Introduction to semiparametric theory

Andrew Yiu

Department of Statistics, University of Oxford

StatML CDT course, Feb-Mar 2024

DEPARTMENT OF
STATISTICS

Recap: general strategy for causal inference

General strategy

1. Determine a causal quantity that will answer the scientific question of interest.
2. Check that the quantity is identifiable from the data you have and assumptions you are willing to make.
3. Choose a statistical estimator based on your data and assumption.

On Monday, you explored various methods for tackling the third objective, including:

- Augmented inverse probability weighting
- Double machine learning
- Targeted maximum likelihood estimation

These approaches are based on semiparametric theory.

Recap: general strategy for causal inference

General strategy

1. Determine a causal quantity that will answer the scientific question of interest.
2. Check that the quantity is identifiable from the data you have and assumptions you are willing to make.
3. Choose a statistical estimator based on your data and assumption.

On Monday, you explored various methods for tackling the third objective, including:

- Augmented inverse probability weighting
- Double machine learning
- Targeted maximum likelihood estimation

These approaches are based on semiparametric theory.

Objectives

Semiparametric theory has a long history and has garnered a reputation for being difficult to get to grips with at first.

Focus: intuition, motivation, big picture ideas, things that I wish were explained to me when I was a PhD student...☹

And please do ask questions!

What does "semiparametric" even mean?

We take a frequentist viewpoint, so a *"model"* is a collection of candidate distributions that could have generated the i.i.d. data $Z_1, \ldots, Z_n$.

Parametric inference

The model $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is indexed by a **finite-dimensional** parameter.
Example: ordinary linear regression with Gaussian errors.

What does "semiparametric" even mean?

We take a frequentist viewpoint, so a *"model"* is a collection of candidate distributions that could have generated the i.i.d. data $Z_1, \ldots, Z_n$.

Parametric inference

The model $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is indexed by a **finite-dimensional** parameter.
Example: ordinary linear regression with Gaussian errors.

Nonparametric inference

The model $\mathcal{P} = \{P_\eta : \eta \in \mathcal{H}\}$ and the target estimand are both **infinite-dimensional**.
Example: density estimation with smoothness assumptions.

What does "semiparametric" even mean?

We take a frequentist viewpoint, so a *"model"* is a collection of candidate distributions that could have generated the i.i.d. data $Z_1, \ldots, Z_n$.

Parametric inference

The model $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is indexed by a **finite-dimensional** parameter.
Example: ordinary linear regression with Gaussian errors.

Nonparametric inference

The model $\mathcal{P} = \{P_\eta : \eta \in \mathcal{H}\}$ and the target estimand are both **infinite-dimensional**.
Example: density estimation with smoothness assumptions.

Semiparametric inference

The model is **infinite-dimensional**, but the target estimand is **finite-dimensional**.

Example 1: partially linear regression

Partially linear regression

Consider

$$Y = A\theta + g(X) + U, \quad \mathbb{E}[U \mid A, X] = 0$$
$$A = m(X) + V, \quad \mathbb{E}[V \mid X] = 0,$$

where Y is the outcome, A is the treatment, and X consists of remaining covariates.

This is an example of a strict semiparametric model. We can partition the parameters into the finite-dimensional target θ and the infinite-dimensional nuisance parameters: $g(x)$, $m(x)$, the laws of U and V.

Example 1: partially linear regression

Partially linear regression

Consider

$$Y = A\theta + g(X) + U, \quad \mathbb{E}[U \mid A, X] = 0$$
$$A = m(X) + V, \quad \mathbb{E}[V \mid X] = 0,$$

where Y is the outcome, A is the treatment, and X consists of remaining covariates.

This is an example of a strict semiparametric model. We can partition the parameters into the finite-dimensional target θ and the infinite-dimensional nuisance parameters: $g(x)$, $m(x)$, the laws of U and V.

Example 2: Cox regression/proportional hazards

Cox regression

Consider $Z = (Y, \Delta, X)$, where $Y = \min(T, C)$, $\Delta = 1(T \leq C)$, T is the event time, C is the right-censoring time, X is a vector of covariates.

Denote the hazard function of T given X by

$$\lambda^{T|X}(t) = e^{\theta^{\mathrm{T}} X} \lambda(t),$$

where λ is the baseline hazard and θ is the target parameter.

Usually, the baseline hazard is completely unrestricted, so again we can partition into the parameters into a finite-dimensional target θ and an infinite-dimensional nuisance parameter λ.

Example 2: Cox regression/proportional hazards

Cox regression

Consider $Z = (Y, \Delta, X)$, where $Y = \min(T, C)$, $\Delta = 1(T \leq C)$, T is the event time, C is the right-censoring time, X is a vector of covariates.

Denote the hazard function of T given X by

$$\lambda^{T|X}(t) = e^{\theta^{\mathrm{T}} X} \lambda(t),$$

where λ is the baseline hazard and θ is the target parameter.

Usually, the baseline hazard is completely unrestricted, so again we can partition into the parameters into a finite-dimensional target θ and an infinite-dimensional nuisance parameter λ.

Example 3: integrated squared density

Integrated squared density

Suppose $Z \sim f$, where f is a Lebesgue density. The target parameter is

$$\psi(f) = \int f(z)^2 \, dz = \mathbb{E}_f[f(Z)].$$

This is a semiparametric problem in a more general sense. We have an infinite-dimensional parameter f, and the target is a one-dimensional functional, i.e. a mapping $\psi : \mathcal{P} \to \mathbb{R}$ from the model to the real line.

The previous strict semiparametric problems are a special case of this: $\mathcal{P} = \{P_{\theta,\eta} : \underbrace{\theta \in \Theta}_{\text{target}}, \underbrace{\eta \in \mathcal{H}}_{\text{nuisance}}\}$ and $\psi(P_{\theta,\eta}) = \theta$.

Example 3: integrated squared density

Integrated squared density

Suppose $Z \sim f$, where f is a Lebesgue density. The target parameter is

$$\psi(f) = \int f(z)^2 \, dz = \mathbb{E}_f[f(Z)].$$

This is a semiparametric problem in a more general sense. We have an infinite-dimensional parameter f, and the target is a one-dimensional functional, i.e. a mapping $\psi : \mathcal{P} \to \mathbb{R}$ from the model to the real line.

The previous strict semiparametric problems are a special case of this: $\mathcal{P} = \{P_{\theta,\eta} : \underbrace{\theta \in \Theta}_{\text{target}}, \underbrace{\eta \in \mathcal{H}}_{\text{nuisance}}\}$ and $\psi(P_{\theta,\eta}) = \theta$.

Example 4: potential outcome mean

Potential outcome mean

Suppose the data takes the form $Z = (X, A, Y)$, where X is a vector of covariates, A is a binary treatment indicator, and Y is the outcome variable of interest.

The target estimand is the potential outcome mean $\psi(P) = \mathbb{E}_P[Y(1)]$, which is identified by

$$\psi(P) = \mathbb{E}_P[\mathbb{E}_P(Y \mid A = 1, X)] \qquad \text{(g-formula)}$$

under the assumptions:

- (conditional exchangeability) $Y(1) \perp\!\!\!\perp A \mid X$.
- (positivity) $0 < \pi(X)$ with P-probability 1, where $\pi(x) = P(A = 1 \mid X = x)$ is called the **propensity score**.

Example 4: potential outcome mean

Potential outcome mean

Suppose the data takes the form $Z = (X, A, Y)$, where X is a vector of covariates, A is a binary treatment indicator, and Y is the outcome variable of interest.

The target estimand is the potential outcome mean $\psi(P) = \mathbb{E}_P[Y(1)]$, which is identified by

$$\psi(P) = \mathbb{E}_P[\mathbb{E}_P(Y \mid A = 1, X)] \qquad \text{(g-formula)}$$

under the assumptions:

- (conditional exchangeability) $Y(1) \perp\!\!\!\perp A \mid X$.
- (positivity) $0 < \pi(X)$ with P-probability 1, where $\pi(x) = P(A = 1 \mid X = x)$ is called the **propensity score**.

Example 4: potential outcome mean

Potential outcome mean

Suppose the data takes the form $Z = (X, A, Y)$, where X is a vector of covariates, A is a binary treatment indicator, and Y is the outcome variable of interest.

The target estimand is the potential outcome mean $\psi(P) = \mathbb{E}_P[Y(1)]$, which is identified by

$$\psi(P) = \mathbb{E}_P[\mathbb{E}_P(Y \mid A = 1, X)] \qquad \text{(g-formula)}$$

under the assumptions:

- (conditional exchangeability) $Y(1) \perp\!\!\!\perp A \mid X$.
- (positivity) $0 < \pi(X)$ with P-probability 1, where $\pi(x) = P(A = 1 \mid X = x)$ is called the **propensity score**.

Example 4: potential outcome mean

Potential outcome mean

Suppose the data takes the form $Z = (X, A, Y)$, where X is a vector of covariates, A is a binary treatment indicator, and Y is the outcome variable of interest.

The target estimand is the potential outcome mean $\psi(P) = \mathbb{E}_P[Y(1)]$, which is identified by

$$\psi(P) = \mathbb{E}_P[\mathbb{E}_P(Y \mid A = 1, X)] \qquad \text{(g-formula)}$$

under the assumptions:

- (conditional exchangeability) $Y(1) \perp\!\!\!\perp A \mid X$.
- (positivity) $0 < \pi(X)$ with P-probability 1, where $\pi(x) = P(A = 1 \mid X = x)$ is called the **propensity score**.

What's so special?

You might be thinking: *since we (usually) have to estimate the infinite-dimensional parameters anyways, what makes this different to nonparametric inference?*

What's so special?

You might be thinking: *since we (usually) have to estimate the infinite-dimensional parameters anyways, what makes this different to nonparametric inference?*

Parametric inference

Often straightforward to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator (e.g., MLE, parametric Bayes).

What's so special?

You might be thinking: *since we (usually) have to estimate the infinite-dimensional parameters anyways, what makes this different to nonparametric inference?*

Parametric inference

Often straightforward to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator (e.g., MLE, parametric Bayes).

Nonparametric inference

Aside from very special cases (like estimating a CDF), the rate of convergence is always slower than $\sqrt{n}$, and there is very limited asymptotic distribution theory.

What's so special?

You might be thinking: *since we (usually) have to estimate the infinite-dimensional parameters anyways, what makes this different to nonparametric inference?*

Parametric inference

Often straightforward to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator (e.g., MLE, parametric Bayes).

Nonparametric inference

Aside from very special cases (like estimating a CDF), the rate of convergence is always slower than $\sqrt{n}$, and there is very limited asymptotic distribution theory.

Semiparametric inference

Similar asymptotic theory to the parametric case if the functional is "differentiable".

Bridging statistics and machine learning

The main reason semiparametric theory currently gets so much attention in causal inference: use machine learning to estimate nuisance parameters and still obtain valid statistical guarantees for the target estimand!

But...

- Unlike the parametric setting, we can't expect to automatically get good inference for all estimands in one swoop. We need to carefully tailor our estimation towards the target.
- Naïve use of machine learning can be disastrous
  - Unclear whether the bootstrap is valid (or whether we have asymptotic normality at all).
  - We might not even have a $\sqrt{n}$ convergence rate.

Bridging statistics and machine learning

The main reason semiparametric theory currently gets so much attention in causal inference: use machine learning to estimate nuisance parameters and still obtain valid statistical guarantees for the target estimand!

But...

- Unlike the parametric setting, we can't expect to automatically get good inference for all estimands in one swoop. We need to carefully tailor our estimation towards the target.
- Naïve use of machine learning can be disastrous
  - Unclear whether the bootstrap is valid (or whether we have asymptotic normality at all).
  - We might not even have a $\sqrt{n}$ convergence rate.

Bridging statistics and machine learning

The main reason semiparametric theory currently gets so much attention in causal inference: use machine learning to estimate nuisance parameters and still obtain valid statistical guarantees for the target estimand!

But...

- Unlike the parametric setting, we can't expect to automatically get good inference for all estimands in one swoop. We need to carefully tailor our estimation towards the target.
- Naïve use of machine learning can be disastrous
  - Unclear whether the bootstrap is valid (or whether we have asymptotic normality at all).
  - We might not even have a $\sqrt{n}$ convergence rate.

Illustration

Average treatment effect

Recall: under conditional exchangeability, the **average treatment effect** is identified as

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

To estimate the right-hand side, it is seemingly natural to posit estimators for:

- The marginal distribution of the covariates $dP(x)$: a simple and convenient choice is the *empirical distribution* $\mathbb{P}_n^X = n^{-1} \sum_{i=1}^n \delta_{X_i}$.
- The "outcome regression function" $\mathbb{E}_P(Y \mid A = a, X = x)$: we will estimate this using *random forests*.

Then we simply "plug-in" these estimators into the identification formula.

Illustration

Average treatment effect

Recall: under conditional exchangeability, the **average treatment effect** is identified as

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

To estimate the right-hand side, it is seemingly natural to posit estimators for:

- The marginal distribution of the covariates $dP(x)$: a simple and convenient choice is the *empirical distribution* $\mathbb{P}_n^X = n^{-1} \sum_{i=1}^n \delta_{X_i}$.
- The "outcome regression function" $\mathbb{E}_P(Y \mid A = a, X = x)$: we will estimate this using *random forests*.

Then we simply "plug-in" these estimators into the identification formula.

Illustration

Average treatment effect

Recall: under conditional exchangeability, the **average treatment effect** is identified as

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

To estimate the right-hand side, it is seemingly natural to posit estimators for:

- The marginal distribution of the covariates $dP(x)$: a simple and convenient choice is the *empirical distribution* $\mathbb{P}_n^X = n^{-1} \sum_{i=1}^n \delta_{X_i}$.
- The "outcome regression function" $\mathbb{E}_P(Y \mid A = a, X = x)$: we will estimate this using *random forests*.

Then we simply "plug-in" these estimators into the identification formula.

Illustration

Average treatment effect

Recall: under conditional exchangeability, the **average treatment effect** is identified as

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

To estimate the right-hand side, it is seemingly natural to posit estimators for:

- The marginal distribution of the covariates $dP(x)$: a simple and convenient choice is the *empirical distribution* $\mathbb{P}_n^X = n^{-1} \sum_{i=1}^n \delta_{X_i}$.
- The "outcome regression function" $\mathbb{E}_P(Y \mid A = a, X = x)$: we will estimate this using *random forests*.

Then we simply "plug-in" these estimators into the identification formula.

Simulation

We simulate some data using the `causl` R package:

$$\begin{aligned} Z &\sim \text{Exponential}(2) \\ A \mid X = x &\sim \text{Bernoulli}(\text{logit}(x)) \\ Y(a) &\sim \mathcal{N}(0.5a, \sigma^2). \end{aligned}$$

So the true ATE = 0.5. We simulate samples of size $n = 500$ across 5000 independent Monte Carlo trials.

For each trial, we compute the plug-in estimator

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\mathbb{E}}(Y \mid A = 1, X_i) - \hat{\mathbb{E}}(Y \mid A = 0, X_i)\},$$

where $\hat{\mathbb{E}}$ is a random forests estimator implemented with the `randomForest` R package.

Simulation

We simulate some data using the `causl` R package:

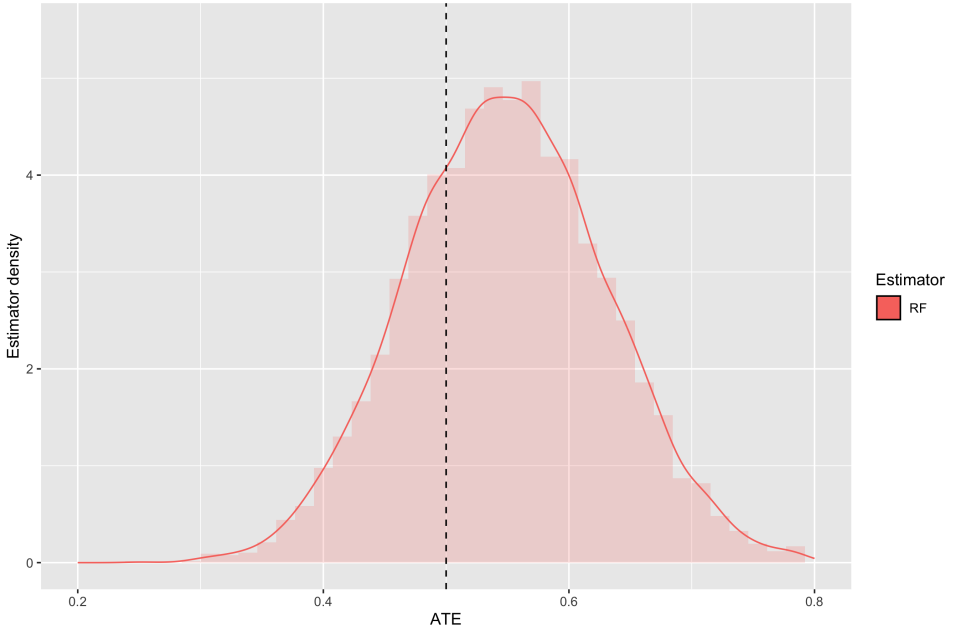
$$\begin{aligned} Z &\sim \text{Exponential}(2) \\ A \mid X = x &\sim \text{Bernoulli}(\text{logit}(x)) \\ Y(a) &\sim \mathcal{N}(0.5a, \sigma^2). \end{aligned}$$

So the true ATE = 0.5. We simulate samples of size $n = 500$ across 5000 independent Monte Carlo trials.

For each trial, we compute the plug-in estimator

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\mathbb{E}}(Y \mid A = 1, X_i) - \hat{\mathbb{E}}(Y \mid A = 0, X_i)\},$$

where $\hat{\mathbb{E}}$ is a random forests estimator implemented with the `randomForest` R package.


Estimator density
4
2
0
0.2
0.4
0.6
0.8
ATE
Estimator
RF

What went wrong?

Random forests is designed to estimate the whole regression surface, and it is optimized for prediction. To perform well at these objectives, it introduces bias (or "regularizes").

This bias bleeds into the estimation of the ATE when we use our naïve plug-in estimator. This is a general phenomenon for nonparametric statistics.
"A good bias-variance trade-off for the whole infinite-dimensional parameter doesn't necessarily translate into a good trade-off for the low-dimensional target estimand."

Semiparametric theory to the rescue! We will use the specific structure of the estimand to remove bias and enable rigorous statistical guarantees (e.g. coverage of confidence intervals, Type I error control etc.)

What went wrong?

Random forests is designed to estimate the whole regression surface, and it is optimized for prediction. To perform well at these objectives, it introduces bias (or "regularizes").

This bias bleeds into the estimation of the ATE when we use our naïve plug-in estimator. This is a general phenomenon for nonparametric statistics.
"A good bias-variance trade-off for the whole infinite-dimensional parameter doesn't necessarily translate into a good trade-off for the low-dimensional target estimand."

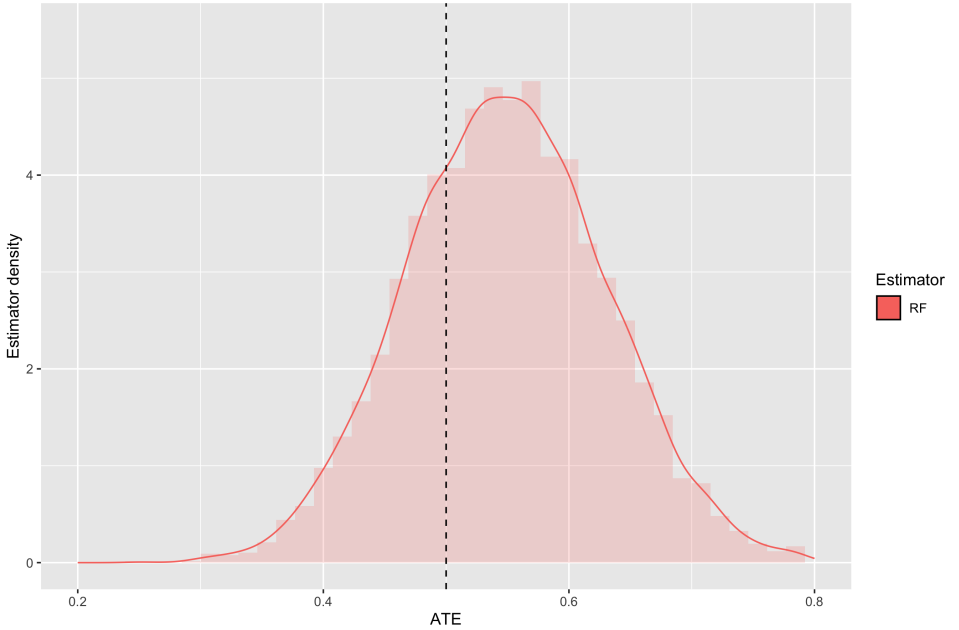
Semiparametric theory to the rescue! We will use the specific structure of the estimand to remove bias and enable rigorous statistical guarantees (e.g. coverage of confidence intervals, Type I error control etc.)

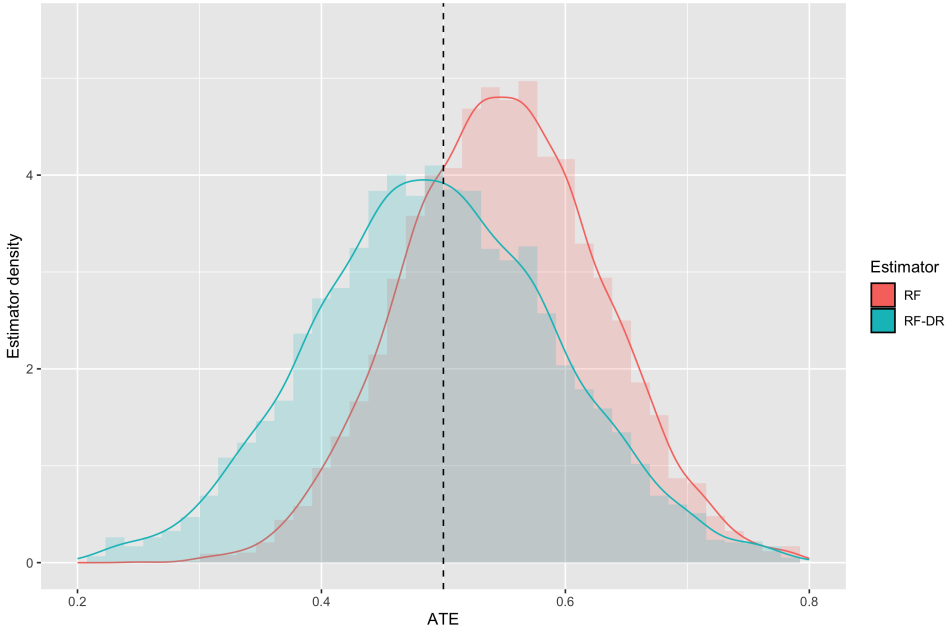
What went wrong?

Random forests is designed to estimate the whole regression surface, and it is optimized for prediction. To perform well at these objectives, it introduces bias (or "regularizes").

This bias bleeds into the estimation of the ATE when we use our naïve plug-in estimator. This is a general phenomenon for nonparametric statistics.
"A good bias-variance trade-off for the whole infinite-dimensional parameter doesn't necessarily translate into a good trade-off for the low-dimensional target estimand."

Semiparametric theory to the rescue! We will use the specific structure of the estimand to remove bias and enable rigorous statistical guarantees (e.g. coverage of confidence intervals, Type I error control etc.)


Estimator density
4
2
0
0.2
0.4
0.6
0.8
ATE
Estimator
RF


Estimator density
4
2
0
0.2
0.4
0.6
0.8
ATE
Estimator
RF
RF-DR

Structure

1. Semiparametric efficiency theory (Today)

- Parametric submodels
- Tangent spaces
- Pathwise differentiability
- The efficient influence function

2. Modern applications and bias correction (Monday)

- The von Mises expansion
- One-step estimation
- Estimating equations, Neyman orthogonality and double ML.

Some historical context

Semiparametric efficiency theory was first conceived by Charles Stein in 1956.

EFFICIENT NONPARAMETRIC TESTING
AND ESTIMATION

CHARLES STEIN
STANFORD UNIVERSITY

The idea was to take existing efficiency theory for parametric models and generalize to semiparametric problems (though the word *"semiparametric"* wasn't coined until later).

Stein's ideas were formalized and extended in the 70's-90's.[1] The application of semiparametric theory to causal inference was pioneered by Robins and Rotnitzky.

[1] Including work by Levit, Koshevnik, Pfanzagl, Bickel, Begun, Klaassen, Wellner, Ritov, Tsiatis, van der Vaart, Murphy, et al.

Some historical context

Semiparametric efficiency theory was first conceived by Charles Stein in 1956.

EFFICIENT NONPARAMETRIC TESTING
AND ESTIMATION

CHARLES STEIN
STANFORD UNIVERSITY

The idea was to take existing efficiency theory for parametric models and generalize to semiparametric problems (though the word *"semiparametric"* wasn't coined until later).

Stein's ideas were formalized and extended in the 70's-90's.[1] The application of semiparametric theory to causal inference was pioneered by Robins and Rotnitzky.

[1] Including work by Levit, Koshevnik, Pfanzagl, Bickel, Begun, Klaassen, Wellner, Ritov, Tsiatis, van der Vaart, Murphy, et al.

Parametric theory

Suppose we have a parametric model $\{P_\theta : \theta \in \Theta\}$ and i.i.d. data $Z_1, \ldots, Z_n \sim P_{\theta_0}$. We are interested in estimating $\psi(P_{\theta_0})$ for some mapping $\psi : \mathcal{P} \to \mathbb{R}$.

Recall:

- Score function $s_\theta(z) = \frac{\partial}{\partial \theta} \log p_\theta(z)$, which has mean zero: $\mathbb{E}_\theta[s_\theta(Z)] = 0$.
- Fisher information $I_\theta = \mathbb{E}_\theta[s_\theta(Z) s_\theta(Z)^{\mathrm{T}}]$.

Cramér-Rao lower bound

Let $\psi(P_\theta) : \Theta \to \mathbb{R}$ be differentiable with derivative ψ'_θ, and let $\hat{\psi} = \hat{\psi}(Z_{1:n})$ be any unbiased estimator of $\psi(P_\theta)$. Then

$$n \operatorname{var}_{\theta_0}(\hat{\psi}) \geq \psi'_{\theta_0} I_{\theta_0}^{-1} \psi'^{\mathrm{T}}_{\theta_0} \quad \text{for all } \theta_0 \in \operatorname{int}(\Theta).$$

Parametric theory

Suppose we have a parametric model $\{P_\theta : \theta \in \Theta\}$ and i.i.d. data $Z_1, \ldots, Z_n \sim P_{\theta_0}$. We are interested in estimating $\psi(P_{\theta_0})$ for some mapping $\psi : \mathcal{P} \to \mathbb{R}$.

Recall:

- Score function $s_\theta(z) = \frac{\partial}{\partial \theta} \log p_\theta(z)$, which has mean zero: $\mathbb{E}_\theta[s_\theta(Z)] = 0$.
- Fisher information $I_\theta = \mathbb{E}_\theta[s_\theta(Z) s_\theta(Z)^{\mathrm{T}}]$.

Cramér-Rao lower bound

Let $\psi(P_\theta) : \Theta \to \mathbb{R}$ be differentiable with derivative ψ'_θ, and let $\hat{\psi} = \hat{\psi}(Z_{1:n})$ be any unbiased estimator of $\psi(P_\theta)$. Then

$$n \operatorname{var}_{\theta_0}(\hat{\psi}) \geq \psi'_{\theta_0} I_{\theta_0}^{-1} \psi'^{\mathrm{T}}_{\theta_0} \quad \text{for all } \theta_0 \in \operatorname{int}(\Theta).$$

Asymptotic efficiency

The Cramér-Rao lower bound informally motivates an asymptotic notion of efficiency.

Asymptotic efficiency

An estimator sequence $\hat{\psi}$ is said to be **asymptotically efficient** for estimating $\psi(P_{\theta_0})$ if

$$\sqrt{n}(\hat{\psi} - \psi(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(Z_i) + o_{P_{\theta_0}}(1).$$

So the best limiting distribution is $\mathcal{N}(0, \psi'_{\theta_0} I_{\theta_0}^{-1} \psi'^{\mathrm{T}}_{\theta_0})$ by the CLT.

We say that $\psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(z)$ is the influence function of the estimator $\hat{\psi}$, *"the error $\hat{\psi} - \psi(P_{\theta_0})$ behaves like the sample average of the influence function"*.

Under regularity conditions, the "plug-in" $\hat{\psi} = \psi(P_{\hat{\theta}_{mle}})$ attains asymptotic efficiency.

Asymptotic efficiency

The Cramér-Rao lower bound informally motivates an asymptotic notion of efficiency.

Asymptotic efficiency

An estimator sequence $\hat{\psi}$ is said to be **asymptotically efficient** for estimating $\psi(P_{\theta_0})$ if

$$\sqrt{n}(\hat{\psi} - \psi(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(Z_i) + o_{P_{\theta_0}}(1).$$

So the best limiting distribution is $\mathcal{N}(0, \psi'_{\theta_0} I_{\theta_0}^{-1} \psi'^{\mathrm{T}}_{\theta_0})$ by the CLT.

We say that $\psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(z)$ is the influence function of the estimator $\hat{\psi}$, *"the error $\hat{\psi} - \psi(P_{\theta_0})$ behaves like the sample average of the influence function"*.

Under regularity conditions, the "plug-in" $\hat{\psi} = \psi(P_{\hat{\theta}_{mle}})$ attains asymptotic efficiency.

Asymptotic efficiency

The Cramér-Rao lower bound informally motivates an asymptotic notion of efficiency.

Asymptotic efficiency

An estimator sequence $\hat{\psi}$ is said to be **asymptotically efficient** for estimating $\psi(P_{\theta_0})$ if

$$\sqrt{n}(\hat{\psi} - \psi(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(Z_i) + o_{P_{\theta_0}}(1).$$

So the best limiting distribution is $\mathcal{N}(0, \psi'_{\theta_0} I_{\theta_0}^{-1} \psi'^{\mathrm{T}}_{\theta_0})$ by the CLT.

We say that $\psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(z)$ is the influence function of the estimator $\hat{\psi}$, *"the error $\hat{\psi} - \psi(P_{\theta_0})$ behaves like the sample average of the influence function"*.

Under regularity conditions, the "plug-in" $\hat{\psi} = \psi(P_{\hat{\theta}_{mle}})$ attains asymptotic efficiency.

Asymptotic efficiency

The Cramér-Rao lower bound informally motivates an asymptotic notion of efficiency.

Asymptotic efficiency

An estimator sequence $\hat{\psi}$ is said to be **asymptotically efficient** for estimating $\psi(P_{\theta_0})$ if

$$\sqrt{n}(\hat{\psi} - \psi(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(Z_i) + o_{P_{\theta_0}}(1).$$

So the best limiting distribution is $\mathcal{N}(0, \psi'_{\theta_0} I_{\theta_0}^{-1} \psi'^{\mathrm{T}}_{\theta_0})$ by the CLT.

We say that $\psi'_{\theta_0} I_{\theta_0}^{-1} s_{\theta_0}(z)$ is the influence function of the estimator $\hat{\psi}$, *"the error* $\hat{\psi} - \psi(P_{\theta_0})$ *behaves like the sample average of the influence function"*.

Under regularity conditions, the "plug-in" $\hat{\psi} = \psi(P_{\hat{\theta}_{mle}})$ attains asymptotic efficiency.

Semiparametric efficiency

Let's return to the more general setting where $\mathcal{P}$ is possibly infinite-dimensional. We want to estimate a functional $\psi : \mathcal{P} \to \mathbb{R}$. How do we define the lower bound at a distribution $P \in \mathcal{P}$?

Suppose $\{P_t\} \subset \mathcal{P}$ is a parametric model that is contained in our model and passes through P. Estimating $\psi(P)$ is at least as hard in $\mathcal{P}$ as it is in $\{P_t\}$.

Idea

Define the semiparametric lower bound to be the greatest Cramér-Rao lower bound across all parametric models that are contained in $\mathcal{P}$ and pass through P.

Semiparametric efficiency

Let's return to the more general setting where $\mathcal{P}$ is possibly infinite-dimensional. We want to estimate a functional $\psi : \mathcal{P} \to \mathbb{R}$. How do we define the lower bound at a distribution $P \in \mathcal{P}$?

Suppose $\{P_t\} \subset \mathcal{P}$ is a parametric model that is contained in our model and passes through P. Estimating $\psi(P)$ is at least as hard in $\mathcal{P}$ as it is in $\{P_t\}$.

Idea

Define the semiparametric lower bound to be the greatest Cramér-Rao lower bound across all parametric models that are contained in $\mathcal{P}$ and pass through P.

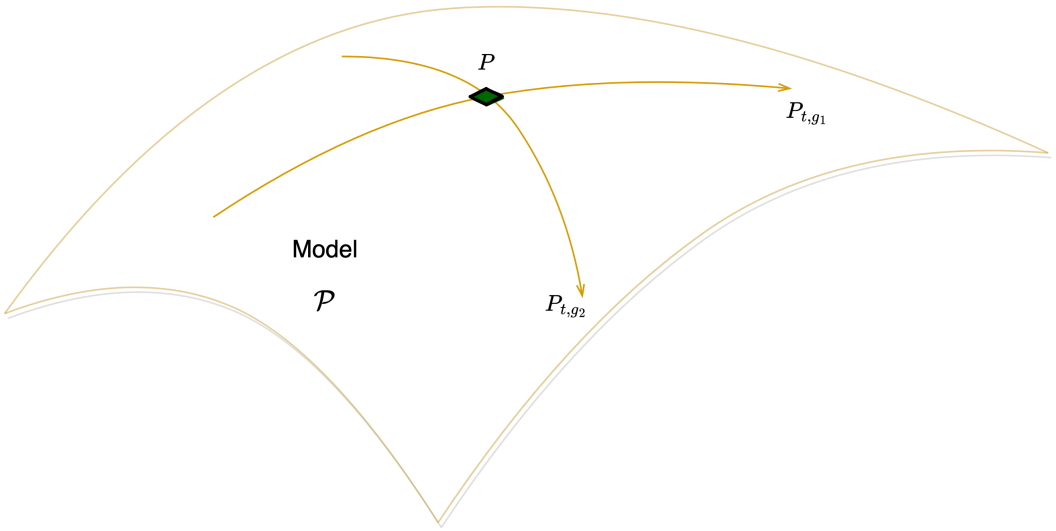
Semiparametric efficiency

Let's return to the more general setting where $\mathcal{P}$ is possibly infinite-dimensional. We want to estimate a functional $\psi : \mathcal{P} \to \mathbb{R}$. How do we define the lower bound at a distribution $P \in \mathcal{P}$?

Suppose $\{P_t\} \subset \mathcal{P}$ is a parametric model that is contained in our model and passes through P. Estimating $\psi(P)$ is at least as hard in $\mathcal{P}$ as it is in $\{P_t\}$.

Idea

Define the semiparametric lower bound to be the greatest Cramér-Rao lower bound across all parametric models that are contained in $\mathcal{P}$ and pass through P.


P
P_{t,g_1}
Model
$\mathcal{P}$
P_{t,g_2}

Parametric submodels

A **parametric submodel** $\{P_{t,g} : t \in (-\varepsilon, \varepsilon)\}$ is a smooth parametric model that passes through P at $t = 0$ with

$$g(z) = \frac{\partial}{\partial t} \log p_{t,g}(z)|_{t=0},$$

i.e. the score function at P is equal to g.

We only care about the value of the score function at $t = 0$ because this is sufficient to compute the C-R lower bound

$$\frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]}.$$

Important: a parametric submodel is not substantively meaningful; it's just a technical device that sets up the theoretical framework.

Parametric submodels

A **parametric submodel** $\{P_{t,g} : t \in (-\varepsilon, \varepsilon)\}$ is a smooth parametric model that passes through P at $t = 0$ with

$$g(z) = \frac{\partial}{\partial t} \log p_{t,g}(z)|_{t=0},$$

i.e. the score function at P is equal to g.

We only care about the value of the score function at $t = 0$ because this is sufficient to compute the C-R lower bound

$$\frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]}.$$

Important: a parametric submodel is not substantively meaningful; it's just a technical device that sets up the theoretical framework.

Parametric submodels

Parametric submodels

A **parametric submodel** $\{P_{t,g} : t \in (-\varepsilon, \varepsilon)\}$ is a smooth parametric model that passes through P at $t = 0$ with

$$g(z) = \frac{\partial}{\partial t} \log p_{t,g}(z)|_{t=0},$$

i.e. the score function at P is equal to g.

We only care about the value of the score function at $t = 0$ because this is sufficient to compute the C-R lower bound

$$\frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]}.$$

Important: a parametric submodel is not substantively meaningful; it's just a technical device that sets up the theoretical framework.

Tangent spaces

Tangent space

The **tangent space** $\dot{\mathcal{P}}_P$ is the collection of all score functions g across the parametric submodels contained in $\mathcal{P}$ that pass through P.

We can interpret a score function as the "direction" in which the submodel passes through P. Then the tangent space is the set of directions in which we can move an infinitesimal distance away from P and still remain in the model.

Thus, the larger the model, the larger the tangent space.

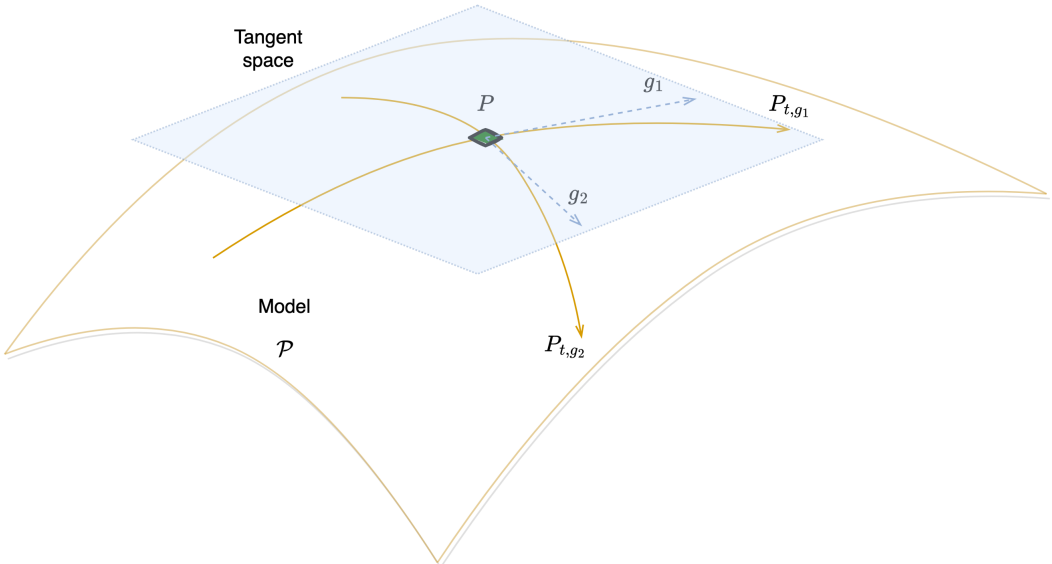
Tangent spaces

Tangent space

The **tangent space** $\dot{\mathcal{P}}_P$ is the collection of all score functions g across the parametric submodels contained in $\mathcal{P}$ that pass through P.

We can interpret a score function as the "direction" in which the submodel passes through P. Then the tangent space is the set of directions in which we can move an infinitesimal distance away from P and still remain in the model.

Thus, the larger the model, the larger the tangent space.


Tangent
space
P
g_1
P_{t,g_1}
g_2
Model
$\mathcal{P}$
P_{t,g_2}

Greatest lower bound

We can now write down our semiparametric lower bound at P:

$$\sup_{g \in \dot{\mathcal{P}}_P} \frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]}.$$

Recall that this is the greatest Cramér-Rao lower bound across all parametric submodels contained in $\mathcal{P}$ that pass through P.

Clearly, we need the mapping $t \mapsto \psi(P_{t,g})$ to be differentiable at $t = 0$ for any parametric submodel. But this is not sufficient; we need a stronger form of differentiability...

Greatest lower bound

We can now write down our semiparametric lower bound at P:

$$\sup_{g \in \dot{\mathcal{P}}_P} \frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]}.$$

Recall that this is the greatest Cramér-Rao lower bound across all parametric submodels contained in $\mathcal{P}$ that pass through P.

Clearly, we need the mapping $t \mapsto \psi(P_{t,g})$ to be differentiable at $t = 0$ for any parametric submodel. But this is not sufficient; we need a stronger form of differentiability...

Pathwise differentiability

Pathwise differentiability

A functional ψ is **(pathwise) differentiable** at P with respect to $\dot{\mathcal{P}}_P$ if:

(a) the mapping $t \mapsto \psi(P_{t,g})$ is differentiable at $t = 0$, and

(b) there exists a fixed function $\phi_P : \mathcal{Z} \to \mathbb{R}$ such that

$$\frac{\partial \psi(P_{t,g})}{\partial t}|_{t=0} = \mathbb{E}_P[\phi_P g]$$

for every $g \in \dot{\mathcal{P}}_P$ and any parametric submodel $\{P_{t,g}\}$ with score function g. We call ϕ_P a **gradient** of ψ at P.

So not only do we need differentiability of $t \mapsto \psi(P_{t,g})$ in the ordinary sense, but we also require a special representation of the derivative.

The canonical gradient

So we need

$$\frac{\partial \psi(P_{t,g})}{\partial t}|_{t=0} = \mathbb{E}_P[\phi_P g].$$

The gradient ϕ_P is not unique; we can always replace ϕ_P by $\phi_P + h$, where h satisfies $\mathbb{E}_P[hg] = 0$ for all $g \in \dot{\mathcal{P}}_P$ (i.e. h is "orthogonal" to the tangent space).

Canonical gradient/Efficient influence function

There is a unique gradient that has minimum variance amongst all mean-zero gradients. We call it the **canonical gradient** or **efficient influence function**, denoted by $\dot{\psi}_P$. It is the *"$L_2(P)$-projection"* of any gradient onto the tangent space, i.e.

$$\dot{\psi}_P = \underset{g \in \dot{\mathcal{P}}_P}{\operatorname{argmin}} \mathbb{E}_P[(\phi_P - g)^2].$$

The canonical gradient

So we need

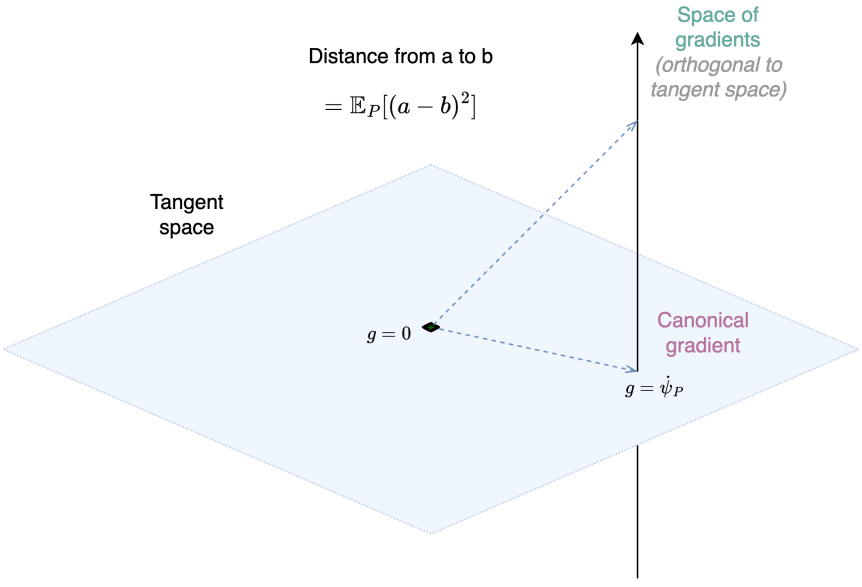
$$\frac{\partial \psi(P_{t,g})}{\partial t}|_{t=0} = \mathbb{E}_P[\phi_P g].$$

The gradient ϕ_P is not unique; we can always replace ϕ_P by $\phi_P + h$, where h satisfies $\mathbb{E}_P[hg] = 0$ for all $g \in \dot{\mathcal{P}}_P$ (i.e. h is "orthogonal" to the tangent space).

Canonical gradient/Efficient influence function

There is a unique gradient that has minimum variance amongst all mean-zero gradients. We call it the **canonical gradient** or **efficient influence function**, denoted by $\dot{\psi}_P$. It is the *"$L_2(P)$-projection"* of any gradient onto the tangent space, i.e.

$$\dot{\psi}_P = \underset{g \in \dot{\mathcal{P}}_P}{\operatorname{argmin}} \mathbb{E}_P[(\phi_P - g)^2].$$


Space of gradients
(orthogonal to tangent space)
Distance from a to b
$= \mathbb{E}_P[(a-b)^2]$
Tangent space
$g = 0$
Canonical gradient
$g = \dot{\psi}_P$

Variance of the canonical gradient

Returning to the lower bound:

$$\sup_{g \in \dot{\mathcal{P}}_P} \frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]} = \sup_{g \in \dot{\mathcal{P}}_P} \frac{\mathbb{E}_P[\dot{\psi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]}$$
$$\leq \mathbb{E}_P[\dot{\psi}_P(Z)^2]$$

by the Cauchy-Schwarz inequality.

But by definition, $\dot{\psi}_P$ lies within the tangent space $\dot{\mathcal{P}}_P$, so we also have the reverse inequality by taking $g = \dot{\psi}_P$, i.e.

$$\mathbb{E}_P[\dot{\psi}_P(Z)^2] \leq \sup_{g \in \dot{\mathcal{P}}_P} \frac{\mathbb{E}_P[\dot{\psi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]}.$$

Thus, the variance lower bound is the variance of the canonical gradient.

Variance of the canonical gradient

Returning to the lower bound:

$$\sup_{g \in \dot{\mathcal{P}}_P} \frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]} = \sup_{g \in \dot{\mathcal{P}}_P} \frac{\mathbb{E}_P[\dot{\psi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]} \leq \mathbb{E}_P[\dot{\psi}_P(Z)^2]$$

by the Cauchy-Schwarz inequality.

But by definition, $\dot{\psi}_P$ lies within the tangent space $\dot{\mathcal{P}}_P$, so we also have the reverse inequality by taking $g = \dot{\psi}_P$, i.e.

$$\mathbb{E}_P[\dot{\psi}_P(Z)^2] \leq \sup_{g \in \dot{\mathcal{P}}_P} \frac{\mathbb{E}_P[\dot{\psi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]}.$$

Thus, the variance lower bound is the variance of the canonical gradient.

Variance of the canonical gradient

Returning to the lower bound:

$$\sup_{g \in \dot{\mathcal{P}}_P} \frac{(\frac{\partial}{\partial t}\psi(P_{t,g})|_{t=0})^2}{\mathbb{E}_P[g(Z)^2]} = \sup_{g \in \dot{\mathcal{P}}_P} \frac{\mathbb{E}_P[\dot{\psi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]} \leq \mathbb{E}_P[\dot{\psi}_P(Z)^2]$$

by the Cauchy-Schwarz inequality.

But by definition, $\dot{\psi}_P$ lies within the tangent space $\dot{\mathcal{P}}_P$, so we also have the reverse inequality by taking $g = \dot{\psi}_P$, i.e.

$$\mathbb{E}_P[\dot{\psi}_P(Z)^2] \leq \sup_{g \in \dot{\mathcal{P}}_P} \frac{\mathbb{E}_P[\dot{\psi}_P(Z)g(Z)]^2}{\mathbb{E}_P[g(Z)^2]}.$$

Thus, the variance lower bound is the variance of the canonical gradient.

Asymptotic efficiency

Semiparametric efficiency

An estimator sequence $\hat{\psi}$ is said to be **asymptotically efficient** for estimating $\psi(P)$ if

$$\sqrt{n}(\hat{\psi} - \psi(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\psi}_P(Z_i) + o_P(1).$$

So the best limiting distribution is $\mathcal{N}(0, \mathbb{E}_P[\dot{\psi}_P^2])$ by the CLT.

This is why $\dot{\psi}_P$ is called the efficient influence function.

Example 4: potential outcome mean

Potential outcome mean

Suppose the data takes the form $Z = (X, A, Y)$, where X is a vector of covariates, A is a binary treatment indicator, and Y is the outcome variable of interest.

The target estimand is $\psi(P) = \mathbb{E}_P[Y(1)] = \mathbb{E}_P[\mathbb{E}_P(Y \mid A = 1, X)]$.

Denote:

- $\pi(x) = P(A = 1 \mid X = x)$ (propensity score)
- $m(x) = \mathbb{E}[Y \mid A = 1, X = x]$.

The efficient influence function is

$$\dot{\psi}_P(Z) = \frac{A(Y - m(X))}{\pi(X)} + m(X) - \psi(P).$$

Summary of semiparametric efficiency theory

- In parametric theory, we quantify the hardness of estimating a parameter at a point using the *Cramér-Rao lower bound*.
- For the infinite-dimensional case, we can generalize this by taking the supremum of the lower bounds across all *parametric submodels* that pass through the distribution.
- We need the functional to be *"pathwise differentiable"*, which means that there exists a pathwise derivative (a *"gradient"*) that takes the score function as input and then outputs the corresponding slope in ψ.
- The unique gradient that lies in the tangent space is called the *"canonical gradient"* or *"efficient influence function"*. Its variance is equal to the semiparametric Cramér-Rao lower bound.

Summary of semiparametric efficiency theory

- In parametric theory, we quantify the hardness of estimating a parameter at a point using the *Cramér-Rao lower bound*.
- For the infinite-dimensional case, we can generalize this by taking the supremum of the lower bounds across all *parametric submodels* that pass through the distribution.
- We need the functional to be *"pathwise differentiable"*, which means that there exists a pathwise derivative (a *"gradient"*) that takes the score function as input and then outputs the corresponding slope in ψ.
- The unique gradient that lies in the tangent space is called the *"canonical gradient"* or *"efficient influence function"*. Its variance is equal to the semiparametric Cramér-Rao lower bound.

Summary of semiparametric efficiency theory

- In parametric theory, we quantify the hardness of estimating a parameter at a point using the *Cramér-Rao lower bound*.
- For the infinite-dimensional case, we can generalize this by taking the supremum of the lower bounds across all *parametric submodels* that pass through the distribution.
- We need the functional to be *"pathwise differentiable"*, which means that there exists a pathwise derivative (a *"gradient"*) that takes the score function as input and then outputs the corresponding slope in ψ.
- The unique gradient that lies in the tangent space is called the *"canonical gradient"* or *"efficient influence function"*. Its variance is equal to the semiparametric Cramér-Rao lower bound.

Summary of semiparametric efficiency theory

- In parametric theory, we quantify the hardness of estimating a parameter at a point using the *Cramér-Rao lower bound*.
- For the infinite-dimensional case, we can generalize this by taking the supremum of the lower bounds across all *parametric submodels* that pass through the distribution.
- We need the functional to be *"pathwise differentiable"*, which means that there exists a pathwise derivative (a *"gradient"*) that takes the score function as input and then outputs the corresponding slope in ψ.
- The unique gradient that lies in the tangent space is called the *"canonical gradient"* or *"efficient influence function"*. Its variance is equal to the semiparametric Cramér-Rao lower bound.

Accessible introductions

Kennedy (2022) "Semiparametric doubly robust..."

Nice general introduction on functional estimation, leaning towards causal problems.

Hines et al. (2022) "Demystifying statistical learning..."

Lots of examples on deriving efficient influence functions.

Newey (1990) "Semiparametric efficiency bounds"

Readable introduction to semiparametric efficiency theory.