Handbook of Big Data

Bühlmann Peter, Drineas Petros, Kane Michael, van der Laan Mark

Targeted Learning for Variable Importance

Publication details
Sherri Rose
Published online on: 18 Feb 2016


PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms.

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
22 Targeted Learning for Variable Importance

Sherri Rose

CONTENTS

22.1 Introduction .................................................. 411
22.2 Literature ......................................................... 412
  22.2.1 Biomarkers and Genomics .................................. 413
    22.2.1.1 Treatment Resistance in HIV ....................... 413
    22.2.1.2 Quantitative Trait Loci Mapping ................. 414
  22.2.2 Complex Data Structures ................................. 415
22.3 Road Map for Estimation and Inference ......................... 416
  22.3.1 Defining the Research Question ......................... 416
    22.3.1.1 Data .................................................. 416
    22.3.1.2 Model and Parameter ............................... 417
    22.3.1.3 Causal Assumptions ................................ 417
  22.3.2 Estimation .................................................. 418
    22.3.2.1 TMLE Algorithm for Quantitative Trait Loci Mapping .... 419
  22.3.3 Inference ................................................... 420
  22.4 Programming .................................................. 421
  22.5 Discussion: Variable Importance and Big Data ................. 422
Acknowledgments .................................................. 423
References .......................................................... 423

22.1 Introduction

Targeted learning methods build machine learning-based estimators of parameters defined as features of the probability distribution of the data, while also providing influence-curve or bootstrap-based confidence internals. The theory offers a general template for creating targeted maximum likelihood estimators for a data structure, nonparametric or semiparametric statistical model, and parameter mapping. The targeted learning framework for efficient estimation was introduced nearly a decade ago [62], following key advances where efficient influence curves were used as estimating functions for effect estimation [27–30,59], and unified loss-based machine learning methods were developed for fitting infinite-dimensional parameters of the probability distribution of the data [54,58]. Targeted maximum likelihood estimation (TMLE) built on the loss-based super learning system such that lower dimensional parameters could be targeted; the remaining bias for the (low-dimensional) target feature of the probability distribution was removed. It also represents the first class of estimators that provide inference in concert with the use of machine learning.
Targeted learning for effect estimation allows for the complete integration of machine learning advances in prediction while providing statistical inference for the target parameter(s) of interest. Targeted maximum likelihood estimators are well-defined (i.e., have one solution) loss-based double robust substitution estimators that respect the global constraints of the statistical model. As noted, the advances in ensembled loss-based learning, namely, (1) defining an infinite-dimensional parameter as a minimizer of the expected loss function and (2) cross-validated estimator selection, were not tailored to lower-dimensional parameters that are often of interest in causal inference and variable importance.

While estimating function approaches may estimate the effect parameters of interest, estimating functions are not loss-based or substitution estimators. Thus, they may have multiple solutions, or produce a solution that falls outside the global constraints of the model. A detailed comparison of the statistical properties of various maximum-likelihood-based estimators, including targeted learning, and estimating-equation-based estimators can be found elsewhere [61]. Therefore, the introduction of targeted learning was a considerable advance, with super learning and TMLE methods labeled “new statistical paradigms” [5].

The targeted learning methodology template involves the following:

1. Starting with a (possibly super-learner-based) initial estimator.
2. Positing a parametric working submodel through this initial estimator, which provides a family of candidate fluctuations of the initial estimator.
3. Holding the initial estimator fixed and choosing the fluctuation that minimizes the selected loss function.
4. Iterating until the updated initial estimator solves the estimating equation defined by the efficient influence curve.

The continued development of targeted learning has led to new solutions for existing problems in many data structures in addition to discoveries in varied applied areas. This has included work in randomized controlled trials [18–20,40,42], parameters defined by a marginal structural model [41], case-control studies [34,35,37,38,50], collaborative TMLE [9,47,55], missing and censored data [36,47], effect modification [25,48], longitudinal data [52,56], networks [53], community-based interventions [51], comparative effectiveness research [13,22], aging [2,32], cancer [25], occupational exposures [3], mental health [12], and HIV [39], as well as many others. A text on targeted learning was also published [61], and contains both introductory and advanced topics.

In this chapter, we focus on the development of targeted learning methods for variable importance problems. Specifically, we are interested in understanding the effect of a list of variables individually on an outcome, such as in a genomic profile, while providing statistical inference that accounts for multiple testing. We discuss a few motivating examples from the published literature on targeted learning for variable importance, and then walk through a road map for estimation and inference in variable importance settings.

22.2 Literature

Some of the earliest works on targeted learning developed methodology for variable importance, and this activity has continued into the present. The articles in this review all consider an experiment where a unit is randomly sampled from a target population and
a list of variables is measured on the unit aside from additional baseline characteristics. An outcome is also measured, and we wish to estimate the effect of each variable in the list on this outcome. Variables in the list may be binary or continuous. In contrast to standard approaches, targeted learning isolates the effect of each variable in the list separately.

22.2.1 Biomarkers and Genomics

There has been considerable work in targeted learning methods for variable importance in biomarker detection and genomics applications. We examine several of these works here, in an effort to highlight the flexibility of the targeted learning framework to address various variable importance research questions while incorporating investigator knowledge. Additional variable importance articles in this area that focus on methodology development, simulations, and data applications include: procedures for semiparametric variable importance and a demonstration in Golub et al. [8], leukemia dataset [49], and sophisticated dimension reduction procedures incorporated into targeted learning for variable importance measures [65].

22.2.1.1 Treatment Resistance in HIV

Bembom et al. [1] studied a list of candidate mutations in an HIV protease enzyme in an effort to identify mutations that reduce the clinical response to antiretroviral treatments. Specifically, they were interested in subjects taking combinations of drugs that contained the protease inhibitor lopinavir who also had a treatment change episode. The ultimate goal was to produce a ranked list of mutations based on their relative contributions to clinical virologic response. In their analysis, it was of particular importance to control for confounding variables that could have been predictive of both the mutation and clinical virologic response through another mechanism aside from the mutation. Without this step, the marginal effect of a mutation in the list may not have reflected an effect due to resistance.

Their study data was derived from two clinical observational data sources: the Stanford Drug Resistance Database and the Kaiser Permanente Medical Care Program, Northern California. They considered a total of 30 candidate protease inhibitor mutations, all of which had been isolated as possibly related to resistance to protease inhibitor drug treatments based on the Stanford HIVdb algorithm. As previously noted, they controlled for a large set of additional baseline variables, including past treatment and prior clinical measurements. The selection of the appropriate variables to control for in the analysis required substantial clinical background, and a complete discussion of these choices, as well as exclusion criteria, can be found in the paper. The list of candidate mutations was treated as a vector of binary variables, where the variable was equal to 1 when the mutation was present in that subject.

The authors implemented targeted learning for variable importance using the deletion/substitution/addition (D/S/A) algorithm [21] to estimate the conditional expectation of the outcome given the list of mutations and other covariates, as well as the conditional distribution of the biomarker given covariates, for each mutation in the list. The estimate of the conditional distribution of the biomarker given covariates was used in the update to the initial estimate of the conditional expectation for the outcome. This is the step where a parametric working submodel is determined to fluctuate the initial estimator. The initial estimator is held fixed while the fluctuation that minimizes the loss function is chosen. This involves using a function of the conditional distribution of the biomarker given covariates as a covariate in the parametric working model. Prior to estimating these components, the authors performed dimension reduction on their set of non-mutation covariates using the unadjusted association of each covariate with the outcome. They retained the top 50 covariates with the smallest $p$-values to include in the D/S/A algorithm.
This targeted maximum likelihood estimator was compared to multiple implementations of g-computation, each using a different initial estimator, in simulations. Their g-computation estimators did not converge to the true known value in their simulations, whereas the targeted learning estimator did. In the HIV data, they implemented an unadjusted marginal estimator, along with g-computation and TMLE. The false discovery rate was used to adjust for multiple testing and control the expected proportion of false positives at 5%.

The targeted learning estimator identified five out of seven known major mutations and three minor mutations. G-computation did not include one of these major mutations, and, along with the unadjusted estimator, also identified two mutations that were not thought to be associated with resistance. The unadjusted estimator identified two major mutations previously thought to be associated with resistance that neither g-computation nor TMLE identified as significant. One of these major mutations, however, lacked variation within the strata of the adjustment covariates (i.e., a positivity violation), and there is in vitro evidence that the other mutation may not play a substantial role in resistance.

22.2.1.2 Quantitative Trait Loci Mapping

Several works have developed targeted learning methods for variable importance in the area of quantitative trait loci mapping. The three Wang et al. articles [63,64,66] present novel semiparametric algorithms for targeted learning, including comparisons to existing approaches, simulations, and three data analyses. The goal of quantitative trait loci mapping is to identify genes that underlie a particular observed trait using genetic markers across the genome. Current standard techniques for quantitative trait loci mapping include interval mapping, composite interval mapping, and multiple interval mapping. Interval mapping methods are fully parametric and require the (unrealistic) assumption that only a single quantitative trait loci accounts for the observed trait. Composite and multiple interval mapping techniques allow for multiple quantitative trait loci, but are also restrictive, relying on parametric assumptions. These parametric models tend to oversimplify the (unknown) underlying genetic mechanisms.

In each work, the authors consider a backcross design to demonstrate the methodology development, although the procedures are general. Backcross is generated by backcrossing the first generation to one of its parental strains, with two possible genotypes at a locus. They propose a semiparametric model that assumes the phenotypic trait changes linearly with the quantitative trait loci. Both TMLEs [63,66] and collaborative TMLEs [63,64] are developed, where collaborative TMLE is an extension of TMLE that tailors the estimation of the nuisance parameter(s) for the target parameter.

The TMLEs in Wang et al. [66] were implemented with adjustment for flanking markers both 10 centimorgans (cM) away and 20 cM away, and compared to univariate regression and composite interval mapping in simulations of 100 markers on 600 backcross subjects, as well as iterative adaptive least absolute shrinkage and selection operator (lasso) and a penalized likelihood approach in additional simulations. In these settings, the TMLE adjusting for flanking markers 10 cM away had the best overall performance. Notably, univariate regression failed to identify the correct quantitative trait loci, composite interval mapping had mixed results, and the TMLE adjusting for flanking markers 20 cM away had a performance between composite interval mapping and the other TMLE. The lasso isolated all the main effects but none of the markers carrying epistatic effects, and effect estimates were biased downward, while the penalized likelihood approach improved on a linear main effects model but did not outperform TMLE. Further simulations results can be found in the paper and supplementary materials. The authors also studied a barely dataset containing 150 doubled haploid lines aiming to identify quantitative trait loci underlying measured
agronomic traits. TMLE was compared to composite interval mapping, with TMLE finding fewer quantitative trait loci.

The collaborative TMLE developed in [64] allowed the authors to use the information in the data to decide which additional markers to include in the updating stage of the estimator. This is the step where, with fewer covariates under consideration, it would not be challenging to simply estimate the conditional distribution of the marker under consideration given the set of other markers. This could then be used to define a parametric working model that codes fluctuations of the initial estimator. However, the number of markers is too numerous, and collaborative TMLE provides an algorithmic procedure to select the most appropriate adjustment set. Readers interested in additional details on theory, implementation, and application of collaborative TMLE are referred to the corresponding works [63,64] as well as other literature [9,47,55,61].

Collaborative TMLE was compared to composite interval mapping in a dataset of 116 female mice that followed their survival time after infection with *Listeria monocytogenes*. They wished to map genetic factors underlying susceptibility to *L. monocytogenes* with the phenotypic trait as time to death in hours. The collaborative TMLE was found to be less noisy than composite interval mapping, with the composite interval mapping identifying many suspicious positives. Another advance in this paper was the proposal of a two-part super learner, which was developed for specific nuances related to the nonnormality of the outcome in this dataset, although we leave the specifics out of our summary [64].

Finally, a genome-wide scan of 119 markers in 633 mice was used to understand the quantitative trait loci involved in wound healing [63]. Estimators included TMLE, collaborative TMLE, and composite interval mapping. The two targeted learning methods were initialized using the D/S/A algorithm. Approximately 400 quantitative trait loci positions were examined in 2 cM increments with *p*-values adjusted using the false discovery rate. The results for TMLE and collaborative TMLE were similar, and they identified the same quantitative trait loci as composite interval mapping, although the targeted learning approaches had improved resolution.

### 22.2.2 Complex Data Structures

Further methodological work for variable importance has been developed in a series of recent works, each with many areas of potential application beyond those considered within the articles. Chambaz et al. [4] also examine the setting where the list of candidate variables has a continuous measure. Their motivating example is that of cancer cells and the continuous *exposure* is DNA copy number. They rank the genes based on the effect of DNA copy number on expression level, controlling for a measure of DNA methylation. Their methods are fully semiparametric and respect a set reference level for the exposure.

Another new contribution to the literature is the manuscript by Sapp et al. [45], which studies interval-censored outcomes. This work provides variable importance estimators for the practical setting where data is not collected continuously, but at specific monitoring times. Along with simulation studies, the authors assessed the importance of a list of covariates among injection drug users in the InC3 cohort for spontaneous viral clearance of HCV. Covariates in the list included gender, age, genotype status, and others.

Finally, Díaz et al. [6] developed techniques for variable importance in longitudinal data with continuous and binary exposures. As the use of longitudinal big data repositories, such as electronic health records, becomes increasingly common, estimating target parameters that respect the longitudinal nature of the data will become more common as well. Here, the authors apply their methods to the ACIT study of severe trauma patients to estimate variable importance measures for the effect of a list of physiological and clinical measurements on death.
22.3 Road Map for Estimation and Inference

The targeted learning framework provides a template for translating variable importance research questions into statistical questions, developing and applying estimators, and assessing uncertainty in the effect measures. We use the motivating examples from the earlier discussed work in quantitative trait loci mapping [63,64,66] to illustrate this road map for estimation and inference, as estimated in [66].

Variable Importance Measures

In this chapter, we focus on a TMLE of the variable importance measure described in Section 22.3 under a semiparametric regression model. This is a flexible definition that can handle both continuous and binary list variables. While we use quantitative trait loci mapping to illustrate the methodology concretely, the applications of these methods are vast. As discussed earlier, the list of variables could involve clinical or epidemiological data [6,45], and these tools also have important implications for testing for possible effect modification (e.g., an intervention modified by the variables in the list) and in controlled randomized trial data [61].

22.3.1 Defining the Research Question

The first step is to define the research question, which includes accurately specifying your data, model, and target parameters. Recall that we are interested in understanding which quantitative trait loci underlie a particular phenotypic trait value. Quantitative trait loci mapping for experimental organisms typically involves crossing two inbred lines that have substantial differences in a trait. The trait is then scored in the segregating progeny. Markers along the genome are genotyped in the segregating progeny, and associations between the trait and the quantitative trait loci are evaluated. The positions and effect sizes of quantitative trait loci are of primary interest. Typical segregating designs include the backcross design, the intercross (F2) design, and the double haploid (DH) design. Backcross is produced by back-crossing the first generation (F1) to one of its parental strains, and there are two possible genotypes, Aa and aa at any locus. For the ease of presentation, as the authors do in the original work [66], we focus most heavily on backcross to demonstrate our method. All the derivations can be readily extended to F2 and other types of experimental crosses.

22.3.1.1 Data

The observed data are given as $n$ i.i.d. realizations of

$$O_i = (Y_i, M_i) \sim P_0 \quad i = 1, \ldots, n$$

Here, $Y$ is the phenotypic trait value and $M$ is a vector of the marker genotypic values, with $i$ indexing the $i$th subject and the 0 subscript, indicating that $P_0$ is the true distribution of the data. The true probability distribution $P_0$ is contained within the set of possible probability distributions that make up the statistical model $\mathcal{M}$.

We introduce the notation $A$ to represent the genotypic value of the quantitative trait loci currently under consideration. $A$ is observed when it lies on a marker, although it can also lie between markers, where it will be unobserved. When $A$ is unobserved, it is imputed using the expected value returned from a multinomial distribution computed from
Defining the Research Question

Data:
n i.i.d. observations of \( O \sim P_0 \).

Model:
Statistical model \( M \) is set of possible probability distributions of \( O \). True \( P_0 \) in \( M \).

Model is statistical model augmented with possible causal assumptions.

Target Parameters:
Parameters \( \Psi(P_0) \) are features of \( P_0 \).
\( \Psi \) maps probability distribution \( P_0 \) into the target parameters.

the locations and genotypes locations of the flanking markers. This is also the approach used in Haley–Knott regression [11]. In this case, the effect is therefore only an estimate of the effect of imputed \( A \) for these locations.

22.3.1.2 Model and Parameter

We use a semiparametric model that assumes that the phenotypic trait changes linearly with the quantitative trait loci. This regression model for the effect of \( A \) at a value \( A = a \) relative to \( A = 0 \), adjusted for the set of other markers, denoted \( M^- \), is

\[
E_0(Y \mid A = a, M^-) - E_0(Y \mid A = 0, M^-) = \beta_0 a \tag{22.1}
\]

Our target parameter is therefore \( \beta_0 \), which is also equivalent to the average marginal effect given by averaging this conditional effect over the distribution of \( M^- \). The target parameter is defined formally as a mapping \( \Psi : M \to \mathbb{R} \) that maps the probability distribution of the data into the (finite dimensional) feature of interest \( \Psi(P_0) = \beta_0 \). Additional discussion of this parameter can be found in earlier literature [49,63].

For our application, the parameter measures the difference in the phenotypic trait outcome \( Y \) when \( A \) shifts from heterozygote to homozygote. This can be understood due to the coding: \( aa \) (homozygote) is given the value 0 and \( Aa \) (heterozygote) is set to 1 in a backcross population and, for an F2 population, the coding is \( (AA,Aa,aa) = (1,0,-1) \).

The linearity assumption discussed above can be seen explicitly in Equation 22.1. It is important to stress that only the effect of our genotypic value \( A \) on the mean outcome of quantitative trait loci \( Y \) is modeled using a parametric form in the semiparametric model. We do not impose any distributional assumptions on the data. We also do not make assumptions about the functional form of all functions \( f(M^-) \) of \( M^- \). We do additionally make the assumption that \( A \) is not a perfect surrogate of \( M^- \) in order for the parameter to be well defined and estimable. Finally, we make the positivity assumption \( 0 < P_0(A = a \mid M^-) < 1 \). The model given in Equation 22.1 is general and may be specified in alternative ways, depending on the target parameter of interest. To include effect modification by markers \( V_j \), we would write: \( a \sum_{j=1}^{\bar{j}} \beta_j V_j \).

22.3.1.3 Causal Assumptions

We do not discuss causal assumptions in detail here for brevity and also given that, in many variable importance settings, these causal assumptions will be violated. In particular, the no unmeasured confounding assumption (also referred to as the randomization assumption) will frequently not hold. Researchers may also not be interested in drawing causal inferences in variable importance settings. However, a case could be made that the exercise of walking through the causal assumptions and articulating the role of endogenous and exogenous
variables in nonparametric structural equations is still useful for a full description of the research question. One could then decide not to augment the statistical model with additional untestable causal assumptions. For a thorough treatment of causal assumptions, nonparametric structural equation models, and directed acyclic graphs, we refer to other literature [23,61].

22.3.2 Estimation

The TMLE procedure builds on the foundation established by maximum likelihood estimation and proceeds in two steps. In the first step, we obtain an ensemble machine learning-based estimator of the data-generating distribution. Super learning is appropriate for this task [32,58,61]. It allows the user to consider multiple algorithms, without the need to select the best algorithm a priori. The super learner returns the best weighted combination of the algorithms considered, selected based on a chosen loss function. Cross-validation is employed to protect against overfitting. The second stage of TMLE fluctuates the initial super learner-based estimator in a submodel focusing on the optimal bias–variance trade-off for the target parameter. This second step can also be thought of as a bias reduction step. We must reduce the bias remaining in the initial estimator for the target parameter, since it was fitted based on a bias–variance trade-off for the data-generating distribution, not the target parameter.

The procedure can also be understood intuitively in the context of our motivating quantitative trait loci example as well. In stage one, the conditional expectation for the phenotypic trait value \( Y \) given the vector \( M \) is not targeted toward our parameter of interest. Here, its bias–variance trade-off is for the overall density. Stage two incorporates the conditional expectation for genotypic value \( A \) of the quantitative trait loci currently being considered to shrink the bias of the conditional expectation of \( Y \), our initial estimate. We now also introduce a subset of \( M \) denoted \( W \) for each \( A \). The vector \( W \) contains the subset of markers that are potential confounders for the effect of genotypic value \( A \) on phenotypic trait \( Y \).

To define our TMLE concretely for this problem, we must begin by calculating the pathwise derivative of our parameter \( \Psi(P) = \beta \) at \( P \) and its corresponding canonical gradient (efficient influence curve) \( D(P,O) \):

\[
D(P,O) = \frac{1}{\sigma^2(A,W)} h(A,W)(Y - Q(A,W))
\]

where

\[
h(A,W) = \frac{d}{d\beta} m(A | \beta) - \frac{E(\frac{d}{d\beta} m(A | \beta) / \sigma^2(A,W) | W)}{E(1/\sigma^2(A,W) | W)}
\]

and \( \sigma^2(A,W) \) is the conditional variance of \( Y \) given \( A \) and \( W \).
The TMLE requires choosing a loss function $L(O, Q)$ for candidate function $Q$ applied to an observation $O$ and then specifying a submodel $\{Q(\epsilon) : \epsilon \} \subset \mathcal{M}$ to fluctuate the initial estimator. Here, we use the squared-error loss function:

$$L(O, Q) = \frac{(Y - Q(A, W))^2}{\sigma^2(A, W)}$$

The submodel $\{Q(\epsilon) : \epsilon \} \subset \mathcal{M}$ through $Q$ at $\epsilon = 0$ is selected such that the linear span of $d/d\epsilon L(Q(\epsilon))$ at $\epsilon = 0$ includes the efficient influence curve in Equation 22.2. The specific steps of the TMLE algorithm for the target parameter $\beta_0$ are enumerated below.

### 22.3.2.1 TMLE Algorithm for Quantitative Trait Loci Mapping

**Estimating $E_0(Y \mid A, M^-) = Q_0(A, M^-)$**. Generate a super learner-based initial estimator that respects the semiparametric model in Equation 22.1 and also takes the form

$$Q_n^0 = \beta_n^0 A + f_n(M^-)$$

We introduce the subscript $n$ to denote estimators and estimates.

**Estimating $E_0(A \mid W) = g_0(W)$.** Recall that we introduced a subset $W$ of $M^-$ for each $A$. Thus, $M^-$ is replaced with $W$ and we can refer to the function $g_0(W) = E_0(A \mid W)$ as a marker confounding mechanism. For the applications considered here, as in Wang et al. [66], the set of markers $W$ are those that lie on the same chromosome as $A$.

However, the choice for $E_0(A \mid W)$, in general, is still a complicated one. The selection of flanking markers to include in the marker confounding mechanism can be further simplified to including only two flanking markers, possibly capturing a good portion of the confounding. But, there is still then the issue of distance for $A$ for the selection of these two flanking markers. Those that are too close to $A$ may be too predictive of $A$, thus failing to isolate the contribution of $A$ when estimating $\beta_0$. On the other hand, if the selected markers are too great a distance from $A$, they may not contribute to reducing bias for the target parameter of interest. Collaborative TMLE, as discussed briefly in our literature review, may also be employed to data-adaptively select the most appropriate adjustment set. We leave further discussion of this issue to other literature [49,63].

**Determine parametric working model to fluctuate initial estimator.** The targeted step uses an estimate $g_n(W)$ of $g_0(W)$ to correct the bias remaining in the initial estimator. This involves defining a so-called clever covariate in a parametric working model coding fluctuations of our initial estimator $Q_n^0$. For our parameter $\beta_0$, the clever covariate is given by

$$h(A, W) = A - g_n(W)$$

the residual of $g_n(W)$, under a condition we describe below.

The clever covariate $h(A, W)$ was defined earlier in Equation 22.3 and derived based on the efficient influence curve in Equation 22.2. When $\sigma^2(A, W)$ is a function of $W$ only, it drops out of the efficient influence curve. We choose to estimate $\sigma^2(A, W)$ with the constant 1, which gives us the simplified clever covariate $h(A, W) = A - g_n(W)$ as above. The estimation of the nuisance parameter $\sigma^2(A, W)$ does not impact the consistency properties of the TMLE, but TMLE will only be efficient if, in addition to estimating $Q_0$ and $g_0$ consistently, $\sigma^2(A, W)$ is in fact only a function of $W$ [49].

**Update $Q_n^0$.** The regression of $Y$ on $h(A, W)$ can be reformulated as

$$Y'' \sim \epsilon h(A, W)$$
where
\[
Y' = Y - Q_0^n(A, M^-)
\]
The estimate of the regression coefficient is denoted \(\epsilon_n\). Our initial estimate \(\beta_0^n\) is updated with \(\epsilon_n\):
\[
\beta_1^n = \beta_0^n + \epsilon_n
\]
Convergence of the algorithm for this target parameter occurs in one step. Since our TMLE is double robust, we have the following properties for this estimator of \(\beta_0\): this TMLE is (1) consistent when either \(Q_0^n\) or \(g_n(W)\) is consistent and (2) is efficient when both \(Q_0^n\) and \(g_n(W)\) are consistent (and \(\sigma^2(A, W)\) is a function of \(W\) only).

### Implementation Summary
The TMLE of the target parameter \(\beta_0\), defined in Equation 22.1, requires an initial fit of \(E_0(Y \mid M)\). Our best fit of \(E_0(Y \mid M)\) will be based on minimizing the chosen error loss function. This initial estimator yields a fit of \(E_0(Y \mid A = 0, M^-)\), which we can map to a first-stage estimator of \(\beta_0\) in our semiparametric model.

We now complete the second-stage targeted updating step. This single update (convergence is achieved in one step) is completed by fitting a coefficient \(\epsilon\) in front of an estimate of \(A - E_0(A \mid W)\) with univariate regression, using the initial estimator of \(E_0(Y \mid A, M^-)\) as an offset. We can show that the TMLE of \(\beta_0\) is \(\beta_0^n + \epsilon_n\).

#### 22.3.3 Inference
The variance \(\sigma^2_n\) for each variable importance measure \(\beta_1^n\) can be calculated using influence-curve-based methods [66], with the variance \(\sigma^2_n\) given by
\[
\sigma^2_n = \frac{\sum_i (Y_i - Q_1^n(A_i, M_i^-))^2 h(A_i, W_i)^2}{(\sum_i A_i h(A_i, W_i))^2}
\]
A detailed discussion of multiple hypothesis testing and inference for variable importance measures is presented in [7]. The authors in the corresponding quantitative trait loci work [66] adjusted for multiple testing using the false discovery rate and interpreted each variable importance measure as a \(W\)-adjusted effect estimate.

In general, variance estimates for TMLE rely on \(\delta\)-method conditions [61,62], and, as such, the asymptotic normal limit distribution of the estimator is characterized by its influence curve. The estimator \(\beta_1^n\) of our target parameter is asymptotically linear; therefore, it behaves as an empirical mean, with bias converging to 0 in sample size faster than a rate of \(1/\sqrt{n}\) and is approximately normally distributed (for sample size \(n\) reasonably large). The variance of the estimator is thus well approximated by the variance of the influence curve divided by \(n\). One can also use the covariance in variable importance questions with a multivariate vector of parameters, where the covariance matrix of the estimator vector is well approximated by the covariance matrix of the corresponding multivariate influence curve divided by \(n\) [61].
22.4 Programming

Practical tools for the implementation of targeted learning methods for variable importance have developed alongside the theoretical and methodological advances. While some work has been done to develop computational tools for targeted learning in proprietary programming languages, such as SAS, the majority of the code has been built in \texttt{R}. TMLE and collaborative TMLE \texttt{R} code specifically tailored to answer quantitative trait loci mapping questions, such as those discussed throughout this chapter, is available in the supplementary material of Wang et al. \cite{Wang}. Each \texttt{R} package discussed in this section is available on The Comprehensive \texttt{R} Archive Network (\url{www.cran.r-project.org}).

Of key importance are the two \texttt{R} packages \texttt{SuperLearner} and \texttt{tmle} \cite{SuperLearner,tmle}. The \texttt{SuperLearner} package, authored by Eric Polley (NCI), is flexible, allowing for the integration of dozens of prespecified potential algorithms as well as a system of wrappers that provide the user with the ability to design their own algorithms, or include newer algorithms not yet added to the package. The package returns multiple useful objects, including the cross-validated predicted values, final predicted values, vector of weights, and fitted objects for each of the included algorithms, among others. The \texttt{tmle} package, authored by Susan Gruber (Reagan-Udall Foundation, Washington, DC), allows for the estimation of both average treatment effects and parameters defined by a marginal structural model in cross-sectional data with a binary intervention. This package also includes the ability to incorporate missingness in the outcome and the intervention, use \texttt{SuperLearner} to estimate the relevant components of the likelihood, and use data with a mediating variable.

The \texttt{multiPIN} package \cite{multiPIN}, authored by Stephan Ritter (Omicia, Inc., Oakland, CA), is designed specifically for variable importance analysis, and estimates an attributable-risk-type parameter using TMLE. This package also allows the use of \texttt{SuperLearner} to estimate nuisance parameters and produces additional estimates using estimating-equation-based estimators and g-computation. The package includes its own internal bootstrapping function to calculate standard errors if this is preferred over the use of influence curves, or influence curves are not valid for the chosen estimator.

Four additional prediction-focused packages are \texttt{casecontrolSL} \cite{casecontrolSL}, \texttt{cvAUC} \cite{cvAUC}, \texttt{subsemble} \cite{subsemble}, and \texttt{h2oEnsemble} \cite{h2oEnsemble}, all primarily authored by Erin LeDell (Berkeley). The \texttt{casecontrolSL} package relies on \texttt{SuperLearner} and performs subsampling in a case-control design with inverse-probability-of-censoring-weighting, which may be particularly useful in settings with rare outcomes. The \texttt{cvAUC} package is a tool kit to evaluate area under the ROC curve estimators when using cross-validation. The \texttt{subsemble} package was developed based on a new approach \cite{subsemble} to ensembling that fits each algorithm on a subset of the data and combines these fits using cross-validation. This technique can be used in datasets of all size, but has been demonstrated to be particularly useful in smaller datasets.
A new implementation of super learner can be found in the Java-based h2oEnsemble package, which was designed for big data. The package uses the H2O R interface to run super learning in R with a selection of prespecified algorithms.

Another TMLE package is ltmle [46], primarily authored by Joshua Schwab (Berkeley). This package mainly focuses on parameters in longitudinal data structures, including the treatment-specific mean outcome and parameters defined by a marginal structural model. The package returns estimates for TMLE, g-computation, and estimating-equation-based estimators.

### 22.5 Discussion: Variable Importance and Big Data

While the development of targeted learning for variable importance has demonstrated promise, its potential has yet to be fully realized. The data we are collecting in biology, social sciences, health care, medicine, business analytics, and ecology, among others, continue to grow in both dimensions (n and p), and are frequently observational in nature [31,43]. Statisticians are armed with a unique set of rigorous and practical tools to tackle these challenges. To face this growth in data going forward, targeted learning provides a framework for incorporating advances in machine learning and TMLE for problems of variable importance [33,60].

It is always important to remember that sophisticated statistical methods will never be able to overcome weak or problematic big data. Misclassification, missingness, and unmeasured confounding are frequently found in these new streams of data. A thorough understanding of the data and associated research questions, often only ascertained by working in interdisciplinary teams, is required before leaping toward analysis. This will not change as technologies continue to advance.

To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.

R.A. Fisher
1938

We will also need to address the increasing computational challenges presented by these data. For example, online targeted learning [57] is a new proposed method for data that arrives sequentially, another common feature of big data applications. Advances will not only be found in statistical theory and methodology development, however. Existing approaches to merge data systems and statistics currently mainly use database systems to serve, for example, R requests in the background. Movement toward integrated native big data systems may be a key component in the adoption of rigorous targeted learning tools for variable importance in massive datasets.

Targeted learning is one of many new statistical innovations that are poised for further theoretical and methodological development in this new era of big data, inspired by these real-world challenges. Advances in dimension-reduction for imaging analyses, for example, will improve our ability to use features of these images as covariates in variable importance
analyses, and also move us toward the ability to estimate variable importance measures of a list of images. The future of statistical and scientific discovery with big data is bright, as we look forward to the creation of automated big data machines that incorporate investigator knowledge, are statistically sound, and can handle the computational burden of our data.

Acknowledgments

The author acknowledges funding from the University of Utah fund P0 163947.

References


