure for testing this hypothesis in Appendix A.4.

Figure 2.7 plots the the share of 1000 replications where the null hypothesis was rejected, as a function of SNR. Our method is shown in blue squares and the procedure based on linear regression is shown in orange crosses. Rejecting H_0 in any of the first two rows is a Type I error, while it is the correct decision in the third row.

For these simulation scenarios, Figure 2.7 suggests that both methods have Type I error rates below the p-value threshold of 0.05. The procedure based on linear regression only has higher power than our method when the truth is linear; for the other three truth types, our method offers a substantial improvement in power.

The observed conservativeness of our test under the null is in line with the discussion in Section 2.3.1. When the null is true and the response-curves between the treatment groups are well-separated, the observed test statistic has a large point mass at 0; a permuted test statistic, on the other hand, is much more likely to be positive in this setting because it is

based on response-curves that are averaged across the two treatment groups and thus are more likely to cross. As the signal-to-noise ratio increases, we would expect the observed test statistic to have more mass at 0 (while the permuted test statistic is still likely to be positive) and thus for conservativeness to increase. One potential saving grace is that at the boundary of our null hypothesis space, we have $\theta_0(x) = \theta_1(x)$ for all x — in which case our test is no longer conservative. We also attempted to fix this conservatism: We considered generating null samples using a parametric bootstrap from our null-constrained estimates as an alternative to permutations. However, we found that this approach was anti-conservative.

Since we used fused lasso as the basis for our estimation, the fitted values for each treatment group are piecewise constant across the range of the feature. The number and location of the knots are chosen data-adaptively instead of being specified beforehand. A simple alternative to this approach would be to pre-specify a fixed number of evenly spaced knots and use the group means in each region as the fitted values. Appendix A.5 compares the power and Type I error of our method to this alternative.

2.7 Conclusion

To test whether treatment is either uniformly superior or uniformly inferior for all patients, we propose a convex, regression-based test for qualitative interactions. We implement our method in Python and share an example of using it to search for a subset of HER2+ breast cancer patients who benefit from adjuvant chemotherapy.

Our work has several notable limitations. One is the long runtime of our Python implementation due to its reliance on a general convex solver. In future work that focuses on computation, we would like to devise a custom solver to work more efficiently with additive models. In addition, our current work only tests a strong null hypothesis. The estimation stage of our framework suggests a subset of patients who benefit from treatment (i.e. the values of the candidate biomarker(s) for which the estimated response in the treatment group is superior to the estimated response in the control group). However, no formal significance test of average treatment effect is run in that subgroup. In addition, the testing stage of

our framework is also stated for pre-specified biomarker candidates and does not perform variable selection among a set of potential candidates.

Although our current implementation estimates the response variable as a piecewise-constant function of each continuous feature (either alone in a univariate model or as part of an additive model) and tests for qualitative interaction, we presented a general underlying framework that accommodates alternative penalty functions that induce different structures on the fitted values. The constraints in the optimization framework can also be modified to test whether the population average benefit from treatment is at least a pre-specified amount.

2.8 Figures and Tables

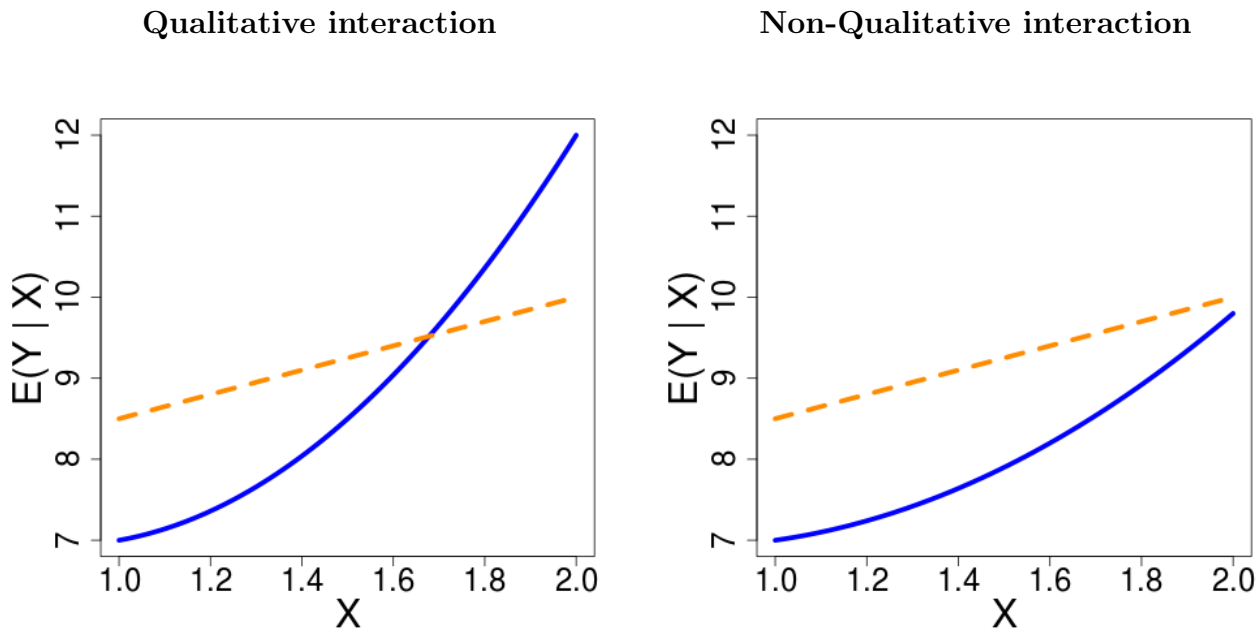


Figure 2.1: Mean of response variable Y, conditional on candidate biomarker X in the treatment group (blue, solid) and control group (orange, dashed).

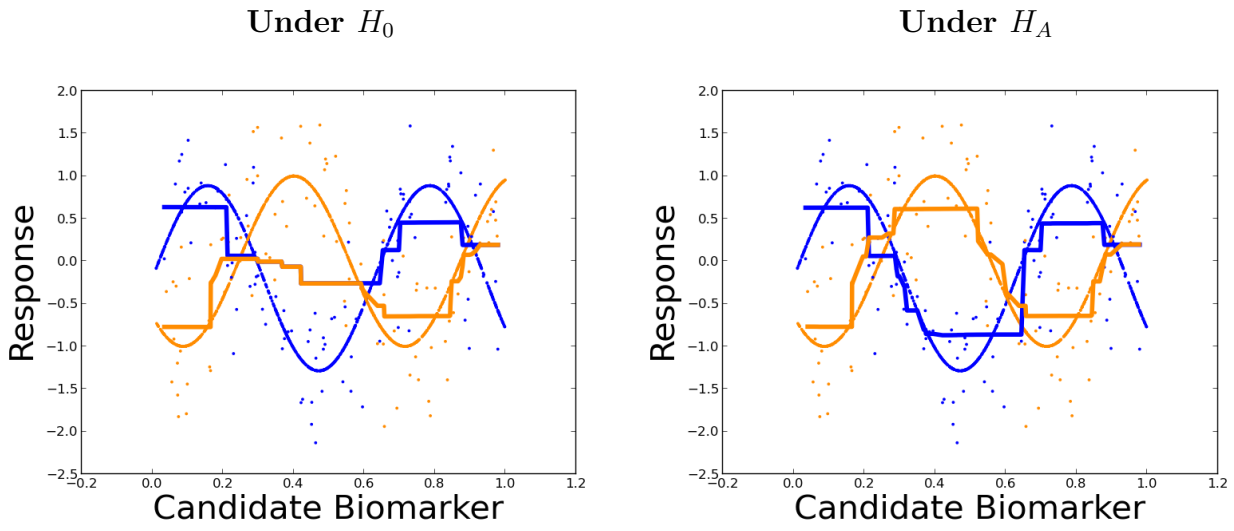


Figure 2.2: Example piecewise-constant solutions to (2.3) and (2.4) using the 1-dimensional fused lasso. The crossing sinusoidal curves show the true relationship between the response and candidate biomarker.

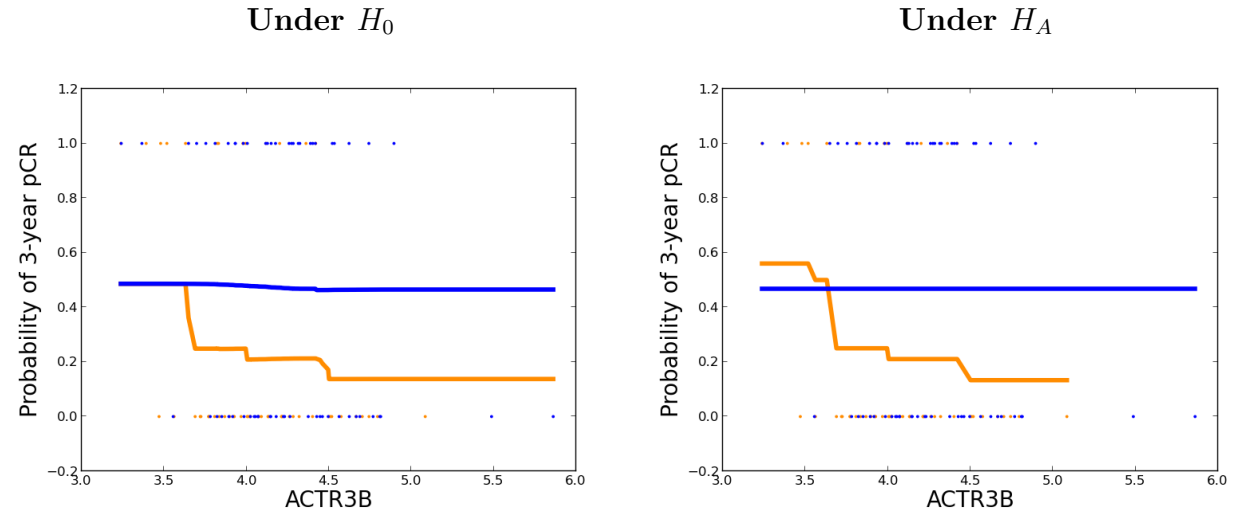


Figure 2.3: Predicted values from our method with ACTR3B as the feature (estimated FDR=0.30). The treatment group is shown in blue and the control group is shown in orange

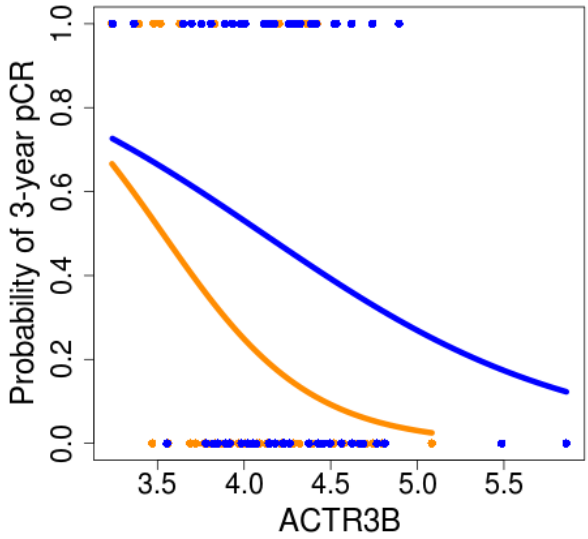


Figure 2.4: Predicted values from logistic regression with ACTR3B, T, and (ACTR3B $\times$ T) as the features. The treatment group is shown in blue and the control group is shown in orange

Under H_0 **Under H_A**

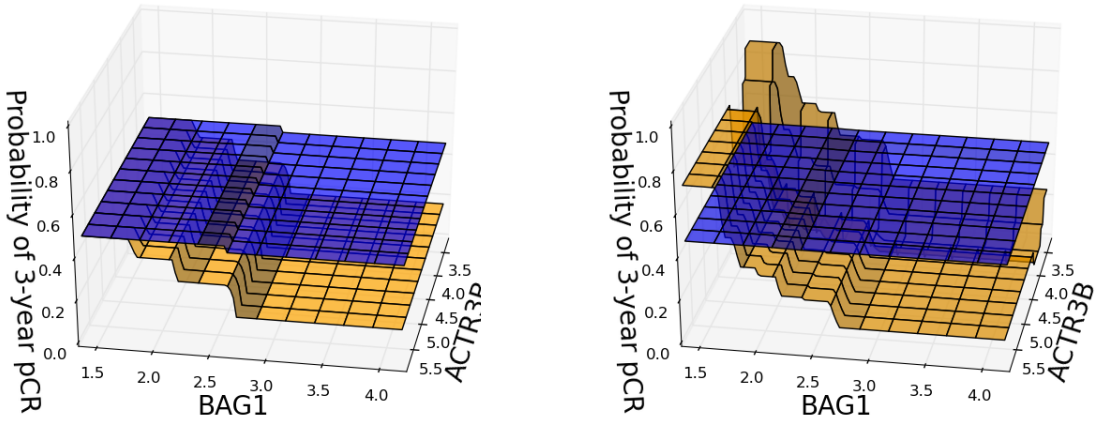


Figure 2.5: Fitted surfaces for the additive model with ACTR3B and BAG1 as features (estimated FDR=0.15). The treatment group is shown in blue and the control group is shown in orange

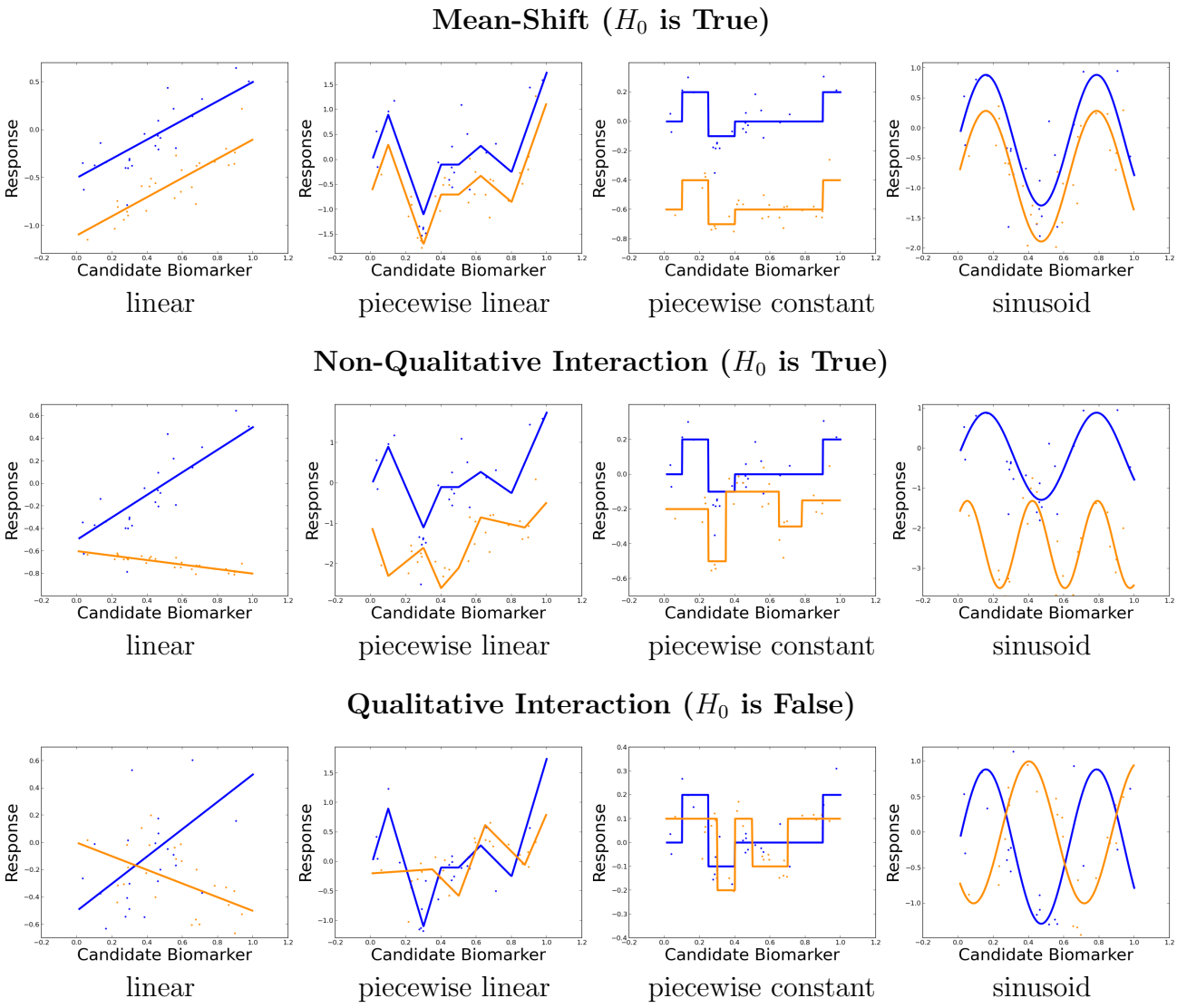


Figure 2.6: The 12 simulation scenarios with SNR = 1.5

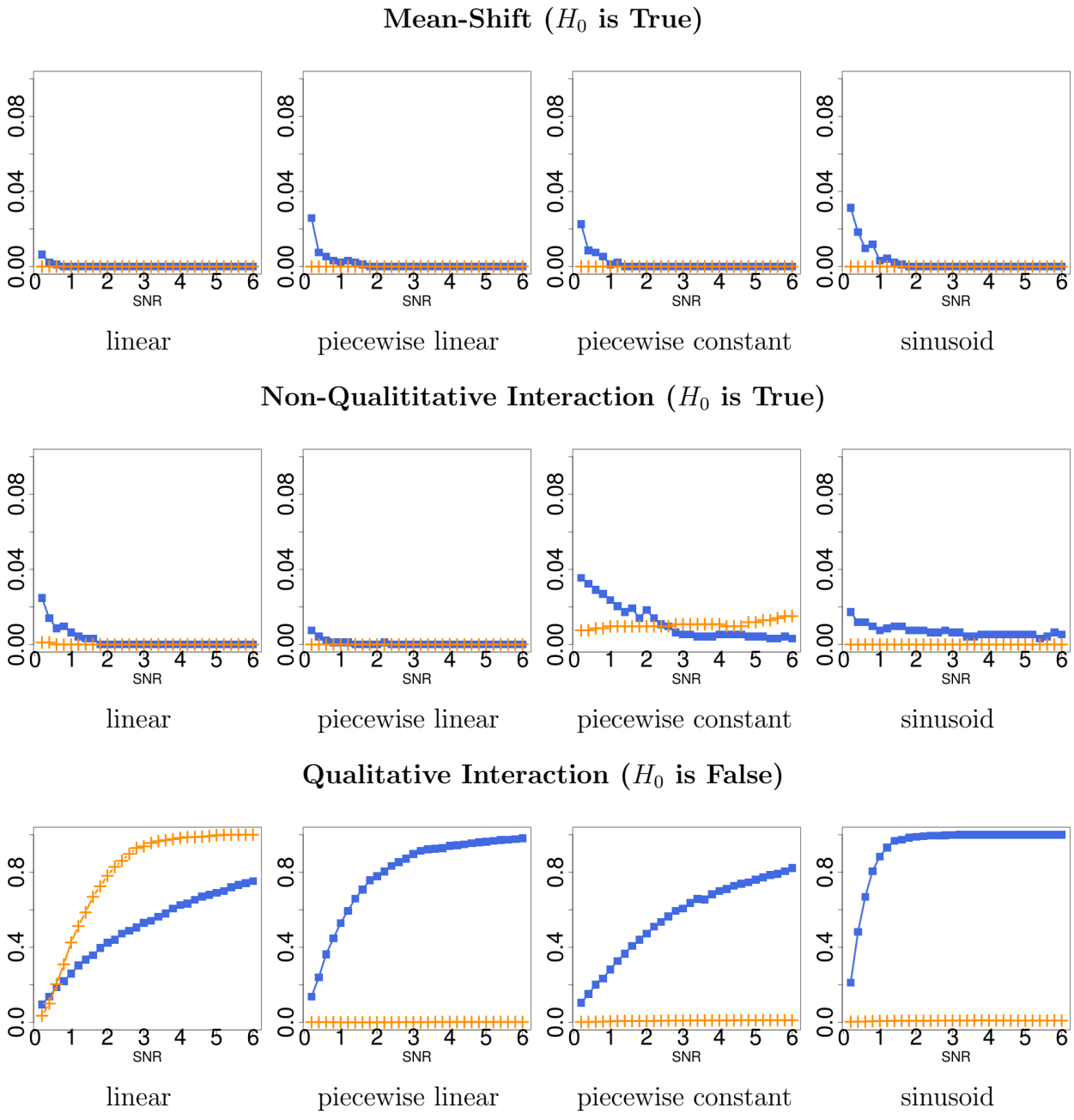


Figure 2.7: Share of 1000 simulations with a p-value below 0.05, as a function of 30 SNR values across each of the 12 scenarios (permutation-based p^* from fused lasso with data adaptive knots in blue squares, p-values from linear regression in orange crosses)

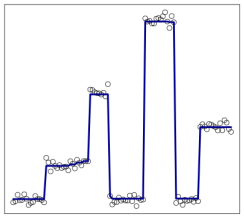
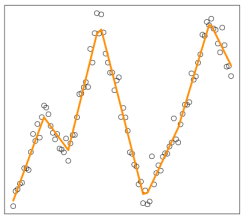
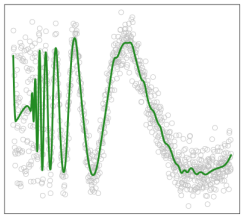
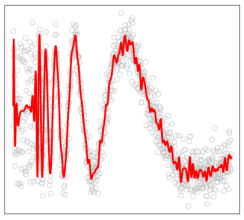
Method	$P(\theta)$	Example Fit
1-dimensional fused lasso (Tibshirani et al., 2005)	$\sum \lvert\theta_{i+1} - \theta_i\rvert$	
1st order trend filtering (Tibshirani, 2013)	$\lvert\lvert D^{(x,2)}\theta\rvert\rvert_1$	
2nd order trend filtering (Tibshirani, 2013)	$\lvert\lvert D^{(x,3)}\theta\rvert\rvert_1$	
Cubic smoothing splines (Hastie et al., 2008)	$\theta^T \Omega_N \theta$	

Table 2.1: Examples of smoothing methods

Chapter 3

USING PROPENSITY SCORES TO DEVELOP AND EVALUATE TREATMENT RULES WITH OBSERVATIONAL DATA

3.1 Introduction

Precision medicine strives to leverage an individual's specific characteristics to determine the most beneficial course of treatment for that individual. The implementation of precision medicine is a challenging task in clinical settings where available information and resources are always limited.

In this chapter, we consider the practice of precision medicine in settings that present two particular sets of challenges: 1) data come from observational studies where individual characteristics may influence treatment assignment; and 2) the data at hand may measure individual characteristics that will not be available in future clinical settings. As we summarize in Table 3.2, there are statistical methods to account for the observational study design in 1) but they do not account for the subtlety regarding variable roles in 2), and they often lack a user-friendly software implementation that would allow practitioners to reliably apply them.

Some characteristics that affect treatment recommendation in a particular study may be unavailable in future clinical settings, while other characteristics that did not directly affect treatment assignment may nonetheless be informative about treatment response in future clinical settings. For example, some observable individual characteristics may not be interpretable in a particular clinical setting due to a dearth of scientific knowledge about their role in a disease pathway (e.g. gene expression levels); such characteristics do not directly affect the treatment decisions of clinicians but nonetheless may be predictive of treatment

benefit. Further, some information that influences treatment assignment (e.g. prognosis) may be measured subjectively in one particular clinical setting and not be measured again in the same manner in other settings; such a variable directly affects the treatment decision of clinicians in one setting but cannot directly influence treatment recommendations in future clinical settings. Each of these variable types should be handled differently by statistical methods that require distinct prediction of treatment assignment and prediction of outcome.

We propose a principled framework (along with a user-friendly implementation in the `R` package `DevTreatRules`) that appropriately handles these distinct variable when developing and evaluating a *treatment rule* based on data from an observational study where treatment is not independent of individual characteristics. A treatment rule is a function that recommends treatment based on individual characteristics; to be useful to practitioners, a treatment rule must appropriately handle individual characteristics as they are actually observed in clinical settings. Our framework yields a treatment rule that accounts for the clinically distinct roles of individual characteristics in an interpretable way by asking two simple questions that reflect clinical rather than statistical expertise:

1. Which characteristics will be measured in the same manner in future settings where the estimated rule would be applied?

2. Which characteristics potentially influence treatment assignment and thus must be accounted for as confounders?

Our framework also places a clear emphasis on developing and evaluating the treatment rule on independent datasets which is also absent from other papers in the treatment rule literature; we elaborate on the novel contributions of this paper in Section 3.1.1 and discuss where this chapter fits in the broader literature in Section 3.2.2.

3.1.1 Contributions of this Chapter

1. **Categorization of individual characteristics that is appropriate for treatment rule development.** We propose partitioning each individual characteristic

collected in an observational study (aside from outcome and treatment variables) into the four clinically distinct categories shown in Table 3.1: C^{TI}, C^{TN}, C^{NI}, or C^{NN}, where the abbreviations in each superscript tell us whether each characteristic (C) affects treatment assignment in the current study (TN), is expected to be observed in independent studies (NI), both (TI), or neither (NN).

Potentially influences treatment in current study	**Observed in independent clinical settings**	
	Yes	No
Yes	C^{TI}	C^{TN}
No	C^{NI}	C^{NN}

Table 3.1: Proposed partitioning of individual characteristics in an observational study

As we discuss in Sections 3.3 and 3.4, the four clinically distinct variable types in Table 3.1 play distinct statistical roles our proposed framework. Assuming no unmeasured confounding, only the C^{TI} and C^{TN} variables are potential confounders of the association between treatment and outcome in the current study.[1]

Additionally, we make explicit (as is not done in previous work) that only the C^{TI} and C^{NI} variables are viable candidates for inclusion in a treatment rule because the remaining C^{TN} and C^{NN} will not be available in future clinical situations where a proposed treatment rule would be implemented. This is a common-sense notion but nonetheless we believe its emphasis is critical to avoid all individual characteristics being treated equivalently as potential predictors (which would be implied by a direct implementation of past approaches) and thus to serve the goal of precision medicine: use the information at hand to obtain a reliable recommendation of which treatment will best serve each individual.

[1]Other variables may still confound the association between treatment and outcome in future independent clinical settings and are still accounted for as described in Section 3.4.3.

Here is a brief example of how each variable type may present itself in a hypothetical observational study:

(a) **Example of $C^{\mathbf{TI}}$: Age.** In the study population, physicians might have been more likely to recommend treatment to older individuals. In independent clinical settings, where an estimated treatment rule would be applied to benefit future individuals, we are confident that age can be reliably collected on the same scale.

(b) **Example of $C^{\mathbf{TN}}$: Center-specific measure of prognosis.** In the study population, clinicians may have been more likely to recommend treatment to individuals with a poor prognosis as estimated by a center-specific set of guidelines (e.g. a hospital's standard rule-of-thumb procedure based on their specific doctors' prior experiences with individuals from the population). In independent clinical settings taking place in different centers, clinicians would not estimate an individual's prognosis with that same center-specific approach.

(c) **Example of $C^{\mathbf{NI}}$: Gene expression levels.** In the study population, individuals' gene expression levels may have been measured but clinicians did not consider the information when recommending treatment due to a lack of scientific knowledge about the genes' roles in disease progression and response to treatment. In independent clinical settings, gene expression levels might still be reliably measured and would be eligible to inform treatment decisions if a newly developed treatment rule suggests their importance or if other scientific knowledge becomes available in the interim.

(d) **Discussion of $C^{\mathbf{NN}}$:** C^{NN} consists of variables in a dataset that are believed to have no role in influencing treatment and cannot be reliably collected in future clinical settings; the variables in C^{NN} are not of interest to development or evaluation of treatment rules. An example could be a study ID variable, where the characteristic would implicitly be discarded even before applying our framework. We may also find observational studies where no variables are classified as C^{NN} .

This category will not be further discussed in this chapter; we just mention it here in the interest of completeness in Table 3.1.

It will be useful in later sections to define: $\mathbf{C}^{\mathrm{I}} \equiv (C^{\mathrm{TI}}, C^{\mathrm{NI}})$ as all observed individual characteristics that are also expected to be observed in independent clinical settings. Also, we define $\mathbf{R}$ as a subset of the variables contained in $\mathbf{C}^{\mathrm{I}}$ that the researcher believes may affect response to treatment and thus are viable candidates for a treatment rule. For example, the height of study participants may be routinely collected in both current and future datasets (so it would be a part of $\mathbf{C}^{\mathrm{I}}$) but a researcher may believe height s is very unlikely to affect an individual's response to treatment, so height would not be included in $\mathbf{R}$. We also define $\mathbf{C}^{\mathrm{T}} \equiv (C^{\mathrm{TI}}, C^{\mathrm{TN}})$ as all individual characteristics that potentially affect treatment in the current observational study. In the `BuildRule()` and `EvaluateRule()` functions from `DevTreatRules`, users are required to provide the characteristics in $\mathbf{C}^{\mathrm{T}}$ with the argument `names.influencing.treatment` and the characteristics in $\mathbf{R}$ with the argument `names.influencing.rule`. That is, individual characteristics are never entered into the package without first being categorized by the user.

In Section 3.4 we explain how, in the context of our proposed categorization of individual characteristics, observation weights implied by the inverse-probability-of-treatment weighting (IPW) approach (Austin, 2011) to estimating treatment effects in observational studies actually depend on a ratio of two distinct propensity scores; this is contrast to using a single propensity score as is done in past approaches that do not partition individual characteristics into into clinically distinct categories as in Table 3.1.[2] This ratio of weights is related to the idea of "stabilized weights" discussed in Robins et al. (2000).

[2]The *propensity score* is the probability of a particular treatment assignment conditional on observed individual characteristics; the propensity score has a long history of importance in comparison of treatment groups in non-randomized studies at least back to Rosenbaum and Rubin (1983).

2. **Separation of development and evaluation of the treatment rule.** At minimum, we advocate using development/evaluation splitting (or development/validation/evaluation splitting if model selection is performed) to partition an observational dataset into independent subsets before developing a treatment rule on the development subset and estimating the benefit of using the rule on the independent evaluation subset; we discuss the importance of data-splitting in Section 3.4.3. Ideally, development and evaluation of the rule would be conducted on datasets collected from two observational studies whose participants are independently sampled from the same population, but data-splitting may be more useful in practical applications because often only one dataset from a single study is available. The `DevTreatRules` package automates the process of data-splitting but also supports the separate development and evaluating the treatment rule on datasets from distinct observational studies if that is available instead.

3. **Implementation of our entire framework in a freely available `R` package.** We would like practitioners to be able to bring the insights discussed in this paper to bear on the observational datasets they have at hand, rather undertaking the time-consuming and error-prone task of trying to carefully implement, from the ground up, a proposed approach that lacks an `R` package. As a result, the software implementation of our approach (including variable categorization and separate development/validation/evaluation of a treatment rule) is freely available as the `R` package `DevTreatRules`. We also share the code to reproduce our simulations on `github.com/jhroth/simulations-split-regression`. Although we are not permitted to share the data used for our data example, we still share the code we used to format and analyze the raw data files so that future researchers who receive access to the data can directly apply our code to the same dataset (`https://github.com/jhroth/data-example-split-regression`).

3.1.2 The Rest of the Chapter

In Section 3.2, we discuss previous research on treatment rule estimation and where this chapter fits in. In Section 3.3, we motivate our approach to building and evaluating a treatment rule in a setting where the association between treatment and outcome is complicated by the presence of individual characteristics with the variable types C^{TI} , C^{TN} , and C^{NI} . In Section 3.4, we summarize the theory (and cite additional detail in the Appendix) that underlies our preferred *split-regression* estimation method and present the explicit estimation recipe that is implemented in `DevTreatRules`. Section 3.5 describes a small simulation study. Section 3.6 presents a data example where we build and evaluate treatment rules on a dataset collected by the Women's Health Initiative Observational Study component. We end with a discussion in Section 3.7.

3.2 Previous Work

The framework presented in this chapter draws on two previous bodies of work: 1) literature on estimating treatment rules, which can be partitioned into *direct* and *indirect* methods as discussed later in this section; and 2) tools from the causal inference literature that describe how to estimate the average treatment effect (ATE) in study designs where treatment is not randomized. There is notable overlap between these two topics even if the connection is not always made explicit in the literature; nearly all methods for estimating treatment effect that are appropriate in settings with non-randomized treatment rely on the inverse-probability-of-treatment weighting (IPW) approach to balance observed confounders across the treatment groups. The propensity score has played a vital role in facilitating reliable comparisons of outcomes across treatment groups in non-randomized studies at least back to Rosenbaum and Rubin (1983). Austin (2011) presents an excellent and accessible overview of the ATE derived from the causal inference literature and how it informs propensity-score-based estimation strategies for non-randomized treatment assignment (e.g. the IPW method).

In this section, we focus on previous work on estimating treatment rules and we defer discussion of the causal inference literature to Section 3.4.1 and Section 3.4.3, where we can more clearly show how it shapes the target parameter of interest that is critical to our preferred approach for developing and evaluating a treatment rule.

3.2.1 Previous Work on Estimating Treatment Rules: Two General Approaches

There is active and exciting statistical research on methods to predict the most beneficial treatment option for a specific individual, in line with the objectives of precision medicine. Existing statistical approaches generally fall into one of two categories, sometimes labeled as *indirect* vs. *direct*.

Direct methods are motivated by optimizing performance of the treatment rule itself in a population of interest rather then optimizing the accuracy of predicted outcomes and then using each individual-level predicted outcome to decide which treatment to assign (which would be the case with an indirect method). That is, these methods *directly* estimate an optimal treatment rule rather than prioritizing estimation of the expected outcome conditional on individual characteristics and treatment assignment (i.e. a regression function), and then taking the additional step of assigning an individual to the treatment with the most desirable predicted outcome (as in a indirect approach).

Direct methods generally seek the treatment rule within a particular class of allowable rules that maximizes an estimate of clinical benefit, and have the appealing motivation of not being vulnerable to mis-specification of the regression function predicting outcome based on individual characteristics (though direct methods must still make other modeling assumptions to which they are sensitive). Lipkovich et al. (2017) present an outstanding recent survey of statistical methods for estimating treatment rules in RCTs; direct methods are part of that review's "optimal treatment regimes" category. Zhang et al. (2015) prespecify that the treatment rule must be a nested sequence of "if, then" statements with one or two individual characteristics involved in each statement. Zhang et al. (2012c) take a similar approach but instead use a regression model to dictate the assumed structure of the

treatment rule. Other direct methods make less restrictive assumptions about the form of the treatment rule at the cost of interpretability of the rule. Zhao et al. (2012) propose outcome-weighted learning (OWL), which defines the optimal treatment rule as the solution to a weighted classification problem whose solution can be approximated using a modified form of support vector machines (SVM), a well-established statistical learning tool (Hastie et al., 2008). As noted by Zhang et al. (2012a), OWL can be viewed as part of a general weighted-classification framework for estimating a treatment rule, so any classification method that accommodates observation weights (e.g. classification trees or penalized regression (Hastie et al., 2008)) is a viable alternative to SVM in the estimation stage of OWL. The interactions-based procedure from Chen et al. (2017) based on the earlier work of Tian et al. (2014) is a recent example of a promising direct approach to estimating treatment rules with even greater flexibility.

Indirect approaches (e.g. Kang et al. (2014); Cai et al. (2011); Lu et al. (2013); McKeague and Qian (2014); Ciarleglio et al. (2015)) typically assume structure on the regression function linking the conditional mean outcome to individual outcomes and a treatment indicator. The ideal treatment assignment for a individual is then inferred by predicting his or her expected outcomes across possible values of the treatment variable using the regression model, and choosing the treatment option with the most desirable predicted outcome (e.g. a larger mean time until relapse or a smaller probability of 5-year relapse). In Lipkovich et al. (2017), indirect approaches are given the label "global outcome modeling". One might consider Chapter 2 in this dissertation an indirect approach (for use with RCT data only) because under its alternative hypothesis it estimates unconstrained conditional mean outcomes for both treatment and standard-of-care groups as a function of characteristics, and thus could be interpreted as recommending treatment to an individual whose predicted outcome conditional on receiving treatment is superior to the prediction under standard-of-care. However, the focus of Chapter 2 was on testing the global null hypothesis that there is no subpopulation that benefits from treatment, rather than on identifying particular subsets of individuals for whom treatment appears beneficial.

Importantly, as mentioned in Lipkovich et al. (2017), some existing direct and indirect methods for estimating treatment rules are adaptable to observational study designs where treatment assignment is not independent of individual characteristics by accommodating the IPW approach (Austin, 2011), which re-weights observations by the inverse of their estimated propensity scores so clinically observed confounders are roughly balanced between the treatment groups. The IPW adjustment thus allows for a sensible direct comparison of mean re-weighted outcomes across treatment groups that is reflective of the underlying population where the rule would be applied in the future.

As of yet, there is no consensus among researchers on whether the direct or indirect approach to estimating treatment rules yields superior results.

3.2.2 Previous Work on Estimating Treatment Rules: Where This Chapter Fits In

To help situate this chapter in the literature, Table 3.2 presents a selection of direct and indirect methods for estimating treatment rules and whether each satisfies five criteria:

1. Does it accommodate observational data?

2. Does it distinguish between observed individual characteristics that will $\left(C^{\mathrm{TI}}, C^{\mathrm{NI}}\right)$ or will not $\left(C^{\mathrm{TN}}\right)$ be observed in independent clinical settings?

    - For indirect methods, failing to make this distinction might lead practitioners to include all potential confounders as predictors in a regression model; if any these potential confounders are unavailable in future clinical settings (i.e. classified as C^{TN}), then in these future clinical settings it would not be possible to form the model-based predictions of conditional mean outcome that are the linchpins of indirect methods.
    - For direct methods, failing to make the distinction would generally lead to one set of individual characteristics $X \equiv \left(C^{\mathrm{TI}}, C^{\mathrm{NI}}, C^{\mathrm{TN}}\right)$ being implicitly defined as both the potential confounders (as predictors for the propensity score model) and

the inputs to the treatment rule, despite application of the rule in future clinical settings being impossible due to the absence of C^{TN} .

- We note that, in a clinical situation where all observed potential confounders are expected to be observed in future settings (i.e. C^{TN} in Table 3.1 is empty), then we would not advocate using our variable classification system and preferred method. Many indirect approaches would immediately be appropriate in that situation, including simply regressing clinical outcome on potential confounders (i.e. C^{TI}) and characteristics that may affect treatment response but not treatment assignment (i.e. C^{NI}) separately in each treatment group (on the development set). Then on an independent evaluation dataset, one could simply assign future individuals to receive the treatment under which their model-based predicted outcome is most desirable. If the form of the regression model was prespecified, then the single set of predictions on the evaluation dataset can be mapped to obtain a trustworthy estimate of treatment rule benefit (e.g. as discussed in Austin (2011)); if different models are considered, model selection should be performed on an additional independent dataset (the validation set) before moving onto the evaluation dataset.

3. Can it accommodate a range of statistical learning methods?

4. Does it offer an R package?

5. Does it share the code needed to reproduce simulations or data application?

Framework	Type of approach	Accommodates observational data	Distinguishes between $(C^{\text{TI}}, C^{\text{NI}})$ and (C^{TN})	Accommodates a range of statistical learning methods	R package	Shares code to reproduce simulations or application
Zhao et al. (2012)	Direct	Yes	No	No	Yes	No
Zhang et al. (2012a,c)	Direct	Yes	No	Yes	No	Yes
Zhang et al. (2015)	Direct	Yes	No	No	No	Yes
Qian and Murphy (2011)	Direct	No	No	No	No	No
Chen et al. (2017)	Direct	Yes	No	Yes	No	Yes
Cai et al. (2011)	Indirect	No	No	Yes	No	No
Lu et al. (2013)	Indirect	No	No	Yes	No	No
Kang et al. (2014)	Indirect	No	No	Yes	No	Yes
McKeague and Qian (2014)	Indirect	No	No	Yes	No	No
Ciarleglio et al. (2015)	Indirect	Yes	No	No	No	Yes
This chapter	**Indirect**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 3.2: Selected methods for estimating treatment rule

As seen in Table 3.2, none of these selected previous methods (and, to our knowledge, no other method in the literature) distinguishes between observed individual characteristics using the clinically meaningful categorization of whether they will be available in future clinical settings (in which case they are sensible candidates for use in the treatment rule) or are not expected to be available in future settings (in which case they should not be used to build the rule). In addition, it appears that most indirect methods in the literature are motivated by and designed for the RCT setting; since there is no consensus in the literature about the superiority of either the direct or indirect approach (either overall or in any particular clinical situation) another useful contribution of this chapter is to provide an indirect approach that is framed by and appropriate for observational studies with non-randomized treatment assignment. The lack of available R packages implementing previous approaches is also re-enforced by Table 3.2.

We believe that our work fills a substantial gap in the literature by providing practitioners with a principled approach to appropriately classify variable types as in Table 3.1 and by providing the user-friendly `DevTreatRules`, which actually requires users to make their clinically informed variable categorizations when they apply the work to real data (using the arguments `names.influencing.treatment` and `names.influencing.rule` in its

`BuildRule()` and `EvaluateRule()` functions).

In addition, although it is not considered in the table, none of those previous methods explicitly integrates data-splitting to ensure that the stages of development/evaluation (or development/validation/evaluation) of a treatment rule are conducted on independent datasets to yield a trustworthy estimate of the rule's population impact; we emphasize data-splitting throughout this Chapter and in the formalized procedure in Section 3.4.2.

3.3 Motivating Example

Suppose an observational study recruits individuals with a particular type of cancer from a single hospital and collects baseline information at the time of recruitment. Further suppose that, for each individual, this observational dataset measures: months until relapse (Y); an indicator of receiving standard-of-care ($T = 0$) or additional chemotherapy in addition to the standard-of-care option ($T = 1$); age; a measure of day-to-day life functioning; and expression levels of p genes. For simplicity, we assume that clinicians decide to treat individuals based only on age and day-to-day life functioning. In contrast, the gene expression levels are unobservable by clinicians due to high cost and are also uninterpretable due to a lack of scientific knowledge about the genes' roles in disease progression and response to treatment. In this illustrative example, we assume no unmeasured confounding.[3]

In practice, there are two alternative scores used to an individual's day-to-day functioning. One is the Eastern Cooperative Oncology Group (ECOG) Performance Status, a grade ranging from 0 to 5 where a lower grade represents fewer restrictions in day-to-day life (Oken et al., 1982). The other is the Karnofsky Performance Status (KPS), an 11-point scale ranging from 0 to 100 where a lower value represents more day-to-day restrictions (Karnofsky, 1949).

[3]In practice, however, gene expression levels may serve as surrogates for unobserved confounders. For instance, it is estimated that 20%-30% of breast cancer cases consist of an over-expression of human epidermal growth factor receptor type 2 (HER2), which can be targeted by additional treatment that inhibits HER2 generation (Joensuu et al., 2006; Hudis, 2007). For types of cancer with unknown subtypes the gene expressions conducted in an observational study may be associated with (but not explicitly define, as with HER2) disease subtypes that may be more or less vulnerable to disruption by the treatment mechanism.

As an added complication, we suppose that in this particular observational study day-to-day functioning is measured using the ECOG score, but that KPS is used instead of ECOG score in future clinical settings where we would like to apply our developed treatment rule in the future.[4] In contrast, we assume an individual's age and expression levels of p genes $(\text{gene}_1, \ldots, \text{gene}_p)$ will be reliably collected in future clinical settings.

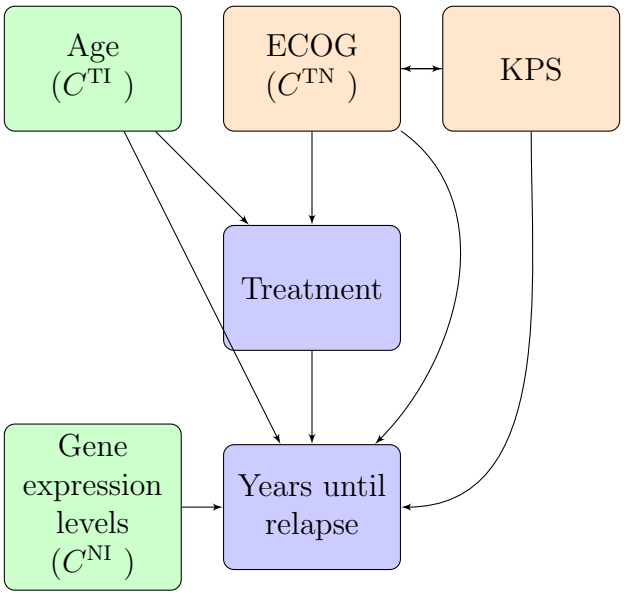


Figure 3.1: Potential data-generating mechanism. Each node shows a individual characteristic (and, for those besides treatment and outcome, its corresponding variable type from Table 3.1 in parentheses)

Figure 3.1 shows our hypothesized data-generating mechanism. We have the following variable types in this example: $C^{\text{NI}} = (\text{gene}_1, \ldots, \text{gene}_p)$, $C^{\text{TI}} = \text{age}$, $C^{\text{TN}} = \text{ECOG}$, $\mathbf{C}^{\text{T}} = (\text{age}, \text{ECOG})$, and $\mathbf{C}^{\text{I}} = (\text{age}, \text{gene}_1, \ldots, \text{gene}_p)$. The KPS variable is a potential confounder only in independent clinical settings (but not the current dataset) and as such KPS plays an important role in evaluation of the rule that we discuss in Section 3.4.3.

[4] Although in practice there are mappings between ECOG and KPS, for illustrative purposes here we treat them as variables representing the same individual characteristic using qualitatively distinct scales (e.g. higher scores mean lower quality of life for KPS but a higher quality of life for ECOG) and suppose they are non-conformable.

We are interested in building and evaluating a classifier, for future individuals, that indicates their optimal treatment based on their gene expression values. However, in constructing this classifier, we must account for a) the confounding influence of age and day-to-day functioning on the relationship between treatment and months until relapse; and b) the distinct approaches to measuring day-to-day functioning across different clinical settings (based on use of ECOG or KPS).

3.4 Method

Here, we provide some theoretical justification for how the individual characteristics C^{TN}, C^{TI}, and C^{NI} should be used to develop a treatment rule and evaluate the rule's benefit. We will refer to this method as the *split-regression* approach to developing a treatment rule. As in Section 3.3, we will interpret T as an indicator of receiving a new treatment in addition to standard-of-care ($T = 1$) or standard-of-care alone ($T = 0$) and outcome Y as months until relapse.

3.4.1 Estimating the Rule

Our goals are 1) to identify a subset of the population – in terms of $\mathbf{R}$, the researcher-chosen subset of individual characteristics that are valid candidates for inclusion in a treatment rule – we expect to benefit from treatment and 2) to estimate the extent of benefit in this subpopulation.

Our estimation strategy begins with a discussion of a foundational parameter of interest, often called the *average treatment effect* (ATE) in the causal inference literature:

$$E[Y^1 - Y^0], \tag{3.1}$$

where, adopting the *potential outcomes* notation from causal inference literature, Y^1 is the months until relapse that would have been observed had an individual received treatment and Y^0 is the months until relapse that would have been observed had the individual received

standard-of-care. So (3.1) tells us the average difference in months until relapse when an individual in the population of interest receives treatment instead of standard-of-care. As detailed in Kennedy (2015), the quantity in (3.1) is identifiable under three assumptions: 1) *consistency*: $T = t$ implies $Y = Y^t$; 2) *no unmeasured confounding*: conditional on observing $\mathbf{C}^{\mathrm{T}}$, T is independent of Y^t; 3) *positivity*: if $P(\mathbf{C}^{\mathrm{T}} > 0)$ then $P(T = t \mid \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}) > 0$, for $t = 0, 1$.

As originally developed by Robins (1986) and discussed in Kennedy (2015), if the consistency, no unmeasured confounding, and positivity assumptions hold, then (3.1) is equivalent to

$$\psi \equiv \int_{\mathbf{C}^{\mathrm{T}}} \left\{ E\left[Y \mid \mathbf{C}^{\mathrm{T}}, T = 1\right] - E\left[Y \mid \mathbf{C}^{\mathrm{T}}, T = 0\right] \right\} dP(\mathbf{C}^{\mathrm{T}}), \tag{3.2}$$

which is known as a "g-computation" formula in the causal inference literature. To adapt (3.1) to the setting of identifying which treatment is expected to offer a superior outcome for an individual with characteristics $\mathbf{R} = \mathbf{r}$, we simply consider the subgroup-specific ATE

$$E[Y^1 - Y^0 \mid \mathbf{R} = \mathbf{r}]. \tag{3.3}$$

The subgroup-specific ATE in (3.3) is also a parameter of interest in Cai et al. (2011), the previous method that we believe is most similar to the one we discuss in this section. The practical limitations of Cai et al. (2011) on which this chapter expands are that it was designed only for the RCT setting, did not explicitly distinguish between individual characteristics $\mathbf{C}^{\mathrm{T}}$ and $\mathbf{C}^{\mathrm{I}}$, was not implemented in an R package, and did not share code to guide users in implementation.

To estimate (3.3) in the setting of non-randomized treatment assignment we want to estimate the target parameter

$$\psi(\mathbf{r}) \equiv \textstyle\int_{\mathbf{C}^{\mathrm{T}}} \left\{ E\left[Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R} = \mathbf{r}, T = 1\right] - E\left[Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R} = \mathbf{r}, T = 0\right] \right\} dP(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R} = \mathbf{r}), \tag{3.4}$$

which estimates the average treatment effect for individuals with characteristics $\mathbf{R} = \mathbf{r}$ that

will be observable in future clinical settings.

From (3.4), we see that we should treat the individuals defined by

$$\Omega^+ = \{\mathbf{r} \mid \psi(\mathbf{r}) > 0\}. \tag{3.5}$$

If $\mathbf{R}$ were a set of gene expression levels, for example, then Ω^+ would be the subset of gene expression levels for which treatment increases the expected number of months until relapse. We note that one could modify (3.5) to require that the expected increase in number of months until relapse is larger than some number $C > 0$. In either case, the expected improvement in months until relapse among the treated subpopulation is

$$\int_{\mathbf{r}\in\Omega^+} \psi(\mathbf{r}) dP(\mathbf{r}). \tag{3.6}$$

One possible approach for estimating the modified ATE in (3.4) would be to separately estimate $E\left[Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T = t\right]$ for $t = 0, 1$ and estimate $dP(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})$, then plug these estimates into (3.4); however, this approach requires estimation of a conditional density function that is only practical when the individual characteristics in $\mathbf{R}$ are perhaps one or two categorical variables with very few levels while in other situations the approach would prescribe a very complicated and highly variable average (as described in greater detail in Chapter 3 of Varadhan and Seeger (2013), for example).

Our goal is now to re-write our estimation target (3.4) as a simple minimization problem that does not involve estimation of the conditional density $dP(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})$. We begin by re-stating (3.4) as

$$\psi(\mathbf{r}) = f_1(\mathbf{r}) - f_0(\mathbf{r}), \tag{3.7}$$

where, for $t = 0, 1$,

$$f_t(\mathbf{r}) = \int_{\mathbf{C}^{\mathrm{T}}} E\left[Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R} = \mathbf{r}, T = t\right] dP(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R} = \mathbf{r}). \tag{3.8}$$

We emphasize that (3.8) is exactly equivalent to (3.4), just with altered notation. By taking the perspective in (3.8) we only need to estimate $f_t(\mathbf{r})$ for $t \in \{0, 1\}$ to obtain an estimate of the $\psi(\mathbf{r})$ in (3.4). As derived in the Appendix, it turns out that $f_t(\mathbf{r})$ can be written as the minimizer

$$f_t(\mathbf{r}) \equiv \arg\min_{f \in \mathcal{F}} \int_{\mathbf{R}} \int_{\mathbf{C}^{\mathrm{T}}} w_t(\mathbf{r}, \mathbf{c}^{\mathrm{T}}) \mathcal{L}(y, f(\mathbf{r})) dP(y, \mathbf{c}^{\mathrm{T}}, \mathbf{r} | T = t), \tag{3.9}$$

where $\mathcal{L}(y, f(\mathbf{r}))$ is any function of y and $\mathbf{r}$ such that its conditional expectation $E[y \mid f(\mathbf{r})]$ is the minimizer – which in practice we may think of as a "canonical" loss function such as squared-error loss for a continuous y or logistic loss for a binary y – and $\mathcal{F}$ is a function class in which the rule is known to lie (one possibility could be $\ell_1(P)$, the space of all absolutely integrable functions over P).

A natural weight function $w_t(\mathbf{r}, l)$ that can be used turns out to be

$$w_t(\mathbf{r}, \mathbf{c}^{\mathrm{T}}) = \frac{P(T = t | \mathbf{R} = \mathbf{r})}{P(T = t | \mathbf{R} = \mathbf{r}, \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}})}. \tag{3.10}$$

It is notable that the observation weight in (3.10) differs from the standard IPW weight, which would either be $\frac{1}{P(T=t|\mathbf{C}^{\mathrm{T}}=\mathbf{c}^{\mathrm{T}})}$ or $\frac{1}{P(T=t|\mathbf{R}=\mathbf{r},\mathbf{C}^{\mathrm{T}}=\mathbf{c}^{\mathrm{T}})}$ depending on whether or not the researcher wants to include $\mathbf{R}$ as predictors of treatment assignment in addition to $\mathbf{C}^{\mathrm{T}}$. The observation weight (3.10) may be thought of as stabilizing the standard IPW weights. In fact, the weight function in (3.10) is very closely related to the "stabilized weights" proposed by Robins et al. (2000) and adopted by Xu et al. (2010) among others, which in this case would be $P(T = t)/P(T = t|\mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}})$. That is, the weight function implied by Robins et al. (2000) differs from (3.10) by the latter's extra conditioning on $\mathbf{R} = \mathbf{r}$, the subset of individual characteristics that are potentially informative inputs to the treatment rule.

The derivation in (3.9) implies a natural estimate of f_t that is not complicated by the dimensionality or continuous/categorical nature of $\mathbf{R}$: the minimizer of the weighted sample

average over individuals in the $T = t$ group

$$\tilde{f}_t \equiv \text{argmin}_{f \in \mathcal{F}} \frac{1}{n_t} \sum_{T_i = t} \left(\frac{\tilde{P}(T = t \mid \mathbf{r}_i)}{\tilde{P}(T = t \mid \mathbf{r}_i, \mathbf{c}^{\mathrm{T}}{}_i)} \right) \mathcal{L}(y_i, f(\mathbf{r}_i)), \tag{3.11}$$

where n_t is the number of individuals in the group $T = t$, $\tilde{P}(T = t \mid \mathbf{r}_i)$ is an estimate of $P(T = t \mid \mathbf{R} = \mathbf{r})$, and $\tilde{P}(T = t | \mathbf{r}_i, \mathbf{c}^{\mathrm{T}}{}_i)$ is an estimate of $P(T = t \mid \mathbf{R} = \mathbf{r}_i, \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i)$. We note that the estimate $\tilde{f}_t$ is only reasonable if $\mathcal{F}$ is a suitably constrained class (e.g. a class with smoothness constraints, like a Sobolev class or class with bounded total variation (van de Geer, 2000)).

However, the formula (3.11) suggests another approach for estimation of f_t: Instead of necessarily solving a formal empirical minimization as in (3.11), one might use *any* predictive modeling method that accommodates observation weights (e.g. generalized linear models, lasso, ridge regression, boosted trees, neural nets, and many others). All of these methods can be written as, either exactly or approximately, minimizing a weighted least-squares-like loss over a, potentially complicated, function class. Many of the methods indicated in Table 3.2 as being flexible in fact only support penalized or non-penalized weighted regression; our framework supports much more flexibility by moving beyond regression-based methods.

Now that we can estimate

$$\tilde{\psi}(\mathbf{r}) = \tilde{f}_1(\mathbf{r}) - \tilde{f}_0(\mathbf{r}), \tag{3.12}$$

we can simply form the treatment rule as

$$\tilde{B}(\mathbf{r}) \equiv I \left[\tilde{f}_1(\mathbf{r}) - \tilde{f}_0(\mathbf{r}) > 0 \right], \tag{3.13}$$

which recommends treatment to an individual with characteristics $\mathbf{R} = \mathbf{r}$ if the estimated months until relapse is higher under $T = 1$ than $T = 0$. We note that, although we needed to observe all of the variables $(Y, T, \mathbf{C}^{\mathrm{T}}, \mathbf{R})$ to estimate $\tilde{f}_1$ and $\tilde{f}_0$, we only need to observe the characteristics $\mathbf{R}$ to actually *apply* the rule in a future dataset.

3.4.2 The Recipe

The work in Sections 3.4.1 yields the following procedure applied to a development dataset (D1), which must be independent of the evaluation dataset (D2). In what follows, we will use the superscripts D1 or D2 to emphasize the dataset on which the accompanying estimate is formed.

1. Use the scientific knowledge underlying D1 to partition observed individual characteristics (aside from outcome and treatment) into the four categories presented in Table 3.1: C^{TI}, C^{TN}, C^{NI}, and C^{NN}. Also form $\mathbf{C}^{\mathrm{T}} = (C^{\mathrm{TI}}, C^{\mathrm{TN}})$, $\mathbf{C}^{\mathrm{I}} = (C^{\mathrm{TI}}, C^{\mathrm{NI}})$, and form the potential inputs for the treatment rule $\mathbf{R} \subseteq \mathbf{C}^{\mathrm{I}}$.

2. For observations $i = 1, \ldots, n$ on D1:

   (a) Choose a prediction method and estimate the propensity scores $\tilde{P}^{\mathrm{D1}}(T = 1 \mid \mathbf{R} = \mathbf{r}_i)$ and $\tilde{P}^{\mathrm{D1}}(T = 1 \mid \mathbf{R} = \mathbf{r}_i, \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i)$.[5]

   (b) Compute the weights $\tilde{W}_t(\mathbf{R} = \mathbf{r}_i, \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i) = \frac{\tilde{P}^{\mathrm{D1}}(T=t|\mathbf{R}=\mathbf{r}_i)}{\tilde{P}^{\mathrm{D1}}(T=t|\mathbf{R}=\mathbf{r}_i, \mathbf{C}^{\mathrm{T}}=\mathbf{c}^{\mathrm{T}}{}_i)}$, for $t = 0, 1$.

   (c) Choose a prediction method that accommodates observation weights (e.g. generalized linear regression, lasso, boosted trees, and many others) and estimate $\tilde{f}_0$ and $\tilde{f}_1$ as suggested by (3.11). For example, weighted linear regression with a continuous response would yield, for observations $i = 1, \ldots, n$ on D1,

$$\tilde{\beta}_0^{\mathrm{D1}} \equiv \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n_0} \sum_{T_i=0} \tilde{W}_0(\mathbf{r}_i, \mathbf{c}^{\mathrm{T}}{}_i)(y_i - \mathbf{r}_i^{\top}\beta)^2, \tag{3.14}$$

$$\tilde{\beta}_1^{\mathrm{D1}} \equiv \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n_1} \sum_{T_i=1} \tilde{W}_1(\mathbf{r}_i, \mathbf{c}^{\mathrm{T}}{}_i)(y_i - \mathbf{r}_i^{\top}\beta)^2, \tag{3.15}$$

[5] In applied work it can be useful to truncate estimated propensity scores so they are not too close to 0 or 1 (which can lead to very large observation weights); a default setting in our software implementation truncates estimated propensity scores to stay between 0.05 and 0.95, but this choice can be overwritten by the user.

where $\tilde{W}_t(\mathbf{R} = \mathbf{r}_i, \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i) = \frac{\tilde{P}^{\mathrm{D1}}(T=t|\mathbf{R}=\mathbf{r}_i)}{\tilde{P}^{\mathrm{D1}}(T=t|\mathbf{R}=\mathbf{r}_i, \mathbf{C}^{\mathrm{T}}=\mathbf{c}^{\mathrm{T}}{}_i)}$ for $t = 0, 1$ and where we define $\tilde{f}_0^{\mathrm{D1}}(\mathbf{r}) \equiv \mathbf{r}^\top \tilde{\beta}_0^{\mathrm{D1}}$ and $\tilde{f}_1^{\mathrm{D1}}(\mathbf{r}) \equiv \mathbf{r}^\top \tilde{\beta}_1^{\mathrm{D1}}$.

3. Form the treatment rule

$$\tilde{B}(\mathbf{r}) \equiv I\left[\tilde{f}_1^{\mathrm{D1}}(\mathbf{r}) - \tilde{f}_0^{\mathrm{D1}}(\mathbf{r}) > 0\right], \tag{3.16}$$

where $I(\cdot)$ is the indicator function.

4. Use scientific knowledge underlying D2 to select the potential confounders $\mathbf{C}^{\mathrm{T,\ eval}}$. [6]

5. For observations $j = 1, \ldots, m$ on D2:

   (a) Assign the recommended treatment with

$$\tilde{B}_j^{\mathrm{D2}} \equiv \tilde{B}^{\mathrm{D2}}(\mathbf{r}_j). \tag{3.17}$$

   (b) As discussed next in Section 3.4.3, form the IPW-based estimators of the ATE in the test-positives and in the test-negatives using (3.26) and (3.27), respectively.

3.4.3 Evaluating the Rule

Recall from Section 3.4.1 that a foundational target parameter in our framework is the average treatment effect (ATE) in a subpopulation of individuals with characteristics $\mathbf{R} = \mathbf{r}$.

$$E[Y^1 - Y^0 \mid \mathbf{R} = \mathbf{r}], \tag{3.18}$$

[6] Ideally, D1 and D2 would be datasets from separate observational studies so that D2 independently samples from the population where future intervention would take place, but D1/D2 may also be a random partition of data from a single observational study; in the latter case, we will have $\mathbf{C}^{\mathrm{T,\ eval}} = \mathbf{C}^{\mathrm{T}}$. The `DevTreatRules` package supports developing and evaluating rules in either situation.

which under the assumptions of consistency, no unmeasured confounding, and positivity, as detailed in Kennedy (2015) can be re-written as

$$\psi(\mathbf{r}) \equiv \int_{\mathbf{C}^{\mathrm{T}}} \left\{ E\left[Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R} = \mathbf{r}, T = 1\right] - E\left[Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R} = \mathbf{r}, T = 0\right] \right\} dP(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R} = \mathbf{r}) \,, \quad (3.19)$$

which implies we should recommend treatment to individuals in the subgroup

$$\Omega^{+} = \{\mathbf{r} \mid \psi(\mathbf{r}) > 0\} \,, \quad (3.20)$$

which is known as the *test-positives* group. If $\mathbf{R}$ were a set of gene expression levels, for example, then Ω^{+} would be the subset of gene expression levels for which treatment increases the expected number of months until relapse. Again, the expected improvement in months until relapse among this treated subpopulation is

$$\int_{\mathbf{r} \in \Omega^{+}} \psi(\mathbf{r}) dP(\mathbf{r}). \quad (3.21)$$

Similarly, treatment should not be recommended to individuals in the *test-negatives* subpopulation defined by

$$\Omega^{-} = \{\mathbf{r} \mid \psi(\mathbf{r}) \leq 0\} \,, \quad (3.22)$$

and

$$\int_{\mathbf{r} \in \Omega^{-}} \psi(\mathbf{r}) dP(\mathbf{r}). \quad (3.23)$$

would yield the average increase in months until relapse for the subpopulation that avoids treatment.

To obtain a trustworthy estimate of how a developed treatment rule will perform for individuals seen in future clinical settings, it is absolutely essential that development of the rule (using the tools of Section 3.4.1) and evaluation of the selected rule are performed independently. Chapter 7 (Model Assessment and Selection) of Hastie et al. (2008) provides an excellent discussion of this topic with regard to the predictive performance of model-based

estimates. In brief, if the development and evaluation datasets coincide then our evaluation of the treatment rule's benefit will be overly optimistic because it will reward the rule for incorrectly classifying noise in the development dataset as signal that would persist when we apply the rule to future individuals.

We can take the treatment rule $\tilde{B}(\mathbf{r})$ as defined in (3.13) – which gives us a mapping from an individual's particular characteristics $\mathbf{R} = \mathbf{r}$ to a treatment recommendation – that was estimated on a *development* dataset and apply it to the independent *evaluation dataset* to yield, for the n_{eval} individuals in the evaluation dataset indexed by $i = 1, \ldots, n_{\text{eval}}$,

$$\tilde{\Omega}^{+} = \left\{ \mathbf{r}_i \mid \tilde{B}(\mathbf{r}_i) = 1 \right\}, \tag{3.24}$$

the test-positives subset of observations in the evaluation dataset, based on $\mathbf{R}$, who are expected to have more months until relapse under treatment than under standard-of-care. Similarly,

$$\tilde{\Omega}^{-} = \left\{ \mathbf{r}_i \mid \tilde{B}(\mathbf{r}_i) = 0 \right\} \tag{3.25}$$

yields the test-negatives subset of observations in the evaluation dataset, based on $\mathbf{R}$, who are expected to have fewer months until relapse under treatment than under standard-of-care.

Now we define $\mathbf{C}^{\mathrm{T,\ eval}}$ as the set of potential confounders of the association between treatment and response *in the evaluation dataset*. We note that if the development and evaluation datasets are partitions of the same observational dataset then the variables in $\mathbf{C}^{\mathrm{T,\ eval}}$ will be identical to those in $\mathbf{C}^{\mathrm{T}}$, but in cases where development/evaluation are carried out using data from separate studies then it need not be the case that $\mathbf{C}^{\mathrm{T}}$ is equivalent to $\mathbf{C}^{\mathrm{T,\ eval}}$. For example, in the motivating example from Section 3.3, we had $\mathbf{C}^{\mathrm{T}}$ =(age, ECOG) but $\mathbf{C}^{\mathrm{T,\ eval}}$ = (age, KPS).

To evaluate the the developed rule $\tilde{B}$, we can use an estimate of the ATE in the test-

positives population with the IPW-based estimator (see e.g. Austin (2011))

$$\widehat{\mathrm{ATE}}^{+} \equiv \frac{1}{N^+} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=1\}} \frac{t_j y_j}{\tilde{P}(T=1|\mathbf{C}^{\mathrm{T,\ eval}}=c^{\mathrm{T,\ eval}})} - \frac{1}{N^+} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=1\}} \frac{(1-t_j) y_j}{\tilde{P}(T=0|\mathbf{C}^{\mathrm{T,\ eval}}=c^{\mathrm{T,\ eval}})}, \quad (3.26)$$

where N^+ is the number of test-positives.

For some intuition explaining the form of (3.26), we note that if the evaluation dataset comes from a randomized controlled trial – where $\tilde{P}(T = t \mid \mathbf{C}^{\mathrm{T,\ eval}} = c^{\mathrm{T,\ eval}})$ is just the sample proportion in each treatment group (n_t^+/N^+) – then (3.26) becomes

$$\begin{aligned} &\frac{1}{N^+} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=1\}} \frac{t_j y_j}{n_1^+/N^+} \quad - \quad \frac{1}{N^+} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=1\}} \frac{(1-t_j) y_j}{n_0^+/N^+} \\ &= \frac{1}{n_1^+} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=1\}} t_j y_j \quad - \quad \frac{1}{n_0^+} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=1\}} (1-t_j) y_j, \end{aligned}$$

which is just the difference between the sample mean for test-positives who received treatment and the sample mean for test-positives who did not receive treatment. This is the usual "difference in means" we would think of when comparing mean outcomes across treatment groups in a clinical trial where there is no systematic confounding.

A small modification to (3.26) also estimates the effect of avoiding treatment among the test-negatives:

$$\widehat{\mathrm{ATE}}^{-} \equiv \frac{1}{N^-} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=0\}} \frac{t_j y_j}{\tilde{P}(T=1|\mathbf{C}^{\mathrm{T,\ eval}}=c^{\mathrm{T,\ eval}})} - \frac{1}{N^-} \sum_{\{j \mid \tilde{B}(\mathbf{r}_j)=0\}} \frac{(1-t_j) y_j}{\tilde{P}(T=0|\mathbf{C}^{\mathrm{T,\ eval}}=c^{\mathrm{T,\ eval}})}, \quad (3.27)$$

where N^- is the number of test-negatives. We note that we would expect (3.27) to be a negative number for a rule that accurately identifies the test-negatives.

We can also form a simple estimator of the average benefit of the rule (ABR) in the population from which our the evaluation dataset is a representative sample with

$$\left(\frac{N^+}{N^+ + N^-}\right) \widehat{\mathrm{ATE}}^{+} + \left(\frac{N^-}{N^+ + N^-}\right) \left(-\widehat{\mathrm{ATE}}^{-}\right),$$

a weighted average of the benefit of receiving treatment the test-positives from (3.26) and the benefit of *avoiding* treatment in the test-negatives from (3.27), respectively.

Alternatively one could estimate ATE with the estimator developed by Robins et al. (1994) that is sometimes called the *doubly-robust* or *augmented* analog of (3.26); Lunceford and Davidian (2004) present an excellent derivation and simulation study comparing different estimators of the ATE. We chose to present only the IPW-based estimators for ease of exposition and to keep emphasis on the methodology that estimates a treatment rule along with subsequent test-positive and test-negative subpopulations, since the evaluation of treatment effect in pre-defined subpopulations has been more extensively studied elsewhere. In practice, one might prefer the doubly-robust estimator to ensure asymptotically correct inference if a method other than regression with generalized linear models is used for the propensity model (e.g. an additive model or a penalized regression model), as discussed in Kennedy (2015).

3.4.4 R Implementation

The `R` package `DevTreatRules` implements the split-regression approach. In particular, the functions `SplitData()`, `BuildRule()`, and `EvaluateRule()` handle, respectively: the development/evaluation partitioning of a dataset (or development/validation/evaluation partitioning if model selection is also performed); the development of the treatment rule (in dataset D1) as in steps 2-4 of Section 3.4.2; and the evaluation of the rule (in dataset D2) as in steps 5-6 of Section 3.4.2. The vignette accompanying `DevTreatRules` walks through an example of building and evaluating a treatment rule using the package, in a situation where model selection is also performed using the `CompareRulesOnValidation()` function.

3.5 Simulations

We construct a simulation scenario where we generate the following data for each observation: a desirable binary outcome Y (e.g. no relapse after 5 years); a binary treatment indicator T (e.g. indicating an additional treatment regimen on top of standard-of-care) that is influenced

only by L and statistical noise; L (e.g. indicator of ECOG score above 2) confounds the association between T and Y; a variable X (e.g. expression level of a gene whose biological role is unknown) that does not influence treatment assignment but is nonetheless predictive of treatment benefit; and another variable G that does affects neither treatment assignment nor treatment benefit, but was nonetheless a candidate for inclusion in the treatment rule (e.g. expression level of another gene whose biological role is also unknown). Formally, we simulate data for a development set of size n with

$$
\begin{aligned}
G_i &\sim \text{Normal}(0, 1), \\
X_i &\sim \text{Uniform}(0, 2), \\
L_i &\sim \text{Bernoulli}(0.5), \\
T_i \mid L_i &\sim \begin{cases} \text{Bernoulli}(0.75), & L_i = 0 \\ \text{Bernoulli}(0.25), & L_i = 1 \end{cases} \\
P(Y_i = 1 \mid X_i, L_i, T_i) &= \begin{cases} \text{expit}\left[\beta_{0,T=0} + \beta_{1,T=0} X_i + \gamma_{T=0} L_i\right], & T_i = 0 \\ \text{expit}\left[\beta_{0,T=1} + \beta_{1,T=1} X_i + \gamma_{T=1} L_i\right], & T_i = 1, \end{cases}
\end{aligned}
$$

for $i = 1, \ldots, n$, where $\beta_{0,T=t}, \beta_{1,T=t}, \gamma_{T=t} \in \mathbb{R}$ for $t = 0, 1$. Using the notation from Section 3.1 we categorize $\mathbf{C}^{\text{T}} = \mathbf{C}^{\text{T, eval}} = L$ and $\mathbf{C}^{\text{I}} = \mathbf{R} = (X, G)$.

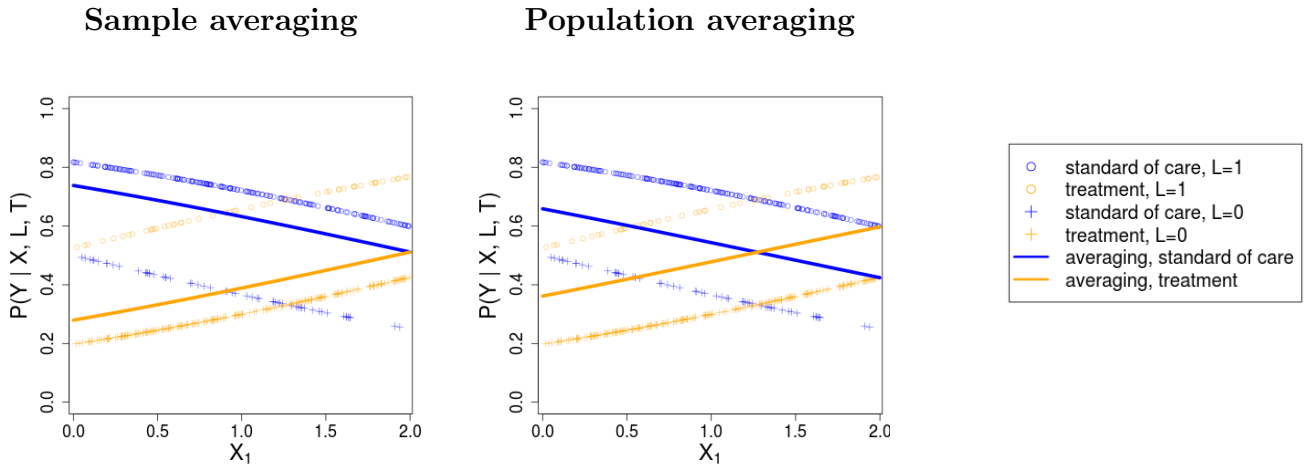


Figure 3.2: Simulation scenario with $n = 500$, $(\beta_{0,T=0}, \beta_{1,T=0}, \gamma_{T=0}) = (0, -0.55, 1.5)$, and $(\beta_{0,T=1}, \beta_{1,T=1}, \gamma_{T=1}) = (-1.4, 0.55, 1.5)$

Figure 3.2 depicts the relationship between $P(Y \mid X, L, T)$ and X, where the true response probability for the $L = 1$ group is shown with circles and for the $L = 0$ group with triangles. The standard-of-care group is shown in blue and the treatment group is shown in orange. As seen in Figure 3.2, the outcome (Y) is more likely under treatment than under standard-of-care for individuals with a value of X about 1.3 in both the $L = 1$ and $L = 0$ groups. Thus, the optimal treatment rule would recommend treatment to an individuals with $X > 1.3$. However, due to the confounding effect of L, an empirical average of the response-curves for each treatment group (solid lines in left panel of Figure 3.2) incorrectly suggests that there is no subset of individuals for whom treatment makes the outcome more likely. On the other hand, the IPW approach (solid lines in right panel of Figure 3.2) uses a weighted average of response-curves, where each weight is the inverse probability of receiving the observed treatment given the observed value of L, to correctly identify the value of X (about 1.3) where the response-curves cross.

In Table 3.3, we present the mean probability of the (desirable) outcome for a range of development set sample sizes, specifying logistic regression (with $\mathbf{R}$ as the predictors) for steps 2 in Section 3.4.2. The first row shows the mean outcome probability for rules built

using the split-regression approach described in Section 3.4.2. The second row reports the mean outcome probability for a modification to the split-regression approach that "naively" uses the incorrect sample averaging shown in the left panel of Figure 3.2 (i.e. it uses identical observation weights in step 2 of Section 3.4.2).

	Sample size in development set				
Type of Rule	50	100	200	500	1000
Split-regression	0.543	0.553	0.562	0.57	0.572
Split-regression (naive, no weights)	0.552	0.554	0.553	0.55	0.549
Optimal rule	**0.574**	**0.574**	**0.574**	**0.574**	**0.574**
Treating all	0.479	0.479	0.479	0.479	0.479
Treating none	0.543	0.543	0.543	0.543	0.543

Table 3.3: Mean outcome probability, as a function of rule type and development set sample size, averaged over 1000 replications and calculated in evaluation sets of size 10000

The estimated outcome probabilities in the first three rows of Table 3.3 can be compared to the benchmark value of 0.574 for the optimal rule (known from the data-generating mechanism) that perfectly assigns treatment to only those who benefit and withholds it from those who do not benefit. We also compare to a rule that recommends everyone receive treatment (which may be the prevailing policy when an available treatment is believed to be uniformly effective) and to a rule that recommends no one receive treatment (which might be the preferred strategy when the effectiveness of a proposed treatment has not yet been established).

Table 3.3 shows the advantage of using split-regression with IPW weights relative to the "naive" uniform weighting: with a large enough sample size the naive approach performs nearly as poorly as the "treat none" strategy that withholds treatment from all individuals in the population even though the roughly 35% of observations with $X > 1.3$ actually would benefit from treatment. The bias of the naive approach manifests itself by forcing the approach to estimate the incorrect non-crossing response-curves shown with the solid lines

in the left panel of Figure 3.2. In contrast, split-regression with the correct IPW weighting approaches the optimal treatment rule with large enough sample size because it is estimating the crossing response-curves in the right panel of Figure 3.2.

3.6 Data Example: WHI-OS

We also illustrate the split-regression approach and compare it to alternatives by applying the `R` package `DevTreatRules` to the Women's Health Initiative Observational Study component (WHI-OS). A detailed description of the study design and summaries of baseline measurements for participants (postmenopausal women between the ages of 49 and 81 who were recruited for the WHI clinical trial but either declined to participate or were later deemed ineligible) are available in Langer et al. (2003). We also present summary tables of the variables we retained for analysis in the appendix. The GitHub page `github.com/jhroth/data-example-split-regression` contains the `R` code needed to go start-to-finish from loading the raw WHI-OS datasets (access to which requires additional permission) to replicating our estimates in the evaluation subset.

Briefly, we would like to build a treatment rule to assign baseline hormone therapy (HRT) – defined as *currently using* unopposed estrogen and/or estrogen plus progesterone at baseline – to postmenopausal women if it will increase a woman's probability of remaining free of coronary heart disease (CHD) after 10 years and, in a separate analysis, if it will increase a woman's probability of remaining free of breast cancer after 10 years. All variables included in our analysis besides the outcomes were measured at baseline. We used the "adjudicated" outcome variables in the WHI-OS as described in Curb et al. (2003).

We classified 4 categorical variables as belonging to C^{TN} : education level, ethnicity, family income, and how each participant heard about the study. We identified 31 self-reported variables to make up C^{TI} and we chose $\mathbf{R} = C^{\text{TI}}$, so there are 31 candidates to be used as inputs in our treatment rule. We did not specify any variables as belonging to C^{NI} . In the Appendix, we present the complete list of these variables and simple summaries including counts of missing values. We intend for this to be primarily an illustrative example

rather than a definitive claim about appropriate variable classifications in this study.

About 17.6% of the 94140 observations in the initial dataset had a missing value of at least one variable in $\mathbf{C}^{\mathrm{T}}$ or $\mathbf{R}$. Instead of conducting a complete-case analysis that would drop observations with missing values, we used an IPW-based adjustment for missingness (Seaman and White, 2013). Our shared code at `github.com/jhroth/data-example-split-regression` shows the details of our adjustment for missingness using the `additional.weights` argument of the `BuildRule()` and `EvaluateRule()` functions in `DevTreatRules`.

	Positives	**Negatives**	**ATE in Positives**	**ATE in Negatives**	**ABR**
Outcome: No breast cancer after 10 years					
Split regression (ridge/lasso)	3544	15758	0.013	-0.051	0.044
Split regression (logistic/logistic)	3692	15610	0.021	-0.053	0.047
Treat no one (logistic/NA)	0	19302	NA	-0.045	0.045

Table 3.4: Summary of Selected Rules in the Validation Set. The selected propensity method/rule method are in parentheses

Table 3.4 presents estimated ATEs and ABR in the validation set for two split-regression specifications using the outcome of no breast cancer after 10 years: one that used ridge regression for the propensity score and lasso for the rule, and another specification that used logistic regression for both the propensity score and rule models. On the validation set, none of the split-regression specifications for the outcome of no CHD after 10 years appeared to be an improvement over the naive strategy of treating no one and thus none of those specifications were chosen. We again emphasize the importance of evaluating a treatment rule on an independent dataset that did not inform any stage of rule development; since the estimated ATEs and ABR in Table 3.4 informed our model selection – in particular, we chose these models because their ABRs in the validation set were relatively high – they do not serve as trustworthy estimates of ATE and ABR in independent samples drawn from this population in the future, which instead must be computed on the evaluation set.

	Positives	Negatives	ATE in Positives	ATE in Negatives	ABR
Outcome: No breast cancer after 10 years					
Split-regression (ridge/lasso)	3408	15894	0.015 (-0.027, 0.068)	-0.045 (-0.067, -0.031)	0.04
Split-regression (logistic/logistic)	3569	15733	0.002 (-0.049, 0.052)	-0.048 (-0.07, -0.035)	0.04
Treat no one (logistic/NA)	0	19302	NA	-0.05	0.05

Table 3.5: Summary of Selected Rules in the Evaluation Set. The selected propensity method/rule method are in parentheses. 95% CIs are based on the basic bootstrap.

Table 3.5 presents the estimated ATEs and ABR in the evaluation set, which do serve as trustworthy estimates of rule performance in future clinical settings that observe individuals from this same population. We see that the rule using ridge/lasso as the propensity/rule models would recommend HRT to about 18% of individuals and, among this treated subpopulation, we estimate that treatment decreases the probability of breast cancer within 10 years by 1.5 percentage points. We also estimate that the same split-regression rule would not assign treatment to the remaining 82% of individuals and, among this non-treated subpopulation, avoiding HRT decreases the probability of 10-year breast cancer by 4.5 percentage points. Also from Table 3.5, we see that the split-regression rule based on using logistic regression for both the propensity and rule methods would recommend HRT to a larger share of individuals with a smaller estimate of treatment benefit in that treated subgroup, so our recommended rule would be the one based on ridge/lasso. Unfortunately, the 95% CI for estimated ATE among the treated population contains 0 for both rules in the evaluation set; as a result, we do not find evidence that either rule has identified a subpopulation of individuals who appear to benefit from treatment.

3.7 Discussion

We outlined a principled approach to classify the roles of variables collected in an observational study into clinically meaningful categories and, using that knowledge, to develop a treatment rule along with a trustworthy estimate of the rule's population benefit. Since this paper is intended to be a practical guide to help practitioners go from start to finish in

estimating and evaluating treatment rules without getting bogged down by the onerous and error-prone tasks of coding the method from scratch in statistical software, we implemented our approach in the R package `DevTreatRules` and shared the code needed to reproduce our simulations and data example on GitHub.

In a simple simulation study, we saw the benefit of estimating a treatment rule using this Chapter's preferred split-regression approach with IPW weighting compared to using uniform observational weights that ignore the observational study design. In the WHI-OS data example, we used the split-regression approach to develop a treatment rule that assigns baseline hormone therapy to postmenopausal women if it is expected to increase their probability of remaining free of coronary heart disease after 10 years or free of breast cancer after 10 years. With the 10-year CHD outcome, split-regression did not estimate a rule with a positive estimate of ATE in the treated subgroup on the validation set. With the 10-year breast cancer outcome, however, split-regression did identify a rule that recommends HRT to about 18% of women and, among this treated subpopulation, had an estimated 1.5 percentage-point decrease in the probability of breast cancer. However, this 1.5 percentage-point decrease lacks statistical significance (95% CI: $-0.027, 0.068$) and, as a result, we would simply recommend not assigning HRT to any women in this population if the goal is to reduce 10-year breast cancer incidence or 10-year CHD incidence. This null finding is, unfortunately, often a typical result in the search for informative treatment rules with observational data.

One notable limitation of this work is that, while the algorithm outlined in Section 3.4.2 is fairly general, the accompanying `R` implementation does not have as much flexibility (e.g. the currently available estimation methods are linear/logistic regression and its lasso/ridge counterparts); in future work we hope to expand the package to support more estimation methods. Another limitation is that our data example is not publicly reproducible because we are unable to share the underlying WHI-OS dataset; however, we do share the code that will reproduce the data example for users who have access to the raw WHI-OS data files.

3.8 Acknowledgements

Thank you to Holly Janes and Susanne May for helpful comments and suggestions.

Chapter 4

ELUCIDATING OUTCOME-WEIGHTED LEARNING AND ITS COMPARISON TO SPLIT-REGRESSION: DIRECT VS. INDIRECT METHODS IN PRACTICE

4.1 Introduction

In Chapter 3, we mentioned the "indirect vs. direct" dichotimization of methods for estimating treatment rules. Indirect approaches are predicated on accurately modeling expected clinical outcome as a function of patient characteristics and treatment assignment, either separately in each treatment group or jointly by including interaction terms (e.g. Kang et al. (2014); Cai et al. (2011); Lu et al. (2013); McKeague and Qian (2014); Ciarleglio et al. (2015); and split-regression in Chapter 3). On the other hand, direct approaches are motivated by seeking a treatment rule that optimizes average clinical outcome in the population (e.g. Qian and Murphy (2011); Zhang et al. (2012a,c, 2015); Chen et al. (2017), among others)

In some sense the justification for an indirect approach is regarded as self-evident: If we can accurately predict clinical outcome as a function of patient characteristics under each possible treatment option, then shouldn't the treatment rule sensibly derived from these predictions – recommending treatment if the predicted outcome under treatment is more desirable than the predicted outcome under standard-of-care – also be accurate? In contrast, the direct method of outcome-weighted learning (OWL) proposed by Zhao et al. (2012) has a very concrete justification for yielding an accurate treatment rule: OWL seeks the treatment rule that maximizes expected clinical outcome in the population. As direct methods in general go after a rule that maximizes an explicit estimate of population benefit, they may be less sensitive to regression model mis-specification than indirect methods (Zhang et al., 2012c), which would be an appealing trait in practice where models are always mis-specified.

We believe the distinction between split-regression (which predicts outcome as a function of patient characteristics under each treatment option) and OWL (which predicts *treatment assignment* as a function of patient characteristics, weighted by outcome) is representative of the general gulf between indirect and direct methods in the literature. We feel indirect methods tend to have an intuitive estimation procedure but a somewhat hazy justification for their use in developing treatment rules, while direct methods tend to have a very convincing justification for use as treatment rules but often imply hazier estimation procedures.

We aim to bridge some of the gap separating the split-regression and OWL methods (as well as indirect and direct methods more generally) in this chapter. Specifically, this chapter makes contributions to the literature that include the following:

- **Interpretation.** We clarify the decision-making underlying OWL by connecting it to the more traditional decision-making of split-regression. With a binary outcome, we offer a straightforward Bayesian interpretation of the rule targeted by OWL through a simple application of Bayes' rule and a set of figures. Interpretation appears more complicated with a continuous outcome, but we frame the continuous setting as an extension of the binary case (with an analogous Bayesian interpretation) where the OWL procedure offers a very convenient computational shortcut for bypassing a cumbersome and variable averaging procedure. For both types of outcomes, there is a connection between OWL and the familiar split-regression procedure.

- **Computation.** We extend the `R` package `DevTreatRules` presented in Chapter 3 to accommodate OWL with greater flexibility and utility than is currently available by leveraging the official and targeted optimization in the `DynTxRegime` implementation of OWL but additionally:

    1. Using a simple computational idea that expands available propensity score estimation in the existing `DynTxRegime` implementation of OWL.
    2. Implementing the broader "OWL framework" also established by Zhao et al.

(2012).

3. Folding the OWL and OWL framework approaches into Chapter 3's principled development and evaluation pipeline and implementing both OWL and OWL framework in `DevTreatRules`.

4. Integrating another promising direct approach for estimating treatment rules proposed by Chen et al. (2017) and Tian et al. (2014), which we refer to as *direct-interactions*, into the principled framework and implementing it in `DevTreatRules`. The direct-interactions approach serves as another option for practitioners who wish to develop a treatment rule and evaluate its benefit on the data they have at hand.

- **Application.** We conduct a simulation study comparing treatment rules developed by the OWL framework (a direct approach) to split-regression from Chapter 3 (an indirect approach) and to direct-interactions (another direct approach) using `DevTreatRules`. We also develop treatment rules using the Women's Health Initiative Observational Study, where it turns out the official implementation of OWL fails due to memory constraints while the OWL framework runs without issues. We share the code needed to reproduce our results using the `DevTreatRules` package at `github.com/jhroth/data-example-elucidating-owl` and `github.com/jhroth/simulations-elucidating-owl`. To the best of our knowledge, `DevTreatRules` is the only available package that offers users the ability to go from start to finish in estimating a treatment rule using both indirect and direct approaches (with the principled handling of variable types as discussed in Chapter 3) and evaluating the benefit of using those rules on an independent dataset.

This chapter is organized as follows. Section 4.2 outlines the OWL procedure, its interpretation as establishing the flexible "OWL framework" class of estimation strategies, and clarifies our focus. Section 4.3 connects the decision-making underlying OWL to the

decision-making underlying split-regression. Sections 4.4 and 4.5 describe how OWL framework and direct-interactions, respectively, fit into the principled development and evaluation framework outlined in Chapter 3. Section 4.6 presents a simulation study comparing OWL framework and direct-interactions to the split-regression approach from Chapter 3, relative to the optimal rule and to naive strategies that assign treatment to everyone or to no one. Section 4.7 outlines the software implementation and describes some subtle but critical issues we encountered during our simulation study that have a huge impact on the performance of OWL and direct-interactions. In Section 4.8, we analyze the Women's Health Initiative Observational Study by using all three methods on a development/validation/evaluation partition of the dataset. Section 4.9 wraps up by discussing lessons we might carry forward and sharing thoughts about future research.

4.2 OWL and Related Work

4.2.1 The OWL Approach

We begin by formally defining the OWL approach developed by Zhao et al. (2012), with notation modified slightly to be consistent with Chapter 3 (e.g. we assume treatment is coded as 0/1 instead of -1/1). Let Y be an outcome where larger values are more desirable, $T \in \{0, 1\}$ be a binary treatment indicator, and $X = (\mathbf{X_1}, \ldots, \mathbf{X_p})$ be a set of patient characteristics that we expect to observe in future clinical settings. Suppose we observe (Y_i, T_i, X_i) for individuals $i = 1, \ldots, n$ in an observational study. Zhao et al. (2012) motivate their OWL procedure by noting that the expected outcome under the treatment rule $\mathcal{D} : X \to \{0, 1\}$ can be written as[1]

$$\mathrm{E}_{Y,T,X}\left[\frac{Y}{T\pi + (1-T)(1-\pi)}\, I[T = \mathcal{D}(X)]\right], \tag{4.1}$$

[1]Further, if the RCT has equal randomization so that $\pi = 1/2$, the expected outcome under the rule in (4.1) would be the even simpler outcome-weighted $\mathrm{E}_{Y,T,X}\left\{2Y\, I[T = \mathcal{D}(X)]\right\}$.

where $I[\cdot]$ is the indicator function and π is defined as the known constant $P(T = 1)$ since the authors assume that data come from the randomized controlled trial (RCT) setting.

Maximizing (4.1) over the class of treatment rules (i.e. finding the treatment rule under which clinical outcome is maximized) is equivalent to finding

$$\arg\min_{\mathcal{D}} \ \mathrm{E}_{Y,T,X}\left[\underbrace{\frac{Y}{T\pi + (1-T)(1-\pi)}}_{\text{“weight”}} \ \underbrace{I[T \neq \mathcal{D}(X)]}_{\text{“misclassification”}} \right]. \tag{4.2}$$

We can see from (4.2) that finding the treatment rule under which expected outcomes are maximized boils down to accurately predicting *treatment assignment*, not clinical outcome, as a function of patient characteristics X in a weighted classification framework, where the weights are a function of outcome and treatment probability. Zhao et al. (2012) mention that OWL encourages the estimation procedure to place greater emphasis on assigning individuals with larger values of the outcome variable to the same treatment they originally received (i.e. larger weights are placed on correctly predicting existing treatment assignment for individuals with larger values of Y).

To make (4.2) more tractable for estimation methods, Zhao et al. (2012) note that $\mathcal{D}(X)$ can be equivalently stated as $\mathrm{pos}(f(X)) \equiv I[f(X) > 0]$ so that optimizing (4.2) is the same as finding

$$\hat{f} \equiv \arg\min_{f \in \mathcal{F}} \ \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{T_i\pi + (1-T_i)(1-\pi)} \, I\left[T_i \neq \mathrm{pos}(f(X_i))\right], \tag{4.3}$$

for a given function class $\mathcal{F}$, and then forming the treatment rule $\widehat{D}(X) \equiv \mathrm{pos}(\hat{f}(X))$. As noted by Zhang et al. (2012b), the problem in (4.3) can be thought of as minimizing weighted classification error with the usual inverse probability of treatment (IPW) weight $\frac{1}{T\pi(X)+(1-T)(1-\pi(X))}$ multiplied by outcome, where $\pi = P(T = 1 \mid X)$. As a result, OWL is indeed appropriate for use in observational data settings where treatment assignment is not independent of patient characteristics.[2]

[2]As described in Austin (2011) and mentioned in Chapter 3, the IPW approach re-weights observations so clinically observed confounders are roughly balanced between the treatment groups, thus facilitating a

Thus, OWL presents a very promising and exciting result because it reduces the unstructured search for a treatment rule that optimizes population benefit to merely a search for the solution to a problem within the familiar estimation framework of weighted classification.

Two major types of modeling assumptions will come into play when approximating the solution to (4.3) using an actual dataset: 1) defining the functional class $\mathcal{F}$ that specifies the allowable structure of $f(X)$, for example through a parametric generalized linear model or through a class that imposes less restrictive smoothness constraints (van de Geer, 2000); and 2) replacing the indicator function with an approximation that turns (4.3) into a convex problem, which is far easier computationally. In their original implementation of the OWL method, Zhao et al. (2012) specify a weighted support vector machines (SVM) method for classification, which essentially replaces the indicator function $I[T \neq \text{pos}(f(X))]$ with the *hinge* loss function and adds a penalty term to tune the trade-off between structure in the estimated $f(X)$ and its proximity to the observed data points (Hastie et al., 2008). It is also recognized, however, that classification methods besides SVM that support observation weights (e.g. standard logistic regression and logistic lasso/ridge regression) are viable alternatives to approximate the solution to (4.3). For the remainder the paper, we write *OWL* to refer to the solution of (4.3) using weighted SVM and write *OWL framework* to refer to the solution of (4.3) using a different classification method that also accommodates observation weights; in both cases, the method is completely credited to Zhao et al. (2012).

4.2.2 Related Work

Past research has proposed other direct approaches for estimating treatment rules. In their review of methods to estimate treatment rules in RCTs, Lipkovich et al. (2017) identify sev-

fair comparison between average outcomes across treatment groups. As detailed in Chapter 3, some of the patient characteristics X may not be available in future clinical settings so it would be incorrect to use them as inputs to the treatment rule $\mathcal{D}(X)$, even though they would still be sensible inputs to the propensity score $\pi(X)$ if they are expected to influence treatment assignment in the current study. Similarly, there may be inputs to the treatment rule that we do not expect to influence treatment assignment in the current study. We clarify this distinction more in Section 4.4 where we show how to integrate OWL into the principled framework we proposed in Chapter 3, but do not make the distinction in this section to ease notation and facilitate comparison with OWL as proposed in Zhao et al. (2012).

eral direct approaches (including OWL) under the label "optimal treatment regimes"; the authors note that OWL is the only method in this category to share a software implementation.

Those past direct methods, along with others such as Zhang et al. (2015) and Zhang et al. (2012c), may differ substantially in the estimation problems they propose solving to find the optimal treatment rule within the class under which structure is assumed. For example, Zhang et al. (2015) enforce a tree-like structure in the treatment rule by requiring it to be a sequence of "if-then" statements, while the structure of the treatment rules permitted in Zhang et al. (2012c) instead comes from the regression model linking outcome to patient characteristics and using an augmented IPW approach as in Bang and Robins (2005). Nonetheless, past direct methods are united through their shared motivation of estimating a treatment rule that directly maximizes some estimate of population benefit.

The aforementioned Zhang et al. (2012b) outlines a general theoretical classification framework whose goal is to correctly classify an observation as either benefiting or not benefiting from treatment while accommodating different choices of observation weights. As part of their work, Zhang et al. (2012b) compute the observation weights such that their general weighted classification framework can reduce to the special case of OWL (if SVM were chosen as the classification method). Unfortunately, Zhang et al. (2012b) do not provide a software implementation for their framework that would facilitate its use by practitioners with the data they have at hand.

More recently, Chen et al. (2017) provide a promising direct approach to estimating treatment rules, based on the earlier work in Tian et al. (2014), that allows for substantial modeling flexibility in its estimation stage and, while not being implemented as an `R` package, does share code to reproduce its simulation results. We implement the method of Chen et al. (2017), which we refer to as *direct-interactions*, in `DevTreatRules` and discuss the method in more detail in Section 4.5 to show how it (like OWL/OWL framework) can be fit into the principled framework from Chapter 3 and can serve as a candidate in a side-by-side comparison with other methods when deciding which approach to use to develop a treatment

rule in practice.

4.2.3 Our Focus

This chapter focuses on OWL/OWL framework because it is a promising approach that we feel is representative of past direct approaches in the literature: Its motivation is quite convincing (try to find the treatment rule that maximizes expected clinical outcome in the target population), but the estimation problem it implies (predict *treatment assignment* rather than clinical outcome) is not as intuitive as the procedure from an indirect approach such as split-regression (predict *clinical outcome* under each treatment option and see which treatment yields the most desirable predicted outcome).

In addition, OWL is implemented in an `R` package (`DynTxRegime`) that has proven useful to practitioners since its introduction. As a result, we believe that providing an interpretation of OWL/OWL framework that connects to the familiar realm of split-regression will also be useful to practitioners by encouraging future use of the procedure without viewing it as a "black-box" method. Further, we believe that implementing OWL/OWL framework and split-regression in the same `DevTreatRules` package (which also appropriately handles data-splitting, model tuning, and the clinically meaningful distinction between patient characteristics described in Chapter 3) provides a user-friendly way to decide between competing treatment rules that were reliably estimated on the same observational dataset in a transparent and reproducible manner.

4.3 Connecting OWL and Split-Regression

4.3.1 Binary Outcome

Consider a clinical setting where we observe a binary outcome variable $Y \in \{0, 1\}$ that indicates the presence of a desirable event (e.g. remaining relapse-free after 5 years). We also observe a set of patient characteristics $X = (\mathbf{X_1}, \ldots, \mathbf{X_p})$ that we would like to use as inputs to a treatment rule that will recommend a treatment option $T \in \{0, 1\}$ based on an

individual's value of X, which we assume contains all the confounding variables.[3]

We saw in Chapter 3 that the split-regression approach recommends $T = 1$ for an individual with characteristics X if

$$P(Y = 1 \mid T = 1, X) \quad > \quad P(Y = 1 \mid T = 0, X), \tag{4.4}$$

which quite intuitively recommends the treatment option under which the individual's probability of the desirable outcome is higher. On the other hand, the treatment rule

$$P(T = 1 \mid Y = 1, X) \quad > \quad P(T = 1 \mid X) \tag{4.5}$$

may look comparatively unintuitive, but it does have a Bayesian interpretation: It recommends $T = 1$ if the "posterior" propensity score $P(T = 1 \mid Y = 1, X)$ exceeds the "prior" propensity score $P(T = 1 \mid X)$, since this indicates a positive association between experiencing the desirable outcome $Y = 1$ and receiving $T = 1$ for a particular individual with characteristics X.

In fact, the rules in (4.4) and (4.5) are equivalent. As shown in Appendix C.1, the equivalence is a direct consequence of Bayes' Rule and thus, in a sense, the "unintuitiveness" of (4.5) is the same "unintuitiveness" of Bayes' Rule. We also show a visualization of this equivalence in Figure 4.1, where we simulated a simple dataset that included a fixed propensity score $P(T = 1 \mid X) = P(T = 1) = 0.67$.[4]

---

[3]For simplicity, we also assume here that the characteristics X can be reliably measured in future clinical settings, so they are sensible inputs for both the rule and the propensity score; a setting where these inputs need not overlap was explored in Chapter 3 and is accommodated in the framework we present in Section 4.4.1.

[4]To simulate the data underlying the plots, we generated a single patient characteristic X from a Uniform$(0, 1)$ distribution, generated binary treatment indicator T as taking the value 1 with probability 0.67 (i.e. a constant propensity score of 0.67), and simulated binary outcome Y using separate linear-logistic models in the $T = 1$ and $T = 0$ group. For more detail please see the code shared at `github.com/jhroth/simulations-elucidating-owl` to simulate this data and reproduce the plots shown in Figure 4.1.

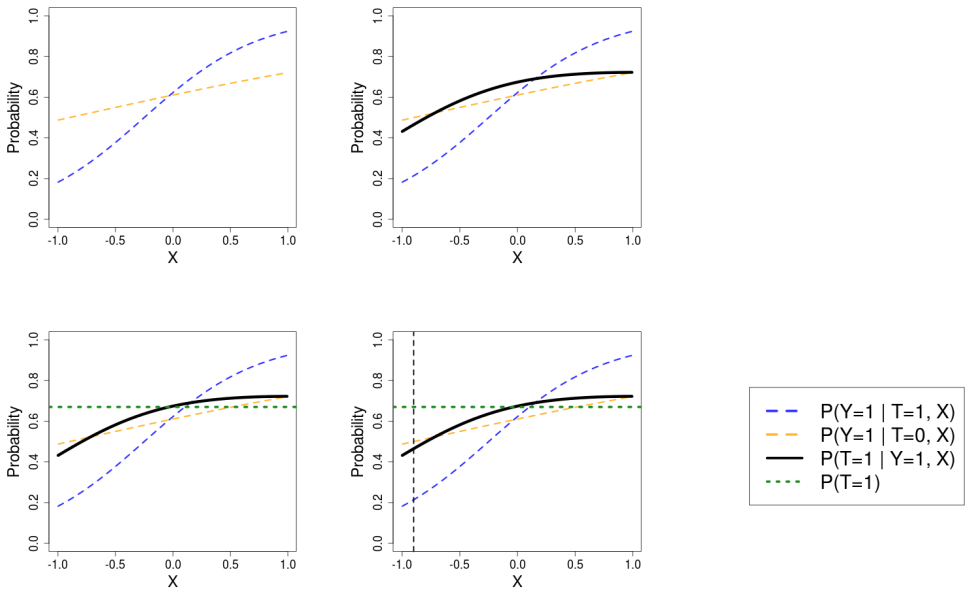


Figure 4.1: Visualization of the treatment rules from the split-regression and OWL approaches ($\pi(X) \equiv P(T = 1) = 0.67$).

The top-left panel of Figure 4.1 shows $P(Y = 1 \mid T = 1, X)$ and $P(Y = 1 \mid T = 0, X)$ plotted as a function of the single characteristic X. We can see that split-regression recommends treatment when $X > 0$ because that is where we begin to see $P(Y = 1 \mid T = 1, X) > P(Y = 0 \mid T = 1, X)$. In the top-right panel of Figure 4.1, the "posterior" propensity score $P(T = 1 \mid Y = 1, X)$ is overlaid in black. The bottom-left panel of Figure 4.1 adds in the "prior" propensity score $P(T = 1 \mid X)$ with the green dashed line. We can see that the solid black curve (posterior propensity score) exceeds the green dashed line (prior propensity score) when $X > 0$, so the resulting rule (treat when $X > 0$) is identical to the rule from split-regression. In the bottom-right panel of Figure 4.1, we add a dashed black vertical line to represent a hypothetical individual with characteristic $X = -0.9$, who would be recommended to not receive treatment under both (equivalent) rules.[5]

[5]Although Figure 4.1 is drawn for a fixed propensity score that is independent of patient characteristics X (e.g. as in an RCT), in an observational study where T is only influenced by X so the propensity score

Why is it helpful to understand the posterior and prior propensity scores, $P(T = 1 \mid Y = 1, X)$ and $P(T = 1 \mid X)$, that form the rule in (4.5) to elucidate the connection between split-regression and OWL? It turns out that just as split-regression directly targets (4.4), OWL targets (4.5). To see this, recall that OWL seeks the treatment rule $\mathcal{D} : X \to \{0, 1\}$ that maximizes

$$\mathrm{E}_{Y,T,X}\left[\frac{I[T = \mathcal{D}(X)]}{P(T = 1 \mid X)\,T + (1 - P(T = 1 \mid X))(1 - T)}\,Y\right], \tag{4.6}$$

which, as detailed in Appendix C.2, very naturally reduces to choosing $\mathcal{D}(X)$ such that

$$\frac{1 - P(T = 1 \mid X, Y = 1)}{1 - P(T = 1 \mid X)} \cdot I[\mathcal{D}(X) = 0] + \frac{P(T = 1 \mid X, Y = 1)}{P(T = 1 \mid X)} \cdot I[\mathcal{D}(X) = 1] \tag{4.7}$$

is maximized. We see that (4.7) is simply a function of the familiar posterior propensity score $P(T = 1 \mid Y = 1, X)$ and prior propensity score $P(T = 1 \mid X)$. In fact, for one Bayesian interpretation we can define a *Bayes factor*

$$K_t(y) = \frac{P(T = t \mid X, Y = y)}{P(T = t \mid X)}, \tag{4.8}$$

a non-negative metric that, for $y = 1$, will have a value above 1 (which supports a positive association between $Y = 1$ and $T = t$) in exactly one treatment group and has a value below 0 (which suggests a negative association between $Y = 1$ and $T = t$) in the other treatment group. Using this Bayes factor notation, the rule targeted by OWL in (4.7) maximizes

$$K_0(1) \cdot I[\mathcal{D}(X) = 0] + K_1(1) \cdot I[\mathcal{D}(X) = 1], \tag{4.9}$$

and will simply recommend the treatment option with the larger Bayes factor (or, equivalently, recommend the treatment option with Bayes factor greater than 1). For a different

---

is a function of X, we would be making the same fundamental visual comparisons as in Figures 4.1: Is the blue curve above the orange curve for split-regression, and is the black curve above the green line/curve, which is now a non-constant function of X, for OWL?

interpretation, inspecting (4.7) also tells us that the rule $\mathcal{D}(X)$ targeted by OWL (i.e. the maximizer) recommends treatment when

$$\frac{P(T=1 \mid X, Y=1)}{P(T=1 \mid X)} > \frac{1-P(T=1 \mid X, Y=1)}{1-P(T=1 \mid X)},$$

or equivalently, in the more intuitive framework of odds (which we recognize from logistic regression and working with odds ratios, for example), when

$$\underbrace{\frac{P(T=1 \mid X, Y=1)}{1-P(T=1 \mid X, Y=1)}}_{\text{“posterior odds”}} > \underbrace{\frac{P(T=1 \mid X)}{1-P(T=1 \mid X)}}_{\text{“prior odds”}}. \tag{4.10}$$

That is, OWL simply targets the rule that recommends treatment when the odds of the posterior propensity score $P(T=1 \mid X, Y=1)$ exceed the odds of the prior propensity score $P(T=1 \mid X)$. Or equivalently, since the odds $\frac{p}{1-p}$ are a monotonic transformation of a probability p, we can transform (4.10) to the probability scale as

$$P(T=1 \mid Y=1, X) \quad > \quad P(T=1 \mid X),$$

which is exactly equal to (4.5) from the beginning of this section.

4.3.2 Continuous Outcome: The Rule Targeted by OWL

What happens when Y is a continuous outcome where larger values are more desirable? As shown in Appendix C.3, the target of OWL quickly reduces to a continuous analog of (4.7):

$$\begin{aligned} &\left\{ \int_Y \frac{1-P(T=1 \mid X, Y=y)}{1-P(T=1 \mid X)} \; y \; dF(Y=y \mid X) \right\} I[\mathcal{D}(X)=0] \\ +&\left\{ \int_Y \frac{P(T=1 \mid X, Y=y)}{P(T=1 \mid X)} \; y \; dF(Y=y \mid X) \right\} I[\mathcal{D}(X)=1]. \end{aligned} \tag{4.11}$$

The similarity of the rule targeted by OWL with a continuous outcome (4.11) to the rule targeted by OWL with a binary outcome (4.7) suggests re-writing (4.11) with Bayes factors

as

$$\left\{\int_Y K_0(y)\ y\ dF(Y=y\mid X)\right\} I[\mathcal{D}(X)=0] \tag{4.12}$$

$$+\left\{\int_Y K_1(y)\ y\ dF(Y=y\mid X)\right\} I[\mathcal{D}(X)=1]. \tag{4.13}$$

As a result, instead of simply recommending the treatment option with the larger Bayes factor (as would be the case with a binary outcome), the rule targeted by OWL recommends the treatment option with the larger Bayes factor, *weighted by Y and averaged over the conditional density of Y given X*.

4.3.3 Continuous Outcome: Plug-In Estimation vs. Minimization

One method for estimating the rule targeted by OWL in (4.13) would be an intuitive "plug-in" approach that substitutes sample means for expectations and directly replaces each of the population distributions with a sample-based model. More precisely, given n_{D1} observations on a development dataset D1 and n_{D2} observations on an evaluation dataset D2 (which may be partitions of a single dataset), the plug-in approach would estimate the rule targeted by OWL with the following procedure:

1. Using observations $(x_1, t_1, y_1), \ldots, (x_{n_{\mathrm{D1}}}, t_{n_{\mathrm{D1}}}, y_{n_{\mathrm{D1}}})$ on D1, form estimates of the posterior propensity score $\widehat{P}^{\mathrm{D1}}(T=1\mid Y,X)$ and of the prior propensity score $\widehat{P}^{\mathrm{D1}}(T=1\mid X)$.

2. Using observations $(x_1, y_1), \ldots, (x_{n_{\mathrm{D1}}}, y_{n_{\mathrm{D1}}})$ on D1, form an estimate of the conditional density function $\widehat{f}^{\mathrm{D1}}(Y\mid X)$.

3. For *each* observation $j = 1, \ldots, n_{\mathrm{D2}}$ on D2:

(i) Compute the plug-in estimate of the integral in (4.12)

$$\mathrm{A}_0(X = x_j) \equiv \frac{1}{n_{\mathrm{D1}}} \sum_{i=1}^{n_{\mathrm{D1}}} \frac{\widehat{P}^{\mathrm{D1}}(T = 0 \mid X = x_j, Y = y_i)}{\widehat{P}^{\mathrm{D1}}(T = 0 \mid X = x_j)} \cdot y_i \cdot \widehat{f}^{\mathrm{D1}}(Y \mid X = x_j).$$

(ii) Compute the plug-in estimate of the integral in (4.13)

$$\mathrm{A}_1(X = x_j) \equiv \frac{1}{n_{\mathrm{D1}}} \sum_{i=1}^{n_{\mathrm{D1}}} \frac{\widehat{P}^{\mathrm{D1}}(T = 1 \mid X = x_j, Y = y_i)}{\widehat{P}^{\mathrm{D1}}(T = 1 \mid X = x_j)} \cdot y_i \cdot \widehat{f}^{\mathrm{D1}}(Y \mid X = x_j).$$

(iii) For that jth individual, define the rule as

$$\mathcal{D}(X = x_j) = I[\mathrm{A}_1(x_j) > \mathrm{A}_0(x_j)].$$

Unfortunately, this is an onerous procedure because it requires us to estimate the conditional density function $f(y \mid x)$, which is a very challenging task fraught with high variability when x has more than a handful of possible values (Varadhan and Seeger, 2013).

In contrast to the plug-in procedure (which, while perhaps more intuitive, is unwieldy computationally), OWL as formulated in Zhao et al. (2012) proposes the vastly simplified recipe:

1. Using observations $(x_1, t_1, y_1), \ldots, (x_{n_{\mathrm{D1}}}, t_{n_{\mathrm{D1}}}, y_{n_{\mathrm{D1}}})$ on D1, form an estimate of the prior propensity score $\widehat{P}^{\mathrm{D1}}(T = 1 \mid X)$.

2. Using observations $(x_1, t_1, y_1), \ldots, (x_{n_{\mathrm{D1}}}, t_{n_{\mathrm{D1}}}, y_{n_{\mathrm{D2}}})$ on D1, use SVM (or another classification method that accommodates observation weights) to predict response variable T using features X and observation weights equal to the usual IPW weights[6] multiplied by Y. Call the estimated classifier $\widehat{T}^{\mathrm{D1}}_{\mathrm{OWL}}(X)$ that maps individual characteristics X to a 0/1 class (transformed from a probability, if necessary).

[6]That is, $\frac{1}{(t_i \widehat{P}^{\mathrm{D1}}(T=1|X=x_j)+(1-t_i)\left(1-\widehat{P}^{\mathrm{D1}}(T=0|X=x_j)\right)}$

3. Given a new set of characteristics such as x_j on D2 for $j = 1, \ldots, n_{\text{D2}}$, evaluate the treatment rule

$$\widehat{D}_{\text{OWL}}(X = x_j) = I\left[\widehat{T}^{\text{D1}}_{\text{OWL}}(x_j) = 1\right] \tag{4.14}$$

That is, OWL offers a dramatic simplification because it completely bypasses the estimation of a conditional density and instead solves a single minimization problem in the combined sample on D1 , in place of a burdensome estimating/averaging process of conditional densities over $2 \times n_{\text{D1}} \times n_{\text{D2}}$ combinations of inputs from D1 and D2.

In Chapter 3 we described how the split-regression procedure also circumvents an unwieldy estimation/averaging procedure over conditional densities by instead solving a minimization problem. Thus, one can view OWL with a continuous outcome Y in a similar fashion as one views split-regression: as a practical method for transforming an intuitive but computationally challenging target parameter into a straightforward minimization framework based on estimation of conditional means, rather than on the convoluted estimation/averaging of conditional densities. It just turns out that when viewed through the minimization lens, split-regression retains a very intuitive form while the form of OWL (i.e. predicting treatment assignment rather than clinical outcome) becomes a bit more distorted.

We also described in Chapter 3 how with a continuous response Y, the split-regression approach can be easily modified to recommend individuals receive treatment if their outcomes are at least C units greater under $T = 1$ than under $T = 0$, for a $C > 0$ that represents a minimum clinically relevant effect of treatment. With the direct approaches of OWL and direct-interactions (and other direct approaches as well) fitted values from estimation are not on the scale of Y and thus the modification is not as natural. However, prior to estimating the treatment rule with such direct approaches, one can simply define the shifted outcome $Y^C \equiv Y + C$ for individuals with $T = 0$ and leave $Y^C \equiv Y$ in the $T = 1$ group, and then use Y^C as the outcome variable during treatment rule development. The test-positives group would then be the subpopulation whose outcome Y is expected to be at least C units larger under $T = 1$ than $T = 0$.

4.4 Integrating OWL Into Chapter 3's Principled Framework

In Chapter 3 we presented a principled framework and software implementation to develop and evaluate treatment rules using the indirect approach of split-regression. In Section 4.4.1, we show how the direct approach of OWL/OWL framework can be folded into that step-by-step framework with only a few small tweaks.

To ease notation, we adopt the language of Chapter 3 where we proposed partitioning patient characteristics as $X = (C^{\text{TI}}, C^{\text{TN}}, C^{\text{NI}}, C^{\text{NN}})$, where the abbreviations in each superscript tell us whether each available characteristic (C) affects treatment assignment in the current study (TN), is expected to be observed in independent studies (NI), both (TI), or neither (NN). When introducing this algorithmic recipe in Chapter 3, we also defined $\mathbf{C}^{\text{T}} = (C^{\text{TI}}, C^{\text{TN}})$ as all the patient characteristics that may affect treatment assignment in the current study and $\mathbf{R}$ as a subset of all the characteristics $\mathbf{C}^{\text{I}} = (C^{\text{TI}}, C^{\text{NI}})$ that the researcher feels may affect treatment response and thus are sensible inputs for the treatment rule (since they will be available in future clinical settings).

4.4.1 The Recipe, Adapted for OWL Framework

What follows is the algorithmic recipe from Chapter 3 adapted to the OWL framework method implied by Zhao et al. (2012) As in Chapter 3, we advocate applying this procedure to a development dataset (D1) that is independent of the evaluation dataset (D2) to ensure trustworthy evaluation of treatment rule benefit.

1. Use the scientific knowledge underlying D1 to partition observed patient characteristics (aside from outcome and treatment) into the four categories presented in Table 3.1: C^{TI}, C^{TN}, C^{NI}, and C^{NN}. Also choose $\mathbf{R} \in \mathbb{R}^{p'}$ as the inputs for the treatment rule and form $\mathbf{C}^{\text{T}} = (C^{\text{TI}}, C^{\text{TN}})$, $\mathbf{C}^{\text{I}} = (C^{\text{TI}}, C^{\text{NI}})$. [7]

[7] In applied work it can be useful to truncate estimated propensity scores so they are not too close to 0 or 1 (which can lead to very large observation weights); a default setting in our software implementation truncates estimated propensity scores to stay between 0.05 and 0.95, but this choice can be overwritten by the user.

2. For observations $i = 1, \ldots, n$ on D1:

   (a) Choose a prediction method and estimate the propensity score $\tilde{P}^{\mathrm{D1}}(T = 1|\mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i, \mathbf{R} = \mathbf{r}_i)$.

   (b) Compute the weights $\tilde{W}_{t_i}(\mathbf{c}^{\mathrm{T}}{}_i, \mathbf{r}_i) = \frac{Y_i}{\tilde{P}^{\mathrm{D1}}(T=t_i|\mathbf{C}^{\mathrm{T}}=\mathbf{c}^{\mathrm{T}}{}_i, \mathbf{R}=\mathbf{r}_i)}$, for $t = 0, 1$.

   (c) Choose a prediction method that accommodates observation weights (e.g. logistic regression, lasso, boosted trees, among others) and predict T with features $\mathbf{R}$ and observation weights $\tilde{W}_{t_i}$. For example, weighted logistic regression would yield, for observations $i = 1, \ldots, n$ on D1,

$$\tilde{\beta}^{\mathrm{D1}} \equiv \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^{p'}}{\arg\min} -\frac{1}{N} \sum_{i=1}^{N} \log \left[\tilde{W}_{t_i} \cdot t_i^{\mathrm{expit}(\beta_0 + \mathbf{r}_i^T \beta)} \cdot (1 - t_i)^{\mathrm{expit}(\beta_0 + \mathbf{r}_i^T \beta)} \right] \tag{4.15}$$

   and we would obtain a predicted risk score as $\tilde{f}^{\mathrm{D1}}_{\mathrm{OWL}}(\mathbf{r}) \equiv \mathbf{r}^\top \tilde{\beta}^{\mathrm{D1}}$. Here in the case of logistic regression we can also predict the probability of $T = 1$ with $\tilde{p}^{\mathrm{D1}}_1(\mathbf{r}) \equiv \mathrm{expit}\left(\tilde{f}^{\mathrm{D1}}_{\mathrm{OWL}}(\mathbf{r})\right)$ and the probability of $T = 0$ with $\tilde{p}^{\mathrm{D1}}_0(\mathbf{r}) = 1 - \tilde{p}^{\mathrm{D1}}_0(\mathbf{r})$.

3. Form the treatment rule

$$\tilde{B}_{\mathrm{OWL}}(\mathbf{r}) \equiv I\left[\tilde{f}^{\mathrm{D1}}_{\mathrm{OWL}}(\mathbf{r}) > 0\right], \tag{4.16}$$

   where $I(\cdot)$ is the indicator function. For logistic regression, (4.16) is equivalent to recommending treatment when the predicted treatment probability exceeds 0.5 (since $\mathrm{expit}(0) = 0.5$); other classification methods may only provide an $\tilde{f}^{\mathrm{D1}}_{\mathrm{OWL}}(\mathbf{r})$ that yields a predicted class (i.e. a 0/1) rather than a score that can be mapped to a probability.

4. Use scientific knowledge underlying D2 to select the potential confounders $\mathbf{C}^{\mathrm{T,\ eval}}$.

5. For observations $j = 1, \ldots, m$ on D2:

   (a) Assign the recommended treatment with $\tilde{B}_{\mathrm{OWL}}(\mathbf{r}_j)$.

(b) As discussed in the "Evaluating the Rule" section of Chapter 3, form the IPW-based estimators of the average treatment effect (ATE) and average benefit of the rule (ABR) based on applying the rule $\tilde{B}_{\text{OWL}}(\mathbf{r}_j)$, accounting for the potential confounders $\mathbf{C}^{\text{T, eval}}$.

4.5 Integrating Direct-Interactions Into Chapter 3's Principled Framework

Another promising and flexible direct approach for estimating treatment rules in observational studies was proposed by Chen et al. (2017), based on the earlier insights in the RCT setting published in Tian et al. (2014). We refer to this method as the *direct-interactions* approach and also implement it in `DevTreatRules` to serve as an alternative to the OWL/OWL framework and split-regression approaches.

4.5.1 Direct-Interactions in a Simplified RCT Setting

Here we describe direct-interactions in the simplified RCT setting considered in Tian et al. (2014): We observe the data (X, Y, T) from an RCT where treatment assignment is coded as $T \in \{-1, 1\}$ and the trial had equal randomization so that $P(T = 1 \mid X) = P(T = -1 \mid X) = 0.5$ for all individual characteristics X.

Now suppose the continuous outcome Y (where higher values are more desirable) is truly generated by the linear model

$$Y = \underbrace{\beta_0^* + \beta^* X}_{\text{main effect}} + \underbrace{(\gamma_0^* + \gamma^* X)T}_{\text{treatment effect}} + \epsilon, \tag{4.17}$$

where ϵ has mean 0 (i.e. there is no unmeasured confounding). In (4.17), $\beta_0^* + \beta^* X$ is the linear main effect of X on Y regardless of T, γ_0^* is the effect of T on Y regardless of X, and $\gamma^* X$ is the effect of X on Y whose sign depends on T. Looking at (4.17), we see that the ideal treatment rule would be

$$\mathcal{D}_{\text{ideal}}(X) = I[\gamma_0^* + \gamma^* X > 0], \tag{4.18}$$

which does not directly depend on β_0^* and β^*.

The "correct" linear model would estimate (γ_0^*, γ^*) by solving

$$(\gamma_0^{\text{adj.}}, \gamma^{\text{adj.}}) \equiv \arg\min_{\gamma_0,\gamma} E_{X,Y,T}\left\{[Y - (\beta_0 + \beta X + (\gamma_0 + \gamma X)T)]^2\right\}. \tag{4.19}$$

where the solutions $(\gamma_0^{\text{adj.}}, \gamma^{\text{adj.}})$ are "adjusted" for the main effect $\beta_0 + \beta X$. The direct-interactions method proposes instead to estimate (γ_0^*, γ^*) with the "incorrect" model

$$(\gamma_0^{\text{unadj.}}, \gamma^{\text{unadj.}}) \equiv \arg\min_{\gamma_0,\gamma} E_{X,Y,T}\left\{[Y - (\gamma_0 + \gamma X)T)]^2\right\}, \tag{4.20}$$

where the solutions $(\gamma_0^{\text{unadj.}}, \gamma^{\text{unadj.}})$ are "unadjusted" for the main effect. The direct-interactions method then simply defines its estimated treatment rule as $\mathcal{D}_{\text{DI}}(X) \equiv I[\gamma_0^{\text{unadj.}} + \gamma^{\text{unadj.}} X > 0]$, to approximate (4.18).

4.5.2 Intuition for Direct-Interactions in a Simplified RCT Setting

How does the "incorrect" minimization problem proposed by direct-interactions in (4.20) compare to the "correct" specification in (4.19)? Since the coefficients of T and XT (as with any regression coefficients) are defined with respect the other variables included in the model, the meanings of (γ_0, γ) under (4.20) and (4.19) are completely different. Nonetheless, in this simplified RCT setting, the *minimizers* $(\gamma_0^{\text{unadj.}}, \gamma^{\text{unadj.}})$ and $(\gamma_0^{\text{adj.}}, \gamma^{\text{adj.}})$ are equivalent. To see the equivalence, let $M \equiv (\beta_0 + \beta X)$ and $C \equiv (\gamma_0 + \gamma X)$ and note the expansion of the "correct" model in (4.19) is

$$\begin{aligned}
&\arg\min_{\gamma_0,\gamma} E_{X,Y,T}\left\{[Y - (M + CT)]^2\right\} \\
&\arg\min_{\gamma_0,\gamma} \left\{E[Y^2 + (M^2 + (CT)^2 + 2MCT) - 2Y(M + CT)]\right\} \\
&\arg\min_{\gamma_0,\gamma} \Big\{ \underbrace{E[(Y - CT)^2]}_{\text{same as (4.20)}} - 2E[MCT] + \underbrace{E[M^2 - 2YM]}_{\text{does not depend on } \gamma_0,\gamma} \Big\},
\end{aligned} \tag{4.21}$$

so we see the minimizers in (4.19) and (4.20) will be equivalent if $E[MCT] = 0$. Now we apply the fact that X is independent of T in this RCT to write

$$E[MCT] \equiv E[(\beta_0 + \beta X)(\gamma_0 + \gamma X)T] = E[(\beta_0 + \beta X)(\gamma_0 + \gamma X)]E[T].$$

Further, we have $E[T] = 0$ because $P(T = 1) = P(T = -1) = 0.5$ in this RCT with equal randomization. As a result, $E[MCT] = 0$ and we see from (4.21) that

$$\arg\min_{\gamma_0,\gamma} E_{X,Y,T}\left\{[Y - (M + CT)]^2\right\} = \arg\min_{\gamma_0,\gamma} E_{X,Y,T}[(Y - CT)^2], \tag{4.22}$$

so indeed the minimizers $(\gamma_0^{\text{unadj.}}, \gamma^{\text{unadj.}})$ and $(\gamma_0^{\text{adj.}}, \gamma^{\text{adj.}})$ are equivalent in this simplified RCT setting.

4.5.3 Intuition for Direct-Interactions More Generally

For intuition justifying the direct-interactions method outside the simplest RCT setting with E[T] = 0, we discuss the reasoning from Tian et al. (2014). In particular, Appendix A of Tian et al. (2014) shows that for a general $f(X)$ (defined so that $f(X)$ always includes a main effect for T, e.g. $f(X) = \gamma_0 + \gamma X$), the minimizer of squared-error loss[8] with the "incorrect" model that ignores the main effect of X,

$$\hat{f}_{\text{DI}} \equiv \arg\min_{f} E_{X,Y,T}[Y - f(X)T]^2, \tag{4.23}$$

is $\hat{f}_{\text{DI}}(X) = \frac{1}{2}(E[Y \mid X, T = 1] - E[Y \mid X, T = -1])$. This tells us that the treatment rule targeted by direct-interactions with $\mathcal{D}_{\text{DI}}(X) \equiv I[\hat{f}_{\text{DI}}(X) > 0]$, of which $I[\gamma_0 + \gamma X > 0]$ is

[8]Chen et al. (2017) describe general optimization criteria that loss functions must satisfy to be suitable choices; other appropriate loss functions discussed in that paper are logistic loss and outcome-weighted loss, among others – please see Chen et al. (2017) for more detail.

one potential choice, will make the very intuitive decision to recommend treatment when

$$E[Y \mid X, T = 1] > E[Y \mid X, T = -1],$$

just like split-regression (but, unlike split-regression, does not have to model the main effect of X).

In our Appendix C.4, we replicate the calculation in Appendix A of Tian et al. (2014) with added detail and description between steps. For the interested reader we present here, as a heuristic, a few selected steps with the choice $f(X) = \gamma_0 + \gamma X$. Please see Appendix C.4 for additional algebra and please consult the original work of Tian et al. (2014) and Chen et al. (2017) for much more context and detail (e.g. general conditions on loss functions for which this approach is valid).

To show exactly where one would apply the assumption of an RCT setting with $P(T = 1 \mid X) = P(T = -1 \mid X) = \frac{1}{2}$, we can show the expectation from the "incorrect" model of direct-interactions in (4.20) is equivalent to

$$\int_X \Bigg\{ \left[\int_Y (Y - (\gamma_0 + \gamma X)T)^2 dF(Y \mid X, T = 1) P(T = 1 \mid X) \right] \tag{4.24}$$

$$+ \left[\int_Y (Y - (\gamma_0 + \gamma X)T)^2 dF(Y \mid X, T = -1) P(T = -1 \mid X) \right] \Bigg\} dF(X). \tag{4.25}$$

Now substituting $P(T = 1 \mid X) = P(T = -1 \mid X) = \frac{1}{2}$ greatly simplifies (4.24) and (4.25) to the sum of conditional expectations

$$\frac{1}{2} E_X \left\{ E_Y \left[(Y - (\gamma_0 + \gamma X)T)^2 \mid X, T = 1 \right] + E_Y \left[(Y - (\gamma_0 + \gamma X)T)^2 \mid X, T = 1 \right] \right\}. \tag{4.26}$$

A bit more algebra (spelled out in Appendix C.4) reveals that

$$
\begin{aligned}
(\gamma_0^{\text{unadj.}}, \gamma^{\text{unadj.}}) &\equiv \arg\min_{\gamma_0,\gamma} E_{X,Y,T}\left\{[Y - (\gamma_0 + \gamma X)T]^2\right\} \\
&= \arg\min_{\gamma_0,\gamma} E_X\left\{\left[\frac{1}{2}\left(E[Y \mid X, T=1] - E[Y \mid X, T=-1]\right) - (\gamma_0 + \gamma X)\right]^2\right\},
\end{aligned}
$$

which tells us that the minimizer is $\hat{f}(X) = \frac{1}{2}\left(E[Y \mid X, T=1] - E[Y \mid X, T=-1]\right)$. As a result, the treatment rule from direct-interactions, $\mathcal{D}_{\text{DI}}(X) \equiv I[\gamma_0^{\text{unadj.}} + \gamma^{\text{unadj.}} X > 0]$, very sensibly recommends $T = 1$ when the individual's expected outcome is higher under $T = 1$ than $T = 0$ (just like the rule from split-regression) and does so without modeling the main effect of X on Y (while split-regression does require us to model this main effect).

Not mentioned in Tian et al. (2014), but used a key component of Chen et al. (2017), is that if X and T were not independent (e.g. in an observational study) one could use the IPW-weighted analogs of (4.24) and (4.25),

$$
\int_X \left\{ \left[\frac{\int_Y (Y - (\gamma_0 + \gamma X)T)^2}{P(T=1 \mid X)} dF(Y \mid X, T=1) P(T=1 \mid X)\right] \right. \tag{4.27}
$$

$$
\left. + \left[\int_Y \frac{(Y - (\gamma_0 + \gamma X)T)^2}{P(T=-1 \mid X)} dF(Y \mid X, T=-1) P(T=-1 \mid X)\right] \right\} dF(X), \tag{4.28}
$$

to achieve the same "cancellation" of $P(T=1 \mid X)$ and $P(T=-1 \mid X)$ that in the RCT setting moved us from (4.24) and (4.25) to the final minimizer $\hat{f}(X) = \frac{1}{2}\left(E[Y \mid X, T=1] - E[Y \mid X, T=-1]\right)$ and highly interpretable treatment rule. That is, performing direct-interactions with observational data simply requires IPW observation weights in place of the uniform observation weights we implicitly used with RCT data.

4.5.4 The Recipe, Adapted for Direct-Interactions

Now we show how direct-interactions (Tian et al., 2014; Chen et al., 2017) can also be folded into the principled development and evaluation framework from Chapter 3 to serve as

another option for estimating treatment rules in addition to split-regression and OWL/OWL framework. What follows is the algorithmic recipe for developing/evaluating treatment rules adapted to the direct-interactions approach.

1. Use the scientific knowledge underlying development dataset D1 to partition observed patient characteristics (aside from outcome and treatment) into the four categories presented in Table 3.1: C^{TI} , C^{TN} , C^{NI} , and C^{NN} . Also choose $\mathbf{R} \subseteq \mathbf{C}^{\mathrm{I}}$ as the inputs for the treatment rule and form $\mathbf{C}^{\mathrm{T}} = (C^{\mathrm{TI}}, C^{\mathrm{TN}})$, $\mathbf{C}^{\mathrm{I}} = (C^{\mathrm{TI}}, C^{\mathrm{NI}})$.

2. For observations $i = 1, \ldots, n$ on D1:

   (a) Re-code the treatment variable as $T^* \in \{-1, 1\}$, instead of $T \in \{0, 1\}$. Choose a prediction method and estimate the propensity score $\tilde{P}^{\mathrm{D1}}(T^* = 1 | \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i, \mathbf{R} = \mathbf{r}_i)$.

   (b) Compute the weights $\tilde{W}_{t_i^*}(\mathbf{c}^{\mathrm{T}}{}_i, \mathbf{r}_i) = \frac{1}{\tilde{P}^{\mathrm{D1}}(T^* = t_i | \mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i, \mathbf{R} = \mathbf{r}_i)}$, for $t^* = -1, 1$.

   (c) Choose a prediction method that accommodates observation weights (e.g. generalized linear regression, lasso, boosted trees, and many others)[9] to estimate $\hat{f}$ by specifying the following for $i = 1, \ldots, n$

   - Observation weights: $\tilde{W}_{t_i^*}(\mathbf{C}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}{}_i)$
   - Response variable: y_i
   - Features: $(t_i^*, t_i^* \mathbf{r}_{i,1}, \ldots, t_i^* \mathbf{r}_{i,p'})$ for $\mathbf{R} \in \mathbb{R}^{p'}$. That is, there are $p' + 1$ predictors: the -1/1 treatment indicator T^* and its product with each of the patient characteristics in $\mathbf{R}$. Note that there is no intercept term if one would normally be included and no main effect terms for the patient characteristics in $\mathbf{R}$.

[9] Other statistical learning methods outside of the regression framework (e.g. boosted trees) can accommodate the list in 2c and thus could also be used in the direct-interactions method, but their theoretical foundation and empirical performance has not yet been evaluated.

For example, weighted linear regression with a continuous response would yield, for observations $i = 1, \ldots, n$ on D1,

$$\tilde{\beta}^{\mathrm{D1}} \equiv \underset{\beta_{T^*}, \beta_1, \ldots, \beta_{p'} \in \mathbb{R}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \tilde{W}_0(\mathbf{c}^{\mathrm{T}}{}_i) \left[y_i - \left(\beta_{T^*} t_i^* + \beta_1 \mathbf{r}_{i,1}^\top t_i^* + \ldots + \beta_{p'} \mathbf{r}_{i,p'}^\top t_i^* \right) \right]^2, \tag{4.29}$$

where we again note that there is no intercept term or main effect terms for the characteristics $\mathbf{r}_{i,1}, \ldots, \mathbf{r}_{i,p'}$. We now obtain a predicted risk score with the function

$$\tilde{f}_{\mathrm{DI}}^{\mathrm{D1}}(1, \mathbf{r}) \equiv (1, \mathbf{r})^\top \tilde{\beta}^{\mathrm{D1}}. \tag{4.30}$$

3. Form the treatment rule

$$\tilde{B}(\mathbf{r})_{\mathrm{DI}} \equiv I\left[\tilde{f}_{\mathrm{DI}}^{\mathrm{D1}}(\mathbf{r}) > 0 \right], \tag{4.31}$$

where $I(\cdot)$ is the indicator function.

4. Use the scientific knowledge underlying independent evaluation dataset D2 to select the potential confounders $\mathbf{C}^{\mathrm{T,\ eval}}$.

5. For observations $j = 1, \ldots, m$ on D2:

   (a) Assign the recommended treatment with $\tilde{B}_{\mathrm{DI}}(\mathbf{r}_j)$.

   (b) As discussed in the "Evaluating the Rule" section of Chapter 3, form the IPW-based estimators of the average treatment effect (ATE) and average benefit of the rule (ABR) based on applying the rule $\tilde{B}_{\mathrm{DI}}(\mathbf{r}_j)$.

4.6 Simulations

4.6.1 Simulation Scenarios

For our simulation study, we generate a range of scenarios by modifying the following data-generating mechanism for observations $i = 1, \ldots, n$:

$$\begin{aligned}
\boldsymbol{X}_i \equiv (X_{i,1}, \ldots, X_{i,p}) &\overset{\text{iid}}{\sim} \text{Uniform}(0, 2) \\
\boldsymbol{Z}_i \equiv (Z_{i,1}, \ldots, Z_{i,q}) &\overset{\text{iid}}{\sim} \text{Normal}(0, 1) \\
L_i &\sim \text{Bernoulli}(0.5) \\
T_i \mid L_i &\sim \begin{cases} \text{Bernoulli}(0.75), & L_i = 0 \\ \text{Bernoulli}(0.25), & L_i = 1 \end{cases} \\
S_i \mid (L_i, T_i, X_i) &\sim \begin{cases} f_0(L_i, X_i), & T_i = 0 \\ f_1(L_i, X_i), & T_i = 1. \end{cases}
\end{aligned}$$

We then either generate a binary outcome as

$$Y_i \sim \text{Bernoulli}\left(\text{expit}\left(S_i\right)\right), \tag{4.32}$$

or a continuous outcome as

$$Y_i \sim \text{Normal}\left(S_i + \mu, \sigma^2\right), \tag{4.33}$$

where $\mu, \sigma > 0$.

Heuristically, we can think of L as an indicator of good prognosis, where an individual with good prognosis is less likely to be assigned treatment than an individual with a poor prognosis (25% compared to 75%). The effect of treatment is then influenced by characteristics X that do not affect treatment assignment (e.g. an uninterpretable gene expression measurement). On the other hand, L has the same influence (a main effect) on outcome for both treatment

groups. That is, some subset of X may be predictive of treatment response while L is purely prognostic (see the discussion in Ballman (2015), for example).

Figure 4.2 shows plots of X_1 versus S in Scenarios I-V, where we specified $p = 1$ and $q = 0$ for I-IV and q=40 for V. Figure 4.3 visualizes Scenarios VI-IX (where $p = 3$) and plots the piecewise-constant $f_0(X_j)$ in blue and $f_1(X_j)$ in orange for $j = 1, 2, 3$.

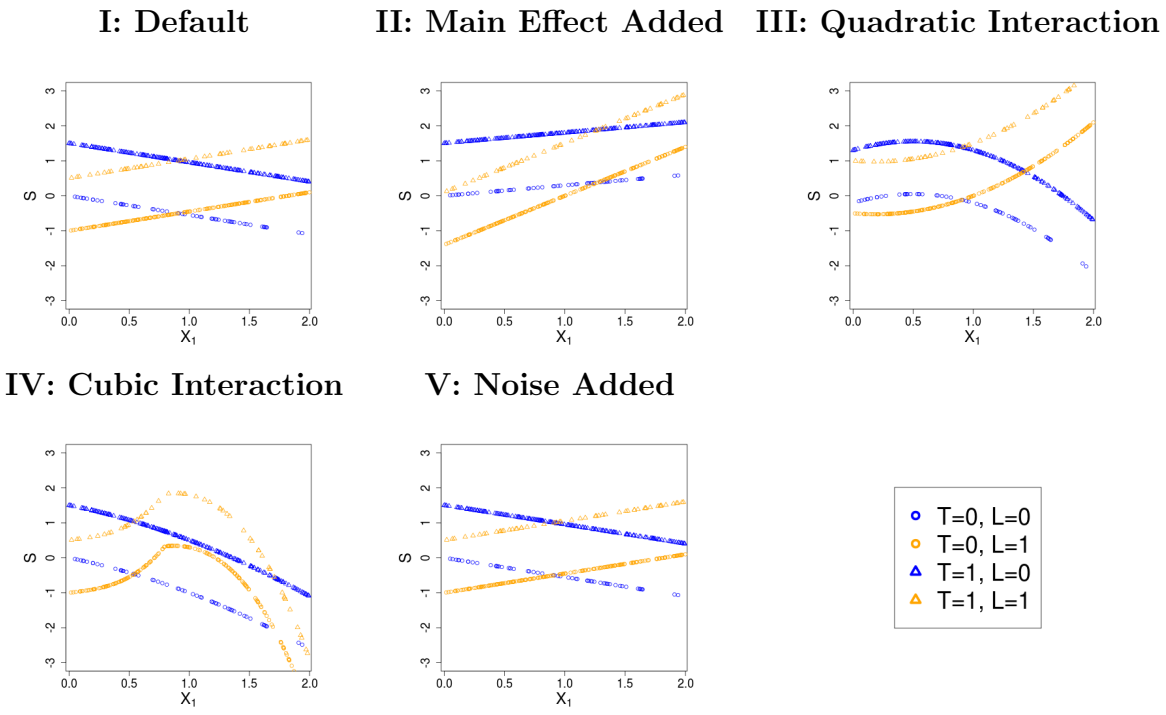


Figure 4.2: Simulation Scenarios 1-5: (p=1)

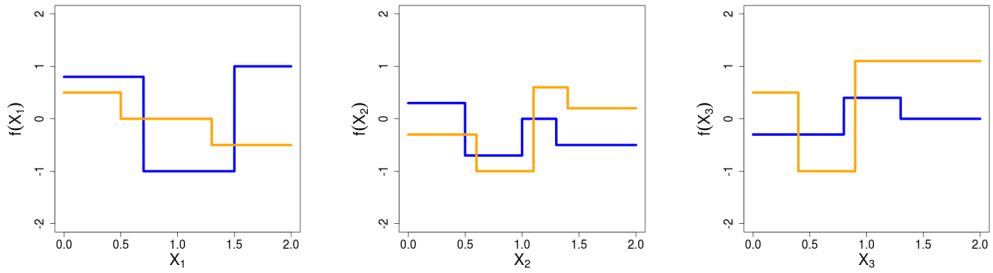


Figure 4.3: The piecewise-constant forms of $f(x_1), f(x_2), f(x_3)$,
used for Scenarios 6-9 (p=3). T=0 is shown in blue and T=1 is shown in orange.

Table 4.1 gives more detail on our eleven simulation scenarios; the corresponding response variable can be either binary or continuous as in (4.32) or (4.33). We share the code needed to reproduce these simulations at `github.com/jhroth/simulations-elucidating-owl`.

	Data Generation		Propensity Estimation		Rule Estimation	
Scenario	$f_0(L,X,Z)$	$f_1(L,X,Z)$	Model	Predictors	Model	Predictors
I: Default	See Fig. 4.2	See Fig. 4.2	Logistic	L	GLM	X_1
II: Main Effect	See Fig. 4.2	See Fig. 4.2	Logistic	L	GLM	X_1
III: Quadratic Interaction	See Fig. 4.2	See Fig. 4.2	Logistic	L	GLM	X_1
IV: Cubic Interaction	See Fig. 4.2	See Fig. 4.2	Logistic	L	GLM	X_1
V: Noise Added	See Fig. 4.2	See Fig. 4.2	Logistic	L	Lasso	$(X_1, Z_1, \ldots, Z_{40})$
VI: Piecewise Constant[1]	$L + \sum_{j=1}^{3} f_0(x_j)$	$L + \sum_{j=1}^{3} f_1(x_j)$	Logistic	L	GLM	(X_1, X_2, X_3)
VII: Two-Way Interactions	$L + \sum_{j=1}^{3} f_0(x_j)$+ two-way interaction[2]	$L + \sum_{j=1}^{3} f_1(x_j)$+ two-way interaction[3]	Logistic	L	GLM	(X_1, X_2, X_3)
VIII: Three-War Interaction	$L + \sum_{j=1}^{3} f_0(x_j)$+ three-way interaction[4]	$L + \sum_{j=1}^{3} f_1(x_j)$+ three-way interaction[5]	Logistic	L	GLM	(X_1, X_2, X_3)
IX: Higher-Order interactions	$L + \sum_{j=1}^{3} f_0(x_j)$+ higher-order interaction[6]	$L + \sum_{j=1}^{3} f_1(x_j)$+ higher-order interaction[7]	Logistic	L	GLM	(X_1, X_2, X_3)
X: High-Dimensional Noise Added	Default from Fig. 4.2	Default from Fig. 4.2	Logistic	L	Lasso	$(X_1, Z_1, \ldots, Z_{500})$
XI: High-Dimensional Noise Added + Prognostic	Default from Fig. 4.2 + prognostic[8]	Default from Fig. 4.2 + prognostic[8]	Logistic	L	Lasso	$(X_1 \ldots, X_6, Z_1, \ldots, Z_{500})$

Table 4.1: Description of simulation scenarios, where the outcome Y can be either binary or continuous.
[1] : See Fig. 4.3 for definitions of $f_0(x_j), f_1(x_j)$ for $j =, 1, 2, 3$
[2] : $-0.2 \cdot f_0(x_1) \cdot f_0(x_2) + 0.2 \cdot f_0(x_1) \cdot f_0(x_3) - 0.3 \cdot f_0(x_2) \cdot f_0(x_3)$
[3] : $0.1 \cdot f_1(x_1) \cdot f_1(x_2) + 0.3 \cdot f_1(x_1) \cdot f_1(x_3) - 0.1 \cdot f_1(x_2) \cdot f_1(x_3)$
[4] : $-0.5 \cdot f_0(x_1) \cdot f_0(x_2) \cdot f_0(x_3)$
[5] : $0.5 \cdot f_1(x_1) \cdot f_1(x_2) \cdot f_1(x_3)$
[6] : $-0.3 \cdot I\left[f_0(x_1) > 1\right] \cdot f_0(x_2)^2 + 0.4 \cdot I\left[f_0(x_2) > 0.5\right] \cdot f_0(x_3)^2 - 0.1 \cdot I\left[f_0(x_1) < 1.5\right] \cdot f_0(x_3)^3$
[7] : $0.4 \cdot I\left[f_1(x_1) > 1\right] \cdot f_1(x_2)^2 + -0.5 \cdot I\left[f_1(x_2) > 0.5\right] \cdot f_1(x_3)^2 - 0.6 \cdot I\left[f_1(x_1) < 1.5\right] \cdot f_1(x_3)^3$
[8] : $0.2 \cdot x_2 + 0.3 \cdot x_3 - 0.3 \cdot x_4 - 0.2 \cdot x_5 + 0.4 \cdot x_6$

In Scenario I, the rule model for split-regression is correctly specified. Scenarios II-V are modifications of Scenario I: Scenario II adds prognostic capacity to X_1 (so it influences outcome equally in both treatment arms) in addition to its previous predictive role; the rule model (GLM with only X_1 as a predictor) is mis-specified in Scenarios III-IV as the treatment effect becomes quadratic and cubic, respectively, in X_1; in Scenario V, the rule model incorrectly includes 40 noise variables as predictors. Our motivation for exploring simulations with/without quadratic interactions and main effects comes from Chen et al. (2017).

Scenario VI departs from Scenarios I-V by linking three variables (X_1, X_2, X_3) to the outcome through piecewise constant functions as shown in Figure 4.3. Scenarios VII-XI are modifications of Scenario VI: VII and VIII introduce two-way and three-way multiplicative interactions; IX relates the outcome to higher-order interactions and indicator functions. The rule models in Scenarios VI-IX were all mis-specified. Scenario X adds 500 noise features to the rule model while the true data-generating model is the same as Scenario VI; Scenario XI

adds 500 noise features and five prognostic variables to the true data-generating model and as predictors to the rule model.

4.6.2 Simulation Results

Figure 4.4 shows the mean ABR across data realizations for a binary outcome, as a function of training set sample size, for the treatment rules estimated by six different approaches: the three candidate methods in `DevTreatRules` (split-regression; OWL framework; and direct-interactions)[10], two naive alternatives (a policy of treating everyone and a policy of treating no one), and the optimal treatment rule computed with knowledge of the data-generating mechanism. We computed ABR on independent test sets with sample size 10000.

Figure 4.4 reveals a few trends. For the scenarios with only a handful of predictors (Scenarios I-IV and VI-IX), the performance of all three approaches in our simulations was almost identical, regardless of the degree of model mis-specification considered. With the 50 noise predictors added in Scenario V, OWL framework has a noticeably lower ABR than split-regression and direct-interactions, and the gap does not close with larger sample size. With the 500 noise predictors in Scenarios X and XI, the OWL framework actually reverted to the naive strategy of treating no-one, even for the largest sample sizes considered. Direct-interactions appears to yield higher ABRs in Scenario X, but loses this edge when the prognostic features are added in Scenario XI. We note the noise-added Scenarios V, X, and XI requires much larger sample sizes than the others in general.

As we might expect, Figure 4.4 also shows that in the fundamentally mis-specified settings (based on piecewise-constant links) of Scenario VI-IX, all three methods fall well short of the optimal ABR. Given the common reasoning that an indirect approach (such as split-regression) should be more sensitive to model mis-specification than a direct approach (such as OWL framework and direct-interactions), it is perhaps surprising that split-regression performs almost identically to those direct approaches. In Scenarios X and XI, which added 500

[10]As detailed in Section 4.7.1, the `DynTxRegmine` implementation of OWL unfortunately did not run successfully for our simulation scenarios so we could not include it in the comparison.

noise features, OWL framework again performs considerably worse than direct-interactions and split-regression.

In Figure 4.5, we present the mean ABRs for a continuous response Y using $(\mu, \sigma) = (50, 1)$ from the data-generating mechanism in (4.33). As elaborated on in Section 4.7.2, we found during the course of our simulation study that OWL framework and direct-interactions performed substantially worse in these scenarios without the following additional "tuning": for OWL framework, shifting Y to have a minimum of slightly above 0; for direct-interactions, mean-centering Y and, when using penalized regression, excluding the coefficient on the treatment variable from the penalty function. In view of this, the results presented in Figure 4.5 were produced after performing this additional tuning for OWL framework and direct-interactions to put them in the best possible light (as far as we could ascertain). The results with continuous outcome in Figure 4.5 are quite similar to the results with binary outcome in Figure 4.4, except that direct-interactions does not outperform split-regression in Scenario X.

In Appendix C.5, we plot the mean difference in ABR evaluated in training set and evaluated in the independent test set for each of the three approaches as an estimate of the methods' optimism bias (Harrell Jr, 2015) in this simulation setting. Direct-interactions appears to have the largest optimism bias for small and intermediate sample sizes in the high-dimensional Scenarios X as well as for small sample sizes in the noise-added Scenario V, for both continuous and binary outcomes.

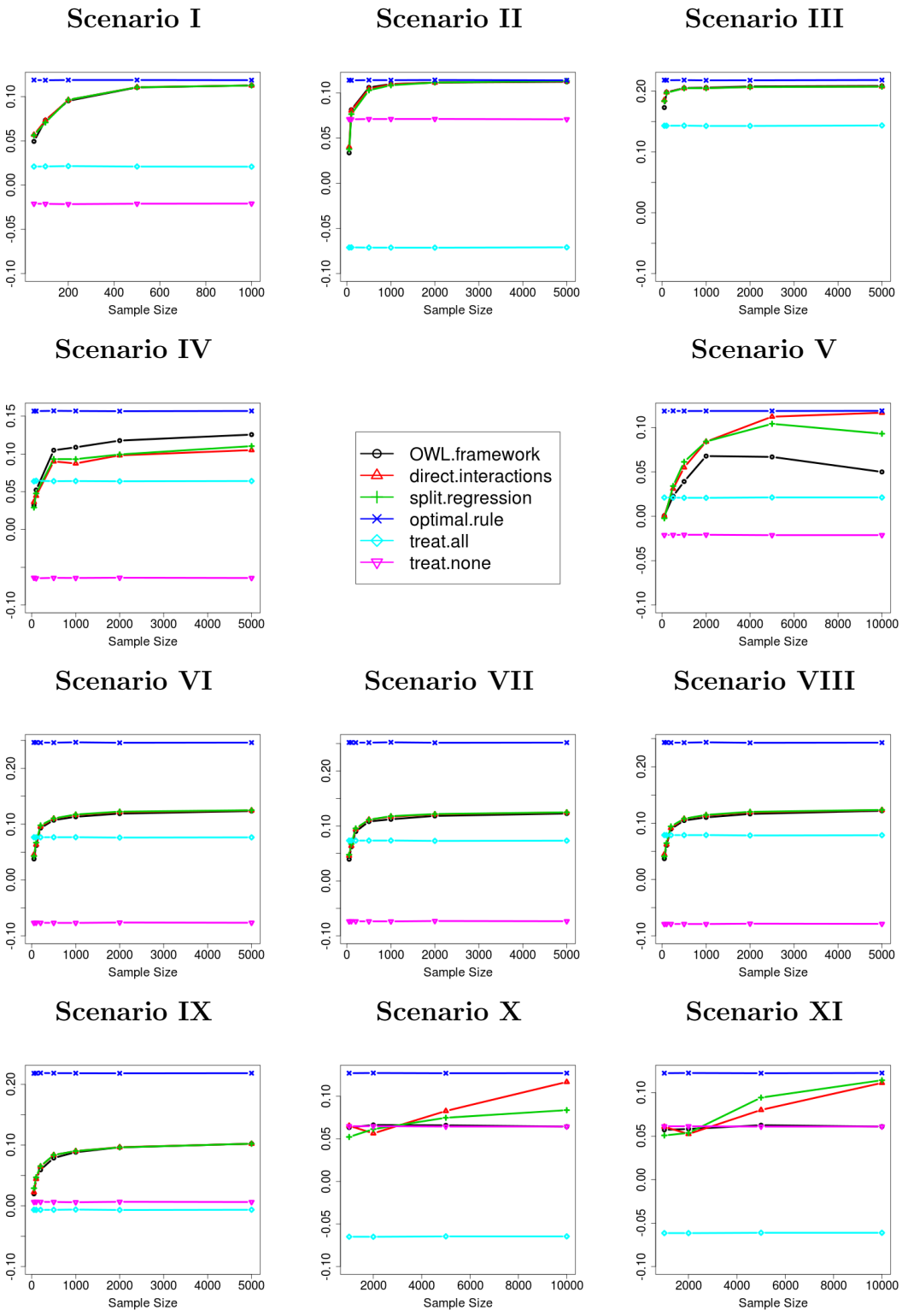


Figure 4.4: Mean ABR (over 50 replications) with binary Y

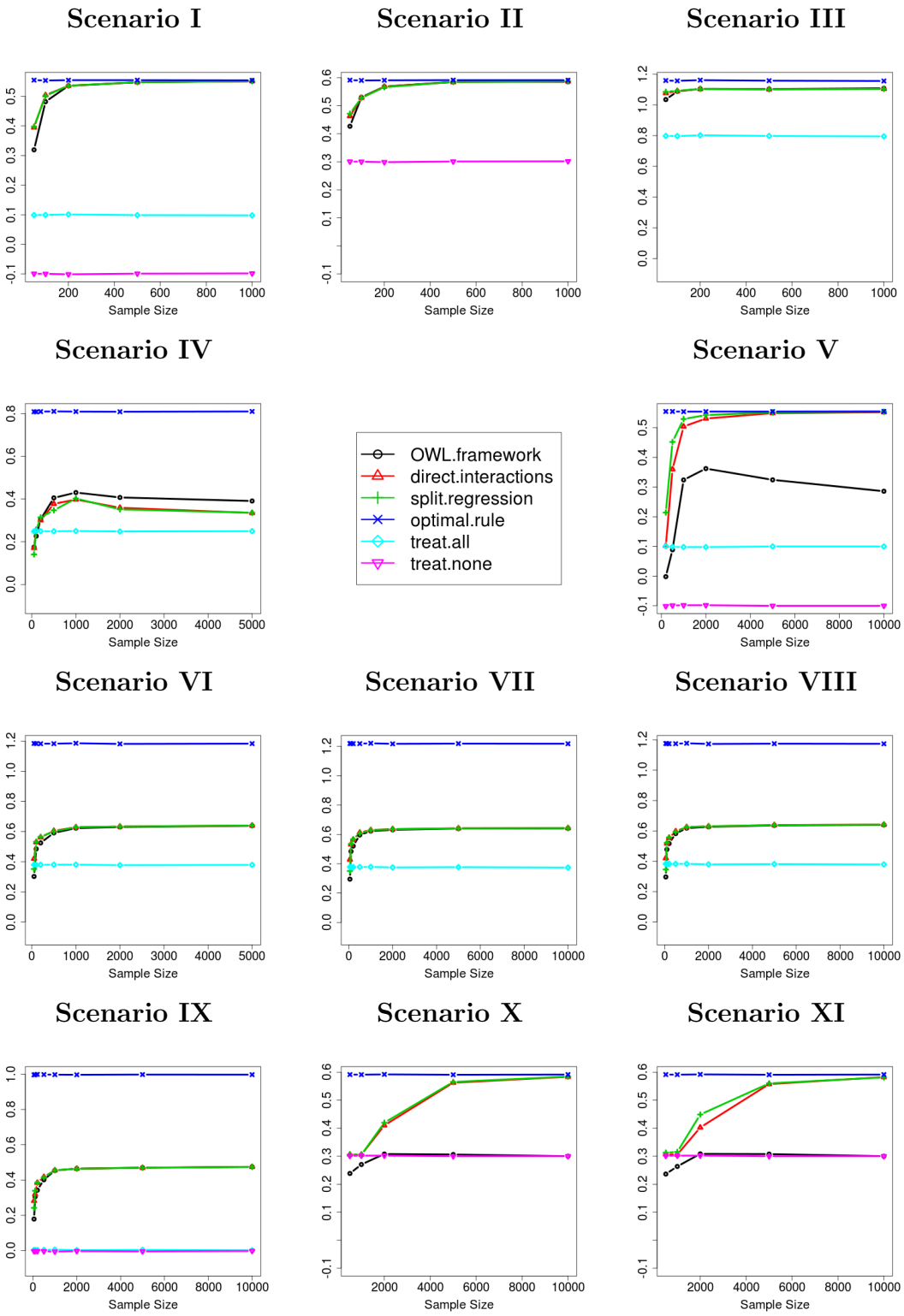


Figure 4.5: Mean ABR (over 50 replications) with continuous Y, $(\mu, \sigma) = (50, 1)$

4.7 Computation with `DevTreatRules`

4.7.1 Expanding `DevTreatRules` *OWL/OWL Framework and Direct-Interactions*

We introduced the `R` package `DevTreatRules` in Chapter 3, where it was the software implementation for our preferred split-regression approach to principled development and evaluation of treatment rules. In this chapter, we leverage the modeling flexibility in that principled framework to substantially expand `DevTreatRules` in the following ways:

1. `DevTreatRules` now supports the flexible interpretation of OWL framework in Zhao et al. (2012) where any classification procedure that uses observation weights is a sensible choice to estimate the solution to the fundamental target in (4.3). To do this, users can specify the argument `prediction.approach='OWL.framework'` in the `BuildRule()` function.[11]

2. `DevTreatRules` now supports the original specification of OWL (i.e. with a weighted SVM estimation procedure) in Zhao et al. (2012) by calling its official implementation in the `DynTxRegime` package. Users can select OWL to develop their treatment rule by specifying the argument `prediction.approach='OWL'` in the `BuildRule()` function.

    - We allow users to use ridge regression or lasso regression (rather than only a generalized linear model (GLM), as required in `DynTxRegime`) to estimate the propensity score.[12]
    - Empirically, we have found that, when performing the essential step of model-tuning in OWL with `DynTxRegime` (by specifying values for the `lambdas` and

[11] `DevTreatRules` currently supports linear regression, logistic regression, and their corresponding ridge and lasso penalized counterparts for its estimation stages.

[12] We do this by simply re-coding the outcome variable as $Y^* = Y \cdot \left(\frac{T_i\hat{\pi}^* + (1-T_i)\hat{\pi}^*}{T_i\hat{\pi} + (1-T_i)\hat{\pi}}\right)$, where $\hat{\pi}^*$ is the more flexible estimate of propensity score and $\hat{\pi}$ is the estimate from a GLM (which is supported by `DynTxRegmine`). This leads to an observation weight of in (4.3) of $\frac{Y}{T_i\hat{\pi}^* + (1-T_i)\hat{\pi}^*}$, rather than $\frac{Y}{T_i\hat{\pi} + (1-T_i)\hat{\pi}}$. This re-coding actually accommodates any method for estimating $\hat{\pi}^*$, but ridge and lasso are what `DevTreatRules` itself currently supports.

`cvFolds` arguments for the `owl()` function, whose defaults are surprisingly set to only fit OWL once using a fixed lambda value of 2), there are several computational issues that do not arise with OWL framework, including: computational infeasibility for n of about 20,000 or greater even with small $p << n$ (often requiring more than 8GB of RAM to store intermediate objects and run-times of over an hour if it does complete successfully); recurring optimization errors occurring in simulation scenarios with n as small as 500 and p as small as 2; and inability to work with $p = 1$. For these reasons, we could not use the OWL implementation in `DynTxRegime` for our simulation study or data example and used only the broader OWL framework definition (again specified in our `BuildRule()` function with `prediction.approach='OWL.framework'`).

  - Nonetheless, users of `DevTreatRules` can still try to use the `DynTxRegime` implementation of OWL (with model tuning through the `cvFolds` and `lambdas` arguments performed by default) by calling `BuildRule()` with the argument `prediction.approach='OWL'`.

3. `DevTreatRules` implements the direct-interactions approach of Chen et al. (2017) and Tian et al. (2014), which does not have an official `R` package, by calling `BuildRule()` with the argument `prediction.approach='direct.interactions'`.

4. `DevTreatRules` uses sensible defaults to incorporate the subtle but critical issues related to model tuning (discussed in Section 4.7.2) for OWL framework and direct-interactions that we came across during our simulation study.

That is, the `DevTreatRules` package implements our principled development and evaluation framework, while supporting three qualitatively distinct approaches for estimating the treatment rule: the indirect split-regression method we described in Chapter 3, as well as the direct methods of OWL/OWL framework and direct-interactions. As a result, `DevTreatRules` allows practitioners to simply see which of the three possible approaches seems to yield the

best treatment rule for their particular observational dataset by:

1. Partitioning their observational data into independent development/validation/evaluation subsets (e.g. using the `SplitData()` function from `DevTreatRules` with `n.sets=3`, the default, specified).

2. On the development dataset, developing treatment rules with all the approaches using `BuildRule()` with the `prediction.approach` argument set to one of `'OWL'`, `'OWL.framework'`, `'split.regression'`, `'direct.interactions'`, and additionally for different choices of underlying regression models by setting the `rule.method` and `propensity.method` arguments to combinations of `'glm.regression'`, `'lasso'`, `'ridge'`.

3. On the validation dataset, estimating ATEs and ABR of each developed treatment rule using the `EvaluateRule()` function.

4. Locking in the treatment rule(s) that appeared the most promising on the validation subset, and obtaining/reporting a trustworthy estimate of performance in a future clinical setting from the same patient population by using `EvaluateRule()` on the evaluation dataset.

More detail on `DevTreatRules` is available in the package documentation and in the accompanying vignette.

4.7.2 Computational Considerations for Indirect vs. Direct Methods

Recall the distinct objectives of an indirect method for estimating a treatment rule (predicting outcome as a function of patient characteristics under each possible treatment option such that a loss function is minimized) and a direct method for estimating a treatment rule (assigning a treatment option to each individual such that an estimate of population benefit is maximized).

The indirect method of split-regression requires no additional model tuning or data formatting beyond what is used to fit the underlying prediction method. Formatting data and performing model tuning for those prediction methods are typically well-studied problems where rules-of-thumb have been established by the research community. With the lasso, for example, it is generally recommended to use 10-fold cross-validation to choose the penalty parameter and to standardize each predictor to have mean 0 and variance 1 (Hastie et al. (2008); Friedman et al. (2009)). The indirect method of split-regression using lasso as the underlying prediction method can thus be carried out in a trustworthy manner by simply predicting outcome using patient characteristics (separately in each treatment group) in accordance with those recommendations.

On the other hand, we found during our simulation study that the direct methods of OWL framework and direct-interactions do not achieve satisfactory performance with a continuous outcome without performing additional steps beyond what would be needed to use split-regression with the same underlying prediction method.

4.7.3 Computational Considerations for OWL Framework

In Figure 4.6, we plot the mean ABR of OWL framework (in black circles) and split-regression (red triangles) as compared to the optimal rule (green crosses). The right panel of Figure 4.6 merely shifts Y by its 99.99% of its minimum value so that the new minimum is just slightly larger than zero[13] while the left panel does not modify Y and does not match the ABR of split-regression and the optimal rule until a sample size of 1000.

[13]Since Y is the numerator of the observation weights in the OWL framework, we are prevented from larger shifts Y (e.g. mean-centering) that would create negative values.

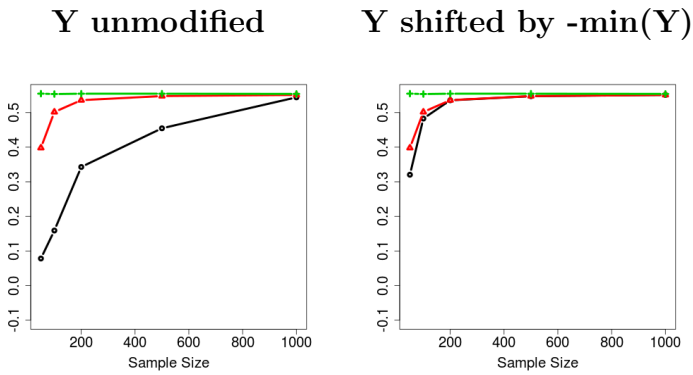


Figure 4.6: Mean ABR of OWL Framework With Shifts in Y (Simulation Scenario I). OWL framework in black circles, split-regression in red triangles, optimal rule in green crosses

Figure 4.7 tells a similar story for the "noise added" Scenario V considered in Section 4.6: Without shifting Y, the mean ABR from OWL framework does not exceed 0.1, while shifting Y yields a mean ABR of about 0.30 for the largest sample sizes considered (5000 and 10000 here). In contrast to the improvement in Scenario I, however, the performance of OWL framework after shifting Y, while still a substantial improvement, remains well short of the optimal mean ABR that is achieved by split-regression beginning at about $n = 2000$.

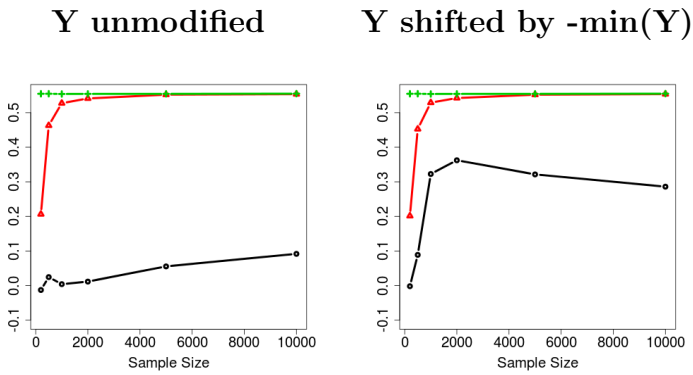


Figure 4.7: Mean ABR of OWL Framework With Shifts in Y (Simulation Scenario V). OWL framework in black circles, split-regression in red triangles, optimal rule in green crosses

4.7.4 Computational Considerations for Direct-Interactions

Our exploration of the direct-interactions approach revealed two computational factors that play a vital role in the efficacy of the method with a continuous outcome Y: (1) mean-centering Y and (2) not penalizing β_{T^*}, the coefficient on the -1/1 treatment indicator in (4.29), if the underlying prediction method performs variable shrinkage or selection (e.g. with the ridge/lasso).

Figure 4.8 shows the dramatic improvements in Scenario I that result from moving from an unmodified continuous Y with $\mu = 50$ (left panel, mean ABR below that of split-regression and optimal rule until $n = 1000$), to shifting by 99.99% of the minimum value (middle panel, mean ABR almost identical to that of split-regression), and finally to mean-centering Y (right panel, mean ABR identical to that of split-regression).

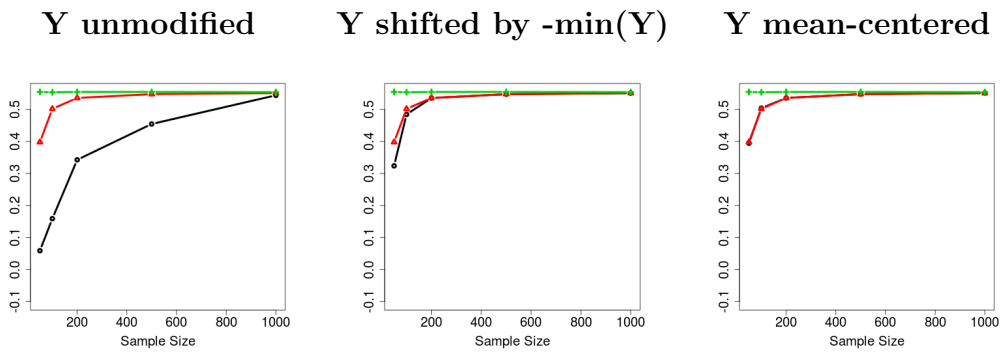


Figure 4.8: Mean ABR of Direct-Interactions With Mean-Centering of Y (Simulation Scenario I). Direct-interactions in black circles, split-regression in red triangles, optimal rule in green crosses

The top row of Figure 4.9 shows that moving from unmodified Y (left panel) to mean-centered Y (right panel), while an enormous improvement, does not quite match the performance of split-regression and the optimal rule while β_{T^*} is included in the lasso penalty function. After moving to an unpenalized β_{T^*} in the bottom row, however, the performance of direct-interactions with a mean-centered Y is almost indistinguishable from split-regression

at $n = 2000$, with both achieving the optimal value by $n = 5000$.

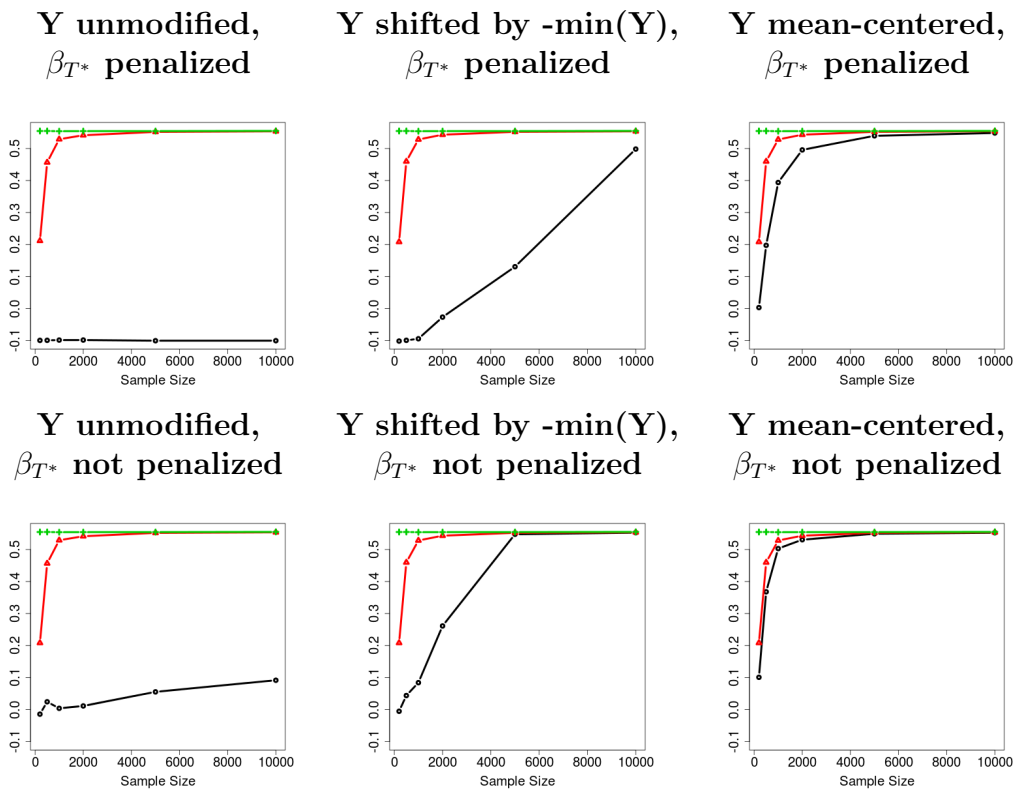


Figure 4.9: Mean ABR of Direct-Interactions With Mean-Centering of Y and Penalization of β_{T^*} (Simulation Scenario V). Direct-interactions in black circles, split-regression in red triangles, optimal rule in green crosses

4.8 Data Example

Please see Chapter 3, Appendix B.2, and Langer et al. (2003) for additional details about the WHI-OS dataset. Here we apply `DevTreatRules` to the WHI-OS dataset to form treatment rules (from the OWL framework, direct-interactions, and split-regression approaches) that recommend hormone therapy (HRT) to postmenopausal women if it appears to reduce probability of coronary heart disease (CHD) within 10 years, and, in a separate analysis, probability of breast cancer within 10 years. We used the "adjudicated" outcome variables as described in Curb et al. (2003).

As mentioned in Chapter 3, 17.6% of observations in the initial dataset had a missing value for at least one variable included as a predictor in either our rule model or propensity score model. In view of this, we used the IPW-based approach as discussed in Seaman and White (2013) to account for missingness of individual characteristics. We share all the code used to perform the data example, including the IPW adjustment for missingness, at `github.com/jhroth/data-example-elucidating-owl`.

Table 4.2 presents estimates of ATE and ABR in the validation set for the most promising two models for each outcome: two OWL framework models for the CHD outcome and two split-regression models for the breast cancer outcome. Since the validation dataset informed our model selection (i.e. we chose these four model specifications because they appeared best on the validation set), the ATE and ABR estimates from the validation set itself are not trustworthy estimates for performance of the rule in an independent sample (e.g. a future clinical setting) – instead we must estimate performance on an independent evaluation set that did not influence model selection, which we do in Table 4.3.

	Positives	**Negatives**	**ATE in Positives**	**ATE in Negatives**	**ABR**
Outcome: No coronary heart disease after 10 years					
OWL framework (logistic/logistic)	4398	14904	-0.001	-0.042	0.033
Treat no one (logistic/NA)	0	19302	NA	-0.029	0.029
Outcome: No breast cancer after 10 years					
Split regression (logistic/logistic)	3692	15610	0.021	-0.053	0.047
Split regression (ridge/lasso)	3544	15758	0.013	-0.051	0.044
Treat no one (logistic/NA)	0	19302	NA	-0.045	0.045

Table 4.2: Summary of Selected Rules in the Validation Set. The selected propensity method/rule method are in parentheses

Table 4.3 presents the estimated ATEs and ABR in the evaluation set, which do serve as trustworthy estimates of rule performance in future clinical settings that observe individuals from this same population. First we note that the bootstrap-based 95% CIs for the ATEs among the treated subgroups all contain 0, so unfortunately we lack evidence that any of the four rules identifies a subpopulation for which treatment is beneficial.

	Positives	Negatives	ATE in Positives	ATE in Negatives	ABR
Outcome: No breast cancer after 10 years					
Split regression (ridge/lasso)	3408	15894	0.015 (-0.027, 0.068)	-0.045 (-0.067, -0.031)	0.04
Split regression (logistic/logistic)	3569	15733	0.002 (-0.049, 0.052)	-0.048 (-0.07, -0.035)	0.04
Outcome: No coronary heart disease after 10 years					
OWL.framework (logistic/logistic)	4477	14825	0.004 (-0.028, 0.036)	-0.033 (-0.057, -0.016)	0.026
Treat no one (logistic/NA)	0	19302	NA	-0.023	0.023

Table 4.3: Summary of Selected Rules in the Evaluation Set. The selected propensity method/rule method are in parentheses. 95% CIs are based on the basic bootstrap.

Table 4.3 shows that, for the outcome of 10-year breast cancer incidence, the split-regression rule using ridge/lasso as the propensity/rule models would recommend HRT to 18% of individuals and, in that treated subgroup, we estimate that treatment decreases the probability of breast cancer within 10 years by an estimated 1.5 percentage points. The split-regression rule using logistic regression for both propensity/rule models recommends treatment for a larger number of individuals than the rule based on ridge/lasso and, in that treated subgroup, the estimated ATE is smaller, so the rule based on ridge/lasso appears to be a superior option.

For the outcome of 10-year CHD, OWL framework using logistic regression recommends HRT to 23% of individuals and, within that treated subgroup, treatment decreases the estimated probability of breast cancer within 10 years by 0.4 percentage points. However, this estimated ATE is quite close to 0 and likely falls short of clinical significance and, since the associated 95% CI contains 0, falls short of statistical significance as well.

4.9 Discussion

4.9.1 Take-Aways from Simulation Study

Our simulation study may suggest a few things in keep in mind when applying the indirect tool of split-regression and the direct tools of OWL framework and direct-interactions to future observational datasets. Three points we consider essential (and we implement as defaults in `DevTreatRules`) are:

1. When using OWL framework with a continuous outcome, it is critical to shift the outcome by a scalar such that its minimum value is just above 0.

2. When using direct-interactions with a continuous outcome, it is critical to mean-center the outcome variable.

3. When using direct-interactions with an underlying method of penalized regression, it is important to exclude the coefficient of the treatment indicator from the penalty function.

Although we advocate further study of this behavior in the future, our understanding is that mean-centering the outcome for direct-interactions is needed due to the method's lack of an intercept. The interpretation of the OWL framework as a special case of direct-interactions (with a specific choice of loss function, as shown in Chen et al. (2017)), indicates that a similar shifting of Y should improve the performance of the OWL framework. Since Y is the numerator of the observation weights in the OWL framework, it cannot take on negative values and thus mean-centering is not permissible. As seen in Figures 4.8 and 4.9, however, the shifting of Y to have a minimum of zero provides direct-interactions with nearly the same improvement as mean-centering Y, which we also believe is due to the method's lack of an intercept. And since OWL framework is a special case of direct-interactions, we believe the dramatic improvement of OWL framework after shifting Y (see Figure 4.6, for example) is also due to that absence of an intercept in that framework.

With less confidence, we also recommend a few additional rules-of-thumb that can be applied in particularly challenging settings where data-splitting must be limited to a development/evaluation partition rather than a development/validation/evaluation partition, and thus simply estimating treatment rules with all three methods using `DevTreatRules` and comparing the rules' performance on a validation set before moving to an evaluation set is infeasible (perhaps due to limited sample size):

1. With only a handful of predictors, all three methods may perform quite similarly.

2. With more than a few predictors such that variable shrinkage/selection is desirable, split-regression and direct-interactions may be better options than OWL framework.

In general, however, we recommend not drawing any firm conclusions based on our simulation study which, as is the case for any simulation study, is capable of revealing information only through its particular (and unrealistic) data-generating context. Rather, we advocate partitioning an observational dataset into independent development/validation/evaluation subsets and trying out *all three* of OWL framework, split-regression, and direct-interactions with different choices of underlying regression models. Specifically, we reiterate our recommendation, which we applied to the WHI-OS data using the `R` scripts shared at `github.com/jhroth/data-example-elucidating-owl`, to:

1. Partition the observational data into independent development/validation/evaluation subsets (using the `SplitData()` function with `n.sets=3`, the default, specified),

2. Decide whether each patient characteristic in the study potentially informs treatment assignment in the current study (specified with the `names.influencing.treatment` argument in `BuildRule()`) and whether it will be available in future clinical situations (specified with the `names.influencing.rule` argument in `BuildRule()`).

3. Develop treatment rules with all three approaches using the function `BuildRule()` with the `prediction.approach` argument set to one of `'OWL.framework'`, `'split.regression'`, `'direct.interactions'`, and the arguments `propensity.method` and `rule.method` set to different combinations of `'glm.regression'`, `'lasso'`, `'ridge'`.

4. Evaluate each developed treatment rule on the validation subset using the `EvaluateRule()` function.

5. Lock in the treatment rule (or perhaps a couple) that appears most promising on the evaluation subset, and obtain a trustworthy estimate of ATE and ABR in future clinical settings by using `EvaluateRule()` in the evaluation dataset.

6. Report all the estimates on the evaluation dataset from item 5 to inform future study.

The vignette accompanying the `DevTreatRules` package also shows how the `CompareRulesOnValidation()` function can automate the above steps 3 and 4.

4.9.2 Precision Variables: Friend or Foe?

Perhaps the defining characteristic of the direct-interactions approach, as highlighted in Section 4.5.1, is the absence of main effect terms for rule inputs $\mathbf{R}$, which researchers commonly interpret as an advantage of the method. If the variables in $\mathbf{R}$ are associated with the outcome Y, however, their corresponding main effect terms play a critical role as *precision variables* (Emerson, 2014) whose inclusion reduces estimation variability in a correctly specified generalized linear model, although the coefficient estimates themselves would be unbiased even if the precision variables were excluded. If the rule inputs $\mathbf{R}$ are not associated with Y (i.e. they are not precision variables), then we agree that direct-interactions – or any direct method that excludes main effects (including OWL which as shown in Chen et al. (2017) is a special case of direct-interactions) – may offer advantages compared to indirect approaches that include the main effect terms for the unnecessary predictors. However, if components of $\mathbf{R}$ are precision variables, we do not view the exclusion of their main effect terms as desirable.

We can see in Figure 4.4 that in Simulation Scenario X – where $\mathbf{R}$ consists of one variable that affects treatment response but does not affect the binary Y and 500 noise variables that affect neither treatment response or Y – direct-interactions does indeed offer a substantial improvement over the indirect approach of split-regression, which includes main effects for the predictors in $\mathbf{R}$. In contrast, in Scenario XI where the one signal variable in $\mathbf{R}$ does

also affect Y (i.e. it is a precision variable), direct-interactions has lower ABR than split-regression, even though direct-interactions does correctly exclude the main effect terms for the 500 noise variables. Curiously, in Figure 4.5 where Y is continuous, we do not see an advantage for direct-interactions in Scenario X but we do see a disadvantage in Scenario XI when the main effect is added in. Going forward, we caution against interpreting the exclusion of main effect terms $\mathbf{R}$ by direct methods such as direct-interactions as uniformly beneficial if $\mathbf{R}$ may contain precision variables, since their exclusion may increase estimation variability. If scientific knowledge indicates that $\mathbf{R}$ is not associated with Y, however, we do agree that exclusion of main effect terms is an advantage of direct-interactions (and other direct approaches that exclude main effect terms), particularly in settings where the number of rule inputs $\mathbf{R}$ is large.

4.9.3 Recap

In this chapter, we began by highlighting the distinction between direct and indirect approaches for estimating a treatment rule. We focused on OWL as a representative of the direct approaches in the literature, since it is popular and its estimation target is a very convincing choice for treatment rule development but its corresponding estimation procedure is somewhat unintuitive. Split-regression from Chapter 3 was our representative indirect approach with a characteristically intuitive estimation procedure but a somewhat fuzzy justification for use in developing a treatment rule.

We offered a Bayesian interpretation to emphasize the straightforward connection between OWL and split-regression with a binary outcome. With a continuous outcome, we showed through an extension of the binary case that there is a more nuanced connection between OWL and split-regression: We can view each method as a tool for transforming an intuitive but computationally challenging target parameter involving complicated estimation/averaging of conditional densities into a much simpler minimization problem.

We also integrated OWL/OWL framework (and direct-interactions, another promising direct approach) into the principled methodological framework of Chapter 3 and the

`DevTreatRules` package. We feel `DevTreatRules` contributes to the field because it offers a principled approach for developing treatment rules (incorporating the subtleties related to variable definitions and data-splitting discussed in Chapter 3 and related to model tuning discussed in Section 4.7.4) for both indirect *and* direct approaches so that we can obtain reliable estimates of their performance in future clinical settings.

In a simulation study, we used `DevTreatRules` to develop treatment rules using all three of OWL framework, split-regression, and direct-interactions, and saw how their performance compared to the optimal rule and to rules from naive alternative approaches. Along the way, we identified three steps that were absolutely essential for OWL framework and direct-interactions to achieve suitable performance with a continuous outcome: for OWL framework, shifting the outcome to have a minimum of just above 0; for direct-interactions, mean-centering the outcome and, if performing penalized regression, excluding the coefficient on the treatment variable from the penalty function. We also listed a few rules-of-thumb to keep in mind when choosing between the three approaches in challenging settings where performing model validation is infeasible (e.g. small sample size) and only a development/evaluation partition is possible.

Our stronger recommendation (which is supported by the design of `DevTreatRules` and described in an accompanying vignette) would be to form a development/validation/evaluation partitioning of the data, develop treatment rules with all three approaches and with different combinations of underlying prediction methods on the development dataset, compare estimated ATEs/ABRs on the validation set, and, based on that exploration, lock in the one or two most promising rules to carry forward to the evaluation set. We followed this recommendation in an example using WHI-OS data and share our code at `github.com/jhroth/data-example-elucidating-owl`. In that WHI-OS data example, we were unable to find a treatment rule that identified a treated subpopulation for which estimated ATE had a 95% CI completely above 0. In this situation (which unfortunately seems quite common in the field) our recommendation would simply be to not offer HRT in the population if the goal is to reduce either 10-year incidence of

breast cancer or 10-year incidence of CHD.

4.9.4 Future Work

More clarity on why mean-centering the outcome (for direct-interactions) or shifting the outcome to have a minimum of just above 0 (for OWL framework) leads to such dramatic improvements might reveal further nuances about the methods. There is also ample opportunity for future research to expand on this work through additions to the `DevTreatRules` package. One pathway would be adding flexibility by offering more choices for the underlying prediction models. Another direction might place greater emphasis on the evaluation of treatment rules (while, as the package name suggests, we instead emphasized development) and provide alternative estimators of the key quantities of ATE and ABR, perhaps based on the *doubly robust* or *augmented* analogs of IPW (Robins et al., 1994; Lunceford and Davidian, 2004). We greatly welcome such contributions and hope that by sharing our entire implementation in an `R` package and by providing the code to reproduce our simulations and data example, we are helping to facilitate future work in both research and clinical settings alike.

4.10 Acknowledgements

Thank you to Holly Janes and Susanne May for helpful comments and suggestions.

BIBLIOGRAPHY

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Ballman, K. V. (2015). Biomarker: predictive or prognostic? *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 33(33):3968–3971.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cai, T., Tian, L., Wong, P. H., and Wei, L. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282.

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

Chen, S., Tian, L., Cai, T., and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209.

Ciarleglio, A., Petkova, E., Ogden, R. T., and Tarpey, T. (2015). Treatment decisions based on scalar and functional baseline covariates. *Biometrics*, 71(4):884–894.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405.

Curb, J. D., Mctiernan, A., Heckbert, S. R., Kooperberg, C., Stanford, J., Nevitt, M., Johnson, K. C., Proulx-Burns, L., Pastore, L., Criqui, M., et al. (2003). Outcomes ascertainment and adjudication methods in the women's health initiative. *Annals of epidemiology*, 13(9):S122–S128.

Diamond, S., Chu, E., and Boyd, S. (2014). CVXPY: A Python-embedded modeling language for convex optimization, version 0.2. http://cvxpy.org/.

Edgar, R., Domrachev, M., and Lash, A. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Emerson, S. S. (2014). Lecture 10 in applied biostatistics: Adjustment for confounders, precision. http://www.emersonstatistics.com/courses/formal/b518_2014/b518L10-2014-02-19-4.pdf.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

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

Gianni, L., Eiermann, W., Semiglazov, V., Manikhas, A., Lluch, A., Tjulandin, S., Zambetti, M., Vazquez, F., Byakhow, M., Lichinitser, M., et al. (2010). Neoadjuvant chemotherapy with trastuzumab followed by adjuvant trastuzumab versus neoadjuvant chemotherapy alone, in patients with her2-positive locally advanced breast cancer (the noah trial): a randomised controlled superiority trial with a parallel her2-negative cohort. *The Lancet*, 375(9712):377–384.

Haque, R., Ahmed, S. A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., Bernstein, L., Enger, S. M., and Press, M. F. (2012). Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiology Biomarkers & Prevention*, 21(10):1848–1855.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, New York, 2nd edition.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Hudis, C. A. (2007). Trastuzumabmechanism of action and use in clinical practice. *New England Journal of Medicine*, 357(1):39–51.

Joensuu, H., Kellokumpu-Lehtinen, P.-L., Bono, P., Alanko, T., Kataja, V., Asola, R., Utriainen, T., Kokko, R., Hemminki, A., Tarkkanen, M., et al. (2006). Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. *New England Journal of Medicine*, 354(8):809–820.

Johnson, N. (2013). A dynamic programming algorithm for the fused lasso and l_0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260.

Kang, C., Janes, H., and Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70(3):695–707.

Karnofsky, D. A. (1949). The clinical evaluation of chemotherapeutic agents in cancer. *Evaluation of chemotherapeutic agents*.

Kennedy, E. H. (2015). Semiparametric theory and empirical processes in causal inference. *arXiv preprint arXiv:1510.04740*.

Kim, C., Tang, G., Pogue-Geile, K., and et al., J. C. (2011). Estrogen receptor (esr1) mrna expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor-positive breast cancer. *Journal of Clinical Oncology*, 29(31):4160–4167.

Kim, S., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360.

Langer, R. D., White, E., Lewis, C. E., Kotchen, J. M., Hendrix, S. L., and Trevisan, M. (2003). The women's health initiative observational study: baseline characteristics of participants and reliability of baseline measures. *Annals of epidemiology*, 13(9):S107–S121.

Lipkovich, I., Dmitrienko, A., and B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196.

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

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

Mandrekar, S. J. and Sargent, D. J. (2009). Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology*, 27(24):4027–4034.

McKeague, I. W. and Qian, M. (2014). Estimation of treatment policies based on functional predictors. *Statistica Sinica*, 24(3):1461.

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

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2016a). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2016b). SCS: Splitting conic solver, version 1.2.6. https://github.com/cvxgrp/scs.

Oken, M. M., Creech, R. H., Tormey, D. C., Horton, J., Davis, T. E., McFadden, E. T., and Carbone, P. P. (1982). Toxicity and response criteria of the eastern cooperative oncology group. *American journal of clinical oncology*, 5(6):649–656.

Oldenhuis, C., Oosting, D., Gietema, J., and Vries, E. D. (2008). Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer*, 4(7):946–953.

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).

Pan, G. and Wolfe, D. A. (1997). Test for qualitative interaction of clinical significance. *Statistics in Medicine*, 16(14):1645–1652.

Parikh, N., Boyd, S. P., et al. (2014). Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239.

Parker, J. S., Mullins, M., and Cheang, M. e. a. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Peto, R. (1982). Statistical aspects of clinical trials. *Treatment of Cancer*, pages 867–871.

Piantadosi, S. and Gail, M. H. (1993). A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine*, 12(13):1239–1248.

Prat, A. and Bianchini, G. e. a. (2014). Research-based pam50 subtype predictor identifies higher responses and improved survival outcomes in her2-positive breast cancer in the noah study. *Clinical Cancer Research*, 20(2):511–521.

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

Ramdas, A. and Tibshirani, R. J. (2014). Fast and flexible admm algorithms for trend filtering. *arXiv*, 1406(2082):1–19.

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

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

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

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.

Roth, J. and Simon, N. (2017). A framework for estimating and testing qualitative interactions with applications to predictive biomarkers. *Biostatistics*.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108.

Tibshirani, R. J. (2013). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

Varadhan, R. and Seeger, J. D. (2013). *Estimation and reporting of heterogeneity of treatment effects*. Agency for Healthcare Research and Quality (US).

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., and Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, 13(2):273–277.

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

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

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

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

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315.

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

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

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168.

Appendix A

APPENDIX FOR CHAPTER 2

A.1 Generalized Gradient Descent with Logistic Loss

Our goal is to find fitted values $\hat{\theta} \in \mathbb{R}^{n_j}$ in treatment group j that maximize the penalized log-likelihood

$$\hat{\theta} = \arg\max_{\theta} \left[y^T\theta - \sum_{i=1}^{n} \log\left(1 + e^{\theta_i}\right) \right] - \lambda P(\theta)$$

or, equivalently, minimize the negative penalized log-likelihood

$$\hat{\theta} = \arg\min_{\theta} \underbrace{\left[\sum_{i=1}^{n} \log\left(1 + e^{\theta_i}\right) - y^T\theta \right]}_{\ell(\theta)} + \underbrace{\lambda P(\theta)}_{h(\theta)}. \tag{A.1}$$

Because equation (A.1) is a minimization over the sum of a smooth function $\ell(\theta)$ and a non-smooth function $h(\theta)$, we appeal to generalized gradient descent to obtain the solution.

We begin by writing the quadratic approximation of $\ell(\theta)$ about the point θ_0:

$$\begin{aligned} \ell(\theta) + h(\theta) &\approx \ell(\theta_0) + (\theta - \theta_0)^T \nabla \ell(\theta_0) + (\theta - \theta_0)^T \nabla^2 \ell(\theta_0)(\theta - \theta_0) + h(\theta) \\ &\leq \ell(\theta_0) + (\theta - \theta_0)^T \nabla \ell(\theta_0) + \frac{L}{2} ||\theta - \theta_0||_2^2 + h(\theta), \end{aligned} \tag{A.2}$$

where $L > 0$ is such that $LI_n \succ \nabla^2 \ell(\theta_0)$ and $L = 1/4$ is a valid choice since it is an upper bound for any eigenvalue of $\nabla^2 \ell(\theta_0)$. Without the penalty term we could minimize (A.2) by taking standard gradient steps of size $1/L$.

Note that completing the square

$$\frac{L}{2}\left|\left|\theta-\theta_0+\frac{1}{L}\nabla\ell(\theta_0)\right|\right|_2^2=\frac{L}{2}\left[||\theta-\theta_0||_2^2+2(\theta-\theta_0)^T\frac{1}{L}\nabla\ell(\theta_0)+\frac{1}{L^2}||\nabla\ell(\theta_0)||_2^2\right]$$
$$=\frac{L}{2}||\theta-\theta_0||_2^2+(\theta-\theta_0)^T\nabla\ell(\theta_0)+\frac{1}{2L}||\nabla\ell(\theta_0)||_2^2$$

allows us to rewrite (A.2) as

$$\ell(\theta)+h(\theta)\approx\ell(\theta_0)+\frac{L}{2}\left|\left|\theta-\theta_0+\frac{1}{L}\nabla\ell(\theta_0)\right|\right|_2^2-\frac{1}{2L}||\nabla\ell(\theta_0)||_2^2+h(\theta).$$

Since the first and third terms do not depend on θ we can rewrite equation (A.1) as

$$\hat{\theta}=\arg\min_\theta\ell(\theta)+h(\theta)\approx\arg\min_\theta\frac{L}{2}\left|\left|\theta-\left(\theta_0-\frac{1}{L}\nabla\ell(\theta_0)\right)\right|\right|_2^2+\lambda P(\theta). \tag{A.3}$$

From here we can call a fused-lasso solver for least-squares loss.

A.2 Estimating False Discovery Rate in Data Example

In our data example, we are considering either a set of 9 candidate biomarkers (when $p=1$) or a set of 36 candidate biomarkers (when $p=2$). For the k-th candidate biomarker, let Z_k be the observed test statistic and let $\{Z_k^{*b}\}_{b=1}^B$ be the permuted test statistics. Then, as suggested by Tusher et al. (2001), we estimate the false discovery rate (Benjamini and Hochberg, 1995) for the k-th candidate biomarker with

$$\text{FDR}_k^*=\frac{\{\#Z_1^{*b}\geq Z_k\}+\ldots+\#\{Z_K^{*b}\geq Z_k\}}{BK}, \tag{A.4}$$

for $k=1,\ldots,K$. We use $B=100$ permutations for each candidate and have either $K=9$ (when $p=1$) or $K=36$ (when $p=2$).

A.3 Conservativeness of Permutation Tests

Consider the simplified case wherein we have piecewise-constant regression functions with prespecified knots where the outcome is Gaussian with equal variance in both treatment arms. Here our fitted values for group j's responses in a particular region r between knots, $\vec{y}_{j,r} \overset{\text{iid}}{\sim} N(\mu_j, \sigma^2)$, are simply that group's sample mean $\overline{y}_{j,r} \sim N(\mu_j, \sigma^2/n_{j,r})$, where $j = 0, 1$. We explore the behavior of our permuted test statistics in this simplified case as a heuristic for thinking about their behavior when fitting with data adaptively chosen knots. Suppose we are testing $H_0 : \mu_0 \leq \mu_1$. In this simplified case the null hypothesis for permutation testing is $(\vec{y}_{0,r}, \vec{y}_{1,r}) \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ for fixed constant σ^2.

Lemma A.3.1 *For $\overline{y}_{0,r} \sim N(\mu_0, \sigma^2/n_{0,r})$, $\overline{y}_{1,r} \sim N(\mu_1, \sigma^2/n_{1,r})$ in any region r under $H_0 : \mu_0 \leq \mu_1$ with fixed $c \in \mathbb{R}^+$,*

$$P_{\mu_0 \leq \mu_1} \left(RSS_{0,r} - RSS_{A,r} > c \right)$$

is dependent on (μ_0, μ_1) only through $(\mu_0 - \mu_1)$ and is maximized at $\mu_0 = \mu_1$. Hence permutation testing in any region r is conservative.

Proof A.3.2 *Let $y_{i,k}$ be the response variable corresponding to the ith individual belonging to group k in a single region r, where $i = 1, \cdots, n_k$, $k = 0, 1$, and n_k is the total number of individuals belonging to group k in region r. Then we can write*

$$RSS_{0,r} = \sum_{k=0}^{1} \sum_{i=1}^{n_k} (y_{i,k} - \overline{y}_k)^2 \cdot 1_{[\overline{y}_0 \leq \overline{y}_1]} + \sum_{k=0}^{1} \sum_{=1}^{n_k} (y_{i,k} - \overline{y})^2 \cdot 1_{[\overline{y}_0 > \overline{y}_1]}$$

$$RSS_{A,r} = \sum_{k=0}^{1} \sum_{i=1}^{n_k} (y_{i,k} - \overline{y}_k)^2$$

so, if $\overline{y}_0 > \overline{y}_1$,

$$\begin{aligned}
RSS_{0,r} - RSS_{A,r} &= \sum_{k=0}^{1}\sum_{i=1}^{n_k}(y_{i,k} - \overline{y})^2 - \sum_{k=0}^{1}\sum_{i=1}^{n_k}(y_{i,k} - \overline{y}_k)^2 \\
&= \sum_k \sum_i \left(y_{i,k}^2 + (\overline{y})^2 - 2y_{i,k}\overline{y}\right) - \sum_k \sum_i \left(y_{i,k}^2 + (\overline{y}_k)^2 - 2y_{i,k}\overline{y}_k\right) \\
&= \sum_k \sum_i y_{i,k}^2 + (n_0 + n_1)(\overline{y})^2 - 2\overline{y}\sum_k \sum_i y_{i,k} \\
&- \sum_k \sum_i y_{i,k}^2 - \sum_k \sum_i (\overline{y}_k)^2 + 2\sum_k \overline{y}_k \sum_i y_{i,k} \\
&= (n_0 + n_1)(\overline{y})^2 - 2(n_0 + n_1)(\overline{y})^2 \\
&- n_0(\overline{y}_0)^2 - n_1(\overline{y}_1)^2 + 2(n_0(\overline{y}_o)^2 + n_1(\overline{y}_1)^2) \\
&= -(n_0 + n_1)(\overline{y})^2 + n_0(\overline{y}_0)^2 + n_1(\overline{y}_1)^2
\end{aligned}$$

which implies

$$\begin{aligned}
RSS_{0,r} - RSS_{A,r} &= -(n_0 + n_1)\left[\frac{n_0}{n_0 + n_1}\overline{y}_0 + \frac{n_1}{n_0 + n_1}\overline{y}_1\right]^2 + n_0(\overline{y}_0)^2 + n_1(\overline{y}_1)^2 \\
&= -\frac{1}{n_0 + n_1}\left[(n_0\overline{y}_0)^2 + (n_1\overline{y}_1)^2 + 2n_0n_1\overline{y}_0\overline{y}_1\right] + \frac{n_0(n_0 + n_1)(\overline{y}_0)^2 + n_1(n_0 + n_1)(\overline{y}_1)^2}{n_0 + n_1} \\
&= \frac{1}{n_0 + n_1}\left[(n_0\overline{y}_0)^2 + n_0n_1(\overline{y}_0)^2 + (n_1\overline{y}_1)^2 + n_0n_1(\overline{y}_1)^2 - (n_0\overline{y}_0)^2 - (n_1\overline{y}_1)^2 - 2n_0n_1\right] \\
&= \frac{n_0n_1}{n_0 + n_1}\left[(\overline{y}_0)^2 + (\overline{y}_1)^2 - 2\overline{y}_0\overline{y}_1\right] \\
&= \frac{n_0n_1(\overline{y}_0 - \overline{y}_1)^2}{n_0 + n_1}
\end{aligned}$$

So for a fixed $c \in \mathbb{R}^+$ *and assuming*

$$\begin{aligned}
z_0 &\equiv (\overline{y}_0 - \mu_0) \sim N\left(0, \sigma^2/n_0\right) \\
z_1 &\equiv (\overline{y}_1 - \mu_1) \sim N\left(0, \sigma^2/n_1\right)
\end{aligned}$$

we can write

$$\begin{aligned} P_{\mu_0\leq\mu_1}\left[RSS_{0,r}-RSS_{A,r}>c\right] &= P_{\mu_0\leq\mu_1}\left[\frac{n_0n_1(\overline{y}_0-\overline{y}_1)^2}{n_0+n_1}>c\right] \\ &= P_{\mu_0\leq\mu_1}\left[(\overline{y}_0-\overline{y}_1)>\sqrt{c\frac{n_0+n_1}{n_0n_1}}\right] \\ &\stackrel{d}{=} P_{\mu_0\leq\mu_1}\left[(z_0-z_1)+(\mu_0-\mu_1)>\sqrt{\frac{c(n_0+n_1)}{n_0n_1}}\right] \\ &= P_{\mu_0\leq\mu_1}\left[z_0-z_1>\sqrt{\frac{c(n_0+n_1)}{n_0n_1}}-(\mu_0-\mu_1)\right] \end{aligned} \tag{A.5}$$

Since (A.5) is monotone increasing in $(\mu_0-\mu_1)$, we conclude $P_{\mu_0\leq\mu_1}\left[RSS_{0,r}-RSS_{A,r}>c\right]$ is dependent on (μ_0,μ_1) only through $(\mu_0-\mu_1)$ and is maximized at $\mu_0=\mu_1$.

Lemma A.3.3 *For For $\overline{y}_{0,r}\sim N(\mu_0,\sigma^2/n_{0,r})$, $\overline{y}_{1,r}\sim N(\mu_1,\sigma^2/n_{1,r})$ in all regions $r=1,\cdots,m$ under $H_0:\mu_0\leq\mu_1$ with fixed $c\in\mathbb{R}^+$,*

$$P_{\mu_0\leq\mu_1}\left(T\equiv\frac{RSS_0-RSS_A}{RSS_A}=\sum_{r=1}^{m}\frac{RSS_{0,r}-RSS_{A,r}}{RSS_{A,r}}>c\right)$$

is maximized at $\mu_0=\mu_1$. Hence a permutation test under $H_0:\mu_0\leq\mu_1$ is conservative over the entire sample range.

Proof A.3.4 *Lemma A.3.1 tells us that for arbitrary region $r\in\{1,\cdots,m\}$,*

$$P_{\mu_0\leq\mu_1}\left(RSS_{0,r}-RSS_{A,r}>c\right)$$

is maximized at $\mu_0=\mu_1$. This implies

$$P_{\mu_0\leq\mu_1}\left(RSS_0-RSS_A=\sum_{r=1}^{m}RSS_{0,r}-RSS_{A,r}>c\right)$$

is also maximized at $\mu_0=\mu_1$. Since $(RSS_0-RSS_A)\perp RSS_A$, it follows that

$$P_{\mu_0 \leq \mu_1}\left(T \equiv \frac{RSS_0 - RSS_A}{RSS_A} = \sum_{r=1}^{m} \frac{RSS_{0,r} - RSS_{A,r}}{RSS_{A,r}} > c\right)$$

is maximized at $\mu_0 = \mu_0$.

A.4 Procedure Based on Linear Regression

The model is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 t_i + \beta_3 (x_i \times t_i) + \epsilon_i, \tag{A.6}$$

where $\epsilon_i \overset{\text{indep}}{\sim} (0, \sigma_i^2)$. We are testing

$$H_0 : \text{there do } \textit{not} \text{ exist } x, x', x'' \in \mathcal{X} \text{ such that } x = -\frac{\beta_2}{\beta_3}, x' < -\frac{\beta_2}{\beta_3}, x'' > -\frac{\beta_2}{\beta_3}.$$

So we reject H_0 IFF for all $\frac{-\beta_2}{\beta_3} \in CI\left(\frac{-\beta_2}{\beta_3}\right)$, there exist $x_A, x', x'' \in \mathcal{X}$ such that

$$x_A = -\frac{\beta_2}{\beta_3}, \quad x' < -\frac{\beta_2}{\beta_3}, \quad x'' > -\frac{\beta_2}{\beta_3}.$$

That is, we reject H_0 if $CI\left(\frac{-\beta_2}{\beta_3}\right) \subset \mathcal{X}$.

If $\beta_3 = 0$, then the fits for the treatment groups share the same slope but have different intercepts so qualitative interaction does not occur. Hence we decide to not reject H_0 if $0 \in CI(\beta_3)$. However, if $0 \notin CI(\beta_3)$ we proceed by noting that, under (A.6),

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\text{d}} N_4\left(0, \Sigma \equiv (\boldsymbol{X}^T \boldsymbol{X})^{-1} \left(\boldsymbol{X}^T \text{Var}(\boldsymbol{Y}) \boldsymbol{X}\right) (\boldsymbol{X}^T \boldsymbol{X})^{-1}\right),$$

so by the delta method with

$$g(\beta_0, \beta_1, \beta_2, \beta_3) = \frac{-\beta_2}{\beta_3}, \qquad \nabla g = \begin{bmatrix} 0 & -\frac{1}{\beta_3} & \frac{\beta_2}{\beta_3^2} & 0 \end{bmatrix}$$

we have

$$\nabla g \Sigma \nabla g^T = \begin{bmatrix} 0 & -\frac{1}{\beta_3} & \frac{\beta_2}{\beta_3^2} & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{bmatrix} \begin{bmatrix} 0 \\ -\frac{1}{\beta_3} \\ \frac{\beta_2}{\beta_3^2} \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{\Sigma_{21}}{\beta_3} + \frac{\beta_2 \Sigma_{31}}{\beta_3^2} & -\frac{\Sigma_{22}}{\beta_3} + \frac{\beta_2 \Sigma_{32}}{\beta_3^2} & -\frac{\Sigma_{23}}{\beta_3} + \frac{\beta_2 \Sigma_{33}}{\beta_3^2} & -\frac{\Sigma_{24}}{\beta_3} + \frac{\beta_2 \Sigma_{34}}{\beta_3^2} \end{bmatrix} \begin{bmatrix} 0 \\ -\frac{1}{\beta_3} \\ \frac{\beta_2}{\beta_3^2} \\ 0 \end{bmatrix},$$

$$= 0 + \frac{\Sigma_{22}}{\beta_3^2} - \frac{\beta_2 \Sigma_{32}}{\beta_3^3} + \frac{\beta_2^2 \Sigma_{33}}{\beta_3^4} - \frac{\beta_2 \Sigma_{23}}{\beta_3^3} + 0,$$

$$= \frac{\Sigma_{22}}{\beta_3^2} + \frac{\beta_2^2 \Sigma_{33}}{\beta_3^4} - \frac{\beta_2 \Sigma_{23} + \beta_2 \Sigma_{32}}{\beta_3^3},$$

$$= \frac{\Sigma_{22}}{\beta_3^2} + \frac{\beta_2^2 \Sigma_{33}}{\beta_3^4} - \frac{2\beta_2 \Sigma_{23}}{\beta_3^3},$$

so

$$\sqrt{n}\left(\frac{-\hat{\beta}_2}{\hat{\beta}_3} - \frac{\beta_2}{\beta_3}\right) \dot{\sim} N_1\left[0, \frac{\Sigma_{22}}{\beta_3^2} + \frac{\beta_2^2 \Sigma_{33}}{\beta_3^4} - \frac{2\beta_2 \Sigma_{23}}{\beta_3^3}\right]. \tag{A.7}$$

Thus an approximate Wald-type $100(1-\alpha)\%$ CI for $\frac{-\hat{\beta}_2}{\hat{\beta}_3}$ is given by

$$\frac{-\hat{\beta}_2}{\hat{\beta}_3} \pm t_{1-\frac{\alpha}{2}, n-4} \sqrt{\frac{\hat{\Sigma}_{22}}{\hat{\beta}_3^2} + \frac{\hat{\beta}_2^2 \hat{\Sigma}_{33}}{\hat{\beta}_3^4} - \frac{2\hat{\beta}_2 \hat{\Sigma}_{23}}{\hat{\beta}_3^3}} \tag{A.8}$$

where

$$\hat{\Sigma} = \left[(\boldsymbol{X}^T \boldsymbol{X})^{-1} \left[\boldsymbol{X}^T \mathrm{diag}(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})^2 \boldsymbol{X} \right] (\boldsymbol{X}^T \boldsymbol{X})^{-1} \right]$$

We conduct this hypothesis test with the following steps:

1. Compute $\hat{\beta}$ and $\hat{\Sigma}$

2. Form $CI(\beta_3)$, a Wald-type $100(1-\alpha)\%$ CI. If $0 \in CI(\beta_3)$, then do not reject H_0

3. If $0 \notin CI(\beta_3)$, then define

$$CI\left(\frac{-\beta_2}{\beta_3}\right) = \frac{-\hat{\beta}_2}{\hat{\beta}_3} \pm t_{1-\frac{\alpha}{2},n-4}\sqrt{\frac{\hat{\Sigma}_{22}}{\hat{\beta}_3^2} + \frac{\hat{\beta}_2^2\hat{\Sigma}_{33}}{\hat{\beta}_3^4} - \frac{2\hat{\beta}_2\hat{\Sigma}_{23}}{\hat{\beta}_3^3}}$$

and reject H_0 if $CI\left(\frac{-\beta_2}{\beta_3}\right) \subset \mathcal{X}$. Do not reject H_0 otherwise.

A.5 Comparison to Fused Lasso with Prespecified Knots

Figure A.1 plots the share of null hypotheses that are rejected as a function of SNR. Our test using the fused lasso is shown in blue filled-in squares, while the tests using piecewise-constant estimates of the regression functions with prespecified evenly-spaced knots are shown in orange crosses (3 knots), orange hollow squares (5 knots), and orange filled-in triangles (7 knots). Rejecting the null hypothesis in either of the first two rows is a Type I error, while it is the correct decision in the third row. We see that the fused-lasso-based test is generally competitive with the best fixed-knot smoother (which in practice will be unknown).

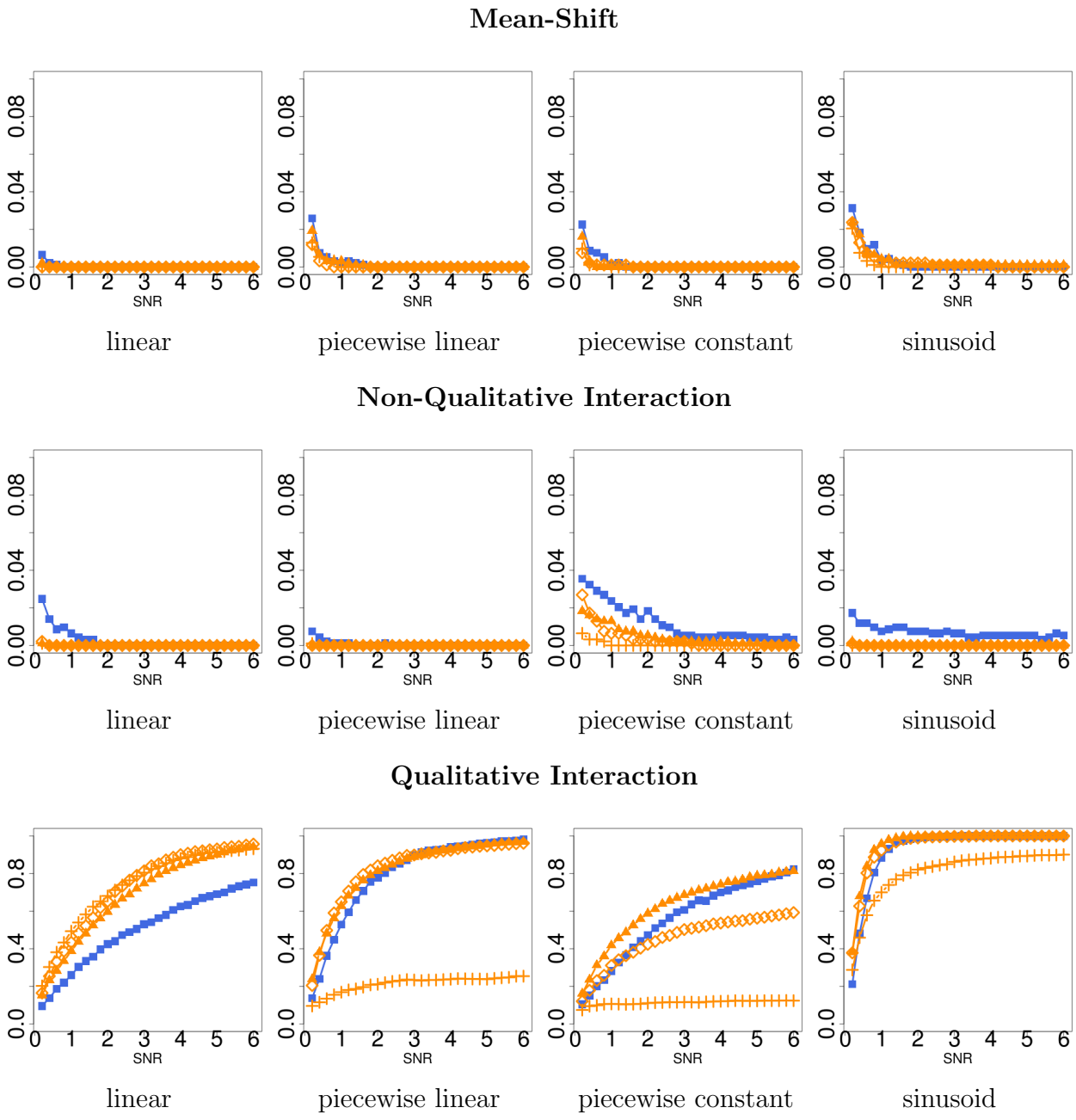


Figure A.1: Share of 1000 simulations with a p-value below 0.05, as a function of 30 SNR values across each of the 12 scenarios (permutation-based p^* from fused lasso with data adaptive knots in blue, permutation-based p-values from fused lasso with 3, 5, 7 prespecified evenly spaced knots in orange crosses, squares, and triangles.

Appendix B

APPENDIX FOR CHAPTER 3

B.1 Motivation for IPW Estimator of Average Treatment Effect

We are starting with

$$\Psi_{\mathbf{R}} \equiv \int_{\mathbf{C}^{\mathrm{T}}} \left\{ E(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1) - E(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=0) \right\} dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R}), \quad \text{(B.1)}$$

and, focusing on the $T=1$ group for a fixed $\mathbf{R}$, we can write

$$\begin{aligned}
\Psi(T=1 \mid \mathbf{R}) &= \int_{\mathbf{C}^{\mathrm{T}}} E(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1)\, dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R}), && \text{(B.2)}\\
&= \int_{\mathbf{C}^{\mathrm{T}}} \min_{f \in \mathcal{F}} \left\{ \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, dP^{\text{actual}}(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1) \right\} dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R}), && \text{(B.3)}\\
&= \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\mathrm{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, dP^{\text{actual}}(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1)\, dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R}). && \text{(B.4)}
\end{aligned}$$

For random $\mathbf{R}$, we have

$$\begin{aligned}
\Psi(T=1, \mathbf{R}) &\equiv \int_{\mathbf{R}} \int_{\mathbf{C}^{\mathrm{T}}} E(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1)\, dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})\, dP^{\text{actual}}(\mathbf{R} \mid T=1)\\
&= \int_{\mathbf{R}} \int_{\mathbf{C}^{\mathrm{T}}} \min_{f \in \mathcal{F}} \left\{ \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, dP^{\text{actual}}(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1) \right\} dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})\, dP^{\text{actual}}(\mathbf{R} \mid T=1)\\
&= \int_{\mathbf{R}} \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\mathrm{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})\, dP^{\text{actual}}(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1)\, dP^{\text{actual}}(\mathbf{R} \mid T=1)\\
&= \int_{\mathbf{R}} \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\mathrm{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})\, dP^{\text{actual}}(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=1)\, dP^{\text{actual}}(\mathbf{R} \mid T=1)\, \frac{dP^{\text{actual}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R}, T=1)}{dP^{\text{actual}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R}, T=1)}\\
&= \int_{\mathbf{R}} \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\mathrm{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, \frac{dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})}{dP^{\text{actual}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R}, T=1)}\, dP^{\text{actual}}(Y, \mathbf{C}^{\mathrm{T}}, \mathbf{R} \mid T=1). && \text{(B.5)}
\end{aligned}$$

Since

$$\begin{aligned} P^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R}, T=1) &= \frac{P^{\text{actual}}(\mathbf{C}^{\text{T}}, \mathbf{R}, T=1)}{P^{\text{actual}}(T=1 \mid \mathbf{R}) P^{\text{actual}}(\mathbf{R})} \\ &= \frac{P^{\text{actual}}(T=1 \mid \mathbf{C}^{\text{T}}, \mathbf{R}) P^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R}) P^{\text{actual}}(\mathbf{R})}{P^{\text{actual}}(T=1 \mid \mathbf{R}) P^{\text{actual}}(\mathbf{R})}, \\ &= \frac{P^{\text{actual}}(T=1 \mid \mathbf{C}^{\text{T}}, \mathbf{R}) P^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R})}{P^{\text{actual}}(T=1 \mid \mathbf{R})}, \\ &\equiv \frac{\pi(\mathbf{C}^{\text{T}}, \mathbf{R}) P^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R})}{\pi(\mathbf{R})}, \end{aligned}$$

we can rewrite (B.5) as

$$\begin{aligned} \Psi(T=1, \mathbf{R}) &\equiv \int_{\mathbf{R}} \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\text{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, \frac{dP^{\text{theoretical}}(\mathbf{C}^{\text{T}} \mid \mathbf{R})}{dP^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R}, T=1)} \, dP^{\text{actual}}(Y, \mathbf{C}^{\text{T}}, \mathbf{R} \mid T=1) \\ &= \int_{\mathbf{R}} \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\text{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, \frac{dP^{\text{theoretical}}(\mathbf{C}^{\text{T}} \mid \mathbf{R}) \pi(\mathbf{R})}{\pi(\mathbf{C}^{\text{T}}, \mathbf{R}) dP^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R})} \, dP^{\text{actual}}(Y, \mathbf{C}^{\text{T}}, \mathbf{R} \mid T=1) \\ &= \int_{\mathbf{R}} \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\text{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, \frac{dP^{\text{theoretical}}(\mathbf{C}^{\text{T}} \mid \mathbf{R})}{dP^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R})} \cdot \frac{\pi(\mathbf{R})}{\pi(\mathbf{C}^{\text{T}}, \mathbf{R})} \, dP^{\text{actual}}(Y, \mathbf{C}^{\text{T}}, \mathbf{R} \mid T=1) \end{aligned}$$

If we assume $dP^{\text{theoretical}}(\mathbf{C}^{\text{T}} \mid \mathbf{R}) = dP^{\text{actual}}(\mathbf{C}^{\text{T}} \mid \mathbf{R})$ (i.e. representative sampling of covariates, conditional on values of the biomarkers), then this becomes

$$\Psi(T=1, \mathbf{R}) = \int_{\mathbf{R}} \min_{f \in \mathcal{F}} \int_{\mathbf{C}^{\text{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \, \frac{\pi(\mathbf{R})}{\pi(\mathbf{C}^{\text{T}}, \mathbf{R})} \, dP^{\text{actual}}(Y, \mathbf{C}^{\text{T}}, \mathbf{R} \mid T=1). \quad \text{(B.6)}$$

If we assume the propensity scores are known, then the plug-in estimator of (B.6) is

$$\min_{f \in \mathcal{F}} \frac{1}{N_1} \sum_{i=1}^{N} I(T_i = 1) \, \frac{\pi(\mathbf{R}_i)}{\pi(\mathbf{C}^{\text{T}}{}_i, \mathbf{R}_i)} \, [Y_i - f(\mathbf{R}_i)]^2, \quad \text{(B.7)}$$

where $N_1 = \sum_{i=1}^{N} I(T_i = 1)$. We note that (B.7) is weighted squared-error loss among $T = 1$ group with weights $w_1(\mathbf{C}^{\text{T}}, \mathbf{R}) \equiv \pi(\mathbf{R})/\pi(\mathbf{C}^{\text{T}}, \mathbf{R})$.

Similarly in the $T = 0$ group, the target parameter is

$$\begin{aligned}\Psi(T=0,\mathbf{R}) &\equiv \int_{\mathbf{R}} E(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=0)\ dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})\, dP^{\text{actual}}(\mathbf{R} \mid T=1) \\ &= \int_{\mathbf{R}} \int_{\mathbf{C}^{\mathrm{T}}} \min_{f\in\mathcal{F}} \left\{ \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2\ dP^{\text{actual}}(Y \mid \mathbf{C}^{\mathrm{T}}, \mathbf{R}, T=0) \right\} dP^{\text{theoretical}}(\mathbf{C}^{\mathrm{T}} \mid \mathbf{R})\, dP^{\text{actual}}(\mathbf{R} \mid T=0) \\ &= \int_{\mathbf{R}} \min_{f\in\mathcal{F}} \int_{\mathbf{C}^{\mathrm{T}}} \int_{\mathcal{Y}} (y - f(\mathbf{R}))^2 \frac{1-\pi(\mathbf{R})}{1-\pi(\mathbf{C}^{\mathrm{T}}, \mathbf{R})}\, dP^{\text{actual}}(Y, \mathbf{C}^{\mathrm{T}}, \mathbf{R} \mid T=0), \end{aligned} \tag{B.8}$$

whose plug-in estimator, assuming the propensity scores are known, is

$$\min_{f\in\mathcal{F}} \frac{1}{N_0} \sum_{i=1}^{N} I(T_i = 0)\, \frac{1-\pi(\mathbf{R}_i)}{1-\pi(\mathbf{C}^{\mathrm{T}}{}_i, \mathbf{R}_i)}\ [Y_i - f(\mathbf{R}_i)]^2 \tag{B.9}$$

where $N_0 = \sum_{i=1}^{N} I(T_i = 0)$. We note that (B.7) is weighted squared-error loss among $T = 0$ group with weights $w_0(\mathbf{C}^{\mathrm{T}}, \mathbf{R}) \equiv (1 - \pi(\mathbf{R}))/(1 - \pi(\mathbf{C}^{\mathrm{T}}, \mathbf{R}))$.

B.2 Data Example: Summary of Dataset

Table B.1: Summary of Outcome and Treatment Variables

	Overall (93676)	**Non-event (3063)**	**Event (90613)**	**N missing**
Outcomes				
No CHD within 10 years of enrollment, n (%)	90613 (97%)	-	-	0
No breast cancer within 10 years of enrollment, n (%)	88883 (95%)	-	-	0
Treatment				
Currently using unopposed estrogen and/or estrogen plus progesterone, n (%)	41630 (44%)	1003 (33%)	40627 (45%)	85

	Overall (93676)	Non-event (3063)	Event (90613)	N missing
Highest grade completed				**767**
None, n (%)	84 (0%)	2 (0%)	82 (0%)	
1-4, n (%)	356 (0%)	11 (0%)	345 (0%)	
5-8, n (%)	1121 (1%)	51 (2%)	1070 (1%)	
9-11, n (%)	3288 (4%)	184 (6%)	3104 (3%)	
High school, n (%)	15122 (16%)	592 (19%)	14530 (16%)	
Vocational, n (%)	9123 (10%)	369 (12%)	8754 (10%)	
Some college, n (%)	24812 (27%)	828 (27%)	23984 (27%)	
College, n (%)	10669 (11%)	277 (9%)	10392 (12%)	
Some post-graduate, n (%)	11018 (12%)	314 (10%)	10704 (12%)	
Master's, n (%)	14732 (16%)	343 (11%)	14389 (16%)	
Doctoral, n (%)	2584 (3%)	67 (2%)	2517 (3%)	
Ethnicity				**265**
American Indian or Alaskan Native, n (%)	421 (0%)	18 (1%)	403 (0%)	
Asian or Pacific Islander, n (%)	2671 (3%)	52 (2%)	2619 (3%)	
Black or African-American, n (%)	7635 (8%)	283 (9%)	7352 (8%)	
Hispanic/Latino, n (%)	3609 (4%)	56 (2%)	3553 (4%)	
White (non-Hispanic), n (%)	78016 (84%)	2611 (86%)	75405 (83%)	
Other, n (%)	0 (0%)	28 (1%)	1031 (1%)	
Heard about study				**1281**
Mailed letter, n (%)	47623 (52%)	1722 (57%)	45901 (51%)	
Brochure, n (%)	9789 (11%)	317 (10%)	9472 (11%)	
TV, n (%)	2731 (3%)	93 (3%)	2638 (3%)	
Radio, n (%)	1017 (1%)	24 (1%)	993 (1%)	
Newspaper or magaize, n (%)	14610 (16%)	415 (14%)	14195 (16%)	
Meeting, n (%)	1158 (1%)	32 (1%)	1126 (1%)	
Friend or relative, n (%)	9408 (10%)	223 (7%)	9185 (10%)	
Other, n (%)	6059 (7%)	198 (7%)	5861 (7%)	
Family income				**4119**
Less than $10,000, n (%)	3917 (4%)	248 (8%)	3669 (4%)	
$10,000 - $19,999, n (%)	10101 (11%)	504 (17%)	9597 (11%)	
$20,000 - $34,999, n (%)	20226 (23%)	838 (29%)	19388 (22%)	
$35,000 - $49,999, n (%)	17430 (19%)	536 (18%)	16894 (19%)	
$50,000 - $74,999, n (%)	17487 (20%)	409 (14%)	17078 (20%)	
$75,000 - $99,999, n (%)	8181 (9%)	169 (6%)	8012 (9%)	
$100,000 - $149,999, n (%)	6034 (7%)	73 (3%)	5961 (7%)	
$150,000 or more, n (%)	3393 (4%)	42 (1%)	3351 (4%)	
Don't know, n (%)	2788 (3%)	99 (3%)	2689 (3%)	

Table B.2: Summary of Variables Influencing Only Treatment Assignment

	Overall (93676)	Non-event (3063)	Event (90613)	N missing
Age, mean (IQR)	63.6 (58, 69)	68.3 (64, 73)	63.5 (57, 69)	0
Angina ever, n (%)	5547 (6%)	584 (19%)	4963 (6%)	708
Aortic aneurysm ever, n (%)	187 (0%)	30 (1%)	157 (0%)	1523
Breast cancer ever, n (%)	5299 (6%)	208 (7%)	5091 (6%)	879
Coronary bypass surgery ever, n (%)	881 (1%)	204 (7%)	677 (1%)	1513
Cancer ever, n (%)	12075 (13%)	481 (16%)	11594 (13%)	752
Cardiac catheterization ever, n (%)	3837 (4%)	453 (15%)	3384 (4%)	1513
Carotid endarterectomy/angioplasty ever, n (%)	344 (0%)	59 (2%)	285 (0%)	1510
Cervix cancer ever, n (%)	1205 (1%)	44 (1%)	1161 (1%)	916
Heart failure ever, n (%)	893 (1%)	134 (4%)	759 (1%)	7
Cadiovascular disease ever, n (%)	17523 (19%)	1206 (40%)	16317 (18%)	2045
Hysterectomy ever, n (%)	39149 (42%)	1415 (46%)	37734 (42%)	87
Diabetes ever, n (%)	5318 (6%)	544 (18%)	4774 (5%)	96
Stroke ever, n (%)	1415 (2%)	142 (5%)	1273 (1%)	56
Have a lot of energy?				**772**
All the time, n (%)	4740 (5%)	108 (4%)	4632 (5%)	
Most the time, n (%)	33633 (36%)	766 (25%)	32867 (37%)	
A good bit, n (%)	20668 (22%)	642 (21%)	20026 (22%)	
Some times, n (%)	19397 (21%)	780 (26%)	18617 (21%)	
A little bit, n (%)	9956 (11%)	481 (16%)	9475 (11%)	
Never, n (%)	4510 (5%)	258 (9%)	4252 (5%)	
General health				**655**
Excellent, n (%)	16576 (18%)	263 (9%)	16313 (18%)	
Very good, n (%)	37684 (41%)	861 (28%)	36823 (41%)	
Good, n (%)	29669 (32%)	1280 (42%)	28389 (32%)	
Fair, n (%)	8210 (9%)	550 (18%)	7660 (9%)	
Poor, n (%)	882 (1%)	82 (3%)	800 (1%)	
High cholesterol requiring pills ever, n (%)	13773 (15%)	748 (25%)	13025 (15%)	2071
Hot flash in past 4 weeks				**733**
No, n (%)	15158 (16%)	386 (13%)	14772 (16%)	
Mild, n (%)	4593 (5%)	139 (5%)	4454 (5%)	
Moderate, n (%)	1267 (1%)	38 (1%)	1229 (1%)	
Severe, n (%)	0 (0%)	0 (0%)	0 (0%)	
Hypertension				**1699**
Never, n (%)	61196 (67%)	1270 (42%)	59926 (67%)	
Yes, untreated, n (%)	7317 (8%)	338 (11%)	6979 (8%)	
Yes, treated, n (%)	23464 (26%)	1392 (46%)	22072 (25%)	
Recent physical/emotional problems socially				**747**
Not at all, n (%)	68565 (74%)	2004 (66%)	66561 (74%)	
Slightly, n (%)	14249 (15%)	549 (18%)	13700 (15%)	
Moderately, n (%)	6121 (7%)	279 (9%)	5842 (6%)	
Quite a bit, n (%)	3217 (3%)	169 (6%)	3048 (3%)	
Extremely, n (%)	777 (1%)	34 (1%)	743 (1%)	
Quality of life (1-10), mean (IQR)	8.3 (8, 9)	8.1 (7, 9)	8.3 (8, 9)	724
Menopause before age 40, n (%)	8352 (9%)	339 (12%)	8013 (9%)	3951
MENPSYMP, n (%)	64608 (71%)	1922 (65%)	62686 (71%)	2425
Limited in daily activities?				**726**
Yes, limited a lot, n (%)	6263 (7%)	442 (15%)	5821 (6%)	
Yes, limited a little, n (%)	23110 (25%)	1132 (37%)	21978 (24%)	
No, not limited at all, n (%)	63577 (68%)	1459 (48%)	62118 (69%)	
One or both ovaries removed				**552**
No, n (%)	65240 (70%)	2019 (66%)	63221 (70%)	
Yes, one taken out, n (%)	6583 (7%)	217 (7%)	6366 (7%)	
Yes, both taken out, n (%)	18890 (20%)	713 (23%)	18177 (20%)	
Yes, unknown number taken out, n (%)	738 (1%)	29 (1%)	709 (1%)	
Yes, part of ovary taken out, n (%)	893 (1%)	30 (1%)	863 (1%)	
Don't know, n (%)	780 (1%)	36 (1%)	744 (1%)	
Osteoporosis ever, n (%)	8282 (9%)	385 (13%)	7897 (9%)	1240
Any part of ovaries removed before age 40, n (%)	8279 (9%)	301 (10%)	7978 (9%)	1120
Peripheral arterial disease ever, n (%)	2084 (2%)	249 (8%)	1835 (2%)	784
Pregnant ever, n (%)	84005 (90%)	2760 (90%)	81245 (90%)	315
Angioplasty of coronary arteries ever, n (%)	1128 (1%)	177 (6%)	951 (1%)	1509
Stroke ever, n (%)	1415 (2%)	142 (5%)	1273 (1%)	56
Health limits vigorous activities				**812**
Yes, limited a lot, n (%)	30022 (32%)	1542 (51%)	28480 (32%)	
Yes, limited a little, n (%)	41367 (45%)	1162 (38%)	40205 (45%)	
No, not limited at all, n (%)	21475 (23%)	328 (11%)	21147 (24%)	

Table B.3: Summary of Variables Influencing Treatment Assignment and Rule

Appendix C

APPENDIX FOR CHAPTER 4

C.1 Equivalence of Rules with Binary Outcome

With a binary outcome Y, the split-regression approach recommends $T = 1$ for an individual with characteristics X

$$P(Y = 1 \mid T = 1, X) \quad > \quad P(Y = 1 \mid T = 0, X), \tag{C.1}$$

which it turns out is equivalent to the rule

$$P(T = 1 \mid Y = 1, X) \quad > \quad P(T = 1 \mid X). \tag{C.2}$$

To see this, we can apply Bayes' Rule to each side of (C.1) and write

$$\begin{aligned} \frac{P(T = 1, X \mid Y = 1)P(Y = 1)}{P(T = 1, X)} &> \frac{P(T = 0, X \mid Y = 1)P(Y = 1)}{P(T = 0, X)} \\ \frac{P(T = 1 \mid Y = 1, X)P(X)P(Y = 1)}{P(T = 1 \mid X)P(X)} &> \frac{P(T = 0, X \mid Y = 1)P(X)P(Y = 1)}{P(T = 0 \mid X)P(X)} \end{aligned}$$

which is equivalent to the rule

$$\begin{aligned} \frac{P(T = 1 \mid Y = 1, X)}{P(T = 1 \mid X)} &> \frac{1 - P(T = 1 \mid Y = 1, X)}{1 - P(T = 1 \mid X)}, \\ \frac{P(T = 1 \mid Y = 1, X)}{1 - P(T = 1 \mid Y = 1, X)} &> \frac{P(T = 1 \mid X)}{1 - P(T = 1 \mid X)}, \end{aligned} \tag{C.3}$$

The left-hand side of (C.3) gives the odds of the "posterior" propensity score $P(T = 1 \mid Y = 1, X)$, while the right-hand side gives the odds of the "prior" propensity score $P(T = 1 \mid X)$.

Since the odds$(p) = p/(1-p)$ is a monotonic transformation of the underlying probability, we know the rule (C.3) is identical to the rule on the probability scale

$$P(T = 1 \mid Y = 1, X) > P(T = 1 \mid X). \tag{C.4}$$

As a result, the split-regression rule $P(Y = 1 \mid T = 1, X) > P(Y = 1 \mid T = 0, X)$ is equivalent to the rule $P(T = 1 \mid Y = 1, X) > P(T = 1 \mid X)$.

C.2 Estimation Target of OWL: Binary Response

As defined in Zhao et al. (2012) and elaborated on in Chapter 4, OWL seeks the rule that maximizes

$$\mathrm{E}_{Y,T,X}\left[\frac{I[T = \mathcal{D}(X)]}{T\pi(X) + (1-T)(1-\pi(X))}\, Y\right], \tag{C.5}$$

or equivalently,

$$\int_X \int_T \int_Y \frac{I[T = \mathcal{D}(X)]}{T\pi(X) + (1-T)(1-\pi(X))}\, Y\, dP(Y, T, X), \tag{C.6}$$

or, since $Y \in \{0, 1\}$,

$$\begin{aligned}
&\int_X \int_T \frac{I[T = \mathcal{D}(X)]}{T\pi(X) + (1-T)(1-\pi(X))} \Big\{0 \cdot dP(T, X \mid Y = 0)dP(Y = 0) + 1 \cdot dP(T, X \mid Y = 1)dP(Y = 1)\Big\} \\
&= \int_X \int_T \frac{I[T = \mathcal{D}(X)]}{T\pi(X) + (1-T)(1-\pi(X))}\, dP(T, X \mid Y = 1)dP(Y = 1).
\end{aligned} \tag{C.7}$$

We can expand (C.7) to yield the following quantity that OWL wishes to maximize (in $\mathcal{D}$)

$$\begin{aligned}
\int_X \Big\{ &\frac{I[\mathcal{D}(X) = 0]}{1 - \pi(X)} dP(T = 0 \mid X, Y = 1) + \\
&\frac{I[\mathcal{D}(X) = 1]}{\pi(X)} dP(T = 1 \mid X, Y = 1)\Big\} dP(Y = 1)\, dP(X \mid Y = 1).
\end{aligned} \tag{C.8}$$

As a result, the optimal treatment rule for OWL, $\mathcal{D}^*_{\text{OWL}}(X)$ will be

$$\mathcal{D}^*_{\text{OWL}}(X) \equiv \begin{cases} 0, & \frac{P(T=0|X,Y=1)}{1-\pi(X)} > \frac{P(T=1|X,Y=1)}{\pi(X)} \\ 1, & \frac{P(T=1|X,Y=1)}{\pi(X)} > \frac{P(T=0|X,Y=1)}{1-\pi(X)}, \end{cases} \tag{C.9}$$

since that maximizes (C.8) for each given value of patient characteristics X.

C.3 Estimation Target of OWL: Continuous Response

As in Appendix C.2, the OWL method Zhao et al. (2012) seeks to maximize

$$\int_X \int_Y \int_T \frac{I[T = \mathcal{D}(X)]}{T\pi(X) + (1-T)(1-\pi(X))}\, Y\, dF(Y, T, X), \tag{C.10}$$

which, since $T \in \{0, 1\}$, is equivalent to

$$\int_X \int_Y \left\{ \frac{I[\mathcal{D}(X) = 0]}{1 - P(T = 1 \mid X)}\, Y dF(T = 0, Y, X) + \frac{I[\mathcal{D}(X) = 1]}{P(T = 1 \mid X)}\, Y dF(T = 1, Y, X). \right\} \tag{C.11}$$

After rearranging a bit and expanding $dF(T = 1, Y, X)$, this becomes

$$\begin{aligned} \int_X \Bigg\{ \Bigg[\int_Y \frac{1 - P(T = 1 \mid Y, X)}{1 - P(T = 1 \mid X)}\, Y dF(Y \mid X) \Bigg] I[\mathcal{D}(X) = 0] \\ + \Bigg[\int_Y \frac{P(T = 1 \mid Y, X)}{P(T = 1 \mid X)}\, Y dF(Y \mid X) \Bigg] I[\mathcal{D}(X) = 1] \Bigg\} dF(X). \end{aligned} \tag{C.12}$$

If we define the Bayes factor

$$K_t(y) = \frac{P(T = t \mid X, Y = y)}{P(T = t \mid X)},$$

then (C.12) is equivalent to

$$\int_X \left\{ \left[\int_Y K_0(y)\, Y dF(Y \mid X) \right] I[\mathcal{D}(X) = 0] + \left[\int_Y K_1(y)\, Y dF(Y \mid X) \right] I[\mathcal{D}(X) = 1] \right\} dF(X).$$

Thus, it is clear that with a continuous Y, OWL targets the rule

$$\mathcal{D}^*_{\text{OWL, continuous}}(X) \equiv \begin{cases} 0, & \int_Y K_1(y)\, Y dF(Y \mid X) < \int_Y K_0(y)\, Y dF(Y \mid X) \\ 1, & \int_Y K_1(y)\, Y dF(Y \mid X) > \int_Y K_0(y)\, Y dF(Y \mid X). \end{cases} \tag{C.13}$$

C.4 Direct-Interactions with Continuous Outcome in RCT Setting

Here we add some detail to the calculation from Appendix A of Tian et al. (2014), filling in some algebra between steps and highlighting the role played by the assumption that T is independent of X, since this assumption will no longer hold in the observational study setting to which we apply the method as presented in Chen et al. (2017). All of the credit for this work goes to Tian et al. (2014) and Chen et al. (2017).

In line with Tian et al. (2014), we assume a continuous outcome Y, a treatment assignment coded as $T \in \{-1, 1\}$, and the RCT setting with $P(T = 1) = P(T = -1) = 0.5$. For the linear loss function $\ell(y, x) \equiv (y - x)^2$, the direct-interactions approach proposes minimizing the following objective function in f:

$$\begin{aligned} \mathcal{L}(f) &= E_{X,Y,T}[\ell(Y, f(X)T)] \\ &\equiv E_{X,Y,T}[(Y - f(X)T)^2] \\ &= \int_X \int_Y \int_T [Y - f(X)T]^2 dF(X, Y, T). \end{aligned}$$

Since $T \in \{-1, 1\}$,

$$\begin{aligned} \mathcal{L}(f) = &\int_X \int_Y [Y - f(X)T]^2 dF(Y \mid X, T = 1) dP(T = 1 \mid X) dF(X) \\ &+ \int_X \int_Y [Y - f(X)T]^2 dF(Y \mid X, T = -1) dP(T = -1 \mid X) dF(X) \qquad \text{(C.14)} \\ = &\int_X \frac{1}{2} \Big\{ \int_Y [Y - f(X)T]^2 dF(Y \mid X, T = 1) \\ &+ \int_Y [Y - f(X)T]^2 dF(Y \mid X, T = -1) \Big\} dF(X), \qquad \text{(C.15)} \end{aligned}$$

where the equality of (C.14) and (C.15) comes from our assumption that $P(T = 1 \mid X) = P(T = 1 \mid X) = 1/2$. If X and T were not independent (e.g. in an observational study) we could simply change the loss function to the IPW-weighted squared error loss

$$\frac{[Y - f(X)T]^2}{I[T = 1]P(T = 1 \mid X) + I[T = -1](1 - P(T = 1 \mid X))} \tag{C.16}$$

to achieve the same "cancellation" of $dP(T = 1 \mid X)$ and $dP(T = -1 \mid X)$ that in the above calculation moved us from (C.14) to (C.15) using independence of T and X. In practice, the propensity score in (C.16) will need to be estimated from the data (Kennedy, 2015).

Next, note that we can write

$$\begin{aligned}
&E_Y\left([Y - f(X)T]^2 \mid X, T = 1\right) \\
&= E_Y\left(Y^2 + f^2(Z) - 2Y f(Z) \mid X, T = 1\right) \\
&= E_Y(Y^2 \mid X, T = 1) + f^2(X) - 2f(X)E[Y \mid X, T = 1],
\end{aligned} \tag{C.17}$$

and similarly

$$\begin{aligned}
&E_Y\left([Y - f(X)T]^2 \mid X, T = -1\right) \\
&= E_Y\left(Y^2 + f^2(Z) + 2Y f(Z) \mid X, T = -1\right) \\
&= E_Y(Y^2 \mid X, T = -1) + f^2(X) + 2f(X)E[Y \mid X, T = -1].
\end{aligned} \tag{C.18}$$

Switching back to the notation of expectations, we thus know that (C.15) is equivalent to

$$\begin{aligned}
\mathcal{L}(f) =& \frac{1}{2} E_Z \left\{ E_Y[(Y - f(X)T)^2 \mid Z, T = 1] + E_Y[(Y - f(X)T)^2 \mid Z, T = 1] \right\} \\
=& \frac{1}{2} E_Z \Big\{ E_Y(Y^2 \mid X, T = 1) + f^2(X) - 2f(X) E_Y[Y \mid X, T = 1] \\
& + E_Y(Y^2 \mid X, T = -1) + f^2(X) + 2f(X) E_Y[Y \mid X, T = -1] \Big\} \\
=& E_Z \Big\{ f^2(X) - f(X) \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right) \Big\} \\
& + \frac{1}{2} E_Z \Big\{ E_Y(Y^2 \mid X, T = 1) + E_Y(Y^2 \mid X, T = -1). \Big\}
\end{aligned}$$

Recalling that our goal is to maximize this objective in f, we can add and subtract $E_Z\{\frac{1}{4}\left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1]\right)^2\}$ to obtain

$$\begin{aligned}
\mathcal{L}(f) =& E_Z \Big\{ f^2(X) - f(X) \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right) \Big\} \\
& + E_Z \Big\{ \frac{1}{4} \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right)^2 \Big\} - E_Z \Big\{ \frac{1}{4} \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right)^2 \Big\} \\
& + \frac{1}{2} E_Z \Big\{ E_Y(Y^2 \mid X, T = 1) + E_Y(Y^2 \mid X, T = -1). \Big\} \\
=& E_Z \Big\{ \frac{1}{4} \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right)^2 + f^2(X) \\
& - f(X) \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right) \Big\} \\
& - E_Z \Big\{ \frac{1}{4} \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right)^2 \Big\} + \frac{1}{2} E_Z \Big\{ E_Y(Y^2 \mid X, T = 1) + E_Y(Y^2 \mid X, T = -1). \Big\} \\
=& E_Z \Big\{ \Big[\frac{1}{2} \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right) - f(X) \Big]^2 \Big\} \\
& - E_Z \Big\{ \frac{1}{4} \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right)^2 \Big\} + \frac{1}{2} E_Z \Big\{ E_Y(Y^2 \mid X, T = 1) + E_Y(Y^2 \mid X, T = -1) \Big\}.
\end{aligned}$$

Note that in the above form of $\mathcal{L}(f)$, only the blue term is a function of $f(X)$. As a result,

$$\arg\min_f \mathcal{L}(f) = \arg\min_f E_Z \Big\{ \Big[\frac{1}{2} \left(E_Y[Y \mid X, T = 1] - E_Y[Y \mid X, T = -1] \right) - f(X) \Big]^2 \Big\}$$

and thus this minimizer is simply

$$f^*(X) \equiv \frac{1}{2}\left(E_Y[Y \mid X, T=1] - E_Y[Y \mid X, T=-1]\right). \tag{C.19}$$

C.5 Simulation Results: Estimates of Optimism

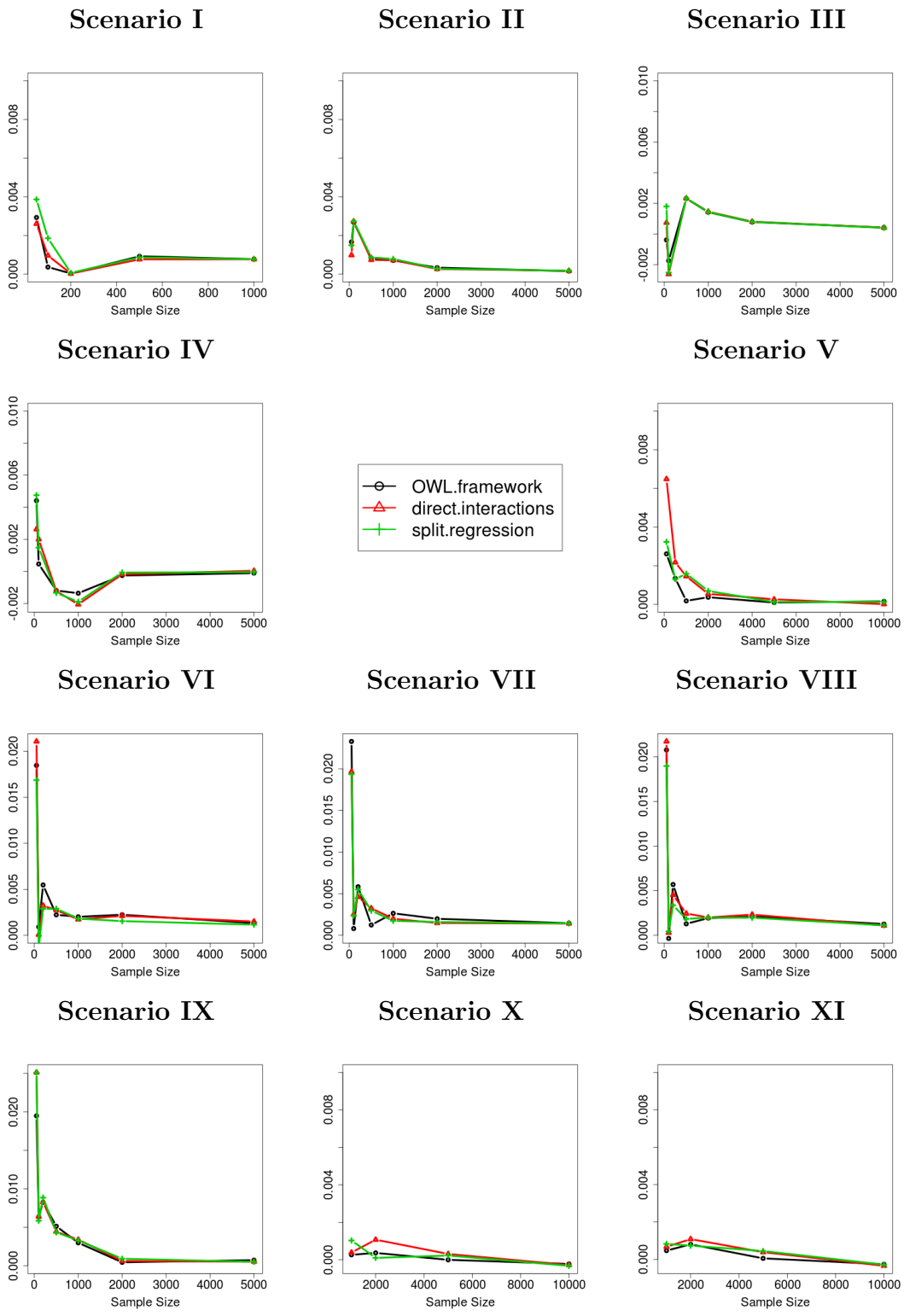


Figure C.1: Mean ABR optimism (over 50 replications) with binary Y

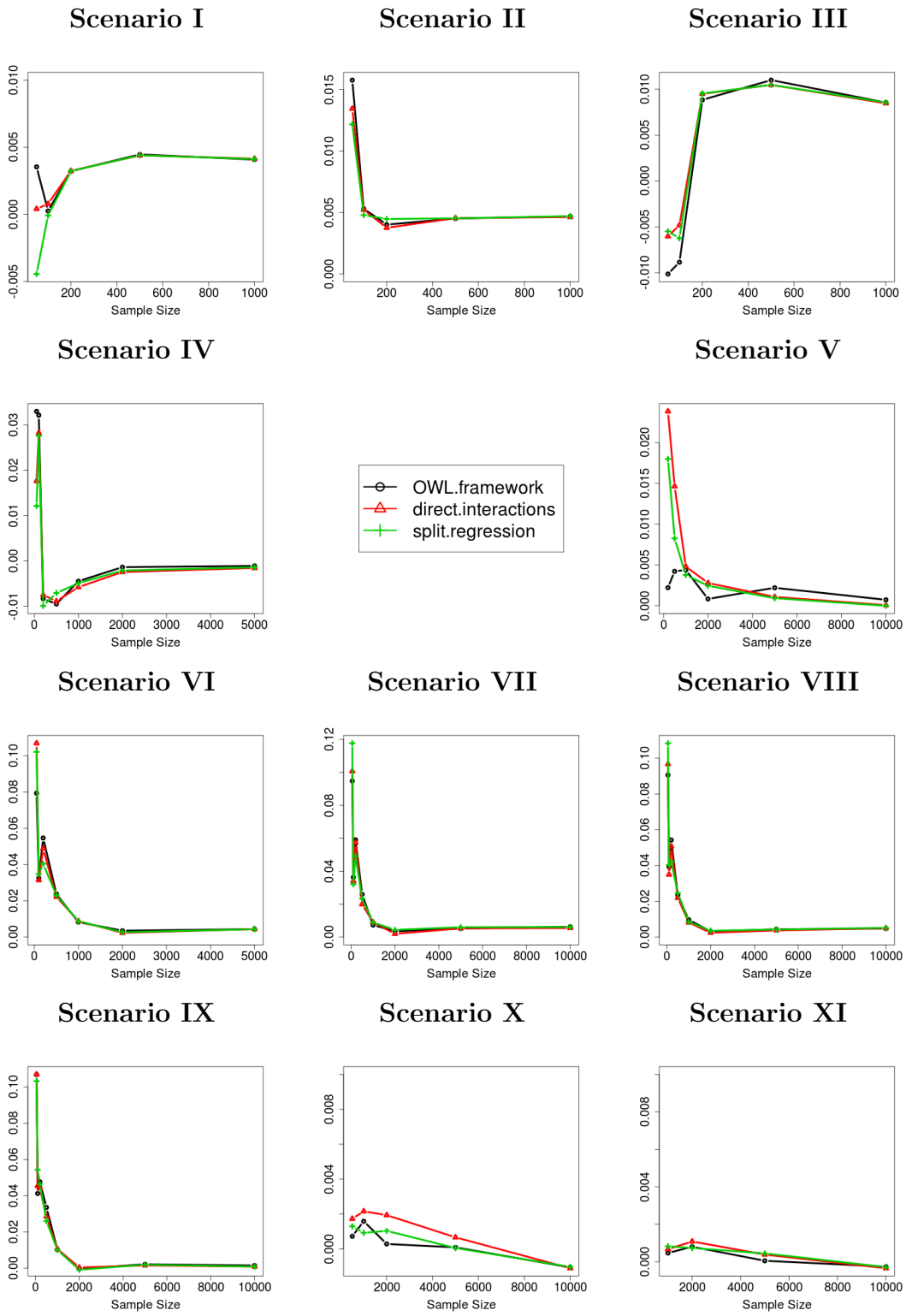


Figure C.2: Mean ABR optimism (over 50 replications) with continuous Y, $(\mu, \sigma) = (50, 1)$