An Introduction to TMLE

Timothée Loranchet timothee.loranchet@inserm.fr

The Third IPLESP Workshop on Causal Inference

June 2026

- Who knows about targeted learning / TMLE?

- Who knows about targeted learning / TMLE?
- Who has ever used TMLE?

Systematic review:

Smith, M. J., Phillips, R. V., Luque-Fernandez, M. A., & Maringe, C. (2023). *Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review.* Annals of Epidemiology, 86, 34–48.

Key takeaway from the conclusion:

> *"The TMLE framework [...] expanded in applied studies via tutorials and user-friendly software. More can be done to reach a wider audience [...] including the development of software packages, tutorial articles, as well as seminars and courses targeted to audiences in specific disciplines."*

So here we are :)

It all starts with a question

You want to answer a causal question

- Does tobacco protect against Parkinson's disease?
- Does hydroxychloroquine reduce the risk of severe COVID-19?
- Do heatwaves have an effect on patients taking antihypertensive medications?
- Does PrEP use affect sexual behavior?
- Does climate change impact treatment adhesion in people living with HIV in Central Africa?
- ...

using a **quantitative** approach.

When it comes to estimation...

The main steps are:

1. Define a **causal question**,
2. Translate it into a **causal estimand** (ATE, ATT, CDE...),
3. Make **causal assumptions** (draw a DAG !) to **identify** this estimand,
4. **Estimate** this *do*-free expression in your data,
5. Publish in Nature !

When it comes to estimation...

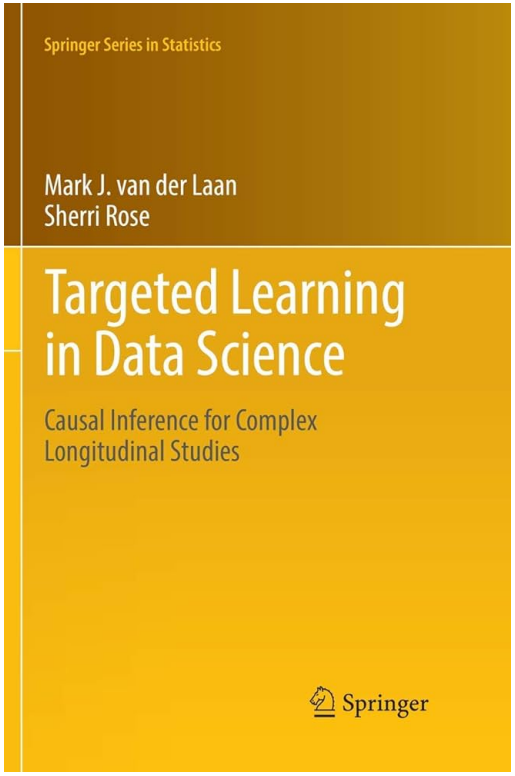
The main steps are:

1. Define a **causal question**,
2. Translate it into a **causal estimand** (ATE, ATT, CDE...),
3. Make **causal assumptions** (draw a DAG !) to **identify** this estimand,
4. **Estimate** this ***do***-free expression in your data,
5. Publish in Nature !

This talk

By the end of this talk, we will:

- Formalize causal estimation problems,
- Understand the motivation for TMLE,
- Build intuition for TMLE,
- Apply TMLE in practice,
- Explore public health applications.



Inserm

iPLesp

SANTÉ SORBONNE UNIVERSITÉ

1

The estimation problem

Statistical and causal models

- $\mathbf{O}$: observed data,

Example

- $\mathbf{O} = (\mathbf{W}, A, Y)$:
  - $A \in \{0, 1\}$: binary treatment
  - $Y \in \{0, 1\}$: binary outcome
  - $\mathbf{W} \in \mathbb{R}^p$: adjustment variables

Statistical and causal models

- $\mathbf{O}$: observed data,
- $\mathbf{O} \sim P_0$: true (unknown) data-generating distribution,

Example

- $\mathbf{O} = (\mathbf{W}, A, Y)$:
  - $A \in \{0, 1\}$: binary treatment
  - $Y \in \{0, 1\}$: binary outcome
  - $\mathbf{W} \in \mathbb{R}^p$: adjustment variables

Statistical and causal models

- $\mathbf{O}$: observed data,
- $\mathbf{O} \sim P_0$: true (unknown) data-generating distribution,
- $\mathcal{P}$: statistical model = set of distributions containing P_0.

Example

- $\mathbf{O} = (\mathbf{W}, A, Y)$:
  - $A \in \{0, 1\}$: binary treatment
  - $Y \in \{0, 1\}$: binary outcome
  - $\mathbf{W} \in \mathbb{R}^p$: adjustment variables
- $\mathcal{P}$: all distributions with finite variance

Statistical and causal models

- $\mathbf{O}$: observed data,
- $\mathbf{O} \sim P_0$: true (unknown) data-generating distribution,
- $\mathcal{P}$: statistical model = set of distributions containing P_0.

Example

- $\mathbf{O} = (\mathbf{W}, A, Y)$:
  - $A \in \{0, 1\}$: binary treatment
  - $Y \in \{0, 1\}$: binary outcome
  - $\mathbf{W} \in \mathbb{R}^p$: adjustment variables
- $\mathcal{P}$: all distributions with finite variance

Causal Model ($\mathcal{M}$) = Statistical Model ($\mathcal{P}$) + Causal Assumptions

The causal estimand

Estimand = the causal quantity we want to learn from the data.

Notation : ψ_0

Example: Average Treatment Effect (ATE)

$$\psi_0 = \mathbb{E}[Y \mid do(A = 1)] - \mathbb{E}[Y \mid do(A = 0)]$$

The causal estimand

Estimand = the causal quantity we want to learn from the data.

Notation : ψ_0

Example: Average Treatment Effect (ATE)

$$\psi_0 = \mathbb{E}[Y \mid do(A = 1)] - \mathbb{E}[Y \mid do(A = 0)]$$

Instantiations may depend on the outcome Y:

- **Binary outcome $Y \in \{0, 1\}$**: risk ratio

$$\psi_0 = P(Y = 1 \mid do(A = 1))/P(Y = 1 \mid do(A = 0))$$

- **Survival outcome $Y \in \mathbb{R}_+$**: difference in survival at time t

$$\psi_0 = P(Y > t \mid do(A = 1)) - P(Y > t \mid do(A = 0))$$

From causal to statistical: identification

From causal to statistical: identification

Example — ATE:
W : valid adjustment set (*e.g.* satisfies the backdoor criterion)

$$\overbrace{\mathbb{E}[Y \mid do(A=1)] - \mathbb{E}[Y \mid do(A=0)]}^{\text{interventional}}$$

$$=$$

$$\underbrace{\mathbb{E}_{\mathbf{W},P_0}[\mathbb{E}_{P_0}[Y \mid A=1,\mathbf{W}] - \mathbb{E}_{P_0}[Y \mid A=0,\mathbf{W}]]}_{\text{observational}}$$

Target Parameter and Estimand

A **target parameter** is a map:

$$\Psi : \mathcal{M} \to \mathbb{R}^d$$

that assigns to each distribution in the model a real-valued quantity (often $d = 1$).

The **estimand** ψ_0 is then defined as the **evaluation of this map at the true data-generating distribution** P_0:

$$\psi_0 := \Psi(P_0)$$

Examples of estimands as target parameters

Mean of an outcome:

$$O \in \mathbb{R}, \quad O \sim P_0$$

$$\mu = \mathbb{E}_{P_0}[O] = \Psi(P_0)$$

Examples of estimands as target parameters

Mean of an outcome:

$$O \in \mathbb{R}, \quad O \sim P_0$$

$$\mu = \mathbb{E}_{P_0}[O] = \Psi(P_0)$$

Average Treatment Effect (ATE):

$$\mathbf{O} = (\mathbf{W}, A, Y) \sim P_0$$

$$ATE = \mathbb{E}_{\mathbf{W},P_0}\Big[\mathbb{E}_{P_0}[Y \mid A = 1, \mathbf{W}] - \mathbb{E}_{P_0}[Y \mid A = 0, \mathbf{W}]\Big] = \Psi(P_0)$$

Examples of estimands as target parameters

Mean of an outcome:

$$O \in \mathbb{R}, \quad O \sim P_0$$

$$\mu = \mathbb{E}_{P_0}[O] = \Psi(P_0)$$

Average Treatment Effect (ATE):

$$\mathbf{O} = (\mathbf{W}, A, Y) \sim P_0$$

$$ATE = \mathbb{E}_{\mathbf{W},P_0}\Big[\mathbb{E}_{P_0}[Y \mid A = 1, \mathbf{W}] - \mathbb{E}_{P_0}[Y \mid A = 0, \mathbf{W}]\Big] = \Psi(P_0)$$

Survival function:

$$O \in \mathbb{R}_+, \quad O \sim P_0$$

$$S(t) = P_0(O > t) = \Psi(P_0)$$

The estimation problem

We have seen that the roadmap to estimation starts with:

1. Defining a causal model $\mathcal{M}$,
2. Defining a target parameter $\Psi(\cdot)$ s.t. $\Psi(P_0)$ is the estimand.

The **estimation problem** is then to construct an **estimator**, i.e. a function of the data $O_1, \ldots, O_n \sim P_0$, taking values in $\mathbb{R}^d$.

This estimator should:

- Respect the assumption that $P_0 \in \mathcal{M}$,
- Provide a "good estimate" of $\Psi(P_0)$,
- Come with valid uncertainty quantification.

2

Standard approaches to estimation

The non-parametric dream

Non-parametric model: $\mathcal{M}$ cannot be indexed by a finite-dimensional parameter.

Example: ATE (binary treatment)

$$\begin{aligned}\psi_0 &= \mathbb{E}[Y \mid do(A=1)] - \mathbb{E}[Y \mid do(A=0)] \\ &= \sum_{\mathbf{w}} \left[\mathbb{E}_{P_0}[Y \mid A=1, \mathbf{W}=\mathbf{w}] - \mathbb{E}_{P_0}[Y \mid A=0, \mathbf{W}=\mathbf{w}] \right] P_0(\mathbf{W}=\mathbf{w})\end{aligned}$$

The non-parametric dream

Non-parametric model: $\mathcal{M}$ cannot be indexed by a finite-dimensional parameter.

Example: ATE (binary treatment)

$$\begin{aligned}\psi_0 &= \mathbb{E}[Y \mid do(A=1)] - \mathbb{E}[Y \mid do(A=0)] \\ &= \sum_{\mathbf{w}} \Big[\mathbb{E}_{P_0}[Y \mid A=1, \mathbf{W}=\mathbf{w}] - \mathbb{E}_{P_0}[Y \mid A=0, \mathbf{W}=\mathbf{w}]\Big] P_0(\mathbf{W}=\mathbf{w})\end{aligned}$$

$$\hat{\psi}_n = \sum_{\mathbf{w}} \Big[\hat{\mu}_n(Y \mid A=1, \mathbf{W}=\mathbf{w}) - \hat{\mu}_n(Y \mid A=0, \mathbf{W}=\mathbf{w})\Big] \hat{P}_n(\mathbf{W}=\mathbf{w})$$

where $\hat{\mu}_n$ = empirical mean ; and $\hat{P}_n$ = empirical proportions.

$\rightsquigarrow$ **Estimate each component of P_0 non parametrically**

Why does it often fail in practice?

Estimating $\hat{\mu}_n(Y \mid A = a, \mathbf{W} = \mathbf{w})$ and $\hat{P}_n(\mathbf{W} = \mathbf{w})$ requires **enough observations per stratum** $(A = a, \mathbf{W} = \mathbf{w})$ and $\mathbf{W} = \mathbf{w}$.

Problems:

- **High-dimensional W**: $\Rightarrow$ most strata are **empty** *Example:* d binary covariates $\Rightarrow 2^{d+1}$ strata.
  With $d = 10$: **2048 strata** to fill.

- **Continuous W**: no two observations share the same value
  $\Rightarrow$ estimator is **undefined**

$\Rightarrow$ **Natural fix**: restrict $\mathcal{M} \rightsquigarrow$ *parametric modeling*

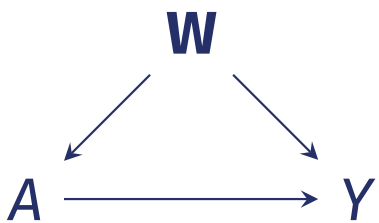
Parametric modeling

A parametric statistical model assumes that the data distribution belongs to a **finite-dimensional family**:

$$\mathcal{M}_\Theta = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^m\}$$

Causal effects are (often) represented by **coefficients**.

Example: linear outcome model



$$Y = \beta A + \gamma^\top \mathbf{W} + \varepsilon, \quad E(\varepsilon \mid A, \mathbf{W}) = 0$$

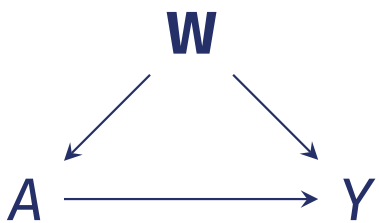
Parametric modeling

A parametric statistical model assumes that the data distribution belongs to a **finite-dimensional family**:

$$\mathcal{M}_\Theta = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^m\}$$

Causal effects are (often) represented by **coefficients**.

Example: linear outcome model



$$Y = \beta A + \gamma^\top \mathbf{W} + \varepsilon, \quad E(\varepsilon \mid A, \mathbf{W}) = 0$$

Under this model, the ATE is summarized by β.
Estimation: regressing Y on $(A, \mathbf{W})$ to obtain $\hat{\beta}$.

Maximum likelihood estimation

Parametric models are commonly estimated via **MLE**.

1. Specify a parametric model $\mathcal{M}_\Theta = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^m\}$
2. Define the likelihood $L(\theta) \overset{(iid)}{=} \prod_{i=1}^n P_\theta(O_i)$
3. Estimate the parameters: $\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta)$
4. Plug-in estimator: $\hat{\psi}^{MLE} = \Psi(P_{\hat{\theta}})$

MLE estimates the full distribution P_θ, then plugs it into Ψ.

Limitations of parametric models

Many standard estimators rely on parametric assumptions:

- Linear regression
- Logistic regression
- Poisson regression
- Cox proportional hazards model

But in reality *(Box, 1976)*:

"All models are wrong."

Limitations of parametric models

Many standard estimators rely on parametric assumptions:

- Linear regression
- Logistic regression
- Poisson regression
- Cox proportional hazards model

But in reality *(Box, 1976)*:

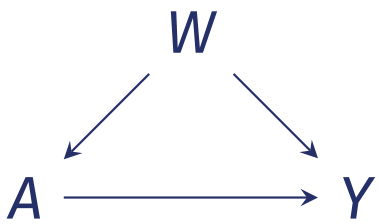
"All models are wrong."

That is: $P_0 \notin \mathcal{M}_\ominus$.

$\hat{\psi}^{MLE}$ may converge to the **wrong quantity**, even as $n \to +\infty$.

Example: ATE in a simple causal model

True causal model over $\mathbf{O} = (W, A, Y) \in \mathbb{R} \times \{0, 1\} \times \mathbb{R}$



$$\begin{cases} W \sim \mathcal{N}(0, 1) \\ A \sim \text{Bernoulli}(\text{logit}^{-1}(5W)) \\ Y = A + 2W^3 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1) \end{cases}$$

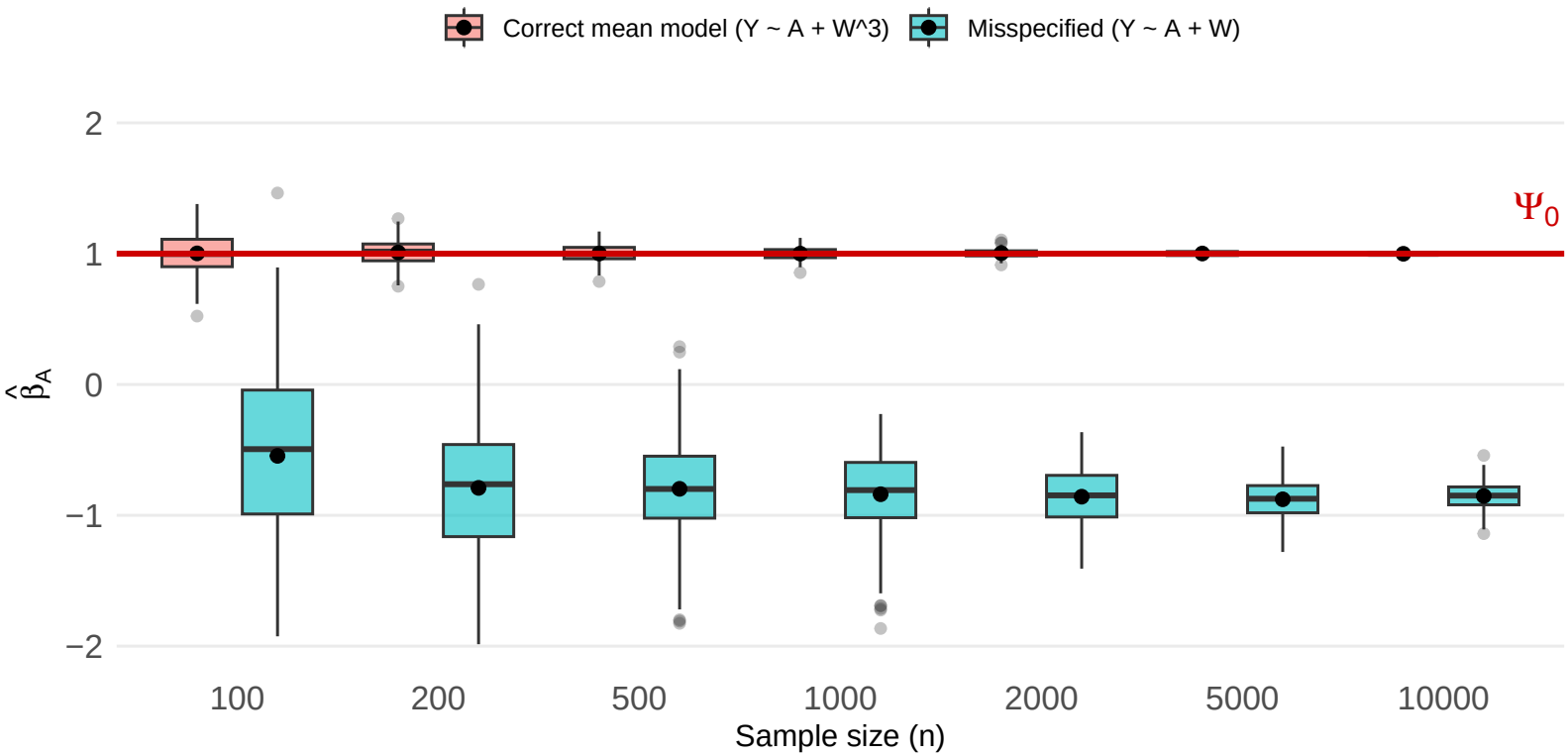
True ATE

$$\Psi(P_0) = \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}[Y \mid A = 1, W] - \mathbb{E}_{P_0}[Y \mid A = 0, W]\right] = 1$$

Example: what if we use a misspecified model?

Working parametric model $\mathcal{M}_\ominus = \{P : \mathbb{E}_P[Y \mid A, W] = \beta A + \gamma W\}$

We fit `lm(Y ~ A + W)` and estimate the ATE by $\hat{\beta}$.



What have we learned?

$\Rightarrow$ We need estimators that are **flexible** yet **statistically efficient**.

Lessons from Targeted Learning:

1. Define causal effects as a **target parameter** Ψ, not as a coefficient in a parametric model.
2. Use a model **large enough** to avoid misspecification and **target** the estimation toward Ψ.

Inserm

iPLesp

SANTÉ SORBONNE UNIVERSITÉ

3

How to construct such an efficient estimator?

Asymptotically Linear Estimators

Setup

- Data: $(O_1, \ldots, O_n) \overset{iid}{\sim} P_0$
- Parameter of interest (estimand): $\Psi(P_0)$
- Estimator: $\hat{\Psi}_n$

Asymptotically Linear Estimators

Setup

- Data: $(O_1, \ldots, O_n) \overset{iid}{\sim} P_0$
- Parameter of interest (estimand): $\Psi(P_0)$
- Estimator: $\hat{\Psi}_n$

$\hat{\Psi}_n$ is asymptotically linear (AL) if

$$\hat{\Psi}_n - \Psi(P_0) = \frac{1}{n} \sum_{i=1}^{n} IC_{P_0}(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

- $IC_{P_0}(\cdot)$: influence curve
- $\mathbb{E}_{P_0}[IC_{P_0}(O)] = 0, \quad \mathbb{E}_{P_0}[IC_{P_0}(O)^2] < +\infty$

Asymptotically Linear Estimators

Setup

- Data: $(O_1, \ldots, O_n) \overset{iid}{\sim} P_0$
- Parameter of interest (estimand): $\Psi(P_0)$
- Estimator: $\hat{\Psi}_n$

$\hat{\Psi}_n$ is asymptotically linear (AL) if

$$\hat{\Psi}_n - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

- $IC_{P_0}(\cdot)$: influence curve
- $\mathbb{E}_{P_0}[IC_{P_0}(O)] = 0, \quad \mathbb{E}_{P_0}[IC_{P_0}(O)^2] < +\infty$

$\Rightarrow$ By the Central Limit Theorem

$$\sqrt{n}\left(\hat{\Psi}_n - \Psi(P_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_0), \quad \sigma_0 = \mathrm{Var}_{P_0}(IC_{P_0}(O))$$

What is the "best" AL estimator?

Let $\hat{\Psi}_n^1$ and $\hat{\Psi}_n^2$ be AL estimators:

$$\sqrt{n}\left(\hat{\Psi}_n^j - \Psi(P_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_0^j), \quad \sigma_0^j = \mathrm{Var}_{P_0}\left(IC_{P_0}^j(O)\right),\ j = 1, 2$$

What is the "best" AL estimator?

Let $\hat{\Psi}_n^1$ and $\hat{\Psi}_n^2$ be AL estimators:

$$\sqrt{n}\left(\hat{\Psi}_n^j - \Psi(P_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_0^j), \quad \sigma_0^j = \mathrm{Var}_{P_0}\left(IC_{P_0}^j(O)\right),\ j = 1, 2$$

$\hat{\Psi}_n^1$ is more efficient than $\hat{\Psi}_n^2$ if $\sigma_0^1 \leq \sigma_0^2$.

Efficient estimator
An estimator $\hat{\Psi}_n^*$ is **efficient** if it minimizes asymptotic variance among all AL estimators:

$$\sigma_0^* = \mathrm{Var}_{P_0}\left(IC_{P_0}^*(O)\right) = \inf \sigma_0$$

- $IC_{P_0}^*$: efficient influence curve (EIC)
- σ_0^*: semiparametric Cramér–Rao lower bound

Update on our Objective

Given:

- a target parameter $\Psi(\cdot)$ such that the estimand is $\Psi(P_0)$,
- a statistical model $\mathcal{M}$ such that $P_0 \in \mathcal{M}$,

our goal is to construct an **efficient AL estimator** $\hat{\Psi}_n^*$:

$$\hat{\Psi}_n^* - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2}),$$

Key ingredient: the efficient influence curve IC_P^*.

Importantly: IC_P^* can be computed for every $P \in \mathcal{M}$!

How do we compute the EIC?

Key ingredients:

- Pathwise derivatives,
- Some integration,
- Hilbert spaces and projections,
- The Riesz representation theorem.

How do we compute the EIC?

Key ingredients:

- Pathwise derivatives,
- Some integration,
- Hilbert spaces and projections,
- The Riesz representation theorem.

Good news: for many common estimands, the EIC has already been derived for us!

Example 1: EIC of a univariate survival function

- $O \in \mathbb{R}_+ \sim P_0 \in \mathcal{M}$
- $\mathcal{M}$: space of all probability densities on $\mathbb{R}_+$
- Target parameter: for all $P \in \mathcal{M}$,

$$\Psi(P) := P(O > t) = \int \mathbb{1}_{[t,+\infty)} \, dP$$

- Estimand (survival function at $t \in \mathbb{R}_+$): $S(t) = \Psi(P_0)$

Example 1: EIC of a univariate survival function

- $O \in \mathbb{R}_+ \sim P_0 \in \mathcal{M}$
- $\mathcal{M}$: space of all probability densities on $\mathbb{R}_+$
- Target parameter: for all $P \in \mathcal{M}$,

$$\Psi(P) := P(O > t) = \int \mathbb{1}_{[t,+\infty)} \, dP$$

- Estimand (survival function at $t \in \mathbb{R}_+$): $S(t) = \Psi(P_0)$

Efficient influence curve for all $P \in \mathcal{M}$:

$$IC_P^*(O_i) = \mathbb{1}_{[t,+\infty)}(O_i) - \Psi(P)$$

Example 2: EIC of the ATE (binary treatment)

- $\mathbf{O} = (\mathbf{W}, A, Y) \sim P_0 \in \mathcal{M}$
- $\mathcal{M}$: space of all probability densities compatible with the causal assumptions
- Target parameter: for all $P \in \mathcal{M}$,

$$\Psi(P) := \mathbb{E}_{\mathbf{W},P}\left[\mathbb{E}_P(Y \mid A = 1, \mathbf{W}) - \mathbb{E}_P(Y \mid A = 0, \mathbf{W})\right]$$

- Estimand: $ATE = \Psi(P_0)$

Example 2: EIC of the ATE (binary treatment)

- $\mathbf{O} = (\mathbf{W}, A, Y) \sim P_0 \in \mathcal{M}$
- $\mathcal{M}$: space of all probability densities compatible with the causal assumptions
- Target parameter: for all $P \in \mathcal{M}$,

$$\Psi(P) := \mathbb{E}_{\mathbf{W},P}\left[\mathbb{E}_P(Y \mid A = 1, \mathbf{W}) - \mathbb{E}_P(Y \mid A = 0, \mathbf{W})\right]$$

- Estimand: $ATE = \Psi(P_0)$

Efficient influence curve for all $P \in \mathcal{M}$:

$$IC^*_P(\mathbf{O}) = \left(\frac{A}{g(1 \mid \mathbf{W})} - \frac{1-A}{g(0 \mid \mathbf{W})}\right)[Y - \bar{Q}(A, \mathbf{W})] + \bar{Q}(1, \mathbf{W}) - \bar{Q}(0, \mathbf{W}) - \Psi(P)$$

where $\bar{Q}(A, \mathbf{W}) := \mathbb{E}_P[Y \mid A, \mathbf{W}]$ and $g(A \mid \mathbf{W}) := P(A \mid \mathbf{W})$.

Estimators and EIC

We have:

- Data $(O_1, \cdots, O_n) \overset{iid}{\sim} P_0 \in \mathcal{M}$
- Target parameter $\Psi(\cdot)$ with estimand $\Psi(P_0)$
- The efficient influence curve IC_P^* for all $P \in \mathcal{M}$

Estimators and EIC

We have:

- Data $(O_1, \cdots, O_n) \overset{iid}{\sim} P_0 \in \mathcal{M}$
- Target parameter $\Psi(\cdot)$ with estimand $\Psi(P_0)$
- The efficient influence curve IC_P^* for all $P \in \mathcal{M}$
- $\hat{P}_n$: a consistent etimator of P_0

Estimators and EIC

We have:

- Data $(O_1, \cdots, O_n) \overset{iid}{\sim} P_0 \in \mathcal{M}$
- Target parameter $\Psi(\cdot)$ with estimand $\Psi(P_0)$
- The efficient influence curve IC^*_P for all $P \in \mathcal{M}$
- $\hat{P}_n$: a consistent etimator of P_0

A first-order expansion of Ψ gives, under regularity conditions:

$$\Psi(\hat{P}_n) - \Psi(P_0) + \frac{1}{n}\sum_{i=1}^{n} IC^*_{\hat{P}_n}(O_i) = \frac{1}{n}\sum_{i=1}^{n} IC^*_{P_0}(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

Solving the EIC equation

Recall: we want $\hat{\Psi}_n^*$ satisfying

$$\hat{\Psi}_n^* - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

From the previous slide:

$$\Psi(\hat{P}_n) - \Psi(P_0) + \frac{1}{n}\sum_{i=1}^{n} IC_{\hat{P}_n}^*(O_i) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

Solving the EIC equation

Recall: we want $\hat{\Psi}_n^*$ satisfying

$$\hat{\Psi}_n^* - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

From the previous slide:

$$\Psi(\hat{P}_n) - \Psi(P_0) + \frac{1}{n}\sum_{i=1}^{n} IC_{\hat{P}_n}^*(O_i) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

Solving the EIC equation

Recall: we want $\hat{\Psi}_n^*$ satisfying

$$\hat{\Psi}_n^* - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

From the previous slide:

$$\Psi(\hat{P}_n) - \Psi(P_0) + \frac{1}{n}\sum_{i=1}^{n} IC_{\hat{P}_n}^*(O_i) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

$\Longrightarrow$ We need to find $\hat{P}_n^*$ satisfying the **EIC equation**:

$$\frac{1}{n}\sum_{i=1}^{n} IC_{\hat{P}_n^*}^*(O_i) = 0$$

Then $\hat{\Psi}_n^* := \Psi(\hat{P}_n^*)$ is an **efficient** AL estimator.

Solving the EIC equation

Recall: we want $\hat{\Psi}_n^*$ satisfying

$$\hat{\Psi}_n^* - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

From the previous slide:

$$\Psi(\hat{P}_n) - \Psi(P_0) + \frac{1}{n}\sum_{i=1}^{n} IC_{\hat{P}_n}^*(O_i) = \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}^*(O_i) + o_{\mathbb{P}}(n^{-1/2})$$

$\Longrightarrow$ We need to find $\hat{P}_n^*$ satisfying the **EIC equation**:

$$\frac{1}{n}\sum_{i=1}^{n} IC_{\hat{P}_n^*}^*(O_i) = 0$$

Then $\hat{\Psi}_n^* := \Psi(\hat{P}_n^*)$ is an **efficient** AL estimator.

TMLE = Procedure to find such an estimator $\hat{P}_n^*$

TMLE IN A NUTSHELL

TMLE = *"Targeted Maximum Likelihood Estimation"*

1. Define a model $\mathcal{M}$ and a target parameter

$$\Psi : \mathcal{M} \to \mathbb{R}^d$$

   such that the estimand is $\Psi(P_0)$.
2. Construct an **initial estimator** $\hat{P}_n^0$ of P_0.
3. **Update** $\hat{P}_n^0$ into $\hat{P}_n^*$ via a **fluctuation submodel** until

$$\frac{1}{n}\sum_{i=1}^{n} IC^*_{\hat{P}_n^*}(O_i) = 0$$

4. Set $\hat{\Psi}_n^{TMLE} = \Psi(\hat{P}_n^*)$.

4

Super Learning for the initial estimator

Do we need to estimate the full distribution?

In many cases, the target parameter depends only on a lower-dimensional component $Q(P_0)$ of the full distribution.

Example: the ATE

$$\psi_0 := \Psi(P_0) = \mathbb{E}_{\mathbf{W},P_0}\Big[\mathbb{E}_{P_0}[Y \mid A = 1, \mathbf{W}] - \mathbb{E}_{P_0}[Y \mid A = 0, \mathbf{W}]\Big]$$

Define:

- $\bar{Q}_0(A, \mathbf{W}) = \mathbb{E}_{P_0}[Y \mid A, \mathbf{W}]$
- $Q_{\mathbf{W},0}$ the marginal distribution of $\mathbf{W}$
- $Q_0 = (\bar{Q}_0, Q_{\mathbf{W},0})$

$$\psi_0 = \int_{\mathbf{W}} [\bar{Q}_0(1, \mathbf{W}) - \bar{Q}_0(0, \mathbf{W})]\, dQ_{\mathbf{W},0}(w) = \Psi(Q_0)$$

We only need to estimate Q_0, not the full distribution P_0.

Choosing an estimator

Setup: $O_1, \ldots, O_n$ i.i.d. from $P_0 \in \mathcal{M}$

Goal: estimate $Q_0 := Q(P_0)$ *(e.g. a conditional expectancy)*

Many candidates:

- Regression models (linear, logistic, Poisson)
- K-nearest neighbors (KNN)
- Kernel regression
- Generalized Additive Models (GAM)
- Smoothing splines
- Decision trees (CART), Random Forest
- Boosting (XGBoost, AdaBoost)
- Penalized regression (Lasso, Ridge, Elastic Net)
- Your favorite model
- ...

Choosing an estimator

Setup: $O_1, \ldots, O_n$ i.i.d. from $P_0 \in \mathcal{M}$

Goal: estimate $Q_0 := Q(P_0)$ *(e.g. a conditional expectancy)*

Many candidates:

- Regression models (linear, logistic, Poisson)
- K-nearest neighbors (KNN)
- Kernel regression
- Generalized Additive Models (GAM)
- Smoothing splines
- Decision trees (CART), Random Forest
- Boosting (XGBoost, AdaBoost)
- Penalized regression (Lasso, Ridge, Elastic Net)
- Your favorite model
- ...

$\Rightarrow$ **Which one is the best for your data ?**

Super Learning: Let the data choose

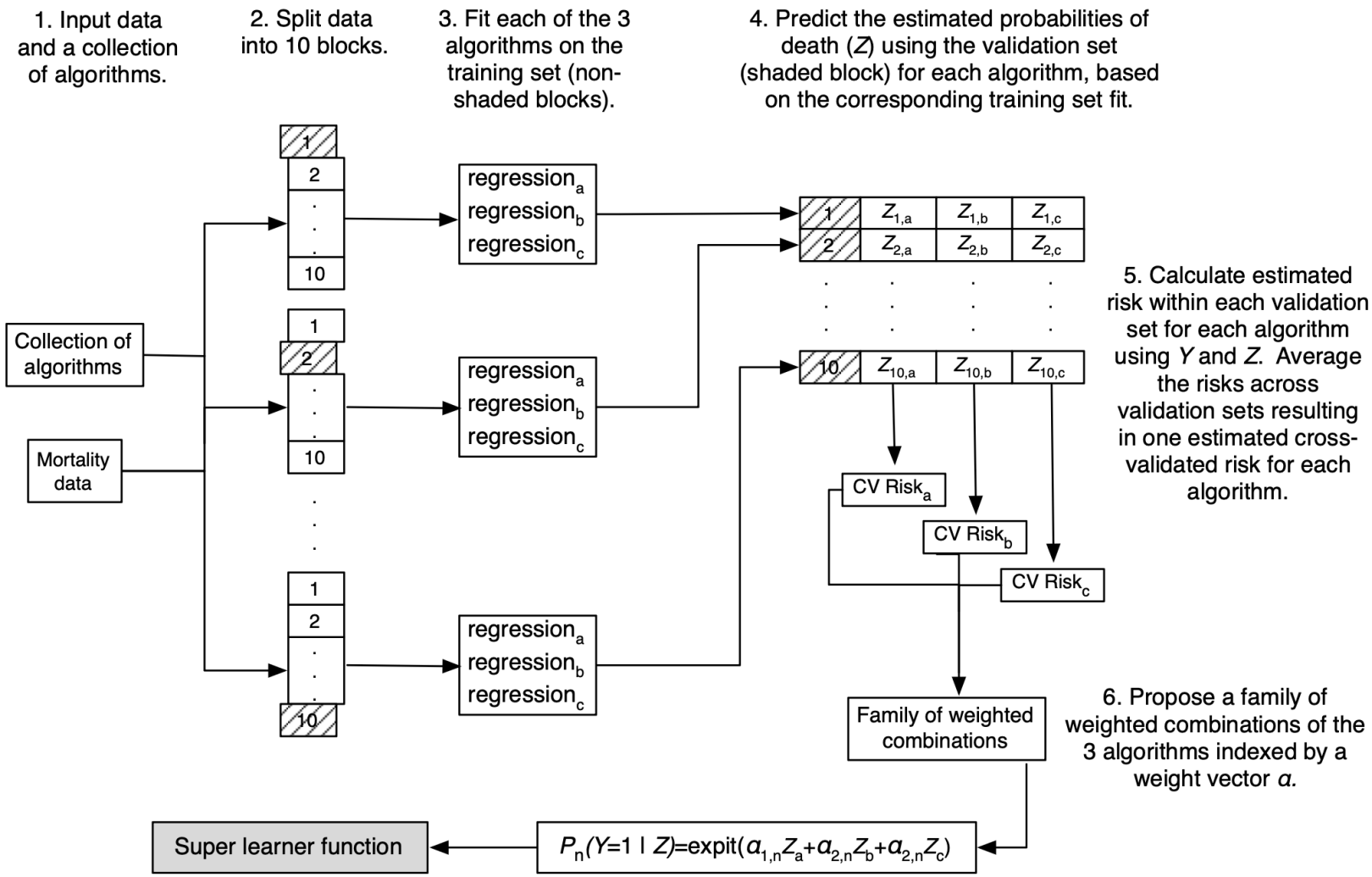
Super Learner algorithm

1. Build a library of candidate learners $\{\hat{f}_1, \ldots, \hat{f}_M\}$
2. Split the data into K folds
3. For each learner and each fold:
   - train on $(K-1)$ folds
   - predict on the held-out fold
4. Find the optimal convex combination minimizing cross-validated risk:

$$\hat{f}_{SL} = \sum_{m=1}^{M} \hat{\alpha}_m \hat{f}_m, \qquad \hat{\alpha}_m \geq 0, \quad \sum_{m=1}^{M} \hat{\alpha}_m = 1$$

Key idea: rather than selecting a single model, **combine learners using cross-validated** predictive performance.

Super Learning: pipeline



Example of Super Learning

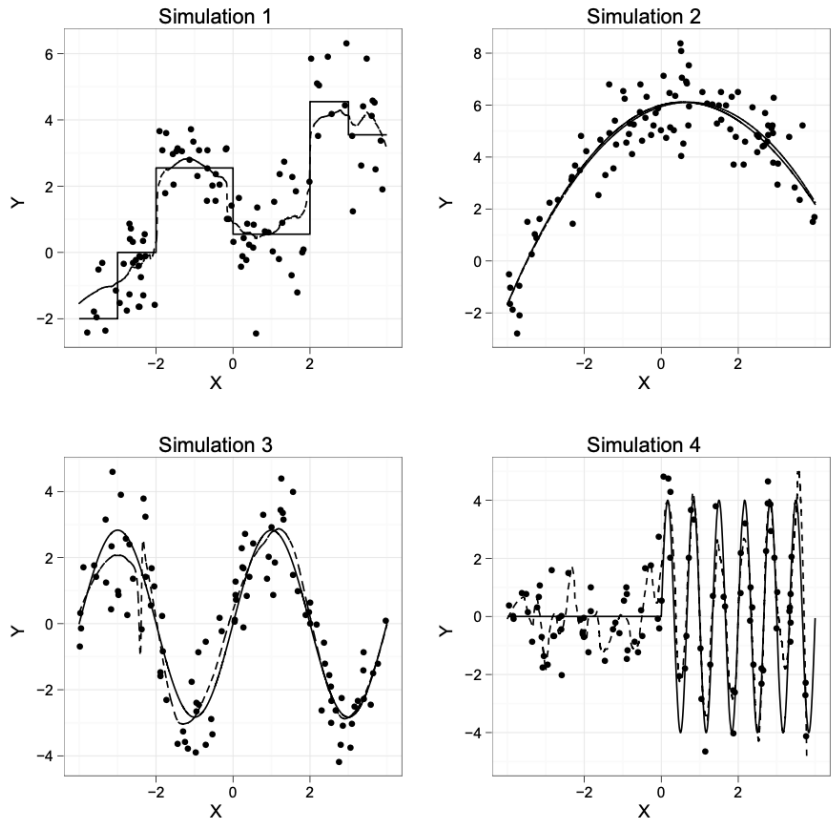
Simulation 1: $Y = -2 \times \mathrm{I}(X < -3) + 2.55 \times \mathrm{I}(X > -2) - 2 \times \mathrm{I}(X > 0) + 4 \times \mathrm{I}(X > 2) - 1 \times \mathrm{I}(X > 3) + U;$

Simulation 2: $Y = 6 + 0.4X - 0.36X^2 + 0.005X^3 + U;$

Simulation 3: $Y = 2.83 \times \sin\left(\frac{\pi}{2} \times X\right) + U;$

Simulation 4: $Y = 4 \times \sin(3\pi \times X) \times \mathrm{I}(X > 0) + U,$

R Algorithm	Description	Source
glm	Linear model	R Development Core Team (2010)
interaction	Polynomial linear model	R Development Core Team (2010)
randomForest	Random forest	Liaw and Wiener (2002)
		Breiman (2001b)
bagging	Bootstrap aggregation of trees	Peters and Hothorn (2009)
		Breiman (1996d)
gam	Generalized additive models	Hastie (1992)
		Hastie and Tibshirani (1990)
gbm	Gradient boosting	Ridgeway (2007)
		Friedman (2001)
nnet	Neural network	Venables and Ripley (2002)
polymars	Polynomial spline regression	Kooperberg (2009)
		Friedman (1991)
bart	Bayesian additive regression trees	Chipman and McCulloch (2009)
		Chipman et al. (2010)
loess	Local polynomial regression	Cleveland et al. (1992)



Source: van der Laan & Rose, Targeted Learning (2011)

Is a good initial estimator enough?

Suppose we have a flexible, consistent estimator $\hat{Q}_n$ of Q_0.

Natural idea: plug it directly into Ψ:

$$\hat{\psi}_n^{plug-in} = \Psi(\hat{Q}_n)$$

Is a good initial estimator enough?

Suppose we have a flexible, consistent estimator $\hat{Q}_n$ of Q_0.

Natural idea: plug it directly into Ψ:

$$\hat{\psi}_n^{plug-in} = \Psi(\hat{Q}_n)$$

Problem: $\hat{Q}_n$ is optimized to fit Q_0 globally, not $\Psi(Q_0)$.

Is a good initial estimator enough?

Suppose we have a flexible, consistent estimator $\hat{Q}_n$ of Q_0.

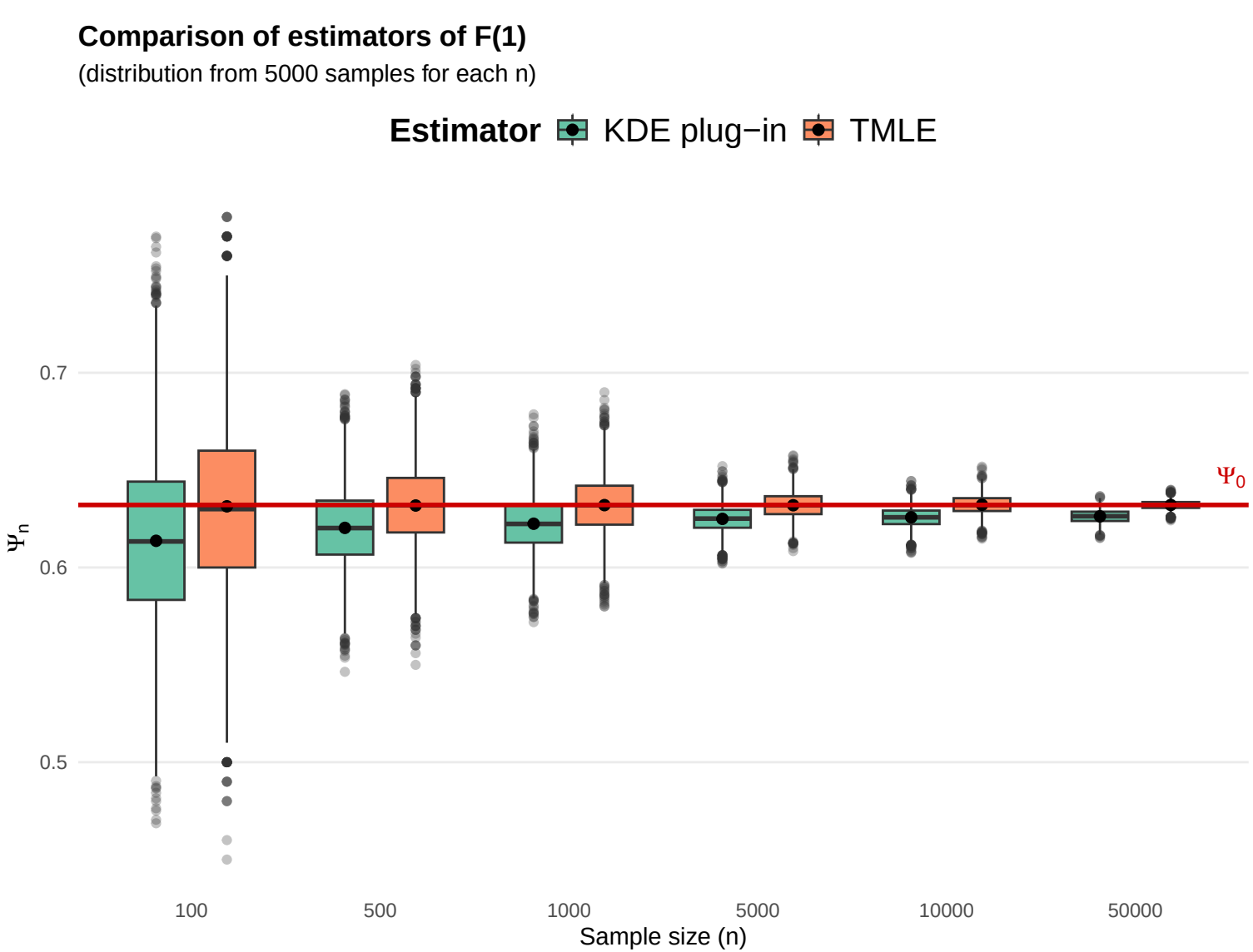
Natural idea: plug it directly into Ψ:

$$\hat{\psi}_n^{plug-in} = \Psi(\hat{Q}_n)$$

Problem: $\hat{Q}_n$ is optimized to fit Q_0 globally, not $\Psi(Q_0)$.

Example: Estimating $\Psi(P_0) = P_0(O \leq 1)$ when $O \sim \exp(\lambda = 1)$
True value: $\psi_0 = \int_{-\infty}^{1} f_0(o)do \approx 0.63$

1. Estimate density $\hat{f}_n$ via KDE
2. Plug in: $\hat{\Psi}_n = \int_{-\infty}^{1} \hat{f}_n(o)\, do$



Good estimation of $Q_0 \neq$ optimal estimation of $\Psi(Q_0)$.

$\Rightarrow$ We need to *target* the fit toward $\Psi \Rightarrow$ TMLE.

Inserm

iPLesp

SANTÉ SORBONNE UNIVERSITÉ

5

TMLE

TMLE IN A NUTSHELL

TMLE = *"Targeted Maximum Likelihood Estimation"*

1. Define a model $\mathcal{M}$ and a target parameter

$$\Psi : \mathcal{M} \to \mathbb{R}^d$$

   such that the estimand is $\Psi(P_0)$.
2. Construct an **initial estimator** $\hat{P}_n^0$ of P_0 (e.g. Super Learner).
3. **Update** $\hat{P}_n^0$ into $\hat{P}_n^*$ via a **fluctuation submodel** until

$$\frac{1}{n}\sum_{i=1}^{n} IC^*_{\hat{P}_n^*}(O_i) = 0$$

4. Set $\hat{\Psi}_n^{TMLE} = \Psi(\hat{P}_n^*)$.

Intuition: navigating the space of distributions

Think of $\mathcal{M}$ as a **space of distributions**.

- P_0: the unknown truth
- $\hat{P}_n^0$: our initial estimate, close to P_0, but not solving $\frac{1}{n}\sum IC^*_{\hat{P}_n^0}(O_i) = 0$
- $\hat{P}_n^*$: the TMLE target = the point in $\mathcal{M}$ that solves the EIC equation

Idea: define a **path** (submodel) through $\mathcal{M}$ starting at $\hat{P}_n^0$, indexed by a scalar ε:

$$\varepsilon \mapsto \hat{P}_n^0(\varepsilon), \qquad \hat{P}_n^0(\varepsilon = 0) = \hat{P}_n^0$$

Then **move along this path** until the EIC equation is solved.

Choice of the submodel

For $P \in \mathcal{M}$, define a parametric submodel $\mathcal{S} = \{P(\varepsilon) : \varepsilon \in (-\delta, \delta)\}$ such that:

- $p(\varepsilon = 0) = p$
- $\frac{d}{d\varepsilon} \log p(\varepsilon)\Big|_{\varepsilon=0} = IC_P^*$ (if $\varepsilon \in \mathbb{R}$)

Choice of the submodel

For $P \in \mathcal{M}$, define a parametric submodel $\mathcal{S} = \{P(\varepsilon) : \varepsilon \in (-\delta, \delta)\}$ such that:

- $p(\varepsilon = 0) = p$
- $\frac{d}{d\varepsilon} \log p(\varepsilon)\Big|_{\varepsilon=0} = IC_P^* \quad (\text{if } \varepsilon \in \mathbb{R})$

Example: if $P \in \mathcal{M}$ admits a density p, define

$$p(\varepsilon) = (1 + \varepsilon \cdot IC_P^*)\, p$$

One can verify:

Choice of the submodel

For $P \in \mathcal{M}$, define a parametric submodel $\mathcal{S} = \{P(\varepsilon) : \varepsilon \in (-\delta, \delta)\}$ such that:

- $p(\varepsilon = 0) = p$
- $\frac{d}{d\varepsilon} \log p(\varepsilon)\Big|_{\varepsilon=0} = IC_P^* \quad (\text{if } \varepsilon \in \mathbb{R})$

Example: if $P \in \mathcal{M}$ admits a density p, define

$$p(\varepsilon) = (1 + \varepsilon \cdot IC_P^*)\, p$$

One can verify:

- $p(\varepsilon = 0) = p$ ✓

Choice of the submodel

For $P \in \mathcal{M}$, define a parametric submodel $\mathcal{S} = \{P(\varepsilon) : \varepsilon \in (-\delta, \delta)\}$ such that:

- $p(\varepsilon = 0) = p$
- $\frac{d}{d\varepsilon} \log p(\varepsilon)\Big|_{\varepsilon=0} = IC_P^* \quad (\text{if } \varepsilon \in \mathbb{R})$

Example: if $P \in \mathcal{M}$ admits a density p, define

$$p(\varepsilon) = (1 + \varepsilon \cdot IC_P^*)\, p$$

One can verify:

- $p(\varepsilon = 0) = p$ ✓
- $\frac{d}{d\varepsilon} \log p(\varepsilon)\Big|_{\varepsilon=0} = \frac{IC_P^*}{1+\varepsilon IC_P^*}\Big|_{\varepsilon=0} = IC_P^*$ ✓

TMLE PROCEDURE

Starting from $\hat{P}_n^0$ and the submodel $\{P(\varepsilon) : \varepsilon\}$:

TMLE PROCEDURE

Starting from $\hat{P}_n^0$ and the submodel $\{P(\varepsilon) : \varepsilon\}$:

1. Compute the MLE for ε:

$$\varepsilon_n^0 := \arg\max_{\varepsilon} \sum_{i=1}^{n} \log \hat{p}_n^0(\varepsilon)(O_i)$$

TMLE PROCEDURE

Starting from $\hat{P}_n^0$ and the submodel $\{P(\varepsilon) : \varepsilon\}$:

1. Compute the MLE for ε:

$$\varepsilon_n^0 := \arg\max_{\varepsilon} \sum_{i=1}^{n} \log \hat{p}_n^0(\varepsilon)(O_i)$$

2. First update: $\hat{P}_n^1 = \hat{P}_n^0(\varepsilon_n^0)$

TMLE PROCEDURE

Starting from $\hat{P}_n^0$ and the submodel $\{P(\varepsilon) : \varepsilon\}$:

1. Compute the MLE for ε:

$$\varepsilon_n^0 := \arg\max_\varepsilon \sum_{i=1}^n \log \hat{p}_n^0(\varepsilon)(O_i)$$

2. First update: $\hat{P}_n^1 = \hat{P}_n^0(\varepsilon_n^0)$
3. Iterate:
   - compute $\varepsilon_n^k := \arg\max_\varepsilon \sum_{i=1}^n \log \hat{p}_n^k(\varepsilon)(O_i)$
   - set $\hat{P}_n^{k+1} = \hat{P}_n^k(\varepsilon_n^k)$

   until $\varepsilon_n^K \approx 0$
   In practice, convergence is often achieved in ***one step****.*

TMLE PROCEDURE

Starting from $\hat{P}_n^0$ and the submodel $\{P(\varepsilon) : \varepsilon\}$:

1. Compute the MLE for ε:

$$\varepsilon_n^0 := \arg\max_\varepsilon \sum_{i=1}^n \log \hat{p}_n^0(\varepsilon)(O_i)$$

2. First update: $\hat{P}_n^1 = \hat{P}_n^0(\varepsilon_n^0)$
3. Iterate:
   - compute $\varepsilon_n^k := \arg\max_\varepsilon \sum_{i=1}^n \log \hat{p}_n^k(\varepsilon)(O_i)$
   - set $\hat{P}_n^{k+1} = \hat{P}_n^k(\varepsilon_n^k)$

   until $\varepsilon_n^K \approx 0$
   In practice, convergence is often achieved in ***one step****.*
4. TMLE: $\hat{P}_n^* = \hat{P}_n^K$

TMLE OF THE ESTIMAND

At convergence we have $\hat{P}_n^*$. The TMLE estimator is:

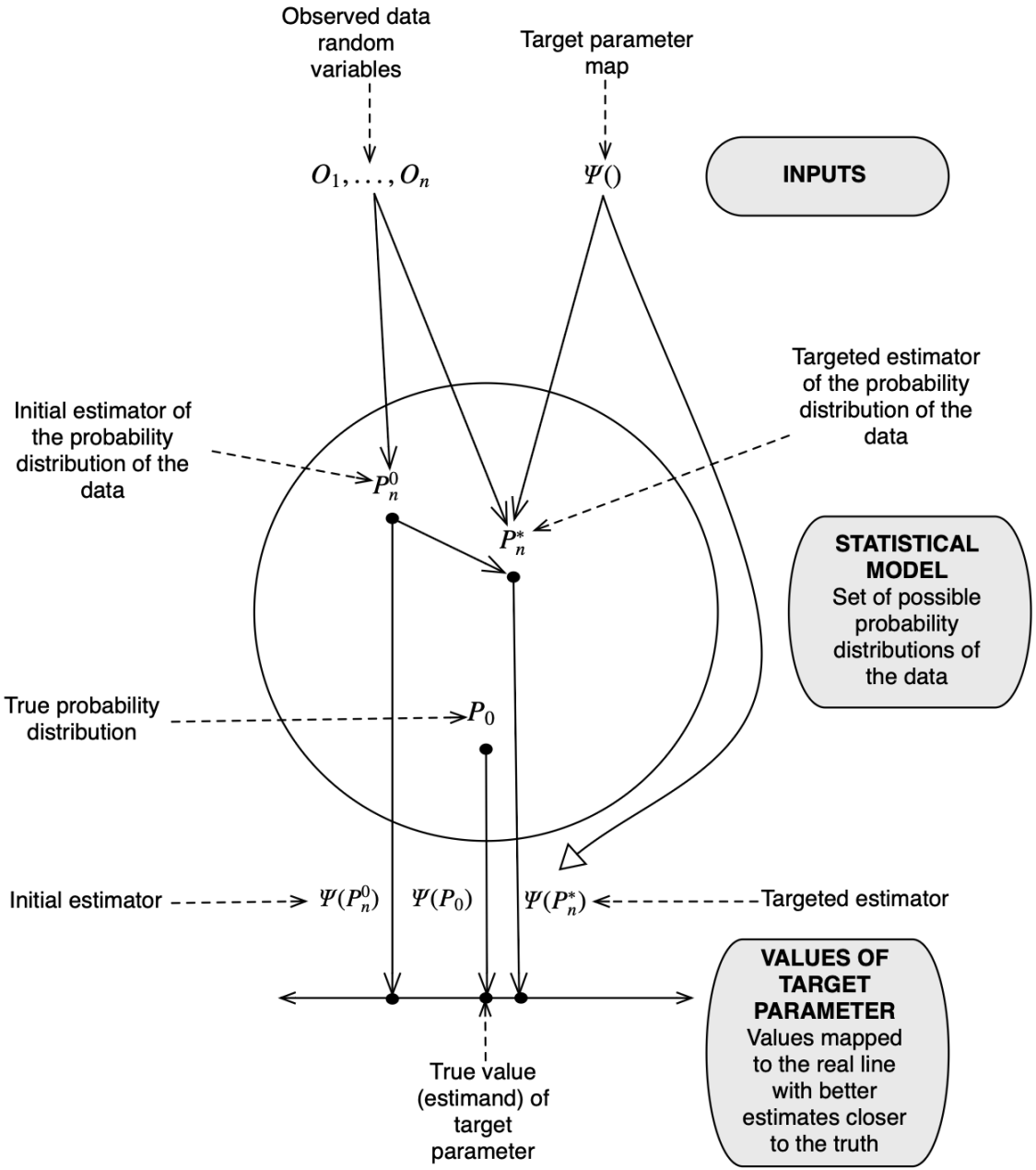
$$\hat{\Psi}_n^* := \Psi(\hat{P}_n^*)$$

Why is it efficient?

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} IC^*_{\hat{P}_n^*}(O_i) &= \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\varepsilon} \log \hat{p}_n^*(\varepsilon)(O_i)\Big|_{\varepsilon=0} && \text{(by construction of the submodel)} \\
&= \frac{d}{d\varepsilon}\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_n^K(\varepsilon)(O_i)\Big|_{\varepsilon=\varepsilon_n^K} && \text{(by definition of TMLE)} \\
&= 0 && \text{(first-order optimality of MLE)}
\end{aligned}$$

$\Rightarrow \hat{P}_n^*$ solves the EIC equation
$\Rightarrow \hat{\Psi}_n^*$ is an efficient AL estimator.

TMLE PROCEDURE SUMMARIZED



Example: TMLE for the ATE

Observed data: $O = (W, A, Y) \sim P_0$

- W: adjustment set
- $A \in \{0, 1\}$: treatment
- $Y \in \{0, 1\}$: outcome

Target parameter:

$$\Psi(P) = \mathbb{E}_{W,P}\Big[\mathbb{E}_P[Y \mid A = 1, W] - \mathbb{E}_P[Y \mid A = 0, W]\Big] \qquad \text{ATE} : \Psi(P_0)$$

Example: TMLE for the ATE (initial estimation)

Define

$$\bar{Q}(A, W) := \mathbb{E}[Y \mid A, W], \qquad g(A \mid W) := P(A = 1 \mid W)$$

Then $\Psi(P) = \mathbb{E}_W[\bar{Q}(1, W) - \bar{Q}(0, W)]$.

Step 1: Initial estimates via Super Learner

$$\bar{Q}_n^0(A, W) \qquad \text{and} \qquad g_n(A \mid W)$$

Note: g_n is not directly targeted but will be used to construct the clever covariate in the fluctuation step.

Example: TMLE for the ATE (EIC)

The EIC of the ATE is:

$$IC^*_{(\bar{Q},g)}(O) = \underbrace{H^*(A,W)}_{\text{clever covariate}} \left[Y - \bar{Q}(A,W)\right] + \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(P)$$

with

$$H^*(A,W) = \frac{A}{g(1 \mid W)} - \frac{1-A}{g(0 \mid W)}$$

TMLE updates $\bar{Q}^0_n$ so that $\frac{1}{n}\sum_i IC^*_{(\bar{Q}^*_n, g_n)}(O_i) \approx 0$.

Example: TMLE for the ATE (fluctuation)

We need $\bar{Q}_n^*$ solving:

$$\frac{1}{n}\sum_{i=1}^{n} H_n^*(A_i, W_i)\left[Y_i - \bar{Q}_n^*(A_i, W_i)\right] = 0$$

(The other term in the EIC cancels automatically.)

Fluctuation submodel:

$$\operatorname{logit} \bar{Q}_n^0(\varepsilon)(A, W) = \operatorname{logit} \bar{Q}_n^0(A, W) + \varepsilon\, H_n^*(A, W)$$

The MLE $\hat{\varepsilon}^*$ satisfies exactly:

$$\frac{1}{n}\sum_{i=1}^{n} H_n^*(A_i, W_i)\left[Y_i - \bar{Q}_n(\hat{\varepsilon}^*)(A_i, W_i)\right] = 0$$

$\Rightarrow$ **One fluctuation step suffices.** Set $\bar{Q}_n^* = \bar{Q}_n^0(\hat{\varepsilon}^*)$.

Example: TMLE for the ATE (final estimation)

The TMLE estimator of the ATE is:

$$\hat{\Psi}_n^* := \Psi(\bar{Q}_n^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)\right]$$

Summary of the ATE-TMLE procedure:

1. Estimate $\bar{Q}_n^0$ and g_n via Super Learner
2. Compute clever covariate $H_n^*(A, W) = \frac{A}{g_n(1|W)} - \frac{1-A}{g_n(0|W)}$
3. Fit logistic regression of Y on H_n^* with offset logit $\bar{Q}_n^0 \to \hat{\varepsilon}^*$
4. Update $\bar{Q}_n^* = \bar{Q}_n^0(\hat{\varepsilon}^*)$
5. Plug in: $\hat{\Psi}_n^* = \frac{1}{n}\sum_i[\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)]$

Inference

Since TMLE is asymptotically linear and efficient:

$$\sqrt{n}\left(\hat{\Psi}_n^* - \Psi(P_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \qquad \sigma^2 = \mathrm{Var}_{P_0}\left(IC_{P_0}^*(O)\right)$$

Estimated by: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(IC_n^*(O_i)\right)^2$

- **Confidence interval** $(1-\alpha)$:

$$\hat{\Psi}_n^* \pm z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$$

- Test $H_0 : \Psi(P_0) = 0$:

$$T = \frac{\hat{\Psi}_n^*}{\hat{\sigma}/\sqrt{n}}, \qquad p\text{-value} = 2\Phi(-|T|)$$

TMLE is Double Robust

$$P_0(W, A, Y) = \underbrace{P_0(Y \mid A, W)P_0(W)}_{Q_0}\underbrace{P_0(A \mid W)}_{g_0} \text{ with}$$

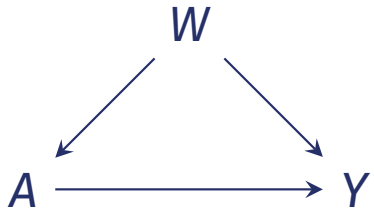
- g_0: treatment mechanism
- Q_0: relevant factor, $\Psi(P_0) \equiv \Psi(Q_0)$

The TMLE $\hat{\Psi}_n^* = \Psi(\hat{Q}_n^*)$ is **consistent** if **either** $\hat{Q}_n^*$ or $\hat{g}_n$ is consistent.

If **both** are consistent $\Rightarrow$ the estimator is **efficient**.

Numerical example: TMLE vs parametric models

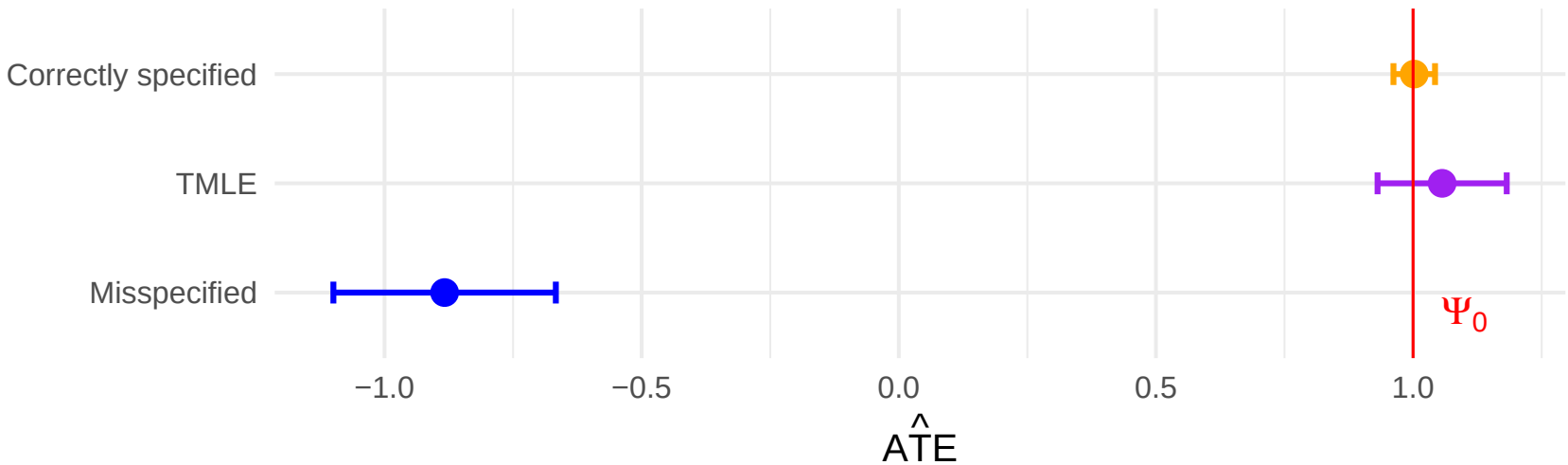
Same causal model as before:



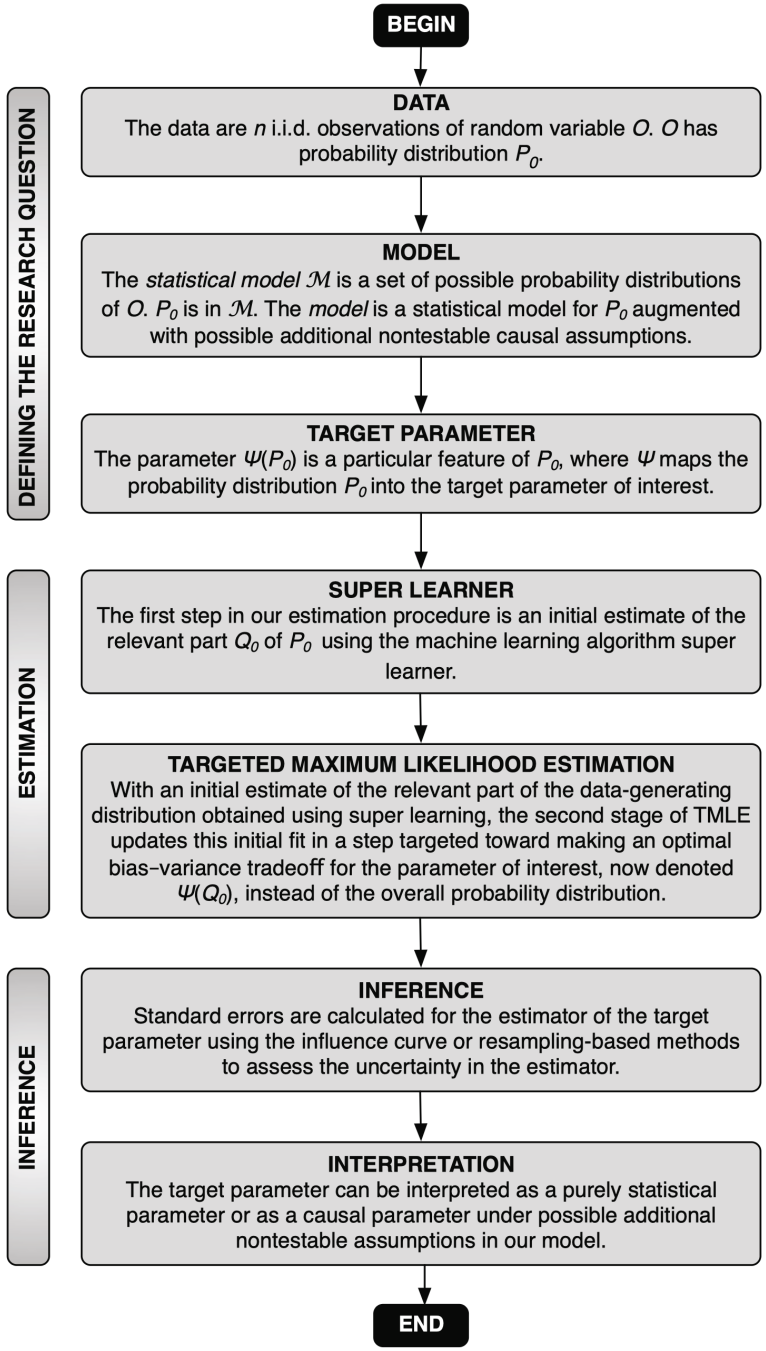
$$\begin{cases} W \sim \mathcal{N}(0,1) \\ A \sim \text{Bernoulli}(\text{logit}^{-1}(5W)) \\ Y = A + 2W^3 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1) \end{cases}$$

Comparison ($n = 5000$):

- **TMLE** with nonparametric initial estimator
- **Misspecified** parametric model: $\mathbb{E}[Y \mid A, W] = \beta A + \gamma W$
- **Correctly specified** parametric model: $\mathbb{E}[Y \mid A, W] = \beta A + \gamma W^3$



ROADMAP


BEGIN
DEFINING THE RESEARCH QUESTION
DATA
The data are n i.i.d. observations of random variable O. O has probability distribution P_0.
MODEL
The *statistical model* $\mathcal{M}$ is a set of possible probability distributions of O. P_0 is in $\mathcal{M}$. The *model* is a statistical model for P_0 augmented with possible additional nontestable causal assumptions.
TARGET PARAMETER
The parameter $\Psi(P_0)$ is a particular feature of P_0, where Ψ maps the probability distribution P_0 into the target parameter of interest.
ESTIMATION
SUPER LEARNER
The first step in our estimation procedure is an initial estimate of the relevant part Q_0 of P_0 using the machine learning algorithm super learner.
TARGETED MAXIMUM LIKELIHOOD ESTIMATION
With an initial estimate of the relevant part of the data-generating distribution obtained using super learning, the second stage of TMLE updates this initial fit in a step targeted toward making an optimal bias–variance tradeoff for the parameter of interest, now denoted $\Psi(Q_0)$, instead of the overall probability distribution.
INFERENCE
INFERENCE
Standard errors are calculated for the estimator of the target parameter using the influence curve or resampling-based methods to assess the uncertainty in the estimator.
INTERPRETATION
The target parameter can be interpreted as a purely statistical parameter or as a causal parameter under possible additional nontestable assumptions in our model.
END

6

A PUBLIC HEALTH APPLICATION

The case study

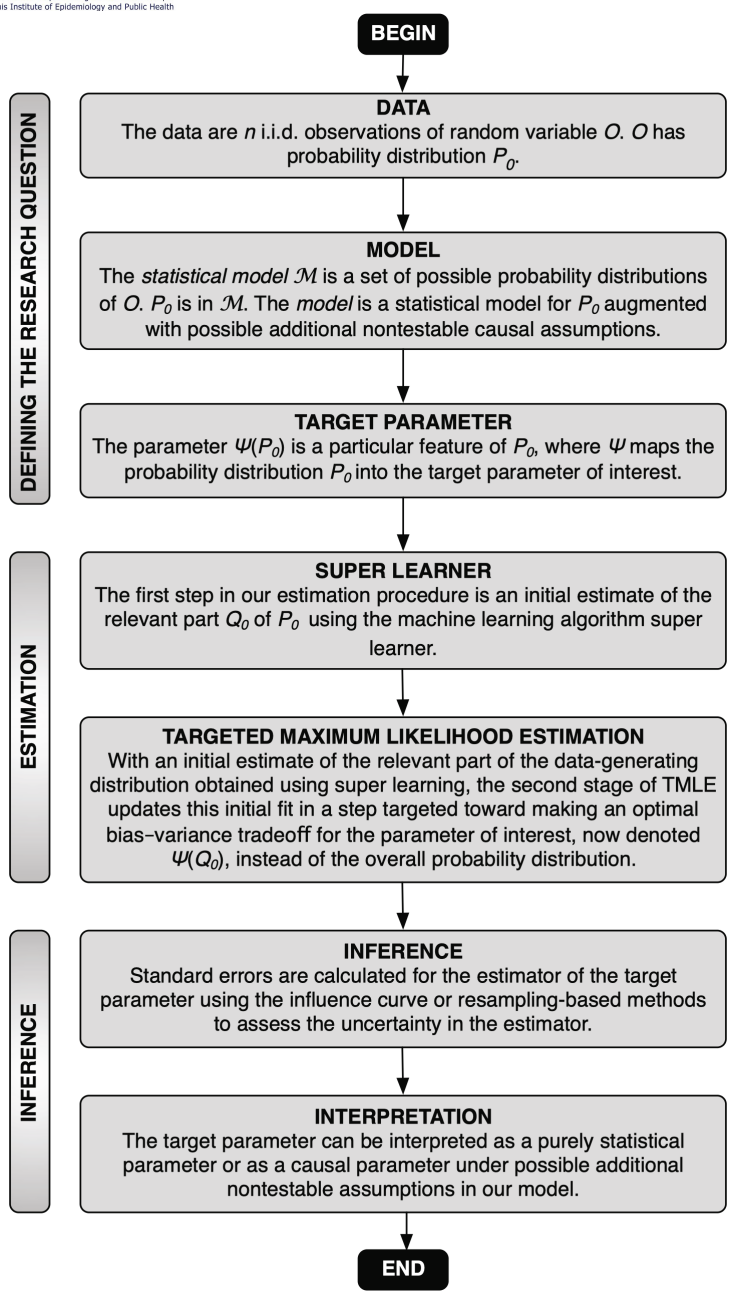
Schnitzer et al. (2014)

Modeling the impact of hepatitis C viral clearance on end-stage liver disease in an HIV co-infected cohort with targeted maximum likelihood estimation

Biometrics, 70(1):144–152.

- **Context**: HIV/HCV co-infected patients have a high risk of progression to end-stage liver disease (ESLD).
- **Causal question**: Does hepatitis C viral clearance reduce the risk of ESLD?
- **Data**: Canadian Co-infection Cohort Study (CCC) (longitudinal)

1. Defining the research question



1. **Data**

   Longitudinal HIV/HCV cohort ($n = 740$)

   - Continuous: age, HIV duration, CD4 count
   - Binary: sex, ARV treatment, alcohol use

2. **Model**

   Semiparametric longitudinal model with:

   - time-varying confounding
   - censoring
   - survival outcome

3. **Target parameter**

   Interventionnal probability of remaining ESLD-free under HCV clearance.

Formalization

We observe patients over K time points.

- **Data structure**

$$\mathbf{O} = (\mathbf{L}_0, \mathbf{A}_0, Y_1, \mathbf{L}_1, \mathbf{A}_1, Y_2, \ldots, \mathbf{L}_{K-1}, \mathbf{A}_{K-1}, Y_K)$$

  - $\mathbf{L}_0$: baseline covariates (age, HIV duration, HCV duration, sex, education)
  - $\mathbf{A}_t = (A_t^{\text{HCV}}, A_t^{\text{cens}})$: treatment and censoring at time t
  - $\mathbf{L}_t$: time-varying confounders (CD_4 count, ARV therapy, HCV treatment, alcohol use)
  - $Y_t = \mathbb{1}(T > t)$: survival indicator

- **Exposure**

$$\bar{a} = (a_0, a_1, \ldots, a_{K-1})$$

  Example: clearance at time 2: $\bar{a} = (0, 1, 1, \ldots, 1)$

- **Outcome**

  T = ESLD-free survival time

- **Target parameter**

$$\Psi^{\bar{a},t}(P) = S_{\bar{a}}(t) = P(T > t \mid do(\bar{a}))$$

  Survival probability under an intervention fixing the clearance pattern.

Efficient Influence Curve (EIC)

Target parameter: survival under treatment regime $\bar{a}$

$$\Psi(P) = S_{\bar{a}}(t) = \mathbb{E}\left[Q_t^{\bar{a}}(1)\right] = \Psi(Q),$$

where

$$Q_j^{\bar{a}}(t) = \mathbb{P}\left(T > t \mid, \bar{A}_{j-1}(1) = \bar{a}_{j-1}, A_{j-1}(2) = 0, \bar{L}_{j-1}, Y_{j-1}\right)$$

Efficient Influence Curve:

$$\boxed{IC^*(\bar{a}, t) = \sum_{j=1}^{t} IC_j(\bar{a}, t)}$$

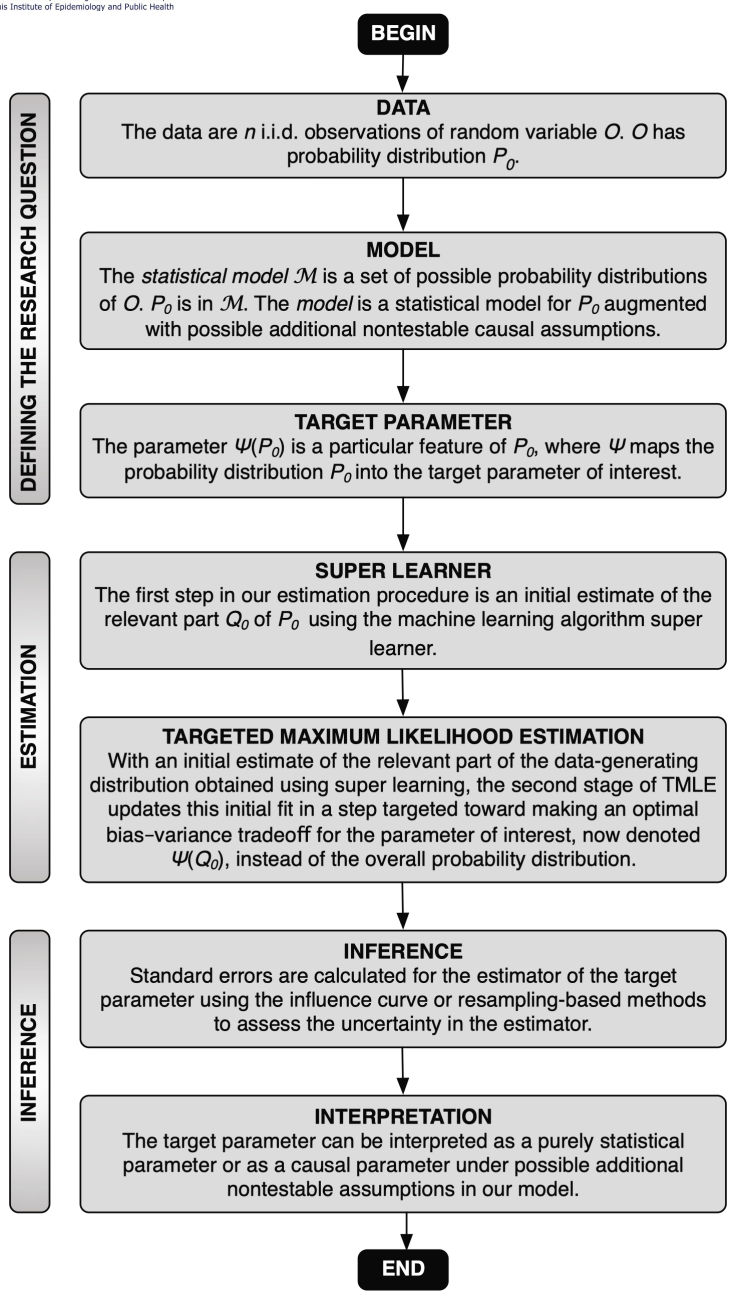
where

$$IC_1(\bar{a}, t) := Q_t^{\bar{a}}(1) - S_{\bar{a}}(t)$$

$$IC_j(\bar{a}, t) := \frac{\mathbb{1}\{\bar{A}_{j-1} = \bar{a}_{j-1},\ A_j(2) = 0\}}{g_{\bar{a}}(j)} \left(Q_j^{\bar{a}}(t) - Q_{j-1}^{\bar{a}}(t)\right)$$

$$g_{\bar{a}}(j) := \prod_{k=1}^{j} P(A_k(1) = a_k \mid \bar{A}_{k-1}, \bar{L}_{k-1}, Y_{k-1} = 1) \cdot P(A_k(2) = 0 \mid \bar{A}_{k-1}, \bar{L}_{k-1}, Y_{k-1} = 1)$$

2. Estimation



1. **Initial estimation (Super Learner)**
   - Estimate outcome regression $\hat{Q}_0$
   - Estimate treatment mechanism $\hat{g}_0$

2. **Targeting step (TMLE fluctuation)**
   Update $\hat{Q}$ using a logistic fluctuation model:

$$\operatorname{logit} Q^*(j) = \operatorname{logit} \hat{Q}(j) + \varepsilon(j)\, H(j)$$

where the clever covariate is:

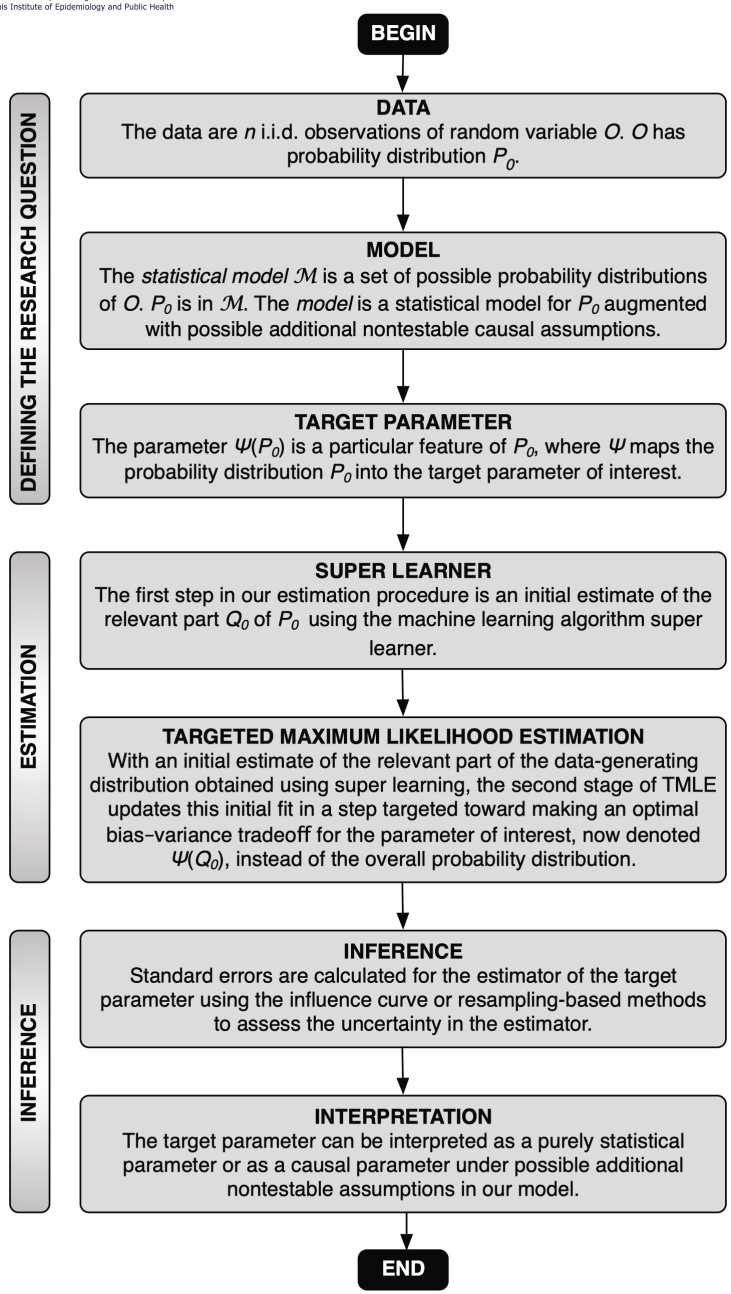
$$H(j) = \frac{\mathbb{1}_{\bar{A}_{j-1}=\bar{a}_{j-1},\, A_j(2)=0}}{\hat{g}_{\bar{a}}(j)}$$

(performed only once for each time j)

$\Rightarrow$ Final estimation :

$$\hat{\Psi}_n^{\bar{a},t} = \hat{S}_{\bar{a}}(t) = \frac{1}{n}\sum_{i=1}^{n}[\hat{Q}_t^{\bar{a},*}(1)]_i$$

3. Inference



1. **Inference**
   - For each $t = 1, \cdots, K$, estimate
   $$\hat{S}_{\bar{a}}(t)$$
   - Compare survival under different exposure patterns
   $$\hat{S}_{\bar{a}_1}(t) \quad \text{vs} \quad \hat{S}_{\bar{a}_2}(t)$$
   - Variance estimated with a sandwich estimator
2. **Interpretation**
   - Estimated survival curves are summarized through a Marginal Structural Model (MSM)
   - Result: clinically relevant but not statistically significant protective effect of HCV clearance on ESLD

7

Comparison to other methods

Competitors

- MLE-based methods
  - **Non-parametric MLE** — Estimates P_0 without parametric assumptions.

Competitors

- MLE-based methods
  - **Non-parametric MLE** — Estimates P_0 without parametric assumptions.
  - **MLE after dimension reduction (propensity score methods)** — Collapses confounders into a single score $g(A \mid W)$ before estimation.

Competitors

- MLE-based methods
  - **Non-parametric MLE** — Estimates P_0 without parametric assumptions.
  - **MLE after dimension reduction (propensity score methods)** — Collapses confounders into a single score $g(A \mid W)$ before estimation.
  - **MLE using a parametric working model** — Specifies a closed-form model for and maximises the likelihood.

Competitors

- MLE-based methods
  - **Non-parametric MLE** — Estimates P_0 without parametric assumptions.
  - **MLE after dimension reduction (propensity score methods)** — Collapses confounders into a single score $g(A \mid W)$ before estimation.
  - **MLE using a parametric working model** — Specifies a closed-form model for and maximises the likelihood.
  - **ML-based super learning** — Data-adaptive ensemble of learners that targets prediction of Q or g.
- Estimating equation methods
  - **IPTW** — Reweights each observation by $1/g(A \mid W)$ to simulate a randomised experiment.

Competitors

- MLE-based methods
  - **Non-parametric MLE** — Estimates P_0 without parametric assumptions.
  - **MLE after dimension reduction (propensity score methods)** — Collapses confounders into a single score $g(A \mid W)$ before estimation.
  - **MLE using a parametric working model** — Specifies a closed-form model for and maximises the likelihood.
  - **ML-based super learning** — Data-adaptive ensemble of learners that targets prediction of Q or g.
- Estimating equation methods
  - **IPTW** — Reweights each observation by $1/g(A \mid W)$ to simulate a randomised experiment.
  - **A-IPTW** — Combines an outcome model Q and a propensity score g via the efficient influence curve.

Some limitations of these methods

- MLE-based methods
  - **Non-parametric MLE** — Intractable in high dimension; does not scale to real-world data structures.

Some limitations of these methods

- MLE-based methods
  - **Non-parametric MLE** — Intractable in high dimension; does not scale to real-world data structures.
  - **Propensity score methods** — Consistent only if $g(A \mid W)$ is correctly specified.

Some limitations of these methods

- MLE-based methods
  - **Non-parametric MLE** — Intractable in high dimension; does not scale to real-world data structures.
  - **Propensity score methods** — Consistent only if $g(A \mid W)$ is correctly specified.
  - **Parametric working model** — Bias is unavoidable under misspecification.

Some limitations of these methods

- MLE-based methods
  - **Non-parametric MLE** — Intractable in high dimension; does not scale to real-world data structures.
  - **Propensity score methods** — Consistent only if $g(A \mid W)$ is correctly specified.
  - **Parametric working model** — Bias is unavoidable under misspecification.
  - **Super learning** — Targets prediction, not Ψ.
- Estimating equation methods
  - **IPTW** — Sensitive to extreme weights; not doubly robust; high variance in finite samples.

Some limitations of these methods

- MLE-based methods
  - **Non-parametric MLE** — Intractable in high dimension; does not scale to real-world data structures.
  - **Propensity score methods** — Consistent only if $g(A \mid W)$ is correctly specified.
  - **Parametric working model** — Bias is unavoidable under misspecification.
  - **Super learning** — Targets prediction, not Ψ.
- Estimating equation methods
  - **IPTW** — Sensitive to extreme weights; not doubly robust; high variance in finite samples.
  - **A-IPTW** — Doubly robust but estimates can fall outside Ψ domain.

8

Useful ressources

R Packages for TMLE

Main package: `tmle`

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.

Other useful TMLE-related packages:

- `ltmle` : longitudinal TMLE for time-dependent treatments and censoring
- `tmle3` : modular and extensible TMLE framework
- `SuperLearner` : machine learning libraries commonly used with TMLE

Because One TMLE Tutorial Was Not Enough

- Books
  - van der Laan, M. J., & Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer.
  - van der Laan, M. J., & Rose, S. (2018). *Targeted Learning in Data Science*. Springer.
- Papers
  - **Original article:** van der Laan, M. J., & Rubin, D. (2006). *Targeted Maximum Likelihood Learning*. The International Journal of Biostatistics.
  - **Tutorial for biostatisticians:** Luque-Fernandez, M. A., Schomaker, M., Rachet, B., & Schnitzer, M. E. (2018). *Targeted maximum likelihood estimation for a binary treatment: A tutorial*. Statistics in Medicine, 37(16), 2530–2546.
  - **TMLE in epidemiology:** Schuler, M. S., & Rose, S. (2017). *Targeted maximum likelihood estimation for causal inference in observational studies*. American Journal of Epidemiology, 185(1), 65–73.
- Online resources
  - Berkeley Course (videos): `https://ctml.berkeley.edu/targeted-learning`
  - TMLE discussion and intuition: `https://stats.stackexchange.com/questions/442569/theory-behind-targeted-maximum-likelihood-estimation-tmle`
  - Interactive tutorial notebook: `https://achambaz.github.io/tlride/`

In Case You Wanted Even More Acronyms

- More on influence curves
  - Fisher, A., & Kennedy, E. H. (2021). *Visually communicating and teaching intuition for influence functions*. The American Statistician, 75(2), 162–172.
  - Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). *Demystifying statistical learning based on efficient influence functions*. The American Statistician, 76(3), 292–304.

- TMLE vs other methods
  - Porter, K. E., Gruber, S., van der Laan, M. J., & Sekhon, J. S. (2011). *The relative performance of targeted maximum likelihood estimators*. The International Journal of Biostatistics, 7(1), 31.
  - Díaz, I. (2020). *Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning*. Biostatistics, 21(2), 353–358.

THANK YOU!

Inserm

iPLesp

SANTÉ
SORBONNE
UNIVERSITÉ