Robust Semiparametric Regression Estimation Using Targeted Maximum Likelihood with Application to Biomarker Discovery and Epidemiology

by

Catherine Ann Tuglus

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Biostatistics and the Designated Emphasis in Computational and Genomic Biology in the GRADUATE DIVISION of the UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Mark J. van der Laan, Chair
Professor Alan E. Hubbard
Professor Adam P. Arkin

Fall 2010
Abstract

Robust Semiparametric Regression Estimation Using Targeted Maximum Likelihood with Application to Biomarker Discovery and Epidemiology

by

Catherine Ann Tuglus
Doctor of Philosophy in Biostatistics
and the Designated Emphasis
in
Computational and Genomic Biology
University of California, Berkeley

Professor Mark J. van der Laan, Chair

In many scientific studies the goal is to determine the effect of a particular feature or variable on a given outcome in order to help understand, identify, and quantify the driving factors behind a particular phenomena. This type of analysis is commonly referred to as variable importance analysis. Parametric methods used to estimate these effects are prone to bias. This bias is often the result of incorrect model specification and improper inference for the parameter of interest. Alternative machine learning techniques, such as Random Forest, often result in abstract measures of importance whose inference depends on a computationally intensive bootstrap analysis. In this thesis, robust estimators for variable importance based on targeted maximum likelihood methodology are presented and developed for three types of outcomes (1) univariate continuous, (2) multivariate continuous, and (3) binary outcome. These estimators are specifically designed to target the effect of a variable of interest on an outcome while adjusting for confounders when the variable of interest is of general form (i.e. continuous or discrete). When the outcome is continuous (1,2), the effect is on an additive scale. When the outcome is binary (3), the effect is on a multiplicative scale, and the importance measure is a relative risk. The estimators are developed under a flexible semiparametric model, in which only components related to the variable of interest must be fully specified, and effect modification can be easily incorporated. Based on targeted maximum likelihood theory, the presented estimators are double robust and locally efficient, and correct inference for the parameter of interest is available using the corresponding influence curve.

In this thesis, the three estimators relating to the three outcomes are derived from targeted maximum likelihood methodology and implemented by adapting standard
statistical regression software. These estimators are applied in both simulation and application. In a simulated biomarker discovery analysis, the robustness of the estimator for a univariate continuous outcome is compared to other common methods of variable importance under increasing correlation among the covariates. In a repeated measures setting, the double robust property of the estimator for a multivariate continuous outcome is demonstrated in simulation, and the estimator is applied in a transcription factor analysis to determine the activity level of transcription factors during the cell cycle in yeast. For a binary outcome, the estimator for the relative risk is applied to estimate the effect of HIV genetic susceptibility scores on viral response. Effect modification is also explored and model selection methodology is introduced.
Rosencrantz: Shouldn’t we be doing something… constructive?
Guildenstern: What did you have in mind?
   A short, blunt human pyramid?
   -Stoppard
# Contents

1 Introduction 1

1.1 Data 4

1.2 Measures of interest 4

1.3 Semiparametric model 5

1.4 Introduction to estimation methods 6

1.4.1 Parametric methods 7

1.4.2 Semiparametric methods 8

1.4.3 Introduction to targeted maximum likelihood 8

1.5 Notation 11

2 TMLE for biomarker discovery - The search for a standard 12

2.1 Introduction 12

2.2 Variable importance 13

2.3 Targeted maximum likelihood estimation 14

2.3.1 Inference and testing 16

2.3.2 TMLE implementation steps 17

2.4 Comparison of variable importance methods 20

2.4.1 Simulated data 21

2.4.2 Methods 22

2.4.3 Results 23

2.4.4 Discussion 30

2.5 Experimental treatment assumptions and consequences of its violation 30

2.6 Application 32

2.6.1 Implementation considerations 32

2.7 Golub et al (1999) 34

2.7.1 Data 34

2.7.2 Analysis 35

2.7.3 Results 35

2.7.4 Discussion 42

2.8 van’t Veer et al (2002) 43

2.8.1 Data 44

2.8.2 Analysis 44
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8.3</td>
<td>Results</td>
<td>44</td>
</tr>
<tr>
<td>2.8.4</td>
<td>Discussion</td>
<td>46</td>
</tr>
<tr>
<td>2.9</td>
<td>Conclusions</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>TMLE for repeated measures</td>
<td>49</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>Repeated measures design</td>
<td>50</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Data</td>
<td>50</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Generalized estimating equations</td>
<td>50</td>
</tr>
<tr>
<td>3.3</td>
<td>Semiparametric variable importance</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>Targeted maximum likelihood estimation</td>
<td>52</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Derivation of TMLE update</td>
<td>53</td>
</tr>
<tr>
<td>3.4.2</td>
<td>TMLE implementation steps</td>
<td>55</td>
</tr>
<tr>
<td>3.5</td>
<td>Additional methods to improve TMLE</td>
<td>58</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Sequential update</td>
<td>58</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Simulation</td>
<td>59</td>
</tr>
<tr>
<td>3.6</td>
<td>Simulation study</td>
<td>59</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Data</td>
<td>60</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Results</td>
<td>61</td>
</tr>
<tr>
<td>3.7</td>
<td>Application - Transcription factor activity</td>
<td>61</td>
</tr>
<tr>
<td>3.8</td>
<td>Model selection for $m(A,V</td>
<td>\beta)$</td>
</tr>
<tr>
<td>3.8.1</td>
<td>Model selection procedure</td>
<td>67</td>
</tr>
<tr>
<td>3.8.2</td>
<td>Simulation</td>
<td>68</td>
</tr>
<tr>
<td>3.8.3</td>
<td>Brief application</td>
<td>70</td>
</tr>
<tr>
<td>3.9</td>
<td>Conclusions</td>
<td>71</td>
</tr>
<tr>
<td>4</td>
<td>TMLE for multiplicative model and relative risk</td>
<td>73</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>73</td>
</tr>
<tr>
<td>4.2</td>
<td>Data</td>
<td>75</td>
</tr>
<tr>
<td>4.3</td>
<td>Multiplicative semiparametric model</td>
<td>75</td>
</tr>
<tr>
<td>4.4</td>
<td>Measure of interest</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Targeted maximum likelihood estimation</td>
<td>76</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Derivation of the TMLE</td>
<td>76</td>
</tr>
<tr>
<td>4.5.2</td>
<td>TMLE implementation steps</td>
<td>80</td>
</tr>
<tr>
<td>4.6</td>
<td>Application</td>
<td>82</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Background</td>
<td>82</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Data</td>
<td>83</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Analysis</td>
<td>83</td>
</tr>
<tr>
<td>4.6.4</td>
<td>Results</td>
<td>85</td>
</tr>
<tr>
<td>4.6.5</td>
<td>Secondary analysis</td>
<td>87</td>
</tr>
<tr>
<td>4.7</td>
<td>Discussion</td>
<td>91</td>
</tr>
</tbody>
</table>
# 5 Conclusion


## Bibliography


## Appendices

### A Additional results for TMLE - univariate continuous case

- **A.1 Double robust property**
  - A.1.1 Updating the "treatment mechanism"
  - A.1.2 Simulation study

- **A.2 A Comparison to nonparametric inverse weighing approach: accessing ETA violation performance**
  - A.2.1 DR-IPTW estimator for V-adjusted variable importance
  - A.2.2 Simulation study

### B Biomarker discovery simulation: additional results

- **B.1 Average importance value**
- **B.2 Average importance rank**

### C The efficient influence curve

- **C.1 Repeated measures TMLE - efficient influence curve derivation outlined**
- **C.2 Influence curve for effect parameter of multiplicative semiparametric model**

### D Variance reduction of the influence curve by weighting

- **D.1 Simulation study**
- **D.2 Results**

### E Sample code

- **E.1 Simple R code example for repeated measures TMLE**
  - E.1.1 Simple simulated data
  - E.1.2 TMLE-RM method
Acknowledgments

There have been numerous people throughout my academic career that have contributed to my success both as a researcher and an individual. As I look back I realize all the people who have contributed to my development at various points of my life, from graduate school, to undergrad, and even back into high school. There have been so many people who have made me who I am and made me capable of achieving this level of academic success. Their support and guidance have been invaluable. I acknowledge them here to honor their contribution to my academic development and its culmination in my Ph.D. dissertation. Please forgive me if any are missing. It would be difficult to mention them all.

I would like to start by acknowledging my Ph.D. advisor, Mark J. van der Laan. Throughout the years, Mark has been a constant guide and supporter of my academic career. He has been continually encouraging and positive and has helped me acquire the skills and resources necessary to succeed. Needless to say, Mark has been instrumental in my development as a researcher and statistician. I would like to thank Mark for all the years of explanations, all the answers to my questions, all the enthusiasm, and for your continual academic support. Without him this work would not have been possible.

I’d like to thank my dissertation committee members: Alan Hubbard and Adam Arkin. Alan was instrumental in my initial recruitment into the Biostatistics Ph.D. Program. He has been a constant mentor to me throughout the years. He is always open to new project directions and was especially supportive of my need to expand my statistical knowledge base beyond my thesis work. Without him I would not have the range of knowledge and experience that makes me the applied statistician I am today.

I would like to thank Adam Arkin for his inspiration and constant interest in working with me throughout the years. He has encouraged me to continually look beyond the math to understand fully the biological implications and meaning behind the algorithms and behind the numbers. He has helped me expand my horizons and allowed me to creatively use statistics in solve problems.

There have been numerous other professors at UC Berkeley who have contributed to my success. I would like to thank Sandrine Dudoit for her support in the early years of my Ph.D. as well as her support as a member of my qualification exams committee. I would like to thank David Brillinger, who has been a great mentor and friend throughout the years. He was instrumental in my first years at UC Berkeley, and has always been willing to lend a hand and a friendly ear throughout my years at the university. I would like to thank Nick Jewel for his amazing teaching and advising, David Freedman for a fantastic consulting class - his teachings will remain with me throughout my career, and Cari Kaufman for our extremely helpful conversations the last couple of years. I would also like to give special thanks to Deb Nolan and Bin Yu for fostering such a wonderful community of women in biostatistics and statistics.
I would also like to thank my numerous collaborators at UC Berkeley. Each has contributed to my development as a statistician and made me into the researcher I am today. In particular I would like to thank Marcin Joachimiak, who has been a collaborator of mine for many years. I would like to thank him for his continual patience and cooperation in the development of our project. It has been a long road but the end is in sight.

In addition I would like to thank and acknowledge my masters advisor from Columbia University, Upmanu Lall. Upmanu was a fantastic supporter and advocate of my academic development, and I would not be where I am today without him. I would also like to thank Andrew Gelman who first introduced me to the world of statistics and its possibilities. It was during his class that I realized I wanted to be a statistician.

Finally, I would like to thank the staff in the biostatistics and statistics department at UC Berkeley, as well as at Columbia University in Earth and Environmental Engineering, for their support and assistance. Everyone has been extremely helpful, and things would not have progressed as smoothly if it were not for you.

Looking back even further, I realize that there were so many other teachers and professors who contributed to my development and made this Ph.D. possible. In particular my undergraduate advisor Chung-Chiun Liu. I’d like to thank all my other professors and mentors at Case Western. I would also like to give a special thanks to the IMSA community, without which I would not have developed into the person and researcher I am today. I would like to especially thank all my professors and instructors at IMSA, in particular, Dr. Pine and Dr. Skinner.

I would also like to acknowledge the financial support of the National Science Foundation Graduate Research Fellowship, the NIH Genomics Training Grant, and the NIH targeted empirical super learning grant, R01 AI074345.

Finally I would like to thank my friends. Their friendship and support has been incredibly important to me. In particular I would like to thank Rana Samuels for all the Sunday afternoons of studying, and my friends in Applied Physics and Mathematics at Columbia U for all their support. I would like to thank all my friends in the Statistics department at UC Berkeley; especially my fellow women statisticians who have helped and supported me throughout these past years. I would like to thank Beja Tinsley, who has provided me with a sense of perspective and balance. I would also like to acknowledge Cal Cycling and the wonderful friends there, who provided a much needed distraction and kept me balanced the last few years of my Ph.D. Finally I’d like to thank my family for their support in my growth and development, and I would like to especially thank my friend and partner Chris Roper for his continuous and unfaltering love and support throughout the years.
Chapter 1

Introduction

Many statistical analyses focus on estimating the effect of a particular variable on an observed outcome. They may be interested in the magnitude of this effect, whether or not this effect is greater than the effect of another variable, and whether this effect is statistically significant. This type of analysis is prevalent in a variety of scientific fields of study where the researcher is interested in understanding the driving variables behind the phenomena. Ideally, the researcher would like to determine the causal mechanism behind a particular phenomena. Unfortunately, causal effects are extremely difficult to measure accurately outside a randomized trail. However, methods which formally target estimation towards the causal effect can still provide useful effect measures as measures of variable importance. This thesis introduces three estimators of variable importance developed under targeted maximum likelihood methodology for three types of outcomes: (1) univariate continuous (2) multivariate continuous and (3) univariate binary. The estimators presented here are developed under a flexible semiparametric model for a general variable of interest (i.e. continuous or binary) and are interpretable as the causal effect under appropriate assumptions. For the purposes of this thesis, the targeted maximum likelihood estimator will be referred to as the TMLE.

More often than many would care to admit, the reported findings of scientific studies are often heavily biased. Articles that address the problem of bias in reported findings have recently appeared in the literature [45, 93]. In particular, John P.A. Ioannidis published a report entitled “Why Most Published Research Findings are False” that addresses the large uncertainty in reported findings and outlines the type of studies most prone to bias, noting that “Simulations show that for most study designs and settings it is more likely for a research claim to be false than true” [45]. This bias arises from multiple sources: sampling bias, biased estimators, and bias in the interpretation of the results. Biased estimators are often the result of a reliance on incorrect models or procedures that are not targeted towards the parameter of interest. False conclusions resulting from this bias can be very expensive, leading to failed clinical trails, improper policy implementation, and incorrect treatment assign-
ment for example. This is particularly true in genomic studies where the number of variables (and tests) is high and very correlated.

The estimators developed in this thesis attempt to prevent this bias through the use of more flexible semiparametric models and a more honest estimation procedure that is targeted towards the parameter of interest. Specifically, targeted maximum likelihood methodology reduces the bias for the targeted parameter by maximizing the likelihood in a direction which corresponds to the best estimate of the targeted parameter [107]. These estimators are robust to model misspecification, and the TMLE is equivalent to the solution of the efficient estimating equation [107]. Therefore, inference targeted towards the parameter of interest can be obtained using the corresponding efficient score equation. The resulting estimates are simple to interpret and can often be obtained using standard statistical software. The estimation procedure also incorporates the use of data-adaptive prediction algorithms, such as super learner [104], which improves the overall performance.

Previous publications on targeted maximum likelihood based variable importance have focused primarily on binary or categorical variables of interest [7, 107, 105]; for instance an indicator for receiving a particular treatment or exposure. However, especially in the world of genomics and epidemiology the variable of interest is often continuous. This thesis focuses on the development of targeted maximum likelihood based estimators under a semiparametric model, which can accommodate both binary and continuous variables of interest as well as incorporate effect modification in a straightforward and interpretable manner. Under this semiparametric model, only model components relating to the variable of interest must be specified, while the remainder of the model is not required to be of any particular form. Variable importance under this model, was first presented in [102], and estimated using TMLE in [107]. Models of this type have also been considered in [79, 81, 120, 13]. Though the consistency of these estimators are less robust to model misspecification than their fully nonparametric counterparts for binary A, they are more robust to violations of the positivity assumption (i.e. sparse support for the variable of interest over all covariates, or high correlation among the covariates; see section 2.5).

Three different estimators for variable importance will be presented in this thesis. Each is developed using targeted maximum likelihood methodology. These three estimators are derived for the following outcomes (1) univariate continuous, (2) multivariate continuous, and (3) binary. A general introduction will be presented first, including an introduction to the observed data structure, an overview of the general semiparametric model form, and the measures of interest defined in terms of the observed data. Then, some of the more common estimation procedures used to estimate these measures will be introduced and a brief overview of their features will be included. A general introduction to targeted maximum likelihood methodology will follow. After this introduction, each estimator will be presented individually in its own chapter. Each chapter will again introduce the measure of interest, describe the estimation theory and procedure, and then show the estimator in action in a
simulation study and/or application.

Specifically in chapter 2, the targeted maximum likelihood variable importance estimator for a general variable of interest given a univariate continuous outcome will be presented. This chapter is motivated specifically by the need for a standardized approach in biomarker discovery analyses in which the goal is to identify specific biomarkers of interest from a long list of possible biomarkers. Biomarker data is high-dimensional and highly correlated making more conventional methods like linear regression, highly susceptible to bias. In this chapter, the derivation of the TMLE is outlined and step-by-step instructions for implementation are provided. A simulation study is presented which compares the performance of common estimators for biomarker importance (i.e linear regression, Random Forest [14], and lasso regression [94]), and the targeted maximum likelihood estimator (TMLE) [107], under increasing levels of correlation among the variables. The methods are compared based on their abilities to distinguish “true” biomarkers from the “decoy” biomarkers. Then, two applications of the TMLE are presented, a leukemia study using the Golub et al. 1999 data [40] and a breast cancer study using the van’t Veer et al. 2001 data [109].

In chapter 3, the targeted maximum likelihood variable importance estimator for a general variable of interest given a multivariate continuous outcome is presented. This estimator is developed for longitudinal or repeated measures studies where the outcome is observed and recorded over time or condition. For this estimator, a more thorough account of targeted maximum likelihood methodology is presented along with implementation instructions using standard statistical software. The robust properties of the TMLE are demonstrated through a small simulation study where the TMLE is compared to the more commonly used generalized estimating equation (GEE) approach. The TMLE is then applied to a transcription factor data analysis. In this analysis, the goal is to identify the relative activity levels of transcription factors at various time points during the cell cycle in yeast (S. cerevisiae). Specifically, the following transcription factors are studied: SWI4, SWI6, MBP1, MCM1, ACE2, FKH2, NDD1, and SWI5. Additionally, a model selection procedure is presented, which adds another level of flexibility to the semiparametric model form and is a useful tool that can be applied to all estimators presented in this thesis.

In chapter four, the targeted maximum likelihood variable importance estimator for general variable of interest given a binary outcome is presented. The TMLE is specifically developed to target estimation towards the relative risk. The relative risk is an intuitive measure of variable importance used in a variety of studies from medicine and epidemiology to policy. In this chapter, the derivation of the TMLE is presented given two forms of the initial density, log-binomial and Poisson. Implementation instructions for the Poisson TMLE are also included. The TMLE is then applied to a study of HIV genetic susceptibility scores. In this analysis, the goal is to determine the effects of different genetic susceptibility scores on viral response. Effect modification by other covariates in the model is also explored. The thesis will conclude with an overall discussion. The work presented in this thesis was done in
collaboration with Mark van der Laan.

1.1 Data

The estimators in this thesis are developed given the following observed data structure

\[ O = (W^*, Y) \sim P_0 \in \mathcal{M} \]

where \( W^* \) is the set of variables (i.e. genes), and \( Y \) is the outcome of interest. The data consists of \( n \) independent and identically distributed observations, sampled from \( P_0 \), the true data generating distribution of \( O \), which is an element of the statistical model \( \mathcal{M} \) presented below in 1.3.

The variable importance measures are defined with respect to a particular \( A = W^*_j \) on outcome \( Y \) controlling for potential confounders among covariate set \( W = W_{-j} \). The three different outcomes presented in this thesis are defined below:

1. Univariate continuous outcome, where \( Y \) is a continuous outcome variable
2. Multivariate continuous outcome, where \( Y \) is the outcome vector of \( T \) repeated continuous measures taken over time on a subject, and \( Y_t \) represents outcome \( Y \) at a specific time point \( t \) for a subject
3. Binary outcome, where \( Y \) is \( \{0, 1\} \)

1.2 Measures of interest

Variable importance analysis is specifically interested in answering the question: What is the effect of variable \( A \) on \( Y \) controlling for other variables \( W \)? The question of interest is defined in terms of the observed data for each of the three outcomes as follows:

1. For a univariate continuous outcome, additive effect is defined as
   \[ \Phi(P_0) = E_0[Y|A = a, W] - E_0[Y|A = 0, W] \]
2. For a multivariate continuous outcome, additive effect is defined as
   \[ \Phi(P_0) = E_0[Y(t)|A = a, W] - E_0[Y(t)|A = 0, W] \]
3. For a binary outcome, multiplicative effect or relative risk is defined as
   \[ \Phi(P_0) = \frac{P_0(Y = 1|A = a, W)}{P_0(Y = 1|A = 0, W)} \]
1.3 Semiparametric model

Targeted maximum likelihood methodology is specifically developed to target a parameter of interest. Given that the variable of interest is continuous, the measures of interest presented above must be parameterized further in order to obtain an estimable and still interpretable measure of the effect of variable $A$ on outcome $Y$ controlling for $W$. The TMLE is therefore developed under the following semiparametric model, $\mathcal{M}$, which parametrizes the measures of interest in terms of the parameter vector $\beta$. The TMLE is then developed to target $\beta$, the parameter of interest, under this model. Variable importance under this model was presented in [102], and estimated using targeted maximum likelihood estimation in [107]. Models of this type have also been considered in [79, 81, 120, 13].

The likelihood of the observed data is factorized as follows

$$P_0(O) = P_0(Y \mid A, W) P_0(A \mid W) P_0(W)$$

No assumptions are made about the distributions of $P_0(A \mid W)$ and $P_0(W)$. The following semiparametric model, $\mathcal{M}$, is assumed for the mean of $Q_0(A, W) = P_0(Y \mid A, W)$,

$$Q_0(A, W) = E_0[Y \mid A, W],$$

as follows

$$b(E_0[Y \mid A, W]) = m(A, V \mid \beta_0) + b_0(W)$$

where the function $b(.)$ is the link function. In this thesis, two link functions are considered: (1) the natural link, resulting in the additive model and (2) the log-link resulting in the multiplicative model. These are defined explicitly below.

In this formulation, $m(A, V \mid \beta_0)$ is the parametric component of the model, defined such that $m(A = 0, V \mid \beta_0) = 0$ for all $(a, v) \in \{A, V\}$, where $V$ is a single or set of effect modifiers of the importance of $A$ on $Y$ such that $V \subset W$. The component $\theta_0(W) = E_0[Y \mid A = 0, W]$ is unspecified and can be estimated nonparametrically. In other words, under this semiparametric model, only components relating to the variable of interest, $A$, must be specified explicitly. The remainder of the model is not required to be of any particular form.

(1) **Additive Model** is used when the outcome, $Y$, is continuous and is the result of a natural link function: $b(x) = x$.

$$E_0[Y \mid A, W] = m(A, V \mid \beta_0) + \theta_0(W)$$

This model form also holds for a multivariate continuous outcome, $Y_t$, where $t$ is an index for each repeated measure

$$E_0[Y_t \mid A, W] = m_t(A, V \mid \beta_0) + \theta_{0,t}(W).$$

The additive model can be rearranged in terms of the measure of interest, the additive effect, as follows

$$E_0[Y \mid A, W] - E_0[Y \mid A = 0, W] = m(A, V \mid \beta_0)$$
**Multiplicative Model** is used when the outcome, $Y$, is binary, and is the result of a log-link function: $b(x) = \log(x)$.

$$\log \left( E_0[Y \mid A, W] \right) = m(A, V \mid \beta_0) + \log(\theta_0(W))$$

Note that when $Y$ is binary, $E_0[Y \mid A, W] = P_0(Y = 1 \mid A, W)$, and the above additive form can be rearranged in terms of the measure of interest, the relative risk, as follows

$$\log \left( \frac{P_0(Y = 1 \mid A, W)}{P_0(Y = 1 \mid A = 0, W)} \right) = m(A, V \mid \beta_0)$$

A major benefit of this model form is that regardless of whether or not the model, $m(A, V \mid \beta_0)$, is correctly specified, the model is still valid under the null hypothesis $H_0 : b(E[Y \mid A, W]) - b(E[Y \mid A = 0, W]) = 0$ or equivalently $H_0 : \beta_0 = 0$. Therefore, hypothesis testing is still valid.

For the purposes of this work, a linear model is assumed for $m(A, V \mid \beta_0)$, which puts the model in the class of generalized partial or partially linear models. We present here examples of possible model formulations.

**Main effect model** is the simplest model formation and is commonly used in biomarker analysis

$$m(A, V \mid \beta) = A\beta$$

**Main effect model with effect modification** is used when one expects the effect of $A$ to depend on another covariate, $V$, where $V \subset W$. This is especially relevant in clinical trials, and the researcher is interested in finding genes which modify the causal effect of a given a particular treatment, $A$, on overall disease response.

$$m(A, V \mid \beta) = A\beta_1 + AV\beta_2 = \beta'[A \mid A: V]$$

**Transformations: Quadratic etc.** can be incorporated. Any form of $A$ or $V$ can be substituted into the model (e.g. $\log(A)$, $A^2$, etc.). Given an additive model (i.e. linear in $\beta$), TMLE is still easily implemented.

### 1.4 Introduction to estimation methods

As shown above, the measures of interest are parameterized by parameter of interest, $\beta$. Ideally estimation should be targeted towards this parameter such that the
bias-variance trade-off is specifically focused on the estimation of $\beta$ and not the entire model form (i.e. $\beta$ and $\theta(W)$). In general this is not the case. A brief discussion of some of the more common methods estimation methods is presented below. These are presented in generalities and are more specifically addressed in the following chapters for each estimator.

There are also alternative measures of variable importance, such as Random Forest, which are not contained within this family of models. These measures can be difficult to interpret and often do not have any formal inference available. A discussion of the importance measures from Random Forest is included in chapter 2.

1.4.1 Parametric methods

The most common parametric method is generalized linear regression [43]. Generalized linear regression is a maximum likelihood based estimator that uses least squares regression methods for estimation (e.g. ordinary least squares, iteratively re-weighted least squares, etc.). In generalized linear regression, $m(A,V|\beta)$ and $\theta(W)$ are assumed to be linear models, and the consistency of the estimator for $\beta$ is dependent on correctly specifying both $m(A,V|\beta)$ and $\theta(W)$. Any misspecification can lead to bias in the estimate of $\beta$. Therefore to obtain an unbiased estimate of $\beta$, all relevant confounders must be included in the model correctly. Standard errors estimated using standard GLR are also highly sensitive to correlation among the covariates, which can result in incorrect inference for $\beta$.

Particularly in genomic studies, where the number of covariates often is much greater than the number of observations, it becomes impossible to use generalized linear regression to control for all confounders. Often researchers will either use simple linear regression where the model includes only the covariate of interest or employ a model selection algorithm. Simple linear regression does not control for the necessary covariates and does not accurately measure the effect of $A$. It is also highly sensitive to correlation among the covariates (see chapter 2). Model selection methods may discount the variable of interest entirely based on the overall predictive ability of the model. This results in an effect measure of zero with no inference, which may not be the case in reality.

Penalized regression methods such as lasso [94] suffer from similar drawbacks. These methods perform a type of shrinkage among the coefficients, which allows the model to incorporate the effect of more covariates than standard regression. However, the model is still a prediction algorithm based on a parametric model, and the consistency of the estimator is still dependent on correctly specifying the full model. The amount of shrinkage is often selected by cross-validation which is based on prediction accuracy. These methods are also constrained and may only include up to $n$ covariates.
1.4.2 Semiparametric methods

There are two basic types of semiparametric models that provide estimates for the parameter $\beta$ as defined above: (1) generalized partial linear model, which is the model defined above in section 1.3 and (2) generalized additive model, which is a less flexible model in which each covariate has a separate additive component of general form (e.g. spline etc.) in the model [43]. Estimation methods include backfitting [43, 16], profile likelihood methods [90, 85, 86], and marginal integration methods [20]. These methods are more flexible than the above parametric models, however estimation still focuses on prediction and the fit of the overall model, not the parameter of interest. Estimation is therefore still prone to bias and any available inference is not correct for the parameter of interest.

The estimating equation approach based on the partial linear model has also been proposed as an estimation method for $\beta$ [120, 102, 13]. The resulting estimator has the double robust property [102] and estimation is targeted towards the parameter of interest. However it is not a substitution estimator and therefore does not respect the global information of the statistical model. It is also can be difficult to implement and can result in multiple solutions with no criteria to accurately select the correct one. Additionally, there is no natural way to integrate targeted selection of estimators for the nuisance parameters. In contrast targeted maximum likelihood methodology results in a targeted likelihood, which provides a natural framework for integrating selection of nuisance parameter estimates as well as subsequent model selection on $m(A, V/\beta)$ [101].

1.4.3 Introduction to targeted maximum likelihood

Maximum likelihood driven methods such as linear regression and lasso regression focus on estimating the entire regression, $E[Y|A, W]$. Therefore, the “plus and chug” maximum likelihood estimate for $\beta$ is based on the bias-variance trade-off for estimating $E[Y|A, W]$, not the parameter of interest, $\beta$. Targeted maximum likelihood methodology reduces the bias for the targeted parameter by maximizes the likelihood in a direction which corresponds to the best estimate of the targeted parameter by using the correct bias-variance trade-off for the parameter of interest [107].

Targeted maximum likelihood estimation uses the general MLE framework and combines it with robust estimation using the efficient influence curve to provide a double robust and locally efficient estimator for the parameter of interest. The resulting estimator is consistent and asymptotically linear if either the mean of the variable of interest as a function of the confounders is correctly modeled (i.e. confounding/treatment mechanism), or if the mean of the outcome as a function of the variables (including variable of interest) is correctly modeled. The targeted maximum likelihood method integrates data-adaptive prediction algorithms such as DSA [88] and super learner [104] to improve overall performance by using these methods to
obtain the initial estimator and the confounding/treatment mechanism used in the targeted update.

The targeted maximum likelihood method is particularly suitable for variable importance analysis. In this thesis, the TMLE is developed under the semiparametric model presented in section 1.3. By using the semiparametric model, the TMLE is more flexible than its parametric counterparts. The semiparametric construction not only provides a flexible model, but nicely handles the effect of continuous variables and also allows the incorporation of effect modification of the variable of interest in a straightforward and interpretable manner. This allows the estimation of not only the variable importance averaged over covariates \( W \), but the importance at a particular level of a covariate (e.g. effect modified by time). Also, the estimation procedure under the semiparametric model does not require inverse weighing by the probability of treatment \( P(A = a|W) \), which is required for nonparametric TMLE based variable importance estimation and can be problematic when the probability of treatment approaches one or zero [7, 8].

The targeted maximum likelihood method

Targeted maximum likelihood methods can be implemented using standard statistical software. Outlined below is a general description of targeted maximum likelihood estimation and inference. This description will be revisited in each chapter, where the estimation procedure will focus on each individual estimator.

Targeted maximum likelihood estimation uses standard maximum likelihood estimation to update the initial estimate in a direction with targets the parameter of interest. Formally, targeted maximum likelihood methodology creates a path through a true density \( p_0 \), represented as the hardest sub-model \( p_0^0(\epsilon) \). The hardest submodel \( p_0^0(\epsilon) \) is selected to only vary \( Q(p)(A,W) = P(Y|A,W) \), and to have a score equal to \( D_{h,Q,g}(p_0^0) \) at \( \epsilon = 0 \), where \( D_{h,Q,g}(p_0^0) \) is the efficient score equation relating to the parameter of interest indexed by choice of \( h \), and densities \( Q \) and \( g \). Given an initial estimate of the density \( p_0^0 = Q^0(A,W) \), and defining the hardest sub-model \( p_0^0(\epsilon|p_0^0) \), \( p_0^0(\epsilon|p_0^0) \) is maximized with respect to \( \epsilon \). Then substituting in the new estimate \( \epsilon_n \), the updated density \( p_1 = p_0^0(\epsilon_n|p_0^0) \), is the new targeted density. In some cases iteration is necessary (substituting the new density estimate as initial density estimate and solving again for \( \epsilon \)). By maximizing \( p_0^0(\epsilon|p_0^0) \) for \( \epsilon \), targeted maximum likelihood estimation maximizes the likelihood in the direction of the parameter of interest, making the final density estimate also the solution to the corresponding efficient estimating equation \( E_n[D_{h,Q,g}(O)] = 0 \) as well.

For this work, the initial density has a corresponding distribution in the semiparametric model presented in section 1.3. To target the measure of interest, the mean of this initial density, \( Q^0(A,W) = E[Y|A,W] \), will be fluctuated to form the class of submodels. The targeted maximum likelihood update is achieved by regressing the outcome on a “clever covariate” while setting the initial estimate, \( Q^0(A,W) \),
as an offset. The “clever covariate” is a function of the “treatment mechanism”, $g(W) = P(A|W)$. This update is iterated until convergence.

**Inference for the TMLE**

As stated previously, the clever covariate is derived such that the converged targeted maximum likelihood estimator is also the solution to the estimating equation corresponding to the efficient influence curve for the parameter of interest\footnote{107}. In other words, the converged TMLE for $\beta_0$ is also the solution to the efficient estimating equation

$$E_n [D_{h,Q,g}(O|\beta_0)] = 0$$

where $D_{h}(O|\beta_0)$ is the efficient score equation. Therefore, formal inference for the TMLE can be estimated using the corresponding efficient influence curve. The following scaled version of the efficient influence curve is used to provide an estimate for the covariance of the TMLE. This is defined for a single subject, $i$, as

$$IC(O) = c^{-1}D_{h,Q,g}(O|\beta_0)$$

given scale factor

$$c = -E \left[ \frac{d}{d\beta} D_{h,Q,g}(O|\beta_0) \right]$$

The efficient influence curve, $IC(O)$ is a 1 by $p$ vector for univariate $Y$ and a $T$ by $p$ matrix for multivariate $Y$, where $\beta_0$ is a vector of length $p$.

Given correctly specified estimates for $Q(A,W)$ and $g(W)$, the covariance for parameter vector estimate $\beta_n$ is asymptotically equivalent to the covariance of $IC(O)$. If $Q(A,W)$ is misspecified, but $g(W)$ is correctly estimated, the above influence curve is known to be conservative\footnote{102}. The empirical estimate of the covariance of $\beta_n$ is

$$\Sigma_n = \frac{1}{n} \sum_i \hat{I}C(O^{(i)})\hat{I}C(O^{(i)})'$$

and the normal approximation

$$\sqrt{n}(\beta_n - \beta_0) \sim N(0, \Sigma_n)$$

can be used for the purpose of statistical inference. This is analogous to the robust sandwich estimator for variance\footnote{60}.  

1.5 Notation

The notation used in this thesis is presented below as a reference.

\(N\) outcome variable
\(A\) variable of interest
\(W\) covariate set, not including the variable of interest
\(W^*\) full covariate set, including the variable of interest
\(W_s\) covariate set, restricted using some initial pre-screening
\(V\) effect modifiers of \(A\), a subset of \(W\)
\(P_0\) True observed data generating distribution from which the observations are i.i.d. samples
\(M\) Class of semiparametric models under which the TMLEs are developed
\(Q\) \(P(Y|A,W)\), probability of \(Y\) given \(A\) and \(W\)
\(\bar{Q}\) \(E[Y|A,W]\), expected value of \(Y\) given \(A\) and \(W\), the mean of \(P(Y|A,W)\)
\(\bar{Q}_0\) i.e \(Q_0\), denotes value under the true data generating distribution \(P_0\)
\(\bar{Q}_n\) i.e \(Q_n\), denotes estimated value given observed data
\(\bar{Q}^*\) i.e \(Q^0\), denotes the initial value/estimate
\(\bar{g}\) \(P(A|W)\), probability of \(A\) given \(W\)
\(\bar{g}_n\) \(E[A|W]\), expectation of \(A\) given \(W\), the mean of \(P(A|W)\)
\(\beta\) Parameter of interest that is targeted
\(m(A,V|\beta)\) Parametric portion of the semiparametric model, used to model the parameter of interest, contains all components of the model containing \(A\)
\(r(W)\) Index used in the derivation of the TMLE to fluctuate the \(\theta(W)\) process
\(\theta(W)\) Estimate of \(E[Y|A=0,W]\) which is equivalent to \(P(Y=1|A=0,W)\) when \(Y\) is binary
\(H^*(A,W)\) clever covariate derived such that TMLE is equivalent to solution of the efficient estimating function
\(D_{h,Q,g}\) Score equation related the the estimating function, indexed by \(Q\) and \(g\) and choice of \(h\) where \(h = h_{opt}\) indicates the efficient score equation
\(IC_{h,Q,g}\) Influence curve indexed by \(Q\) and \(g\) and choice of \(h\) where \(h = h_{opt}\) indicates the efficient score equation
\(\Phi(P)\) Measure of interest according the the data generating distribution \(P\)
\(\Psi(P)\) General notation for the parameter of interest according the the data generating distribution \(P\), this is equivalent to \(\beta\) for these estimators
\(E_0\) Expectation taken with respect to \(P_0\), the true data generating distribution
\(E_n\) Expectation taken with respect to \(P_n\), the empirical distribution of the sample data
\(I\) Indicator (i.e. \(I\{A = 0\}\))
\(A: B\) Interaction term of \(A\) and \(B\), mathematically \(A \times B\)
\([A A: B]\) matrix with columns(s) \(A\) appended onto matrix of columns(s) \(A: B\)
Chapter 2

TMLE for biomarker discovery -
The search for a standard

2.1 Introduction

The use of biomarkers in disease diagnosis and treatment has grown rapidly in recent years, as microarray and sequencing technologies capable of detecting biological signatures have become more effective and efficient research tools. In an attempt to create a level of quality assurance with respect to biological and more specifically biomarker research, the FDA has called for the development of a standard protocol for biomarker qualification [100]. Such a protocol would define “evidentiary” standards for biomarker usage in areas of drug development and disease treatment and provide a standardized assessment of a biomarker’s significance and biological interpretation. This is especially relevant for clinical trials, where the protocol would prohibit the use of unauthenticated biomarkers to determine treatment regime; resulting in safer and more reliable treatment decisions [99]. Consequently, identifying accurate and flexible analysis tools to assess biomarker importance is essential. This chapter presents a measure of variable importance estimated with targeted maximum likelihood methodology based on a flexible semiparametric model as a standardized measure for biomarker importance.

Many biomarker discovery methods only measure the association between the marker and the biological outcome. However a significant association is often difficult to interpret and does not guarantee that the biomarker will be a suitable and reliable drug candidate or diagnostic surrogate. This is especially true with genomic data, where genes are often present in multiple pathways and can be highly correlated amongst themselves. Applying association-based methods to this data will often lead to a long and ambiguous listing of biomarkers, which can be expensive to analyze.

Ideally, biomarker discovery analyses want to identify markers that systematically effect the outcome through a biological pathway or mechanism; in other words markers
causally related to the outcome of interest. Once these markers are identified, they can be further analyzed and eventually applied as potential drug targets or prognostic markers. Due to the complex nature of the human genome, this is not a straightforward task and certain assumptions are required to identify a causal effect.

In general, causal effects are often difficult if not impossible to estimate correctly, especially in high-dimensional and highly correlated genomic data. The specific assumptions they require (randomized treatment, experimental treatment assignment, etc.) are often only fully realized in randomized trials, making their utility in a standard protocol limited. However, measures which are causally interpretable in randomized trials, can still be biologically interpretable in observation data as measures of importance.

Here, the typical representation of a marginal causal effect is presented as a potential measure of biomarker importance

$$\mu_0(a) = E_0[E_0[Y|A = a, W] - E_0[Y|A = 0, W]]$$

Given the observed data $O = (A, W, Y) \sim P_0$, this measure corresponds to the marginal effect of a biomarker ($A$) on the outcome ($Y$), adjusting for confounders ($W$). Here, $A$ can represent a single biomarker or set of biomarkers. This chapter will focus on the univariate case.

This chapter presents the TMLE for the variable importance measure above under a semiparametric model, which can accommodate continuous treatment or exposure variables often seen in biomarker analyses. This TMLE is then introduced as a candidate standardized measure of biomarker importance. Its efficacy and functionality are demonstrated in simulation. Simulations provide a performance assessment of our estimated measure under increasing levels of correlation. The simulations demonstrate the accuracy with which TMLE can detect “true” variables from amongst increasingly correlated “decoy” variables. Additionally, the accuracy of three commonly used methods for biomarker discovery are also evaluated under the same conditions, univariate linear regression, lasso regression [31], and Random Forest [14, 15]. The TMLE is then applied to two data studies, a Leukemia analysis [40] and a breast cancer analysis [109].

2.2 Variable importance

The measure of interest is defined in terms of the semiparametric model presented in section 1.3 as follows

$$\Phi(P_0) = E_0[Y|A = a, W] - E_0[Y|A = 0, W]$$

$$= m(a, V|\beta_0) + \theta(W) - m(a = 0, V|\beta_0) - \theta(W)$$

$$= m(a, V|\beta_0)$$
Given this semiparametric form with user supplied \( m(.,) \), the marginal variable importance of a particular \( A \) is defined generally as

\[
\mu_0(a) = E[m(a, V|\beta_0)]
\]

under the constraint \( m(A = 0, V|\beta_0) = 0 \) for all \( \beta_0 \) and \( V \). This semiparametric form allows one to be flexible with how \( m(A = 0, V|\beta_0) \) is specified, allowing the measure to be more realistic given the data. However it is important to remember that the optimal TMLE estimator is developed under the assumption that \( m(.,) \) is correct [102].

Given an estimator \( \beta_n \) of \( \beta_0 \), an estimate of this parameter of interest at a particular \( A = a \) is defined as

\[
\mu_n(a) = \frac{1}{n} \sum_{i=1}^{n} m(a, V|\beta_n)
\]

If the model \( m(A, V|\beta) \) is defined as linear in \( A \) (i.e. \( m(A, V|\beta) = A\beta'V \)), where \( A \) is continuous, the importance can be represented as the linear curve in terms of \( A \) as follows, \( E(m(A = a, V|\beta)) = a\beta E(V) \). Given a linear representation of \( m(.,) \), the parameter of interest becomes a simple linear combination (i.e. \( c'\beta \)) and formal inference can be estimate by applying the delta method.

This chapter focuses on the simplest linear case \( m(A, V|\beta) = A\beta \), where the marginal importance of \( A \) can be represented by single coefficient value \( \beta \). This is the effect measure commonly reported in biomarker applications. Using this simple model allows a direct comparison with alternative measures of importance obtained from univariate and multivariate regression methods.

### 2.3 Targeted maximum likelihood estimation

The parameter \( \beta \) is estimated using targeted maximum likelihood methodology [107]. The targeted maximum likelihood method specific for estimating \( \beta \) under the semiparametric model presented in section 1.3 updates an initial estimate for \( \bar{Q}(A, W) = m(A, V|\beta) + \theta(W) \) in a direction which targets the parameter of interest, \( \beta \). This updates occurs using standard linear regression, where \( Y \) is regressed on a “clever covariate” while setting the initial fit of \( \bar{Q}(A, W) \) as an offset. The “clever covariate” is derived such that the targeted maximum likelihood solution is equivalent to the solution of the efficient estimating equation corresponding to this parameter of interest. Therefore the TMLE is both double robust and locally efficient and correct inference for \( \beta \) can be obtained from the corresponding efficient influence curve. The “clever covariate” specific to this parameter of interest is derived as follows. This derivation has been previously presented in [107].
Assuming an initial normal density, \( f^N \), with conditional mean \( Q^0(A,W) \) and conditional variance \( \sigma^2(A,W) \), a class of submodels is defined by fluctuating \( Q^0(A,W) = E[Y|A,W] \) through parameter \( \epsilon \) as follows

\[
\bar{Q}(\epsilon) = m(A,V|\beta(\epsilon)) - \theta(\epsilon)(W)
\]

where \( \beta(\epsilon) = \beta^0 + \epsilon \) and \( \theta(\epsilon)(W) = \theta^0(W) + \epsilon r(W) \)

The form of \( r(W) \) is chosen such that \( \bar{Q}(\epsilon = 0)(A,W) = \bar{Q}^0(A,W) \), and the score of the normal likelihood with conditional mean \( \bar{Q}(\epsilon = 0)(A,W) \) at \( \epsilon = 0 \) is equivalent to the efficient score equation for this parameter of interest. The efficient score equation, \( D_{h_{\text{opt}},Q,g}(O) \), is defined below [102]. The efficient score equation is indexed by \( Q \) and \( g \), where \( g(W) = P(A|W) \) and is commonly referred to as the “treatment mechanism.”

\[
D_{h_{\text{opt}},Q,g}(O) = h_{\text{opt}}(A,W)(Y - m(A,W|\beta) - \theta(W))
\]

where

\[
h_{\text{opt}} = \frac{1}{\sigma^2(A,W)} \left\{ \frac{d}{d\beta} m(A,V|\beta) - \frac{E \left[ \frac{1}{\sigma^2(A,W)} \frac{d}{d\beta} m(A,V|\beta) \big | W \right]}{E \left[ \frac{1}{\sigma^2(A,W)} \big | W \right]} \right\}
\]

It follows that the proper form of \( r(W) \) is then

\[
r(W) = \frac{E \left[ \frac{1}{\sigma^2(A,W)} \frac{d}{d\beta} m(A,V|\beta) \big | W \right]}{E \left[ \frac{1}{\sigma^2(A,W)} \big | W \right]} \]

With only a small loss in efficiency, one can assume, \( \sigma^2(A,W) = \sigma^2(W) \), reducing the above to the following

\[
r(W) = E \left[ \frac{d}{d\beta} m(A,V|\beta) \big | W \right]
\]

Assuming a linear form for \( m(A,V|\beta) = A\beta'V \), the update can be arranged in terms of the initial estimate \( Q^0(A,W) \) as follows

\[
\bar{Q}(\epsilon)(A,W) = (\beta + \epsilon')(AV) + \theta^0(W) + \epsilon' r(W)
\]

\[
= (\beta'(AV) + \theta^0(W)) + \epsilon'((AV) + r(W))
\]

\[
= \bar{Q}^0(A,W) + \epsilon'((AV) + r(W))
\]

The “clever covariate” is the covariate associated with the fluctuation parameter \( \epsilon \). This is defined for a general linear \( m(A,V|\beta) \) as

\[
H^*(A,W) = \frac{d}{d\beta} m(A,V|\beta) - E \left[ \frac{d}{d\beta} m(A,V|\beta) \big | W \right]
\]
The update parameter $\epsilon$ can be estimated using simple linear regression, regressing $Y$ onto the “clever covariate” with an offset equal to the initial estimate $\bar{Q}^0(A,W)$. This process (i.e. the update) is iterated until $\epsilon = 0$, which maximizes the original likelihood in the direction which results in the best estimate for the parameter of interest. For this TMLE, convergence is achieved in one step, therefore the new TMLE for $\beta$ is defined as $\beta^* = \beta^0 + \epsilon$. The resulting TMLE for $\beta$ is double robust and locally efficient. In other words, given correct model specification for either $\bar{Q}(A,W) = E[Y|A,W]$ or $\bar{g}(W) = E[A|W]$, TMLE is consistent and asymptotically normal and linear. The TMLE is efficient when both models are correctly specified (a.k.a. “locally efficient” ). Therefore, improving the estimates of $E[Y|A,W]$ and $E[A|W]$ using data-adaptive or super learning algorithms is beneficial and will improve the overall consistency and efficiency of the estimate. A discussion of the double robust property of the TMLE and a simulation study exploring the double robust nature is presented in appendix A.1.

2.3.1 Inference and testing

Asymptotically, the TMLE is equivalent to the solution of the efficient estimating equation corresponding to the efficient score equation for the parameter of interest

$$\mathbb{E}[D_{h_{opt},Q,g}(O|\beta)] = 0$$

Therefore, formal inference for the TMLE can be estimated using the efficient influence curve. An estimate of the $p$ by $p$ covariance matrix, $\Sigma_n$, for parameter vector $\beta_n$, of length $p$, is obtained as outlined in section 1.4.3.

Using the estimated covariance matrix, hypothesis tests can be performed for a single parameter $\beta_n(j)$, where $j = 1, \ldots, p$, under the null hypothesis $H_0 : \beta_n(j) = 0$ using a standard test statistic to obtain $p$-values, with estimated variance $\Sigma_n(j,j)$.

$$T_n(j) = \frac{\sqrt{n}\beta_n(j)}{\sqrt{\Sigma_n(j,j)}} \overset{n \to \infty}{\sim} \text{Normal}(0,1)$$

Likewise the hypothesis $H_0 : c'\beta_n = 0$ can also be tested using a standard Wald test, where the covariance of $c'\beta_n$ is $c'\Sigma_nc$. This allows one to obtain inference for $\mu(a)$ directly, when $m$ is linear. In practice the parameter of interest may be redefined as the effect at a specific value of effect modifier $V$ instead of the mean effect as implied by the definition in section 2.2. See section 1.4.3 for details. Alternatively, the covariance can also be estimated by bootstrap estimates of $\beta$, but this would require extra computational time.
2.3.2 TMLE implementation steps

The steps for applying targeted MLE methodology for variable importance under the semiparametric model are outlined below and presented in the flowchart depicted in Figure 2.1.

In biomarker discovery analyses the variable importance is estimated for all variables within the data matrix \( W^* = \{A, W\} \), where \( W^* \) is a matrix of genes, SNPs, or other biological variables of interest. The TMLE estimation method is outlined below for a single \( A = W^*_j \), defining the possible covariate set as \( W = W^*_{-j} \) and in practice is applied over all variables in \( W^* \).

There are three initial components necessary for applying targeted maximum likelihood methodology to estimate the parameter of interest.

1. A model \( m(A, V|\beta) \) satisfying \( m(0, V | \beta) = 0 \) for all \( \beta \) and \( V \), where \( V \subset W \). In this case, it is defined as \( m(A, V|\beta) = \beta A \)

2. An initial regression estimate \( \bar{Q}_0(A, W) = E[Y|A, W] \) of the form \( E[Y|A, W] = m(A, V|\beta_0) + \theta(W) \)

3. An estimate of the “treatment mechanism” \( \bar{g}(W) = E[A|W] \)

The initial regression estimate of proper form may be obtained from semiparametric methods (e.g. \[43\]), or by using methods such as DSA \[88\] which allow the user to fix a portion of the model. However, we recommend more flexible approach which allows one to use a wider range of data-adaptive software. This approach is outlined as follows

- Obtain an initial regression estimate for \( \bar{Q}(A, W) \) of general model form using data-adaptive machine learning algorithms such as super learner \[104\] or DSA \[88\]. This is valid for all \( A \in \{W^*\} \).

- Estimate \( \bar{Q}(A = 0, W) \)

- Using standard linear regression, solve for the initial estimate,

\[
\bar{Q}_n^0(A, W) = m(A, V|\beta_n^0) + \alpha \bar{Q}(A = 0, W)
\]

by specifying model \( m(.) \) and treating \( \bar{Q}(A = 0, W) \) as a covariate.

This provides us with initial parameter estimates for \( \beta \) as well as an initial density estimate of the correct form which can then be updated using TMLE. This is an update from the original method outlined in Tuglus and van der Laan 2008 \[96\], which improves computational efficiency by only requiring a single data-adaptive estimate for \( \bar{Q}(A, W) \) of general model form for all \( A \).

Given these three components, TMLE can easily be applied in the following steps
1. Estimate the “clever covariate” which will allow us to update the initial regression in a direction which targets the parameter of interest. In this case the clever covariate is defined as:

\[ H^*(A, W) = \frac{d}{d\beta} m(A, V|\beta) - E \left[ \frac{d}{d\beta} m(A, V|\beta) \right| W \] 

which for this particular \( m(A, V|\beta) = \beta A \) simplifies to \( H^*(A, W) = A - \bar{g}(W) \)

2. Compute the fitted values for the initial estimate \( \bar{Q}^0_n(A, W) \)

3. Project \( Y \) onto \( H^*(A, W) \) with offset \( = \bar{Q}^0_n(A, W) \) and define the resulting coefficient as \( \epsilon \). This is done using standard software (lm in R) setting the offset, and projecting onto the model \( Y \sim \epsilon' H^*(A, W) + \text{offset} \). Note there is no intercept in the model, only the offset value.

4. Update the initial estimate \( \beta^1_n = \beta^0_n + \epsilon \) and overall density \( \bar{Q}^1_n(A, W) = \bar{Q}^0_n(A, W) + \epsilon_n H^*(A, W) \). These are now the single-step targeted estimates. Since this is a simple linear model, the single step solution is the final solution (i.e. \( \bar{Q}^* = \bar{Q}^1_n \) and \( \beta^*_n = \beta^1_n \)).

5. Obtain standard error and inference for \( \beta_n \) using the empirical estimate of the conservative influence curve as outlined in section 1.4.3.

6. Using the estimated covariance, test the hypothesis \( H_0 : \beta_n(j) = 0 \), using a standard test statistic to obtain p-values.

\[ T_n(j) = \frac{\sqrt{n} \beta_n}{\sqrt{\Sigma_n(j, j)}} \sim \text{Normal}(0, 1) \]

Also note that inference for linear combinations can be obtained by applying the delta method \[96\].
### Table: General Steps for Applying TMLE for Variable Importance

<table>
<thead>
<tr>
<th>Parameter of Interest</th>
<th>Initial Estimate</th>
<th>Targeted Update</th>
<th>Inference</th>
</tr>
</thead>
</table>
| Identify variable of interest, A \[W^* = \{A = W^*_1, W = W^*_2\}\] | **Estimate \(Q(A,W)\) of general form** (i.e. non-parametric fit from data-adaptive algorithm, e.g. super learner) \[
m(A,V|\beta) = A(\beta'V)
\] | **Calculate clever covariate** \(H^*(A,W)\) i.e. \[H^*(A,W) = V(A - \hat{g}(W))\] | **Calculate estimate of influence curve** i.e. \[IC_n = c^{-1} (H^*(A,W)|R)\] \[
R = Y - Q(A,W)
\] \[
c = (H^*(A,W)|AV)
\] |
| Choose model form for \(m(A,V|\beta)\) i.e. \(m(A,V|\beta) = A(\beta'V)\) | **Update** \[
\hat{Q}(A,W) = \hat{e} H^*(A,W) + \hat{Q}(A,W)
\] \[
\hat{\beta}^I = \hat{\beta}^0 + \hat{e}
\] | **Calculate estimate of variance** \[
\hat{\sigma}_n = \frac{1}{n} \sum_{i=1}^{n} IC_i IC_n^T
\]
| \[
\hat{\beta}_n \sim N(\hat{\beta}_n, \frac{1}{n} \hat{\sigma}_n)
\] | **Perform hypothesis testing** \[
T_n(J) = \frac{\sqrt{\hat{\sigma}_n} \delta_n}{\sqrt{\hat{\sigma}_{\hat{\delta}_n}(J,J)}}
\] | **Figure 2.1:** Flowchart depicts the general steps for applying TMLE for variable importance under a semiparametric model. Examples for each step are according to a linear model of the form \(m(A,V|\beta) = A(\beta'V)\)
2.4 Comparison of variable importance methods

In this section, TMLE is compared to three other methods commonly used for determining variable importance in biomarker discovery analyses: univariate linear regression, lasso regression with cross-validation based model-selection \cite{31} using R package \texttt{lars} \cite{30}, and Random Forest \cite{14,15} using R package \texttt{randomForest} \cite{61}. The four methods are listed below. Note that given any estimate, bootstrap sampling may be used to provide standard error estimates and p-values. However in this analysis we choose to compare the methods based on their current merits and accessible implementation, not on any additional processing. Also, in biomarker discovery there are thousands of genes and bootstrap sampling is computationally expensive and impractical.

**Univariate Linear Regression (LM):** Marginal variable importance is represented by the coefficient and p-value resulting from the univariate linear regression fit, $E[Y|A] = \beta A$. P-values are calculated using a standard t-test and are subjected to the Benjamini & Hochberg step-up FDR controlling procedure \cite{11} to control for multiple testing. This method does not account for any confounding and will often misclassify genes correlated with the “true” genes as significant. In most situations this importance measure can not be interpreted in as a causal effect.

**Penalized Regression, LASSO (Q):** Marginal Variable Importance is represented by the coefficient of $A$ in lasso regression main term fit of $\hat{Q}(A, W) = E[Y|A, W]$, where $W_s \subset W$ representing the subset of $W$ found significant according to their univariate regression on $Y$. Lasso regression is applied using R package \texttt{lars} \cite{30}, which does not provide any formal inference therefore p-values are not recorded. Results are compared based on the variable importance measure and its rank. Lasso regression does attempt to account for confounding, but will only allow for n-1 non-zero coefficient values, making its applicability to high dimensional data limited \cite{94}. Lasso regression is also maximum likelihood method which focuses on estimating the overall distribution $E[Y|A, W]$ and not the parameter of interest.

**Targeted Maximum Likelihood (TMLE):** Marginal Variable Importance measure is obtained from applying targeted MLE to the initial density estimate provided by LASSO fit for $\hat{Q}(A, W)$. Coefficient of $A$ is targeted directly, and p-values are provided based in the covariance estimate of the conservative empirical influence curve. P-values are calculated using a standard t-test and are subjected to the Benjamini & Hochberg step-up FDR controlling procedure \cite{11} to control for multiple testing. The importance measure will be represented and compared in terms of the coefficient $\beta_n$. 

Random Forest (RF1 and RF2): Two measures of importance, RF1 and RF2, are provided by the R function `randomForest`. The function is applied directly to the full data matrix $W$ using R package `randomForest` [61], using the default setting with 500 trees. Random Forest provides two measure of importance based on the effect perturbing the variable of interest has on overall classification error and node splits.

- **RF1**: Random Forest importance measure based on “out-of-bag” error rate [14, 15, 61] (no p-values provided)
- **RF2**: Random Forest importance measure based on accuracy of node split [14, 15, 61] (no p-values provided)

Random Forest (RF) is a tree-based algorithm developed by [14, 15] commonly used in biomarker discovery analyses, though it does not estimate the same measure as LM, LASSO, or TMLE. Due to the nature of Random Forest, there is no guarantee that all biomarkers will receive a measure of importance. Also no formal inference is available; therefore no p-values are recorded.

Data is simulated to compare the four approaches under increasing correlation levels using a diagonal block correlation structure. The structure of the simulated data allows us to study the effects that both correlated and uncorrelated variables have on the reported importance of the true variables. For each approach, the biomarkers will be ranked by the resulting importance measure and p-value (when available). The sensitivity and specificity of methods will be compared based on both p-value and rank-based cut-off values, and will be summarized using ROC plots. This analysis will determine the ability of each approach to identify the true variables and each variables true importance rank by comparing the length of list required to label all true variables as “important.”

### 2.4.1 Simulated data

The data is defined as $O = (W^*, Y) \sim P_0$, with covariate matrix $W^*$ and outcome $Y$. Covariate matrix $W^*$ consists of $J=100$ variables with $n=300$ observations simulated from a multivariate normal distribution with block diagonal correlation structure and mean vector created by randomly sampling mean values from \{0.1, 0.2, ..., 9.9, 10.0, 10.1, ...., 50\}, resulting in $K=10$ independent sets of variables, each correlated according to an exchangeable correlation structure with variance, $\sigma_Y$, and specified correlation $\rho_{\text{TRUE}}$. This forms a $J$ by $n$ matrix where each set of ten is correlated among themselves but independent from all other variables.

Outcome $Y$ is simulated from a main effect linear model using one variable from each of the $K$ sets. These $K$ variables are designated as “true variables.” The importance of a variable is determined by its coefficient value in simulation. Two sets
of values are used: a constant value ($\{\beta_k = 4 : k = 1, \ldots, 10\}$) and an increasing set ($\{\beta_k = k : k = 1, \ldots, 10\}$). A normal error with mean zero and variance $\sigma_Y$ is added as noise.

Simulations are run for $\rho_{TRUE} = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ and $\sigma_Y = 1, 10, 20$ using both sets of coefficient values. At $\sigma_Y = 1$ all methods perform very well, resulting in p-values much below zero. At $\sigma_Y = 20$ all methods became largely erratic and overcome by noise. Simulations at $\sigma_Y = 10$ had enough variation to highlight the different strengths of each method and are considered the most realistic noise scenario. For these reasons, only $\sigma_Y = 10$ results are presented in full.

### 2.4.2 Methods

In this analysis, we consider each variable in $W^*$ such that, $A = W_j^*$ and the remaining variables are defined as the covariate set $W = W_{-j}^*$, for all $j = 1 \ldots J$. Importance measures according to the five methods outlined previously (LM, LASSO(Q), TMLE, RF1, and RF2) are calculated for each individual variable (i.e. biomarker), $A$.

We first apply univariate regression to all $J$ biomarkers, estimating $E[Y|A] = \beta_{LM}^A A$. We record each $\beta_{LM}^A$, as the LM importance measure along with its associated p-value, and adjust for multiple testing using Benjamini & Hochberg step-up FDR controlling procedure \cite{benjamini1995controlling} applied using the `mt.rawp2adjp()` R function in package `multtest` \cite{yoon2005multtest}.

To facilitate lasso regression estimation, we first reduce the possible covariate set $W$, to only those variables which are univariate significant with marginal LM adjusted p-value less than $\alpha = 0.05$. We define this reduced set for a given $A$ as $W_s$, and apply lasso penalized regression to the covariate set $\{A,W_s\}$, giving us an initial estimate $\hat{Q}(A,W_s) = E[Y|A,W_s]$. The coefficient of $A$ from the lasso regression fit is recorded as the LASSO (Q) importance measure, and this fit is then used as the initial estimate for TMLE. We use the R library `lars` implementation of lasso regression \cite{tibshirani1996regression,tibshirani1997least}, which does not provide formal inference. Therefore p-values are not recorded.

We estimate $\hat{g}(W) = E[A|W]$ using lasso regression as well. The additive main effect form of a lasso derived model accurately reflects the correlation structure of the data giving us a correct estimate of $\hat{g}(W)$. This guarantees under minimal ETA violations that we will obtain a consistent estimate due to the double robust nature of the TMLE measure \cite{van2007 doubly}. We record the updated TMLE measure as well as its respective p-values. All p-values are adjusted for multiple testing using the Benjamini - Hochberg step-up FDR controlling procedure \cite{benjamini1995controlling}.

Random Forest is applied directly to the full data $W$, and importance measures RF1 and RF2 are calculated internally. Importance measures for Random Forest cannot be directly compared because they are not on the same scale as LM, LASSO (Q), or TMLE estimate. Instead we compare based on importance rank.
2.4.3 Results

For each $\{\rho, \sigma_Y\}$ set, simulations of 100 are completed. Recorded importance measures and p-values are translated into a list of ranks, and the ranks are averaged over the 100 iterations. A rank of one being the largest importance value or smallest p-value. Sensitivity and Specificity calculations for each simulation are also determined for each individual iteration and averaged across the 100 iterations to produce the final estimates. Simulation results are summarized here in terms of “area Under the curve (AUC)” and “length of list”.

Analysis found no appreciable difference when ranking by measure or p-value for LM and TMLE in these simulations, therefore results in terms of measure will be presently in more detail allowing us to include LASSO (Q) and Random Forest measures in all comparisons.

Area under the curve (AUC)

The simulations are set up to test the ability of each method to detect (or classify) the true important variables. The overall performance of a classifier is often summarized in terms of the AUC, the Area Under the Curve derived from the basic ROC curve, which plots the true positive rate (Sensitivity) by the false positive rate (1-Specificity) [35]. Under pure noise conditions $AUC = 0.5$, indicating that at any threshold the false positive and true positive rate are equal (random classifier). The more convex the curve becomes, the higher the AUC, and the better the classifier, and a perfect classifier will have AUC=1. Here, we use the R function `AUCi()` from R package `ROC` which uses `integrate()` to calculate the AUC [18]. The calculated AUC values are plotted versus correlation for each of the five methods using importance measure importance rank, and p-values when available for correlations, $\rho_{TRUE} = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ (Figure 2.2).
Figure 2.2: AUC value from ROC curves by $\rho = 0, ..., .9$ completed for (top) ranking by measure (middle) ranking by p-value, and (bottom) p-value cut-off. The later two only contain values for linear regression and TMLE ($\sigma_Y = 10$). Note: minimum AUC is 0.5, maximum and optimum is AUC=1. Simulation is done with $\sigma_Y = 10$ for $n=300$ with total number of variables at 100 of which 10 are truly related to the outcome. At zero correlation, LASSO (Q), TMLE, and LM perform perfectly with AUC=1. Plots are shown for constant $\beta = 4$, but results are comparable when $\beta = \{1, ..., 10\}$.

From Figure 2.2 we can see that TMLE performs well up to $\rho = 0.6$, performing only marginally better than Q for $\rho > 0.2$, but with AUC visibly greater than Random Forest and LM as correlation increases. As expected LM is most susceptible
to increases in correlation, performing perfectly when correlation is zero, but falling consistently as correlation increases, reaching below 0.8 by $\rho = 0.5$.

**Average length of list**

We can also compare the methods based on the average length of list required to detect all “true” variables. Having a short and accurate list allows the biologist to spend money analyzing the top genes with confidence, knowing that the most important genes are at the top of the list.

The average required list length to find all 10 “true” variables is plotted versus correlation for all five measures and two p-value average ranked lists. These plots are shown for both constant $\beta_{\text{true}} = 4$, and $\beta_{\text{true}} = \{1 \ldots 10\}$. Additional results and plots of the average ranks and importance values can be found in [96] and appendix [B].
Figure 2.3: Total length of list required to have all ten true variables in the list by $\rho = 0, \ldots, 0.9$, ranking by importance measure. ($\sigma_Y = 10$) Results for univariate regression (LM), LASSO (Q), TMLE and two Random Forest based importance measures (RF1, RF2) are shown. Here $\beta_{TRUE}$ is constant at 4.
Figure 2.4: Total length of list required to have all ten true variables in the list by $\rho = 0, ..., 9$, ranking by importance measure. ($\sigma_Y = 10$) Results for univariate regression (LM), LASSO (Q), TMLE and two Random Forest based importance measures (RF1, RF2) are shown. Here $\beta_{TRUE}$ is set at $\{1, ..., 10\}$. 
Figure 2.5: Total length of list required to get all ten true variables by $\rho = 0, \ldots, 9$, ranking by p-value. ($\sigma_Y = 10$) Results for univariate regression (LM), and TMLE are shown. Here $\beta_{TRUE}$ is constant at 4.
Figure 2.6: Total length of list required to get all ten true variables by $\rho = 0, \ldots, 9$, ranking by p-value. ($\sigma_Y = 10$) Results for univariate regression (LM), and TMLE are shown. Here $\beta_{TRUE}$ is set at $\{1, \ldots, 10\}$.

Length of list is a direct reflection of Type I error or false discovery rate. We see that overall TMLE performs well up to correlations of 0.9, though the improvement over LASSO is less clear when $\beta_{TRUE}$ is constant (Figure 2.3, Figure 2.5). In the case where $\beta_{TRUE} = \{1, \ldots, 10\}$ (Figure 2.4, Figure 2.6), the improvement of TMLE over LASSO is more pronounced, but detection of the first variable (with the lowest $\beta$ value) is difficult for all methods. When ranking by measure or p-value, all methods have their lowest list length around 20 variables while the total number of variables
expected is 10. In contrast, when $\beta$ was constant at value 4, the lowest list length was near its minimum at 10 (Figure 2.3, Figure 2.5). The shift in list length most likely due to the importance value for the variable associated with $\beta = 1$. At such a high noise level ($\sigma_Y = 10$), the lower importance values are more difficult to distinguish from the noise. This is apparent by comparing the average importance rank and average importance value for the variable with $\beta = 1$ (see [96] and appendix B). The rank is much higher than 10, but the value is close to one as it should be.

In general, TMLE has the shortest list and is less affected than any other methods by increases in correlation. Increases above the minimal list length were seen for both Random Forest and univariate regression methods at correlation values greater than 0.4, with Random Forest faring better at higher correlations. When ranking by p-value and similar trend for univariate regression was apparent.

### 2.4.4 Discussion

These simulations address each methods ability to accurately identify the causally related genes as the correlation among variables increases. Though TMLE performs better than the three other methods, it is still sensitive to more extreme correlations (0.7-0.9). Our simulations show a small increase in bias for the measure of the true variables at higher correlations (see [96] and appendix B). However, in practice, high correlation can adversely effect the TMLE estimate due to violation of the experimental treatment assumption. The increased length of the variable list when ranked by importance measure at correlation 0.8 and 0.9 indicates that TMLE cannot distinguish the true variable from among a group of variables when correlation is very high.

### 2.5 Experimental treatment assumptions and consequences of its violation

When variables are highly correlated with the variable of interest $A$, experimental treatment assumption (ETA) violations often occur, which reduces the ability to estimate the effect accurately. Consider a drug trial for breast cancer. If the drug is assigned to people who also undergo radiation treatment and individuals in the placebo group generally do not undergo radiation, then the effect of the drug and radiation on the cancer cannot be distinguished. Formally, the experimental treatment assumption states that the probably of $A$ given $W$ must always be positive for all possible sets $(a, W)$, $(P(A|W) > 0 \forall (a, W))$ [117]. In other words all values of $A$ must be possible given any observed set of values $W$, and no $W$ can be a perfect predictor of $A$. If either is invalid, estimation of the effect of $A$ will require extrapolation. This introduces bias and, if the ETA violation is extreme enough, can result in a non-identifiable importance estimate.
If our semiparametric model is correct, extrapolation is less of a concern. However, in practice, we cannot assume a correct semiparametric model. In the more realistic case, where we view our importance parameter as a projection onto a working semiparametric model, violations of ETA can result in a highly sensitive estimate of $Q(A, W)$ leading to instability in the importance (parameter) estimate.

Variable importance measures are also affected by ETA violations through the form of the empirical influence curve used in targeted maximum likelihood estimation. The methods for binary $A$ presented in [9] use inverse weights of the treatment mechanism $\left(\frac{1}{P(A=a|W)}\right)$ for the TMLE update and inference calculation. When the $P(A = a|W)$ becomes very small from ETA violation, these weights explode leading to unreliable importance estimates. In comparison, the effect of ETA violation on the semiparametric variable importance presented here is less extreme, but still a concern. It’s influence curve is weighted by $(A - E[A|W])$ (for the univariate case), which effectively downplays observations responsible for ETA violations. Under large ETA violation, the measure is only accounting for a small subset of the observations making it a less applicable and interesting importance.

ETA violations can often be avoided if the “problem” variables (the variables highly correlated with the gene of interest $A$), are removed from the set of confounders ($W$). One simple method is to apply a correlation cut-off, where all $W$ whose correlation with $A$ is greater than a particular correlation ($\rho_\delta$), are removed from the set of possible confounders for variable $A$ prior to the application of TMLE method. We explored this briefly through simulation.

In simulation study analogous to the previous set-up, a correlation cut-off was applied to subset $W_s$ for each $A$ before LASSO analysis. In this scenario, $W_s$ is restricted to all $W_i \in W_s$ where $\text{cor}(A, W_i) < \rho_\delta$, for various cut-offs $\rho_\delta = \{0.5, 0.75, 0.9, 1\}$. We applied this method to our simulated data from the previous section. Results showed that such a restriction resulted in the elimination of relevant $W_i$ from the estimate of $E(Y|A, W_s)$. In other words when $A_d$ is a decoy variable highly correlated with a true variable $W_t$. Restrictions on the covariate set remove $W_t$ from the possible covariate set for $A_d$, resulting in $A_d$ having a higher and more significant importance that it would have otherwise.

The restriction of $\rho_\delta$ results in the algorithm identifying all true variables as well as variables whose correlation with the true variables is higher than $\rho_\delta$. Once we select $\rho_\delta$, we are conceding that variables with correlations greater than $\rho_\delta$ cannot be teased apart to determine the true underlying (important) variable. By applying the correlation cut-off we are redefining our parameter. It is no longer the singular effect of $A$. Instead, we admit that given the data, the true important variable cannot be targeted when the data is highly correlated and redefine our measure as a correlation-based $W_\delta$-adjusted importance where $W_\delta$ is a newly defined subset of $W$ based on the correlation cut-off. Given this new definition of the parameter, important variables according to the $W_\delta$-adjusted method include all important variables as well as all
variables whose correlation to a important variable is greater than a particular delta cut-off.

Therefore we must be careful when selecting $\rho$, it must high enough to reduce bias from ETA violation, but low enough to acquire all information on the causal effects allowed by the data, which maintains the greatest level of reproducibility. If $\rho$ is higher than necessary, the list will contain decoy variables that could have been discounted using the available data. This would decrease the reproducibility of the measures in other populations. The relationship between the decoy variables and the causal variables (distribution of $W$) is not necessarily constant across populations while the causal mechanism (distribution of $Y|W$) can be assumed to be (i.e. the mechanisms of disease are consistent across all populations). Including decoy variables that could otherwise have been discounted adds unnecessary uncertainty when applying the final results to other populations. A method was proposed in [6], which defines an analytical formula for identifying these “problem” variables data-adaptively for each $A$. This reduces the bias while detecting the most accurate gene set allowed by the data, maintaining reproducibility.

In the next section, we apply the correlation cut-off ($\rho = \{0.5, 0.75\}$) to a leukemia application, where the truth is unknown, and the data is noisy. In practice it is reasonable to label all potentially relevant variables as important when their effects cannot be disentangled. Setting a correlation cut-off explicitly specifies and acknowledges the method’s threshold to detect the important variables among highly correlated confounders. We recommend that future applications use a larger set of $\rho$ values and provide importance measures and rankings for all variables given each $\rho$, or data-adaptively select $\rho$ using the methods outlined in [6].

### 2.6 Application

In practice, when applying any method to high dimensional and highly correlated biomarker data, often it is beneficial to employ pre-screening to reduce the initial data set and/or covariate screening to increase the stability of importance estimates. We first briefly present discuss ramifications of pre-screening and cut-off choice with respect to TMLE as well as variable importance methods in general. We then follow with a biomarker discovery application of TMLE to the Golub Leukemia data [40]. Results are discussed in terms of their biological relevance.

#### 2.6.1 Implementation considerations

**Pre-screening**

Biomarker data is generally high dimensional and highly correlated, therefore certain pre-screening is necessary prior to performing a biomarker analysis. We are primarily concerned with screening the potential covariate set $W$. Reducing this set
to relevant biomarkers can not only decrease computation time, but also can result in better estimates from data-adaptive algorithms.

We want to reduce this set of covariates to include only potential confounders. Potential confounders for a given $A$ are any $W$ which are related to both $Y$ and $A$. However, it can be time consuming to screen for every $A$ separately. To save time, we recommend screening only in terms of $Y$. This can be accomplished by discounting any $W$ not significantly associated with $Y$ using simple univariate regression, or a combination of results from multiple methods (e.g. all variables significant according to at least one of: linear regression, Random Forest, or lasso).

The above screening can also serve to reduce the number of variables for which we estimate variable importance (i.e. variables $A$). Removing these variables presumes they have insignificant importance. If the data is reduced based on the outcome $Y$, this reduction must be accounted for in any subsequent multiple testing procedures. An easy way to accomplish this is, after estimating the importance measure and calculating the associated p-values for one’s subset of the full variable set, automatically assign all pre-screened variables (i.e. variables with no estimate) a p-value of one. Then apply Benjamini & Hochberg step-up FDR controlling procedure [11] to the full set of p-values as usual and apply p-value cut-off to identify significant genes. Assuming the pre-screening has only discounted variables which would have had p-values greater than the cut-off, the procedure will retain the type I and II control of the Benjamini & Hochberg step-up FDR controlling procedure. This is referred to as the Modified FDR procedure. See [97] for more details on the theory behind this procedure.

$\delta$ cut-off

The experimental treatment assumption (ETA) dictates that the probability of any $A = a$, must be positive for any $W = w$ (see section 2.5). If $A$ is highly correlated with a particular covariate in $W$, for instance $W_1$, this will result in ETA violations because the support of $A = a$ does not extend over all possible values of $W_1$. In other words, under perfect correlation, all observations with $A = a$ will be restricted to $W_1 = w_1 = \rho A$, where $\rho$ is the correlation of $A$ and $W_1$. High enough correlation will lead to instabilities in the importance estimate. This is more of a concern with the non-parametric variable importance presented in [9] which requires inverse weighing by the treatment mechanism. However it can still lead to instabilities in the TMLE estimate (see simulations in section 2.4).

To combat this effect, we can reduce the covariate set for a given $A$, by applying a simple correlation cut-off to the covariate set before applying the TMLE procedure. For instance, one can remove any $W$ which has correlation with $A$ greater than $\rho_c$. However, if $A$ is significantly and causally related to $Y$, any removed $W$ which are confounders of $A$ will register as significantly important regardless of there individual causality (see section 2.5). In other words, given a correlation cut-off, TMLE will
identify all causally related variables as well as all variables the data is unable to disentangle due to the high correlation structure. Therefore caution must be taken with how low of a correlation to use. In application, we use multiple correlation cut-offs (i.e. $\rho_c = \{0.5, 0.75\}$) and compare results, referring to the final ranking as the realistic ranking of the variables given the data.

2.7 Golub et al (1999)

2.7.1 Data

The data set from Golub et al. 1999 [40] has been used in many papers for methodological comparison, due to its relevance, limited gene set, and biological interpretability. One goal in the original study was to identify differentially expressed genes in patients with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In Table 2.1 we summarize the basic differences between ALL and AML type leukemia.

<table>
<thead>
<tr>
<th>Traits</th>
<th>AML</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effected population</td>
<td>Most common in adults</td>
<td>Most common in children</td>
</tr>
<tr>
<td>Biological characteristics</td>
<td>Identified with the myeloid line of white blood cells, which includes any leukocyte that is not a lymphocyte</td>
<td>Identified with abnormal lymphocytes, B-cells, T-cells, and natural killer (NK) cells</td>
</tr>
<tr>
<td>Clinical characteristics</td>
<td>Anemia, fatigue, weight loss, easy bruising, thrombocytopenia, and granulocytopenia with bacterial infections</td>
<td>Anemia, fatigue, weight loss, easy bruising, thrombocytopenia, and granulocytopenia with bacterial infections, bone pain, lymphadenopathy and hepatosplenomegaly</td>
</tr>
<tr>
<td>Relative overall survival rate</td>
<td>Low</td>
<td>Higher in both adults and children</td>
</tr>
<tr>
<td>Additional characteristics</td>
<td></td>
<td>Can spread to the nervous system</td>
</tr>
</tbody>
</table>

Gene expression levels were measured using Affymetric oligonucleotide arrays with 6,817 human genes for n=38 patients (27 ALL, 11 AML). The gene expression set was pre-processed and reduced to 3,051 genes according to methods described in [28]. This data set was obtained from the R package multtest, data set golub [39].
2.7.2 Analysis

This analysis mirrors the procedure implemented in the previous simulations. We first apply univariate linear regression to all genes and control for multiple testing using Benjamini & Hochberg step-up FDR controlling procedure [11]. In this application, we chose all genes with p-value less than 0.1 as the initial covariate set \( W \) for each \( A \). This set contains 876 genes initially. To minimize bias due to ETA violations, a simple correlation cut-off of \( \rho_c = \{0.5, 0.75\} \) is applied. In application where we do not known the truth and the data is especially noisy with a complex correlation structure, often the effect of an individual gene can not be disentangled. Applying the correlation cut-off results in labeling all potentially relevant genes as important.

As in simulation we model the importance as \( m(A, V | \beta) = \beta A \) for all \( A \). For the initial \( \bar{Q}^0(A, W_s) \) and \( \bar{g}(W_s) \) we use a polynomial spline fit. We recommend using this or a similar data-adaptive algorithm such as super learner [104] over lars/lasso in application, since in reality the structure of \( E[Y|A, W] \) and \( E[A|W] \) may have more than just additive main effects.

In this application, the outcome is binary ( ALL (\( Y=0 \)) vs. AML (\( Y=1 \)) ), therefore we can interpret the parameter of interest as an estimate of excess risk.

\[
m(A = a, V | \beta) = E[Y|A = a, W_s] - E[Y|A = 0, W_s] = \beta a
\]

The model-based approach outlined in this paper must use standard gaussian regression for our estimate and update of \( E[Y|A, W_s] \). However we believe the final list of ranked importance measures and p-values are still relevant regardless.

TMLE updated measures and p-values from standard t-tests are recorded. We adjust for multiple testing using Benjamini & Hochberg (1995) step-up FDR controlling procedure [11]. We recommend selecting all genes with adjusted p-values less than or equal to an appropriate cut-off (we use a standard cut-off of 0.05), and then ranking this set of genes by their absolute importance measures to achieve the final importance ranking of genes. The same method is used to rank genes according to the LM measures and p-values. RF1 and RF2 importance measures are simply ranked. We compare the results in Tables 2.2 - 2.6.

2.7.3 Results

Using a p-value cut-off of 0.05, TMLE results in 272 significant genes at \( \rho_c = 0.5 \) and 225 significant genes at \( \rho_c = 0.75 \), while LM identifies 681 significant genes. We plot the number of significant genes for a range of p-value cut-offs in Figure 2.7. The overall trend showing LM to be much more conservative than the TMLE.
Figure 2.7: The number of significant genes (y-axis) given a p-value cut-off (x-axis) is plotted for LM and TMLE results for $\rho_c = 0.5, 0.75$
Table 2.2: Univariate Linear regression (LM): Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

<table>
<thead>
<tr>
<th>Gene Name/Gene Symbol</th>
<th>Mapped IDs</th>
<th>LM rankp</th>
<th>LM rankp (0.75)</th>
<th>TMLE rankp</th>
<th>TMLE rankp (0.5)</th>
<th>RF1 rank</th>
<th>RF2 rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystatin C</td>
<td>M27891.at</td>
<td>0.258</td>
<td>1</td>
<td>13</td>
<td>17</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Cystatin A</td>
<td>D88422.at</td>
<td>0.341</td>
<td>2</td>
<td>521</td>
<td>466</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Zyxin</td>
<td>X95735.at</td>
<td>0.345</td>
<td>3</td>
<td>287</td>
<td>534</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Macmarcks</td>
<td>HG1612-HT1612.at</td>
<td>-0.619</td>
<td>4</td>
<td>1041</td>
<td>1768</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>CD33 antigen (differentiation antigen)</td>
<td>M23197.at</td>
<td>0.517</td>
<td>5</td>
<td>906</td>
<td>28</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td>C-myb gene extracted from Human (c-myb) gene, v-myb myeloblastosis viral oncogene homolog (avian)</td>
<td>U22376.cds2.s.at</td>
<td>-0.403</td>
<td>6</td>
<td>69</td>
<td>99</td>
<td>40</td>
<td>28</td>
</tr>
<tr>
<td>ELA2: Elastase 2, neutrophil</td>
<td>M27783.s.at</td>
<td>0.334</td>
<td>7</td>
<td>104</td>
<td>1970</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>DF: D component of complement (adipsin)</td>
<td>M84526.at</td>
<td>0.262</td>
<td>8</td>
<td>175</td>
<td>145</td>
<td>96</td>
<td>149</td>
</tr>
<tr>
<td>Retinoblastoma Binding Protein P48</td>
<td>X74262.at</td>
<td>-0.431</td>
<td>9</td>
<td>291</td>
<td>266</td>
<td>57</td>
<td>31</td>
</tr>
<tr>
<td>LTC4S: Leukotriene C4 synthase gene</td>
<td>U50136.rna1.at</td>
<td>0.725</td>
<td>10</td>
<td>146</td>
<td>2110</td>
<td>38</td>
<td>60</td>
</tr>
</tbody>
</table>
Table 2.3: RF1: Top 10 ranked genes according to their importance measures

<table>
<thead>
<tr>
<th>Gene Name/Gene Symbol</th>
<th>Mapped IDs</th>
<th>RF1 rank</th>
<th>RF2 rank</th>
<th>TMLE rankp</th>
<th>TMLE rankp</th>
<th>LM rankp</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAH: Fumarylacetoacetate</td>
<td>M55150.at</td>
<td>0.953 1</td>
<td>1 588</td>
<td>234 52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zyxin</td>
<td>X95735.at</td>
<td>0.823 2</td>
<td>2 287</td>
<td>534 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCF3: Transcription factor 3</td>
<td>M31523.at</td>
<td>0.718 3</td>
<td>6 155</td>
<td>400 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADM: Adrenomedullin</td>
<td>D14874.at</td>
<td>0.693 4</td>
<td>5 329</td>
<td>2136 57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTX3: Pentaxin-related gene, rapidly induced by IL-1 beta</td>
<td>M31166.at</td>
<td>0.691 5</td>
<td>33 33</td>
<td>201 28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CST3: Cystatin C (amyloid angiopathy and cerebral hemorrhage)</td>
<td>M27891.at</td>
<td>0.682 6</td>
<td>3 13</td>
<td>17 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOP2B: Topoisomerase (DNA) II beta (180kD)</td>
<td>Z15115.at</td>
<td>0.654 7</td>
<td>4 1</td>
<td>2 33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCND3: Cyclin D3</td>
<td>M92287.at</td>
<td>0.621 8</td>
<td>10 481</td>
<td>924 19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macmarcks</td>
<td>HG1612-HT1612.at</td>
<td>0.613 9</td>
<td>9 1041</td>
<td>1768 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APLP2: Amyloid beta (A4) precursor-like protein 2</td>
<td>L09209.s.at</td>
<td>0.610 10</td>
<td>7 160</td>
<td>408 25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.4: RF2: Top 10 ranked genes according to their importance measures

<table>
<thead>
<tr>
<th>Gene Name/Gene Symbol</th>
<th>Mapped IDs</th>
<th>RF2 rank</th>
<th>RF1 rank</th>
<th>TMLE rankp</th>
<th>TMLE rankp</th>
<th>LM rankp</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAH: Fumarylacetoacetate</td>
<td>M55150.at</td>
<td>0.426</td>
<td>1</td>
<td>1</td>
<td>588</td>
<td>234</td>
</tr>
<tr>
<td>Zyxin</td>
<td>X95735.at</td>
<td>0.282</td>
<td>2</td>
<td>2</td>
<td>287</td>
<td>534</td>
</tr>
<tr>
<td>CST3: Cystatin C (amyloid angiopathy and cerebral hemorrhage)</td>
<td>M27891.at</td>
<td>0.218</td>
<td>3</td>
<td>6</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>TOP2B: Topoisomerase (DNA) II beta (180kD)</td>
<td>Z15115.at</td>
<td>0.208</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ADM: Adrenomedullin</td>
<td>D14874.at</td>
<td>0.200</td>
<td>5</td>
<td>4</td>
<td>329</td>
<td>2136</td>
</tr>
<tr>
<td>TCF3: Transcription factor 3</td>
<td>M31523.at</td>
<td>0.186</td>
<td>6</td>
<td>3</td>
<td>155</td>
<td>400</td>
</tr>
<tr>
<td>APLP2: Amyloid beta (A4) precursor-like protein 2</td>
<td>L09209.s.at</td>
<td>0.183</td>
<td>7</td>
<td>10</td>
<td>160</td>
<td>408</td>
</tr>
<tr>
<td>Cystatin A</td>
<td>D88422.at</td>
<td>0.171</td>
<td>8</td>
<td>12</td>
<td>521</td>
<td>466</td>
</tr>
<tr>
<td>Macmarcks</td>
<td>HG1612-HT1612.at</td>
<td>0.164</td>
<td>9</td>
<td>9</td>
<td>1041</td>
<td>1768</td>
</tr>
<tr>
<td>CCND3: Cyclin D3</td>
<td>M92287.at</td>
<td>0.159</td>
<td>10</td>
<td>8</td>
<td>481</td>
<td>924</td>
</tr>
</tbody>
</table>
Table 2.5: TMLE using correlation cut-off of $\rho_c = 0.5$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

<table>
<thead>
<tr>
<th>Gene Name/Gene Symbol</th>
<th>Mapped IDs</th>
<th>TMLE rankp</th>
<th>TMLE rankp</th>
<th>LM rankp</th>
<th>RF1 rank</th>
<th>RF2 rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.5)</td>
<td>(0.75)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOP2B: Topoisomerase (DNA) II beta (180kD)</td>
<td>Z15115.at</td>
<td>-0.973</td>
<td>1</td>
<td>2</td>
<td>33</td>
<td>7</td>
</tr>
<tr>
<td>CHRNA7: Cholinergic receptor, nicotinic, alpha polypeptide 7 corneodesmosin</td>
<td>X70297.at</td>
<td>0.839</td>
<td>2</td>
<td>1</td>
<td>48</td>
<td>69</td>
</tr>
<tr>
<td>BCL3: B-cell CLL/lymphoma 3</td>
<td>L20815.at</td>
<td>0.338</td>
<td>3</td>
<td>3</td>
<td>1875</td>
<td>1846</td>
</tr>
<tr>
<td>KTN1: kinectin 1 (kinocin receptor)</td>
<td>U05681.s.at</td>
<td>0.314</td>
<td>4</td>
<td>4</td>
<td>477</td>
<td>558</td>
</tr>
<tr>
<td>CaM: kinase II isoform mRNA</td>
<td>Z22551.at</td>
<td>-0.311</td>
<td>5</td>
<td>18</td>
<td>373</td>
<td>2967</td>
</tr>
<tr>
<td>TCF7: transcription factor 7 (T-cell specific, HMG-box)</td>
<td>U81554.at</td>
<td>0.272</td>
<td>6</td>
<td>81</td>
<td>367</td>
<td>476</td>
</tr>
<tr>
<td>PTTG1IP: pituitary tumor-transforming 1 interacting protein</td>
<td>X59871.at</td>
<td>-0.159</td>
<td>7</td>
<td>6</td>
<td>569</td>
<td>635</td>
</tr>
<tr>
<td>MCL1: myeloid cell leukemia sequence 1 (BCL2-related)</td>
<td>Z50022.at</td>
<td>0.310</td>
<td>8</td>
<td>5</td>
<td>2753</td>
<td>2674</td>
</tr>
<tr>
<td>PI3K: Phosphatidylinositol 3-kinase</td>
<td>L08246.at</td>
<td>0.293</td>
<td>9</td>
<td>2406</td>
<td>61</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Z46973.at</td>
<td>-0.172</td>
<td>10</td>
<td>113</td>
<td>734</td>
<td>772</td>
</tr>
</tbody>
</table>
When comparing the four lists it is difficult to determine which list is better. Especially when the lists include hundreds of genes. In this analysis we compare the top 10 of each list in an effort to compare their biological relevance. In any given list we include the top 10 genes of the particular method along with their ranks for all other methods. For many of the genes, these ranks vary greatly over the different methods. By consulting the literature, we hope to gain insight on the biologically validity of each list. The top 10 genes according to their importance ranking for LM, RF1, RF2, and TMLE ($\rho_c = 0.5, 0.75$) are shown in Tables 2.2 - 2.6.

Among the top 10 genes according to LM results, CSTA, CD33, MYB, and ELA2 have all been associated with various types of cancer in the literature in previous quan-

Table 2.6: TMLE using correlation cut-off of $\rho_c = 0.75$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

<table>
<thead>
<tr>
<th>Gene Name/Gene Symbol</th>
<th>Mapped IDs</th>
<th>TMLE rankp</th>
<th>LM rankp</th>
<th>RF1 rank</th>
<th>RF2 rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(0.75)</td>
<td>(0.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHRNA7: Cholinergic receptor, nicotinic, alpha polypeptide 7</td>
<td>X70297.at</td>
<td>1.260</td>
<td>1</td>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>TOP2B: Topoisomerase (DNA) II beta (180kD) corneodesmosin</td>
<td>Z15115.at</td>
<td>-0.946</td>
<td>2</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>BCL3: B-cell CLL/lymphoma 3</td>
<td>L20815.at</td>
<td>0.327</td>
<td>3</td>
<td>3</td>
<td>1875</td>
</tr>
<tr>
<td>Surface glycoprotein</td>
<td>U05681.s.at</td>
<td>0.181</td>
<td>4</td>
<td>4</td>
<td>477</td>
</tr>
<tr>
<td>TCF7: Transcription factor 7 (T-cell specific)</td>
<td>Z50022.at</td>
<td>0.310</td>
<td>5</td>
<td>8</td>
<td>2753</td>
</tr>
<tr>
<td>CAT: Catalase (EC 1.11.1.6) 5’flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)</td>
<td>X50871.at</td>
<td>-0.175</td>
<td>6</td>
<td>7</td>
<td>569</td>
</tr>
<tr>
<td>E2F4: transcription factor Dp-2 (E2F dimerization partner 2)</td>
<td>X04085.rna1.at</td>
<td>0.163</td>
<td>7</td>
<td>21</td>
<td>92</td>
</tr>
<tr>
<td>UGP2 Uridine diphosphoglucose pyrophosphorylase mRNA</td>
<td>U18422.at</td>
<td>-0.256</td>
<td>8</td>
<td>42</td>
<td>1752</td>
</tr>
<tr>
<td>SELL: Leukocyte adhesion protein beta subunit</td>
<td>U27460.at</td>
<td>-0.244</td>
<td>9</td>
<td>14</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>M15395.at</td>
<td>0.183</td>
<td>10</td>
<td>43</td>
<td>340</td>
</tr>
</tbody>
</table>
titative analyses. CSTA has been proposed as a diagnostic and prognostic biomarker for cancer [52]. CD33 antigen has been shown in vitro to induce apoptosis in AML cells [110]. MYB is the homolog of an avian viral oncogene [22], and ELA2 has been related to acute promyelocytic leukemia [54].

Among the top 10 genes according to RF1 and RF2, all genes in RF1 were also in the top 10 for RF2 except CBX1, it was replaced by CSTA in the list for RF1. CSTA was also in the top 10 of LM. Out of the top 10 the following genes have been associated with various cancers: TCF3, TOP2B, CCND3, and CSTA. Chromosomal abnormalities in TCF3 have been linked to T-cell and B-cell ALL [44]. TOP2B is a current drug target having been linked to drug resistant cancers [72, 46]. CCND3 is a cyclin D. In the absence of cyclin D’s cells have shown increased resistance to oncogenic transformation in mouse models [53].

There are marked differences and similarities between the TMLE based results using a correlation cut-off of 0.5 and 0.75. There are 5 genes that are common between the two lists, 4 of which have some cancer-related association: TOP2B, CHRNA7, BCL3, and TCF7. Directional relationships remain consistent between the two lists, but the magnitudes shift due to the different covariate sets. TOP2B, a current drug target [72, 46], was also identified by Random Forest. BCL3 is a proto-oncogene biologically associated with B-cell ALL [68]. TCF7 is a known biomarker for T-cell ALL, and is rarely expressed in AML cancer cells [75]. CHRNA7 has recently found to inform the role of nicotine in colon cancer [118]. It is also important to note that CHRNA7 is highly correlated with CD33. Cancer relevant genes found only in Table 2.5 ($\rho_c = 0.5$) are PTTG1IP, MCL1, PI3K, and CAMK2G. PTTG1IP has been consistently found overly expressed in human tumors [77, 76, 37, 123]. MCL1 is related to BCL2 and is a negative regulator of apoptosis [47]. PI3K is activated by cellular agents known to stimulate B and T cells [36]. CAMK2G has an active role in cell growth control and has tumor cell-specific variants [95]. Cancer relevant genes found only in Table 2.6 ($\rho_c = 0.75$) are CAT and E2F4. CAT regulates BCL-2 and is often under-expressed in ALL tissues [84, 50]. E2F4 has an essential role in cell proliferation and cell fate decisions [4] as well as activation of tumor suppressor proteins [58].

### 2.7.4 Discussion

When comparing the four lists it is difficult to determine which list is better. By shifting through the literature, we had hoped to gain a sense of biologically validity associated with each list.

Using simple univariate linear regression 681 genes were significant at the 0.05 level after adjusting for multiple-testing. However we know from general knowledge and our simulations, that univariate linear regression is highly sensitive to correlation among the variables, leading to large increases in type I error rate. Given this and a set of 681 genes, attempting to further analyze the lists to identify and biologically verify
the relevant genes seems a nearly impossible and very expensive task. Attempting to control type I error by adding additional covariates requires model selection methods that are geared towards prediction.

Random Forest is a prediction and classification method which includes a type of model selection. The importance measures provided by Random Forest are difficult to interpret. Given an importance value of 0.612 the relationship between the variable and the outcome is unclear - is it highly expressed in AML or ALL? We only know that that variable is more “important” than a variable with a value of 0.611. Also out of the top 10 lists for RF1 and RF2 (12 genes total), only four genes were found to be biologically associated with cancer and only one specifically relating to ALL/AML distinction, TCF3. Why TCF3 is rated second for RF1 and sixth for RF2 is unclear. In comparison, LM found four related to cancer, two of which specifically related to AML/ALL.

The TMLE measure provides directionality and is less sensitive to increases in correlation (see section 2.4). Given a importance measure of -0.175, we can conclude that this particular gene is up-regulated in ALL patient when compared to AML patients. This particular measure is for TCF7 using a correlation cut-off of 0.75. TCF7 is rarely expressed in AML and often highly expressed in ALL patients (esp. T-cell related). Out of the 6 cancer related genes in the top 10 list for 0.75 cut-off, 3 are biologically related to the AML/ALL distinction. When the cut-off is 0.5, there are 8 cancer related genes, 3 related to the AML/ALL distinction. For all three AML/ALL related genes the directionality of the relationship is biologically correct.

The TMLE results do have a greater number of cancer-related genes and a greater number of specifically AML/ALL related genes. However the increase over LM is small, and the comparison only includes the top 10 genes. Further support for TMLE is gained from the previous simulations where we demonstrated its resistance to increases in correlation and its control of type I error, while still being an interpretable and meaningful measure of importance.

2.8 van’t Veer et al (2002)

The response to standard chemotherapy among breast cancer patients can drastically vary even among women with a common stage of breast cancer at initial diagnosis. Chemotherapy is a very long and difficult treatment process, and though it is known to reduce the occurrence of metastases in 70-80% of patients, for the remaining 30-20% there is little or no response. Knowing a priori a probability of response to a given patient would aid doctors in determining a more optimal and efficient treatment plan, reducing patient discomfort and the cost of expensive trail-and-error treatment regimes. This is reflective of the current trend towards the development of individualized or “patient-tailored” treatments.

The study in [109] attempts to develop a classifier predicting treatment response
to adjuvant chemotherapy among breast cancer patients based on their pre-treatment (at diagnosis) genetic profile. Given that there are over 20,000 protein-coding genes in the human genome, developing a predictor requires first reducing the data to a set of relevant genes. Here we present an application of the TMLE as a method to identify these genes. Unlike linear regression and other data mining algorithms (Random Forest, etc.), the TMLE targets the causal effect instead of estimating only an association based on a predictive fit. We propose using the TMLE to determine this subset of genes prior to the application of the prediction algorithm super learner [104].

2.8.1 Data

The initial data set contains 98 patients with similar stages of breast cancer at the time they enter the study. All patients are exposed to adjuvant chemotherapy. It is unknown if any other treatment methods (i.e. radiation, surgery, etc.) are applied and to what extent. For the purposes of the van’t Veer analysis, the patients are assumed to be part of the same treatment arm. We continue with that assumption. Of the 98 patients, 34 develop metastases within 5 years (bad responders, Y=1), while 44 remain disease free (good responders, Y=0) [109].

2.8.2 Analysis

For computation considerations we reduced our data set to genes whose raw p-values from univariate linear regression were less than or equal to 0.05 (2254 genes) or those which had a Random Forest importance value greater than zero. We also did not include genes with more than 80% of their values missing. This left us with a total of 4446 genes. Its important to note that once we adjusted for multiple testing, there were no adjusted significant p-values at the 0.05 level among these genes. All missing data is imputed with the column mean (average gene expression over all patients). The maximum number of missing values for any gene was five.

We apply the TMLE using correlation cutoffs of 0.5 and 0.75 as outlined above. The covariate set prior to correlation cut-off included all genes among the 4446 whose raw univariate linear regression p-value was less than or equal to 0.01 (540 genes). Genes significant at the 0.05 level are used as input to super learner these results are outlined in [104]. Here we explore the relevance of the genes obtained using targeted maximum likelihood. We again rank the significant genes by their importance values.

2.8.3 Results

There were no statistically significant genes (at the 0.05 level) once the univariate linear regression p-values were adjusted for multiple testing, while for TMLE
VIM there were 197 and 204 genes when correlation cut-off was set at 0.5 and 0.75 respectively. We show the top 10 significant genes in Table 2.7 and Table 2.8.

Table 2.7: TMLE using correlation cut-off of $\rho_c = 0.5$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

<table>
<thead>
<tr>
<th>GeneID</th>
<th>Description/Function</th>
<th>TMLE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GALNT14 (AA165698)</td>
<td>UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 14</td>
<td>6.455</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>AIP</td>
<td>aryl hydrocarbon receptor interacting protein</td>
<td>6.164</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>LRTM1</td>
<td>leucine-rich repeats and transmembrane domains</td>
<td>3.517</td>
<td>13.03E-05</td>
</tr>
<tr>
<td>ZBTB22 (AI524306)</td>
<td>zinc finger and BTB domain containing 22</td>
<td>3.125</td>
<td>2.33E-07</td>
</tr>
<tr>
<td>VAMP3</td>
<td>vesicle-associated membrane protein 3 (cellubrevin)</td>
<td>-2.714</td>
<td>1.58E-06</td>
</tr>
<tr>
<td>ERGIC1 (AI248720)</td>
<td>endoplasmic reticulum-golgi intermediate compartment 1(ERGIC)</td>
<td>-2.590</td>
<td>2.26E-02</td>
</tr>
<tr>
<td>CALCOCO1</td>
<td>sarcoma antigen nysar3</td>
<td>2.564</td>
<td>3.27E-03</td>
</tr>
<tr>
<td>NRG2</td>
<td>neuregulin 2</td>
<td>2.546</td>
<td>4.38E-02</td>
</tr>
</tbody>
</table>

Table 2.8: TMLE using correlation cut-off of $\rho_c = 0.75$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

<table>
<thead>
<tr>
<th>GeneID</th>
<th>Description/Function</th>
<th>TMLE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GALNT14 (AA165698)</td>
<td>UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 14</td>
<td>6.455</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>AIP</td>
<td>aryl hydrocarbon receptor interacting protein</td>
<td>5.906</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>LRTM1</td>
<td>leucine-rich repeats and transmembrane domains</td>
<td>3.703</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>ZBTB22</td>
<td>unknown 3.23E-08</td>
<td>3.609</td>
<td>3.23E-08</td>
</tr>
<tr>
<td>FUBXO41 (AA524093)</td>
<td>F-box protein 41</td>
<td>-2.843</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>VAMP3</td>
<td>vesicle-associated membrane protein 3 (cellubrevin)</td>
<td>-2.714</td>
<td>1.58E-06</td>
</tr>
<tr>
<td>ERGIC1 (AI248720)</td>
<td>endoplasmic reticulum-golgi intermediate compartment 1(ERGIC)</td>
<td>-2.590</td>
<td>2.26E-02</td>
</tr>
<tr>
<td>SH2D3C</td>
<td>SH2 domain containing 3C</td>
<td>2.950</td>
<td>6.94E-04</td>
</tr>
<tr>
<td>CTLA4</td>
<td>cytotoxic T-lymphocyte-associated protein 4 gamma, 1</td>
<td>2.719</td>
<td>1.05E-04</td>
</tr>
</tbody>
</table>

There are 6 common genes between the two top 10 lists, of these 6 genes, 4 (GALNT4, AIP, ZBTB22, and FBXO41) have been associated with chemotherapy
resistance. Beyond these 4, there are 3 other relevant genes in Table 2.7 with correlation cut-off at 0.5 (VAMP3, CALCOCO1, and NRG2) and 3 others in Table 2.8 with correlation cut-off at 0.75 (EIF4G1, SH2D3C, and CTLA4) making it 7 out of 10 relevant genes in both lists.

2.8.4 Discussion

GALNT14 is listed first (with the highest TMLE) in both tables and there is no variation in its estimate indicating no additional genes are removed from its covariate set when reducing the correlation cut-off from 0.75 to 0.5. GALNT14 has been recently acknowledged as an informative biomarker for Apo2/TRAIL-based cancer therapy. The Apo2/TRAIL-based cancer therapy falls into the class of apoptosis activating therapies - therapies which activate or enforce programmed cell death. Apoptosis regulates cell number in normal tissues. When apoptosis is no longer active, the tissue is considered malignant. Alternatively anthracycline, a common drug used in adjuvant chemotherapy, inhibits the topoisomerase II-alpha religation reaction leading to cytotoxic cell damage and death; while the taxane class drugs (also common in adjuvant chemotherapy) inhibits cell division [25]. A major benefit of the Apo2/TRAIL ligand is that it preferentially induces apoptosis in cancer cells over normal cells [34]. A recent study, [113], has shown that GALNT14 levels determine the sensitivity of tumor cells to apoptosis induced by Apo2L/TRAIL ligand. Increased expression of GALNT14 increases tumor cell response to this ligand making it a beneficial biomarker for sensitivity to Apo2/TRAIL-based cancer therapy. Among the patients in this study, exposed to adjuvant chemotherapy, we find GALNT14 up-regulated among the “bad-responders.” Given the results of [113] this could indicate that a Apo2/TRAIL-based cancer therapy may have been more beneficial for these patients. In addition to GALNT14, our results indicate that AIP, which is also known to reduce apoptosis [111,12], is up-regulated among “bad responders” and has the second highest VIM values in both lists.

Beyond the apoptosis-related genes, we also see various indicators of drug resistance. ZBTB22 binds to Cul3 forming a complex in the Ubiquitin system and elevated Cul3 has been identified as an indicator of drug resistance [122]. The over-expression of EIF4G1 has been directly identified as an indicator of chemotherapy resistance [113]. SH2D3C interacts with BCAR and partially responsible for resistance to anti-estrogen therapy in breast cancer cells [71]. Our results indicate that all three are elevated in bad responders. In addition, CALCOCO1 has been identified as a potential target for cancer vaccines [57]. Antibodies of CTLA-4 activate anti-tumor response in breast cancer cells. - drugs targeting this mechanism are in clinical trails [51]. NRG2 interacts with the Erbb family (including the HER-2 receptor) and induces cell growth among breast cancer cells [92]. All three again are found elevated in bad responders in our analysis. Also, FBXO41 has been found to be significant and important in numerous other biomarker discovery analyses, including ours, as an indicator of good
Another interesting, though confusing result is the elevated expression of VAMP3 among “good responders.” Past research has identified VAMP3 as an indicator of drug resistance [70]. It’s possible that the specific chemotherapy treatment chosen was correct for patients with elevated VAMP3. Specifics are unknown.

2.9 Conclusions

Variable Importance results vary widely leading to long lists and confusion, which list to use? In this chapter we propose using targeted maximum likelihood estimation as a standard method for biomarker discovery. In simulation it has proven resilient to increases in correlation, controlling type I error. It also provides an interpretable and meaningful measure of importance, which given an appropriate study design is interpretable as a causal effect. In comparison, common univariate linear regression is highly susceptible to increases in type I due to increased correlation. And though lasso/lars provides some improvement, using targeted maximum likelihood estimation to update its estimate increases the accuracy in importance measure and rank and provides the correct asymptotic inference.

By targeting the causal effect, the measures obtained by TMLE are less sensitive to changes in the covariate distribution and therefore more reproducible in any population given it has the same conditional distribution of $Y|W$. For instance, this allows the TMLE to be generalizable across microarray platforms that may have different noise levels. This reproducibility is essential for any standardized method, increasing confidence in diagnostic and treatment decisions based on these measures. In other words, if the causal effect between gene A and the response is correctly estimated in a population, it will be applicable to other populations. If instead we attribute the effect to gene B which is highly correlated to the causal gene A in the first population the correlation between gene B and gene A is not necessarily consistent in the other populations making the measure effect inapplicable in those populations. For instance if people in the second population have a cold, and gene B is related to immune response, it’s levels may be much higher and no longer correlated in the same degree with the level of gene A; making inferences on the disease state from the level of gene B erroneous.

The TMLE is an interpretable measure with interpretable inference. In the analysis of the Golub 1999 AML/ALL data set, linear regression results in a list of 681 genes. Among those genes, there are ones that are biologically relevant and possibly causally related, however determining the relevant from the irrelevant is an impossible task given regression alone. Random Forest results are even more ambiguous; while they may have a high importance value, the directionality and the meaning of the value is unclear. Comparing among the top 10 for each method, the TMLE results in more relevant genes which are related to AML./ALL.

The resulting top 10 tables from the Breast Cancer analysis are even more promis-
ing. Not only do we identify genes biologically related to chemotherapy resistance, we also identify genes which indicate a possible mechanism of treatment for “poor responders” based on up-to-date biological information. The relevance of the gene list supports the use of the TMLE not only for biomarker discovery but also as a pre-screening method for prediction.

Applying a correlation cut-off in practice reduces the bias in the TMLE estimate due to potential ETA violations. However, the difference between the lists for the TMLE with correlation cut-off 0.5 and 0.75 affirm the need for a method which identifies the proper cut-off for a given gene. Having too low of a cut-off neglects controlling for the appropriate genes to achieve an estimate of the causal effect, decreasing its reproducibility across populations. Having to high a cut-off leads to ETA violations which increase bias in our importance estimate. In [6], they propose a method which analytically choses the cut-off for each variable given a binary A, tailoring the controlling variable set for each A. This method has been recently extended for a general (i.e. continuous) A and will be implemented in the future [6].

Variable Importance estimated using targeted maximum likelihood estimation is a robust, locally efficient, and interpretable measure of importance with formal inference. It is simple to implement and understand. It’s accuracy, reproducibility, practically, and flexibility make it an ideal standardized method for biomarker discovery.
Chapter 3

TMLE for repeated measures

3.1 Introduction

Longitudinal data analysis, or more generally repeated measures analysis, has become increasingly popular in epidemiological and medical studies. Often the main goal of these studies is to determine the effect, or importance, of a particular variable on the outcome over time, for instance the effect of a drug on disease prognosis over the course of a clinical trial. In most cases, the repeated measures are observations on subjects at multiple time points or under multiple conditions. More recently, repeated measures analysis has been applied in computational biology, where the experimental unit is now a gene or protein that is observed over time, condition, or even species. Similarly, in these analyses the goal is to determine the importance of biological features (i.e. variables) with respect to the observed repeated measures outcome. This chapter presents a new tool to estimate variable importance for a repeated measures outcome based on targeted maximum likelihood methodology under a flexible semiparametric model.

In this chapter, a semiparametric repeated measures regression model is presented in which the parametric component models the effect of a specific variable of interest and any effect modification by other covariates. The targeted maximum likelihood estimator is developed for the effect parameters of this model using targeted maximum likelihood methodology as presented in and introduced in section. Prior applications of targeted maximum likelihood have shown great promise and applicability in the epidemiological and medical fields, in particular, for biomarker discovery (see chapter and ). The TMLE presented here builds upon previous variable importance methodology, adapting it for repeated measures data and incorporating updates on the methodology to increase efficiency and computational speed.
3.2 Repeated measures design

As indicated above, in repeated measures experimental designs multiple observations are recorded for each subject over a set of conditions and/or time (e.g. longitudinal). Though this experimental design is attractive in that it reduces the variance among observations and can increase the power of the analysis, statistical methods, such as regression, must account for the correlation among the observations on a single subject. Ignoring this dependence can lead to biased standard error estimates for regression parameters ([116] among others). A popular method to account for the correlation among the observations in parametric regression models is generalized estimating equations (GEE). GEE methods were introduced in 1986 by Zeger and Liang [60] and are an extension of generalized linear regression using a quasi-likelihood approach, which weights the residuals according to the correlation structure of the observations on each subject. More flexible semiparametric extensions of the GEE method, such as generalized partially linear models [121, 86, 33] model covariate effects non-parametrically, but require complicated estimation methods to fit both the parametric and nonparametric portions of the model. These methods can produce inconsistent and/or inefficient estimates of the model parameters [62, 59].

3.2.1 Data

Given a repeated measures experiment taken over times \( t = 1, \ldots, T \), the observed data is defined as \( O = (W^*, Y) \sim P_0 \), where \( P_0 \) is the true data generating distribution. Here, \( W^* \) is a vector of \( p \) variables, and \( Y \) is the outcome vector of \( T \) repeated measures taken over time on a subject, where \( Y_t \) represents outcome \( Y \) at a specific time point \( t \) for a subject. The TMLE is defined for a particular variable of interest \( A = W^*_j \), controlling for confounders \( W = W^*_{-j} \).

3.2.2 Generalized estimating equations

One of the most common approaches for modeling repeated measures data is generalized estimating equation methodology. Introduced by Liang and Zeger in 1986 [60], generalized estimating equations uses a quasi-likelihood approach, which weights the residuals in a generalized regression score function according to a working correlation matrix. Specifically, GEE estimates of the parameter \( \beta \) for a Gaussian model are the solution to

\[
\sum_{i=1}^{n} (D^{(i)})'(V^{(i)})^{-1}(Y^{(i)} - \bar{Q}(W^*(i) | \beta)) = 0
\]

where, for subject \( i \) in \( i = 1 \ldots n \), \( Y^{(i)} \) is a vector of observations over time \( t = 1, \ldots, T \), with \( T \) by \( T \) covariance matrix, \( V^{(i)} \). Here \( \bar{Q}(W^*(i) | \beta) = E[Y^{(i)}|W^*(i)] = \beta'W^*(i) \) is
the vector of fitted values for subject i, and \( D^{(i)} = \left[ \frac{dQ(W^{*}(i)|\beta)}{d\beta} \right] \).

The parameter estimates are obtained using iteratively reweighted least squares estimation. More robust estimates are obtained by iterating this with the re-estimation of the covariance parameters in \( V^{(i)} \) as a function of \( \beta \). This robust method is applied in R library `geepack` [119]. Standard GEE regression parameter estimates remain consistent given an incorrect correlation structure [60].

The GEE approach does not require the specification of the joint distribution of the observations over time for a given subject, only the marginal distribution for each time point and a working correlation matrix. Assuming independence among the subjects and a correctly specified model \( \beta'W \), parameter estimates \( \hat{\beta}_n \) are consistent and given true parameter \( \beta_0 \),

\[
n^{1/2}(\hat{\beta}_n - \beta_0) \sim MVN(0, \Sigma_{gee})
\]

such that given \( U^{(i)} = (D^{(i)})'V^{(i)}^{-1}D^{(i)} \) and \( R^{(i)} = Y^{(i)} - Q(W^{*}(i)|\beta) \),

\[
\Sigma_{gee} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (U^{(i)})^{-1} ((D^{(i)})'V^{(i)}^{-1}R^{(i)}(R^{(i)})'(V^{(i)})^{-1}D^{(i)})^{-1} (U^{(i)})^{-1}
\]

This is referred to as the sandwich variance estimator [42].

In this chapter, the R implementation of GEE in library `geepack`, function `geeglm` [119] is used. In simulation GEE is allowed to update the correlation parameters. However for computation ease in the application in section 3.7, a fixed correlation matrix estimate is provided based on the residuals of an initial GEE estimate under independent correlation structure [42]. The semiparametric model for the TMLE is a more non-parametric analogue of the standard GEE repeated measures regression model, and the targeted maximum likelihood updates is easily implemented using standard GEE software.

### 3.3 Semiparametric variable importance

The variable importance for a repeated measures design is defined for a particular variable \( A = W_j^* \) and time, \( t \), controlling for confounders \( W = W_{-j}^* \) as

\[
E[Y_t|A = a, W] - E[Y_t|A = 0, W] = m_t(a, V|\beta_t)
\]

The model \( m_t(A, V|\beta_t) \) is referred to as a semiparametric regression model for the effect of \( A \) on \( Y_t \). This can also be clearly represented in a more standard semiparametric representation of the mean outcome process \( E[Y|A, W] \), where

\[
E[Y|A = a, W] = m(A = a, V|\beta) + \theta(W)
\]

such that \( m(A = 0, V|\beta) = 0 \) for all \( \beta \) and \( V \), and \( \theta(W) \) is unspecified. When \( m(A = 0, V|\beta) \) is linear, this model is often referred to as a partially linear model.
The marginal variable importance of a specific \( A = W_j^* \) controlling for confounders \( W = W_{-j}^* \) can be defined generally as follows for a particular time \( t \).

\[
\mu_t(a) = E[m_t(a, V|\beta_t)]
\]
or this can be represented in vector form for all \( t \)

\[
\mu(a) = E[m(A, V|\beta)]
\]

for a user supplied model \( m \), which models the effect

\[
m(A = a, V|\beta) = E[Y|A = a, W] - E[Y|A = 0, W]
\]

under the constraint \( m(A = 0, V|\beta) = 0 \) for all \( \beta \) and \( V \).

In model formulations for a repeated measures outcome, effect modification by time is very natural. Time can be treated as a continuous variable \( t \), like \( V \) above, or treated as a set of indicator variables. The time indicator is defined as \( t_*^t = I\{t^* = t\} \), where \( t = 1, \ldots, T \). The model can then be written as

\[
m(A, V|\beta) = A(\beta t_1^*) + \ldots + A(\beta t_T^*)
\]

\[
E[Y|A, W] = A(\beta t_1^* + \ldots + \beta t_T^*) + \theta(W) \text{ where } \theta(W) \text{ is unspecified}
\]

Multiple levels of effect modification as also possible, and relevant in repeated measures where you might consider the effect of \( A \) at time \( t \) is modified by the level of variable \( V \)

\[
m(A, V|\beta) = At_1 \beta_1 + AtV \beta_2
\]

\[
E[Y|A, W] = A(\beta_1 t + \beta_2 tV) + \theta(W) \text{ where } \theta(W) \text{ is unspecified}
\]

### 3.4 Targeted maximum likelihood estimation

Given a user defined model, \( m(A, V|\beta) \), with parameter \( \beta \), the TMLE is developed using targeted maximum likelihood methodology to estimate of parameter \( \beta \). The initial regression estimate is defined as \( \hat{Q}(A, W) = E[Y|A, W] \), of a form respecting \( m(0, V|\beta) = 0 \), and the "treatment mechanism" or confounding mechanism is defined as \( g(W) = E[A|W] \).

From targeted maximum likelihood theory, it can be shown that this estimate is asymptotically consistent and linear given that either \( \hat{Q}(A, W) \) or \( g(W) \) is correctly specified, making the estimate doubly robust \cite{107}. The TMLE is also efficient when both \( \hat{Q}(A, W) \) and \( g(W) \) are correctly specified \cite{107}, while it can easily be super-efficient if \( \hat{Q}(A, W) \) is correctly specified, and \( g(W) \) is misspecified by not incorporating all \( W \) \cite{103}. The double robust nature of the TMLE makes the methodology ideal for use in randomized trials when the treatment mechanism \( (E[A|W]) \) is known. The derivation of the targeted maximum likelihood update is outlined below.
3.4.1 Derivation of TMLE update

The TMLE method updates an initial density estimate \( p^0(Y|A,W) \) in the direction which targets the parameter of interest using standard MLE and a “clever covariate” defined such that the TMLE solves the efficient score equation. In the case of repeated measures, the initial density is defined as the normal density \( f_N ) \) such that

\[
p^0(Y|A,W) = f_{\bar{Q}^0,\Sigma}(Y|A,W)
\]

where \( Y \) is an 1 by \( T \) vector and \( \bar{Q}^0(A,W) = E[Y|A,W] \). Here, \( \Sigma(A,W) \), is defined as a \( T \) by \( T \) covariance matrix corresponding to the covariance among the \( t = 1 \ldots T \) observations for a single subject.

The form of \( \bar{Q}^0(A,W) \) is as follows, \( \bar{Q}^0(A,W) = m(A,V|\beta^0) + \theta^0(W) \), where the model \( m(A,V|\beta^0) \) is defined given the constraint \( m(A=0,V|\beta^0) = 0 \) for all \( \beta^0 \) and \( V \), and \( \theta^0(W) = \bar{Q}^0(A=0,W) \). The update to the initial density is defined as its hardest submodel in terms of update parameter vector \( \epsilon \) as follows

\[
p(\epsilon)(Y|A,W) = f_{\bar{Q}(\epsilon),\Sigma}(Y|A,W)
\]

where \( \bar{Q}(\epsilon)(A,W) = m(A,V|\beta(\epsilon)) + \theta(\epsilon)(W) \) in which \( \beta(\epsilon) = \beta^0 + \epsilon \), and \( \theta(\epsilon)(W) = \theta^0(W) + \epsilon r(W) \).

The form of \( r(W) \) is defined such that the score of \( p(\epsilon)(Y|A,W) \) at \( \epsilon = 0 \) is equivalent to the efficient score equation for the parameter \( \beta \) in \( \mu(a) = E[m(A,V|\beta)] \). The efficient score equation is presented below and derived in appendix C.1

\[
D_{h_{opt},Q,g}(O) = h_{opt}(A,W)(Y - m(A,V|\beta) - \bar{Q}(0,W))
\]

with

\[
h_{opt} = \Sigma^{-1} \left( \frac{d}{d\beta} m(A,V|\beta) - E \left[ \Sigma^{-1} \bigg| W \right]^{-1} E \left[ \Sigma^{-1} \frac{d}{d\beta_0} m(A,V|\beta) \bigg| W \right] \right)
\]

where \( \Sigma^{-1} \) is shorthand for \( \Sigma^{-1}(A,W) \).

This is the multivariate extension of the efficient score equation for univariate \( Y \) presented in [102], [96], and in chapter 2.

It follows that the correct form of \( \theta(W) \) is

\[
r(W) = E \left[ \Sigma(A,W)^{-1} \bigg| W \right] E \left[ \Sigma(A,W)^{-1} \frac{d}{d\beta} m(A,V|\beta) \bigg| W \right]
\]

The expectations can be approximated by discretizing \( A \) and calculating

\[
E \left[ \Sigma(A,W)^{-1} \bigg| W \right] = \sum_{a \in A} \Sigma(a,W)^{-1} p(A = a|W)
\]
and

\[ E \left[ \Sigma(A, W)^{-1} \frac{d}{d\beta} m(A, V|\beta) \bigg| W \right] = \sum_{a \in A} \Sigma(a, W)^{-1} \frac{d}{d\beta} m(A = a, W|\beta)p(A = a|W) \]

Using standard MLE, one can solve for \( \epsilon \) and calculate the updated regression estimate \( \bar{Q}^1(A, W) = m(A, V|\beta(\epsilon)) + \theta^0(W) + \epsilon r(W) \). The procedure is iterated, and at convergence (i.e. \( \epsilon = 0 \)), the final regression estimate is the solution to the robust estimating equation corresponding to the efficient score equation for observed data \( O = \{O^{(i)} : i = 1 \ldots n\} \), for \( n \) subjects

\[ \frac{1}{n} \sum_{i=1}^{n} [D_{h_{opt},Q^*_n,g_n}(O^{(i)}|\beta^*_n)] = 0 \]

such that \( Q^*_n \) and \( \beta^*_n \) are the converged estimates of \( Q \), and \( \beta \) for the observed data. The TMLE solution therefore inherits the double robust properties of the solution to the efficient score equation and allows us to use the efficient score equation to estimate the correct covariance and inference for the parameter of interest (see section 2.3.1).

**Linear case**

Given a linear model for \( m(A, V|\beta) \), the update can be written as

\[ \bar{Q}^1(A, W) = \bar{Q}^0(A, W) + \epsilon H^*(A, W) \] where \( H^*(A, W) \) is referred to as the “clever covariate” and defined as follows

\[ H^*(A, W) = \frac{d}{d\beta} m(A, V|\beta) - E \left[ \Sigma(A, W)^{-1} \bigg| W \right]^{-1} E \left[ \Sigma(A, W)^{-1} \frac{d}{d\beta} m(A, V|\beta) \bigg| W \right] \]

In the linear case, this update can be achieved using standard software by regressing \( Y \) onto the covariate \( H^*(A, W) \), setting \( \bar{Q}^0(A, W) \) as an offset. If \( f_N \) is defined such that \( \Sigma(A, W) = \Sigma(W) \), the function \( h_{opt} \) will simplify to

\[ h_{opt}^* = \Sigma(W)^{-1} \left( \frac{d}{d\beta} m(A, V|\beta) - E \left[ \frac{d}{d\beta} m(A, V|\beta) \bigg| W \right] \right) \]

and the “clever covariate” simplifies to

\[ H^*(A, W) = \left( \frac{d}{d\beta} m(A, V|\beta) - E \left[ \frac{d}{d\beta} m(A, V|\beta) \bigg| W \right] \right) \]

Note that if the true covariance is a function of \( A \), estimation using the simplified covariate form will lose efficiency but will still remain double robust.

Given the simplified form of the “clever covariate” with linear model for \( m(A, V|\beta) \), the TMLE estimate is a closed form solution and can be calculated without iteration. The linear semiparametric form allows us to introduce time and/or any additional
covariate as effect modifiers of the importance of A in a straightforward interpretable fashion. Consider the following possible model, which incorporates effect modification of time indicator variable $t^*_t = I\{t^* = t\}$.

$$m(A, V|\beta) = A(\beta' t^*_1) + \ldots + A(\beta' t^*_T)$$

When $m(.)$ becomes large it is beneficial to update the coefficient terms sequentially until convergence (i.e. targeting one at a time) instead of completing an update of the full coefficient vector in one step. Updating the model sequentially in this fashion has been shown to improve the overall stability of the updated estimates (see section 3.5.1 for details).

**Inference**

Since the TMLE solution solves the double robust estimating function implied by the efficient score equation \cite{107}, one can use the influence curve corresponding with this double robust estimating function to provide an estimate of the covariance for TMLE estimated $\beta_n$. See section 1.4.3 for details. For this estimator the efficient influence curve, $IC(O)$, is a $T$ by $p$ matrix for a parameter vector $\beta$ of length $p$. Given correctly specified estimates for $\bar{Q}(A, W)$ and $g(W)$, the covariance for parameter vector estimate $\beta_n$ is asymptotically equivalent to the covariance of $IC(O)$ regardless of the form of $\Sigma(A, W)$. The covariance can also be estimated by bootstrap estimates of $\beta_n$, but this would require extra computational time and any sampling would need to respect the repeated measures design. If $E[A | W]$ is estimated consistently, then the variance estimates based on the influence curve are consistent or asymptotically conservative. Using the estimated covariance matrix, hypothesis tests can be performed for a single parameter $\beta_n(j)$, where $j = 1, \ldots, p$, under the null hypothesis $H_0 : \beta_n(j) = 0$ using a standard test statistic to obtain p-values, with estimated variance $\Sigma_n(j, j)$.

$$T_n(j) = \frac{\sqrt{n}\beta_n(j)}{\sqrt{\Sigma_n(j, j)}} \sim N(0, 1)$$

Likewise the hypothesis $H_0 : c'\beta_n = 0$ can also be tested using a standard Wald test, where the covariance of $c'\beta_n$ is $c'\Sigma_n c$. This allows one to obtain inference for $\mu(a)$ directly, when $m$ is linear. In practice the parameter of interest may be redefined as the effect at a specific value of effect modifier $W$, or time $t$, instead of the mean effect as implied by the definition in section 1.2.

**3.4.2 TMLE implementation steps**

Below the basic procedure for estimating the TMLE is outlined given a fixed correlation matrix.
Components

There are three initial components necessary for applying targeted maximum likelihood methodology to estimate TMLE for repeated measures.

1. Model \( m(A, V | \beta) \) satisfying \( m(A = 0, V | \beta) = 0 \) for any \( \beta \) and \( V \subset W \)

2. An estimate for \( \bar{g}(W) = E[A|W] \): Estimating this data-adaptively is recommended.

3. An initial estimate for \( \bar{Q}(A, W) = E[Y | A, W] \), \( \bar{Q}_0(A, W) \), containing valid model \( m(A, W | \beta) \): This provides an initial estimate for the parameter \( \beta, \beta_0 \), and must be defined such that \( Y|A, W \sim Normal(\bar{Q}(A, W), \Sigma(W)) \), with an empirically estimated correlation.

The initial regression estimate of proper form may be obtained from semiparametric methods such as those of [121, 33, 115] among others, or by using methods such as DSA [88] which allow the user to fix a portion of the model. However, a more flexible approach is adopted which allows one to use a wider range of data-adaptive software, providing that any internal cross-validation respects the repeated measures nature of the data. An initial regression estimate with proper semiparametric form is obtained by updating a data-adaptively estimate for \( \bar{Q}(A, W) \) of general model form using data-adaptive machine learning algorithms such as super learner [104] or DSA [88]. It is outlined as follows

- Obtain an initial regression estimate for \( \bar{Q}(A, W) \) of general model form using data-adaptive machine learning algorithms such as super learner [104] or DSA [88]. This is valid for all \( A \in \{W^*\} \).
- Solve for \( \bar{Q}(A = 0, W) \)
- Using standard GEE regression, solve for the initial estimate,

\[
\bar{Q}^0(A, W) = m(A, W | \beta^0) + \alpha \bar{Q}(A = 0, W)
\]

by specifying model \( m(.) \) and treating \( \bar{Q}(A = 0, W) \) as a covariate.

This provides us with initial parameter estimates for \( \beta \) as well as an initial density estimate of the correct form. This is an update from the original method outlined in [96], which improves computational efficiency by only requiring a single data-adaptive estimate for \( \bar{Q}(A, W) \) of general model form for all \( A \).

Using data-adaptive algorithms such as super learner [104] and DSA [88] will provide a better estimate for the initial \( \bar{Q}(A, W) \), which improves the performance of the TMLE. It is noted that these methods do not account for the correlation among the repeated measures and only require that any cross-validation within the algorithm
respects the repeated measure structure of the data. The asymptotic covariance matrix for the TMLE of $\beta$ is based on the update of a GEE quasi-likelihood, which allows for the specification of a more accurate covariance structure (i.e., $\Sigma(A,W)$ in the definition of the efficient score equation). In this manner the targeted MLE can still fully utilize the covariance structure of the repeated measures and potentially be asymptotically linear with efficient influence curve identified by the true $\Sigma(A,W)$ without a risk of being inconsistent. The overall consistency of the estimator relies on correct specification of either the estimate of $E[A|W]$ or of $E[Y|A,W]$. This is addressed further in section 2.4.

**Step-by-step**

Given the three components, TMLE is applied using the following steps. Sample R code for a simple example is provided in appendix E.

1. Estimate the “clever covariate” which will allow us to update the initial regression in a direction which targets the parameter of interest. For a linear model the clever covariate is:

   $$H^*(A,W) = \frac{d}{d\beta} m(A,V|\beta) - E \left[ \frac{d}{d\beta} m(A,V|\beta)|W \right]$$

2. Compute the fitted values for your initial estimate, $\bar{Q}_0^n(A,W)$

3. Project $Y$ onto $H^*(A,W)$ with $offset = \bar{Q}_0^n(A,W)$, define the resulting coefficient as $\epsilon$. This is done using generalized estimating equations with fixed correlation (geeglm in R [119]) by fitting the model $Y \sim H^*(A,W) + offset$. Note there is no intercept in the model, only the offset value.

4. Update initial estimate $\beta_n = \beta_0^n + \epsilon$ and overall density $Q_n(A,W) = \bar{Q}_n(A,W) + \epsilon H^*(A,W)$. These are now your single-step targeted estimates. Since this is a simple linear model, the single step solution is the final solution.

5. Obtain standard error and inference for $\beta_n$ using the influence curve as outlined in section 1.4.3

Given that the number of possible covariates for both $\bar{Q}_0^n(A,W)$ and $\bar{g}(W)$ can be quite large and include main effects, interactions among the covariate set $W$, and interactions with time, reducing the set of possible covariates using basic univariate linear regression is recommended. As in the previous implementation (see chapter 2 and [96, 8]), the instability in the estimate from ETA (Experimental Treatment Assumption) violations can be reduced by restricting the covariate set using a $\delta$ cut-off based on some measure of dependence between $A$ and $W$. This removes variables in $W$ which may be highly correlated with $A$ [6].
Alternative methods of getting your initial $\bar{Q}$ estimate

In the procedure outlined above, the initial density estimate for TMLE is a GEE model with covariate $\bar{Q}(0, W)$, which is obtained from a data-adaptive fit of $\bar{Q}(A, W)$ using a data-adaptive prediction algorithm such as DSA [88] or super learner [104]. Both of these methods respect the repeated measures nature of the data by allowing the user to specify a subject ID to use in sampling and cross-validation, but apply an independent correlation structure for the sake of estimation. If the true correlation structure is not independent, there might be a finite sample loss in efficiency by using this structure. However, by using GEE model with a correlation matrix closer to the truth to carry out the targeted MLE update, this loss is asymptotically negligible. Nevertheless, an alternative initial estimate is proposed that potentially already takes into account correlation structure between the repeated measures. Given an outcome of repeated measures, one can transform the observations prior to implementing DSA or super learner, and then transform back the predicted values using an estimate of their covariance matrix. This is outlined here.

For a fixed working covariance matrix $\Sigma(A, W)$, the quasi-likelihood has the equivalent loss function

$$L(O) = (Y - \bar{Q}(A, W))\Sigma(A, W)^{-1}(Y - \bar{Q}(A, W))^\prime$$

This can be rewritten as the euclidean norm

$$\left\|\Sigma(A, W)^{-\frac{1}{2}}(Y - \bar{Q}(A, W))\right\|$$

which can be restructured in the equivalent form

$$\left\|\Sigma(A, W)^{-\frac{1}{2}}Y - \Sigma(A, W)^{-\frac{1}{2}}\bar{Q}(A, W)\right\|$$

Therefore if $Y$ is transformed into $Y_r = \Sigma(A, W)^{-\frac{1}{2}}Y$, then $E[Y_r|A, W] = \bar{Q}_r(A, W)$ and the non-transformed predicted values can be regained as follows.

$$\bar{Q}(A, W) = \Sigma(A, W)^{\frac{1}{2}}\bar{Q}_r(A, W)$$

This method can be applied to any machine learning algorithm as long as any sampling or cross-validation respects the repeated measures structure.

3.5 Additional methods to improve TMLE

3.5.1 Sequential update

Targeted maximum likelihood methodology was initially developed around a low dimensional update of an initial density estimate. For the semiparametric TMLE
presented here, the dimension of the update increases with the size of the model. This is especially relevant for repeated measures TMLE which can easily have high dimensional model for even a one dimensional A. In an effort to avoid any potential instability in the high dimensional update we propose using a sequential targeted update which updates each component of $\epsilon$ sequentially iterating until convergence.

The results of a small simulation show that the sequential update is as good or better than the standard targeted update.

### 3.5.2 Simulation

A set of 20 possible covariates, $W$, is simulated from a multivariate normal with random mean between 0 and 50, a constant variance $\rho$, and zero correlation. The variable of interest, $A$, is also simulated from a normal distribution. Three different simulation set ups are used.

1. Uncorrelated: Variables in $W$ and variable of interest, $A$, are uncorrelated

2. Correlated $W$: Variables in $W$ are correlated with $\rho = 0.8$ and $A$ is still independent of all variables in $W$

3. $A$ dependent on $W$: Variables in $W$ are correlated with $\rho = 0.8$ and $A$ is still a linear function of two variables from $W$ with mean zero variance 0.1 error

We model the outcome, $Y$, as a linear function of $A$: $W$ interactions using 12 different variables from $W$ with normal mean zero variance one error. All interaction terms have coefficients equal to four. The average mean square error for the three scenarios are compared based on 100 simulations and 500 observations.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Standard Update</th>
<th>Iterative Update</th>
<th>Percent Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrelated</td>
<td>0.10950</td>
<td>0.10766</td>
<td>1.7 %</td>
</tr>
<tr>
<td>Correlated $W$</td>
<td>0.01001</td>
<td>0.00917</td>
<td>8.4 %</td>
</tr>
<tr>
<td>$A$ dependent on $W$</td>
<td>0.20454</td>
<td>0.20052</td>
<td>2.0 %</td>
</tr>
</tbody>
</table>

### 3.6 Simulation study

In simulation, the robust features of the TMLE method are demonstrated under a known data generating distribution with model misspecification, confounding, and
varying levels of overall noise. Results are compared with those of standard GEE applied using `geeglm` R function from library `geepack` [119]. The `geeglm` function is allowed to update the correlation structure which is simulated and modeled correctly as AR(1). The variable of interest is univariate so sequential updating is not used for the TMLE estimate, however pre-weighting of the initial density estimate to improve overall efficiency is applied (See appendix D).

### 3.6.1 Data

Simulated data is drawn for n=50, n=100, and n=500 subjects with 4 replicates (e.g. time points) from a linear model $Y \sim 1 - 2A + 0.5W + \gamma$, where $Y$ is a vector $\{Y_t : t = 1, \ldots, 4\}$ and the error, $\gamma$, is normal with AR(1) covariance structure within replicates for each subject given a true lag-1 correlation of 0.667 and standard deviation $\sigma_Y = 1, 10$. Variable $A$ is simulated both independent of $W$, and as a function of $W$ (e.g. under confounding), where $A \sim N(2, 1)$ or $A \sim N(W + 2, 1)$ respectively, with $W \sim N(3, 1)$.

For each case, the importance parameter for $A$ is measured using both basic GEE methods and TMLE as described in section 2.3, under both correct and incorrect model specification, $Y \sim A + W$ and $Y \sim A$ respectively. Note that in all cases the treatment mechanism ($E[A|W]$) is correctly modeled.
### 3.6.2 Results

Table 3.2: Simulation results comparing GEE and TMLE-RM with n=50, 100, 500 and $\sigma_y = 1, 10$: provided are the mean value ($\mu_\beta$) and standard error ($SE_\beta$) for $\beta$ estimates over the 500 iterations, the mean standard error estimate ($\mu_SE$) for the influence curve based standard error estimate from 500 iterations, and the percent of time the true $\beta$ value is included in the 95% confidence interval ($CI_{95\%}$) based on the standard error estimate over the 500 iterations.

<table>
<thead>
<tr>
<th>n=50, $\sigma_y = 1$</th>
<th>TMLE-RM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q Confounding $\mu_\beta$ $SE_\beta$ $\mu_SE$ $CI_{95%}$ $\mu_\beta$ $SE_\beta$ $\mu_SE$ $CI_{95%}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true N -1.997 0.080 0.075 0.944 -1.997 0.080 0.075 0.942</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true Y -1.997 0.080 0.076 0.942 -1.997 0.080 0.075 0.942</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong N -2.005 0.079 0.083 0.962 -2.005 0.079 0.083 0.956</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong Y -2.000 0.080 0.078 0.950 -1.735 0.055 0.057 0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=50, $\sigma_y = 10$</td>
<td>TMLE-RM</td>
<td>GEE</td>
</tr>
<tr>
<td>true N -1.990 0.251 0.238 0.944 -1.990 0.251 0.238 0.942</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true Y -1.990 0.251 0.241 0.942 -1.990 0.251 0.238 0.942</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong N -1.999 0.250 0.241 0.944 -1.999 0.250 0.241 0.940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong Y -1.993 0.253 0.242 0.944 -1.734 0.173 0.169 0.622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=100, $\sigma_y = 1$</td>
<td>TMLE-RM</td>
<td>GEE</td>
</tr>
<tr>
<td>true N -1.998 0.048 0.051 0.956 -1.998 0.048 0.047 0.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true Y -1.998 0.048 0.052 0.956 -1.998 0.048 0.047 0.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong N -2.003 0.049 0.057 0.970 -2.003 0.049 0.052 0.952</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong Y -1.999 0.048 0.053 0.960 -1.760 0.034 0.036 0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=100, $\sigma_y = 10$</td>
<td>TMLE-RM</td>
<td>GEE</td>
</tr>
<tr>
<td>true N -1.993 0.153 0.162 0.956 -1.993 0.153 0.148 0.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true Y -1.993 0.153 0.164 0.956 -1.993 0.153 0.148 0.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong N -1.997 0.154 0.164 0.960 -1.997 0.154 0.150 0.938</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong Y -1.994 0.153 0.164 0.960 -1.756 0.109 0.109 0.364</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=500, $\sigma_y = 1$</td>
<td>TMLE-RM</td>
<td>GEE</td>
</tr>
<tr>
<td>true N -2.000 0.023 0.023 0.936 -2.000 0.023 0.023 0.934</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true Y -2.000 0.023 0.023 0.930 -2.000 0.023 0.023 0.934</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong N -1.990 0.024 0.026 0.962 -1.990 0.024 0.025 0.960</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong Y -1.995 0.023 0.023 0.946 -1.746 0.016 0.017 0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=500, $\sigma_y = 10$</td>
<td>TMLE-RM</td>
<td>GEE</td>
</tr>
<tr>
<td>true N -2.001 0.074 0.074 0.936 -2.001 0.074 0.072 0.934</td>
<td></td>
<td></td>
</tr>
<tr>
<td>true Y -2.001 0.074 0.072 0.930 -2.001 0.074 0.072 0.934</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong N -1.990 0.074 0.074 0.944 -1.990 0.074 0.073 0.942</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wrong Y -1.996 0.073 0.072 0.936 -1.747 0.050 0.050 0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.7 Application - Transcription factor activity

The biological pathways and mechanisms of an organism are regulated by a network of transcription factors, which control a gene’s expression by binding to specific regulatory motifs upstream of the gene’s coding sequence. Activity of a transcription
factor (TF) is reflected in the gene expression profile, and given a TF to gene mapping, this information can be used to determine which transcription factors are active under various stimuli or gene conditions.

The simple approach introduced by [17] sets the expression profile as an outcome and regresses it onto a set of covariates, representing motif or TF to gene association measures. The association measures are generally determined from the presence of regulatory motifs upstream of the gene’s coding sequence. Often, the association measure is an affinity or matching score that is determined experimentally and/or using algorithms to detect motifs and assign probabilities to each gene-TF pairing [38, 114, 23]. For this analysis a simple binary TF-gene mapping obtained from [67] is used. This mapping is based on a combination of experimental ChIP-Chip data and algorithm findings. In this covariate matrix a value of one indicates that the TF has been shown to regulate that particular gene according to the strictest conservation and binding thresholds provided by [67]. In the original analysis [17], the association measure is the number of known binding motif occurrences upstream of the gene. An alternative analysis using similar regression methods focuses on the regulatory motif importance, using the motif-gene mapping as a covariate set to score potential motifs and then relate them back to the transcription network [49, 48, 23, 64].

Using this regression approach, TMLE can be used to determine the importance of a specific transcription factor in relation to a set of gene expression profiles. In this case, the repeated measures gene expression outcome is a time series of yeast gene expression over two cell cycles [21]. The model-based semiparametric nature of TMLE allows us to determine the importance of a TF at specific time points by specifying time indicators as potential effect modifiers of the TF. The goal is to identify the active phases of a given transcription factor during the cell cycle based on the estimated TMLE importance values.

For simplicity in the application, the binary TF-gene mapping provided by [67] is used with the simple linear model \( m_t(A, V | \beta_t) = \beta_t A t^*_t \) for \( t = 0, 10, \ldots, 160 \), where \( t^*_t = \{t^*_t = t\} \). For this model, the parameter of interest is \( \mu_t(A = 1) = \beta_t P(A_t t^*_t = 1) \). Note that for each time point, \( P(A_t t^*_t = 1) \) is equivalent. Therefore, the importance of \( A \) at time \( t^*_t \) is represented by the coefficient \( \beta_t \), and only these coefficients and their inference will be reported. Estimates for the initial \( \tilde{Q}(A, W) \) and \( barg(W) \) are obtained using DSA [88].

Data

In this analysis the outcome is the cell cycle gene expression profile for yeast from [21]. It consists of 17 time points, which is approximately two cell cycles. Data was obtained from the Yeast Cell Cycle Analysis Project website [1]. The cell cycle consists of four phases G1, S, G2, M. A brief description of each phase along with its corresponding time points is presented in Table 3.3.
Table 3.3: Description of stages of cell cycle. Note there are three major checkpoints at which cell cycle may arrest [24]

<table>
<thead>
<tr>
<th>Cell Cycle Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>Growth phase, decision to proceed through division made,</td>
</tr>
<tr>
<td></td>
<td>checkpoint: Enough nutrients present and cell health</td>
</tr>
<tr>
<td>S</td>
<td>DNA synthesis occurs</td>
</tr>
<tr>
<td>G2</td>
<td>Checkpoint: Cell is critical size and DNA synthesis and repair are complete</td>
</tr>
<tr>
<td>M</td>
<td>Mitosis occurs, checkpoint on chromosome alignment before cell division</td>
</tr>
</tbody>
</table>

The covariate set consists of 117 binary transcription factor-gene mappings provided by MacIsaac et al. 2006 [67]. Though the transcription regulatory network for yeast is not completely known, it is widely accepted that the cell cycle involves the following transcription factors: SWI4, SWI6, MBP1, MCM1, ACE2, FKH2, NDD1, and SWI5 [41]. Therefore this analysis will focus on these 8 transcription factors. Their known phase associations and reported active time points in [21] cell cycle data are shown in Table 3.4.

Table 3.4: Association of transcription factor with cell cycle phase [21]

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Cell Cycle Phase</th>
<th>Approx. Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWI4-SWI6, MBP1-SWI6</td>
<td>G1 phase,</td>
<td>0-30, 80-110</td>
</tr>
<tr>
<td></td>
<td>G1 to S transition</td>
<td></td>
</tr>
<tr>
<td>MCM1, (MCM1-ACE2) FKH2, NDD1</td>
<td>G2 phase,</td>
<td>40-70, 130-150</td>
</tr>
<tr>
<td></td>
<td>G2 to M transition</td>
<td></td>
</tr>
<tr>
<td>MCM1, SWI5, (SWI5-MCM1-FKH2-NDD1) ACE2</td>
<td>M phase,</td>
<td>70-90, 150-160,0</td>
</tr>
<tr>
<td></td>
<td>M to G1 transition</td>
<td></td>
</tr>
</tbody>
</table>

The TMLE method is applied to the 8 TFs listed above, and importance estimates are provided along with standard error derived from the influence curve. It’s important to note that though the current covariate set is binary, this method can also be applied to continuous variables and can be extended to using a score-based mapping of binding motifs such as presented in [49].

In order to improve computation speed, the yeast gene set has been reduced by removing genes with variance across time less than 0.10. This reduces the data set to 3135 genes for 17 time points. The transcription factor data set is also constrained to TFs with at least 10 related genes. TFs with less than 10 related genes are problematic for cross-validation splits used in data-adaptive algorithms. This reduces the number of potential TF confounders to 112. For this application, the initial density estimates are not weighted as discussed in section 3.6 and appendix D however in practice it is possible to apply weighting to improve the overall efficiency.
Prescreening

Confounders of variable of interest, A, must be significantly related to the outcome, Y, therefore the initial TF data matrix is screened using simple regression which should improve the performance of model selection methods [6]. All individual TF effects and all TF:time interactions are considered using univariate regression, where interactions are treated as a single main effect. The cut-off is p-value of less than or equal to 0.05 based on standard t-test. Prescreening in this fashion reduces the potential covariate set to 92 TF main effects and 481 TF:time interactions.

For each TF, separate subsequent individual screening on the covariate set was completed based on the correlation between the covariates and the TF of interest. Any covariates with correlation greater than 0.5 were removed. Such a cut-off aims to reduce bias in the final estimate by excluding variables highly correlated with the variable of interest from the possible covariate set, avoiding ETA (experimental treatment assumption) violations [6]. This cut-off is user supplied. Currently the appropriate cut-off is chosen a priori to the application of TMLE, and in practice results are reported over a range of delta values allowing the researcher to see the full compendium of results [6]. In previous studies it has been shown that TMLE methods remain stable up to correlations of 0.8 (see chapter 2 and [96]). Here, a delta of 0.5 is chosen based on knowledge from previous studies and computational constraints [96, 10].
Results and discussion

![Graph showing marginal variable importance over time for various transcription factors](image)

Figure 3.1: The TMLE importance measures over time with 95% confidence intervals for (top to bottom) MBP1, SWI4, SWI6, MCM1, FKH2, NDD1, ACE2, and SWI5

The resulting importance measures ($\beta_t$) for the 8 transcription factors are presented in Figure 3.1 for each time point (0 min - 160 min) calculated according to the equation in section 3.4.1. Error bars are included, representing the 95% confidence interval for each estimate using the standard error derived from the influence curve as outlined in section 3.4.1.
Many of the trends in Figure 3.1 coincide well with the expected temporal trends outlined in Table 3.4. MBP1 and SWI6 correspond especially well with a clear periodic trend peaking at 20 and 100 minute within the two G1 phase periods. MCM1 peaks around 70 minutes, then decreases before increasing again around 150 minutes. This approximately corresponds to decreasing during G1 phase, which is the only phase MCM1 is not active. FKH2 and NDD1 peak at 70 and 150 minutes, which corresponds well to G2 phase and G2-M transition, their more active phases.

ACE2, SWI5, and SWI4 do not correspond as well with their expected behavior. ACE2 and SWI5 have similar trends, which remain fairly constant during the first cell cycle (0-80 minutes) and then increase around 90-100 minutes, at the G1 to S transition of the second cycle. They then slightly decrease only to increase again at 150 minutes before decreasing at the end of the cycle. SWI4 only shows a slight periodic trend with no significant time points.

Inconsistencies in the behavior could be due to modeling the effects of the single TF and not the full complex. To explore this briefly the importance of the SWI4-SWI6 complex is estimated using TMLE, allowing for effect modification by time. Note that in this model, transcription factor complexes are not adjusted for, only single TFs and TF:time interactions. Results are shown in Figure 3.2.

![Figure 3.2: The TMLE importance measures over time with 95% confidence intervals for SWI4-SWI6 transcription factor composite](image)

In Figure 2, the expected periodic trend is present, with peaks during G1 phases. It also shows the confidence intervals are smaller than when the importance of SWI6 was measured individually. Additional improvements may be obtained by allowing TF complexes as covariates.

Inconsistencies in the findings may come from a number of sources including the use and accuracy of the binary TF-gene mapping for the covariate set, incomplete knowledge of the yeast cell phases, as well as not providing model selection for the working model, which includes all time interactions. The current application is also fairly simplistic, and though it does show the method has promise for these types of applications, a more extensive and comprehensive study, including a thorough study of complexes, is necessary to obtain more conclusive findings.
3.8 Model selection for $m(A, V | \beta)$

The targeted maximum likelihood method for variable importance presented in this thesis is specifically developed and implemented for a model, $m(A, V | \beta)$, which is linear in parameter $\beta$. This form is beneficial in that it retains the simple interpretation of a linear regression parameter and provides some flexibility by allowing the incorporating of effect modifiers. The simplest effect modifier form is a first order interaction, $A: V$. However, higher level polynomials such as $A^2$, $A: V^2$, etc. are also valid model terms for $m(A, V | \beta)$. A model selection method is presented here which selects among possible models for $m(A, V | \beta)$. Specifically developed using the DSA algorithm [SS], this methodology selects among possible polynomials of $A$ and $V$ using cross validation based on the targeted risk function.

Targeted maximum likelihood estimators have the added benefit of having an associated targeted likelihood and therefore a natural targeted loss and risk function. This targeted risk function can be used as a criteria in cross-validation. Here, a targeted model selection procedure is presented to select among candidate models for $m(A, V | \beta)$ using cross-validation based on the targeted risk. It is outlined in terms of the TMLE for a univariate continuous outcome, but is easily extended for a repeated measures analysis. This extension is completed by including time as a possible effect modifier and respecting the repeated measures design when splitting the data for cross-validation and calculating the final influence curve. This methodology is also applicable to the TMLE under the multiplicative semiparametric model presented in chapter 4.

3.8.1 Model selection procedure

This procedure relies on the DSA algorithm [SS]. However it can be adapted to any model selection algorithm which allows the user to specify at least one permanent model term. In this implementation, the variable of interest $A$ is orthogonalized by the fitted values of a data-adaptive fit for $\bar{g}(W) = E[A|W]$. The variable $A$ is then replaced by its orthogonalized version, $A_{ortho} = (A - \bar{g}(W))$, and the fitted values for $\bar{g}(W)$ are also included as a covariate (i.e. $\bar{Q}(A, W) = \bar{Q}(A_{ortho}, \bar{g}(W), W)$). The model selection procedure is outlined below.

1. Estimate $\bar{Q}(A, W) = \bar{Q}(A_{ortho}, \bar{g}(W), W)$ and record predicted values for $\bar{Q}(A, W)$ and $\bar{Q}(A = 0, W)$

2. Split the observed data $O$ and fitted values $\bar{Q}(A, W)$ and $\bar{Q}(A = 0, W)$ into $V = 10$ validation and training sets, where for $b = 1, \ldots, V$, the training set is defined as $\{O_{b}, \bar{Q}_{b}(A, W), \bar{Q}_{b}(A = 0, W)\}$ with corresponding validation set $\{O_{b}, \bar{Q}_{b}(A, W), \bar{Q}_{b}(A = 0, W)\}$

3. For each $b=1,\ldots,10$
(a) Use DSA \[88\] to fit the following model 
\[ Y_{\cdot b} \sim m_{\cdot b}(A, V|\beta) + Q_{\cdot b}(0, W) \]
on the \(b^{th}\) training set, setting \(Q_{\cdot b}(0, W)\) as a permanent covariate, and
allowing DSA to select among polynomials of \(A\) and \(V\) for \(m(A, V|\beta)\). The
simplest case, and the one implemented here, is to allow all \(A\) main effect
and interaction terms with a specified subset \(V \in W\). DSA will choose the
best model for each model size up to the specified maximum model size,
\(K\).

(b) Record \(K\) candidate models, \(m_{\cdot b,k}(A, V|\beta)\), of sizes \(k = 1 \ldots K\) given by
the DSA output

(c) Applied the targeted maximum likelihood update to the training set, up-
dating \(\bar{Q}_{\cdot b}(A, W)\) for each of the \(k = 1 \ldots K\) models \(m_{\cdot b,k}(A, V|\beta)\). Using
the resulting update \(\epsilon\) vector, predict the targeted maximum likelihood
update for the validation set, \(\bar{Q}_b(A, W), \bar{Q}^*_{b,k}(A, W)\), for the \(b^{th}\) training
set and \(k^{th}\) model.

(d) Calculate the risk for each \((b,k)\) combination
\[ Risk(b, k) = \sum_{i=1}^{n} (Y_{\cdot b} - \bar{Q}^*_{b,k}(A, W))^2 \]

4. Sum the risk over the 10 validation sets, and select the model size with the
smallest summed risk
\[ k_s = \arg\min_k \left( \sum_{b=1}^{10} Risk(b, k) \right) \]

5. Given selected model size \(k_s\), repeat step 3 on the full data, again allowing DSA
to choose models up to size \(K\) among the specified terms of \(A\). Choose the
model of size \(k_s\), \(m_{k_s}(A, V|\beta)\). This is final selected model.

6. Estimate the TMLE given model \(m_{k_s}(A, V|\beta)\) on full data.

3.8.2 Simulation

Data

The model selection procedure is demonstrated in a simple simulation. The true
model, \(m_{true}(A, V|\beta)\), is defined as
\[ m_{true}(A, V|\beta) = \beta'[A: W_4 \ A: W_{15}] + r(W) \]
where \(r(W) = \beta'_W[W_3 \ W_4 \ W_{15} \ W_{17}]\). In this simulation all \(\beta\) and \(\beta_W\) are equal to
4. The full covariate matrix, \(W^*\) is \(n\) by \(p\), where \(n = 500\) observations and \(p = 20\)
variables. The matrix is simulated from a multivariate normal with constant variance equal to one and means sampled randomly from the vector \((1 : 500)/60\). The variable of interest, \(A\) is set equal to \(W_i^*\) and \(W = W_{-1}^*\). This simulation is completed in R with random seed set at 2349. Simulations are run for 500 iterations. In this simulation the maximum model size allowed for \(m(A, V|\beta)\) is 10.

**Results**

The sensitivity and specificity are calculated for the model resulting from each simulation draw. The sensitivity is calculated as the number of accurately selected model components (true positives) divided by the total number of components selected (total positive). The specificity is calculated as the number of accurately discounted components (true negatives) over the total number of components not allowed into the model (total negatives). The values are averaged over the 500 simulations. All models contained the true model leading to a calculated sensitivity of 1. The average specificity is calculated as 0.9578.

To further explore the results, the percentage of models among the 500 resulting models with a given number of false positives (i.e. incorrect components in the model) are plotted versus the number of false positives (Figure 3.3). This gives an indication of the overall distribution of model sizes selected as well.
Results show that the model selection procedure works well and is able to detect the true model with a satisfactory level of accuracy.

### 3.8.3 Brief application

Model selection is a very useful tool for the targeted maximum likelihood estimation method presented in this thesis. A small application example is presented here. This example is a follow up to the transcription factor activity analysis presented in the previous section, but focuses on a single time point, t=10 (G1 phase). In this example, the above model selection procedure is applied to select among effect modifiers for the MPB1 transcription factor (i.e. MPB1:TF components).

In this example, a single time point, t=10 (G1 phase) is isolated, and potential effect modifiers of the MPB1 transcription factor are considered. The model selection
algorithm selects among interaction terms with the MPB1 variable,

\[ m_{full}(A_{MPB1}, V|\beta) = A_{MPB1}(\beta_A + \beta_VV_1 + \ldots + \beta_VV_k) \]

The algorithm selects the following model

\[ m_s(A_{MPB1}, V|\beta) = A_{MPB1}(\beta_A + \beta_1V_{SWI6} + \beta_2V_{STE12}) \]

and the final TMLE for the model parameters are presented in Table 3.5. The variable importance estimate, \( E[m(A,V|\beta)] \), is calculated using the delta method and setting the values for \( V_{SWI6} \) and \( V_{STE12} \) to their respective empirical expectations (e.g. \( P(V_{SWI6} = 1) \)). This results in a final variable importance estimate for MBP1 of 0.1763 with an associated p-value of 0.0319.

Table 3.5: Resulting TMLE for the coefficients of the selected model, \( m(A,V|\beta) = A_{MPB1}(\beta_A + \beta_1V_{SWI6} + \beta_2V_{STE12}) \)

<table>
<thead>
<tr>
<th>Component</th>
<th>( \beta )</th>
<th>( SE_{IC} )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBP1</td>
<td>0.183</td>
<td>0.067</td>
<td>0.007</td>
</tr>
<tr>
<td>MBP1:SWI6</td>
<td>-0.444</td>
<td>0.149</td>
<td>0.003</td>
</tr>
<tr>
<td>MBP1:STE12</td>
<td>0.212</td>
<td>0.143</td>
<td>0.136</td>
</tr>
</tbody>
</table>

The resulting model is a reasonable model. It is known that MBP1 does form a complex with SWI6 which is active during the G1 cell cycle phase [21] (also see Table 3.4). MBP1 and STE12 have also been selected as co-regulators in other analyses [26], and STE12 does play a role in the regulation of mating specific genes, which peak in activity during the G1 phase [73].

The application presented here is very brief and only serves as an example of how the model selection algorithm could be used in practice. This algorithm presents a powerful tool for the variable importance estimators presented in this thesis and will be used regularly in future applications.

### 3.9 Conclusions

The TMLE method is a robust and targeted method for variable importance in repeated measures analysis. The targeted maximum likelihood step in the TMLE method is easily carried out with standard GEE, which allows the user to implement it with standard readily available software. The nature of the update provides a locally efficient and double robust estimate, which remains consistent given that either the initial density estimate (\( E[Y|A,W] \)), or treatment mechanism (\( E[A|W] \)) is specified correctly. This is demonstrated in simulation, showing the consistency and
efficiency of the TMLE method under incorrect model specification and confounding. In general, TMLE performs as well or better than the standard GEE approach assuming a parametric regression model.

In this chapter, TMLE is applied to yeast cell cycle data, measuring the importance of 8 transcription factors with respect to gene expression outcome over two cell cycles. The results are promising, showing significant importance trends during the appropriate time periods. Its applicability for TF complexes is demonstrated in a follow up analysis. Future work will focus on the development of targeted model selection methods which will allow selection among TF and time effect modifiers for the TF of interest. The analysis is a simple case using a binary TF-gene mapping. However the targeted method can easily be extended for more sophisticated analyses such as binding motif discovery [48] and phylogenetic associations [87], where the TF-gene association may be a continuous measures. The model selection methodology introduced in this chapter will also be a useful tool to further analyze this data.

Targeted variable importance for repeated measures data provides a powerful new tool for biological studies interested in understanding the driving force behind a mechanism over time and/or experimental condition. This method has a wide range of applicability and will be useful in computational biology as demonstrated here, as well as epidemiology and randomized clinical trials, where the TMLE based methods have been shown to be especially powerful [8, 96].
Chapter 4

TMLE for multiplicative model and relative risk

4.1 Introduction

The ability to accurately estimate the conditional relative risk is important to many fields of research. When the outcome is binary, analysts will commonly use the conditional odds ratio regardless of the study design because its easily estimated with logistic regression. However, for prospective studies, the RR is often considered a more intuitive measure of variable importance [83]. Defined as the relative increase in the probability of the outcome given one is in the exposed group versus the non-exposed group, the interpretation of the RR is straightforward allowing the researcher to communicate results to a variety of audiences. When the outcome is rare, the conditional RR can be approximated using the conditional odds ratio (OR). However, when the outcome is common, estimating the conditional RR can be problematic, especially when the exposure or covariates are continuous [69, 5, 66]. In this chapter, a new estimator for the conditional relative risk (RR) for a common outcome based on targeted maximum likelihood methodology is presented [107]. This estimator is developed under a flexible multiplicative semiparametric model and targets estimation towards the parameter of interest.

In the literature the most prevalent methods for estimating the conditional relative risk for a common outcome are parametric methods based on log-linear (e.g. [112], [89] and [124]), Poisson (e.g. [125] and [69]), or Cox regression models (e.g. [56] and [55]). Overviews and comparisons of the three regression methods can be found in papers by [69], [5] and [66]. These three methods do not imply three unique estimators, however. As the above mentioned papers point out, the estimators implied by using Poisson regression and Cox regression are equivalent. Though some have also suggested methods to covert the OR to the RR, these methods are susceptible to bias in both their estimates and confidence intervals [65, 69].
Given the model is correctly specified, these parametric methods will provide consistent estimates of the conditional relative risk. Log-binomial regression estimates the conditional RR directly using maximum likelihood estimation. However, when continuous covariates are included in the model, log-binomial regression becomes highly susceptible to convergence issues and requires modifications to achieve more stable parameter estimates (e.g. [112], [82] and [66]). Poisson regression is not plagued by these convergence issues and provides consistent estimates of the conditional relative risk when the model is correct [91, 125, 19]. However, the standard errors provided by Poisson regression are overestimated, and alternative methods such as the sandwich estimator [125, 19] or bootstrap [5] must be used to obtain correct inference.

Due to their dependence on the accuracy of a fully specified model, parameter estimates from the above parametric methods are often bias in observation studies. More flexible alternatives presented in the literature include semiparametric counterparts to the log-linear and Poisson regression models. Typically these are built under either generalized partial linear models or generalized additive partial linear models, the latter being a less flexible approach where each covariate has a separate additive component in the model [43]. Estimation methods for the parameter under a partial linear model include profile likelihood methods [85, 86] and backfitting algorithms [43]. Under the additive partial linear model, estimation methods include backfitting [16] and marginal integration [20]. Though these estimation methods are less dependent on model specification, they do not target estimation towards the parameter of interest and can still result in biased estimates and improper inference as discussed in chapter [1].

Previous applications of target maximum likelihood have focused on the additive variable importance (see chapters [2] and [3] as well as [7]). In this chapter, the targeted maximum likelihood estimator of conditional RR under the flexible generalized partial linear model is presented. This model is also referred to as a multiplicative semiparametric regression model. This model formulation only requires specification of the model terms relating to the variable of interest (i.e. the exposure). Under this multiplicative semiparametric model, the TMLE for conditional relative risk can be derived as a fluctuation and targeted update of either a log-binomial or a Poisson density.

In each case, the TMLE is developed to target estimation towards the effect parameter relating to the variable of interest. Inference for the TMLE is then provided using the sandwich variance estimator based on the corresponding influence curve for the effect parameter under the assumed semiparametric multiplicative model and chosen density [107]. Derivation of the TMLE under either density results in consistent estimators for the conditional relative risk, and both TMLE are double robust and asymptotically linear. Under the log-binomial density, the TMLE respects the binary form of the outcome variable. Therefore its corresponding influence curve is also the efficient influence curve for the parameter of interest given a binary outcome, resulting in a locally efficient estimator for the conditional relative risk. Alternatively, under
a Poisson density, the TMLE assumes a count (Poisson) outcome, and though it can be applied effectively to a binary outcome to obtain an estimate for the conditional relative risk, the corresponding influence curve is not the efficient influence curve. Therefore the resulting TMLE is double robust and asymptotically linear, but is not efficient. However, as discussed previously, log-binomial regression is susceptible to convergence issues. Therefore in practice, the Poisson based TMLE is recommended and applied here.

This chapter first reintroduces the data structure (section 4.2), the model (section 4.3), and the parameter of interest (section 4.4) for the TMLE for conditional relative risk. Then in section 4.5, the TMLE for conditional relative risk under the multiplicative semiparametric model is introduced and derived for both log-binomial and Poisson outcomes. This is followed by step-by-step implementation instructions which focus on the Poisson case in section 4.5.2. This chapter concludes with an application of the TMLE to a HIV viral response data set in section 4.6, followed by an overall conclusion in section 4.7. This work was completed in collaboration with Kristin Porter.

### 4.2 Data

The observed data is defined as \( O = (W^*, Y) \sim P_0 \), where \( P_0 \) is the true data generating distribution. Here, \( W^* \) is a vector of \( p \) variables, and \( Y \) is the binary outcome. We define the TMLE measure for a particular variable of interest \( A = W^*_j \), controlling for confounders \( W = W_{-j}^* \).

### 4.3 Multiplicative semiparametric model

The following semiparametric multiplicative model form is assumed for \( E_0[Y|A,W] = P_0(Y = 1|A,W) \):

\[
P_0(Y = 1|A,W) = e^{m(A,V|\beta_0)}P_0(Y = 1|A = 0,W)
\]

or on the additive scale

\[
\log(P_0(Y = 1|A,W)) = m(A,V|\beta_0) + \log(P_0(Y = 1|A = 0,W)),
\]

where \( m(A,V|\beta_0) \) is a specified function of \( A \) and effect modifiers \( V \subset W \), and \( \theta(W) = P_0(Y = 1|A = 0,W) \) is a non-parametric model of the conditional expectation of \( Y \) given baseline covariates \( W \) and \( A = 0 \) (i.e no exposure). The model \( m(A,V|\beta_0) \), can be any form such that \( m(A,V|\beta_0) = 0 \) for all values of \( A \) and \( V \). In this thesis, a linear model is specified for \( m(A,V|\beta_0) \). Two forms are used (1) simple main effect: \( m(A,V|\beta_0) = \beta_0 A \) or (2) V-adjusted \( m(A,V|\beta_0) = \beta'_0 [A \cdot A : V] \), where the effect of exposure \( A \) is modified by covariate \( V \).
4.4 Measure of interest

The measure of interest, the conditional relative risk (RR), is defined in terms of the observed data generating distribution, $P_0$ and parameter vector $\beta_0$ as follows

$$RR(P_0) = \frac{P(Y = 1|A, W)}{P(Y = 1|A = 0, W)} = \frac{\bar{Q}_0(A, W)}{\theta_0(W)} = e^{m(A,V|\beta_0)} \quad (4.1)$$

or on the log scale:

$$\log \left( \frac{\bar{Q}_0(A, W)}{\theta_0(W)} \right) = m(A, V|\beta_0).$$

The measure of interest is a function of the parameter $\beta_0$, therefore targeted maximum likelihood estimation is developed such that estimation is targeted towards $\beta_0$. Note, that under the simple model $m(A,V|\beta_0) = \beta_0 A$, the parameter $\beta_0$ can be interpreted directly as the change in the conditional relative risk with respect to a one unit change in variable $A$. This is referred to as the increment conditional relative risk. Under the effect modification model, $m(A,V|\beta_0) = \beta_0[A \cdot A: V]$, the increment conditional relative risk is defined at a particular value of $V = V_q$.

4.5 Targeted maximum likelihood estimation

The TMLE method provides a targeted update of an initial estimate of the parameter of interest using standard MLE methods and a “clever covariate” defined such that the TMLE solves the efficient score equation and thus acquires the double robust properties of the solution to the estimating equation associated with the score. The TMLE of conditional relative risk under the semiparametric regression model is presented below. The derivation of the TMLE is outlined and implementation steps are also included.

4.5.1 Derivation of the TMLE

The likelihood of the observed data can be factorized as


The parameter of interest only involves $P(Y|A,W)$ therefore no assumptions are made on the distributions of $P(A|W)$ and $P(W)$. When the parameter of interest is the relative risk two natural distribution choices for $P(Y|A,W)$ are log-binomial and Poisson. TMLE estimates can be achieved using either distribution. Both derivations are shown here. In practice the Poisson is more computationally stable and therefore is subsequently presented for implementation and application.
In both cases, an initial density estimate \( P^0(Y|A,W) \), with mean, \( E[Y|A,W] = P(Y = 1|A,W) = \tilde{Q}^0(A,W) \) is defined in terms of the semiparametric model presented above, where
\[
\log \tilde{Q}^0(A,W) = m(A,V|\beta^0) + \log \theta^0(W)
\]
A class of submodels fluctuated with parameter \( \epsilon \) is defined as \( P(Y|A,W)(\epsilon) \) with corresponding mean
\[
\log \tilde{Q}(A,W)(\epsilon) = m(A,V|\beta(\epsilon)) + \log \theta(W)(\epsilon)
\]
where \( m(A,V|\beta(\epsilon)) = m(A,V|\beta^0 + \epsilon) \) and \( \log \theta(W)(\epsilon) = \log \theta^0(W) + \epsilon r(W) \). The form of \( r(W) \) is determined such that at \( \epsilon = 0 \), \( Q(A,W)(\epsilon = 0) = Q^0(A,W) \) and the linear span of the score equation of the likelihood for \( Q(A,W)(\epsilon) \) with respect to \( \epsilon \) at \( \epsilon = 0 \) is equal to the efficient score equation \[107\].

**Log-binomial**

Under the Log-binomial distribution, the initial density is a binomial density defined as
\[
P(Y|A,W) = \tilde{Q}(A,W)^Y(1 - \tilde{Q}(A,W))^{1-Y}
\]
where \( \tilde{Q}(A,W) = P(Y = 1|A,W) = \theta(W)e^{m(A,V|\beta)} \) with the associated fluctuation
\[
P(\epsilon)(Y|A,W) = \tilde{Q}(\epsilon)(A,W)^Y(1 - \tilde{Q}(\epsilon)(A,W))^{1-Y}
\]
given \( \tilde{Q}(A,W)(\epsilon) = \theta(W)(\epsilon)e^{m(A,V|\beta(\epsilon))} \)

The associated score equation for the above likelihood in \( \epsilon \) at \( \epsilon = 0 \) is as follows
\[
S(r) = \frac{1}{1 - \tilde{Q}(A,W)} \left\{ \frac{d}{d\beta} m(A,V|\beta) + r(W) \right\} (Y - \tilde{Q}(A,W))
\]

The score equation associated with the estimating function for effect parameter \( \beta \) under a Log-binomial density given the semiparametric model shown in section [1.3] is defined below.
\[
D_{h_{opt},Q,\beta}(O) = h_{opt}(A,W)(Y - \tilde{Q}(A,W))
\]
where
\[
h_{opt}(A,W) = \frac{1}{1 - \tilde{Q}(A,W)} \left\{ \frac{d}{d\beta} m(A,V|\beta) - \frac{E\left[ \tilde{Q}(A,W) \frac{d}{d\beta} m(A,V|\beta) \right] W}{E\left[ \tilde{Q}(A,W) \right] W} \right\}
\]
This is the efficient score equation for the conditional relative risk defined as the effect parameter under the multiplicative semiparametric model of section [1.3] for binary outcome. The form of this estimating function is within the general class
of estimating functions for effect parameter $\beta$ under the multiplicative semiparametric model (see appendix C.2).

The proper form of the fluctuation function $r(W)$ is therefore as follows

$$r(W) = -\frac{E\left[\frac{\hat{Q}(A,W)}{1-\hat{Q}(A,W)} \frac{d}{d\beta} m(A,V|\beta) \right] W}{E\left[\frac{\hat{Q}(A,W)}{1-\hat{Q}(A,W)} \right] W}$$

Given a model form linear in $\beta$ for $m(A,V|\beta)$, the model,

$$\log \hat{Q}(A,W)(\epsilon) = m(A,V|\beta(\epsilon)) + \log \theta(W)(\epsilon)$$

can be rearranged as an update to the initial fit

$$\log \hat{Q}(A,W)(\epsilon) = \log \hat{Q}^0(A,W) + \epsilon \frac{d}{d\beta} m(A,V|\beta) + \epsilon r(W)$$

Therefore the update can be achieved by estimating $\epsilon$ with standard maximum likelihood estimation. The update can be completed using log-binomial regression setting the initial estimate, $\hat{Q}^0(A,W)$, as an offset and regressing $Y$ onto the following “clever covariate”.

$$H^*(A,W) = \frac{d}{d\beta_0} m_{\beta_0}(A,V) - \frac{E\left[\frac{\hat{Q}(A,W)}{1-\hat{Q}(A,W)} \frac{d}{d\beta} m(A,V|\beta) \right] W}{E\left[\frac{\hat{Q}(A,W)}{1-\hat{Q}(A,W)} \right] W}$$

The update process is iterated until convergence ($\epsilon \approx 0$). The final estimate is the solution to the robust estimating equation corresponding to the efficient score equation

$$\frac{1}{n} \sum_{i=1}^{n} \left[ D_{\text{opt},Q^*,g^*}(O_i|\beta^*) \right] = 0$$

such that $Q^*$, $g^*$, and $\beta^*$ are the converged TMLE of $Q$, $g$, and $\beta$ for the observed data. Therefore, the TMLE inherits the double robust properties of the solution to the efficient estimating equation and the efficient score equation can be used to estimate the correct covariance and inference for the parameter of interest $\beta$.

**Poisson**

Under the Poisson distribution, the initial density is a Poisson density defined as

$$P(Y|A,W) = \frac{\hat{Q}(A,W)^Y}{Y!} e^{-\hat{Q}(A,W)}$$

with the associated fluctuation

$$P(\epsilon)(Y|A,W) = \frac{\hat{Q}(\epsilon)(A,W)^Y}{Y!} e^{-\hat{Q}(\epsilon)(A,W)}$$
The score equation associated with the estimating function for effect parameter \( \beta \) under a Poisson density given the semiparametric model shown in section 1.3, is defined below. Note again, that this form is within the general class of estimating functions for the effect parameter \( \beta \) under the multiplicative semiparametric model (see appendix C.2) and is efficient for a Poisson outcome.

\[
D_{h_{opt},Q,g}(O) = h_{opt}(A,W)(Y - \bar{Q}(A,W))
\]

where

\[
h_{opt}(A,W) = \left\{ \frac{d}{d\beta} m(A,V|\beta) - \frac{E\left[\frac{d}{d\beta} m(A,V|\beta) e^{m(A,V|\beta)|W}\right]}{E[e^{m(A,V|\beta)|W}]} \right\}
\]

Therefore the proper form for \( r(W) \) is

\[
r(W) = -\left\{ \frac{d}{d\beta} m(A,V|\beta) - \frac{E\left[\frac{d}{d\beta} m(A,V|\beta) e^{m(A,V|\beta)|W}\right]}{E[e^{m(A,V|\beta)|W}]} \right\}
\]

Given a simple linear model for \( m(A,V|\beta) = \beta'(A) \) the above can rewritten as

\[
r(W) = -\left\{ A - \frac{E[Ae^{\beta A}|W]}{E[e^{\beta A}|W]} \right\}
\]

Similar to the log-binomial case, the update can be achieved by estimating \( \epsilon \) with standard maximum likelihood estimation. The update is completed using Poisson regression with an offset equal to the initial fit and “clever covariate” defined as

\[
H^* = \frac{d}{d\beta_0} m_{\beta_0}(A,V) - \left\{ \frac{E\left[\frac{d}{d\beta} m(A,V|\beta) e^{m(A,V|\beta)|W}\right]}{E[e^{m(A,V|\beta)|W}]} \right\}
\]

Again, the update process must be iterated until convergence ( \( \epsilon \approx 0 \)). The final estimate is the solution to the corresponding robust estimating equation. Therefore, the TMLE inherits the double robust property and the corresponding influence curve can be used to estimate the correct covariance and inference for the parameter of interest \( \beta \). However, as mentioned previously, the corresponding influence curve is not the efficient influence curve and the resulting TMLE will be double robust and asymptotically linear but will not be locally efficient.

In practice, the Poisson based TMLE is recommended due to its computational stability. Therefore, subsequent implementation instructions and applications in this chapter will be presented in terms of the Poisson TMLE.

**Estimating the “clever covariate”**

Unlike the TMLE for a continuous outcome presented in chapters 2 and 3, the “clever covariate” presented here is not a simple function of the treatment mechanism.
(i.e. \( g(W) \)). When \( A \) is continuous, the “clever covariate” can be estimated using a data-adaptive algorithm to estimate the expectations in the numerator and denominator of the second term. This is the method that is used in the following application in section 4.6.

Given a binary or categorical \( A \) with \( L \) levels, the “clever covariate” can also be directly estimated as follows

\[
H^*(A, W) = A - \sum_{l=1}^{L} \frac{d}{d\beta} m(A = A_l, V | \beta) e^{m(A = A_l, V | \beta)} g_n(A = A_l | W) \frac{g_n(A = A_l | W)}{\sum_{l=1}^{L} e^{m(A = A_l, V | \beta)} g_n(A = A_l | W)}
\]

This method requires the estimation of \( g_n(A = A_l | W) \) for all \( l \). It is recommended that these values are estimated data-adaptively. This method can also be used to approximate the expectation given a continuous \( A \) if the analyst is willing to discretize \( A \) solely for the purpose of estimating this covariate.

**Inference**

As in the previous chapters, inference for the TMLE can be obtained from the influence curve as outlined in section 1.4.3. Covariance may also be estimated using the bootstrap. However, this TMLE requires iteration and can be computationally intensive. Therefore, bootstrap estimates are not recommended.

An estimate of the \( p \) by \( p \) covariance matrix, \( \Sigma_n \), for parameter vector \( \beta_n \), of length \( p \), is obtained as outlined in section 1.4.3. Using the estimated covariance matrix, hypothesis tests can be performed for a single parameter \( \beta_n(j) \), where \( j = 1, \ldots, p \), under the null hypothesis \( H_0: \beta_n(j) = 0 \) using a standard test statistic to obtain \( p \)-values, with estimated variance \( \Sigma_n(j, j) \).

\[
T_n(j) = \frac{\sqrt{n} \beta_n(j)}{\sqrt{\Sigma_n(j, j)}} \xrightarrow{n \to \infty} \text{Normal}(0, 1)
\]

Likewise the hypothesis \( H_0: c'\beta_n = 0 \) can also be tested using a standard Wald test, where the covariance of \( c'\beta_n \) is \( c'\Sigma_n c \). In the following application, effect modification is tested at various levels of \( V \). In this situation the vector \( c \) for the Wald test is \( \{1, V_q\} \), where \( V_q \) is a specific quantile of effect modifier \( V \).

**4.5.2 TMLE implementation steps**

Implementation steps are presented in terms of the TMLE under a Poisson distribution. Given a linear model form for our working model, \( m(A, V | \beta) = A(\beta'V) \), the update at each iteration can be achieved using generalized linear regression for Poisson distribution with natural log link (R function \texttt{glm}). At each iteration, response \( Y \) is regressed onto the “clever covariate” defined for the linear model as
\[ H^*(A, W) = AV + \frac{E[AV e^{(A(\beta V))} | W]}{E[e^{(A(\beta V))} | W]} \]

There are three initial components necessary for applying targeted maximum likelihood methodology to estimate TMLE

1. Model \( m(A, V|\beta) \) satisfying \( m(A = 0, V|\beta) = 0 \) for any \( \beta \) and \( V \)

2. An estimate for \( g(W) = P(A|W) \) or data adaptive algorithm to estimate \( E[AV e^{(A(\beta V))} | W] \) and \( E[e^{(A(\beta V))} | W] \) directly

3. An initial estimate for \( \bar{Q}(A, W) = P(Y = 1|A, W) \), \( \bar{Q}_n^0(A, W) \), containing valid model \( m(A, V|\beta) \) providing an initial estimate for parameter \( \beta, \beta_n^0 \)

The initial regression estimate of proper form may be obtained from semiparametric methods such as those of \[90, 85, 43\] among others, or by using methods such as DSA \[88\] which allow the user to fix a portion of the model. However, a more flexible approach is adopted. This method is presented previously in chapters 2 and 3 and is summarized again here. After obtaining an initial estimate for \( \bar{Q}(A, W) \) of general form using using data-adaptive machine learning algorithms such as super learner \[104\] or DSA \[88\], solve for \( \bar{Q}(A = 0, W) \) using this general fit. Then regress \( Y \) onto the model \( m(A, W|\beta^0) \) and treating \( \bar{Q}(A = 0, W) \) as a covariate. This provides initial estimates \( \bar{Q}_n^0(A, W) \) and \( \beta_n^0 \). This update improves computational efficiency by only requiring a single data-adaptive estimate for \( \bar{Q}(A, W) \) of general model form for all \( A \).

Given the three components, TMLE is applied using the following steps

1. Estimate the “clever covariate” which will allow the update of the initial regression in a direction which targets the parameter of interest. In this case the clever covariate is:

\[ H^*(A, W) = AV + \frac{E[AV e^{(A(\beta^0 V))} | W]}{E[e^{(A(\beta^0 V))} | W]} \]

2. Compute the fitted values for the initial estimate, \( \bar{Q}_n^0(A, W) \)

3. Project \( Y \) onto \( H^*(A, W) \) with offset \( = \bar{Q}_n^0(A, W) \). The resulting coefficient is \( \epsilon \). This is done using generalized Poisson regression and setting the offset, and projecting onto the model \( Y \sim H^*(A, W) + \text{offset} \). Note there is no intercept in the model, only the offset value.

4. Update initial estimate \( \beta_n^1 = \beta_n^0 + \epsilon \) and overall fit

\[ \log(\bar{Q}_n^1(A, W)) = \log(\bar{Q}_n^0(A, W)) + \epsilon H^*(A, W) \]

These are now the first-step targeted estimates.
5. Iterate steps 1 through 4. At each $k^{th}$ iteration, set $\beta^k = \beta^{k-1} + \epsilon$ and $\log(\hat{Q}_n^k(A,W)) = \log(\hat{Q}_n^{k-1}(A,W)) + \epsilon H^*(A,W)$, until convergence ($\epsilon \approx 0$)

6. Obtain standard error and inference for the final estimate $\beta^*$ using the influence curve as outlined in section 1.4.3

4.6 Application

Genotypic resistance testing has become a powerful tool for clinicians in determining the appropriate treatment regimen for people with HIV [29, 98]. However interpretation of the resistance testing can be difficult. Over the years multiple interpretation algorithms have been developed to provide more straightforward measures of resistance [78]. A study completed by [78] assessed the predictive ability of four genotypic resistance test interpretation algorithms (ANRS [27], HIVdb [63], Rega [108], and ViroSeq [32]) to determine virologic response (VR), adjusting for additional baseline covariates. In this analysis this data is reanalyzed using targeted maximum likelihood methodology to determine the importance of each algorithm with respect to its predictive fit. Then, the genotypic algorithm with the highest importance measure is analyzed further by estimating the modification of its effect by each of the covariates included in the original model.

4.6.1 Background

For individuals infected with human immunodeficiency virus (HIV) effective antiretroviral (ARV) treatments carry the promise of a longer and more gratifying life. HIV infects the body’s immune system and progressively destroys and impairs its ability to fight off infection. ARV treatments are designed to slow down HIV reproduction and stall its debilitating effects. A properly designed and administered treatment can prolong survival and increase overall quality of life [74].

There are many ARV drugs available. Some of the more common classes of ARV drugs are boosted protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs) [74]. Patients are generally placed on a combination of several drugs from multiple classes called a treatment regimen. However, HIV is a rapidly evolving virus and commonly develops drug resistant mutations rendering initially effective treatment regimens useless [74]. The rapid rate of mutation has made genotypic resistance testing essential to determining the appropriate treatment regimen for an individual patient [29, 98]. To facilitate the interpretation of these tests, interpretation algorithms have been developed. The algorithms are drug-specific and applied to a patient’s baseline genotype [78].
4.6.2 Data

Study subjects were selected according to the eligibility requirements outlined in [78] from 16 clinics of the Kaiser-Permanente Medical Care Program, Northern California. In the original study 734 valid treatment change episodes (TCEs) were recorded for 641 patients. In this study a valid TCE occurs when an individual has undergone a change in treatment regimen within 24 weeks of a genotypic resistance test and has received at least four weeks of a new salvage regimen. Though the original study uses all TCEs, to simplify our analysis, we randomly select only one TCE per patient.

Virologic response is measured according to plasma HIV-1 RNA levels. High plasma HIV-1 RNA levels indicate strong viral activity. The original study classifies VR into three classes: sustained, transient, and absent. Sustained VR is achieved when two subsequent tests show plasma RNA levels below the limit of quantification (BLQ). Transient refers to cases where only 1 subsequent test was at BLQ level, and absent when no subsequent test is BLQ. For this analysis sustained and transient classes are merged into a single class (VR=1) (see [78] for more details).

The original study focused on four interpretation algorithms: ANRS [27], HIVdb [63], Rega [108], and ViroSeq [32]. Each algorithm is applied to an individual patient’s baseline genotype to determine the drug specific genotypic susceptibility scores (GSS’s) for each ARV. GSS measures range from 0 to 1, where a GSS of 1 indicates full susceptibility of HIV-1 to the particular ARV and a GSS of 0 indicates full resistance. The drug-specific GSS’s are then combined into regimen specific GSS’s (rGSS’s) through three alternative weighing schemes: “boosted PI weighted”, “comprehensive weighted”, and “unweighted”. The “unweighted” rGSS is the addition of all drug-specific GSS’s weighted equally with 1.0. The “boosted PI weighted” rGSS increases the weight of the drug-specific GSS’s for boosted PIs to 1.5. The “comprehensive weighted” rGSS increases the weight of the drug-specific GSS’s for boosted PIs to 2.0 and decreases the drug-specific GSS’s for NRTIs to 0.5. This results in a total of 12 rGSS’s. We standardize each rGSS by subtracting its mean and dividing by its standard deviation to allow direct comparison of the importance measures.

Additional covariates are also included in the original prediction analysis including individual demographics (age, sex, and race), features of ARV treatment prior to the TCE (i.e. duration of therapy, number of ARV’s, etc.), and features of the salvage regimen (i.e. number of new ARV’s, new ARV drug classes, etc.), as well as plasma HIV-1 RNA level and CD4 count at baseline.

4.6.3 Analysis

Targeted maximum likelihood estimation is first applied to estimate the variable importance of each rGSS with respect to its own prediction of VR. These estimates provide a measure of how much each variable changes the probability of VR on a rela-
tive scale. As stated previously in section 4.5, TMLE for variable importance updates an initial regression estimate to target the parameter of interest. In this case, the initial regression estimate for a specific rGSS is the super learner estimate, predicting VR using the rGSS and additional covariates. The covariate set is consistent across all rGSS fits and is defined using univariate logistic regression. Covariates associated with VR with a p-value of 0.1 or less are included. This analysis focuses on the overall effect of the rGSS, therefore the model is the simple effect model: $m(A, V|\beta) = \beta A$. For each importance measure, inference is obtained using the influence-curved based estimate of the standard error.
4.6.4 Results

Figure 4.1: TMLE estimated coefficients for 4 algorithms, ANRS, HIVdb, Rega, and ViroSeq, under three different weighting schemes: “unweighted” (W0), “boosted PI weighted” (W1), and “comprehensive weighted” (W3). The brackets represent the 95% confidence intervals according the influence curve based estimate of the standard error.
Figure 4.2: TMLE estimated increment RR for 4 algorithms, ANRS, HIVdb, Rega, and ViroSeq, under three different weighting schemes: “unweighted” (W0), “boosted PI weighted” (W1), and “comprehensive weighted” (W3). The brackets represent the 95% confidence intervals according the influence curve based estimate of the standard error using the delta method.

Results of the initial analysis show significant importance values for all rGSS algorithms as expected. Note that though standardizing the scores allows us to directly compare the importance measures, the increment increase in RR is now relative to an increase of one in the z-score of the rGSS or correspondingly an increase of one.
standard deviation of the original rGSS measure. From Figures 4.1 and 4.2, we see that in general weighting did have an effect on overall importance. Weighting scheme 1 seems to increase the coefficient for each algorithm over the unweighted, and weighting scheme 2 increases it even more over that. Under any weighting scheme, ANRS seemed to have the highest coefficient and corresponding increment RR (e.g. the change in the probability of sustained virologic response for one unit increase in the zscore of the rGSS). Of the twelve, the highest importance is attributed to ANRS with “comprehensive weighting” with a coefficient $\beta = 0.206$ corresponding to an increment RR of 1.23 and associated p-value of 3.80e-06 before adjusting for multiple testing.

4.6.5 Secondary analysis

As a secondary analysis the GSS with the highest importance, comprehensively weighted ANRS, is chosen and V-adjust variable importance analysis is performed. In this analysis, the importance measure adjusted by each of the other covariates is estimated. The model for this analysis is as follows

$$m(A, V|\beta) = A(\beta_A + \beta_V V)$$

where the parameter of interest is now measured as a function of two parameters with respect to a particular covariate, $V$. The results are shown below.

The individual coefficient values estimated using TMLE with their respective confidence intervals are shown in Figure 4.3. Given only the coefficients it is difficult to interpret the results. To clarify, the increment RR change is calculated at varying levels, $V_q$ of each effect modifier $V$. This is defined as follows for value $V_q$ of any $V$ as

$$RR = e^{\beta_A + \beta_V V_q}$$

Calculating the corresponding standard error is achieved by first calculating the standard error of the linear combination $c'\beta = \beta_A + \beta_V V_q$, as $c\Sigma_V c'$, where $c = \{1, V_q\}$, and $\Sigma_V$ is the influence curved based covariance estimate for $\{\beta_A, \beta_V\}$. Then, the delta method is used to calculate the corresponding standard error estimate for the exponential of this linear combination. The quantiles of $V$ (min, 25%, 50%, 75%, max) are chosen for $V_q$. Note that in some cases the covariate in binary and there are only two possible values, and therefore only two points. The results are shown in Figures 4.4 and 4.5.
Figure 4.3: TMLE coefficients ($\beta$) for ANRS algorithm under comprehensive weighting adjusted by the other covariates. Left plot is the coefficient for the main effect, $A$, and the right plot is the coefficient of the effect modification or interaction term, $A: V$. The brackets represent the 95% confidence intervals according to the influence curve based estimate of the standard error. The covariates adjusted are listed from top to bottom: history of virologic suppression prior to baseline (prev_vr_suppress), number of PIs in new regimen (NumPIs), number of regimens received prior to baseline (NumHXRegimens), number of PIs received prior to baseline (NumHXPIs), number of NRTIs received prior to baseline (NumHXNNRTIs), number of non-HAART regimens received prior to baseline (NumHXNonHAART), number of NNRTIs received prior to baseline (NumHXXNRTIs), number of HAART regimens received prior to baseline (NumHXHAART), number of new ARVs in new regimen (num_newdr_newReg), number of ARVs received prior to baseline (num_dr_pastReg), number of ARVs in new regimen (num_dr_newReg), number of new ARV classes in new regimen (newdr_class), duration of new regimen in weeks (duration_newReg), plasma HIV-1 RNA level at baseline in log-copies/ml (bl_vload), CD4 count at baseline in cells/ml (bl_cd4), and age at baseline in years (age_baseline).
Figure 4.4: TMLE increment RR for ANRS algorithm under comprehensive weighting modified by covariate $V$ at quantile levels $V_q = \{0\%, 25\%, 50\%, 75\%, 100\%\}$ for each covariate $V$. The brackets represent the 95% confidence intervals according to the influence curve based estimate of the standard error. Covariate $V$ are as follows (left to right, top to bottom): duration of new regimen in weeks (duration_newReg), number of HAART regimens received prior to baseline (NumHXHAART), number of NNRTIs received prior to baseline (NumHXNNRTIs), number of non-HAART regimens received prior to baseline (NumHXNonHAART), number of NRTIs received prior to baseline (NumHXNRTIs), number of PIs received prior to baseline (NumHXPIs), number of regimens received prior to baseline (NumHXRegimens), and number of PIs in new regimen (NumPIs).
Figure 4.5: TMLE increment RR for ANRS algorithm under comprehensive weighting modified by covariate \( V \) at quantile levels \( V_q = \{0\%, 25\%, 50\%, 75\%, 100\%\} \) for each covariate \( V \). The brackets represent the 95% confidence intervals according the influence curve based estimate of the standard error. Covariate \( V \) are as follows (left to right, top to bottom): age at baseline in years (age\_baseline), CD4 count at baseline in cells/ml (bl\_cd4), plasma HIV-1 RNA level at baseline in log-copies/ml (bl\_vload), number of new ARV classes in new regimen (new\_dr\_class), number of ARVs in new regimen (num\_dr\_newReg), number of ARVs received prior to baseline (num\_dr\_pastReg), number of new ARVs in new regimen (num\_newdr\_newReg), history of virologic suppression prior to baseline (prev\_vr\_suppress).
From the results (Figures 4.4 and 4.5), it can be seen that the increase in the risk of sustained virologic response with respect to change in the rGSS score of ANRS under comprehensive weighting is modified by many of the covariates. This is not surprising due to the complexity of body’s response to HIV. Virologic response is without a doubt a combination of genetics, current viral load, as well as current and past treatment regimens. For instance, it is logical that increased baseline viral load would result in increased risk of virologic response, but through this type of analysis, it can also be seen that increased baseline viral load seems to modify the effect of the genetic score on the relative risk of VR (Figures 4.4 and 4.5). This type of analysis helps elucidate and interpret the complex set of interactions that results in the body’s virologic response. Having a method which targets the effect and provides consistent and locally efficient estimates of the effect with formal inference is key to further the understanding and treatment of diseases such as HIV.

4.7 Discussion

This paper has introduced a TMLE for an important and readily interpretable parameter of interest in medical and epidemiology studies, the conditional relative risk. The most commonly employed methods, especially in the field of epidemiology rely on parametric models and the manipulation of statistical software to obtain a stable estimate. These methods also do not target the parameter of interest and are not double robust. In contrast, the TMLE of relative risk is double robust, in that it is unbiased if either the estimator for $\bar{Q}_0(A,W)$ or the estimator for $g_0(A,W)$ is consistent. This feature is very important in a world where models are often incorrect. The TMLE is also locally efficient if both estimators are consistent and it is estimated under the log-binomial distribution.

The utility of TMLE for real analyses is shown through the application. In this chapter, the estimation is targeted towards the main effect of a particular variable, in this case a genotypic score. This type of analysis is particularly useful in biomarker discovery and variable importance analyses when one is interested in accurately testing the effect of many variables (see chapter 2 and [96]). The analysis was taken one step further and the model was augmented to allow TMLE to test modifications of an effect by a single or set of covariates. Here, the effect of a genotypic score on the relative risk of virologic response was modified by the baseline viral load. Accurate and interpretable effect modification analysis such as this can be useful in clinical trials to test how the effect of a treatment ($A$) is modified by a particular gene expression for instance.
Chapter 5
Conclusion

In this thesis, three estimators for three types of outcomes are developed under a semiparametric partial linear model using targeted maximum likelihood estimation. These estimators are developed specifically to target measures of variable importance for analyses where the question of interest is primarily concerned with determining the effect of a particular variable on a given outcome while controlling for potential confounders. Alternative methods often used in variable importance analyses include parametric methods such as linear regression and data mining algorithms such as Random Forest. These alternative methods focus estimation and inference on the bias-variance trade-off associated with the overall distribution, not the specific parameter of interest. This introduces unnecessary bias and can result in expensive false positives. Estimators of variable importance based on targeted maximum likelihood estimation target estimation and inference specifically towards the parameter of interest. The resulting estimators of variable importance are double robust and locally efficient and have the added benefit of an associated targeted likelihood which allows one to implement standard model selection approaches such as cross validation.

Targeted maximum likelihood estimation is a robust estimation method that updates an initial density estimate in a direction which reduces the bias for the parameter of interest. This targeted estimation method combined with data-adaptive prediction algorithms, such as super learner [104], provides a unified framework for variable importance analysis, allowing one to control for all potential confounders while still targeting an interpretable parameter of interest. The measures of importance presented in this thesis are causally interpretable parameters under the appropriate causal assumptions. By targeting these measures specifically, the resulting variable importance measures retain a causal interpretation when applied to clinical trials or specifically designed experiments where the appropriate randomization and time ordering assumptions hold. When these assumptions do not hold, these measures are still useful as measures of variable importance.

In contrast to previously published targeted maximum likelihood estimators, this thesis focuses on estimators developed under a semiparametric partial linear model.
This model makes fewer assumptions than a full parametric model and only requires specification of the model components relating to the parameter of interest. It accommodates both binary and continuous variables of interest while providing a straightforward and interpretable way to incorporate effect modifiers of the variable of interest. Unlike other some TMLE presented in the past \[105\], the TMLE under the semiparametric model does not depend on inverse probability weighting and is therefore more robust to ETA violations \[96\].

For each of the three outcomes (1) univariate continuous (2) multivariate continuous and (3) binary, the targeted maximum likelihood estimator is developed and implemented in both simulation and application. This thesis presents the complete derivation of the TMLE for each of these outcomes under the semiparametric partial linear model. The basic methodology behind targeted maximum likelihood estimation is also reviewed for the reader. Step by step implementation instructions are provided for each estimator using standard statistical software packages. In general all analysis are completed in R and some sample R code is provided.

The applications presented in this thesis involve purely observational data. In these cases, the TMLE must rely on the accuracy of the initial fit for \(E[Y|A,W]\) or the fit of the treatment/confounding mechanism, \(E[A|W]\). Though not presented here, the TMLE is ideally suited for use in estimating the effects of the “treatment” in a randomized trial. For instance, a clinical trial for a new AIDS drug would be interested in the average effect of the drug on CD_4 counts over time. In other words, \(E[E[CD_4|Drug_A, time] − E[CD_4|placebo, time]] = E[\beta_{Drug_A}]\), where \(\beta\) represents the effect of drug A over time. Given a randomized experimental design, the estimate of \(E[A|W]\) is consistent and therefore the TMLE is a consistent estimate of \(\beta\).

In chapter 2 the TMLE for a univariate continuous outcome was presented as a candidate standard measure for biomarker discovery analyses. In biomarker analyses, there are thousands of variables to test and the data is highly correlated. The TMLE is a especially useful tool in high-dimensional data sets in that each individual variable can be targeted separately and receive its own importance value with accurate inference while still adjusting for potential confounders. The capabilities of the TMLE is demonstrated in a simulation study where the correlation among the covariates is steadily increased. The TMLE as able to effectively detect the “true” variables and often outperforms the alternative methods, linear regression, lasso regression, and Random Forest.

In chapter 3 the TMLE for a continuous multivariate outcome was developed. Specifically designed for a repeated measures analysis, the TMLE was implemented as a targeted update to an initial fit based on generalized estimating equations. The associated efficient influence was also determined and its proof is outlined in appendix C.1. This TMLE is applied to a transcription factor analysis. In this analysis, it was able to accuracy detect relevant time points of transcription factor activity in the yeast cell cycle. A model selection procedure was also introduced in this chapter and applied in practice to this analysis to choose among potential effect modifiers of
a given transcription factor.

In chapter 4, the method is extended to a generalized partial linear model. In this case, the TMLE is developed for a binary outcome, and the model has a multiplicative or log-linear semiparametric form. The targeted measure of interest is the adjusted relative risk. This TMLE is developed as an update to either a log-binomial or Poisson regression estimate, and proper inference was achieved through targeted maximum likelihood theory using the associated influence curve. The Poisson based TMLE was then applied to an analysis focused on measuring the effectiveness of genetic susceptibility scores on predicting virologic response in HIV.

Targeted maximum likelihood has been shown to be robust and practical method for estimating variable importance measures. In this thesis, the flexibility of the method to accommodate continuous variables of interest for multiple types of outcomes is demonstrated through the use of a semiparametric partial linear model form. The resulting estimators are straightforward to implement and easy to interpret making this methodology applicable to a wide range of data types in a variety of research disciplines as demonstrated through the applications presented in this thesis. The targeted maximum likelihood methodology is an up-and-coming area of statistical research and is continuously being developed and advanced. This thesis presents a culmination of work on TMLEs under the semiparametric model. It is important to note that there are an ever increasing number of targeted maximum likelihood based estimators including methods developed for survival outcomes as well as case-control studies [105].
Bibliography


[58] G. Leone, R. Sears, E. Huang, R. Rempel, F. Nuckolls, C.H. Park, P. Gian-
grande, L. Wu, H.I. Saavedra, S.J. Field, M.A. Thompson, H. Yang, Y. Fu-
distinct e2f activities to induce s phase and apoptosis. *Molec. Cell*, 8:105–113,

[59] J.L. Li, Y.C. Xia, M. Palta, and A. Shankar. Impact of unknown covari-
ance structures in semiparametric models for longitudinal data: An applica-
tion to wisconsin diabetes data. *Computational Statistics and Data Analysis*,


[61] A. Liaw and M. Wiener. *randomforest: Classification and regression by ran-

generalized estimating equations. *Journal of the American Statistical Associa-


[64] Y. L. Liu, M. W. Taylor, and H. J. Edenberg. Model-based identification of

[65] A.R. Localio, D.J. Margolis, and J.A. Berlin. Relative risks and confidence
intervals were easily computed indirectly from multivariate logistic regression.

[66] T. Lumley, R. Kronmal, and S. Ma. Relative risk regression in medical research:
Models, contrasts, estimators, and algorithms. Technical Report 293, University

E. Fraenkel. An improved map of conserved regulatory sites for saccharomyces

[68] J.I. Martin-Subero, R. Ibbotson, W. Klapper, L. Michaux, E. Callet-Bauchu,
F. Berger, M.J. Calasanz, C. De Wolf-Peeters, M.J. Dyer, P. Felman, A. Gar-
diner, R.D. Gascoyne, S. Gesk, L. Harder, D.E. Horsman, M. Kneba, R. Kup-
pers, A. Majid, N. Parry-Jones, M. Ritgen, M. Salido, F. Sole, G. Thiel, H.H.
Wacker, D. Oscier, I. Wlodarska, and R. Siebert. A comprehensive genetic and
histopathologic analysis identifies two subgroups of b-cell malignancies carrying
a t(14;19)(q32;q13) or variant bcl3-translocation. *Leukemia*, 21(7):1532–1544, 2007.


Appendix A

Additional results for TMLE - univariate continuous case

A.1 Double robust property

In this section, the double robust property of the TMLE for a univariate continuous outcome under the partial linear model is explored. A simulation study is presented which compares the performance of the TMLE to basic linear regression as well as estimation of the parameter of interest using the estimating equation approach \[\text{[106]}\]. Additionally the TMLE estimator may also incorporate an update to the “treatment mechanism”. This update is presented below and compared to the standard TMLE in simulation as well.

A.1.1 Updating the “treatment mechanism”

The “treatment mechanism” may also be updated in a similar fashion to the standard update presented in section \text{[2.3]} First, note that the efficient influence curve may be decomposed as follows.

\[ D_0(P) = D_0(P) - \Pi(D_0(P)|T_{CAR}) \]

where \( D_0(P) = (h(P)(A|W) - E(h(P)|W))(Y - m(A,V|\beta)) \) and \( D_{CAR} = \Pi(D(P)|T_{CAR}) \) is the projection of \( D(P_0) \) onto \( T_{CAR} \) in the Hilbert space \( L_0^2(P) \) which is equal to \( (h(P)(A|W) - E(h(P)|W))(\theta(P)(W)) \).

Note that in the case of a linear model \( m(A,V|\beta) = A\beta^V' \) and assuming \( \sigma(A,W) = \sigma(W) \), \( D_{CAR} \) reduces to \( V(A - E[A|W])\theta(W) \). Also note that the update to \( \tilde{Q}(A,W) \), \( r(W) \), reduces to \( r(W) = VE[A|W] \). Therefore updates on the “treatment mechanism” \( g(W) = P(A|W) \) will influence the value of \( r(W) \) and therefore also the overall density update.
In this case, $A$ is continuous therefore, a normal density model, $p_g(A|W)$, with mean $g(W) = E[A|W]$ and variance $\sigma_g(W)$ is assumed. The corresponding hardest submodel is then defined as the normal density $p_g(\epsilon_g)(A|W)$ with mean 

$$\bar{g}(\epsilon_g)(A|W) = g^0(W) - \epsilon_g r_g(W)$$

The score of this density in terms of $\epsilon_g$ at $\epsilon_a = 0$, is 

$$\frac{d}{d\epsilon_g} p_g(\epsilon_g)(A|W)\bigg|_{\epsilon_g=0} = -\frac{1}{\sigma_g(W)} (A - \bar{g}(W)) r_g(W)$$

Equating the above score to $D_{CAR}$ gives 

$$r_g(W) = -\sigma_g^2(W)\theta(W)W$$

The likelihood for $p_g(\epsilon_g)(A|W)$ can now be maximized in terms of $\epsilon_g$, providing an update for $\bar{g}(W)$. This can also be completed using weighted linear least squares.

### A.1.2 Simulation study

A simulation study was completed to compare TMLE to simple linear regression, where $E[Y \mid A = 1, W] - E[Y \mid A = 0, W] = \beta$. The basic TMLE update is also compared to TMLE with the “treatment mechanism” update and the alternative estimation of $\beta$ using estimating function methodology [102].

Data is simulated with linear $m(A, W|\beta) = A(3 + 1.5W_1 + 2W_2)$ and $\theta(W) = 4W_2^2 + 5.6W_3^4$. Linear regression estimate is linear in all terms and modeled as $Y \sim A + A: W_1 + A: W_2 + W_1 + W_2$. The covariate matrix $W$ is simulated from a multivariate normal with mean vector $\{5, -1, 3\}$ and independent covariance matrix with constant variance equal to 1. Variable of interest $A$ is simulated from a normal distribution with mean 4 and variance 1. In this simulation, $n = 1000$ observations were simulated $n_{sim} = 1000$ times.

Table A.1: Comparing the estimation of $\beta$ values given simulation from a model with non-linear $\theta(W)$. Given $m(A, V|\beta) = A(3 + 1.5W_1 + 2W_2)$, data is simulated using $\theta(W) = 4W_2^2 + 5.6W_3^4$. The average MSE with respect to $m(A, V|\beta)$ is 1.0067 for TMLE, 1.0066 for TMLE with treatment update, 3.8768 with the estimating function based method, and 5.6621 with linear regression.

<table>
<thead>
<tr>
<th>Actual</th>
<th>TMLE Estimate</th>
<th>Std Er</th>
<th>TMLE w/g(W) update Estimate</th>
<th>Std Er</th>
<th>Est. Fxn Estimate</th>
<th>Std Er</th>
<th>Linear Regression Estimate</th>
<th>Std Er</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0000</td>
<td>3.0012</td>
<td>0.0521</td>
<td>3.0074</td>
<td>0.0516</td>
<td>1.8383</td>
<td>0.8809</td>
<td>1.6504</td>
<td>0.7490</td>
</tr>
<tr>
<td>1.5000</td>
<td>1.4939</td>
<td>0.0193</td>
<td>1.4931</td>
<td>0.0191</td>
<td>1.9389</td>
<td>0.3814</td>
<td>2.0663</td>
<td>0.2340</td>
</tr>
<tr>
<td>2.0000</td>
<td>1.9963</td>
<td>0.0362</td>
<td>2.0018</td>
<td>0.0359</td>
<td>1.7975</td>
<td>0.5372</td>
<td>1.9034</td>
<td>0.5365</td>
</tr>
</tbody>
</table>
Linear Regression, assuming the linearity of all terms, incorrectly specifies the model form, and thus the estimates of $\beta$ are inconsistent with the simulated values. The TMLE on the other hand does not require specification of $g(W)$ to estimate $\beta$ consistently. The average MSE with respect to $m(A,V|\beta)$ increases from 1.0067 (or 1.0066 with treatment update) with TMLE to 5.6621 with linear regression. The MSE for the estimating function based method is 3.8768. Note that since the treatment mechanism is set at the “truth,” a treatment update will be minimal with minimal effect. The simulation shows that the TMLE is more robust to model misspecification than standard linear regression, and also outperforms the estimating function based estimation method as well when given accurate knowledge about $E[A|W]$.

A.2 A Comparison to nonparametric inverse weighing approach: accessing ETA violation performance

In this section, the performance of the variable importance estimator under a semiparametric model is compared to a proposed method for estimating a similar variable importance measure for continuous $A$ using the nonparametric DR-IPTW as presented in [102]. One of the benefits of the semiparametric model based estimating function over the DR-IPTW is that it does not require inverse weighing by the “treatment mechanism” and avoids instabilities due to this weighing. The methods are compared under increasing levels of ETA violation in simulation. Both estimators are obtained using estimating function approach. Note, “V-adjusted” refers to the ability of the estimators to include modification of the effect of $A$ by covariate(s) $V$.

A.2.1 DR-IPTW estimator for V-adjusted variable importance

An alternative approach for V-adjusted variable importance is purposed in [102] for general $A$. If $A$ is continuous, first discretize $A$ into $\{a : a \in A_d\}$ and calculate the IPTW-DR estimating function $D_{h,DR-IPTW}(O)$ for each $A$, where

$$D_{h,DR-IPTW}(O) = \frac{1(A=1)}{P(A=1|W)}(Y - E[Y|A=1,W]) - \frac{1(A=0)}{P(A=0|W)}(Y - E[Y|A=0,W]) + E[Y|A=1,W] - E[Y|A=0,W]$$
This creates a counterfactual set \( D_{h,d} = \{ D_{h,DR-IPTW}(Y, A = a, W) : a \in A_d \} \). Regress \( D_{h,d} \) on \( V \) using a specified model form \( m(A, V | \beta) \). Coefficients of this regression will provide \( V \)-adjusted variable importance measures.

Here we show the result that the estimating function associated with this method of \( V \)-adjusted variable importance belongs in the broader class of estimating functions for known \( g(A|W) \) provided in [120]. The smaller subspace for unknown \( g(A|W) \) shown in Theorem 1 containing the efficient influence curve, is contained in this more general subspace. However this new \( V \)-adjusted variable importance is not contained within the class of estimating functions in Theorem 1 showing that though it is a valid estimating function it is not efficient. The result is shown for binary \( A \), but is applicable to any discrete \( A \).

The estimating function of the alternative approach is simply the least squares estimation function of the regression of \( D_{h,d} \) on \( V \) with model \( m(A, V | \beta) \)

\[
D_{IPTW-DR,VI}(0) = \frac{d}{d\beta} m(A, V \beta) \left( \frac{I\{A=0\}}{P(A=1|W)} (Y - E[Y|A = 1, W]) \right.
- \left. \frac{I\{A=0\}}{P(A=0|W)} (Y - E[Y|A = 0, W]) \right.
+ E[Y|A = 1, W] - E[Y|A = 0, W] - m(A, V | \beta) \right)
\]

Simple algebra equates this to the more general class of estimating functions proved in [120] Theorem 2.1 shown below for known \( g(A|W) \), indexed by vector valued function \( h_1, h_2 \).

\[
D_{h_1,h_2}(O) = (Y - m)(h_1 - E[h_1|W]) + h_2 - E[h_2|W]
\]

Simple algebra shows the two are equivalent when

\[
h_1 = \frac{d}{d\beta} m(A, V \beta) \left( \frac{I\{A = 1\}}{P(A = 1|W)} - \frac{I\{A = 0\}}{P(A = 0|W)} \right)
\]

\[
h_2 = \frac{d}{d\beta} m(A, V \beta) \left\{ \left( \frac{I\{A = 1\}}{P(A = 1|W)} - \frac{I\{A = 0\}}{P(A = 0|W)} \right) (m(A, V | \beta) - E[Y|A, W]) \right.
\]

We propose that not only is the semiparametric model based method more efficient, it is also more robust to ETA violations due to the lack of inverse weighing. Here we provide simulations comparing it to DR-IPTW under varying levels of ETA violations to demonstrate the increased robustness of the semiparametric model based method over double robust IPTW for a binary variable of interest, \( A \).
A.2.2 Simulation study

Simulations comparing DR-IPTW and the semiparametric model based approach under varying levels of ETA violations to demonstrate the increased robustness of the semiparametric method over double robust IPTW for a binary variable of interest, $A$. The variable of interest, $A$, is simulated as binomial with mean $\text{logit}(\beta_I + \beta_W)$. Values of $\beta_W$ are increased to incorporate increasing levels of ETA violation. Simulations are completed for $n = 1000$ observations and $n_{sim} = 1000$ iterations.
Table A.2: Binary Variable of Interest simulated as normally distributed with mean \( \bar{y}(W) = \beta^T W \) with intercept 0.001 and coefficients for (a) \( W=(0,0) \), (b) \( W=(2,2) \), (c) \( W=(-2,2) \), (d) \( W=(-4,4) \), (e) \( W=(-6,7) \). Here, \( SE_\beta \) is the standard error of the simulated \( \beta \) and \( SE_{IC} \) is the asymptotic standard error calculated from the influence curve.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Semiparametric Method</th>
<th>DR-IPTW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta )</td>
<td>( SE_\beta )</td>
</tr>
<tr>
<td>3.0000</td>
<td>3.0234</td>
<td>0.4771</td>
</tr>
<tr>
<td>1.5000</td>
<td>1.4754</td>
<td>0.5052</td>
</tr>
<tr>
<td>2.0000</td>
<td>1.9975</td>
<td>0.4512</td>
</tr>
<tr>
<td></td>
<td>(a)</td>
<td></td>
</tr>
<tr>
<td>3.0000</td>
<td>3.0321</td>
<td>0.5823</td>
</tr>
<tr>
<td>1.5000</td>
<td>1.4614</td>
<td>0.6405</td>
</tr>
<tr>
<td>2.0000</td>
<td>2.0274</td>
<td>0.8092</td>
</tr>
<tr>
<td></td>
<td>(b)</td>
<td></td>
</tr>
<tr>
<td>3.0000</td>
<td>2.9669</td>
<td>0.5214</td>
</tr>
<tr>
<td>1.5000</td>
<td>1.5362</td>
<td>0.5648</td>
</tr>
<tr>
<td>2.0000</td>
<td>2.0047</td>
<td>0.5790</td>
</tr>
<tr>
<td></td>
<td>(c)</td>
<td></td>
</tr>
<tr>
<td>3.0000</td>
<td>2.9675</td>
<td>0.6902</td>
</tr>
<tr>
<td>1.5000</td>
<td>1.5369</td>
<td>0.8479</td>
</tr>
<tr>
<td>2.0000</td>
<td>1.9876</td>
<td>0.9063</td>
</tr>
<tr>
<td></td>
<td>(d)</td>
<td></td>
</tr>
<tr>
<td>3.0000</td>
<td>2.9723</td>
<td>1.0915</td>
</tr>
<tr>
<td>1.5000</td>
<td>1.5448</td>
<td>1.2595</td>
</tr>
<tr>
<td>2.0000</td>
<td>1.9643</td>
<td>1.5089</td>
</tr>
<tr>
<td></td>
<td>(e)</td>
<td></td>
</tr>
</tbody>
</table>

Under ETA, the estimates of both methods are consistent. Upon increasing the values of \( \beta_W \), the estimates \( \beta_{DR-IPTW} \) become less stable (with increased standard error) until the estimates themselves are no longer consistent with the simulated values. The semiparametric model based method provides fairly consistent estimates even under the largest ETA violation.
Appendix B

Biomarker discovery simulation: additional results

Additional performance measures used to analyze the simulation results from section 2.4 include comparing the average simulated importance value and rank to the truth. Previous performance measures, length of list and AUC, are focused on determining how well the methods rank the “true” variables with respect to all variables. Average Importance measures showcase the ability of each method to not only distinguish “true” variables from decoys, but also properly determine the magnitude of importance accurately. This is completed here only for the case when $\beta = 1, \ldots, 10$.

The average importance value is plotted versus actual value for LM, LASSO(Q), and TMLE methods at each correlation level. This is only relevant for LM, LASSO(Q), and TMLE, which are on the same scale as the simulated importance measures. To allow the incorporation of Random Forest measures, the average importance rank is also analyzed. When $\beta = 1, \ldots, 10$, average rank and importance should lie on the x=y line when plotting average measure or rank by “true” importance value. The difference between the true measure/rank versus the estimated average measure/rank is summarized by calculating the mean squared deviation of the estimated values from the true values. These measures are plotted versus correlation providing a visual representation of the effect of correlation on the overall accuracy of each method.

B.1 Average importance value

The results are shown below for each of the 10 correlation values. Also shown is a plot of the mean square error between the average measured values and the true values.
Figure B.1: Average importance value for each of ten true variables with importance values $= 1, ..., 10$. Plots included for all $\rho = 0, ..., 9$. Only linear regression (LM), LASSO(Q), and TMLE are analyzed since RF values are not necessarily on the same scale as the true level of importance. ($\sigma_Y = 10$)
Figure B.2: Mean square error difference between average importance values and true values at $\rho = 0, ..., .9$.

TMLE measures the actual importance values accurately even at higher correlation values. After correlation of 0.2 LASSO(Q) measures began to deviate from the true values increasing above linear regression estimates at 0.8 and above. Linear regression does approximately the same across all correlations.
B.2 Average importance rank

In order to compare Random Forest with the other methods, the average ranks are plotted along with the mean squared error between the average measured ranked and the true rank.

Figure B.3: Average importance rank for each of ten true variables with actual ranks = 1,...,10. Plots included for all $\rho = 0,...,9$, ranking by measure. ($\sigma_Y = 10$)
Figure B.4: Mean square error difference between average importance ranks and true ranks at $\rho = 0, ..., .9$, when ranking by measure

We can see clearly that the variables with $\beta = 1$ and 2 are harder for all methods to pinpoint and rank accurately, however TMLE performs the best overall correlation values as seen by the MSE plot.
Appendix C

The efficient influence curve

C.1 Repeated measures TMLE - efficient influence curve derivation outlined

Given observed data for a single subject $O \sim (W^*, Y = \{Y_t : t = 1, \ldots, T\}) \sim P_0$, where $W^*$ is the set of $p$ covariates and $Y$ is the set of repeated measures outcome taken over time, we define the TMLE importance effect for a particular $A = W^*_j$ and time, $t$, controlling for confounders $W = W^* - j$ as

$$E[E[Y_t|A = a, W] - E[Y_t|A = 0, W]] = E[m_t(a, V|\beta_t)]$$

We propose the following form for the efficient influence curve for the model parameters $\beta$ of the parameter of interest presented above.

$$D_{h_{opt, Q, G}} = h_{opt}(A, W)\Sigma(A, W)^{-1}(Y - m(A, V|\beta) - \theta(W))$$

with the optimal scaling factor

$$h_{opt} = \left(\frac{d}{d\beta} m(A, V|\beta) - r(W)\right)$$

where $\theta(W) = Q(0, W)$ and

$$r(W) = E\left[\Sigma(A, W)^{-1}|W\right]^{-1} E\left[\Sigma(A, W)^{-1} \frac{d}{d\beta_0} m(A, V|\beta)\right| W]$$

We propose that the multivariate extension of the semiparametric TMLE influence curve \cite{102, 96} is indeed the efficient influence curve for the semiparametric targeted variable importance for repeated measures. Given the following properties (i) it is a score (ii) it is orthogonal to all nuisance scores

- Scores of the form $s(W)$ for tangent space of $p(W)$. 
• Scores of the form $s(A|W)$ for tangent space of $p(A|W)$
• Scores of the form $(Y - \bar{Q}(A,W))\Sigma(A,W)^{-1}(Y - \bar{Q}(A,W))'$ for tangent space of $\Sigma(A,W)$
• Nuisance scores of the form $r(W)\Sigma^{-1}(Y - \bar{Q}(A,W))$ for tangent space of $\theta = \bar{Q}(0,W)$ given fixed $\beta$

Given this, we conclude it is efficient influence curve.

1. It is straightforward to see that the influence curve above is indeed a score in the multivariate normal model space, where the multivariate normal model is defined here as

$$p(Y|A,W) = f_N^\Sigma(Y|A,W)$$

where $f_N$ is the multivariate normal density with scores of the form

$$L(O) = h_{opt}(A,W)\Sigma(A,W)^{-1}(Y - \bar{Q}(A,W))$$

2. It must be shown that the above form is orthogonal to the above nuisance scores

• It can be shown that $D_{h_{opt},Q,g}$ is orthogonal to scores of the form $s(W)$ in that

$$E[D_{h_{opt},Q,g}s(W)] = E[E[D_{h_{opt},Q,g}s(W)]|A,W] = 0$$

• It can be shown that $D_{h_{opt},Q,g}$ is orthogonal to scores of the form $s(A|W)$ in that

$$E[D_{h_{opt},Q,g}s(A|W)] = E[E[D_{h_{opt},Q,g}s(A|W)]|W] = 0$$

• It can be shown that $D_{h_{opt},Q,g}$ is orthogonal to scores of the form

$$s(\Sigma) = (Y - \bar{Q}(A,W))\Sigma(A,W)^{-1}(Y - \bar{Q}(A,W))'$$

under the assumption of a multivariate normal density model, in that we require $E[(Y - \bar{Q}(A,W))^3] = 0$. Given this, it follows

$$E[D_{h_{opt},Q,g}s(\Sigma)] = E[E[D_{h_{opt},Q,g}s(\Sigma)|A,W]] = 0$$

• It follows that $D_{h_{opt},Q,g}$ is orthogonal to scores of the form

$s(\theta) = r(W)\Sigma^{-1}(Y - \bar{Q}(A,W))$ in that $r(W)$ is defined such that

$$E[h_{opt}(A,W)\Sigma(A,W)^{-1}(Y - \bar{Q}(A,W))r(W)\Sigma^{-1}(Y - \bar{Q}(A,W))] = 0$$
C.2 Influence curve for effect parameter of multiplicative semiparametric model

Assuming a multiplicative semiparametric model of the following form

\[ \bar{Q}(A,W) = m^*(A,V|\beta)\theta(W) \]

under the following constraints: \( m^*(A = 0, V|\beta) = 1 \) and \( 0 \leq m(A = 0, V|\beta) \leq 1 \) for all \( A, W \), where \( m^*(A, V|\beta) = e^{-m(A,V|\beta)} \). The effect parameter of interest is defined as \( \Psi(P) = \beta \), where \( \Psi(P_0) = \beta_0 \) is the true parameter defined under the true data generating distribution.

As presented in [102], the orthogonal complement of the nuisance tangent space for estimation of \( \beta \) is found to be of the form.

\[
T_{\nuis}^\perp(P_0) = \{h(A,W) - E_0(h(A,W)|W)\}(Ym^*(A,V|\beta) - \theta(W))
\]

with class of estimating functions

\[
(O, \beta, \theta, \Pi) \to D_{h,Q,g}(O|\beta) \equiv \{h(A,W) - E_g(h(A,W)|W)\}(Ym^*(A,V|\beta) - \theta(W))
\]

for \( \beta_0 \) indexed by \( h(.) \), where \( \theta_0 = E_0(Y|A = 0, W) \), and \( g = P(A|W) \). The corresponding influence curve is defined as

\[
IC_h(O) = -\frac{D_{h,Q_0,g_0}(O|\beta_0)}{d\beta E_0[D_{h,Q_0,g_0}(O|\beta_0)]}
\]

The optimal choice of \( h, h_{opt} \), is that which for any vector \( c \),

\[
c^T \text{Cov}(IC_{h_{opt}})c \leq c^T \text{Cov}(IC_h)c
\]

for all possible \( h(A,W) \), thus providing the most efficient estimating function. For effect parameter \( \beta \) under the presented multiplicative semiparametric model, \( h_{opt} \) is defined as follows.

Define the following terms.

\[
H_0(O|\beta_0) \equiv Ym^*(A,V|\beta_0)
\]

\[
\epsilon(\beta_0) \equiv H_0(O|\beta_0) - E_0(H_0(O|\beta_0)|W)
\]

\[
\epsilon'(\beta_0|A,W) \equiv \frac{d}{d\beta} E_0(\epsilon(\beta)|A,W) \bigg|_{\beta=\beta_0}
\]

\[
\sigma^2(A,W) \equiv E_0(\epsilon^2(\beta_0)|A,W)
\]

where,

\[
h_{opt}(A,W) = \frac{1}{\sigma^2(A,W)} \left\{ \epsilon'(\beta_0|A,W) - \frac{\int \epsilon'(\beta_0|A,W) dP_0(a|W)}{\int \frac{1}{\sigma^2(A,W)} dP_0(a|W)} \right\}
\]
or

\[ h_{opt}(A, W) = \frac{1}{\sigma^2(A, W)} \left\{ \epsilon'(\beta_0 | A, W) - \frac{E[\epsilon'(\beta_0 | A, W)]}{E[\frac{1}{\sigma^2(A, W)} | W]} \right\} \]

This can be rewritten as

\[ h_{opt}(A, W) = \frac{1}{\sigma^2(A, W)} \left\{ Q(A, W) \frac{d}{d\beta} m^*(A, V | \beta_0) - \frac{E[Q(A, W) \frac{d}{d\beta} m^*(A, V | \beta_0)]}{E[\frac{1}{\sigma^2(A, W)} | W]} \right\} \]

Resulting in the general form for the efficient estimating equation

\[ D_h(A, W) = h_{opt}(A, W)(Y m^*(A, V | \beta) - \theta(0)) \]

Note that \( \sigma^2(A, W) = m^*(A, V | \beta_0)^2 Var(Y | A, W) \). Up until this point no distributional assumptions are made about \( Y \). The above efficient estimating function can be simplified further by assuming a form for \( Var(Y | A, W) \). If one assumes a bernoulli or binomial outcome, \( Var(Y | A, W) = Q(A, W)(1 - Q(A, W)) \), and the above estimating function reduces to the estimating function used in chapter 4 for the targeted maximum likelihood update of an initial log-binomial density. If one assumes a Poisson count outcome, \( Var(Y | A, W) = Q(A, W) \), and the above estimating function reduces to the estimating function used in chapter 4 for the targeted maximum likelihood update of an initial Poisson density.
Appendix D

Variance reduction of the influence curve by weighting

In addition to the standard targeting of TMLE, steps can be taken to further increase the efficiency of the estimate. We can weigh the initial fit for \( \bar{Q}(A,W) = E[Y|A,W] \) in such a way that reduces the variance of the influence curve. To determine the correct weights we refer to the form of the variance of the influence curve shown below for the linear model \( m(A,V|\beta) = A\beta \).

\[
\text{Var} \left( (A - E[A|W])(Y - \bar{Q}(A,W)) \right) = (A - E[A|W])^2 \text{Var} \left( (Y - \bar{Q}(A,W)) \right)
\]

Therefore by specifying the weights of \((A - E[A|W])^2\) for our initial fit of \( \bar{Q}(A,W) \) we should be able to effectively increase the efficiency. We show this in practice through a small simulation under increasing levels of ETA violation comparing the efficiency of the TMLE from the following estimation methods for \( \bar{Q}(A,W) \).

1. Weighted \( \bar{Q}(A,W) \) where weights=\((A - E[A|W])^2\)
2. Unweighted \( \bar{Q}(A,W) \)
3. Unadjusted (and unweighted) \( \bar{Q}(A) \)

D.1 Simulation study

Percent of complete ETA violation (i.e. perfect prediction of \( A \) by \( W \)) was set at \( p_w = \{10, 20, 30, 40, 50, 60, 70, 80, 90\} \). For percent \( p_w \) of the total number of observations of \( A \), \( A \) is perfectly predicted by \( W \). For \((1 - p_w)\) percent of the observations \( A \) is not a function of \( W \). Here we simulate \( A,W, \) and \( Y \) as continuous variables. This was completed for 500 simulations with \( n=500 \) and 100 observations using perfect
confounding between $A$ and $W$ over a set fraction of the observations, $p_w$. The data was simulated as follows:

$$W \sim \text{Normal}(2, 1)$$

$$A[W \geq q_1] = 2W$$

$$A[W < q_1] \sim \text{Norm}(5, 1)$$

where, $q_1$ is the $p_w^{th}$ quantile of $W$. The true treatment mechanism model is $A \sim W + I\{W < q_1\} - 1$, and is fitted using standard \texttt{lm} function in R. We add an additional covariate $W_2 \sim \text{Norm}(2A, 1)$, which is correlated with $A$, creating an incorrect model specification for $\bar{Q}(A, W)$. The true $Y$ is simulated as follows where $\beta_1 = 4$, $\beta_2 = 2$, $\beta_3 = 2$:

$$Y = \beta_1 A + \beta_2 W + \beta_3 W_2 + \epsilon$$

$$\epsilon \sim \text{Normal}(0, 1)$$

### D.2 Results

The following tables compare the standard error averaged over the 500 simulations.

<table>
<thead>
<tr>
<th>$p_w$</th>
<th>cor(A,W)</th>
<th>with weights</th>
<th>without weights</th>
<th>percent decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.3042</td>
<td>0.3966</td>
<td>0.3967</td>
<td>0.0257</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4736</td>
<td>0.2689</td>
<td>0.2730</td>
<td>1.5091</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4186</td>
<td>0.3075</td>
<td>0.3087</td>
<td>0.3851</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6005</td>
<td>0.2356</td>
<td>0.2398</td>
<td>1.7828</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5498</td>
<td>0.2721</td>
<td>0.2738</td>
<td>0.6346</td>
</tr>
<tr>
<td>0.6</td>
<td>0.5548</td>
<td>0.3211</td>
<td>0.3324</td>
<td>3.3866</td>
</tr>
<tr>
<td>0.7</td>
<td>0.5733</td>
<td>0.2196</td>
<td>0.2217</td>
<td>0.9594</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6686</td>
<td>0.1612</td>
<td>0.1634</td>
<td>1.2968</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8448</td>
<td>0.0534</td>
<td>0.0616</td>
<td>13.4010</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$p_w$</th>
<th>cor(A,W)</th>
<th>with weights</th>
<th>without weights</th>
<th>percent decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.3310</td>
<td>0.1869</td>
<td>0.1872</td>
<td>0.1647</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3941</td>
<td>0.1783</td>
<td>0.1812</td>
<td>1.5614</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4593</td>
<td>0.1780</td>
<td>0.1796</td>
<td>0.8770</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5357</td>
<td>0.1527</td>
<td>0.1532</td>
<td>0.3380</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5546</td>
<td>0.1540</td>
<td>0.1551</td>
<td>0.6753</td>
</tr>
<tr>
<td>0.6</td>
<td>0.5674</td>
<td>0.1259</td>
<td>0.1260</td>
<td>0.0878</td>
</tr>
<tr>
<td>0.7</td>
<td>0.6381</td>
<td>0.1004</td>
<td>0.1007</td>
<td>0.3099</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6981</td>
<td>0.0776</td>
<td>0.0778</td>
<td>0.2872</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7886</td>
<td>0.0538</td>
<td>0.0551</td>
<td>2.1913</td>
</tr>
</tbody>
</table>
Appendix E

Sample code

E.1 Simple R code example for repeated measures
TMLE

Below is code for implementing TMLE using a simple main effect working model
\( m(A, V|\beta) = A\beta \).

E.1.1 Simple simulated data

```r
library(geepack) # loads package geepack
nobs<-40 # number of subjects
nt<-4 # number of replicates/time points
visit <- rep(1:nt, nobs)
id <- gl(nobs, nt, nt*nobs)
W <- rnorm(nobs,3,1)
A <- runif(nobs, 0, 1)
# creating AR(1) structure
phi <- 1
rhomat <- 0.667 ^ outer(1:nt, 1:nt, function(x, y) abs(x - y))
chol.u <- chol(rhomat)
noise <- as.vector(sapply(1:nobs, function(x) chol.u %*% rnorm(nt,0,1)))
e <- sqrt(phi) * noise
# True Model
y <- 1+3 * W - 2 * A + e
dat <- data.frame(y, id, visit, W, A)
A=dat[,5] # variable of interest
```

E.1.2 TMLE-RM method

Initialization

##Initial fit for Q(A,W) and g(W)
GW<-predict(lm(A~W,data=dat),newdata=dat)
wts1<-(-A-GW)^2 #create weights
fW<-W #Though this can be Q*(0,W) from a data-adaptive fit
AW1<-matrix(A)
dat1 <- data.frame(y, id, visit, fW, AW1)
geeQf<-geeglm(y ~ AW1+fW, id = id, weights=wts1,data = dat1,
family=gaussian,corstr ="ar1")
# The above can also include interactions A:W
#covariance matrix estimate (below)
covY<-cov((matrix(residuals(geeQf),ncol=nt)))
geeQ<-predict(geeQf,newdata=dat1)
bint<-coefficients(geeQf)[2] #initial parameter est.

TMLE update

##apply tMLE update
Scov<-(A-GW) #solve for simple clever covariate
geeUpQ<-geeglm(y~Scov+offset(geeQ)-1, id = id, data = dat,
family=gaussian,corstr ="ar1")#,zcor=zcor1)
bn<-bint+coefficients(geeUpQ) #updated tMLE estimate
geeQn<-predict(geeUpQ)

Covariance estimation

#Calc std error estimates and p-values using influence curve
Scov1<-array(Scov,dim=c(nt,nobs,1))
Vs<-solve(covY)
VScov1<-Scov1
for(vs in 1:nobs) VScov1[,vs,]<-Vs%*%Scov1[,vs,]
VScov11<-array(VScov1,dim=c(nt*nobs,dim(Scov)[2]))

dDh<-(1/(nt*nobs))*t(VScov11)%*%(AW1)
AY<-(matrix(y)-geeQn) #recently switched from t(bout)
Dh<-as.matrix(VScov11)*AY #apply((VAWmat1),2,function(x){x*AY})
IC<-apply(Dh,1,function(x){x%*%solve(dDh)})
spI<-split(1:(nt*nobs),1:(nt))
ICrep<-array(IC,dim=c(nt,(nobs),1))
for(ic in 1:nt) ICrep[ic,,]=IC[spI[[ic]]]

ICrep1<-apply(ICrep,c(2,3),mean)
SigmaAWn<-(1/nobs)*(1/nobs)*t(ICrep1)%*%(ICrep1)

Simple hypothesis test

###Complete simple hypothesis test
SE<-sqrt(diag(CVest))
tests<-bn/sqrt(diag(CVest))
Pval<-2*(1-pnorm(abs(tests)))