Integrating machine learning into causal inference: the Targeted Maximum Likelihood Estimation approach

Scott Grey, PhD

April 12, 2016

Overview

1. Background on the development of targeted learning
2. Theory of TMLE
3. Application of TMLE in R
4. Extensions of TMLE

This presentation, the data (with documentation) and R code is available at: https://github.com/sfgrey/Super-Learner-Presentation.git

Background

Background

"Essentially, all models are wrong, but some are useful"
- George Box, 1979

Mantra of statisticians regarding the development of statistical models for many years

In the 1990s an awareness developed among statisticians (Breiman, Harrell) that this approach was wrong

- Parametric model assumptions rarely met
- Large number of variables makes it difficult to correctly specify a model

Simultaneously, computer scientists and some statisticians developed the machine learning field to address the limitations of parametric models

Targeted learning

Combines advanced machine learning with efficient semiparametric estimation to provide a framework for answering causal questions from data

- Developed by Mark van der Laan and his research group at UC Berkeley
- Started with the seminal 2006 article on targeted maximum likelihood estimation

Central motivation is the belief that statisticians treat estimation as *Art* not **Science**

- This results in misspecified models that are data-adaptively selected, but this part of the estimation procedure is not accounted for in the variance

Estimation is a Science, *Not an Art*

Specific definitions required

1. **Data**: realizations of random variables with a probability distribution
2. **Model**: actual knowledge about the data generating probability distribution
3. **Target Parameter**: a feature of the data generating probability distribution
4. **Estimator**: an a priori-specified algorithm, benchmarked by a dissimilarity-measure (e.g., MSE) w.r.t. target parameter

Theory of TMLE

Data

Random variable O, observed n times, defined in a simple case as $O = (A, W, Y) \sim P_0$ if we are without common issues such as missingness and censoring

- A: exposure or treatment
- W: vector of covariates
- Y: outcome
- P_0: the true probability distribution

This data structure makes for an effective example, but data structures found in practice are much more complicated

Model

General case: Observe n i.i.d. copies of random variable O with probability distribution P_0

The data-generating distribution P_0 is also known to be an element of a statistical model M : $P_0 \in M$

A **statistical** model M is the set of possible probability distributions for P_0; it is a collection of probability distributions

If all we know is that we have n i.i.d. copies of O, this can be our statistical model, which we call a non-parametric statistical model

Model

A statistical model can be augmented with additional non-testable assumptions, allowing one to enrich the interpretation of $\Psi(P_0)$; This does not change the **statistical model**

We refer to the statistical model augmented with a possibly additional assumptions as a **causal model**

In the Neyman-Rubin causal inference framework, assumptions include

- $(A \perp Y_a|W)$; randomization
- Stable unit treatment value assumption (SUTVA); no interference between subjects and consistency assumption
- Positivity; each possible exposure level of A occurs with some positive probability within each stratum of W

A (very) brief review of the Neyman-Rubin causal inference framework

Potential outcomes: every individual i has a different potential outcome depending on their treatment "assignment"

- $Y_i\left(A=1\right)$ and $Y_i\left(A=0\right)$
- The "fundamental problem with causal inference" is that we can only observe one of these potential outcomes
- If we randomly assign i to receive A, then the groups will be equivalent and causal inference can be inferred:

$$E\left(Y_{i1}|A_i=1\right)-E\left(Y_{i0}|A_i=0\right)$$

- This framework has been extended to observational data through propensity score matching

Target Parameters

Define the parameter of the probability distribution P as function of P : $\Psi(P)$

In a causal inference framework, a target parameter for the effect of A could be

$$\Psi(P_0)_{RD} = E_{W,0}\left[E_0\left(Y|A=1,W\right) - E_0\left(Y|A=0,W\right)\right]$$

Or, if we wish to use a ratio instead of a difference:

$$\Psi(P_0)_{OR} = E_{W,0}\left[O\left[Y|A=1,W\right]/O\left[Y|A=0,W\right]\right]$$

Where $O[.] = E[.]/1 - E[.]$

Estimators

The target parameter $\Psi(P_0)$ depends on P_0 through the conditional mean $\overline{Q}_0(A,W) = E_0(Y|A,W)$ and the marginal distribution $Q_{W,0}$ of W; or

$$\overline{Q}(A,W) = E(Y|A,W) / \overline{Q}(W) = E(Y|W)$$

Where $\overline{Q}$ is an **estimator** of $\overline{Q}_0(A,W)$, shortened to $\overline{Q}_0$

An **estimator** is an algorithm that can be applied to any empirical distribution to provide a mapping from the empirical distribution to the parameter space

- But which algorithm?

Effect Estimation vs. Prediction

Both **effect** and **prediction** research questions are inherently *estimation* questions, but they are distinct in their goals

- **Prediction**: Interested in generating a function to input covariates and predict a value for the outcome: $E_0\left(Y|W\right)$
- **Effect**: Interested in estimating the true effect of exposure on outcome adjusted for covariates, $\Psi\left(P_0\right)$, the **targeted estimand**
- Targeted maximum likelihood estimation (TMLE), is an iterative procedure that updates an initial (super learner) estimate of the relevant part $\overline{Q}_0$ of the data generating distribution P_0
- See second presentation given on April 14 to the Ann Arbor R User Group

Some effect estimators

Maximum-likelihood-based substitution estimators will be of the type

$$\Psi\left(Q_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{\overline{Q}_n\left(A=1, W_i\right) - \overline{Q}_n\left(A=0, W_i\right)\right\}$$

where this estimate is obtained by plugging in $Q_n = \left(\overline{Q}_n, Q_{W,n}\right)$ into the mapping Ψ

Estimating-equation-based function is a function of the data O and the parameter of interest. If $D(\psi)(O)$ is an estimating function, then $\Psi\left(Q_n\right)$ is a solution that satisfies:

$$0 = \sum_{i=1}^{n} D\left(\psi\right)\left(O_i\right)$$

Targeted Maximum Likelihood Estimation

It is an iterative procedure that:

1. Generates an initial (super learner) estimate of the relevant part $\overline{Q}_0$ of the data generating distribution P_0, noted as $\overline{Q}_n^0$
2. Updates an initial estimate, possibly using an estimate of a nuisance parameter, g_0

Produces a well-defined, unbiased, efficient **substitution estimator** of target a parameter Ψ
- Is semi-parametric, no need to make assumptions about P_0
- Uses machine learning techniques to get initial estimates

TMLE steps

Step 1: Use the super learner procedure to generate an initial estimate $\overline{Q}_n^0$

Step 2: Estimate g_0, the conditional distribution of A given W (a propensity score, called a nuisance parameter if A is randomized), denoted g_{n}

Step 3: Construct a "clever covariate" that will be used to fluctuate the initial estimate

$$H_n^*(A,W) \equiv \left(\frac{I(A=1)}{g_n(1|W)}\right) - \left(\frac{I(A=0)}{g_n(0|W)}\right)$$

TMLE steps

Step 4: Use maximum likelihood to obtain ε_n, the estimated coefficient of $H_n^*(A,W)$ in:

$$\text{logit}\,\overline{Q}_n^1(A,W) = \text{logit}\,\overline{Q}_n^0(A,W) + \varepsilon_n H_n^*(A,W)$$

Step 5: plug-in the substitution estimator using updated estimates $\overline{Q}_n^1(A=1,W_i)$ and $\overline{Q}_n^0(A=1,W_i)$ and the empirical distribution of W into formula:

$$\psi_{TMLE,n} =$$
$$\Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{\overline{Q}_n^1(A=1,W_i) - \overline{Q}_n^1(A=0,W_i)\right\}$$

Step 6: Inference using an infuence curve (IC)

The Infuence Curve (IC)

IC is a function that describes estimator behavior under slight perturbations of the empirical distribution.

IC has mean 0 at the true parameter value, so it can be used as an estimating equation:

$$IC_n\left(O_i\right) = H_n^*\left(A,W\right)\left(Y - \overline{Q}_n^1\left(A_i,W_i\right)\right)$$
$$+\overline{Q}_n^1\left(A=1,W_i\right) - \overline{Q}_n^1\left(A=0,W_i\right) - \psi_{TMLE,n}$$

The empirical mean of IC for regular asymptotically linear (RAL) estimator provides a linear approximation of estimator. Thus, VAR(IC) provides asymptotic variance of estimator

The Infuence Curve (IC)

We then calculate the sample variance of the estimated influence curve values:

$$S^2\left(IC_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(IC_n\left(o_i\right) - \bar{I}\overline{C}_n\right)^2$$

After which standard errors, confidence intervals and p-values can be calculated in the standard fashion

Also possible to utilize bootstrapping to calculate standard errors, but computationally expensive

Application of TMLE in R

TMLE package in R

Created by Susan Gruber in collaboration with Mark van der Laan

```
library(tmle)

effA1 <- tmle(Y=Y,
              A=A,
              W=W,
              Q.SL.library = c(),
              g.SL.library = c(),
              family = "binomial",
              cvQinit = TRUE,
              verbose = TRUE)
```

TMLE Arguments

- `Y` - The outcome
- `A` - Binary treatment indicator, 1 treatment, 0 control
- `W` - A matrix of covariates
- `Q.SL.library` - a character vector of prediction algorithms for initial Q
- `g.SL.library` - a character vector of prediction algorithms for g
- `family` - 'gaussian' or 'binomial' to describe the error distribution
- `cvQinit` - estimates cross-validated predicted values for initial Q, if TRUE

Additional TMLE Arguments

- `id` - Subject or group identifier if observations are related. Causes corrected standard errors to be calculated
- `verbose` - helpful to set this to `TRUE` to see the progress of the estimation
- `Delta` - Indicator of missing outcome or treatment assignment
- `Z` - Binary mediating variable

Using super learner with TMLE

Permits the use of multiple machine learning algorithms to generate the initial estimate of Q

- Should use cross validation as SL will easily overfit
- The better the initial estimate of Q, the easier it is to calculate the updated estimates

Currently, SL should not be used to estimate g

- Often creates violations of the positivity assumption
- Best to use standard GLM or LASSO

TMLE example

Does placing a right heart catheter change 30 day mortality?

The ARF dataset has 2490 patients admitted to an ICU and 47 variables including:

- **Demographic characteristics**, including age, gender and race
- **Patient medical history**, 12 variables for medicial conditions: MI, COPD, stroke, cancer, etc.
- **Current condition variables**, that provide information about the patient's current health status: diagnostic scales, vital statistics
- **RHC status**, The placement of a right heart catheter (RHC) is controversial as there is no empirical evidence that benefits patients

Preparing data for TMLE

Only works with numeric matrices; can be specified in-line, i.e. `Y= dataset$Y`

Data must be pre-processed:

- Can only handle missingness in the outcome `Y`, `X` must be removed/imputed
- Continuous variables must be appropriately re-scaled
- Categorical variables must be appropriately dummy coded

Preparing data for TMLE

```
# Impute missing X values #
library("VIM")

# Scale cont vars #
library(arm)
cont <- c("age","edu","das2d3pc","aps1","scoma1","meanbp1","wblc1","hrt1",
          "resp1","temp1","pafi1","alb1","hema1","bili1","crea1","sod1",
          "pot1","paco21","ph1","wtkilo1")
arf[,cont] <- data.frame(apply(arf[cont], 2, function(x)
  {x <- rescale(x, "full")})); rm(cont) # standardizes by centering and
                                        # dividing by 2 sd

# Create dummy vars #
arf$rhc <- ifelse(arf$swang1=="RHC",1,0)
arf$white <- ifelse(arf$race=="white",1,0)
arf$swang1 <- arf$race <- NULL
```

Run TMLE

```
system.time({
  eff <- tmle(Y=arf$death,
              A=arf$rhc,
              W=arf[1:44],
              Q.SL.library = c("SL.gam","SL.knn","SL.step"),
              g.SL.library = c("SL.glmnet"),
              family = "binomial",
              cvQinit = TRUE, verbose = TRUE)
  })[[3]] # Obtain computation time
```

TMLE results

Run time on laptop: 15.43 min.

```
print(eff)
```

Odds Ratio
Parameter Estimate: 1.207
p-value: 0.063956
95% Conf Interval: (0.98914, 1.4728)

Interpretation: Right heart catheterization does not appear to change 30 day mortality

- Note that causal assumptions require non-testable assumptions previously outlined

Advantages of the TMLE approach

Incorporates machine learning so the limitations of parametric methods are avoided

Is “double robust” meaning that estimates are asymptotically unbiased if either the initial SL estimate or the propensity score is correctly specified

- As a result, TMLE works very well with rare outcomes

Can be extended to a variety of situations

- **Missing outcomes**: can account for missing outcomes in a MAR way
- **Controlled direct effect estimation**: can account for mediators in the relationship between A and Y
- **Marginal structural models**: flexible framework for handling issues of time-dependent confounding

Extensions of TMLE being developed in new R packages

- **ltmle**: Longitudinal TMLE permits the evaluation of interventions over time using a marginal structural model
- **multiPIM**: variable importance analysis that estimates an attributable-risk-type parameter
- **tmle.npvi**: permits modeling an intervention variable that is a continuous variable
- **CTMLE**: collaborative TMLE accounts for the relationship between Q and g

Thank you!

References

- van der Laan, M.J. and Rubin, D. (2006), Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1). http://www.bepress.com/ijb/vol2/iss1/11/
- van der Laan, M.J. and Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, Berlin Heidelberg New York, 2011. http://www.targetedlearningbook.com/
- M.J. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Stat Appl Genet Mol*, 6(1): Article 25, 2007.
- Gruber, S. and van der Laan, M.J. (2012), tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software*, 51(13), 1-35. http://www.jstatsoft.org/v51/i13/
- Sekhon, Jasjeet (2007). "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods" (PDF). *The Oxford Handbook of Political Methodology*. http://sekhon.berkeley.edu/papers/SekhonOxfordHandbook.pdf
- F.R. Hampel. “The influence curve and its role in robust estimation” *JASA*, 69(346): 383-393, 1974.

Software and online resources

- *tmle: Targeted Maximum Likelihood Estimation* https://cran.r-project.org/web/packages/tmle/index.html
- *SuperLearner: Super Learner Prediction* https://cran.r-project.org/web/packages/SuperLearner/index.html
- M. Petersen and L. Balzer. *Introduction to Causal Inference*. UC Berkeley, August 2014. http://www.ucbbiostat.com/
- This presentation, the data (with documentation) and R code is available at: https://github.com/sfgrey/Super-Learner-Presentation.git