Day 2: Introduction to semiparametric theory

Chris Holmes and Andrew Yiu

Department of Statistics, University of Oxford

Leuven, September 2024

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. Perform **statistical inference** for the identified quantity based on your data and assumptions.

Yesterday, we explored various concepts for tackling the second objective of *identification*. Our focus now is on the third objective of *inference*. In particular, for this lecture, we will study approaches based on semiparametric theory.

Objectives

Objectives:

- Provide some hints about the kinds of issues that motivate methods such as double-robustness, double machine learning, targeted maximum likelihood estimation etc.
- Understand the challenges of using nonparametric/machine learning algorithms for causal inference (including nonparametric Bayesian algorithms) and how we might deal with these challenges.

Objectives

Objectives:

- Provide some hints about the kinds of issues that motivate methods such as double-robustness, double machine learning, targeted maximum likelihood estimation etc.
- Understand the challenges of using nonparametric/machine learning algorithms for causal inference (including nonparametric Bayesian algorithms) and how we might deal with these challenges.

Frequentist Bayes

We will work under a frequentist set-up, so a *"model"* is a collection of candidate distributions that could have generated the i.i.d. data $Z_1, \ldots, Z_n$.

This won't stop us from considering Bayesian estimation procedures—it's just that we will be evaluating the performance of our procedures with respect to frequentist metrics like *bias*, *root mean squared error*, *coverage*, etc. And we will study frequentist asymptotic properties like *asymptotic normality* and *efficiency*.

This is sometimes referred to as frequentist Bayes (e.g., by Aad van der Vaart and collaborators).

Frequentist Bayes

We will work under a frequentist set-up, so a *"model"* is a collection of candidate distributions that could have generated the i.i.d. data $Z_1, \dots, Z_n$.

This won't stop us from considering Bayesian estimation procedures—it's just that we will be evaluating the performance of our procedures with respect to frequentist metrics like *bias*, *root mean squared error*, *coverage*, etc. And we will study frequentist asymptotic properties like *asymptotic normality* and *efficiency*.

This is sometimes referred to as frequentist Bayes (e.g., by Aad van der Vaart and collaborators).

Frequentist Bayes

Why should we care about frequentist properties if we're Bayesians?

- A frequentist perspective is ubiquitous in causal inference: applications/collaborators will usually demand frequentist guarantees, e.g., can you test a treatment effect away from zero with Type I error control, can you provide confidence intervals with (approximately) nominal coverage?
- Arguably, causal inference offers a more natural interpretation of frequentism than many other fields. This is because we usually think about a *target superpopulation*.
- An assumption like overlap/positivity is inherently frequentist. Testing/handling overlap (or lack thereof) requires thinking from a frequentist point-of-view. (There's more of this in the second practical.)

Frequentist Bayes

Why should we care about frequentist properties if we're Bayesians?

- A frequentist perspective is ubiquitous in causal inference: applications/collaborators will usually demand frequentist guarantees, e.g., can you test a treatment effect away from zero with Type I error control, can you provide confidence intervals with (approximately) nominal coverage?
- Arguably, causal inference offers a more natural interpretation of frequentism than many other fields. This is because we usually think about a *target superpopulation*.
- An assumption like overlap/positivity is inherently frequentist. Testing/handling overlap (or lack thereof) requires thinking from a frequentist point-of-view. (There's more of this in the second practical.)

Frequentist Bayes

Why should we care about frequentist properties if we're Bayesians?

- A frequentist perspective is ubiquitous in causal inference: applications/collaborators will usually demand frequentist guarantees, e.g., can you test a treatment effect away from zero with Type I error control, can you provide confidence intervals with (approximately) nominal coverage?
- Arguably, causal inference offers a more natural interpretation of frequentism than many other fields. This is because we usually think about a *target superpopulation*.
- An assumption like overlap/positivity is inherently frequentist. Testing/handling overlap (or lack thereof) requires thinking from a frequentist point-of-view. (There's more of this in the second practical.)

Why Bayes?

Some potential benefits of nonparametric Bayes:

- The prior offers a natural approach for regularization, as well as a route for incorporating expert knowledge to increase precision.
- Bayesian algorithms have been shown to provide competitive or even state-of-the-art predictive performance for many problems, e.g. BART, Gaussian processes.
- A Bayesian model can automatically adapt to unknown regularity/complexity parameters.

We want to take advantage of these attractive features of nonparametric Bayes while still obtaining good inference for our target estimands.

Why Bayes?

Some potential benefits of nonparametric Bayes:

- The prior offers a natural approach for regularization, as well as a route for incorporating expert knowledge to increase precision.
- Bayesian algorithms have been shown to provide competitive or even state-of-the-art predictive performance for many problems, e.g. BART, Gaussian processes.
- A Bayesian model can automatically adapt to unknown regularity/complexity parameters.

We want to take advantage of these attractive features of nonparametric Bayes while still obtaining good inference for our target estimands.

Why Bayes?

Some potential benefits of nonparametric Bayes:

- The prior offers a natural approach for regularization, as well as a route for incorporating expert knowledge to increase precision.
- Bayesian algorithms have been shown to provide competitive or even state-of-the-art predictive performance for many problems, e.g. BART, Gaussian processes.
- A Bayesian model can automatically adapt to unknown regularity/complexity parameters.

We want to take advantage of these attractive features of nonparametric Bayes while still obtaining good inference for our target estimands.

Why Bayes?

Some potential benefits of nonparametric Bayes:

- The prior offers a natural approach for regularization, as well as a route for incorporating expert knowledge to increase precision.
- Bayesian algorithms have been shown to provide competitive or even state-of-the-art predictive performance for many problems, e.g. BART, Gaussian processes.
- A Bayesian model can automatically adapt to unknown regularity/complexity parameters.

We want to take advantage of these attractive features of nonparametric Bayes while still obtaining good inference for our target estimands.

What does "semiparametric" even mean?

So in the frequentist set-up, the data $Z_1, \ldots, Z_n$ are drawn iid from an unknown distribution belonging to a model $\mathscr{P}$.

Parametric inference

The model $\mathscr{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is indexed by a **finite-dimensional** parameter. Example: logistic regression.

What does “semiparametric” even mean?

So in the frequentist set-up, the data $Z_1, \ldots, Z_n$ are drawn iid from an unknown distribution belonging to a model $\mathscr{P}$.

Parametric inference

The model $\mathscr{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is indexed by a **finite-dimensional** parameter.
Example: logistic regression.

Nonparametric inference

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

What does “semiparametric” even mean?

So in the frequentist set-up, the data $Z_1, \ldots, Z_n$ are drawn iid from an unknown distribution belonging to a model $\mathscr{P}$.

Parametric inference

The model $\mathscr{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is indexed by a **finite-dimensional** parameter.
Example: logistic regression.

Nonparametric inference

The model $\mathscr{P} = \{P_\eta : \eta \in \mathscr{H}\}$ 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: linear regression with heteroscedastic errors

Linear regression

Consider

$$Y = T\theta + \beta^{\mathrm{T}} X + U, \quad \mathbb{E}[U \mid T, X] = 0,$$

where Y is the outcome, T is the treatment, and X consists of remaining covariates. The target parameter is the treatment effect θ.

This is an example of a strict semiparametric model. The nuisance parameters are β and the conditional distribution of U given (T, X)—the latter is an infinite-dimensional parameter without further model restrictions.

Example 1: linear regression with heteroscedastic errors

Linear regression

Consider

$$Y = T\theta + \beta^{\mathrm{T}} X + U, \quad \mathbb{E}[U \mid T, X] = 0,$$

where Y is the outcome, T is the treatment, and X consists of remaining covariates. The target parameter is the treatment effect θ.

This is an example of a strict semiparametric model. The nuisance parameters are β and the conditional distribution of U given (T, X)—the latter is an infinite-dimensional parameter without further model restrictions.

Example 2: partially linear regression

Partially linear regression

Consider

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

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

This is again a strict semiparametric model. We can partition the parameters into the finite-dimensional target θ and the infinite-dimensional nuisance parameters: $g(x)$ and the law U given (T, X).

For those familiar with survival analysis: many popular models are strict semiparametric models, e.g., proportional hazards, proportional odds, accelerated failure time.

Example 2: partially linear regression

Partially linear regression

Consider

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

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

This is again a strict semiparametric model. We can partition the parameters into the finite-dimensional target θ and the infinite-dimensional nuisance parameters: $g(x)$ and the law U given (T, X).

For those familiar with survival analysis: many popular models are strict semiparametric models, e.g., proportional hazards, proportional odds, accelerated failure time.

Example 3: integrated squared density

Integrated squared density

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

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

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 $\chi : \mathscr{P} \to \mathbb{R}$ from the model to the real line.

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

Example 3: integrated squared density

Integrated squared density

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

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

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 $\chi : \mathscr{P} \to \mathbb{R}$ from the model to the real line.

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

Example 4: average treatment effect

Average treatment effect

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

The target estimand is the ATE $\chi(P) = \mathbb{E}_P[Y^1 - Y^0]$, which is identified by

$$\chi(P) = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)]$$

under the assumptions:

- (unconfoundedness) $Y^t \perp\!\!\!\perp T \mid X$ for $t = 0, 1$.
- (overlap/positivity) $0 < \pi(X) < 1$ with P-probability 1, where $\pi(x) = P(T = 1 \mid X = x)$ is the **propensity score**.

Example 4: average treatment effect

Average treatment effect

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

The target estimand is the ATE $\chi(P) = \mathbb{E}_P[Y^1 - Y^0]$, which is identified by

$$\chi(P) = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)]$$

under the assumptions:

- (unconfoundedness) $Y^t \perp\!\!\!\perp T \mid X$ for $t = 0, 1$.
- (overlap/positivity) $0 < \pi(X) < 1$ with P-probability 1, where $\pi(x) = P(T = 1 \mid X = x)$ is the **propensity score**.

Example 4: average treatment effect

Average treatment effect

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

The target estimand is the ATE $\chi(P) = \mathbb{E}_P[Y^1 - Y^0]$, which is identified by

$$\chi(P) = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)]$$

under the assumptions:

- (unconfoundedness) $Y^t \perp\!\!\!\perp T \mid X$ for $t = 0, 1$.
- (overlap/positivity) $0 < \pi(X) < 1$ with P-probability 1, where $\pi(x) = P(T = 1 \mid X = x)$ is the **propensity score**.

Example 4: average treatment effect

Average treatment effect

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

The target estimand is the ATE $\chi(P) = \mathbb{E}_P[Y^1 - Y^0]$, which is identified by

$$\chi(P) = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)]$$

under the assumptions:

- (unconfoundedness) $Y^t \perp\!\!\!\perp T \mid X$ for $t = 0, 1$.
- (overlap/positivity) $0 < \pi(X) < 1$ with P-probability 1, where $\pi(x) = P(T = 1 \mid X = x)$ is 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 estimation procedures with nice properties like asymptotic normality (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 estimation procedures with nice properties like asymptotic normality (e.g., MLE, parametric Bayes).

Nonparametric inference

Aside from very special cases (like estimating a CDF), we rarely have any asymptotic distribution theory, so no uncertainty quantification.

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 estimation procedures with nice properties like asymptotic normality (e.g., MLE, parametric Bayes).

Nonparametric inference

Aside from very special cases (like estimating a CDF), we rarely have any asymptotic distribution theory, so no uncertainty quantification.

Semiparametric inference

Similar asymptotic theory to the parametric case (if the functional is sufficiently "smooth").

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 nonparametric algorithms/machine learning can be disastrous
  - Potentially huge biases.
  - Unclear whether the bootstrap works (or whether we have asymptotic normality at all). So our uncertainty quantification might be misleading.

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 nonparametric algorithms/machine learning can be disastrous
  - Potentially huge biases.
  - Unclear whether the bootstrap works (or whether we have asymptotic normality at all). So our uncertainty quantification might be misleading.

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 nonparametric algorithms/machine learning can be disastrous
  - Potentially huge biases.
  - Unclear whether the bootstrap works (or whether we have asymptotic normality at all). So our uncertainty quantification might be misleading.

Illustration

Average treatment effect

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

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)].$$

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

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

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

Illustration

Average treatment effect

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

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)].$$

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

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

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

Illustration

Average treatment effect

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

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)].$$

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

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

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

Illustration

Average treatment effect

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

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 1, X)] - \mathbb{E}_X[\mathbb{E}_{Y|T,X}(Y \mid T = 0, X)].$$

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

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

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

Simulation

We simulate some data as follows:

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

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

$$\hat{\chi} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\},$$

where $\hat{\mu}$ is a random forests estimator regression Y on (T, X) implemented with the `randomForest` R package.

Simulation

We simulate some data as follows:

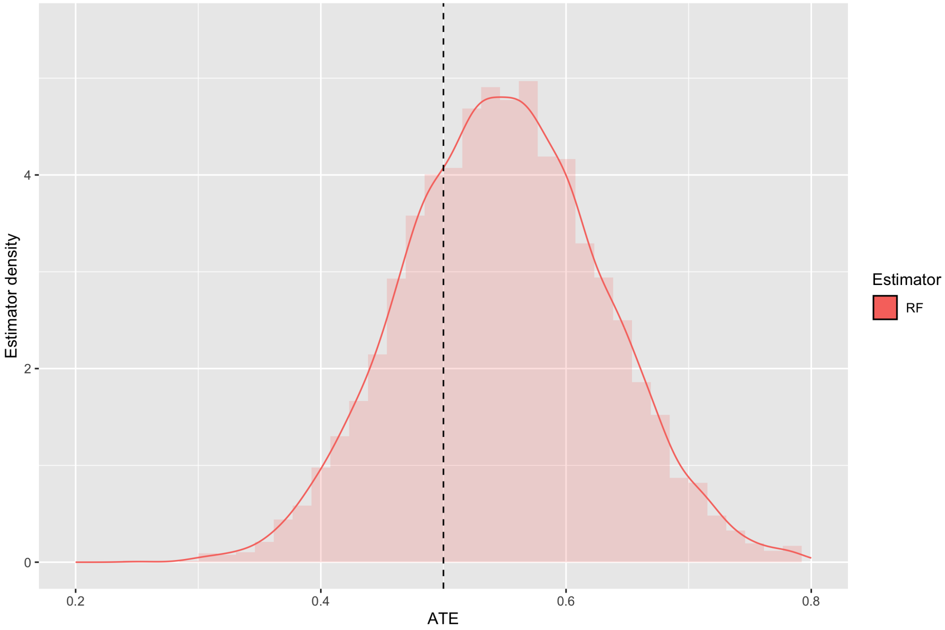
$$\begin{aligned} X &\sim \text{Exponential}(2) \\ T \mid X = x &\sim \text{Bernoulli}(\text{logit}(x)) \\ Y^t &\sim \mathcal{N}(0.5t, \sigma^2). \end{aligned}$$

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

$$\hat{\chi} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\},$$

where $\hat{\mu}$ is a random forests estimator regression Y on (T, X) implemented with the `randomForest` R package.


Estimator density
0
2
4
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."

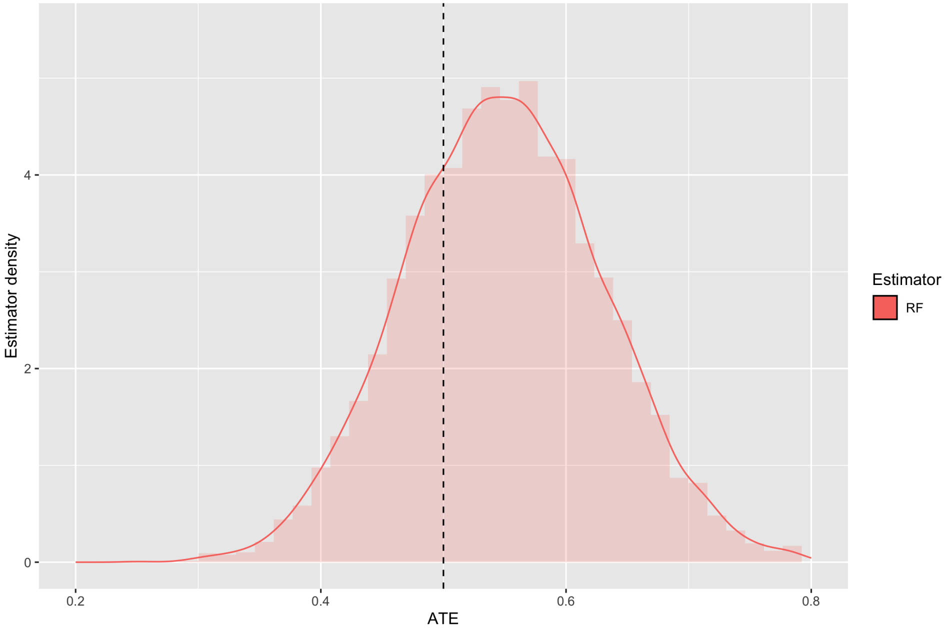
Semiparametric theory to the rescue! We will use the specific structure of the estimand to remove bias and enable rigorous statistical guarantees (e.g. coverage of confidence intervals, Type I error control etc.)

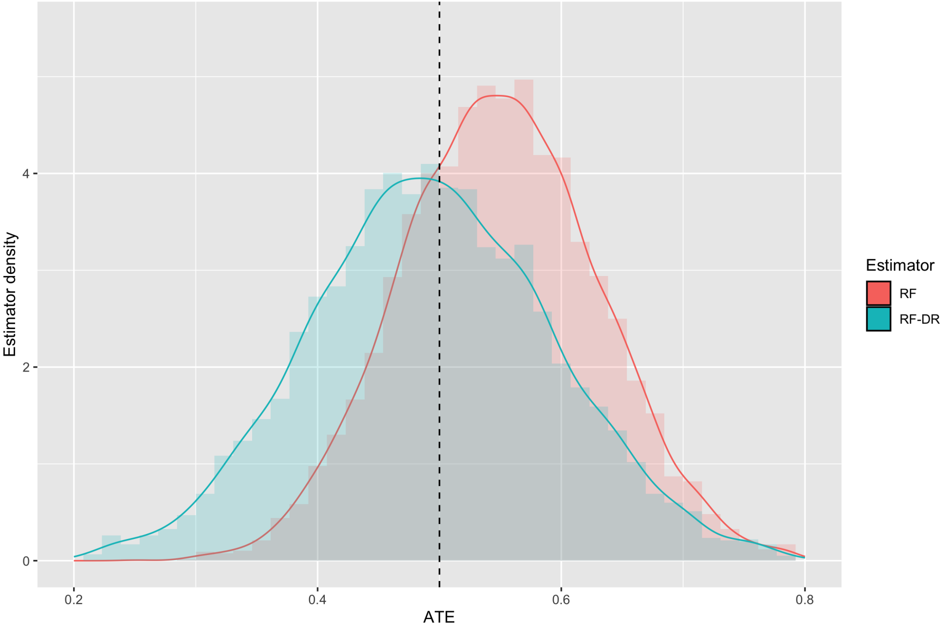
What went wrong?

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
0
2
4
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

The Bayesian case

How does this related to Bayes? It turns out that Bayesian inference is also a plug-in method! And therefore we run into the same types of issues if we (naïvely) use nonparametric Bayesian models to estimate average treatment effects.

For a Bayesian analysis for the ATE, we need:

- A posterior for the marginal distribution of the covariates P_X (e.g. the Bayesian bootstrap)
- A posterior for the outcome regression function $\mu(t,x)$ (e.g. linear regression, Bayesian additive regression trees, Gaussian processes etc.)

Then this automatically gives us the marginal posterior of the ATE:

$$\chi = \mathbb{E}_{P_X}[\mu(1,X) - \mu(0,X)].$$

The Bayesian case

How does this related to Bayes? It turns out that Bayesian inference is also a plug-in method! And therefore we run into the same types of issues if we (naïvely) use nonparametric Bayesian models to estimate average treatment effects.

For a Bayesian analysis for the ATE, we need:

- A posterior for the marginal distribution of the covariates P_X (e.g. the Bayesian bootstrap)
- A posterior for the outcome regression function $\mu(t,x)$ (e.g. linear regression, Bayesian additive regression trees, Gaussian processes etc.)

Then this automatically gives us the marginal posterior of the ATE:

$$\chi = \mathbb{E}_{P_X}[\mu(1,X) - \mu(0,X)].$$

The Bayesian case

How does this related to Bayes? It turns out that Bayesian inference is also a plug-in method! And therefore we run into the same types of issues if we (naïvely) use nonparametric Bayesian models to estimate average treatment effects.

For a Bayesian analysis for the ATE, we need:

- A posterior for the marginal distribution of the covariates P_X (e.g. the Bayesian bootstrap)
- A posterior for the outcome regression function $\mu(t,x)$ (e.g. linear regression, Bayesian additive regression trees, Gaussian processes etc.)

Then this automatically gives us the marginal posterior of the ATE:

$$\chi = \mathbb{E}_{P_X}[\mu(1,X) - \mu(0,X)].$$

The Bayesian case

How does this related to Bayes? It turns out that Bayesian inference is also a plug-in method! And therefore we run into the same types of issues if we (naïvely) use nonparametric Bayesian models to estimate average treatment effects.

For a Bayesian analysis for the ATE, we need:

- A posterior for the marginal distribution of the covariates P_X (e.g. the Bayesian bootstrap)
- A posterior for the outcome regression function $\mu(t,x)$ (e.g. linear regression, Bayesian additive regression trees, Gaussian processes etc.)

Then this automatically gives us the marginal posterior of the ATE:

$$\chi = \mathbb{E}_{P_X}[\mu(1,X) - \mu(0,X)].$$

Standard Bayesian algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples $\{P_X^{(1)},\dots,P_X^{(B)}\}$ and $\{\mu^{(1)},\dots,\mu^{(B)}\}$.
2. Compute

$$\chi^{(b)} = \mathbb{E}_{P_X^{(b)}}[\mu^{(b)}(1,X)-\mu^{(b)}(0,X)]$$

   for each $b = 1,\dots,B$.
3. Output $(\chi^{(1)},\dots,\chi^{(B)})$, which is a sample from the marginal posterior of the ATE.

Notice that we don't model the treatment assignment/propensity score model $\mathbb{P}(T \mid X)$.

Standard Bayesian algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples $\{P_X^{(1)}, \dots, P_X^{(B)}\}$ and $\{\mu^{(1)}, \dots, \mu^{(B)}\}$.
2. Compute
$$\chi^{(b)} = \mathbb{E}_{P_X^{(b)}}[\mu^{(b)}(1,X) - \mu^{(b)}(0,X)]$$
for each $b = 1, \dots, B$.
3. Output $(\chi^{(1)}, \dots, \chi^{(B)})$, which is a sample from the marginal posterior of the ATE.

Notice that we don't model the treatment assignment/propensity score model $\mathbb{P}(T \mid X)$.

Standard Bayesian algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples $\{P_X^{(1)}, \dots, P_X^{(B)}\}$ and $\{\mu^{(1)}, \dots, \mu^{(B)}\}$.
2. Compute
$$\chi^{(b)} = \mathbb{E}_{P_X^{(b)}}[\mu^{(b)}(1, X) - \mu^{(b)}(0, X)]$$
for each $b = 1, \dots, B$.
3. Output $(\chi^{(1)}, \dots, \chi^{(B)})$, which is a sample from the marginal posterior of the ATE.

Notice that we don't model the treatment assignment/propensity score model $\mathbb{P}(T \mid X)$.

Standard Bayesian algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples $\{P_X^{(1)}, \dots, P_X^{(B)}\}$ and $\{\mu^{(1)}, \dots, \mu^{(B)}\}$.
2. Compute
$$\chi^{(b)} = \mathbb{E}_{P_X^{(b)}}[\mu^{(b)}(1,X) - \mu^{(b)}(0,X)]$$
for each $b = 1, \dots, B$.
3. Output $(\chi^{(1)}, \dots, \chi^{(B)})$, which is a sample from the marginal posterior of the ATE.

Notice that we don't model the treatment assignment/propensity score model $\mathbb{P}(T \mid X)$.

Standard Bayesian algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples $\{P_X^{(1)}, \dots, P_X^{(B)}\}$ and $\{\mu^{(1)}, \dots, \mu^{(B)}\}$.
2. Compute
$$\chi^{(b)} = \mathbb{E}_{P_X^{(b)}}[\mu^{(b)}(1, X) - \mu^{(b)}(0, X)]$$
for each $b = 1, \dots, B$.
3. Output $(\chi^{(1)}, \dots, \chi^{(B)})$, which is a sample from the marginal posterior of the ATE.

Notice that we don't model the treatment assignment/propensity score model $\mathbb{P}(T \mid X)$.

The one-step posterior

To avoid similar issues to the random forests example earlier, we're going to post-process our posterior.

What this means is that we can fit our Bayesian model as usual and then we add an extra step at the end that "corrects" the marginal posterior of our target parameter. This correction will make the new posterior debiased/doubly robust.

This time, we need:

- A posterior for the outcome regression function $\mu(t,x)$ (same as before)
- A posterior for the treatment assignment/propensity score $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ (e.g. logistic regression, probit-link BART etc.)

So we **don't** need to model the marginal covariate distribution this time round.

The one-step posterior

To avoid similar issues to the random forests example earlier, we're going to post-process our posterior.

What this means is that we can fit our Bayesian model as usual and then we add an extra step at the end that "corrects" the marginal posterior of our target parameter. This correction will make the new posterior debiased/doubly robust.

This time, we need:

- A posterior for the outcome regression function $\mu(t,x)$ (same as before)
- A posterior for the treatment assignment/propensity score $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ (e.g. logistic regression, probit-link BART etc.)

So we **don't** need to model the marginal covariate distribution this time round.

The one-step posterior

To avoid similar issues to the random forests example earlier, we're going to post-process our posterior.

What this means is that we can fit our Bayesian model as usual and then we add an extra step at the end that "corrects" the marginal posterior of our target parameter. This correction will make the new posterior debiased/doubly robust.

This time, we need:

- A posterior for the outcome regression function $\mu(t, x)$ (same as before)
- A posterior for the treatment assignment/propensity score $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ (e.g. logistic regression, probit-link BART etc.)

So we **don't** need to model the marginal covariate distribution this time round.

The one-step posterior

To avoid similar issues to the random forests example earlier, we're going to post-process our posterior.

What this means is that we can fit our Bayesian model as usual and then we add an extra step at the end that "corrects" the marginal posterior of our target parameter. This correction will make the new posterior debiased/doubly robust.

This time, we need:

- A posterior for the outcome regression function $\mu(t, x)$ (same as before)
- A posterior for the treatment assignment/propensity score $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ (e.g. logistic regression, probit-link BART etc.)

So we **don't** need to model the marginal covariate distribution this time round.

The one-step posterior

To avoid similar issues to the random forests example earlier, we're going to post-process our posterior.

What this means is that we can fit our Bayesian model as usual and then we add an extra step at the end that "corrects" the marginal posterior of our target parameter. This correction will make the new posterior debiased/doubly robust.

This time, we need:

- A posterior for the outcome regression function $\mu(t, x)$ (same as before)
- A posterior for the treatment assignment/propensity score $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ (e.g. logistic regression, probit-link BART etc.)

So we **don't** need to model the marginal covariate distribution this time round.

Correction algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples of the propensity score $\{\pi^{(1)}, \dots, \pi^{(B)}\}$ and the outcome regression function $\{\mu^{(1)}, \dots, \mu^{(B)}\}$.
2. Pair each posterior sample $(\pi^{(b)}, \mu^{(b)})$ with an independent draw of $(W_1^{(b)}, \dots, W_n^{(b)})$ from $\mathrm{Dir}(n; 1, \dots, 1)$ (these are Bayesian bootstrap weights).
3. Evaluate

$$\tilde{\chi}^{(b)} = \sum_{i=1}^{n} W_i^{(b)} \left[\mu^{(b)}(1, X_i) - \mu^{(b)}(0, X_i) + \underbrace{\frac{T_i(Y_i - \mu^{(b)}(1, X_i))}{\pi^{(b)}(X_i)} - \frac{(1 - T_i)(Y_i - \mu^{(b)}(0, X_i))}{1 - \pi^{(b)}(X_i)}}_{\text{double-robust correction term}} \right]$$

   for every posterior sample.
4. Output $(\tilde{\chi}^{(1)}, \dots, \tilde{\chi}^{(B)})$, which is a sample from the one-step posterior.

Correction algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples of the propensity score $\{\pi^{(1)}, \ldots, \pi^{(B)}\}$ and the outcome regression function $\{\mu^{(1)}, \ldots, \mu^{(B)}\}$.
2. Pair each posterior sample $(\pi^{(b)}, \mu^{(b)})$ with an independent draw of $(W_1^{(b)}, \ldots, W_n^{(b)})$ from $\mathrm{Dir}(n; 1, \ldots, 1)$ (these are Bayesian bootstrap weights).
3. Evaluate

$$\tilde{\chi}^{(b)} = \sum_{i=1}^{n} W_i^{(b)} \left[\mu^{(b)}(1, X_i) - \mu^{(b)}(0, X_i) + \underbrace{\frac{T_i(Y_i - \mu^{(b)}(1, X_i))}{\pi^{(b)}(X_i)} - \frac{(1 - T_i)(Y_i - \mu^{(b)}(0, X_i))}{1 - \pi^{(b)}(X_i)}}_{\text{double-robust correction term}} \right]$$

   for every posterior sample.
4. Output $(\tilde{\chi}^{(1)}, \ldots, \tilde{\chi}^{(B)})$, which is a sample from the one-step posterior.

Correction algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples of the propensity score $\{\pi^{(1)}, \ldots, \pi^{(B)}\}$ and the outcome regression function $\{\mu^{(1)}, \ldots, \mu^{(B)}\}$.
2. Pair each posterior sample $(\pi^{(b)}, \mu^{(b)})$ with an independent draw of $(W_1^{(b)}, \ldots, W_n^{(b)})$ from $\mathrm{Dir}(n; 1, \ldots, 1)$ (these are Bayesian bootstrap weights).
3. Evaluate

$$\tilde{\chi}^{(b)} = \sum_{i=1}^{n} W_i^{(b)} \left[\mu^{(b)}(1, X_i) - \mu^{(b)}(0, X_i) + \underbrace{\frac{T_i(Y_i - \mu^{(b)}(1, X_i))}{\pi^{(b)}(X_i)} - \frac{(1 - T_i)(Y_i - \mu^{(b)}(0, X_i))}{1 - \pi^{(b)}(X_i)}}_{\text{double-robust correction term}} \right]$$

   for every posterior sample.
4. Output $(\tilde{\chi}^{(1)}, \ldots, \tilde{\chi}^{(B)})$, which is a sample from the one-step posterior.

Correction algorithm

1. Run your posterior computation algorithm of choice (e.g. MCMC, HMC) to obtain posterior samples of the propensity score $\{\pi^{(1)}, \dots, \pi^{(B)}\}$ and the outcome regression function $\{\mu^{(1)}, \dots, \mu^{(B)}\}$.
2. Pair each posterior sample $(\pi^{(b)}, \mu^{(b)})$ with an independent draw of $(W_1^{(b)}, \dots, W_n^{(b)})$ from $\mathrm{Dir}(n; 1, \dots, 1)$ (these are Bayesian bootstrap weights).
3. Evaluate

$$\tilde{\chi}^{(b)} = \sum_{i=1}^{n} W_i^{(b)} \left[\mu^{(b)}(1, X_i) - \mu^{(b)}(0, X_i) + \underbrace{\frac{T_i(Y_i - \mu^{(b)}(1, X_i))}{\pi^{(b)}(X_i)} - \frac{(1 - T_i)(Y_i - \mu^{(b)}(0, X_i))}{1 - \pi^{(b)}(X_i)}}_{\text{double-robust correction term}} \right]$$

   for every posterior sample.
4. Output $(\tilde{\chi}^{(1)}, \dots, \tilde{\chi}^{(B)})$, which is a sample from the one-step posterior.

Conclusions

- If we want to use flexible, nonparametric methods like BART to estimate causal effects, we need to take extra care—a naïve application of these methods will generally result in large biases, misleading uncertainty quantification etc.
- What we can do is introduce a post-processing step that attaches onto any existing posterior sampling implementation without the need to modify the original algorithm.
- The correction will debias and correct the shape of the marginal posterior for our target parameter.
- Preprint: *"Semiparametric posterior corrections"*
  `https://arxiv.org/abs/2306.06059`
  Seminar recording:
  `https://www.youtube.com/watch?v=uMQmHCKQ-zw`

Conclusions

- If we want to use flexible, nonparametric methods like BART to estimate causal effects, we need to take extra care—a naïve application of these methods will generally result in large biases, misleading uncertainty quantification etc.
- What we can do is introduce a post-processing step that attaches onto any existing posterior sampling implementation without the need to modify the original algorithm.
- The correction will debias and correct the shape of the marginal posterior for our target parameter.
- Preprint: *"Semiparametric posterior corrections"*
  `https://arxiv.org/abs/2306.06059`
  Seminar recording:
  `https://www.youtube.com/watch?v=uMQmHCKQ-zw`

Conclusions

- If we want to use flexible, nonparametric methods like BART to estimate causal effects, we need to take extra care—a naïve application of these methods will generally result in large biases, misleading uncertainty quantification etc.
- What we can do is introduce a post-processing step that attaches onto any existing posterior sampling implementation without the need to modify the original algorithm.
- The correction will debias and correct the shape of the marginal posterior for our target parameter.
- Preprint: *"Semiparametric posterior corrections"*
  `https://arxiv.org/abs/2306.06059`
  Seminar recording:
  `https://www.youtube.com/watch?v=uMQmHCKQ-zw`

Conclusions

- If we want to use flexible, nonparametric methods like BART to estimate causal effects, we need to take extra care—a naïve application of these methods will generally result in large biases, misleading uncertainty quantification etc.
- What we can do is introduce a post-processing step that attaches onto any existing posterior sampling implementation without the need to modify the original algorithm.
- The correction will debias and correct the shape of the marginal posterior for our target parameter.
- Preprint: *"Semiparametric posterior corrections"*
  `https://arxiv.org/abs/2306.06059`
  Seminar recording:
  `https://www.youtube.com/watch?v=uMQmHCKQ-zw`