G-estimation of Dynamic Treatment Regimes in the Presence of Shared Parameters

Shouao Wang

Master of Science

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal, Quebec

June 2017

A thesis submitted to McGill University in partial fulllment of the requirements of the degree of Master of Science

Dedication

I would like to dedicate this thesis to my dearest parents, their unconditionally emotional and financial support and constant encouragement have always been, and will always be, my greatest source of strength and inspiration.

Acknowledgments

I would like to express my most sincere gratitude to my supervisor, Professor Erica Moodie, for all the guidance, help and support throughout the work of this thesis. I am especially grateful to Dr. Moodie for holding weekly meeting for the past year and spending her own time revising this thesis. Her office door was always open whenever had a question about my research, and I have learned so much from her. I can not imagine finishing this thesis without her help and I am just so lucky for having her as my supervisor.

I would also like to express my appreciation to my thesis examiner Professor Yanqing Yi for her valuable comments and suggestions which have led to significant improvement on the quality of this thesis.

I would like to thank to all the professors who have taught and helped me.

I also wish to thank all my friends for their help and support during my years in McGill.

I want to send my special thanks to my girlfriend Lisa for her companionship and emotional support.

Abstract

Personalized medicine is gaining attention as a promising avenue for improved healthcare, and has received increased research interest in many domains. A *dynamic treatment regime* (DTR) is one approach to personalized medicine, which has as its basis sequential (in terms of treatment stages) decision rules that are based on a patient's personal, and evolving, medical history. In this work, I focus on G-estimation, a regression-based approach to estimating the parameters of a DTR, in the specific setting where treatment decision rule parameters may be shared across different stages of the treatment sequence.

In this thesis, a new computational method is introduced to perform shared-parameter G-estimation. The new method shares similar theoretical properties with the original, "unshared" sequential G-estimation: the new approach retains the double-robustness property, which ensures consistent estimation as long as one of (i) the expected treatment-free outcome model or (ii) the treatment model is correctly specified. Simulation studies are conducted to test the validity and performance of the shared G-estimation. In addition, comparisons between unshared and shared Q-learning, unshared sequential G-estimation, and shared-parameter G-estimation are made in terms of bias and variance. The shared parameter G-estimation method is applied to the data from the the STAR*D (NIMH Sequenced Treatment Alternatives to Relieve Depression) randomized trial to estimate the optimal shared-parameter DTR aimed at reducing symptoms of depression.

Résumé

La médecine personnalisée attire l'attention comme une avenue prometteuse pour l'amélioration des soins de santé et a suscité un intérêt croissant pour la recherche dans de nombreux domaines. A *régime de traitement dynamique* (DTR) est une approche de la médecine personnalisée, qui a comme base les règles de décision séquentielles (en termes de niveaux de traitement) basées sur l'histoire médicale personnelle et évolutive d'un patient. Dans ce travail, je me concentre sur l'estimation G, une approche basée sur la régression pour estimer les paramètres d'une DTR, dans le cadre spécifique où les paramètres de la règle de décision de traitement peuvent être partagés entre différents stades de la séquence de traitement.

Dans cette thèse, une nouvelle méthode de calcul est introduite pour effectuer une estimation G partagée. La nouvelle méthode partage des propriétés théoriques similaires avec l'estimation séquentielle initiale "non partagée": La nouvelle méthode conserve la propriété double robustesse, ce qui garantit une estimation constante aussi longtemps que l'un des (i) le modèle de résultat sans traitement attendu ou (ii) le modèle de traitement est correctement spécifié. Des études de simulation sont menées pour tester la validité et la performance de l'estimation G partagée. En outre, les comparaisons entre l'apprentissage Q non partagé et partagé, l'estimation G séquentielle non partagée et l'estimation G partagée sont faites en termes de biais et de variance. La méthode d'estimation G du paramètre partagé est appliquée aux données provenant de l'essai randomisé STAR*D (NIMH Sequenced Treatment Alternatives to Relieve Depression) pour estimer le DTR optimal des paramètres partagés visant à réduire les symptômes de la dépression.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Precision Medicine

Precision medicine is a medical model that proposes the customization of healthcare through the use of medical decisions that are tailored to the individual patient. Precision medicine is also referred to as personalized medicine which conveys the same notion of assigning different therapies to different patients based on their covariates such as genetic information and personal disease-course histories. It is gaining attention as a promising avenue for improved healthcare, and has received increased research interest in many domains as an alternative to the traditional "one size fits all" approach. Attention to the personalized medicine is not confined to academia: former U.S. President Barack Obama stated his intention to fund a United States national "precision medicine initiative" [1] in his 2015 State of the Union.

The essential motivation and potential superiority of personalized medicine over traditional approaches is based on the fact that patients often respond to a medical treatment differently, in terms of both the therapeutic effect and side effects. This inherent heterogeneity across patients in response to many treatments has prompted many health researchers to call for evidence-based implementations

[1] https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

of personalized medicine [1]. The tailoring of treatments in personalized medicine need not be targeted at genes, but rather could also tailor treatment on factors such as diet, exercise and smoking history as well. Compared to the traditional approach, the benefits of precision medicine may include better treatment efficacy, fewer side effects and a reduction of the overall cost of health care [2].

Personalized treatments can be viewed as realizations of a set of decision rules which indicate what to do in a given state (e.g. personal history, genetic information) of a patient. A simple way to view it is as a rule book, which ask physicians to apply certain therapy given the patient's information. However, the reality of decision-making in healthcare often involves complex choices with more than one stage, where decisions made at one stage may affect those to be made at other stages. Dynamic treatment regimes (DTRs) are introduced for these more complex cases, where the treatment rules can account for heterogeneity across patients and within patients over time.

1.2 Dynamic Treatment Regimes

A dynamic treatment regime (also known as "adaptive intervention" or "adaptive treatment strategy") is one approach to personalized medicine with sequential decision making, i.e. in a setting where treatments are given in stages. A DTR takes patient information as input and outputs a recommended treatment. By tailoring treatment decisions to a patient's characteristics, the DTR is able to formalize personalized medicine and improve long-term outcomes compared to the traditional non-tailored approaches [3].

We are interested in finding the optimal DTR which optimizes the mean long-term outcome, observed at the end of the final stage of intervention [4]. Hence it is required to know or estimate the outcome[2] to be able to identify a optimal DTR.

There are numerous statistical approaches to estimating a DTR. As I will detail in sections 2.2 - 2.5, regression-based approaches that rely on structural nested models [5, 6, 7] offer both flexibility and interpretability, and hence are attractive analytic choices.

[2]The outcome here is a function of other variables including patient's treatment and covariate history.

1.3 Thesis Aims and Structure

Numerous methods have been proposed for estimating an optimal DTR, and most can be classified into two general types: regression-based methods and value search methods. The regression-based approaches estimation rely on either structural nested mean models (SNMMs), or conditional expectations of the primary outcome. SNMMs parameterizes the difference between the conditional expectations under different treatment options. By estimating the parameters in a SNMM or the conditional expectation of the outcome, one can then identify the optimal DTR which is the sequence of treatment decisions that maximizes the expected outcome. There are a variety of regression-based methods in DTR estimation, including Q-learning [8, 9], dynamic weighted least squares [7], and G-estimation [10], the last of which is the focus of this thesis.

Most methods aimed at estimating an optimal DTR focus on different decision rules across stages, although it is in some settings reasonable to have the same decision rule for more than one stage, e.g., when the decision rule at each stage is a function of the same time-varying covariates at multiple stages. We refer this kind of DTR as a *shared parameter DTR*, i.e. the parameter of some or all of the time-varying covariates are shared across different stages.

Theory and implementations of the existing G-estimation approach have largely focused on unshared-parameter DTRs, a setting where sequential G-estimation can be applied in stages, typically with closed form solutions for the estimators available. Inspired by shared parameter Q-learning [11], this thesis will implement similar computational approach and thereby expand the application of G-estimation to the realm of shared parameter DTRs using a computationally tractable and stable algorithm.

I will first review some well known regression-based methods for estimating DTRs in Chapter 2. I then propose a new computational approach to G-estimation for shared parameters in Chapter 3. The proposed approach will be evaluated via simulations in Chapter 4, and then applied to the STAR*D (NIMH Sequenced Treatment Alternatives to Relieve Depression) data in Chapter 5. The conclusion and further discussions are presented in Chapter 6.

Chapter 2

Literature Review

In this chapter, I will first briefly describe the type of data needed for estimating an optimal DTR. Then I will introduce the notations that will be used throughout the thesis along with the necessary assumptions and framework. Finally, some popular regression-based approaches to estimate an optimal DTR will be explained in detail.

For estimating an optimal DTR, we need data to provide sufficient information, which includes well defined treatments, the outcomes and the measurements of covariates which are thought to influence the outcome and treatment decisions. The data could be either observational or experimental; however analyses using the observational data need more assumptions to make valid causal inference.

There are mainly two types of approaches to estimating the optimal DTRs: regression-based methods and value search methods. The regression-based methods of estimating the optimal DTRs typically proceed by first modeling the conditional mean outcomes or the contrast between optimal treatment and observed treatment for different stages, and then finding the treatments that optimize the estimated mean or contrast for different stages [4]. There are three common approaches of regression-based methods: Q-learning, G-estimation and dynamic weighted ordinary least squares; all three will be reviewed in this chapter.

In contrast to the regression-based approaches, value search estimator targets estimation of the decision rule parameters directly (rather than indirectly through the outcome mean or contrast). Value

search methods include the inverse probability weighted estimator [12], the augmented inverse probability weighted estimator [13], and outcome weighted learning along with its variants [14, 15]; since the focus of the research is on G-estimation, a regression-based method, I will not detail those value search methods here.

For simplicity, all methods introduced in the thesis will be illustrated with linear models.

2.1 Data for Estimating a DTR

To be able to construct a DTR, there are several components of information that are needed in the data [3]:

- Treatment options, which could include not only medications or drugs, but also dosage, modes of delivery (e.g., intravenous, injection or oral), behavioral intervention, etc.
- Critical decision points at which treatment is assessed and decisions are made; for example, when researchers decide to continue, stop, add, or subtract treatment.
- Tailoring variables, which are the covariates (available up to the decision point) used for making current treatment decisions; these usually include: previous treatment, response to previous treatment, demographic and genetic information and test results.
- Predictive variables, which are the variables used for predicting the outcome other than tailoring variables. These variables are particularly of use in regression-based methods, where the outcome is modeled directly; they are also useful for control of confounding.
- Measurements on other covariates that might predict treatment decisions, which are used for control of confounding.
- Outcomes, measurements for evaluating the effect of a set of sequential treatments.

The above elements are essential to conduct a DTR analysis and then make data-based clinical decisions. I will detail the use of above information further in the following parts of this chapter for some common

estimation methods[1]. Note that in particular, the variables that are predictive of both outcome and treatment choices, i.e. the confounders, are required for all approaches to DTR estimation.

2.1.1 Longitudinal Observational Studies

A longitudinal study refers to a type of design that repeatedly records the observations of the same subject (e.g., people) over a period of time. And observational study is the study where researchers observe the effect of a risk factor or treatment without trying to change who is or who is not exposed to it. Longitudinal observational data may be drawn from a variety of sources including cohort studies; randomized trials of a particular intervention that is not the treatment of interest; electronic health records such as provincial billing claims, private insurance billing claims, or disease-specific registries.

The advantages of longitudinal observational data include the possibility of obtaining a large sample size of data at relatively low cost, greater heterogeneity in the participant pool, and a more generalizable study population. However, in terms of detecting causal relationship, a longitudinal observational study is usually less reliable than a randomized experimental study due to the possibility of confounding.

Although the data from observational studies may run the risk of hidden biases and confounding[2], causal inference based on observational data is still possible under certain assumptions. Those assumptions will be further discussed in section 2.2.3. The importance of randomized trial in detection of average causal effect (ACE) has been pointed out by many researchers, hence if a randomized trial is available, it is often preferred for more accurate estimation and stronger statistical inference [17, 18]. While it is crucial to generate meaningful and suitable data for estimating DTR, this is often beyond the scope of typical randomized control trials. Thus a special class of design named sequential multiple assignment randomized trial has been proposed for providing data to develop optimal DTRs.

2.1.2 Sequentially Randomized Studies

The sequential multiple assignment randomized trial (SMART) was first introduced by Lavori and Dawson with the name "biased coin adaptive within-subject design" [19], and then Murphy proposed

[1]The second last one is not required for Q-learning methods, but needed for G-estimation.

[2]As Jerzy Neyman use to say, "without randomization an experiment has little value irrespective of the subsequent treatment" [16].

the general framework of the SMART design [20]. SMART designs initially randomize patients to the possible first stage treatments, and follow by re-randomization at each subsequent stage of all or part of the patients to another set of available treatments at that stage. The re-randomization at each subsequent stage depends on the information collected in previous stages, prior to the new treatment. It is common for subsequent randomizations to depend on whether or not a patient responds to the treatment given in the previous stage.

There are different types of SMART designs in terms of the extent of multiple randomizations:

- both responders and non-responders are re-randomized;
- only the non-responders are re-randomized;
- the re-randomization depends on both response status and previous treatment.

I present the first and second types of SMART with a two stage case in Figure 2.1 and 2.2 respectively. And the STAR*D trial that is analyzed in Chapter 5 is an example of the third type of SMART.

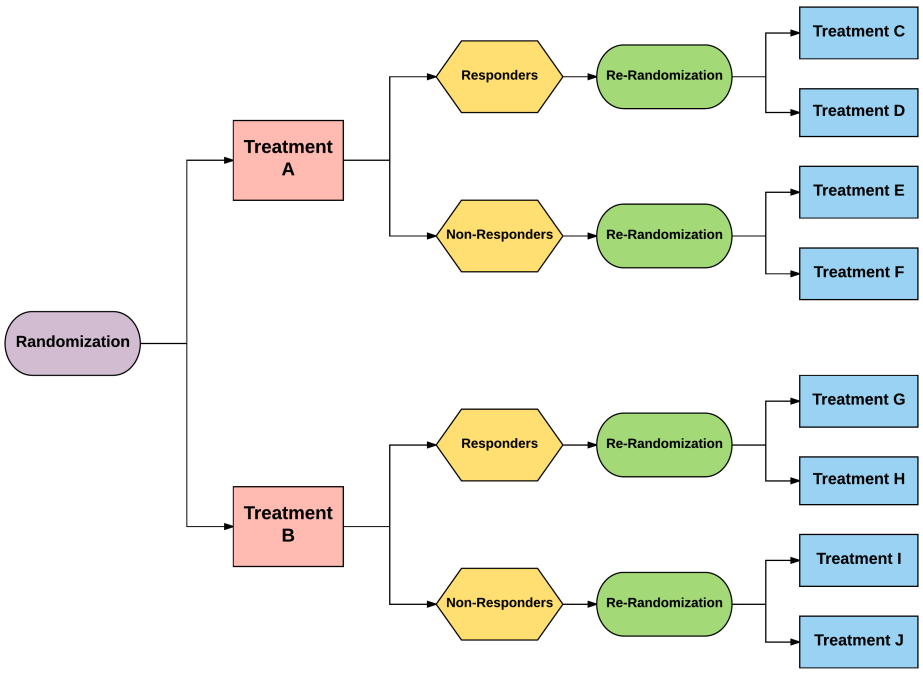

Figure 2.1: SMART design in which both responders and non-responders are re-randomized

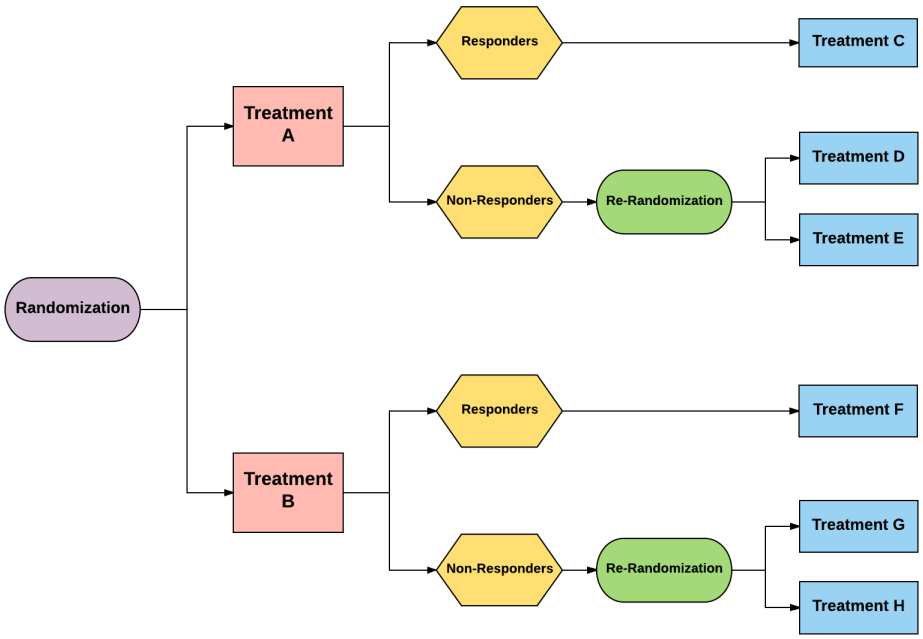

Figure 2.2: SMART design in which only non-responders are re-randomized

A SMART design is ideal for generating data for developing DTRs. There is a separate stage for each of the critical decision points involving decision making, and at each stage, all the research subjects are randomized into different treatments. Thanks to the randomization, patients at each treatment decision node are comparable with respect to all confounders, whether measured or not. There are often several embedded DTRs in one SMART design, which allows researchers to develop optimal DTRs by comparing them. Beyond those, the SMART design could detect the delayed effects and decrease the impact of cohort effects [20].

2.2 Notation and Assumptions

2.2.1 Notation

The following notations and setting will be used throughout this paper; the realized values will be indicated via lower case while upper case is used for random variables.

- Y_j: Subject's outcome at the end of stage j, after the j^{th} treatment; Y will be used to denote the final outcome or reward that we wish to maximize; it is frequently a function of $(Y_1, ..., Y_J)$ where J denotes the total number of stages;
- X_j: The covariates measured prior to treatment at the beginning of the j^{th} stage;
- A_j: The treatment assigned at the j^{th} stage subsequent to observing X_j, coded as binary {0,1}, (usually, 1 for taking the treatment, 0 for not);
- H_j: The subject history, $H_j = (X_1, ..., X_j, A_1, ..., A_{j-1})$; $H_1 = (X_1)$, $H_2 = (X_1, A_1, X_2)$, $H_3 = (X_1, A_1, X_2, A_2, X_3)$, etc;
- d_j: The decision rules, which is a function that projects from subject's history space $\mathcal{H}_j$ to treatment space $\mathcal{A}_j$: $a_j = d_j(h_j)$;
- $\mathbb{P}_n$: Denotes the empirical average function.

A DTR is a set of decision rules for all stages $(d_1, d_2, ..., d_j)$. In this thesis, I will restrict my attention to 2 and 3 stage problems.

2.2.2 The Potential Outcomes Framework

In order to estimate the optimal DTR, one needs to assess the effects of potential treatments without bias due to confounding effects, for example, sicker patients may be more likely to receive more intensive treatment. Thus, before we get any further in the discussion of estimation methods, a causal inference framework is required to control the confounders. There are several popular causal frameworks including the potential outcomes framework [17], which include the structural causal model [21] and the structural

equation models [22], and the "competing" predictive modeling approach of Dawid [23]. I will work under the potential outcomes frameworks in this thesis. The potential outcomes framework, also termed the counterfactual framework or Rubin's causal model, defines a subject's outcome when following a particular treatment regime, which could differ from what he takes in reality, and builds estimands based on "forced" interventions.

An individual level causal effect of a regime then could be viewed as the difference in outcomes if a subject followed one regime compared to another (reference regime usually), denote the causal effect as $Y(a) - Y(a')$. Unfortunately, unless there exists a parallel universe (and there exists an identical research subject), for a individual subject, one cannot observe $Y(a)$ and $Y(a')$ at the same time, if a differs from a'. This is the so-called fundamental problem of causal inference.

However, with randomization, perfect compliance and no missingness, the population level causal effect or average causal effect can be identified. For example, let $A = 0, 1$ denote treatment $a = 0, 1$, respectively. The outcome of a subject then can be expressed as $Y = A \cdot Y(1) + (1 - A) \cdot Y(0)$. Under the premise of randomization of A, we could still identify the average causal effect (ACE):

$$ACE(A \to Y) = E[Y(1) - Y(0)],$$

since

$$\begin{aligned} ACE(A \to Y) =& E[Y(1)] - E[Y(0)] \\ =& E[Y(1)|A = 1] - E[Y(0)|A = 0] \\ =& E[Y|A = 1] - E[Y|A = 0] \end{aligned}$$

because $A \perp \{Y(1), Y(0)\}$ is assured by randomization.

Without randomization, i.e. an observational study or randomized trial with imperfect compliance, more assumptions are needed to identify the causal effect.

2.2.3 Assumptions

The essential requirement of the potential outcomes framework is the axiom of consistency, which states that the potential outcome under the observed treatment equals to the observed outcome. This axiom is encapsulated in the expression $Y = A \cdot Y(1) + (1 - A) \cdot Y(0)$ given above. In addition to

consistency, three assumptions are needed for unbiased estimation of a DTR, which I will explain in the two-stage context.

- Stable unit treatment value assumption (SUTVA): A subject's outcome $Y(a_1, a_2)$[3] is not in influenced by other subject's treatment [24]. This assumption sometimes is also referred to as no interaction between units or no interference between units.

- No unmeasured confounders (NUC) or sequential ignorability, or conditional exchangeability: The treatment assignment A_j is independent of all future potential outcomes conditional on the covariates and treatment history. i.e. $A_1 \perp \{Y(a_1), Y(a_1, a_2)|H_1\}$ and $A_2 \perp \{Y(a_1, a_2)|H_2\}$ where $a_j \in \mathcal{A}_j$. The NUC assumption always holds under sequential randomization. This assumption may also be true for observational studies when all the relevant confounders have been measured and taken into consideration [25].

- Positivity: There are both treated and untreated individuals at every level of the treatment and covariate history. $P(A_1 = a|H_1) > 0$ and $P(A_2 = a|H_2) > 0, \forall a_j \in \mathcal{A}_j$. The positivity assumption assures there are subjects through all the possible combinations of treatments, from where could gain information. This assumption may be violated if a particular stratum of subjects has very few receiving treatment. Another possible reason of violation of this assumption is the study design prohibits certain subjects from receiving a particular treatment. While estimation may be possible in the presence of positivity violations, it must be acknowledged that results are being extrapolated -- and may not in fact hold -- in strata where the data do not have both treated and untreated individuals.

There is an additional strong assumption which is not necessary for DTR estimation but allow us to make counterfactual interpretations of various quantities.

- Additive local rank preservation: $Y(a) - Y(a') = E[Y(a) - Y(a')] = constant$[4]. The individual causal effect equals to the average causal effect.

[3]Here we explain the assumptions with a 2-stage longitudinal setting, where $Y(a_1)$ is the potential outcome at the end of the first stage and $Y(a_1, a_2)$ is the potential outcome at the end of second stage. A_j denotes the treatment assignment/decision at stage j, where $j = 1, 2$; and $a = 0, 1$.

[4]Note the first term $Y(a) - Y(a')$ is the realization instead of the random variable here.

This assumption states that the difference between any two individual's outcomes will be the same under all treatment patterns. The subject who do best under one regime will also do so under any another regime, and in fact the ranking of each individual's outcome will remain unchanged whatever the treatment pattern received [4].

With the above assumptions, all the methods that I will discuss in this thesis also rely on specifying components of the longitudinal distribution (i.e. models) of Y_j, A_j and H_j. Each method requires at least some if not all of the them to be correctly specified. I will detail those model specifications for each approach in later sections.

2.2.4 Value Functions and Optimal DTRs

The primary goal of personalized medicine is to estimate the optimal DTR from the data, which can be viewed as a multistage decision making problem. The optimal DTR is the one that has the greatest possible outcome value (i.e. expected outcome under that regime). The stage j value function given a regime d is defined as follows:

$$V_j^d(h_j) = E_d[\sum_{k=j}^{J} Y(H_k, A_k, X_{k+1})|H_j = h_j],\ 1 \leq j \leq J.$$

This gives the total expected future reward from stage j onward. This value function could be recursively expressed as:

$$V_j^d(h_j) = E_d[Y_j(H_j, A_j, X_{j+1}) + V_{j+1}^d(H_{j+1})|H_j = h_j],\ 1 \leq j \leq J.$$

Hence the optimal stage j value function with history h_j is $V_j^{opt}(h_j) = \max_{d \in \mathcal{D}} V_j^d(h_j)$, again it can be expressed recursively as:

$$V_j^{opt}(h_j) = \max_{a_j \in \mathcal{A}_j} E[Y_j(H_j, A_j, X_{j+1}) + V_{j+1}^{opt}(H_{j+1})|H_j = h_j, A_j = a_j].\ ^{5}$$

It is natural to directly build a model for the value function and then estimate its covariates associated parameters; we are especially interested in those covariates that interacting with the treatments together to influence the outcome. Once we have estimated the parameter related to those covariates, we could then deduce the optimal DTR, and this is the essential concept of Q-learning method.

[5] Obviously, estimating the optimal DTR is equivalent to finding the DTR that returns the greatest value here.

2.3 Q-Learning

One common regression-based approach to estimate an optimal DTR is to use Q-learning. Q-learning was first proposed by Watkins [8] under the topic of reinforcement learning in computer science, and then was applied as a reinforcement learning-based approach to estimating optimal DTRs due to the natural similarities between the backwards induction used to estimated DTRs and reinforcement learning for batch data[6].

Instead of estimating the value functions for all possible regime directly, Q-learning works with the following Q-function:

$$Q_j^d(h_j, a_j) = E[Y_j(H_j, A_j, X_{j+1}) + V_{j+1}^d(H_{j+1})|H_j = h_j, A_j = a_j],$$

which is the total expected future reward starting from stage j with history h_j, taking treatment a_j, and followed by regime d thereafter.

Then the optimal stage j Q-function is

$$Q_j^{opt}(h_j, a_j) = E[Y_j(H_j, A_j, X_{j+1}) + V_{j+1}^{opt}(H_{j+1})|H_j = h_j, A_j = a_j],$$

or re-expressed in a recursive fashion:

$$Q_j^{opt}(h_j, a_j) = E[Y_j(H_j, A_j, X_{j+1}) + Q_{j+1}^{opt}(H_{j+1})|H_j = h_j, A_j = a_j].$$

2.3.1 Unshared Q-Learning

From the definition of a Q-function, for a J stages scenario, the Jth stage Q-function is $Q_J(H_J, A_J) = E[Y_J|H_J, A_J]$; the jth stage Q-function is $Q_j(H_j, A_j) = E[Y_j + \max_{a_{j+1}} Q_{j+1}(H_{j+1}, a_{j+1})|H_j, A_j]$, $j = J-1, ..., 1$, where Y_j and Y_J are the intermediate and final outcome observed at the end of jth and Jth stage, respectively.

A special case of it is the one with single primary outcome, where $Y_j = 0$ and $Y_J = Y$. Then the Q-function can be simplified as:

$$Q_J(H_J, A_J) = E[Y|H_J, A_J],$$

[6]Both can be viewed as techniques that deal with problems involving multi-stage, sequential decisions making

$$Q_j(H_j, A_j) = E\big[\max_{a_{j+1}} Q_{j+1}(H_{j+1}, a_{j+1}) | H_j, A_j\big].$$

Whether or not intermediate outcomes are measured, we find that the optimal DTR is $d_j(h_j) = \arg\max_{a_j} Q_j(h_j, a_j)$, $\forall j$.

The true Q-function is unknown, thus we need to posit a model for the Q-function and estimate its parameters from data. It is reasonable to model the Q-function as following since it is in the form of a conditional expectation. Let the stage-j Q-function be:

$$Q_j(H_j, A_j; \beta_j, \psi_j) = \beta_j^T H_j^\beta + (\psi_j^T H_j^\psi) A_j \tag{2.1}$$

Here H_j^β and H_j^ψ are possibly different components of the history H_j, where H_j^β denotes the "main effect of history", termed as predictive variables; and H_j^ψ [7] denotes the "treatment effect of history", termed as tailoring or prescriptive variables [4]. These two elements H_j^β and H_j^ψ of subject history H_j serves as core for all the regression-based methods throughout this paper.

The prime interest here is to estimate all the ψ_js in the model, since β_js only affect the outcome but not the treatment decisions; once we estimate all the $\hat{\psi}_j$s, we will be able to estimate the optimal DTRs as well. Let ψ be a vector containing J stage-specific ψ_js, and ψ_js are different across J stages, referred to as **unshared** parameters. In unshared Q-learning, at each stage except the last, the pseudo-outcome is calculated using plug-in estimates of parameters found at the previous stages. The lack of sharing permits recursive estimation.

Algorithm

Let x_+ denotes the positive part of x. That is, $x_+ = x \cdot \mathrm{I}[x > 0]$. Then the recursive Q-learning algorithm is given by the following:

[7] Intercepts are needed in both cases so there is a leading column of ones in both H_j^β and H_j^ψ.

Algorithm 2.1 Unshared Q-Learning

- Step 1: Stage-J regression, Compute:

$$(\hat{\beta}_J, \hat{\psi}_J) = \arg\min_{\beta_J, \psi_J} \sum_{i=1}^{n} (Y_{Ji} - Q_J(H_{Ji}, A_{Ji}; \beta_J, \psi_J))^2$$

using ordinary least squares (OLS) regression with Y_{Ji} as the response and H_{Ji}, A_{Ji} as covariates to estimate β_J and ψ_J. For simplicity, we sometimes use Y_j instead of Y_{ji} as its vector form, same of H_j and A_j.

- Step 2: Calculate the stage-j pseudo-outcome:

$$\hat{Y}_j = \max_{a_{j+1}} Q_{j+1}(H_{j+1}, a_{j+1}; \hat{\beta}_{j+1}, \hat{\psi}_{j+1}) = \hat{\beta}_{j+1} H_{j+1}^{\beta} + ((\hat{\psi}_{j+1} H_{j+1}^{\psi}) \cdot A_{j+1})_+.$$

Notice the above expression involves calculating the maximum of Q_{j+1}. For $j = J-1$, $\hat{Y}_{J-1} = \max_{a_J} Q_J(H_J, a_J; \hat{\beta}_J, \hat{\psi}_J) = \hat{\beta}_J H_J^{\beta} + ((\hat{\psi}_J H_J^{\psi}) \cdot A_J)_+$, where $\hat{\psi}_J$ and $\hat{\beta}_J$ are estimated from the previous regression step.

- Step 3: Stage-j regression;

$$(\hat{\beta}_j; \hat{\psi}_j) = \arg\min_{\beta_j, \psi_j} \sum_{i=1}^{n} (\hat{Y}_{ji} - Q_j(H_{ji}, A_{ji}; \beta_j, \psi_j))^2$$

Use OLS with $\hat{Y}_{ji}/\hat{Y}_j$ from the previous stage-j pseudo-outcome step as the outcome.

- Step 4: Repeat Step 2 and Step 3, for $j = J-1, ..., 1$.

After implementing the above unshared Q-Learning algorithm, we will have estimates of parameter vector β (from $\hat{\beta}_J$ to $\hat{\beta}_1$), and ψ (from $\hat{\psi}_J$ to $\hat{\psi}_1$).

Then the estimated optimal DTR is $(\hat{d}_1^{opt}, ..., \hat{d}_J^{opt})$, where

$$\hat{d}_j^{opt}(h_j) = \arg\max_{a_j} Q_j(h_j, a_j; \hat{\beta}_j, \hat{\psi}_j).$$

To obtain the consistent estimators, the Q-function models need to be correctly specified, which may depend on some external information to help formulating the correct models.

Example

Here we elaborate how Q-learning works with $J = 3$ stages:

- Stage 3:

Propose a stage 3 Q-function model:

$$Q_3(H_3, A_3) = E[Y|H_3, A_3] = H_3^\beta \beta_3 + (H_3^\psi \psi_3) A_3.$$

Then estimate $\hat{\beta}_3$ and $\hat{\psi}_3$:

$$(\hat{\beta}_3, \hat{\psi}_3) = \arg\min_{\beta_3, \psi_3} \mathbb{P}_n (Y - H_3^\beta \beta_3 - (H_3^\psi \psi_3) A_3)^2,$$

which is solved by OLS regression.

- Stage 2:

First, calculate the stage-2 pseudo-outcome:

$$\hat{Y}_2 = \max_{a_3} Q_3(H_3, A_3) = H_3^\beta \hat{\beta}_3 + (H_3^\psi \hat{\psi}_3 A_3)_+.$$

Secondly, propose the stage 2 Q-function model:

$$Q_2(H_2, A_2) = E[\max_{a_3} Q_3(H_3, A_3)|H_2, A_2] = E[\hat{Y}_2|H_2, A_2] = H_2^\beta \beta_2 + (H_2^\psi \psi_2) A_2.$$

With the linear regression model $E[\hat{Y}_2|H_2, A_2] = H_2^\beta \beta_2 + (H_2^\psi \psi_2) A_2$ and the pseudo-outcome $\hat{Y}_2$, again use OLS to estimate $\hat{\beta}_2$ and $\hat{\psi}_2$:

$$(\hat{\beta}_2, \hat{\psi}_2) = \arg\min_{\beta_2, \psi_2} \mathbb{P}_n (\hat{Y}_2 - H_2^\beta \beta_2 - (H_2^\psi \psi_2) A_2)^2.$$

- Stage 1:

First, calculate stage-1 pseudo-outcome:

$$\hat{Y}_1 = \max_{a_2} Q_2(H_2, A_2) = H_2^\beta \hat{\beta}_2 + (H_2^\psi \hat{\psi}_2 A_2)_+.$$

Secondly, propose the stage 1 Q-function model:

$$Q_1(H_1, A_1) = E[\max_{a_2} Q_2(H_2, A_2)|H_1, A_1] = H_1^\beta \beta_1 + (H_1^\psi \psi_1) A_1.$$

Estimate $\hat{\beta}_1$ and $\hat{\psi}_1$:

$$(\hat{\beta}_1, \hat{\psi}_1) = \arg\min_{\beta_2, \psi_2} \mathbb{P}_n (\hat{Y}_1 - H_1^\beta \beta_1 - (H_1^\psi \psi_1) A_1)^2.$$

With $\hat{\psi}_3$, $\hat{\psi}_2$ and $\hat{\psi}_1$, the estimated optimal DTR is:

$$\hat{d}_j^{opt}(h_j) = \arg\max_{a_j} Q_j(h_j, a_j; \hat{\beta}_j, \hat{\psi}_j) \equiv \mathrm{I}[H_j^\psi \hat{\psi}_j > 0], \text{ for } j = 1, 2, 3$$

2.3.2 Shared Q-Learning

It is sometimes reasonable to assume that a treatment has the same effect for a subject across different stages, or that the decision rules or some component of the rules are same across different stages. In this case, the corresponding features in ψ are the same from 1 to J, i.e., $\psi_J = \psi_{J-1} = ... = \psi_1$[8]. Thus, as opposed to the unshared Q-learning, a shared Q-learning approach has been proposed [11].

In the unshared Q-learning algorithm, there are J regression equations $Q_j = \beta_j^T H_j^\beta + (\psi_j^T H_j^\psi)A_j$, and we solve them backwards from $j = J$ to 1 recursively. However, when the parameter ψ is shared through stages, we can not proceed through the algorithm sequentially as in the unshared setting.

Now let $\theta_j^T = (\beta_j^T, \psi^T)$ be the parameter of interest. With ψ being shared, let $Z_j = (H_j^\beta, H_j^\psi A_j)$ be the matrix of relevant covariate history; then $Q_j = \beta_j^T H_j^\beta + (\psi^T H_j^\psi)A_j$ can be rewritten as $Q_j = Z_j\theta_j$. The shared Q-learning approach solves the J equations simultaneously rather than recursively. This could be done by minimizing $\|Y_J - Z_J\theta_J\|^2$ and $\|Y_j(\theta_{j+1}) - Z_j\theta_j\|^2$, $\forall j < J$. We only observe $Y_J = Y$ (the primary outcome); the other outcomes $Y_j(\theta_{j+1})$, are unobserved as they depend on the unknown parameters θ_{j+1}. However, these can be replaced by the estimates of pseudo-outcome, those estimates are functions of $\hat{\theta}_{j+1}$. Similar to the unshared Q-learning, $\hat{Y}_j = \hat{\beta}_{j+1} H_{j+1}^\beta + (\hat{\psi} H_{j+1}^\psi A_{j+1})_+$. Now let:

$$Y^*(\theta) = (Y, \hat{Y}_{J-1}(\theta_J), ..., \hat{Y}_1(\theta_2))^T,$$

$$\theta = (\beta_J, \beta_{J-1}, \cdots, \beta_1, \psi)^T,$$

$$Z = \begin{pmatrix} H_J^\beta & 0 & \cdots & 0 & H_J^\psi \\ 0 & H_{J-1}^\beta & \cdots & 0 & H_{J-1}^\psi \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & H_1^\beta & H_1^\psi \end{pmatrix}.$$

Then the regression model can be written as $Y^*(\theta) = Z\theta$, with

$$\hat{\theta} = \arg\min_\theta \|Y^*(\theta) - Z\theta\|^2.$$

By doing so, we summarize all stages Q-functions in one system of equations. $J(\theta) = \|Y^*(\theta) - Z\theta\|^2$ can be identified as the (approximate) squared Bellman error [26]. Due to the non-smooth maximization

[8] All or part of decision rule parameters can be assumed to be the same, since they are vectors with sometimes different lengths.

operation used in defining the pseudo-outcomes, algorithms that directly minimize the squared Bellman error can be unstable and may not converge [9, 27]. Hence instead of minimizing the squared Bellman error, the estimating equation with Bellman residual proposed by Chakraborty et al. [11] is: $Z^T \cdot (Y^*(\theta) - Z\theta) = 0$. Since $Y^*(\theta)$ depends on the parameters θ via the maximization, the estimating equation is non-linear, thus an iterative method needs to be used.

Algorithm

Algorithm 2.2 Shared Q-Learning

- Step 1: Propose initial values for θ, denoted $\hat{\theta}^{(0)} = (\hat{\beta}_J^{(0)}, \hat{\beta}_{J-1}^{(0)}, \cdots, \hat{\beta}_1^{(0)}, \hat{\psi}^{(0)})^T$, and set $t = 1$.
- Step 2: Calculate $Y^*(\hat{\theta}^{(t)})$ using $\hat{\theta}^{(t-1)}$.
- Step 3: Solve the estimating equation $Z^T \cdot (Y^*(\hat{\theta}^{(t)}) - Z\theta) = 0$ for θ to obtain $\hat{\theta}^{(t+1)}$, increment t.
- Step 4: Repeat step 2 and step 3 until convergence: $\|\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}\| < \epsilon$.

The choice of initial value of θ could be taken from estimates of unshared Q-learning, denoted as $\hat{\theta}_j^{(0)}$, $j = J, ..., 1$, or just using fixed values such as 0 throughout. Note that for shared parameter ψ, we need to map J distinct estimates $\hat{\psi}_j$ into one $\hat{\psi}$: $\hat{\psi}^{(0)} = f(\hat{\psi}_J^{(0)}, \hat{\psi}_{J-1}^{(0)}, ..., \hat{\psi}_1^{(0)})$ if we choose to use unshared Q-learning to obtain the initial value.

The following are some possible choice of initial estimates:

1. Simple Average:

$$\hat{\theta}^{(SA)} = (\hat{\beta}_J^{(0)}, ..., \hat{\beta}_1^{(0)}, \hat{\psi}^{(SA)}),$$

   where $\hat{\psi}^{(SA)} = \frac{1}{J} \sum_{j=1}^{J} \hat{\psi}_j^{(0)}$.

2. Inverse Variance Weighted Average:

$$\hat{\theta}^{(IVWA)} = (\hat{\beta}_J^{(0)}, ..., \hat{\beta}_1^{(0)}, \hat{\psi}^{(IVWA)}),$$

   where $\hat{\psi}^{(IVWA)} = \sum_{j=1}^{J} \frac{\hat{\psi}_j^{(0)}}{\hat{\sigma}_j^2} / \sum_{j=1}^{J} \frac{1}{\hat{\sigma}_j^2}$, and $\hat{\sigma}_j^2$ is the estimated variance of $\hat{\psi}_j^{(0)}$.

3. Maximum:

$$\hat{\theta}^{(MAX)} = (\hat{\beta}_J^{(0)}, ..., \hat{\beta}_1^{(0)}, \hat{\psi}^{(MAX)}),$$

where $\hat{\psi}^{(MAX)} = \max\{\hat{\psi}_1^{(0)}, ..., \hat{\psi}_J^{(0)}\}$.

4. Minimum:

$$\hat{\theta}^{(MIN)} = (\hat{\beta}_J^{(0)}, ..., \hat{\beta}_1^{(0)}, \hat{\psi}^{(MIN)}),$$

where $\hat{\psi}^{(MIN)} = \min\{\hat{\psi}_1^{(0)}, ..., \hat{\psi}_J^{(0)}\}$.

5. Zeros:

$$\hat{\theta}^{(ZERO)} = (0, ..., 0, 0).$$

Chakraborty et al. found that there was no dependence of the estimates on the initial values, however the algorithm converged faster when the initial values were closer to the true values [11].

2.4 G-estimation

Nearly two decades ago, Robins proposed a new method for finding optimal DTRs called G-estimation using the structural nested mean model (SNMM) [6]. Murphy [5] proposed the first semi-parametric approaches, which was later shown to be a special case of G-estimation [28].

As opposed to Q-learning, G-estimation models contrasts of the conditional expectation of outcomes rather than modeling the conditional expectation of the outcomes themselves. G-estimation is similar to Q-learning in concept, but requires additional modeling; in return it provides increased robustness to model miss-specification.

2.4.1 Structural Nested Mean Models (SNMM)

The SNMM proposed by Robins is used to model the effect of a treatment as a function of the covariate history up to that stage. In a two stage setting, $E[Y(a_1, a_2) - Y(a_1, a_2')|H_2]$ is a SNMM; it only models the difference in outcomes under different treatment regimes rather than outcomes. The SNMM has variety of forms, and the blip model is a one of them.

Define the optimal blip-to-reference function:

$$\gamma_j(h_j, a_j) = E[Y(\bar{a}_j, \underline{d}_{j+1}^{opt}) - Y(\bar{a}_{j-1}, d_j^{ref}, \underline{d}_{j+1}^{opt})|H_j = h_j].$$

When the "zero" regime is taken as reference regime, it becomes the optimal blip-to-zero function:

$$\gamma_j(h_j, a_j) = E[Y(\bar{a}_j, \underline{d}_{j+1}^{opt}) - Y(\bar{a}_{j-1}, 0_j, \underline{d}_{j+1}^{opt})|H_j = h_j],$$

where the "zero" treatment refers to some meaningful treatment such as placebo or standard care.

The optimal blip-to-zero function measures the expected difference between the average outcome for someone who received treatment a_j at stage j and someone who is given the "zero" treatment at stage j. Both subjects have the same covariate and treatment history, and will be treated optimally from stage $j+1$ onward. This can be viewed as the expected effect of treatment a_j[9].

Another variant of the blip function is the regret function, defined as:

$$\mu_j(h_j, a_j) = E[Y(\bar{a}_{j-1}, \underline{d}_j^{opt}) - Y(\bar{a}_j, \underline{d}_{j+1}^{opt})|H_j = h_j],$$

or equivalently: $\mu_j(h_j, a_j) = \max \gamma_j(a_j) - \gamma_j(a_j)$. The regret function is the negative of the optimal blip-to-reference function where the optimal treatment is taken to be the reference regime. Without further specification, blip function refers to the optimal blip-to-zero function in this thesis.

2.4.2 Unshared G-estimation

Having now discussed the necessary preliminaries of G-estimation, I will now give the details of it. G-estimation estimates the parameters ψ of the optimal blip function via a combination of regression models and estimating equations [29]. Define

$$G_j(\psi) = Y - \gamma_j(h_j, a_j; \psi) + \textstyle\sum_{k=j+1}^{J}[\gamma_k(h_k, d_k^{opt}; \psi) - \gamma_k(h_k, a_k; \psi)].$$

where $\gamma_k(h_k, d_k^{opt}; \psi) = E[Y(\bar{a}_{k-1}, d_k^{opt}, \underline{d}_{k+1}^{opt}) - Y(\bar{a}_{k-1}, 0_k, \underline{d}_{k+1}^{opt})|H_k = h_k]$, is the blip function interpreted as a subject's outcome adjusted by the difference between the expected outcome for someone who received optimal regime and someone received the zero regime at stage k, both of them have the same treatment and covariate history up to stage $k-1$, and treated with optimal regime from stage

[9]This blip can be interpreted as the expected (counterfactual) difference in outcome due to treatment a_j relative to zero.

$k+1$ onward. Similarly, $\gamma_k(h_k, a_k; \psi) = E[Y(\bar{a}_k, a_k, \underline{d}^{opt}_{k+1}) - Y(\bar{a}_k, 0_k, \underline{d}^{opt}_{k+1})|H_k = h_k]$ is the difference in expected outcome caused by taking observed regime a_k instead of zero regime at stage k. Thus $\gamma_k(h_k, d^{opt}_k; \psi) - \gamma_k(h_k, a_k; \psi)$ is the difference in expected outcome caused by taking the optimal regime over observed regime a_k at stage k. $G_j(\psi)$ then can be interpreted as subject's outcome adjusted by the expected effect of taking optimal regime from stage $j+1$ and onward. Imagine the patient can actually time travel back to the time point j, if he now makes "perfect" treatment decisions (optimal regime) for the current stage and all the following stages, then $G_j(\psi)$ will be his outcome under this scenario.

Robins has proposed the following estimating equation:

$$U(\psi) = \sum_j^J (G_j(\psi) - E[G_j(\psi)|H_j; \beta_j])(S_j(A_j) - E[S_j(A_j)|H_j; \alpha_j]),$$

where $S_j(A_j)$ is a vector-valued function that contains variables thought to interact with treatment to effect a difference in expected outcome, here let $S_j(A_j) = \frac{\partial \gamma_j}{\partial \psi_j}$ [10]. Here we assume the blip $\gamma_j(H_j, A_j; \psi_j)$ is always correctly specified, under this premise, Robins proved that the estimators have the double robustness property, i.e. the estimators of ψs are consistent if either the expected treatment-free outcome model $E[G_j(\psi)|H_j, \beta_j]$ or the treatment model $p_j(A_j = 1|H_j; \alpha_j)$, used to compute $E[S_j(A_j)|H_j; \alpha_j]$, is correctly specified [10].

Algorithm

The unshared G-estimation algorithm can be described as follows:

[10] As we will see later, $S_j(A_j) = H_j^{\psi} A_j$, and it is closely related to the treatment model $p_j(A_j = 1|H_j; \alpha_j)$, also note this probability is in fact the propensity score.

Algorithm 2.3 Unshared G-estimation

- Step 1: Propose the optimal blip-to-zero function model:

$$\gamma_j(H_j, A_j; \psi) = E[Y(\bar{a_j}, a_j, \underline{d}_{j+1}^{opt}) - Y(\bar{a}_{j-1}, 0_j, \underline{d}_{j+1}^{opt})|H_j = h_j] = (\psi_j^T H_j^{\psi})A_j.$$

- Step 2: Set the recursive G-function:

$$G_j(\psi) = Y - \gamma_j(H_j, a_j; \psi_j) + \sum_{k=j+1}^{J} [\gamma_k(H_k, d_k^{opt}; \psi_k) - \gamma_k(H_k, a_k; \psi_k)].$$

- Step 3: Propose the expected treatment free outcome model:

$$E[G_j(\psi_j)|H_j = h_j; \beta_j] = \beta_j^T H_j^{\beta},$$

and β_j is estimated as a function of ψ_j from data using OLS. Denote this as $\hat{\beta}_j(\psi_j)$, obtained from the estimating equation of $\hat{\beta}_j$: $E[H_j^{\beta}(Y_j - \psi_j^T H_j^{\psi} A_j - \beta_j^T H_j^{\beta})] = 0$.

- Step 4: Propose a treatment model:

$$p(A_j|H_j; \alpha_j).$$

Set $S_j(A_j) = \frac{\partial \gamma_j}{\partial \psi_j} = H_j^{\psi} A_j$, then $E[S_j(A_j)|H_j; \alpha_j] = E[H_j^{\psi} A_j|H_j^{\psi}; \alpha_j] = H_j^{\psi} \cdot E[A_j|H_j; \alpha_j]$ is a function of $p_j(A_j = 1|H_j; \alpha_j)$, and α_j is usually estimated from the data using logistic regression.

- Step 5: Construct the estimating function:

$$U_j(\psi_j) = (G_j(\psi_j) - E[G_j(\psi_j)|H_j = h_j; \beta_j])(S_j(A_j) - E[S_j(A_j)|H_j = h_j; \alpha_j]).$$

Plugging in $G_j(\psi_j)$, $E[G_j(\psi_j)|H_j = h_j; \beta_j]$, $S_j(A_j)$ and $E[S_j(A_j)|H_j = h_j; \alpha_j]$ into the function, we have:

$$U_j(\psi_j) = \mathbb{P}_n[(\hat{Y}_j - \psi_j^T H_j^{\psi} A_j) - \beta_j^T H_j^{\beta}] \cdot [H_j^{\psi} A_j - E[H_j^{\psi} A_j|H_j; \alpha_j]],$$

where $\hat{Y}_J = Y$.

- Step 6: Estimate $\hat{\psi}_j$ by solving the equation system $U_j(\psi_j) = 0$ with substitution of the estimates of $\hat{\beta}_j$ and $\hat{\alpha}_j$.

- Step 7: Move one stage backwards, calculate the stage specific pseudo-outcome:

$\hat{Y}_j = Y + \sum_{k=j+1}^{J}[-\gamma_k(H_k, A_k; \hat{\psi}_k) + \gamma_k(H_k, d_k^{opt}; \hat{\psi}_k)]$, or

$$\hat{Y}_j = Y + \sum_{k=j+1}^{J} [-H_k^{\psi}\hat{\psi}_k A_k + (H_k^{\psi}\hat{\psi}_k A_k)_+],$$

is the observed outcome "taking off" the expected effect of observed treatment, then adding back the expected effect of the optimal treatment.

- Step 8: Repeat from step 1 to step 7 till reached stage 1.

Remark 1: Here we proposed three models in total:

- optimal blip-to-zero models: $\gamma_j(H_j, A_j; \psi_j) = (\psi_j^T H_j^\psi)A_j$,
- expected treatment free models: $E[G_j(\psi_j)|H_j = h_j; \beta_j] = \beta_j^T H_j^\beta$, and
- treatment models: $p(A_j|H_j; \alpha_j)$ or $E[A_j|H_j; \alpha_j]$.

The optimal-to-zero blip models are always assumed to be correctly specified. As long as one of the last two models is correctly specified, the estimators $\hat{\psi}_j$ are consistent. The first two models together are equivalent to the Q-function models; i.e., the Q-function is $E[Y|H, A; \psi, \beta] = \beta^T H^\beta + A \cdot (\psi^T H^\psi) = E[G(\psi)|H; \beta] + \gamma(H, A; \psi)$.

Remark 2: I have mentioned pseudo-outcome both in Q-learning and G-estimation, but they are different. In Q-learning, the pseudo-outcome is the predicted outcome under the optimal treatment $\mathbb{E}[Y|A^{opt}]$. In G-estimation, the pseudo-outcome is the observed outcome taking off the expected effect of observed treatment, then adding in the expected effect of the optimal treatment. In addition, the pseudo-outcome in Q-learning relies on the estimates of β_j, and the pseudo-outcome in G-estimation does not:

$$\hat{Y}_j^Q = \max_{a_{j+1}} Q_{j+1}(H_{j+1}, a_{j+1}; \hat{\beta}_{j+1}, \hat{\psi}_{j+1}) = \hat{\beta}_{j+1} H_{j+1}^\beta + (\hat{\psi}_{j+1} H_{j+1}^\psi)_+.$$

$$\hat{Y}_j^G = Y + \textstyle\sum_{k=j+1}^J [-H_k^\psi \hat{\psi}_k A_k + (H_k^\psi \hat{\psi}_k A_k)_+];$$

or with recursive fashion $\hat{Y}_j^G = \hat{Y}_{j+1} + [-H_{j+1}^\psi \hat{\psi}_{j+1} A_{j+1} + (H_{j+1}^\psi \hat{\psi}_{j+1} A_{j+1})_+]$. Note also that with the usage of the pseudo-outcome, we could simplify the notation for the stage j G-function to:

$$G_j(\psi) = \hat{Y}_j - \gamma_j(h_j, a_j; \psi_j),$$

or equivalently, $G_j(\psi) = Y + \sum_{k=j}^J \mu_k(\psi)$ with the regrets parameterization. From this point forward, the term pseudo-outcome will be used without specifying its definition, with the appropriate choice for each of Q-learning and G-estimation being assumed.

Example

Here we consider an example with 3 stages:

- Stage 3:

Step 1: Propose the optimal blip-to-zero model:

$$\gamma_3(H_3, a_3; \psi_3) = E[Y(\bar{a}_3) - Y(\bar{a}_2, 0_3)|H_3] = H_3^{\psi}\psi_3 A_3.$$

Step 2: Set the G-function $G_J(\psi_J) = Y - \gamma_J(H_J, a_J; \psi_J)$:

$$G_3(\psi_3) = Y - \gamma_3(H_3, a_3; \psi_3) = Y - H_3^{\psi}\psi_3 A_3.$$

Step 3: Propose the expected treatment free outcome model:

$$E[G_3(\psi_3)|H_3 = h_3; \beta_3] = E[Y - \gamma_3(H_3, a_3; \psi_3)] = H_3^{\beta}\beta_3 + H_3^{\psi}\psi_3 A_3 - H_3^{\psi}\psi_3 A_3 = H_3^{\beta}\beta_3.$$

$\hat{\beta}_3(\psi_3) = ((H_3^{\beta})^T H_3^{\beta})^{-1}(H_3^{\beta})^T(Y - H_3^{\psi}\psi_3 A_3)$ is the OLS estimator with response as $Y - H_3^{\psi}\psi_3 A_3$, and the covariate is H_3^{β}.

Step 4: Propose the treatment model: $E[A_3|H_3; \alpha_3]$.

Since $S_j(A_j) = \frac{\partial \gamma_j}{\partial \psi_j} = H_j^{\psi} A_j$, $S_3(A_3) = H_3^{\psi} A_3$. Then we have $E[S_3(A_3)|H_3; \alpha_3] = H_3^{\psi} E[A_3|H_3; \alpha_3]$, and $\hat{\alpha}_3$ is estimated using some possibly non-parametric methods.

Step 5: Construct the stage 3 estimation function:

$$U_3(\psi_3) = \mathbb{P}_n[(Y - H_3^{\psi}\psi_3 A_3) - H_3^{\beta}\hat{\beta}_3] \cdot [H_3^{\psi} A_3 - H_3^{\psi} E[A_3|H_3; \hat{\alpha}_3]],$$

Step 6: Solve $U_3(\psi_3) = 0$ to estimate ψ_3.

Step 7: Compute the stage 2 pseudo-outcome $\hat{Y}_2 = Y - H_3^{\psi}\hat{\psi}_3 A_3 + (H_3^{\psi}\hat{\psi}_3 A_3)_+$.

- Stage 2:

Step 1: Propose the optimal blip-to-zero model:

$$\gamma_2(H_2, a_2; \psi_2) = E[Y(\bar{a}_2, d_3^{opt}) - Y(a_1, 0_2, d_3^{opt})|H_2] = H_2^{\psi}\psi_2 A_2.$$

And from stage-3 blip model, we have:

$$\gamma_3(H_3, d_3^{opt}; \psi_3) = E[Y(\bar{a}_2, d_3^{opt}) - Y(\bar{a}_2, 0_3)|H_3] = (H_3^{\psi}\psi_3 A_3)_+.$$

Step 2: Set the G-function [11]:

$$G_2(\psi_2) = Y - \gamma_2(h_{2,}a_2;\psi_2) + [\gamma_3(h_3, d_3^{opt};\psi_3) - \gamma_3(H_3, a_3;\psi_3)]$$

$$= Y - H_2^{\psi}\psi_2 A_2 - H_3^{\psi}\psi_3 A_3 + (H_3^{\psi}\psi_3 A_3)_+$$

With the substitution of $\hat{\psi}_3$ and $\hat{Y}_2$,

$$G_2(\psi_2) = Y - H_2^{\psi}\psi_2 A_2 - H_3^{\psi}\hat{\psi}_3 A_3 + (H_3^{\psi}\hat{\psi}_3 A_3)_+$$

$$= \hat{Y}_2 - H_2^{\psi}\psi_2 A_2$$

Step 3: Propose the expected treatment free outcome model:

$$E[G_2(\psi_2)|H_2;\beta_2] = H_2^{\beta}\beta_2.$$

Hence $\hat{\beta}_2(\psi_2) = ((H_2^{\beta})^T H_2^{\beta})^{-1}(H_2^{\beta})^T(\hat{Y}_2 - H_2^{\psi}\psi_2 A_2)$ via linear regression.

Step 4: Propose the treatment model: $E[A_2|H_2;\alpha_2]$. Since $S_2(A_2) = H_2^{\psi}A_2$, then $E[S_2(A_2)|H_2;\hat{\alpha}_2] = H_2^{\psi}E[A_2|H_2;\hat{\alpha}_2]$.

Step 5: Construct the stage 2 estimating function:

$$U_2(\psi_2) = \mathbb{P}_n[(\hat{Y}_2 - H_2^{\psi}\psi_2 A_2) - H_2^{\beta}\hat{\beta}_2] \cdot [H_2^{\psi}A_2 - H_2^{\psi}E[A_2|H_2;\hat{\alpha}_2]]$$

Step 6: Solve $U_2(\psi_2) = 0$ to estimate ψ_2.

Step 7: Calculate the stage 1 pseudo-outcome $\hat{Y}_1 = Y - H_3^{\psi}\hat{\psi}_3 A_3 + (H_3^{\psi}\hat{\psi}_3 A_3)_+ - H_2^{\psi}\hat{\psi}_2 A_2 + (H_2^{\psi}\hat{\psi}_2 A_2)_+$ [12].

- Stage 1:

Step 1: Propose the optimal blip-to-zero model:

$$\gamma_1(H_1, a_1;\psi_1) = E[Y(a_1, \underline{d}_2^{opt}) - Y(0_1, \underline{d}_2^{opt})|H_1] = H_1^{\psi}\psi_1 A_1.$$

And from the stage-2 blip model, we have:

$$\gamma_2(H_2, d_2^{opt};\psi_2) = E[Y(a_1, \underline{d}_2^{opt}) - Y(a_1, 0_2, d_3^{opt})] = (H_2^{\psi}\psi_2 A_2)_+.$$

Step 2: Set the G-function:

$$G_1(\psi_1) = \hat{Y}_1 - \gamma_1(H_1, a_1;\psi_1).$$

[11] With $A_j \in \{0,1\}$, $(H_3^{\psi}\psi_3 A_3)_+ = \frac{H_3^{\psi}\psi_3 + |H_3^{\psi}\psi_3|}{2}$

[12] If express $\hat{Y}_1$ with regrets, $\hat{Y}_1 = Y + \mu_3(H_3, A_3) + \mu_2(H_2, A_2)$.

Then,

$$G_1(\psi_1) = Y - H_3^{\psi}\hat{\psi}_3 A_3 + (H_3^{\psi}\hat{\psi}_3 A_3)_+ - H_2^{\psi}\hat{\psi}_2 A_2 + (H_2^{\psi}\hat{\psi}_2 A_2)_+ - H_1^{\psi}\psi_1 A_1.$$

Step 3: Propose the expected treatment free outcome model:

$$E[G_1(\psi_1)|H_1;\beta_1] = H_1^{\beta}\beta_1.$$

Estimate β_1 using OLS : $\hat{\beta}_1(\psi_1) = ((H_1^{\beta})^T H_1^{\beta})^{-1}(H_1^{\beta})^T(\hat{Y}_1 - H_1^{\psi}\psi_1 A_1)$.

Step 4: Propose the treatment model: $E[A_1|H_1;\alpha_1]$. With the estimate of α_1, $S_1(A_1) = H_1^{\psi}A_1$, $E[S_1(A_1)|H_1;\hat{\alpha}_1] = H_1^{\psi}E[A_1|H_1;\hat{\alpha}_1]$.

Step 5: Construct the stage 2 estimation function:

$$U_1(\psi_1) = \mathbb{P}_n[(\hat{Y}_1 - H_1^{\psi}\psi_1 A_1) - H_1^{\beta}\hat{\beta}_1] \cdot [H_1^{\psi}A_1 - H_1^{\psi}E[A_1|H_1;\hat{\alpha}_1]].$$

Step 6: Solve $U_1(\psi_1) = 0$ to estimate ψ_1.

Finally, we find that d_1^{opt} is 1 when $H_1\hat{\psi}_1 \geq 0$ and 0 otherwise, and similarly for d_2^{opt} and d_3^{opt}. The optimal DTR then is $d = \{d_1^{opt}, d_2^{opt}, d_3^{opt}\}$.

2.5 Dynamic Weighted Ordinary Least Squares

From section 2.3 and 2.4, we see that the Q-learning approach is relatively easy to implement but suffers from a lack of robustness; G-estimation offers double robustness property but is sometimes difficult to understand and implement.

Wallace and Moodie [7] proposed an new approach to DTR estimation: dynamic weighted ordinary least squares (dWOLS). The dWOLS combines the intuitiveness of Q-learning and double robustness of G-estimation together with only some minor pre-computations and implementation of standard weighted ordinary least squares regression.

With the same setup as Q-learning and G-estimation, the algorithm of dWOLS is:

Algorithm 2.4 dWOLS

For each step $j = J...1$

- Propose the stage j regret function, or equivalently the blip function:

$$\mu_j(H_j, A_j; \psi_j) = \gamma_j(H_j, A_j^{opt}; \psi_j) - \gamma_j(H_j, A_j; \psi_j) = (H_j^{\psi}\psi_j A_j)_+ - H_j^{\psi}\psi_j A_j$$

- Propose the stage j treatment free outcome model:

$$\mathbb{E}[\hat{Y}_j | H_j; \beta_j, \psi_j] = \beta_j H_j^{\beta} + \psi_j H_j^{\psi} A_j$$

- Calculate the stage j pseudo-outcome:

$$\hat{Y}_j = Y + \sum_{k=j+1}^{J} \mu_k(H_k, A_k; \hat{\psi}_k),$$

where $\hat{\psi}_k$ are taken from previous stages of estimation, notice this is identical to the pseudo-outcome of G-estimation.

- Perform a weighted regression of $\hat{Y}_j$ on $\{H_j^{\beta}, H_j^{\psi} A_j\}$ with weights $w_j(a_j, H_j)$ that satisfy criteria $\pi(H)w(1, H) = (1 - \pi(H))w(0, H)$ to estimate $\hat{\psi}_j$, where $\pi(H) = P(A = 1|H)$.
- Move one stage back, till reached stage 1.

For the selection of w_j, Wallace and Moodie have proposed some possible choices of weights w_j:

1. $w_{1i} = |a_i - P(A_i = 1|H_i)|$,
2. $w_{2i} = [P(A_i = a_i|H_i)]^{-1}$,
3. $w_{3i} = 1_{a_i=1} + 1_{a_i=0} \frac{P(A_i=1|H)}{1-P(A_i=1|H)}$, or
4. $w_{4i} = 1_{a_i=0} + 1_{a_i=1} \frac{1-P(A_i=1|H)}{P(A_i=1|H)}$.

All of above weights satisfy $\pi(H)w(1, H) = (1-\pi(H))w(0, H)$, which will provide consistent estimators of ψ.

The dWOLS takes the form of Q-learning which is relatively easy to understand and perform, and assures the double robustness property by introducing specialized weights w. Hence, the results from dWOLS can compete, and sometimes out-perform G-estimation in terms of efficiency [7].

2.6 Summary

I have reviewed three regression-based approaches for DTR estimation: Q-learning, G-estimation and dWOLS. Q-learning models the conditional mean outcomes, while G-estimation and dWOLS models the contrast between optimal (counterfactual) treatment and observed treatment. In contrast, the value search approaches target directly to the parameters of the treatment rule itself rather than the parameters of the mean outcome model or the contrast model, and these have not been detailed here.

The Q-learning approach is easy to understand and implement, but it is a singly-robust method, the consistent estimators are only guaranteed by the correct specification of the Q-function models. G-estimation and dWOLS are doubly-robust, but requires additional modeling of the treatment.

As already showed in section 2.3, Q-learning can be implemented in the presence of shared parameters of decision rules. Due to the natural resemblance of Q-learning and G-estimation, I want to extend G-estimation to the case with shared decision rules. I will fill this gap in the DTR literature in the next chapter.

Chapter 3

Shared G-estimation

3.1 Proposed Approach

In the previous chapter, the unshared parameter G-estimation was introduced. Here I show that it can be extended to the shared parameter case similar to Q-learning. The shared parameter G-estimation has previously been implemented by minimizing the squared Bellman error $J(\theta) = \|Y^*(\theta) - Z\theta\|^2$ [28]. As discussed in 2.3.2, an algorithm aimed directly at minimizing the squared Bellman error could be unstable. Thus I will proceed with an alternative approach.

As I have already addressed in shared Q-learning, due to the fact that the parameter ψ is shared through stages, we have to solve for $\hat{\psi}$ simultaneously across all stages rather than solving $\hat{\psi}_j$ backwards for each stage separately.

I propose the following estimating system of equations, instead of J different estimating equations:

$$U(\psi) = \begin{cases} \mathbb{P}_n(G_J(\psi) - E[G_J(\psi)|H_J; \hat{\beta}_J])(S_J(A_J) - E[S_J(A_J)|H_J; \hat{\alpha}_J]) & = 0 \\ \vdots & \\ \mathbb{P}_n(G_j(\psi) - E[G_j(\psi)|H_j; \hat{\beta}_j])(S_j(A_j) - E[S_j(A_j)|H_j; \hat{\alpha}_j]) & = 0 \\ \vdots & \\ \mathbb{P}_n(G_1(\psi) - E[G_1(\psi)|H_1; \hat{\beta}_1])(S_1(A_1) - E[S_1(A_1)|H_1; \hat{\alpha}_1]) & = 0 \end{cases}$$

where

$G_j(\psi) = \hat{Y}_j - \gamma_j(H_j, a_j; \psi) = \hat{Y}_j - \psi H_j^{\psi} A_j$ [1] ,

$E[G_j(\psi)|H_j = h_j; \beta_j] = \beta_j H_j^{\beta}$,

$S_j(A_j) = \frac{\partial \gamma_j}{\partial \psi} = H_j^{\psi} A_j$, and

$E[S_j(A_j)|H_j; \alpha_j] = E[H_j^{\psi} A_j | H_j^{\psi}; \alpha_j] = H_j^{\psi} \cdot E[A_j|H_j; \alpha_j]$.

Similar to unshared G-estimation, β_j can be expressed as a function of ψ via regression, and $E[A_j|H_j; \alpha_j]$ can be estimated separately from the treatment model. Unfortunately, unlike in unshared G-estimation, $\hat{Y}_j$ cannot be calculated by plugging in the estimates of ψ from subsequent stages since it is shared through all stages. Thus there is no closed form solution for $\hat{Y}_j$, and an iterative procedure needs to be employed instead.

The $\hat{\psi}$ could be estimated with the following fashion:

[1] $\hat{Y}_j = \hat{Y}_{j+1} + [-H_{j+1}^{\psi}\hat{\psi}A_{j+1} + (H_{j+1}^{\psi}\hat{\psi}A_{j+1})_+]$ or $\hat{Y}_j = Y + \sum_{k=j+1}^{J}[-H_k^{\psi}\hat{\psi}A_k + (H_k^{\psi}\hat{\psi}A_k)_+]$ is the pseudo-outcome.

Algorithm 3.1 Shared G-estimation

- Step 1: Propose initial value of ψ, denote as $\hat{\psi}^{(t=0)}$.
- Step 2: Propose the model of optimal blip-to-zero function for all stages ($j = 0, 1, \cdots, J$):

$$\gamma_j(H_j, A_j; \psi) = (\psi^T H_j^{\psi})A_j.$$

- Step 3: Calculate the pseudo-outcome for all stages:

$$\hat{Y}_J = Y, \text{ and}$$
$$\hat{Y}_j = \hat{Y}_{j+1} - H_{j+1}^{\psi}\hat{\psi}^{(t)}A_{j+1} + (H_{j+1}^{\psi}\hat{\psi}^{(t)}A_{j+1})_+$$

- Step 4: Define $G_j(\psi) = Y - \gamma_j(H_j, A_j; \psi) + \sum_{k=j+1}^{J}[\gamma_k(H_k, d_k^{opt}; \psi) - \gamma_k(H_k, A_k; \psi)]$, or equivalently:

$$G_j(\psi) = \hat{Y}_j - \gamma_j(H_j, A_j; \psi).$$

- Step 5: Propose the expected treatment-free outcome model $E[G_j(\psi)|H_j; \beta_j] = \beta_j H_j^{\beta}$, where β_j can be estimated as a function of $\psi^{(t)}$ with the estimating equation $E[H_j^{\beta}(Y_j - \psi^{(t)}H_j^{\psi}A_j - \beta_j H_j^{\beta})] = 0$.
- Step 6: Propose the treatment model: $E[S_j(A_j)] = E[H_j^{\psi}A_j|H_j]$ which is equivalent to proposing a model for $P_j(A_j = 1|H_j; \alpha_j)$, and estimate its parameters for all j.
- Step 7: Construct the estimating system of equations, which is constituted by the J stage-specific equations below,

$$U_j(\psi) = \mathbb{P}_n[G_j(\psi) - E[G_j(\psi)|H_j; \beta_j]] \cdot [S_j(A_j) - E[S_j(A_j)|H_j; \alpha_j]] = 0, \text{ or}$$
$$U_j(\psi) = \mathbb{P}_n[\hat{Y}_j - \gamma_j(H_j, A_j; \psi) - E[\hat{Y}_j - \gamma_j(H_j, A_j; \psi)|H_j]] \cdot [H_j^{\psi}A_j - E[H_j^{\psi}A_j|H_j; \alpha_j]] = 0.$$

- Step 8: Increment t, solve the step 7 system of equations for $\hat{\psi}$, and use it as the values of $\hat{\psi}^{(t+1)}$.
- Step 9: Re-calculate the pseudo-outcome using $\hat{\psi}^{(t+1)}$, and repeat the above procedures (from step 3 to step 8) till convergence: $\|\hat{\psi}^{(t+1)} - \hat{\psi}^{(t)}\| < \epsilon$.

With the estimates of ψ, we can identify the optimal DTRs: $d_j^{opt} = 1$ when $H_j\hat{\psi} \geq 0$ and $d_j^{opt} = 0$ otherwise. Just as in the shared Q-learning, there are various choices of the initial values that could be used for shared G-estimation, including a zero vector $\hat{\psi}^{(ZERO)}$, or estimates from other methods like unshared G-estimation: $\hat{\psi}^{(SA)}$, $\hat{\psi}^{(IVWA)}$, $\hat{\psi}^{(MAX)}$ or $\hat{\psi}^{(MIN)}$.

3.2 Example

Here are the details of shared G-estimation with a 3-stage example illustrated using a matrix formulation. Wallace (2016) [30] showed a more friendly presentation of G-estimation as follows:

Rewrite the estimating equation as

$$U_j(\psi) = \sum_{i=1}^{n} \{H_{ji}^{\psi}[\hat{Y}_{ji} - \psi^T H_{ji}^{\psi} A_{ji} - \beta_j^T H_{ji}^{\beta}] \cdot [A_{ji} - E[A_{ji}|H_{ji}; \hat{\alpha}_j]]\} = 0$$

Calculate $\hat{\beta}_j$ using OLS:

$$\textstyle\sum_{i=1}^{n} [H_{ji}^{\beta}(Y_{ji} - \psi^T H_{ji}^{\psi} A_{ji} - \beta_j^T H_{ji}^{\beta})] = 0.$$

Combining the above two equations together gives:

$$\begin{pmatrix} H_j^{\beta} \\ H_j^{\psi}(A_j - E[A_j|H_j^{\alpha}; \hat{\alpha}_j]) \end{pmatrix} (Y_j - A_j H_j^{\psi} \psi - H_j^{\beta} \beta_j) = 0 \tag{3.1}$$

where the predicting variables, tailoring variables and treatments: H_j^{β}, H_j^{ψ} and A_j are matrices with n rows for stage j.

For simplicity, re-defining $H_j^{\delta} = (H_j^{\beta}, A_j H_j^{\psi})$, $H_j^{\omega} = (H_j^{\beta}, (H_j^{\psi})^T (A_j - E[A_j|H_j^{\alpha}; \hat{\alpha}_j]))$, and $\delta_j = (\beta_j, \psi)$, rewrite (3.1) as:

$$(H_j^{\omega})^T (Y_j - H_j^{\delta} \delta_j) = 0.$$

Solve for δ_j: $\delta_j = ((H_j^{\omega})^T H_j^{\delta})^{-1} (H_j^{\omega})^T \hat{Y}_j$.

Now merge all the single-stage calculations into one, which could be done by the following:

With the estimates $\hat{\psi}^k$ (k starts from zero), we can calculate all three stages pseudo-outcome for kth iterations:

$\hat{Y}_3 = Y$,

$\hat{Y}_2 = Y - H_3^{\psi} \hat{\psi}^{(k)} A_3 + (H_3^{\psi} \hat{\psi}^{(k)})_+ = Y - H_3^{\psi} \hat{\psi}^{(k)} A_3 + \frac{1}{2}(H_3^{\psi} \hat{\psi}^{(k)} + |H_3^{\psi} \hat{\psi}^{(k)}|)$,

$\hat{Y}_1 = Y - H_3^{\psi} \hat{\psi}^{(k)} A_3 + (H_3^{\psi} \hat{\psi}^{(k)})_+ - H_2^{\psi} \hat{\psi}^{(k)} A_2 + (H_2^{\psi} \hat{\psi}^{(k)})_+$.

Denote

$$Y^* = \begin{pmatrix} \hat{Y}_3 \\ \hat{Y}_2 \\ \hat{Y}_1 \end{pmatrix},$$

$$H^\beta = \begin{pmatrix} H_1^\beta, 0, 0 \\ 0, H_2^\beta, 0 \\ 0, 0, H_3^\beta \end{pmatrix},$$

$$H^\psi = \begin{pmatrix} H_1^\psi \\ H_2^\psi \\ H_3^\psi \end{pmatrix}.$$

If vector ψ has different length among stages, just fill the blank in H^ψ up with 0, to ensure the shared parts are aligned. Then

$$AH^\psi = \begin{pmatrix} A_1 * H_1^\psi \\ A_2 * H_2^\psi \\ A_3 * H_3^\psi \end{pmatrix}, \text{ and}$$

$$H^\psi(A - E[A|\alpha]) = \begin{pmatrix} H_1^\psi * (A_1 - E[A_1|\alpha_1]) \\ H_2^\psi * (A_2 - E[A_2|\alpha_2]) \\ H_3^\psi * (A_3 - E[A_3|\alpha_3]) \end{pmatrix},$$

Let $H^\delta = (H^\beta, AH^\psi)$, $H^\omega = (H^\beta, H^\psi(A - E[A|\alpha]))$ and $\delta = (\beta_1, \cdots, \beta_J, \psi)$, then finally, $\hat{\delta} = ((H^\omega)^T H^\delta)^{-1}(H^\omega)^T Y^*$ will give the estimates of β_j and ψ.

3.3 Simulated Example

I will further illustrate this newly proposed shared G-estimation approach with a simple simulated example. But before that, I will introduce two characteristics that quantified the closeness between

estimated optimal DTR and true optimal DTR[2]. First, define the stage specific matching rate $M_j = P[d_j^{\hat{\psi}}(H_j) = d_j^{\psi}(H_j)]$, e.g., $M_1 = \mathbb{P}[(\hat{\psi}_0 + \hat{\psi}_1 \cdot X_1 > 0) = (\psi_0 + \psi_1 \cdot X_1 > 0)]$. Averaging M_j over stages then we have $\overline{M} = \frac{\sum_{j=1}^{J} M_j}{J}$. Secondly, define the overall matching rate over all stages $\tilde{M} = P[d^{\hat{\psi}} = d^{\psi}]$, an overall matching implies the treatment decision from the estimated DTR agrees with the true DTR for *all* stages for a subject. Thus the stage specific matching rate $\overline{M}$ and overall matching rate $\tilde{M}$ can be used for measuring the proportion of subjects which their decision trajectories guided by estimated optimal DTR agreed with those guided by the true unknown optimal DTR.

Data Generation

The longitudinal data have 3 stages, and no subject will drop out at either stages, i.e. we will have the same number of subjects throughout all three stages. The data contains 200, 500 or 2000 subjects, and we will generate 1000 samples of data for mean of the estimates and variance. The data are simulated with the following setup:

- Covariates X_j where $j = 1, 2, 3$

$$X_1 \sim \text{Normal}(10, 5),$$
$$X_2 \sim \text{Normal}(1.25 \cdot X_1, 5),$$
$$X_3 \sim \text{Normal}(X_1 + X_2, 5).$$

- Treatments $A_j \in \{0, 1\}$,

$$A_j \sim \text{Bernoulli}(\pi_j),$$

where

$$\text{logit}(\pi_1) = 0.05 \cdot X_1,^3$$
$$\text{logit}(\pi_2) = -0.05 \cdot X_2,$$
$$\text{logit}(\pi_3) = 0.02 \cdot X_3.$$

[2]Of course, we could always compare $\hat{\psi}$ with ψ, but that is less intuitive and sometimes could not fully reflect how the estimated optimal DTR differs from the true one in terms of each decision making.

- Blips γ_j:

$$\gamma_1 = A_1 \cdot (\psi_0 + \psi_1 X_1),$$

$$\gamma_2 = A_2 \cdot (\psi_0 + \psi_1 X_2 + \psi_2 A_1),$$

$$\gamma_3 = A_3 \cdot (\psi_0 + \psi_1 X_3 + \psi_{_2} A_2 + \psi_3 A_1 X_1).$$

The regrets μ_j:

$$\mu_j = \max \gamma_j - \gamma_j.$$

- Values of ψs: $(\psi_0,\ \psi_1,\ \psi_2,\ \psi_3) = (8,\ -1.2,\ 8,\ 3)$.

- Outcome:

$$Y = \text{Normal}(\beta_{10} + \beta_{11} X_1, 60) - \mu_1 - \mu_2 - \mu_3, \ ^4$$

where $\beta_{10} = 30$, $\beta_{11} = 3$. The observed dataset contains only X_1, A_1, X_2, A_2, X_3, A_3 and Y.

Model Specification

Three models need to be specified at each stage:

1. Blip model: $\gamma(H, A; \psi)$.

2. Treatment-free outcome model: $G(H; \beta)$.

3. Treatment model: $E[A|H; \alpha]$

Table 3.1 summarizes two scenarios for model specifications considered in the analysis of the simulated data. Because the shared parameter G-estimation is a doubly robust method, analysis 1 will have unbiased estimates of ψ and analysis 2 will not. We will use simulation to verify that [5].

[3]This is equivalent to $A_1 \sim \text{Bernoulli}(\text{expit}(0.05 \cdot X_1))$.

[4]This is from $G_1(\psi) = Y + \mu_1 + \mu_2 + \mu_3$ with $E[G_1|X_1] = \beta_{10} + \beta_{11} X_1$.

[5]We could also have an analysis 3 with correct $G(H; \beta)$ and incorrect $\mathbb{E}[A|H; \alpha]$. However, with the way we generate the data and primary outcome Y, we could only know the true form of stage 1 treatment-free outcome model $\mathbb{E}[G_1(H_1; \beta_1)] = \beta_{10} + \beta_{11} X_1$, but not for stage 2 and 3. It is difficult to derive the true model for these under realistic data-generating scenarios [11].

Table 3.1: Two model specifications of G-estimation

	$\gamma(H, A; \psi)$	$G(H; \beta)$	$\mathbb{E}[A\|H; \alpha]$	ψ
Analysis 1	✓	×	✓	✓
Analysis 2	✓	×	×	×

For both analysis 1 and 2, the blip models are correct are specified as follows:

$$\gamma_1 = A_1 \cdot (\psi_0 + \psi_1 X_1),$$
$$\gamma_2 = A_2 \cdot (\psi_0 + \psi_1 X_2 + \psi_2 A_1),$$
$$\gamma_3 = A_3 \cdot (\psi_0 + \psi_1 X_3 + \psi_2 A_2 + \psi_3 A_1 X_1).$$

Analysis 1: Correct treatment, incorrect treatment-free outcome

Treatment models were correctly specified. And the treatment-free outcome models were assumed with the following covariate specifications: $H_j^\beta = (1, X_j)$ for $j = 1, 2, 3$.

Analysis 2: Incorrect treatment, incorrect treatment-free outcome

The incorrect treatment model was assumed, with the analyst positing $E[A_j] = 0.1$ for all three intervals. As in Analysis 1, the following specification was used for the treatment-free models: $H_j^\beta = (1, X_j)$ for $j = 1, 2, 3$.

Results

The analysis results are summarized in Table 3.2, and both analyses takes zero as initial value[6].

From Table 3.2, it is clear that analysis 1, the one with correct treatment model has unbiased estimates and higher matching rate in contrast to analysis 2; even though the treatment-free outcome model is incorrect. This shows the shared parameter G-estimation has the robustness as we desire. In addition to that, larger sample size can lower the bias of estimates and the variance as well. With

[6]The standard deviations are showed in brackets after their estimates.

Table 3.2: Estimates and concordance of estimated and true optimal treatments averaged across stages ($\overline{M}$) or in all stages ($\tilde{M}$).

	Sample Size	$\hat{\psi}_0$ ($\psi_0 = 8$)	$\hat{\psi}_1$ ($\psi_1 = -1.2$)	$\hat{\psi}_2$ ($\psi_0 = 8$)	$\hat{\psi}_3$ ($\psi_3 = 3$)	$\overline{M}$	$\tilde{M}$
Analysis 1	n=200	9.717 (10.616)	-1.163 (0.706)	5.450 (14.424)	2.871 (1.548)	0.933	0.810
	n=500	8.517 (7.180)	-1.185 (0.463)	7.068 (9.676)	2.962 (0.981)	0.973	0.922
	n=2000	8.023 (3.614)	-1.201 (0.231)	7.700 (5.246)	2.981 (0.470)	0.981	0.942
Analysis 2	n=200	10.500 (8.805)	-0.897 (0.558)	1.155 (9.639)	1.799 (0.882)	0.810	0.500
	n=500	8.950 (5.798)	-0.868 (0.368)	1.561 (6.114)	1.777 (0.533)	0.854	0.634
	n=2000	8.062 (2.963)	-0.868 (0.186)	1.676 (3.286)	1.804 (0.245)	0.870	0.668

the correct treatment model, 500 subjects are enough for gaining a 92.2% correct decision rate for this specific case. We further note that, although the first stage treatment-free model is correctly specified, parameter estimates are biased due to contamination due to the incorrectly-specified second and third stage models in analysis 2.

3.4 Summary

In this chapter, I have discussed a new computational approach to estimate the optimal DTRs using G-estimation in the presence of shared parameters ψ. The estimators $\hat{\psi}$ are consistent with either the treatment-free outcome model $E[G_j(\psi)|H_j = h_j; \beta_j]$ or treatment model $p(A_j|H_j; \alpha_j)$ correctly specified; thus it is doubly robust [31]. Note there could be other good choice of $S_j(A_j)$ besides $S_j(A_j) = \frac{\partial \gamma_j}{\partial \psi}$ to provide more efficient estimates, but the optimal form of $S_j(A_j)$ depends on the knowledge of the variance of the outcome which is usually difficult to know [11]. This new approach permits one to estimate the optimal DTR with a doubly robust method, which is superior to shared Q-learning.

As I noted in Chapter 2, for G-estimation (unshared or shared), one can split the expected outcome (Q-function model) into the sum of two components: $\beta^T H^\beta + A \cdot (\psi^T H^\psi)$, the first component $\beta^T H^\beta$ is the impact of patient history in absence of treatments, named as treatment-free outcome model; and the second component $A \cdot (\psi^T H^\psi)$ is the impact of treatment and named as blip model. In Q-learning one

aim to find the optimal DTRs that maximize the Q-function. However, in G-estimation, researchers are looking for the optimal DTRs that maximize $A \cdot (\psi^T H^\psi)$. Of course, under correct specification of the treatment free model, the same parameters ψ will maximize both the Q-function and the blip function. Q-learning (unshared or shared) is a singly robust method, where consistency of the estimators relies on the specification of the Q-function model. In contrast, G-estimation (unshared or shared) is a doubly robust method: as long as one of the treatment-free outcome model or treatment model is correct[7], the estimators are consistent. I have showed this property through a simple simulation by comparing results from different model specifications.

In this chapter, I have implemented and demonstrated shared G-estimation on the simulated data. In the next chapter, performance of the estimators will be compared among different methods.

[7]We always assume the blip model is correct in this thesis.

Chapter 4

Simulation Study

In this chapter, I will conduct several simulations comparing different DTR estimating methods including unshared Q-learning, shared Q-learning, unshared G-estimation and shared G-estimation with different model specifications. For simplicity, the simulation will consider two stages, although it is straightforward to implement the methods in more stages problem as shown in chapter 3.

4.1 Data Generation

The simulated data have 2 stages, and no subject is lost to follow-up. The dataset contains 200, 500 or 2000 subjects respectively, and I will generate 1000 samples of data for mean and variance of the estimates. The data are generated as follows:

- Covariates X_1 and X_2:

$$X_1 \sim \text{Normal}(10, 5)$$

$$X_2 \sim \text{Normal}(1.25 \cdot X_1, 5)$$

- Treatments A_1 and A_2, $A_j \in \{0, 1\}$:

$$A_1 \sim \text{Bernoulli}(\text{expit}(0.05 \cdot X_1))$$

$$A_2 \sim \text{Bernoulli}(\text{expit}(-0.05 \cdot X_2))$$

- Blips γ_1 and γ_2:

$$\gamma_1 = A_1 \cdot (\psi_0 + \psi_1 X_1)$$

$$\gamma_2 = A_2 \cdot (\psi_0 + \psi_1 X_2 + \psi_2 A_1)$$

The regrets μ_j:

$$\mu_j = \max \gamma_j - \gamma_j$$

- Values of ψs: $\psi_0 = 8$, $\psi_1 = -1.2$ and $\psi_2 = 8$.

- Outcome:

$$Y = \text{Normal}(\beta_{10} + \beta_{11} X_1, 60) - \mu_1 - \mu_2, \text{ where } \beta_{10} = 30 \text{ and } \beta_{11} = 3.$$

The dataset contains the following variables: X_1, A_1, X_2, A_2 and Y.

4.2 Model Specification

Both analysis 1 and analysis 2 are based on the correct blip models and treatment models, but with different treatment-free outcome models, just as in the first analysis of the previous chapter. However, now we consider two incorrect forms of the treatment-free model at the second interval, where one is more flexible and hence possible closer to the true model:

- Correct blip models:

$$\gamma_1 = A_1 \cdot (\psi_0 + \psi_1 X_1)$$

$$\gamma_2 = A_2 \cdot (\psi_0 + \psi_1 X_2 + \psi_2 A_1)$$

- Correct treatment models [1]:

$$\text{logit}(\pi(A_j = 1)) = \log(\tfrac{\pi(A_j=1)}{1-\pi(A_j=1)}) = \alpha_j X_j,\ j = 1, 2.$$

[1] The treatment models are only provided for G-estimation but not Q-learning

Analysis 1:

- Simple, linear (non-flexible) treatment-free outcome models have the covariate specifications as follows:

$$H_1^{\beta} = (1, X_1)$$

$$H_2^{\beta} = (1, X_2)$$

- Q-Model:

$$Q_1 = \beta_{10} + \beta_{11} \cdot X_1 + A_1 \cdot (\psi_0 + \psi_1 \cdot X_1)$$

$$Q_2 = \beta_{20} + \beta_{21} \cdot X_2 + A_2 \cdot (\psi_0 + \psi_1 \cdot X_2 + \psi_2 \cdot A_1)$$

Analysis 2:

- Flexible treatment-free outcome models have the covariate specifications as follows:

$$H_1^{\beta} = (1, X_1)$$

$$H_2^{\beta} = (1, X_1^2, A_1 X_1^2, X_2, X_2^2)$$

- Q-Model:

$$Q_1 = \beta_{10} + \beta_{11} \cdot X_1 + A_1 \cdot (\psi_0 + \psi_1 \cdot X_1)$$

$$Q_2 = \beta_{20} + \beta_{21} \cdot X_1^2 + \beta_{22} \cdot A_1 \cdot X_1^2 + \beta_{23} \cdot X_2 + \beta_{24} \cdot X_2^2 + A_2 \cdot (\psi_0 + \psi_1 \cdot X_2 + \psi_2 \cdot A_1)$$

In analysis 2, I add more explanatory variables to fit the treatment-free outcome model, to assess whether it might increase the accuracy of the estimates. In both analysis 1 and analysis 2, the Q-models are incorrectly specified (See the supplement of [11] for the correct functional form of Q-models, which depends on non-linear functions of the blip model parameters).

4.3 Results

In this section, I first display the convergence plot for the shared G-estimation with different initial values, those initial values are the zero, simple average (SA) and inverse variance weighted average (IVWA) estimates from unshared G-estimation.

Figure 4.1 shows that the estimates of shared G-estimation procedure with different initial values converge after only a few iterations. Thus, we will only display the estimates of shared G-estimation with simple average as initial values in Table 4.1 and Table 4.2[2].

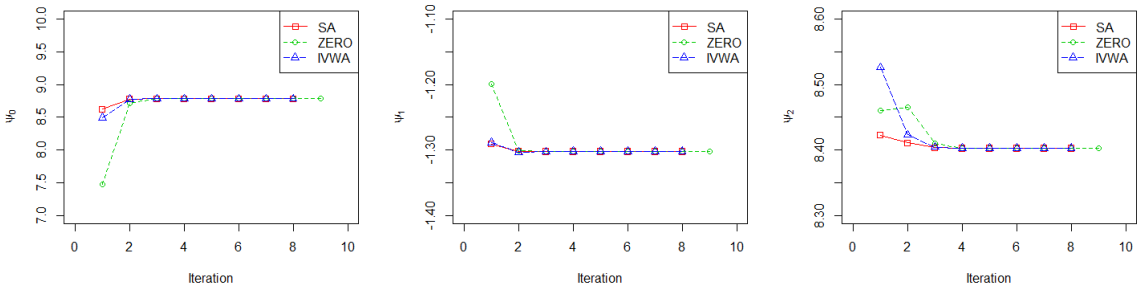

Figure 4.1: Convergence patterns of ψ_0, ψ_1 and ψ_2 for shared G-estimation with three different initial values

For unshared Q-learning and unshared G-estimation, I display their simple average estimates instead of stage specific estimates. For shared Q-learning, the estimates are calculated with the initial value zero. And for shared G-estimation, I choose their simple average estimates from unshared G-estimation for the initial value. The estimated results are showed in Tables 4.1 and 4.2[3].

[2]Actually, the estimates of shared G-estimation for this simulated example are exactly the same with different initial values.

[3]The standard deviations are showed in brackets after their estimates.

Table 4.1: Analysis 1: Non-flexible treatment-free outcome model, $\psi_0 = 8$, $\psi_1 = -1.2$ and $\psi_2 = 8$

Sample Size	Method	$\hat{\psi}_0$	$\hat{\psi}_1$	$\hat{\psi}_2$	$\overline{M}$	$\tilde{M}$
n=200	unshared Q	7.087 (14.404)	-1.003 (1.082)	3.032 (15.411)	0.935	0.872
	shared Q	8.151 (9.577)	-0.706 (0.823)	0.907 (9.144)	0.764	0.583
	unshared G	6.969 (14.858)	-1.127 (1.110)	8.706 (19.317)	0.973	0.947
	shared G	9.160 (15.171)	-1.125 (1.210)	6.885 (14.100)	0.925	0.855
n=500	unshared Q	6.416 (9.063)	-0.947 (0.690)	2.875 (9.493)	0.942	0.884
	shared Q	6.896 (5.457)	-0.658 (0.523)	0.920 (5.857)	0.801	0.650
	unshared G	6.670 (9.260)	-1.106 (0.705)	8.311 (11.678)	0.973	0.947
	shared G	7.738 (8.694)	-1.151 (0.687)	7.656 (7.993)	0.997	0.994
n=2000	unshared Q	7.358 (4.523)	-1.013 (0.351)	2.698 (4.644)	0.924	0.852
	shared Q	6.815 (3.109)	-0.680 (0.260)	1.260 (2.864)	0.823	0.686
	unshared G	7.932 (4.544)	-1.197 (0.358)	7.932 (6.056)	0.997	0.995
	shared G	8.183 (4.151)	-1.212 (0.329)	7.996 (4.304)	0.996	0.993

Table 4.2: Analysis 2: Flexible treatment-free outcome model, $\psi_0 = 8$, $\psi_1 = -1.2$ and $\psi_2 = 8$

Sample Size	Method	$\hat{\psi}_0$	$\hat{\psi}_1$	$\hat{\psi}_2$	$\overline{M}$	$\tilde{M}$
n=200	unshared Q	7.907 (14.982)	-1.177 (1.137)	7.243 (16.441)	0.991	0.982
	shared Q	9.891 (10.424)	-1.235 (1.005)	6.729 (10.993)	0.941	0.886
	unshared G	7.089 (14.866)	-1.139 (1.112)	8.426 (19.186)	0.979	0.958
	shared G	9.201 (15.208)	-1.136 (1.206)	6.901 (13.780)	0.928	0.861
n=500	unshared Q	7.531 (9.340)	-1.146 (0.730)	6.849 (9.955)	0.984	0.969
	shared Q	8.840 (6.546)	-1.174 (0.626)	6.331 (6.707)	0.958	0.918
	unshared G	6.760 (9.170)	-1.115 (0.704)	8.289 (11.466)	0.975	0.950
	shared G	7.760(8.605)	-1.156 (0.685)	7.666 (7.968)	0.998	0.996
n=2000	unshared Q	8.543 (4.707)	-1.224 (0.378)	6.775 (5.002)	0.975	0.950
	shared Q	8.842 (3.423)	-1.208 (0.312)	6.834 (3.452)	0.968	0.936
	unshared G	7.937 (4.519)	-1.198 (0.357)	7.945 (5.995)	0.997	0.994
	shared G	8.174 (4.136)	-1.212 (0.327)	8.004 (4.285)	0.997	0.994

As one can see from Tables 4.1 and 4.2, larger sample size can provide less estimation bias and lower variance for all methods, although the shared Q-learning in particular remains biased with the simple, incorrectly specified Q-function. In addition, larger sample size also leads to higher averaged stage specific and overall matching rates $\overline{M}$ and $\tilde{M}$. In this specific simulated example, the choice of initial values for the shared G-estimation did not have any impact on the final estimates, but the initial values that were closer to the true values can result in less computation time. In general, the G-estimation methods outperform the Q-learning methods in terms of both bias and matching rates, due to the incorrect specifications of Q-models and the robustness of G-estimation. Although the unshared and shared Q-learning are biased procedures in this case, we can still reduce the bias by implementing a more flexible outcome model.

Other than the general discussions above, there is also one interesting finding I think worth to bring up in here: the unshared Q-learning methods performs better than shared Q-learning in this case. Under the proposed data generation scheme, the Q_1 function is correctly specified since both treatment-

free outcome model and blip model are correct, and Q_2 function is not due to the unknown stage 2 treatment-free outcome model. Because the unshared Q-learning solves Q_1 and Q_2 separately and the shared Q-learning solves them simultaneously, the incorrect Q_2 might compromises the overall estimation in the shared Q-learning method. The G-estimation has assured consistent estimates, although the extra treatment model could lead to small increase in variance of the estimators.

4.4 Summary

In this chapter, I have conducted several simulation studies with different settings and showed the superiority of the newly proposed shared G-estimation algorithm as compared to Q-learning and unshared G-estimation in the presence of shared parameters. By introducing the averaged stage specific matching rate $\overline{M}$ and overall matching rate $\tilde{M}$, I compared different DTR estimation methods not only through the estimates themselves, but also via those matching rates. In the next chapter I will implement the above-considered methods on a real dataset.

Chapter 5

Data Analysis

In the previous chapters, I have conducted simulations with known data generation mechanism. However, the true blip models, treatment-free outcome models, and treatment models are often unknown with real data.

As I introduced earlier, the SMART design was developed to provide high-quality data to construct DTRs. By randomizing patients multiple times, researchers are able to assess effectiveness of treatment for each stage without fear of confounding. In this chapter, I will apply the proposed shared G-estimation method to data from a SMART design.

5.1 The Sequenced Treatment Alternatives to Relieve Depression Study

The sequenced treatment alternatives to relieve depression study, or STAR*D, funded by NIMH is a SMART aimed to assess the effectiveness of depression treatments in patients diagnosed with major depressive disorder under different treatment regimes. The study was conducted over a seven-year period, enrolled 4,041 participants, and all the participants were diagnosed with major depressive disorder[1].

[1] https://www.nimh.nih.gov/funding/clinical-research/practical/stard/index.shtml

The scheme for treatment assignment is given in Figure 5.1. At level 1, all the patients were treated with citalopram (CIT). Those who responded well to the CIT treatment remained on this treatment[2]; and those who did not respond well to the CIT treatment moved to level 2. At level 2, depending on their preference, the participants could either choose to switch or augment. Patients who chose switch were randomly assigned to one of four treatments: bupropion (BUP), cognitive psychotherapy (CT), sertraline (SER), or venlafaxine (VEN). Those who chose augmentation were randomly assigned to one of three options: CIT+BUP, CIT+buspirone (BUS) or CIT+CT. For those who received CT or CIT+CT at level 2, if the response were unsatisfactory, they moved to the supplementary level 2a. At level 2a, they were randomized to either BUP or VEN. Participants who did not respond well at level 2 or level 2a would move to level 3. At level 3, based on their preference, patients were randomly assigned to switch to either mirtazapin (MIRT) or nortriptyline (NTP); or randomly assigned to augment their previous treatment with lithium (Li) or thyroid hormone (THY). Participants unsatisfied with level 3 treatment continued to level 4 treatments, with options of either tranylcypromine (TCP) or MIRT+VEN. The severity of depression was assessed based on the Quick Inventory of Depressive Symptomatology (QIDS). The effectiveness of treatments was deemed to be satisfied when the patients achieved remission (QIDS< 5) [4].

The selective serotonin reuptake inhibitors (SSRIs) are the most common class of antidepressants prescribed to the patients suffered from depression, however few studies have focused on the treatment regimes [32]. Here we want to conduct an analysis to study the optimal DTRs with SSRIs and non-SSRIs for treating the major depressive disorder. For this purpose, we classified all the level 2 treatments into two categories: (i) treatment involving SSRI (alone or combined): SER, CIT+BUP, CIT+BUS and CIT+CT or (ii) treatment with non-SSRI (one or more): VEN, BUP or CT alone. Similar for the treatments on level 3, (i) treatment with SSRIs: augmentation of any SSRI level 2 treatment with either Li or THY and (ii) treatment with non-SSRI: MIRT, NTP or augmentation of any non-SSRI level 2 treatment with either Li or THY. We summary the scheme of study design in Figure 5.1.

[2]This is the same throughout the trial: patients who are benefiting from a treatment are not switched away from that treatment for ethical reasons.

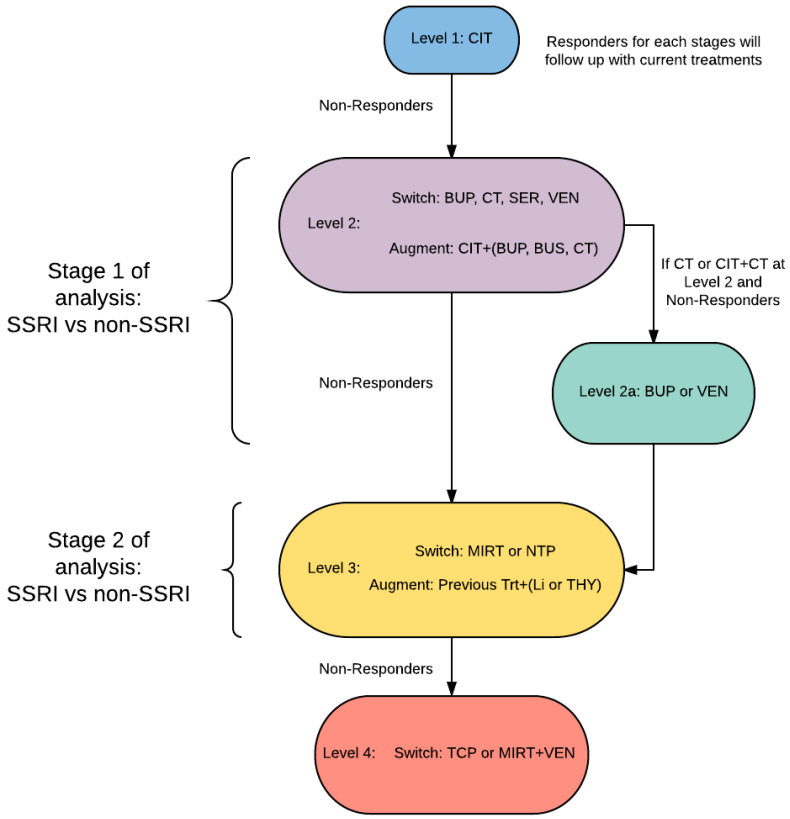

Figure 5.1: The scheme for treatment assignment in the STAR*D study

5.2 Models and Analysis

Our model framework contains two stages[3]: level 2 and level 2a together as stage 1, and level 3 as stage 2. Participants who took an SSRI treatment at level 2 but a non-SSRI at level 2a were considered receiving SSRI at stage 1.

In addition, the following notations are introduced:

- The stage specific outcomes: Y_1, Y_2 are negative QIDS scores at the end of stage 1 and 2.

[3]Level 4 data were not included in this analysis. Level 1 data were not included because all patients received the same treatment.

- The primary outcomes: $Y = R_1 \cdot Y_1 + (1 - R_1) \cdot (\frac{Y_1+Y_2}{2})$, where $R_1 = 1$ if the patient achieved remission (QIDS$\leq$5) at the end of stage 1, and $R_1 = 0$ otherwise.
- QS_1, QS_2 are QIDS scores at the start of stage 1 and 2 (so $QS_2 = -Y_1$).
- S_1, S_2 are QIDS slopes over the preceding interval (change in score/time).
- A_1, A_2 are treatments with 0 = non-SSRI, 1 = SSRI.
- P_1, P_2 are preferences, coded 1 for preference to switch and 0 otherwise, the preferences can also be viewed as the side effects in this study[4].
- Other personal data of the participants: age, sex, race, years of schooling completed, employment category and private insurance (yes/no); denote them as Age, Sex, Race, School, Emplcat and Privins, respectively.

Here I think it is reasonable to believe QS_1 and QS_2 share the same parameter ψ_1 between two stages, and similar for S_1 and S_2. Thus it seem appropriate to use shared parameter method.

I proposed the following models for Q-functions:

$$Q_1 = \beta_{10} + \beta_{11} \cdot QS_1 + \beta_{12} \cdot S_1 + \beta_{13} \cdot P_1 + A_1 \cdot (\psi_0 + \psi_1 \cdot QS_1 + \psi_2 \cdot S_1 + \psi_3 \cdot P_1),$$

$$Q_2 = \beta_{20} + \beta_{21} \cdot QS_2 + \beta_{22} \cdot S_2 + \beta_{23} \cdot P_2 + \beta_{24} \cdot A_1 + A_2 \cdot (\psi_0 + \psi_1 \cdot QS_2 + \psi_2 \cdot S_2).$$

Thus the models for blip function are:

$$\gamma_1 = A_1 \cdot (\psi_0 + \psi_1 \cdot QS_1 + \psi_2 \cdot S_1 + \psi_3 \cdot P_1),$$
$$\gamma_2 = A_2 \cdot (\psi_0 + \psi_1 \cdot QS_2 + \psi_2 \cdot S_2),$$

and the treatment-free outcome models are:

$$E[G_1|H_1; \beta_1] = \beta_{10} + \beta_{11} \cdot QS_1 + \beta_{12} \cdot S_1 + \beta_{13} \cdot P_1,$$
$$E[G_2|H_2; \beta_2] = \beta_{20} + \beta_{21} \cdot QS_2 + \beta_{22} \cdot S_2 + \beta_{23} \cdot P_2 + \beta_{24} \cdot A_1.$$

[4]Since patients who experienced side effects usually tended to switch to other treatments instead of to augment.

For G-estimation, we also need to specify the treatment models.

Treatment models 1:

$$\text{logit}(\pi_1) = \alpha_{11} \cdot \text{Age} + \alpha_{12} \cdot \text{Sex} + \alpha_{13} \cdot QS_1 + \alpha_{14} \cdot S_1 + \alpha_{15} \cdot P_1,$$

$$\text{logit}(\pi_2) = \alpha_{21} \cdot \text{Age} + \alpha_{22} \cdot \text{Sex} + \alpha_{23} \cdot QS_2 + \alpha_{24} \cdot S_2 + \alpha_{25} \cdot P_2.$$

Also we could add more explanatory covariates to the above logistic regressions;

Treatment models 2:

$$\text{logit}(\pi_1) = \alpha_{11}\text{Age}+\alpha_{12}\text{Sex}+\alpha_{13}\text{Race}+\alpha_{14}\text{School}+\alpha_{15}\text{Emplcat}+\alpha_{16}\text{Privins}+\alpha_{17}QS_1+\alpha_{18}S_1+\alpha_{19}P_1$$

$$\text{logit}(\pi_2) = \alpha_{21}\text{Age}+\alpha_{22}\text{Sex}+\alpha_{23}\text{Race}+\alpha_{24}\text{School}+\alpha_{25}\text{Emplcat}+\alpha_{26}\text{Privins}+\alpha_{27}QS_2+\alpha_{28}S_2+\alpha_{29}P_2.$$

With the above models, I will implement four different methods to estimate the optimal DTRs: unshared Q-learning, unshared G-estimation, shared Q-learning with zero as initial values and shared G-estimation with zero as initial values.

5.3 Results

With the necessary manipulation and cleaning of the original data, there are 1,159 patients in stage 1 and 273 patients in stage 2 left in the study. This large reduction in sample size from the original 4,041 is primarily due to a great portion of patients responding to treatment in Levels 1 and 2 of the STAR*D study. In order to calculate the variance of the estimates, I will bootstrap the data with the same sample size 1000 times.

The parameters estimates and their bootstrap standard errors are summarized in Table 5.1. Just as I did in Chapter 4, the parameter estimates of unshared methods are only displayed through their simple average transformations; it is easier to compare estimates from unshared methods and those from shared methods by doing so.

Table 5.1: STAR*D analysis with SSRI and non-SSRI treatments at two stages

		$\hat{\psi}_0$	$\hat{\psi}_1$	$\hat{\psi}_2$	$\hat{\psi}_3$
unshared Q		-0.254 (0.850)	0.056 (0.069)	0.263 (0.389)	-2.644 (1.111)
shared Q		-0.674 (1.026)	-0.451 (0.018)	0.568 (0.113)	0.2173 (0.131)
unshared G	Treatment model 1	-0.674 (1.026)	0.086 (0.081)	-0.039 (0.489)	-2.593 (1.175)
	Treatment model 2	-0.633 (1.044)	0.088 (0.084)	-0.097 (0.492)	-2.681 (1.182)
shared G	Treatment model 1	-0.829 (1.012)	0.093 (0.080)	-0.208 (0.507)	-2.652 (1.149)
	Treatment model 2	-0.776 (1.042)	0.098 (0.082)	-0.225 (0.527)	-2.702 (1.150)

Overall, the results of the four G-estimation methods are broadly similar, with the sign and magnitude of each estimate agreeing quite well. In contrast, the unshared Q-learning estimate of ψ_0 is less than half the magnitude of all other estimates, and the estimate of ψ_1 and ψ_3 for shared Q-learning differs in sign from all other approaches.

The estimated optimal DTRs suggested by the shared G-estimation (with treatment model 1) are: For a patient with depression, if $-0.829+0.093\cdot QS_1-0.208\cdot S_1-2.652\cdot P_1 \geq 0$ at stage 1, then prescribe the treatments with SSRI; otherwise prescribe the treatments with non-SSRI. If $-0.829+0.093\cdot QS_2-0.208\cdot S_2 \geq 0$ at stage 2, then prescribe the treatments with SSRI; otherwise prescribe the treatments with non-SSRI.

5.4 Summary

In this chapter, I demonstrated an application of the newly proposed shared G-estimation method, along with three existing methods to a real dataset, to examine the optimal treatment strategy for depressive disorder. For G-estimation (unshared and shared), different treatment models resulted in slightly different estimates in this study, though given the randomized nature of the study and the similarity of the estimates, it is likely that the double robustness property has assured consistent estimates. The estimates of unshared G-estimation and shared G-estimation were not far from to each other. How-

ever the estimates of shared Q-learning were quite different from those of other three methods; which agreed with what I have found in the simulation study: even mis-specifying the Q-model from one stage can lead to poorer performance of the shared Q-learning overall relative to the unshared Q-learning.

From the optimal DTRs estimated by shared G-estimation (with treatment model 1), we can conclude that, if the QIDS score is high, the optimal DTRs tended to suggest treatments with SSRI; and if the QIDS score changed dramatically over the preceding stage or the patients suffered from side effect, the rule tended to suggest treatments without SSRI.

Chapter 6

Conclusion and Discussion

In this thesis, I have reviewed some popular methods for estimating optimal DTRs, and then extended G-estimation for unshared parameters to the shared parameter setting. By doing so, this new proposed approach allows the parameters of blip functions to be the same through different stages; this is especially useful when the different stage's blip functions have similar structure and components.

From the simulation study and data analysis, I have shown that the newly proposed shared G-estimation algorithm produces consistent estimators under assumptions of no unmeasured confounding and certain model specifications being correct. Further, I observed that shared G-estimation performs better than shared Q-learning in general with the extra modeling. The key reason for this phenomenon is the double robustness property of G-estimation. But on the other hand, the G-estimation could sometimes suffer from small increase in variance of the estimators over Q-learning with finite samples, because of the extra treatment models.

In the process of completing this thesis, there were some interesting findings may provide avenues for further research.

I have discussed shared parameters ψ in the blip models, and I posit that it will not be too complicated to take it one step further, which is considering shared parameters β in the treatment-free outcome model. While the β parameters are nuisance parameters and not of direct interest, it may be of value to estimate these parsimoniously. In this thesis, I have only used the linear model for both blip model

and treatment-free outcome model, future work would extend this to the non-linear case. As I have introduced in section 2.5, dWOLS is another regression-based method for estimating the optimal DTR, I would like to extend the shared parameter method to dWOLS as well.

Bibliography

[1] Frank Emmert-Streib. Personalized medicine: Has it started yet? A reconstruction of the early history. *Frontiers in Genetics*, 3:313, 2012.

[2] F. Randy Vogenberg, Carol Isaacson Barash, and Michael Pursel. Personalized Medicine: Part 1: Evolution and Development into Theranostics. *Pharmacy and Therapeutics*, 35(10):560–562, 565–567, 576, 2010.

[3] Kelley M. Kidwell. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*, chapter 2: DTRs and SMARTs: Definitions, designs, and applications. In Chakraborty and Moodie [33], 2013.

[4] Bibhas Chakraborty and Erica E.M. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer, New York, NY, 2013.

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

[6] James M. Robins, Miguel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

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

[8] Watkins Chris. *Learning from Delayed Rewards*. PhD thesis, Kings College, 1989.

[9] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

[10] James M. Robins. *Optimal Structural Nested Models for Optimal Sequential Decisions*, volume 179, pages 189–326. Springer, New York, NY, 2004.

[11] Bibhas Chakraborty, Palash Ghosh, Erica E.M. Moodie, and Augustus John Rush. Estimating optimal shared-parameter dynamic regimens with application to a multistage depression clinical trial. *Biometrics*, 72(3):865–76, 2016.

[12] James M. Robins. *Marginal Structural Models versus Structural nested Models as Tools for Causal inference*, volume 116, pages 95–133. Springer, New York, NY, 2000.

[13] Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68:1010–1018, 2012.

[14] Yingqi Zhao, Donglin Zeng, Augustus John Rush, and Michael R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(449):1106–1118, 2012.

[15] Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R. Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.

[16] Constance Reid. *Neyman*, page 45. Springer, New York, NY, 1998.

[17] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, page 688, 1974.

[18] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

[19] Philip W. Lavori and Ree Dawson. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38, 2000.

[20] Susan A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481, 2005.

[21] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition edition, 2009.

[22] Arthur S. Goldberger. Structural equation models in the social sciences. *Econometrica: Journal of the Econometric Society*, 40(6):979–1001, 1972.

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

[24] Donald B. Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[25] Sander Greenland and James M. Robins. Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives and Innovations*, 6(1):4, 2009.

[26] Geoffrey J. Gordon. *Approximate Solutions to Markov Decision Processes*. PhD thesis, Carnegie Mellon University, 1999.

[27] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, Apr 2008.

[28] Erica E. M. Moodie, Thomas S. Richardson, and David A. Stephens. Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455, 2007.

[29] Michael P. Wallace and Erica E. M. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*, chapter 6: Analysis in the single-stage setting: An overview of estimation approaches for dynamic treatment regimes. In Chakraborty and Moodie [33], 2013.

[30] Michael P. Wallace, Erica E.M. Moodie, and David A. Stephens. Model assessment in dynamic treatment regimen estimation via double robustness. *Biometrics*, 72:855–64, 2016.

[31] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121, 1995.

[32] J. Craig Nelson. Safety and tolerability of the new antidepressants. *Journal of Clinical Psychiatry*, 58(6):26–31, 1997.

[33] Michael R. Kosorok and Erica E.M. Moodie. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. ASA–SIAM, 2016.