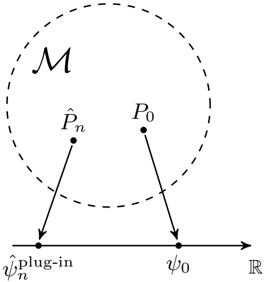
Targeted nonparametric inference

Targeted nonparametric inference

Key idea:

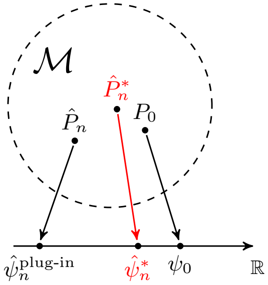
we should focus estimation on our target $\psi_0 = \Psi(P_0)$ specifically



Targeted nonparametric inference

Key idea:

we should focus estimation on our target $\psi_0 = \Psi(P_0)$ specifically



Targeted nonparametric inference

Key (technical) concepts we cannot avoid talking about:

* asympotically linear estimation
* efficient influence curve

Targeted nonparametric inference

Key (technical) concepts we cannot avoid talking about:

* asympotically linear estimation
* efficient influence curve
* von Mises expansions

Targeted nonparametric inference

Key (technical) concepts we cannot avoid talking about:

* asympotically linear estimation
* efficient influence curve
* von Mises expansions
* second-order remainders

Targeted nonparametric inference

Key (technical) concepts we cannot avoid talking about:

* asympotically linear estimation
* efficient influence curve
* von Mises expansions
* second-order remainders

Targeted nonparametric inference

Key (technical) concepts we cannot avoid talking about:

* asympotically linear estimation
* efficient influence curve
* von Mises expansions
* second-order remainders

While TMLE can be applied without knowing about these concepts ...

... we cannot understand the purpose of TMLE without knowing a bit about the efficient influence curve.

... to understand what conditions are required for TMLE, we need to understand the second-order remainder.

Targeted nonparametric inference

Recap notation:[1]

- For a function $h : \mathcal{O} \to \mathbb{R}$ and distribution P

$$Ph = \mathbb{E}_P[h(O)] = \int h dP = \int_{\mathcal{O}} h(o) dP(o)$$

  where $\mathcal{O} = \mathbb{R}^d \times \{0, 1\} \times \{0, 1\}$ is the sample space of $O = (X, A, Y)$.

- For the empirical measure $\mathbb{P}_n$ of the sample $O_1, \ldots, O_n$:

$$\mathbb{P}_n h = \int h d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} h(O_i);$$

  note: the right-hand-side is really just the empirical average.

- $X_n = o_P(1)$ means that $X_n \xrightarrow{P} 0$; $X_n = o_P(n^{-1/2})$ means that $n^{1/2} X_n \xrightarrow{P} 0$.

[1] van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

Asymptotic linearity

A very desirable property —

[2] $o_P(1)$ denotes a sequence which is converges to zero in probability.

Asymptotic linearity

The empirical measure $\mathbb{P}_n$ of the sample $O_1, \ldots, O_n$:
$$\mathbb{P}_n h = \int h d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} h(O_i).$$

A very desirable property —

An estimator $\hat{\psi}_n$ is $\sqrt{n}$-consistent and asymptotically linear with influence function $\phi(P_0)(O)$ if [2]

$$\sqrt{n}\left(\hat{\psi}_n - \psi_0\right) = \sqrt{n}\,\mathbb{P}_n\phi(P_0) + o_P(1),$$

where $\mathbb{E}_{P_0}[\phi(P_0)(O)] = 0$ and $\mathbb{E}_{P_0}[\{\phi(P_0)(O)\}^2] < \infty$.

---

[2] $o_P(1)$ denotes a sequence which is converges to zero in probability.

Asymptotic linearity

The empirical measure $\mathbb{P}_n$ of the sample $O_1, \ldots, O_n$:

$$\mathbb{P}_n h = \int h d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} h(O_i).$$

A very desirable property —

An estimator $\hat{\psi}_n$ is $\sqrt{n}$-consistent and asymptotically linear with influence function $\phi(P_0)(O)$ if [2]

$$\sqrt{n}\big(\hat{\psi}_n - \psi_0\big) = \sqrt{n}\,\mathbb{P}_n\phi(P_0) + o_P(1),$$

where $\mathbb{E}_{P_0}[\phi(P_0)(O)] = 0$ and $\mathbb{E}_{P_0}[\{\phi(P_0)(O)\}^2] < \infty$.

Then CLT + Slutsky implies:

$$\hat{\psi}_n \overset{as}{\sim} N(\Psi(P_0), \mathrm{Var}(\phi(P_0))/n).$$

The estimator behaves asymptotically as an average of the influence function.

[2] $o_P(1)$ denotes a sequence which is converges to zero in probability.

Asymptotic linearity

Simple example: Estimator for the mean $\psi_0 = \mathbb{E}[X]$:

$$\hat{\psi}_{n,0} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then

$$\sqrt{n}(\hat{\psi}_n - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} = \sqrt{n}\mathbb{P}_n\phi(P_0)$$

$\hat{\psi}_{n,0}$ is linear and thus asymptotically linear.

Asymptotic linearity

Simple example: Estimator for the mean $\psi_0 = \mathbb{E}[X]$:

$$\hat{\psi}_{n,1} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}$$

Then

$$\sqrt{n}(\hat{\psi}_n - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} + \frac{\sqrt{n}}{n} = \sqrt{n}\mathbb{P}_n\phi(P_0) + \underbrace{\frac{1}{\sqrt{n}}}_{=o(1)}$$

$\hat{\psi}_{n,1}$ is asymptotically linear.

Asymptotic linearity

Simple example: Estimator for the mean $\psi_0 = \mathbb{E}[X]$:

$$\hat{\psi}_{n,2} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n^{1/2+0.1}}$$

Then

$$\sqrt{n}(\hat{\psi}_n - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} + \frac{\sqrt{n}}{n^{1/2+0.1}} = \sqrt{n}\mathbb{P}_n\phi(P_0) + \underbrace{\frac{1}{n^{0.1}}}_{=o(1)}$$

$\hat{\psi}_{n,2}$ is asymptotically linear.

Asymptotic linearity

Simple example: Estimator for the mean $\psi_0 = \mathbb{E}[X]$:

$$\hat{\psi}_{n,3} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n^{1/2-0.1}}$$

Then

$$\sqrt{n}(\hat{\psi}_n - \psi_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{(X_i - \psi_0)}_{=\phi(P_0)(O_i)} + \frac{\sqrt{n}}{n^{1/2-0.1}} = \sqrt{n}\mathbb{P}_n\phi(P_0) + \underbrace{n^{0.1}}_{\to\infty}$$

$\hat{\psi}_{n,3}$ is **not** asymptotically linear.

Asymptotic linearity

An estimator $\hat{\psi}_n$ has rate of convergence $r_n \to \infty$ if [3]

$$r_n(\hat{\psi}_n - \psi_0) = O_P(1), \quad \text{i.e.,} \quad \hat{\psi}_n - \psi_0 = O_P(1/r_n).$$

The convergence rate r_n tells us how fast $\hat{\psi}_n$ centers around ψ_0, with the difference $\hat{\psi}_n - \psi_0$ behaving like $1/r_n$.

- One wants negligible bias such as to obtain reliable confidence intervals for ψ_0.
- The bias of an asymptotically linear estimator converges to zero at a rate faster than $1/\sqrt{n}$.

Data-adaptive machine learning estimators rarely achieve this rate.

[3] $O_P(1)$ denotes a sequence which is bounded in probability.

Asymptotic linearity

$$\sqrt{n}\hat{\psi}_{n,1} = \underbrace{\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n}}_{\to 0}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,1} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,2} = \underbrace{\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2+0.1}}}_{\to 0}, \quad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,3} = \underbrace{\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overset{P}{\to}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2-0.1}}}_{\to \infty}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) \overset{P}{\to} \infty.$$

Asymptotic linearity

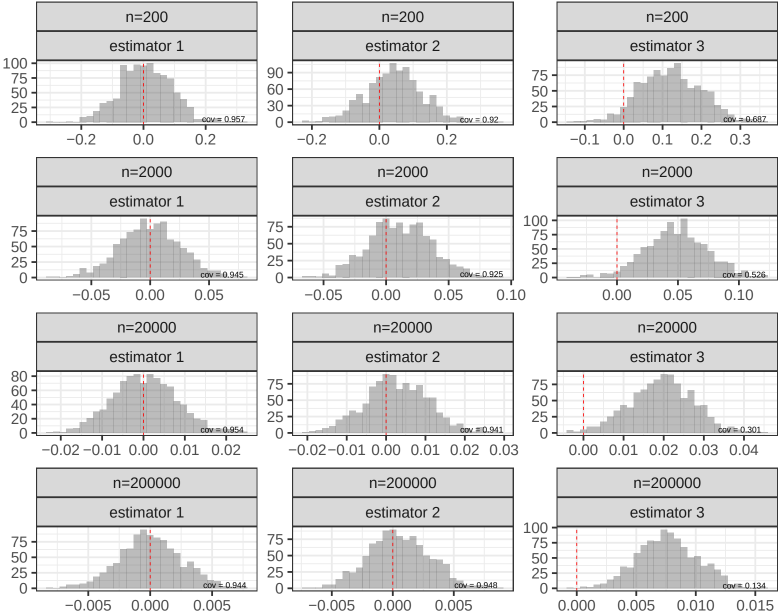
$$\sqrt{n}\hat{\psi}_{n,1} = \underbrace{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i}_{\xrightarrow{P}\psi_0} + \underbrace{\frac{\sqrt{n}}{n}}_{\to 0}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,1} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,2} = \underbrace{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i}_{\xrightarrow{P}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2+0.1}}}_{\to 0}, \quad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) = o_P(1).$$

$$\sqrt{n}\hat{\psi}_{n,3} = \underbrace{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i}_{\xrightarrow{P}\psi_0} + \underbrace{\frac{\sqrt{n}}{n^{1/2-0.1}}}_{\to\infty}, \qquad \text{i.e.,} \quad \sqrt{n}(\hat{\psi}_{n,3} - \psi_0) \xrightarrow{P} \infty.$$

[The remainder term that determines the asymptotic bias the estimator].

Asymptotic linearity



A quick run-through of the theoretical basis for targeted nonparametric inference

The decomposition that guides the construction of TMLE.

The von Mises expansion and the efficient influence curve

A key component in constructing a $\sqrt{n}$-consistent and asymptotically linear estimator, *even when using machine learning estimation*, is the so-called the efficient influence function (also known as the canonical gradient).

The von Mises expansion and the efficient influence curve

The von Mises expansion:

Suppose the functional (the target parameter) $\Psi : \mathcal{M} \to \mathbb{R}$ is sufficiently smooth (as a map from distributions to the real line), in the sense that it admits a certain distributional Taylor expansion

$$\Psi(P) - \Psi(P') = \int \phi(P)(o) d(P - P')(o) + R_2(P, P'), \qquad (1)$$

for distributions $P, P' \in \mathcal{M}$ for a function ϕ satisfying $P\phi(P) = 0$ (mean zero) and $P\phi(P)^2 < \infty$ (finite variance).

Intuitively, the von Mises expansion is just a distributional analogue of a Taylor expansion, with the function $\phi(P)$ acting as a usual derivative term; it describes how the functional Ψ changes locally when the distribution changes from P to P'.

The von Mises expansion and the efficient influence curve

When the model $\mathcal{M}$ is assumed properly nonparametric, there exists *one* function $\phi(P)$. This is called the efficient influence curve; we also denote it $\phi^*(P)$.[4]

The efficient influence curve in nonparametric models indicates how to construct asymptotically linear (and efficient) estimators.

[4] This may be confusing here, but it is useful in restricted (semi)parametric models, where multiple ϕ's can satisfy (1). For these situations we by the way have that $P_0\phi(P_0)^2 \geq P_0\phi^*(P_0)^2$.

The von Mises expansion and the efficient influence curve

When the model $\mathcal{M}$ is assumed properly nonparametric, there exists *one* function $\phi(P)$. This is called the efficient influence curve; we also denote it $\phi^*(P)$.[4]

The efficient influence curve in nonparametric models indicates how to construct asymptotically linear (and efficient) estimators.

Also note that the efficient curve has many names (influence function, pathwise derivative, Neyman orthogonal score, canonical gradient).

[4] This may be confusing here, but it is useful in restricted (semi)parametric models, where multiple ϕ's can satisfy (1). For these situations we by the way have that $P_0\phi(P_0)^2 \geq P_0\phi^*(P_0)^2$.

Goal: $\sqrt{n}$-consistency and asymptotically linearity

An estimator $\hat{\psi}_n$ is $\sqrt{n}$-consistent and asymptotically linear with influence function $\phi(P_0)(O)$ if

$$\sqrt{n}\big(\hat{\psi}_n - \psi_0\big) = \sqrt{n}\,\mathbb{P}_n\phi(P_0) + o_P(1),$$

where $\mathbb{E}_{P_0}[\phi(P_0)(O)] = 0$ and $\mathbb{E}_{P_0}[\{\phi(P_0)(O)\}^2] < \infty$.

Then CLT + Slutsky implies:

$$\hat{\psi}_n \overset{as}{\sim} N(\Psi(P_0), \mathrm{Var}(\phi(P_0))/n).$$

The estimator behaves asymptotically as an average of the influence function.[5]

[5] One may also note that the efficient influence curve characterizes the estimator with the smallest variance.

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n^*$ and the true data-generating P_0:

$$\Psi(\hat{P}_n^*) - \Psi(P_0) = (\hat{P}_n^* - P_0)\phi^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0)$$

$(*1)$

$(*2)$

$(*3)$

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n^*$ and the true data-generating P_0:

$$\begin{aligned}\Psi(\hat{P}_n^*) - \Psi(P_0) &= (\hat{P}_n^* - P_0)\phi^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0) \\ &= -P_0\phi^*(\hat{P}_n^*) + R_2(P_0, \hat{P}_n^*)\end{aligned}$$

$(*1)$

$(*2)$

$(*3)$

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n^*$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n^*) - \Psi(P_0) &= (\hat{P}_n^* - P_0)\phi^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0) \\
&= -P_0\phi^*(\hat{P}_n^*) + R_2(P_0, \hat{P}_n^*) \\
&\qquad + \mathbb{P}_n\phi^*(\hat{P}_n^*) - \mathbb{P}_n\phi^*(\hat{P}_n^*) \\
&\qquad + (\mathbb{P}_n - P_0)\phi^*(P_0)
\end{aligned}$$

$(*1)$

$(*2)$

$(*3)$

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n^*$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n^*) - \Psi(P_0) &= (\hat{P}_n^* - P_0)\phi^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0) \\
&= -P_0\phi^*(\hat{P}_n^*) + R_2(P_0, \hat{P}_n^*) \\
&\qquad + \mathbb{P}_n\phi^*(\hat{P}_n^*) - \mathbb{P}_n\phi^*(\hat{P}_n^*) \\
&\qquad + (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&\qquad + (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n^*) - \phi^*(P_0)\big) && (*1) \\
&\qquad + R_2(\hat{P}_n^*, P_0) && (*2) \\
&\qquad - \mathbb{P}_n\phi^*(\hat{P}_n^*) && (*3)
\end{aligned}$$

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n^*$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n^*) - \Psi(P_0) &= (\hat{P}_n^* - P_0)\phi^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0) \\
&= -P_0\phi^*(\hat{P}_n^*) + R_2(P_0, \hat{P}_n^*) \\
&\qquad + \mathbb{P}_n\phi^*(\hat{P}_n^*) - \mathbb{P}_n\phi^*(\hat{P}_n^*) \\
&\qquad + (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= \mathbb{P}_n\phi^*(P_0) \\
&\qquad + (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n^*) - \phi^*(P_0)\big) && (*1) \\
&\qquad + R_2(\hat{P}_n^*, P_0) && (*2) \\
&\qquad - \mathbb{P}_n\phi^*(\hat{P}_n^*) && (*3)
\end{aligned}$$

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n^*$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n^*) - \Psi(P_0) &= (\hat{P}_n^* - P_0)\phi^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0) \\
&= -P_0\phi^*(\hat{P}_n^*) + R_2(P_0, \hat{P}_n^*) \\
&\qquad + \mathbb{P}_n\phi^*(\hat{P}_n^*) - \mathbb{P}_n\phi^*(\hat{P}_n^*) \\
&\qquad + (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= \mathbb{P}_n\phi^*(P_0) + o_P(n^{-1/2}) \\
&\qquad + (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n^*) - \phi^*(P_0)\big) && (*1) \\
&\qquad + R_2(\hat{P}_n^*, P_0) && (*2) \\
&\qquad - \mathbb{P}_n\phi^*(\hat{P}_n^*) && (*3)
\end{aligned}$$

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

Evaluating the von Mises expansion in an estimator $\hat{P}_n^*$ and the true data-generating P_0:

$$\begin{aligned}
\Psi(\hat{P}_n^*) - \Psi(P_0) &= (\hat{P}_n^* - P_0)\phi^*(\hat{P}_n^*) + R_2(\hat{P}_n^*, P_0) \\
&= -P_0\phi^*(\hat{P}_n^*) + R_2(P_0, \hat{P}_n^*) \\
&\qquad + \mathbb{P}_n\phi^*(\hat{P}_n^*) - \mathbb{P}_n\phi^*(\hat{P}_n^*) \\
&\qquad + (\mathbb{P}_n - P_0)\phi^*(P_0) \\
&= \mathbb{P}_n\phi^*(P_0) + o_P(n^{-1/2}) \\
&\qquad + (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n^*) - \phi^*(P_0)\big) && (*1) \\
&\qquad + R_2(\hat{P}_n^*, P_0) && (*2) \\
&\qquad - \mathbb{P}_n\phi^*(\hat{P}_n^*) && (*3)
\end{aligned}$$

i.e., need $(*1)$–$(*3)$ to be $o_P(n^{-1/2})$.

Estimator expansion

An estimator $\hat{\psi}_n$ is asymptotically linear if,

$$\sqrt{n}\big(\hat{\psi}_n - \Psi(P_0)\big) = \sqrt{n}\,\mathbb{P}_n\phi^*(P_0) + o_P(1). \quad (*)$$

$$\begin{aligned}
\Psi(\hat{P}_n^*) - \Psi(P_0) = \mathbb{P}_n\phi^*(P_0) + o_P(n^{-1/2}) & \\
+ (\mathbb{P}_n - P_0)\big(\phi^*(\hat{P}_n^*) - \phi^*(P_0)\big) & \qquad (*1)\\
+ R_2(\hat{P}_n^*, P_0) & \qquad (*2)\\
- \mathbb{P}_n\phi^*(\hat{P}_n^*) & \qquad (*3)
\end{aligned}$$

- $(*1)$ is an empirical process term.
- $(*2)$ second-order bias term.
- $(*3)$ is called the efficient influence curve equation.

Estimator expansion

... about the empirical process term $(*1)$:

1. can be handled by empirical process theory, if $(\phi^*(P) : P \in \mathcal{M})$ is assumed Donsker.[6]
2. otherwise can handled by extra sample splitting.

[6] Lemma 19.24 of van der Vaart, A. W. (2000): Asymptotic statistics yields then that $(\mathbb{P}_n - P_0)(\phi^*(\hat{P}_n) - \phi^*(P_0)) = o_P(n^{-1/2})$.

Estimator expansion

Side note: Usually, we will assume the Donsker class condition.

- this is a way of nonparametrically characterizing the complexity of nuisance parameters.
- classes of functions that are Donsker: Indicator functions, bounded monotone functions, Lipschitz parametric functions, smooth functions, ...

Donsker classes also include traditional parametric functions.

We will not discuss this further. For a nice intro see Sections 4.2 and 4.3 of Kennedy, E. H. (2016): Semiparametric theory and empirical processes in causal inference.

Estimator expansion

That is it.

Estimator expansion

That is it.

Conditions (asymptotic linearity and efficiency)

(C1) Solve the efficient influence curve equation: $\mathbb{P}_n \phi^*(\hat{P}_n) = o_P(n^{-1/2})$.

(C2) Remainder $R(\hat{P}_n, P_0) = o_P(n^{-1/2})$.

(C3) Donsker class conditions for $\{\phi^*(P) : P \in \mathcal{M}\}$.

Then: $\Psi(\hat{P}_n) \overset{as}{\sim} N(\Psi(P_0), P_0\phi^*(P_0)^2/n)$.

Construction of estimators

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

For **a given target parameter** $\Psi : \mathcal{M} \to \mathbb{R}$, we need to

1. Know the efficient influence curve, so that we can solve the efficient influence curve equation.
2. Analyze the remainder $R(P, P_0) := \Psi(P) - \Psi(P_0) + P_0 \phi^*(P)$.

NB: These are solely properties of the estimation problem, but also tell us how to construct estimators such as TMLE.

Example: ATE estimation

ATE estimation

EXAMPLE: Average treatment effect (ATE)

Observed data $O = (X, A, Y) \in \mathbb{R}^d \times \{0,1\} \times \{0,1\} = \mathcal{O}$

* $X \in \mathbb{R}^d$ are covariates
* $A \in \{0,1\}$ is a binary exposure variable (treatment decision)
* $Y \in \{0,1\}$ is a binary outcome variable

$O \sim P_0$ where P_0 assumed to belong to nonparametric model $\mathcal{M}$.

We are interested in estimating the ATE:

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

ATE estimation

EXAMPLE: Average treatment effect (ATE)

1. The efficient influence function:

$$\begin{aligned}\phi^*(P)(O) &= \tilde{\phi}^*(f,\pi)(O) \\ &= \left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big) + f(1,X) - f(0,X) - \Psi(P)\end{aligned}$$

ATE estimation

EXAMPLE: Average treatment effect (ATE)

1. The efficient influence function:

$$\begin{aligned}\phi^*(P)(O) &= \tilde{\phi}^*(f,\pi)(O)\\ &= \left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big) + f(1,X) - f(0,X) - \Psi(P)\end{aligned}$$

2. The remainder:

$$\begin{aligned}R(P,P_0) &= \tilde{R}(f,\pi,f_0,\pi_0)\\ &= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big) d\mu_{0,X}(x)\end{aligned}$$

ATE estimation

$$f(A,X) = \mathbb{E}_P[Y \mid A, X],\ \pi(A \mid X) = P(A = a \mid X)$$
$$f_0(A,X) = \mathbb{E}_{P_0}[Y \mid A, X],\ \pi_0(A \mid X) = P(A = a \mid X)$$

$$R(P, P_0) := \Psi(P) - \Psi(P_0) + P_0\phi^*(P).$$

2. Deriving the remainder for the ATE:

$$\begin{aligned}
R(P,P_0) &= \mathbb{E}_P[f(1,X) - f(0,X)] - \mathbb{E}_{P_0}[f_0(1,X) - f_0(0,X)] \\
&+ \mathbb{E}_{P_0}\left[\left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big)\right] \\
&+ \mathbb{E}_{P_0}\big[f(1,X) - f(0,X)\big] - \Psi(P) \\
&\overset{*}{=} \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\left(\frac{\pi_0(a \mid x)}{\pi(a \mid x)} - 1\right)\big(f_0(a,x) - f(a,x)\big) d\mu_{0,X}(x) \\
&= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big) d\mu_{0,X}(x)
\end{aligned}$$

the equality marked by $*$ is detailed on the next slide.

ATE estimation

We used that:

$$
\begin{aligned}
&\mathbb{E}_{P_0}\left[\left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big)\right] \\
&= \mathbb{E}_{P_0}\left[\frac{2A-1}{\pi(A \mid X)}\big(Y - f(A,X)\big)\right] \\
&= \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}\left[\frac{2A-1}{\pi(A \mid X)}\big(Y - f(A,X)\big)\,\middle|\, A, X\right]\right] \\
&= \mathbb{E}_{P_0}\left[\frac{2A-1}{\pi(A \mid X)}\big(f_0(A,X) - f(A,X)\big)\right] \\
&= \int_{\mathbb{R}^d} \sum_{a=0,1} \frac{2a-1}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big)\pi_0(a \mid x)\, d\mu_{0,X}(x) \\
&= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\frac{\pi_0(a \mid x)}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big)\, d\mu_{0,X}(x)
\end{aligned}
$$

ATE estimation

The remainder determines the asymptotic bias.

For the ATE, the remainder has a really nice structure!

$$
\begin{aligned}
R(P, P_0) &= \tilde{R}(f, \pi, f_0, \pi_0) \\
&= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1) \frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)} \big(f_0(a, x) - f(a, x) \big) d\mu_{0,X}(x)
\end{aligned}
$$

A "double robust" structure, which has some important implications.

ATE estimation

$$
\begin{aligned}
|R(P, P_0)| &= |\tilde{R}(f, \pi, f_0, \pi_0)| \\
&\leq \sum_{a=0,1} \int_{\mathbb{R}^d} \frac{|\pi_0(a \mid x) - \pi(a \mid x)|}{\pi(a \mid x)} |f_0(a, x) - f(a, x)| d\mu_{0,X}(x)
\end{aligned}
$$

ATE estimation

$$
\begin{aligned}
|R(P, P_0)| &= |\tilde{R}(f, \pi, f_0, \pi_0)| \\
&\leq \sum_{a=0,1} \int_{\mathbb{R}^d} \frac{|\pi_0(a \mid x) - \pi(a \mid x)|}{\pi(a \mid x)} |f_0(a, x) - f(a, x)| d\mu_{0,X}(x) \\
&\overset{*}{\leq} \sum_{a=0,1} \frac{1}{\pi(a \mid x)} \sqrt{\int_{\mathbb{R}^d} \left\{\pi_0(a \mid x) - \pi(a \mid x)\right\}^2 d\mu_{0,X}(x)} \\
&\qquad\qquad \times \sqrt{\int_{\mathbb{R}^d} \left\{f_0(a, x) - f(a, x)\right\}^2 d\mu_{0,X}(x)}
\end{aligned}
$$

$*$ uses Cauchy-Schwarz.

ATE estimation

$$
\begin{aligned}
|R(P, P_0)| &= |\tilde{R}(f, \pi, f_0, \pi_0)| \\
&\leq \sum_{a=0,1} \int_{\mathbb{R}^d} \frac{|\pi_0(a \mid x) - \pi(a \mid x)|}{\pi(a \mid x)} |f_0(a, x) - f(a, x)| d\mu_{0,X}(x) \\
&\overset{*}{\leq} \sum_{a=0,1} \frac{1}{\pi(a \mid x)} \sqrt{\int_{\mathbb{R}^d} \left\{\pi_0(a \mid x) - \pi(a \mid x)\right\}^2 d\mu_{0,X}(x)} \\
&\qquad \times \sqrt{\int_{\mathbb{R}^d} \left\{f_0(a, x) - f(a, x)\right\}^2 d\mu_{0,X}(x)}
\end{aligned}
$$

Thus, if $\pi(a \mid X) > \delta > 0$ a.s., then:

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

$*$ uses Cauchy-Schwarz.

ATE estimation

What does this imply for estimation?

Double robustness in consistency

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

then $\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = o_P(1)$.

Asymptotic linearity (easier to establish due to double robust structure)

$$|\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)| \leq \sum_{a=0,1} \delta^{-1} \underbrace{\|\pi_0(a \mid \cdot) - \hat{\pi}_n(a \mid \cdot)\|_{\mu_0}}_{=o_P(n^{-1/4})} \underbrace{\|f_0(a \mid \cdot) - \hat{f}_n^*(a \mid \cdot)\|_{\mu_0}}_{=o_P(n^{-1/4})}$$

i.e., $\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) = o_P(n^{-1/2})$.

I.e., bias is converging at fast enough rate for reliable confidence intervals.

ATE estimation

Side note: Showing the double robustness in consistency ...

Say we have estimators $(\hat{f}_n^*, \hat{\pi}_n)$;

- converging to (f, π)
- solving the efficient influence curve equation.

Per definition, $\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) = \tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) + P_0\tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n)$.

I.e.,
$$\begin{aligned}\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) &= -P_0\tilde{\phi}^*(\hat{f}_n^*, \hat{\pi}_n) + \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) \\ &= (\mathbb{P}_n - P_0)\phi^*(\hat{f}_n^*, \hat{\pi}_n) + \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0)\end{aligned}$$

The first term is an empirical process term which equals $(\mathbb{P}_n - P_0)\tilde{\phi}^*(f, \pi)$ plus an $o_P(n^{-1/2})$-term.

This then gives

$$\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = \underbrace{(\mathbb{P}_n - P_0)\tilde{\phi}^*(f, \pi)}_{\text{LLN applies}} + \tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) + o_P(n^{-1/2})$$

which yields that $\tilde{\Psi}(\hat{f}_n^*) - \tilde{\Psi}(f_0) = o_P(1)$ if $\tilde{R}(\hat{f}_n^*, \hat{\pi}_n, f_0, \pi_0) = o_P(1)$.

Analysis of a concrete estimation problem

EXAMPLE: Average treatment effect (ATE)

1. The efficient influence function:

$$\phi^*(P)(O) = \tilde{\phi}^*(f,\pi)(O)$$
$$= \left(\frac{A}{\pi(A \mid X)} - \frac{1-A}{\pi(A \mid X)}\right)\big(Y - f(A,X)\big) + f(1,X) - f(0,X) - \Psi(P)$$

2. The remainder:

$$R(P,P_0) = \tilde{R}(f,\pi,f_0,\pi_0)$$
$$= \int_{\mathbb{R}^d} \sum_{a=0,1} (2a-1)\frac{\pi_0(a \mid x) - \pi(a \mid x)}{\pi(a \mid x)}\big(f_0(a,x) - f(a,x)\big)d\mu_{0,X}(x)$$

Deriving these is done once for a given target parameter $\Psi : \mathcal{M} \to \mathbb{R}$.

TMLE

Conditions (asymptotic linearity and efficiency)

(C1) Solve the efficient influence curve equation: $\mathbb{P}_n \phi^*(\hat{P}_n) = o_P(n^{-1/2})$

(C2) Remainder $R(\hat{P}_n, P_0) = o_P(n^{-1/2})$

(C3) Donsker class conditions for $\{\phi^*(P) : P \in \mathcal{M}\}$

Then: $\Psi(\hat{P}_n) \overset{as}{\sim} N(\Psi(P_0), P_0\phi^*(P_0)^2/n)$

TMLE is a two-step procedure:

- Step 1 Construct initial estimator $\hat{P}_n$ for P such that $R(\hat{P}_n, P_0) = o_P(n^{-1/2})$.
- Step 2 Update the estimator $\hat{P}_n \mapsto \hat{P}_n^*$ such that $\hat{P}_n^*$ solves the efficient influence curve equation.[7]

[7] TMLE and the estimating equation estimator both solve the efficient influence curve equation ... that is why they have the same asymptotic (large-sample) properties.

TMLE

- The role of the targeting step (Step 2):
  - Gaining double robustness in consistency.
  - Easier to get 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.

Practical 2: Continued explorations based on simulated data

In this exercise we continue the simulation setting of Practical 1.
Now we want to explore —

1. Inference for estimators based on the efficient influence curve;
   - TMLE;
   - Estimating equation (EE) estimator;
2. Variance estimation and coverage of confidence intervals;
3. Small-sample properties, particularly under positivity violations.

This is described in detail in: **day1-practical2.pdf**.

Practical 2: Continued explorations based on simulated data

NB —

- these exercises emphasize the asymptotic equivalence of TMLE and estimating equation (EE) estimation;

Practical 2: Continued explorations based on simulated data

NB —

- these exercises emphasize the asymptotic equivalence of TMLE and estimating equation (EE) estimation;
- there may be small-sample differences in performance (often argued in favor of TMLE);

Practical 2: Continued explorations based on simulated data

NB —

- these exercises emphasize the asymptotic equivalence of TMLE and estimating equation (EE) estimation;
- there may be small-sample differences in performance (often argued in favor of TMLE);
- otherwise, the differences are not so important for the ATE estimation problem. BUT, for other problems (e.g., in survival analysis), the substitution property of TMLE may be crucial!

Practical 2: Continued explorations based on simulated data

NB —

- these exercises emphasize the asymptotic equivalence of TMLE and estimating equation (EE) estimation;
- there may be small-sample differences in performance (often argued in favor of TMLE);
- otherwise, the differences are not so important for the ATE estimation problem. BUT, for other problems (e.g., in survival analysis), the substitution property of TMLE may be crucial!
- also, the ideas of TMLE, and targeting, extend in various important (technical) directions, where estimating equations estimators do not.