# Targeted Minimum Loss-based Estimation (TMLE) for Causal Inference

Helene Charlotte Wiese Rytgaard (hely@sund.ku.dk)

Thomas Alexander Gerds (tag@biostat.ku.dk)

Anders Munch (a.munch@sund.ku.dk)

Day 1, Lecture 1

# Introduction: The roadmap of targeted learning

# Overview: The roadmap of targeted learning

**Theoretical angle** The roadmap of targeted learning

- data as a random variable having a probability distribution, scientific knowledge represented by a large statistical model, a statistical target parameter representing an answer to the question of interest.

**Applied angle** The roadmap of targeted learning / causal inference

- translation from real-world data applications to a mathematical and statistical formulation of the relevant estimation problem.
- statistical analysis tailored towards answering that question.

Opposed to choosing a model for the data-generating process and using that model to answer all questions.

# The roadmap (theoretical)

1. Data is a random variable $O$ with a probability distribution $P_0$
2. $P_0$ belongs to a statistical model $\mathcal{M}$
3. Our target is a parameter $\Psi : \mathcal{M} \to \mathbb{R}$
4. Construct estimator $\hat{P}_n$ for (relevant part of) $P_0$ and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$
5. Quantify uncertainty for the estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$

# The roadmap (theoretical)

1. Data is a random variable $O$ with a probability distribution $P_0$

$O_1, \ldots, O_n \overset{iid}{\sim} P_0$

$O_i$ is the observation for individual $i$ of the dataset

For example, $O$ consists of

- Covariates: $X \in \mathcal{X} \subseteq \mathbb{R}^d$
- Exposure/treatment: $A \in \{0, 1\}$
- Outcome: $Y \in \{0, 1\}$ or $Y \in \mathbb{R}$

# The roadmap (theoretical)

1. Data is a random variable $O$ with a probability distribution $P_0$

$O_1, \ldots, O_n \overset{iid}{\sim} P_0$

$O_i$ is the observation for individual $i$ of the dataset

For example, $O$ consists of

- Covariates: $X \in \mathcal{X} \subseteq \mathbb{R}^d$
- Exposure/treatment: $A \in \{0, 1\}$
- Outcome: $Y \in \{0, 1\}$ or $Y \in \mathbb{R}$

This is the data structure we stick to for now.

# The roadmap (theoretical)

2. $P_0$ belongs to a statistical model $\mathcal{M}$

What do we know about the probability distribution of the data?

The statistical model $\mathcal{M}$ is the set of all possible probability distributions for our observed data.

Limited statistical knowledge? $\Rightarrow$ $\mathcal{M}$ should be large to reflect that.

# The roadmap (theoretical)

Consider a parametric[1] model for the distribution of $Y \in \{0, 1\}$ given $X \in \mathbb{R}^d$ and $A \in \{0, 1\}$:

[1] i.e., distribution can be characterized by a finite number of parameters.

# The roadmap (theoretical)

Consider a parametric[1] model for the distribution of $Y \in \{0, 1\}$ given $X \in \mathbb{R}^d$ and $A \in \{0, 1\}$:

$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \alpha_0 + \alpha_A A + \alpha_X^\top X \tag{M1}$$

- assumption of convenience?

[1] i.e., distribution can be characterized by a finite number of parameters.

# The roadmap (theoretical)

Consider a parametric[1] model for the distribution of $Y \in \{0,1\}$ given $X \in \mathbb{R}^d$ and $A \in \{0,1\}$:

$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \alpha_0 + \alpha_A A + \alpha_X^\top X \tag{M1}$$

- assumption of convenience?

Another parametric model could be

$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \gamma_0 + \gamma_A A + \gamma_X^\top X + \gamma_{A,X}^\top AX \tag{M2}$$

- (M1) and (M2) cannot be true at the same time (except if $\gamma_{A,X} = 0$).

[1] i.e., distribution can be characterized by a finite number of parameters.

# The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2, 2] \times \{0, 1\} \times \{0, 1\}$
- True model is

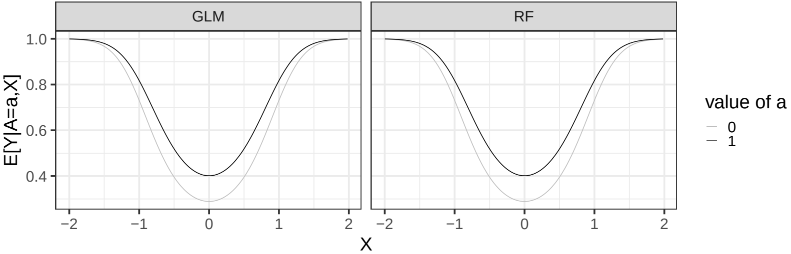
$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \beta_0 + \beta_A A + \beta_X X^2$$

# The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2, 2] \times \{0, 1\} \times \{0, 1\}$
- True model is

$$\operatorname{logit} \mathbb{E}[\,Y \mid A, X\,] = \beta_0 + \beta_A A + \beta_X X^2$$

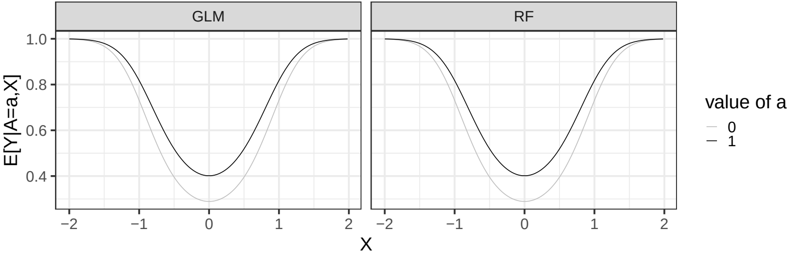



[ Truth shown with solid lines ]

# The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2, 2] \times \{0, 1\} \times \{0, 1\}$
- True model is
$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \beta_0 + \beta_A A + \beta_X X^2$$




[ Truth shown with solid lines ]

GLM: $\operatorname{logit} \mathbb{E}[Y \mid A, X] = \alpha_0 + \alpha_A A + \alpha_X X$

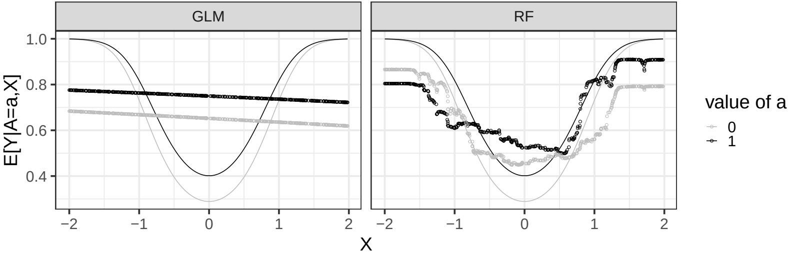
RF: Random forest (untuned)

# The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2, 2] \times \{0, 1\} \times \{0, 1\}$
- True model is

$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \beta_0 + \beta_A A + \beta_X X^2$$




[ Truth shown with solid lines ]

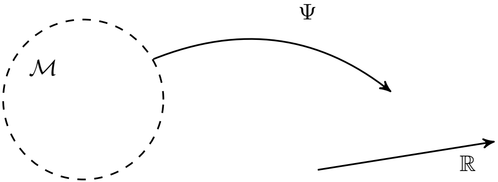
GLM: $\operatorname{logit} \mathbb{E}[Y \mid A, X] = \alpha_0 + \alpha_A A + \alpha_X X$

RF: Random forest (untuned)

# The roadmap (theoretical)

3. Our target is a parameter (a functional) $\Psi : \mathcal{M} \to \mathbb{R}$

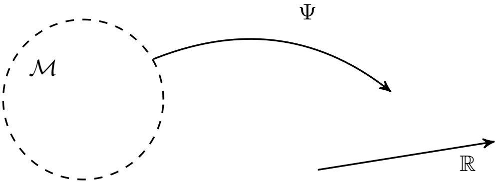
What are we trying to learn from the data?



# The roadmap (theoretical)

3. Our target is a parameter (a functional) $\Psi : \mathcal{M} \to \mathbb{R}$

What are we trying to learn from the data?




EXAMPLE: Average treatment effect (ATE)

- $O = (X, A, Y) \in \mathbb{R}^d \times \{0, 1\} \times \{0, 1\}$
- The ATE is defined for $P \in \mathcal{M}$ as

$$\Psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y \mid A = 1, X] - \mathbb{E}_P[Y \mid A = 0, X]]$$

# The roadmap (theoretical)

**EXAMPLE: Average treatment effect (ATE)**

- $O = (X, A, Y) \in \mathbb{R}^d \times \{0,1\} \times \{0,1\}$
- The ATE is defined for $P \in \mathcal{M}$ as

$$\Psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y \mid A = 1, X] - \mathbb{E}_P[Y \mid A = 0, X]]$$

The ATE can also be written, for $P \in \mathcal{M}$:

$$\Psi(P) = \tilde{\Psi}(\mu_X, f) = \int_{\mathbb{R}} \big(f(1,x) - f(0,x)\big) d\mu_X(x),$$

where $f(a,x) := \mathbb{E}_P[Y \mid A = a, X = x]$ and $\mu_X$ is the marginal distribution of $X$

$f, \mu_X$ are called *nuisance parameters*

# The roadmap (theoretical)

This suggests a straightforward two-step estimation strategy:

1. estimate the nuisance parameters
2. plug estimates into the expression for the target parameter

A straightforward estimate of the ATE would be

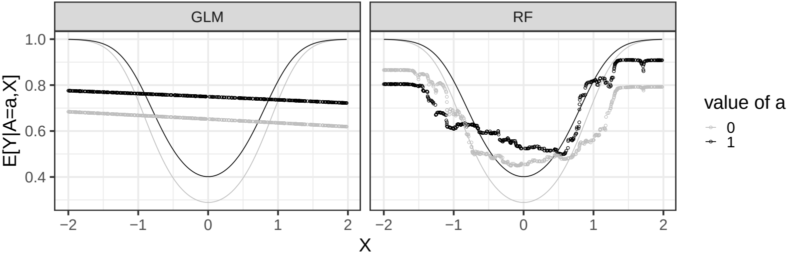
$$\hat{\psi}_n^{\mathrm{ATE}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{f}_n(1, X_i) - \hat{f}_n(0, X_i) \right\}$$

where $\hat{f}_n$ denotes some estimator for $f(a, x) = \mathbb{E}_P[\, Y \mid A = a, X = x \,]$

→ logistic regression, random forest, neural network, lasso, ...

# The roadmap (theoretical)

In the previous example we had two different estimators for $f(a,x) = \mathbb{E}_P[Y \mid A = a, X = x]$




$$\hat{\psi}_n^{\text{ATE,GLM}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_n^{\text{GLM}}(1, X_i) - \hat{f}_n^{\text{GLM}}(0, X_i)\right\} = 0.0975$$

$$\hat{\psi}_n^{\text{ATE,RF}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_n^{\text{RF}}(1, X_i) - \hat{f}_n^{\text{RF}}(0, X_i)\right\} = 0.0551$$

# The roadmap (theoretical)

Contrast this to fitting a logistic regression model

$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \beta_0 + \beta_A A + \beta_X^\top X \tag{1}$$

to estimate the conditional odds ratio $\exp(\beta_A)$

- valid interpretation when model is correct
- statistical inference when model is correct
- *conditional* interpretation (crude and adjusted models target different parameters)

# The roadmap (theoretical)

Contrast this to fitting a logistic regression model

$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \beta_0 + \beta_A A + \beta_X^\top X \tag{1}$$

to estimate the conditional odds ratio $\exp(\beta_A)$

- valid interpretation when model is correct
- statistical inference when model is correct
- *conditional* interpretation (crude and adjusted models target different parameters)

... and: (1) must be a priori specified (the same data cannot be used for testing and for fitting the final model).

# The roadmap (theoretical)

4. Construct estimator $\hat{P}_n$ for (relevant part of) $P_0$ and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$

A priori specified algorithm that maps the data to an estimate in the parameter space for the target parameter

- a pre-specified logistic regression model
- a random forest
- cross-validated selection between a pre-specified library of different models ("super learning")

# The roadmap (theoretical)

4. Construct estimator $\hat{P}_n$ for (relevant part of) $P_0$ and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$

A priori specified algorithm that maps the data to an estimate in the parameter space for the target parameter

- a pre-specified logistic regression model
- a random forest
- cross-validated selection between a pre-specified library of different models ("super learning")

+ "targeting" to yield the an estimator with improved properties

# The roadmap (theoretical)

4. Construct estimator $\hat{P}_n$ for (relevant part of) $P_0$ and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$

A priori specified algorithm that maps the data to an estimate in the parameter space for the target parameter

"Initial estimation":

- a pre-specified logistic regression model
- a random forest
- cross-validated selection between a pre-specified library of different models ("super learning")

+ "targeting" to yield the an estimator with improved properties

# The roadmap (theoretical)

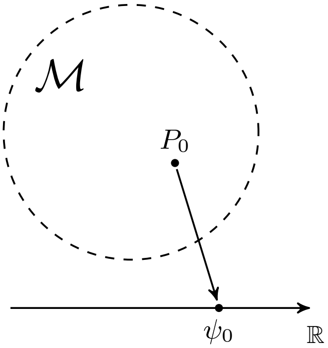
## Estimation paradigm

1. Minimal amount of parametric assumptions on $P_0$ (goal)
2. Go after optimal/efficient estimation of $\psi_0 = \Psi(P_0)$

# The roadmap (theoretical)

## Estimation paradigm

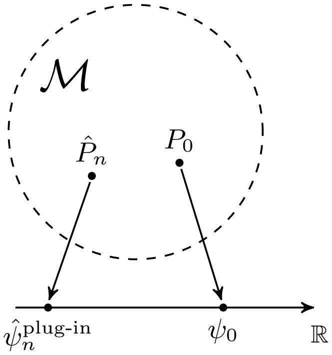
1. Minimal amount of parametric assumptions on $P_0$ (goal)
2. Go after optimal/efficient estimation of $\psi_0 = \Psi(P_0)$



# The roadmap (theoretical)

## Estimation paradigm

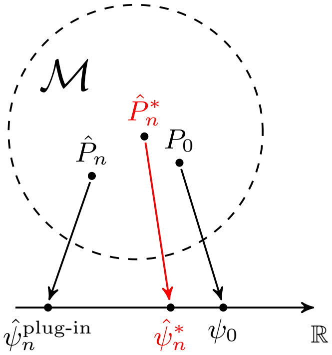
1. Minimal amount of parametric assumptions on $P_0$ (goal)
2. Go after optimal/efficient estimation of $\psi_0 = \Psi(P_0)$



# The roadmap (theoretical)

## Estimation paradigm

1. Minimal amount of parametric assumptions on $P_0$ (goal)
2. Go after optimal/efficient estimation of $\psi_0 = \Psi(P_0)$




Tools from semiparametric efficiency theory and empirical process theory tell us how to construct an optimal estimator

# The roadmap (theoretical)

5. Quantify uncertainty for the estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$

If we repeat the experiment of drawing $n$ observations we would every time end up with a different realization of our estimator.

Across the repetitions, the estimator has a sampling distribution that we wish to quantify.

Under some conditions, we may use the asymptotic distribution

$$\hat{\psi}_n \overset{as}{\sim} N(\psi_0, \sigma^2/n)$$

to provide statistical inference.

# Overview

**Theoretical angle** The roadmap of targeted learning

- data as a random variable having a probability distribution, scientific knowledge represented by a large statistical model, a statistical target parameter representing an answer to the question of interest.

**Applied angle** The roadmap of targeted learning / causal inference

- translation from real-world data applications to a mathematical and statistical formulation of the relevant estimation problem.
- statistical analysis tailored towards answering that question.

Opposed to choosing a model for the data-generating process and using that model to answer all questions.

# The roadmap (applied)

1. Observed data
2. Causal model
3. Causal question and target causal estimand
4. Identifiability
5. Stating the statistical estimation problem
6. Estimate
7. Interpret results

# The roadmap (applied)

1. Observed data
2. Causal model
3. Causal question and target causal estimand
4. Identifiability
5. Stating the statistical estimation problem
6. Estimate
7. Interpret results

... putting things into the right boxes.

... make the statistical analysis about the targeted scientific question (and not the other way around).

... focus on statistical parameters that have a meaningful interpretation.

# The roadmap (applied)

A formal causal framework can help us[2]

- designing a statistical analysis that come as close as possible to answering scientific/causal questions.
- understand how far away from a causal conclusion we may be.

[2] The output of the analysis is not causal just because we use causal inference methods.

# The roadmap (applied)

A formal causal framework can help us[2]

- designing a statistical analysis that come as close as possible to answering scientific/causal questions.
- understand how far away from a causal conclusion we may be.

Clearly defining what an EFFECT is and WHAT effect we are interested in

[2] The output of the analysis is not causal just because we use causal inference methods.

# The roadmap (applied)

A formal causal framework can help us[2]

- designing a statistical analysis that come as close as possible to answering scientific/causal questions.
- understand how far away from a causal conclusion we may be.

Clearly defining what an EFFECT is and WHAT effect we are interested in

- this gets even more relevant when we deal with time-varying settings.

[2] The output of the analysis is not causal just because we use causal inference methods.

# The roadmap (applied)

At the consultation service at the Section of Biostatistics:

> *" I need help to choose the right statistical method to analyze my data ... I have a binary outcome and a lot of covariates ... I tried to run a logistic regression ... "*

No mentioning of what scientific question is actually of interest.

# The roadmap (applied)

At the consultation service at the Section of Biostatistics:

> *" I need help to choose the right statistical method to analyze my data ... I have a binary outcome and a lot of covariates ... I tried to run a logistic regression ... "*

No mentioning of what scientific question is actually of interest.

No clear distinction between "the statistical estimation part" and the "scientific question part".

# The roadmap (applied)

1. Observed data — $O = (X, A, Y)$
2. Causal model — what we know/believe/assume about directions of effects
3. Causal question and target causal estimand — formulating the scientific question as a contrast between counterfactual outcomes (e.g., in terms of ideal hypothetical experiment)
4. Identifiability — is data sufficient to estimate the causal effect?

# The roadmap (applied)

1. Observed data — $O = (X, A, Y)$
2. Causal model — what we know/believe/assume about directions of effects
3. Causal question and target causal estimand — formulating the scientific question as a contrast between counterfactual outcomes (e.g., in terms of ideal hypothetical experiment)
4. Identifiability — is data sufficient to estimate the causal effect?

. . . the rest is purely statistics.

# Summary — roadmap of targeted learning

## Statistical theory for parametric models

- meant for settings where the model is known a priori
  - the model is rarely known a priori
  - theory does not reflect how data are in fact analyzed (e.g., due to use of model selection strategies)
- the model is chosen for its simplicity and convenience
  - simple summary measures of associations

## Targeted learning

- translating scientific question into predefined model-free target parameter
- machine learning based estimators can be constructed and still combined with valid/honest inference (allowing full prespecification of the statistical analysis)