Day 2, Lecture 2

A tiny overview

Summary of TMLE

In this lecture, our goal is to:

1. Summarize the role of the targeting step and the initial estimation steps of the TMLE procedure, and particularly the requirements for performance and consistency of initial estimation for valid inference.

2. Differentiate between applying an existing implementation of TMLE for a specific practical setting or research question, and the steps required for constructing TMLE to address novel research questions or practical scenarios.

Summary of TMLE

"Classical" causal inference estimators:

- consistency of g-formula estimators rely on correct specification of the outcome regression $f(a, x) = \mathbb{E}[Y \mid A = a, X = x]$
- consistency of ipw estimators rely on correct specification of the propensity score $\pi(a \mid x) = P(A = a \mid X = x)$
- both types of estimators work poorly with machine learning (cross-validation does not help)

TMLE (and other estimators based on the efficient influence curve):

- built-in bias correction
- double robustness in consistency
- inference based on the efficient influence curve
- even when incorporating machine learning (under conditions!!)

Summary of TMLE

TMLE is a two-step procedure:

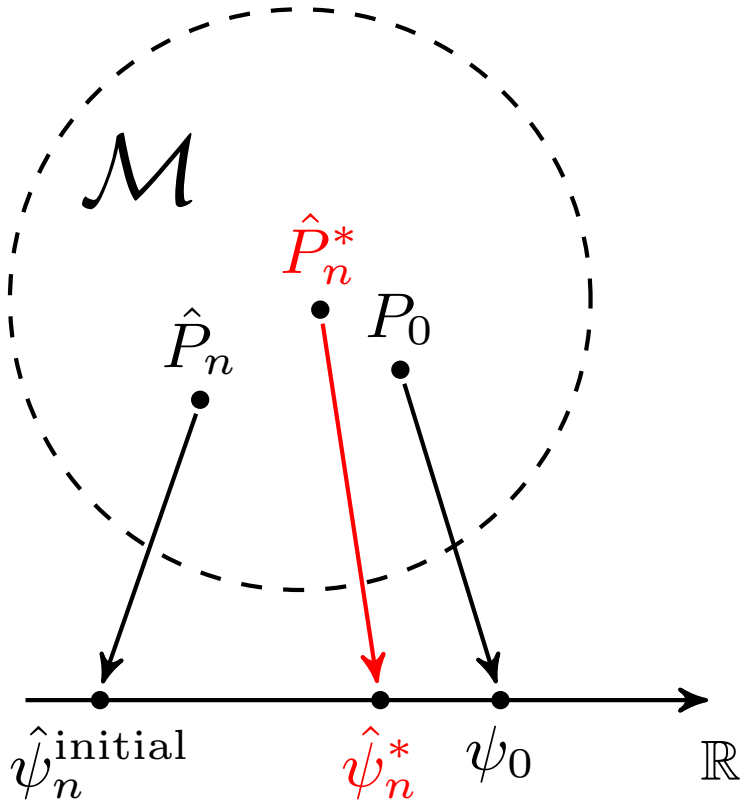
- Step 1 Construct initial estimator $\hat{P}_n$ for P.
- Step 2 Update the estimator $\hat{P}_n \mapsto \hat{P}_n^*$ such that $\hat{P}_n^*$ solves the efficient influence curve equation, i.e.,

$$\mathbb{P}_n \phi^*(\hat{P}_n^*) = \frac{1}{n} \sum_{i=1}^{n} \phi^*(\hat{P}_n^*)(O_i) \approx 0.$$

Step 1 = "initial estimation step"

Step 2 = "targeting step"

Summary of TMLE



- in contrast to the estimating equation (EE) estimator (and the IPW estimator), TMLE is a substitution estimator.
- TMLE *may* show better small-sample performance in settings with unstable weights.

Summary of TMLE

$$\Psi(\hat{P}_n) - \Psi(P_0) = \mathbb{P}_n \phi^*(P_0) + o_P(n^{-1/2})$$
$$+ R(\hat{P}_n, P_0)$$
$$- \mathbb{P}_n \phi^*(\hat{P}_n)$$

- The role of the targeting step (Step 2):
  - Gain double robustness in consistency.
  - Easier to achieve asymptotic lineariy (amounts to getting rid of second-order remainder).

- The role of the initial estimation step (Step 1):
  - This should be done well enough to get rid of the second-order remainder.
  - The second-order remainder tells us if/how machine learning estimators can be incorporated while still achieving asymptotic linearity.

Summary of TMLE

Specifically for the ATE:

$$\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = \mathbb{P}_n \tilde{\phi}^*(f_0, \pi_0) + o_P(n^{-1/2})$$

$$+ \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) - \underbrace{\mathbb{P}_n \tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n)}_{=0, \text{ by targeting.}}$$

When $\mathbb{P}_n \tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n) = 0$, recall that:

$$|\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)| \leq \sum_{a=0,1} \delta^{-1} \left\| \pi_0(a \mid \cdot) - \hat{\pi}_n(a \mid \cdot) \right\|_{\mu_0} \left\| f_0(a \mid \cdot) - \hat{f}_n^*(a \mid \cdot) \right\|_{\mu_0}$$

What this tells us:

- Asymptotic linearity when π_0 and f_0 are estimated at rate at least $n^{-1/4}$.

Summary of TMLE

Specifically for the ATE:

$$\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = \mathbb{P}_n \tilde{\phi}^*(f_0, \pi_0) + o_P(n^{-1/2})$$

$$+ \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) - \underbrace{\mathbb{P}_n \tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n)}_{=0, \text{ by targeting.}}$$

When $\mathbb{P}_n \tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n) = 0$, recall that:

$$|\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)| \leq \sum_{a=0,1} \delta^{-1} \left\| \pi_0(a \mid \cdot) - \hat{\pi}_n(a \mid \cdot) \right\|_{\mu_0} \left\| f_0(a \mid \cdot) - \hat{f}_n^*(a \mid \cdot) \right\|_{\mu_0}$$

What this tells us:

- Asymptotic linearity when π_0 and f_0 are estimated at rate at least $n^{-1/4}$.

We often see this structure of the second-order remainder, but note that it has to be verified on a case-by-case basis.

Summary of TMLE

How can we perform estimation of π_0 and f_0 such as to achieve rate at least $n^{-1/4}$?

- Correctly specified parametric models
  - although consistency is guaranteed, inference cannot be based on the efficient influence curve when one is misspecified!
- There are no results on this being the case for generic implementations of, for example, random forests.
- Lasso, highly adaptive lasso (HAL), ...
- Loss-based "super learning"
  - oracle property: the super learner achieves the rate of convergence of the *best* estimator in its library.

Summary of TMLE

How can we perform estimation of π_0 and f_0 such as to achieve rate at least $n^{-1/4}$?

- Correctly specified parametric models
  - although consistency is guaranteed, inference cannot be based on the efficient influence curve when one is misspecified!
- There are no results on this being the case for generic implementations of, for example, random forests.
- Lasso, highly adaptive lasso (HAL), ...
- Loss-based "super learning"
  - oracle property: the super learner achieves the rate of convergence of the *best* estimator in its library.
  - this is about minimizing expected loss; tuning is still important.

Summary of TMLE

$$f(A, X) = \mathbb{E}_P[Y \mid A, X]$$

A **loss function** $\mathscr{L}(f)(O)$ measuring the distance between an estimator f and the observed outcome Y, e.g., the negative log-likelihood:

$$\mathscr{L}(\hat{f}_n)(Y_i, A_i, X_i) = -(Y_i \log(\hat{f}_n(A_i, X_i)) + (1 - Y_i) \log(1 - \hat{f}_n(A_i, X_i))).$$

- The estimator $\hat{f}_n$ closest to the true f_0 minimizes the risk:

$$\mathbb{E}_{P_0}[\mathscr{L}(\hat{f}_n)(Y_i, A_i, X_i)].$$

- Loss-based super learning: Minimizing the cross-validated empirical risk with respect to the loss function $\mathscr{L}$ over the statistical model.

Summary of TMLE

$f(A, X) = \mathbb{E}_P[Y \mid A, X]$

A **loss function** $\mathscr{L}(f)(O)$ measuring the distance between an estimator f and the observed outcome Y, e.g., the negative log-likelihood:

$$\mathscr{L}(\hat{f}_n)(Y_i, A_i, X_i) = -(Y_i \log(\hat{f}_n(A_i, X_i)) + (1 - Y_i) \log(1 - \hat{f}_n(A_i, X_i))).$$

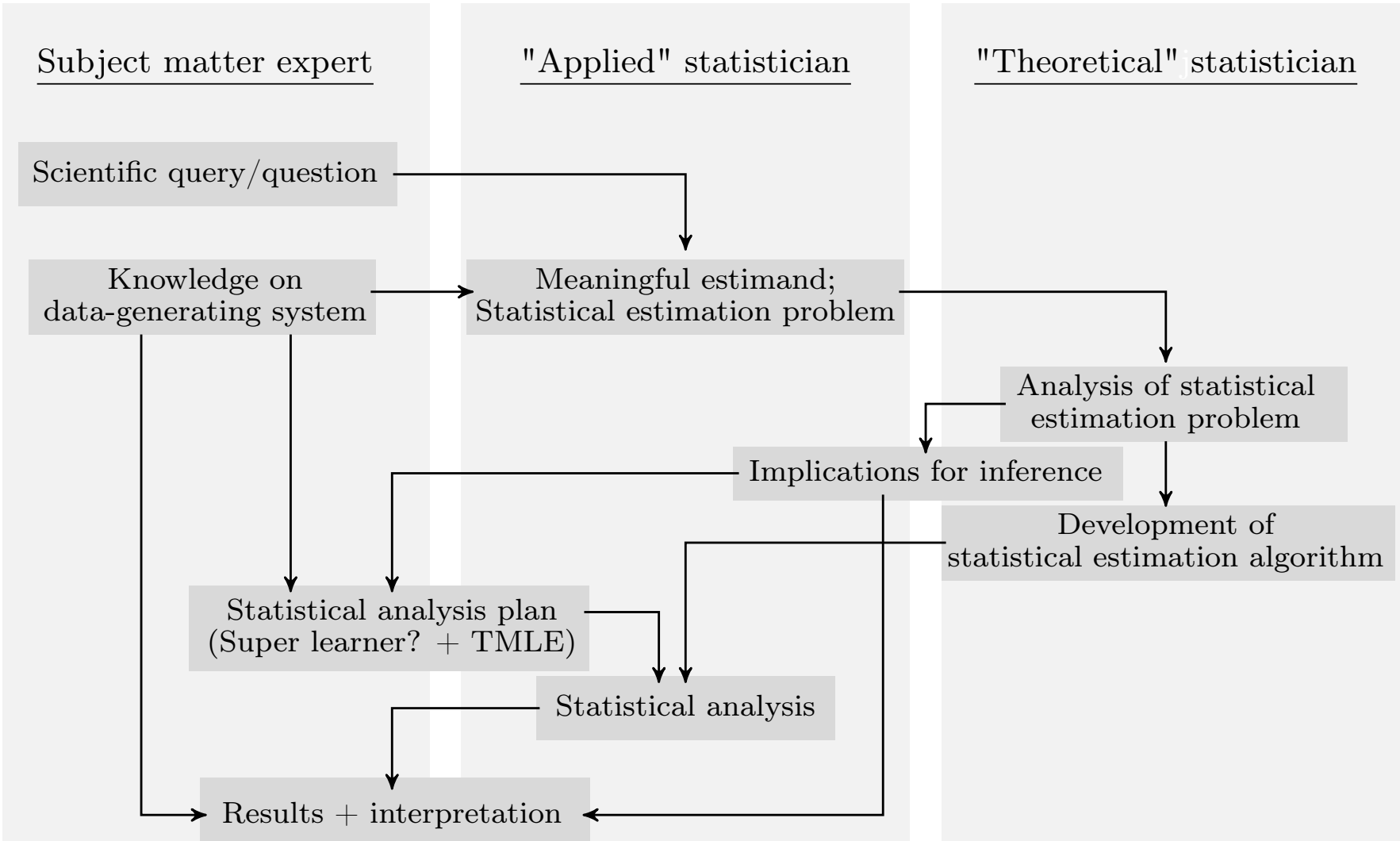
- The estimator $\hat{f}_n$ closest to the true f_0 minimizes the risk:

$$\mathbb{E}_{P_0}[\mathscr{L}(\hat{f}_n)(Y_i, A_i, X_i)].$$

- Loss-based super learning: Minimizing the cross-validated empirical risk with respect to the loss function $\mathscr{L}$ over the statistical model.

Again, this does not usually by simple plug-in yield a good estimator for the particular feature of interest (the target parameter). But, combined with targeting, it is used to get rid of the second-order remainder.

Summary of targeted learning and TMLE



Summary of targeted learning and TMLE

1. Scientific question $\Rightarrow$ causal parameter
2. Causal parameter $\Rightarrow$ statistical parameter
3. Statistical estimation problem = statistical parameter + statistical model
   - ▷ Efficient influence function
   - ▷ Second-order remainder
4. Identify relevant components that need targeting
   - ▷ Submodel + loss function
   - ▷ Targeting algorithm
5. Construct strong initial learners!!
   - ▷ *a priori* specified
6. Inference based on the efficient influence function

Summary of targeted learning and TMLE

... from a theoretical perspective, for any new type of problem.

1. Scientific question $\Rightarrow$ causal parameter
2. Causal parameter $\Rightarrow$ statistical parameter
3. Statistical estimation problem = statistical parameter + statistical model
   - ▷ Efficient influence function
   - ▷ Second-order remainder
4. Identify relevant components that need targeting
   - ▷ Submodel + loss function
   - ▷ Targeting algorithm
5. Construct strong initial learners!!
   - ▷ *a priori* specified
6. Inference based on the efficient influence function

Summary of targeted learning and TMLE

... from a theoretical perspective, for any new type of problem.

1. Scientific question $\Rightarrow$ causal parameter
2. Causal parameter $\Rightarrow$ statistical parameter
3. Statistical estimation problem = statistical parameter + statistical model
   - ▷ Efficient influence function
   - ▷ Second-order remainder
4. Identify relevant components that need targeting
   - ▷ Submodel + loss function
   - ▷ Targeting algorithm
5. Construct strong initial learners!!
   - ▷ *a priori* specified
6. Inference based on the efficient influence function

Summary of targeted learning and TMLE

... from a more applied perspective.

1. Scientific question ⇒ causal parameter
2. Causal parameter ⇒ statistical parameter
3. Statistical estimation problem = statistical parameter + statistical model
   - ▷ Efficient influence function
   - ▷ Second-order remainder
4. Identify relevant components that need targeting
   - ▷ Submodel + loss function
   - ▷ Targeting algorithm
5. Construct strong initial learners!!
   - ▷ *a priori* specified
6. Inference based on the efficient influence function

Comment — substitution estimation

The g-formula estimator and the TMLE are substitution estimators:

$$\hat{\psi}_n^{\text{g-formula}} = \tilde{\Psi}(\hat{f}_n, \hat{\mu}_X) = \int_{\mathbb{R}^d} \big(\hat{f}_n(1,x) - \hat{f}_n(0,x)\big) d\hat{\mu}_X(x)$$

$$\hat{\psi}_n^{\text{tmle}} = \tilde{\Psi}(\hat{f}_n^*, \hat{\mu}_X) = \frac{1}{n}\sum_{i=1}^{n} \big\{\hat{f}_n^*(1,X_i) - \hat{f}_n^*(0,X_i)\big\}$$

The IPW estimator and the EE estimator are <u>not</u> substitution estimators:

$$\hat{\psi}_n^{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{\hat{\pi}_n(A_i \mid X_i)} - \frac{(1-A_i) Y_i}{\hat{\pi}_n(A_i \mid X_i)} \right\}$$

$$\hat{\psi}_n^{\text{ee}} = \hat{\psi}_n^{\text{ee}} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{A_i}{\hat{\pi}_n(A_i \mid X_i)} - \frac{(1-A_i)}{\hat{\pi}_n(A_i \mid X_i)} \big(Y_i - \hat{f}_n(A_i, X_i)\big) + \hat{f}_n(1,X_i) - \hat{f}_n(0,X_i) \right\}$$

Comment — substitution estimation

The g-formula estimator and the TMLE are both based on the parametrization of the ATE as follows:

$$\tilde{\Psi}(f, \mu_X) = \mathbb{E}_P[f(1, X) - f(0, X)] = \int_{\mathbb{R}^d} \big(f(1, x) - f(0, x)\big)\mu_X(x) \quad (1)$$

corresponding to an average outcome under the post-interventional distribution.

- This will always have a statistical interpretation as a covariate-adjusted difference of treatment-specific risks (but may have no causal validity).
- Estimators based on plugging estimators for f (respecting the model constraints for f) into (1) will always respect the global constraints of the observed data model.

Comment — substitution estimation

The g-formula estimator and the TMLE are both based on the parametrization of the ATE as follows:

$$\tilde{\Psi}(f, \mu_X) = \mathbb{E}_P[f(1, X) - f(0, X)] = \int_{\mathbb{R}^d} \big(f(1, x) - f(0, x)\big)\mu_X(x) \quad (1)$$

corresponding to an average outcome under the post-interventional distribution.

- This will always have a statistical interpretation as a covariate-adjusted difference of treatment-specific risks (but may have no causal validity).
- Estimators based on plugging estimators for f (respecting the model constraints for f) into (1) will always respect the global constraints of the observed data model. = the substitution property