Interpretable and Stable Machine Learning for Causal Inference

Jasjeet Sekhon

UC Berkeley

September 22, 2017

Causal Inference and Big Data

- Measuring human activity has generated massive datasets with granular population data: e.g.,
  - Browsing, search, and purchase data from online platforms
  - Internet of things
  - Electronic medical records, genetic markers
  - Administrative data: schools, criminal justice, IRS
- Big in size and breadth: wide datasets
- Many inferential issues: e.g., heterogeneity, targeting optimal treatments, interpretable results, stability
- Team: Peter Bickel and Bin Yu along with Nicolai Meinshausen and Bernhard Schölkopf

ML Prediction versus Causal Inference

- Causal Inference is like a prediction problem: but predicting something we don't directly observe and possibly cannot estimate well in a given sample
- ML algorithms are good at prediction, but have issues with causal inference:
  - Interventions imply counterfactuals: response schedule versus model prediction
  - Validation requires estimation in the case of causal inference
  - Identification problems not solved by large data
  - Predicting the outcome mistaken for predicting the causal effect
    - targeting based on the lagged outcome

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

1 (classical) statistical theory:
it works because we have **relevant** theory that tells us it should
Hopefully, this is not simply: "Assume that the data are generated by the following model ..." (Brieman 2001)

2 Training/test loop:
it works because we have validated against ground truth and it works

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

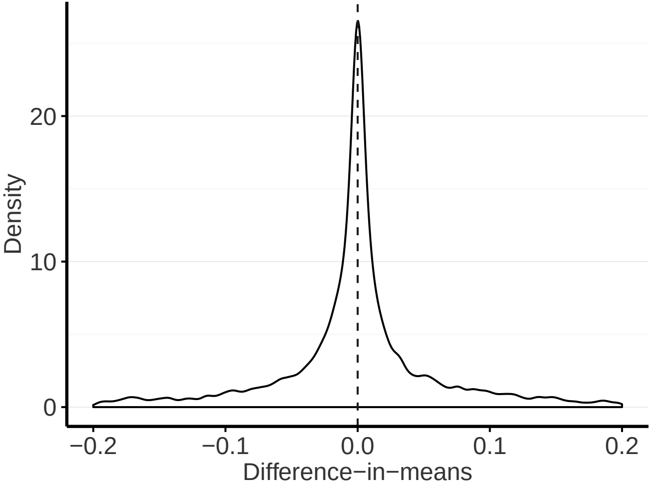
1 (classical) statistical theory:
it works because we have **relevant** theory that tells us it should
Hopefully, this is not simply: "Assume that the data are generated by the following model ..." (Brieman 2001)

2 Training/test loop:
it works because we have validated against ground truth and it works

On the **normal distribution**:

> *"Everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact." — Henri Poincaré (quoted by de Finetti 1975)*

Distribution of Treatment Effects



Shem-Tov and Sekhon (2017)

Conditional Average Treatment Effect (CATE)

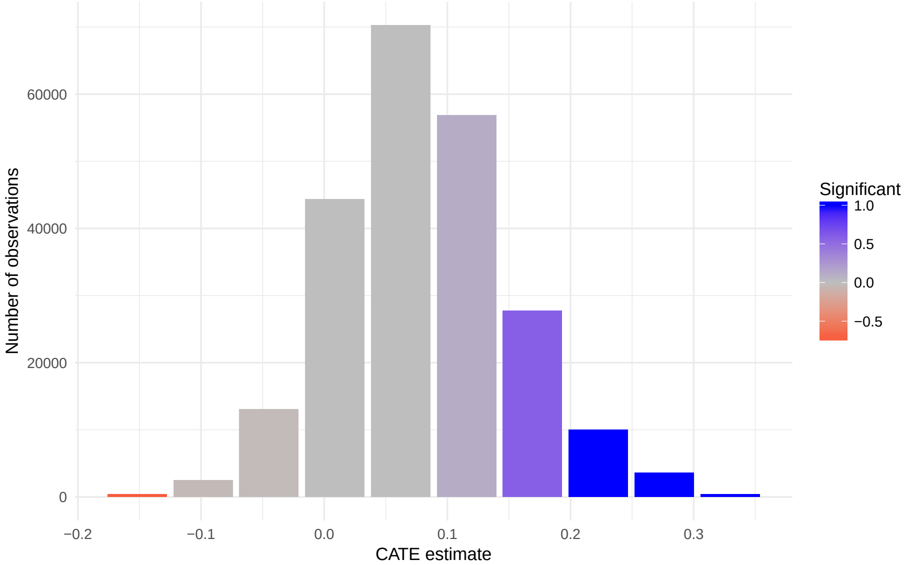
Individual Treatment Effect (ITE): $D_i := Y_i(t) - Y_i(c)$

Let $\hat{\tau}_i$ be an estimator for D_i

$\tau(x_i)$ is the **CATE** for all units whose covariate vector is equal to x_i:

$$CATE := \tau(x_i) := \mathbb{E}\Big[D \Big| X = x_i\Big] = \mathbb{E}\Big[Y(t) - Y(c) \Big| X_i = x_i\Big]$$

GOTV: Social pressure (Gerber, Green, Lairmer, 2008)



How to estimate the CATE?

Meta–learners

A meta–learner decomposes the problem of estimating the CATE into several sub–regression problems. The estimator which solve those sub–problems are called **base–learners**

- Flexibility to choose base–learners which work well in a particular setting
- Deep Learning, (honest) Random Forests, BART, or other machine learning algorithms

How to estimate the CATE?

$$
\begin{aligned}
\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\
&= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\
&= \mu_1(x) - \mu_0(x)
\end{aligned}
$$

How to estimate the CATE?

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mu_1(x) - \mu_0(x)\end{aligned}$$

T–learner

1.) Split the data into control and treatment group,

2.) Estimate the response functions separately,

$$\begin{aligned}\hat{\mu}_1(x) &= \hat{\mathbb{E}}[Y^{obs}|X = x, W = 1] \\ \hat{\mu}_0(x) &= \hat{\mathbb{E}}[Y^{obs}|X = x, W = 0],\end{aligned}$$

3.) $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$

How to estimate the CATE?

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mu_1(x) - \mu_0(x)\end{aligned}$$

T–learner

1.) Split the data into control and treatment group,

2.) Estimate the response functions separately,

$$\hat{\mu}_1(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 1]$$
$$\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = 0],$$

3.) $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$

S–learner

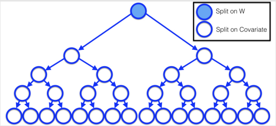
1.) Use the treatment assignment as a usual variable without giving it any special role and estimate

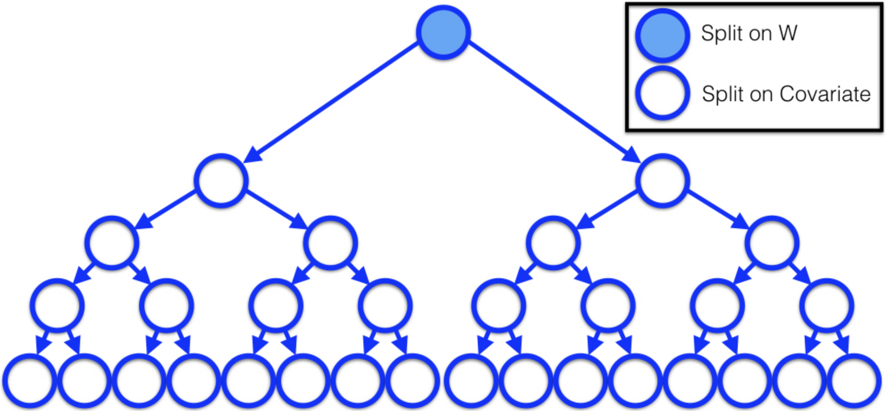
$$\hat{\mu}(x, w) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = w]$$

2.) $\hat{\tau}(x) := \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$

CATE $:= \hat{\tau}(x) = f(x, w=1) - f(x, w=0)$

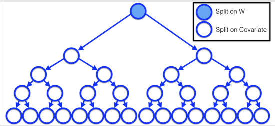
T-Learner



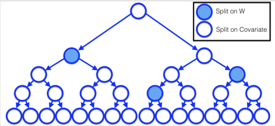


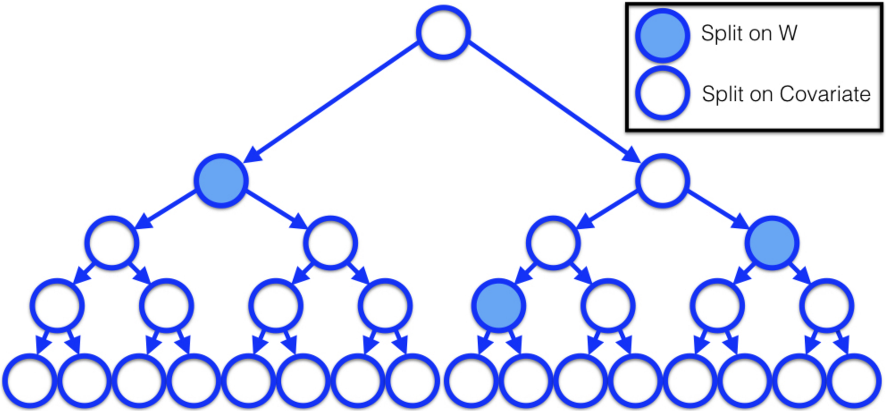
CATE $:= \hat{\tau}(x) = f(x, w = 1) - f(x, w = 0)$

T-Learner

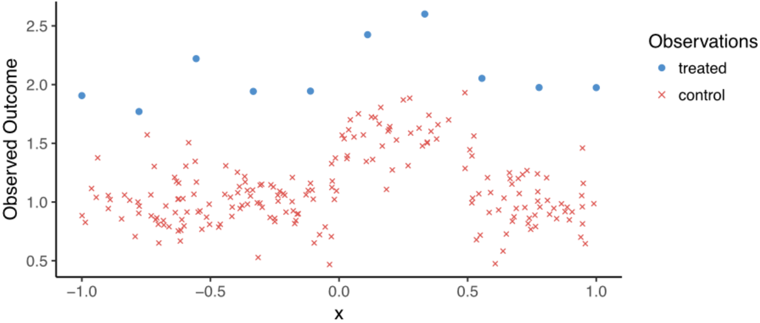


S-Learner

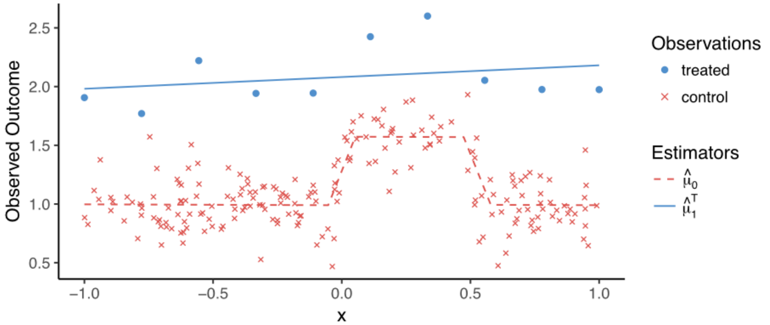




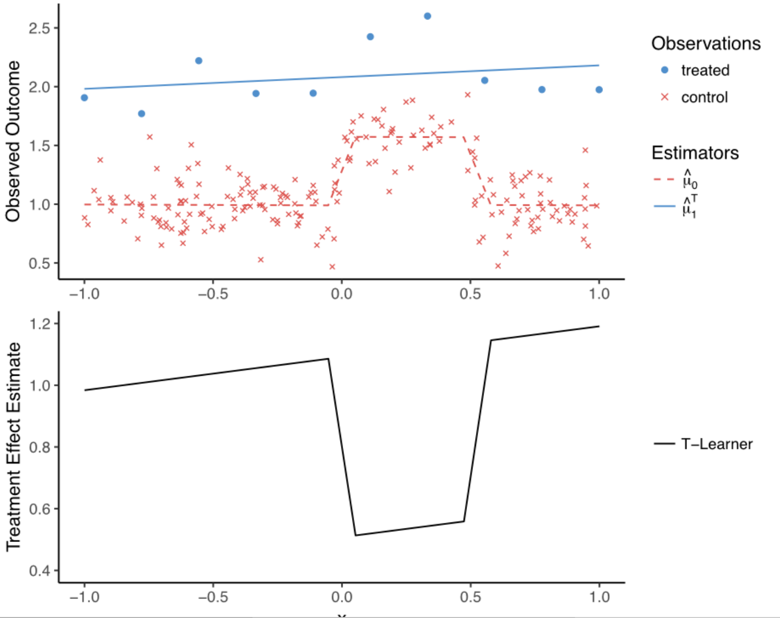
Motivating X–learner



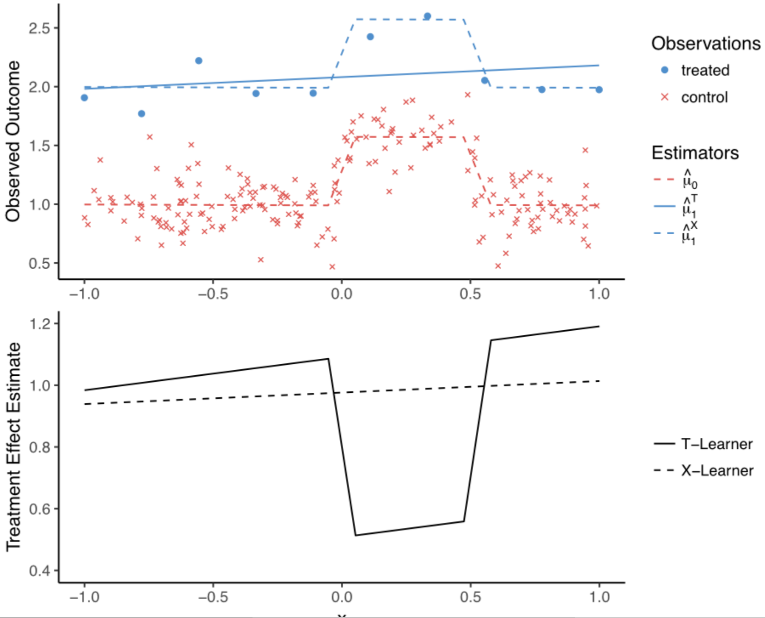
Motivating X–learner



Motivating X–learner



Motivating X–learner



Formal definition of the X–learner

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x]\\ &= \mathbb{E}[Y(1) - \mu_0(x)|X = x]\end{aligned}$$

with $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$.

X–learner

1.) Estimate the control response function,

$$\hat{\mu}_0(x) = \hat{\mathbb{E}}[Y(0)|X = x],$$

2.) Define the **pseudo residuals**,

$$\tilde{D}_i^1 := Y_i(1) - \hat{\mu}_0(X_i(1)),$$

3.) Estimate the CATE,

$$\hat{\tau}(x) = \hat{\mathbb{E}}[\tilde{D}^1|X = x].$$

X–learner in algorithmic form

procedure $\textsc{X–Learner}(X, Y^{obs}, W)$

$\hat{\mu}_0 = M_1(Y^0 \sim X^0)$ ▷ Estimate response function
$\hat{\mu}_1 = M_2(Y^1 \sim X^1)$

$\tilde{D}^1_i := Y^1_i - \hat{\mu}_0(X^1_i)$ ▷ Compute pseudo residuals
$\tilde{D}^0_i := \hat{\mu}_1(X^0_i) - Y^0_i$

$\hat{\tau}_1 = M_3(\tilde{D}^1 \sim X^1)$ ▷ Estimate CATE
$\hat{\tau}_0 = M_4(\tilde{D}^0 \sim X^0)$

$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$ ▷ Average

Algorithm 1: X–learner

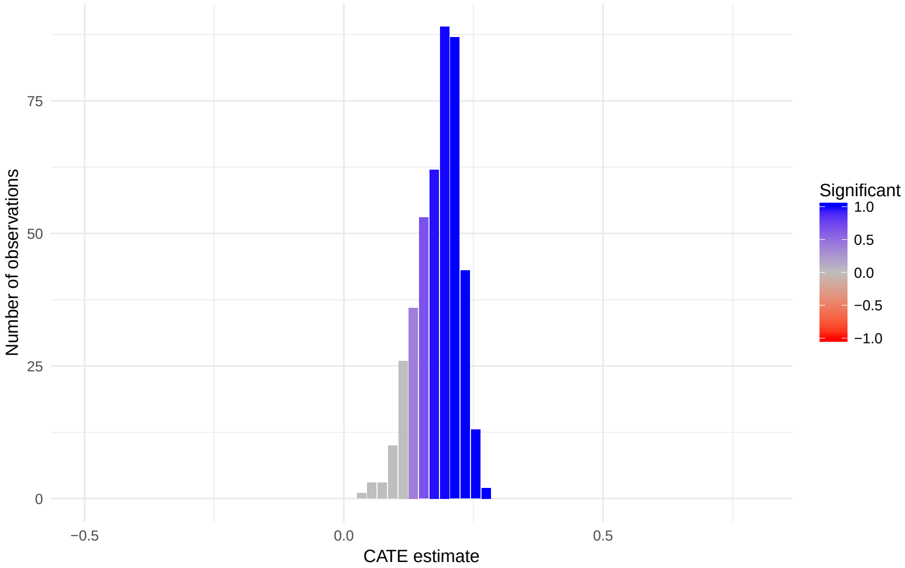


Figure: Reducing Transphobia: X–RF

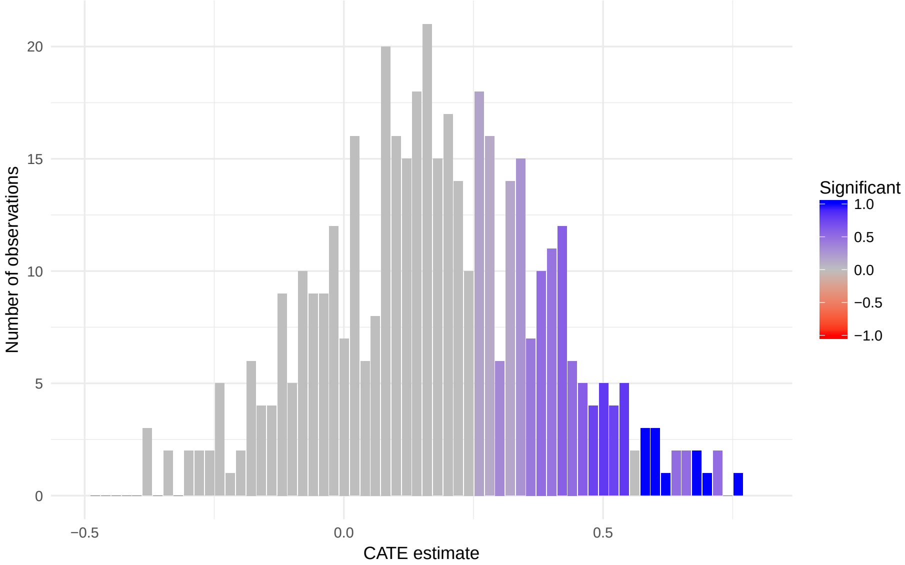


Figure: Reducing Transphobia: T–RF

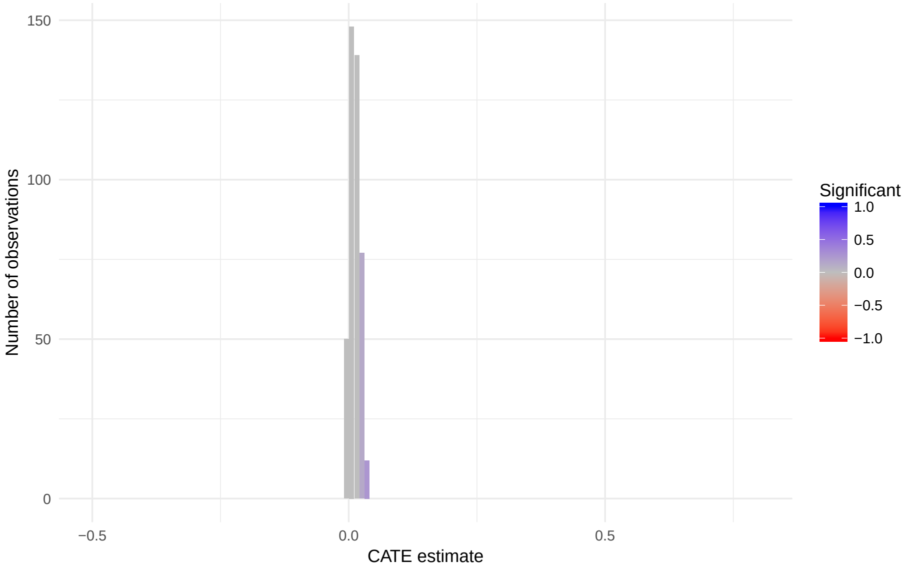


Figure: Reducing Transphobia: S–RF

Dimensionality and Identification

Ignorability	Overlap
$(Y(0), Y(1)) \perp\!\!\!\perp W \mid X$	$0 < e(X) < 1$ w.p. 1
More Plausible (with exceptions, e.g., M-bias)	Less Plausible
"Blessing"	"Curse"

Information accumulates in columns

Sample Size (n)

Nuisance Size (p)

	bookletnumber	sex	daicu	sedpar24	ahsurv	shanew	pac	tc	new_qals	icnno	type	unitsize	bestrate
[illegible]	1	F	10/21/2001	1	1	5	1	24570.084	0.000000	H05	1	3	0.0094850948
[illegible]	2	M	10/22/2001	1	1	5	0	20690.597	0.000000	H05	1	3	0.0094850948
[illegible]	3	F	10/27/2001	1	1	5	1	2586.325	0.000000	H05	1	3	0.0094850948
[illegible]	4	F	1/5/2002	1	1	[illegible]	[illegible]	[illegible]	0.000000	H05	1	3	0.0094850948
[illegible]	5	M	5/19/2002	1	0	5	0	40974.166	3.033727	H05	1	[illegible]	0.0094850948
[illegible]	6	F	7/21/2002	1	0	5	1	78764.935	5.014604	H05	1	3	0.0094850948
[illegible]	9	M	10/2/2002	1	0	5	0	36000.293	9.562864	H05	1	3	0.0094850948
8	10	F	10/12/2002	0	1	5	1	2586.325	0.000000	H05	1	3	0.0094850948
9	12	F	1/21/2002	1	0	5	1	30068.207	12.177965	H34	0	1	0.0116279069
10	13	M	7/11/2002	1	1	5	0	29742.733	0.000000	H34	0	1	0.0116279069

Intuition in High Dimensions

Information accumulates in columns **and rows**.

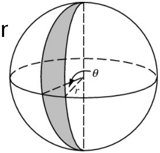
	bookletnumber	sex	daicu	sedpar24	ahsurv	shanew	pac	tc	new_qals	icnno	type	unitsize	bestrate
1	1	F	10/21/2001	1	1	5	1	24570.084	0.000000	H05	1	3	0.0094850948

Stochastic Process Framing

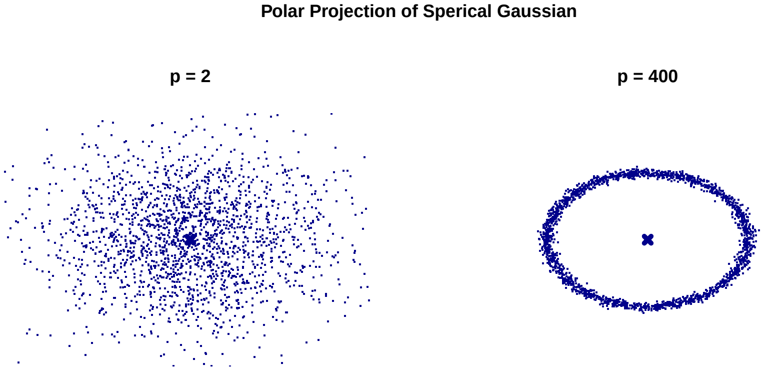
- Let $(X^{(k)})_{k>0}$ be a stochastic process.
- Covariate vector $X_{1:p}$ is a sample of length p from this process.
- Statistics of $X_{1:p}$ can concentrate as p grows.
- Drives counterintuitive behavior of high-dimensional random vectors.

Counterintuitive: Shell Concentration

$\|X_{1:p} - \mu_{1:p}\|$ is function of a sum of p variables. Concentrates in mar



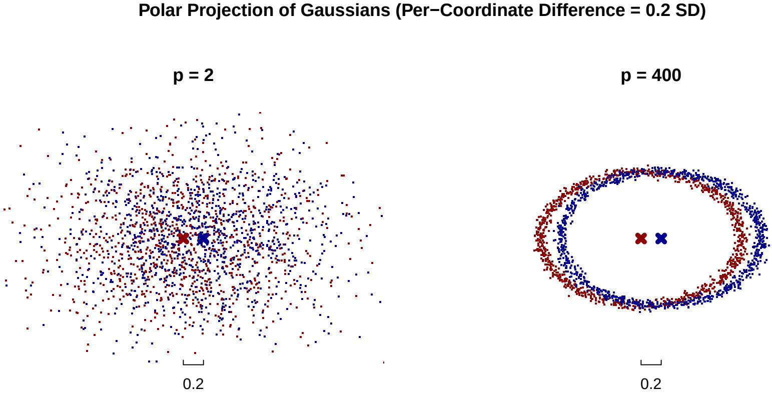
$\|X_{1:p} - \mu_{1:p}\|$ concentrates $\Rightarrow X_{1:p}$ concentrates on a **shell**.



Structure can accumulate in high dimensions. Projection Details

Concentration of Discriminating Information

Likewise, small covariate-wise discrepancies accumulate.



Need to recalibrate intuition for overlap for high dimensions. Projection Details

Analytical Framework

Covariate vector $X_{1:p}$ selected from covariate sequence $(X^{(k)})_{k>0}$.

View covariates **generatively**.

Define **control** and **treated** covariate probability measures, for all p:

$$P_0(X_{1:p} \in A) := P(X_{1:p} \in A \mid W = 0),$$
$$P_1(X_{1:p} \in A) := P(X_{1:p} \in A \mid W = 1).$$

Correspondence with propensity score

$$\frac{e(X_{1:p})}{1 - e(X_{1:p})} = \frac{P(W = 1)}{P(W = 0)} \frac{dP_1(X_{1:p})}{dP_0(X_{1:p})}.$$

Strict Overlap

In practice, make **strict overlap** assumption with bound η.

$$\eta < e(X_{1:p}) < 1 - \eta \quad \text{w.p. } 1.$$

When $P(W = 1) = 0.5$, equivalent statement:

$$\frac{\eta}{1-\eta} < \frac{dP_1(X_{1:p})}{dP_0(X_{1:p})} < \frac{1-\eta}{\eta} \quad \text{w.p. } 1.$$

(For remainder of results, assume $P(W = 1) = 0.5$. Paper includes general case.)

Necessary condition for bounded semiparametric efficiency bound.

Implications: Gaussian Case

Suppose P_0 and P_1 are Gaussian measures

$$X_{1:p} \mid W = 1 \sim N\left(\mu_{1,1:p}, \Sigma_{1,1:p}\right) \quad \text{and} \quad X_{1:p} \mid W = 0 \sim N\left(\mu_{0,1:p}, \Sigma_{0,1:p}\right).$$

Theorem (Gaussian Mean Mahalanobis Distance Bound)

Strict overlap with bound η implies that the Mahalanobis distance with respect to $\Sigma_{1,1:p}$ between the means $\mu_{0,1:p}$ and $\mu_{1,1:p}$ is bounded by

$$\left\|\Sigma_{1,1:p}^{-1/2}(\mu_{0,1:p} - \mu_{1,1:p})\right\| \leq \sqrt{2\left|\log \frac{\eta}{1-\eta}\right|}. \tag{1}$$

Implications: Gaussian Case and p

For large p, strict overlap implies

most covariate means are arbitrarily close together

if the largest eivenvalue of $\Sigma_{1,1:p}$ doesn't grow too fast.

Corollary

Let $\|\Sigma_{1,1:p}^{1/2}\|_{op}$ *be the operator norm of* $\Sigma_{1,1:p}^{1/2}$.

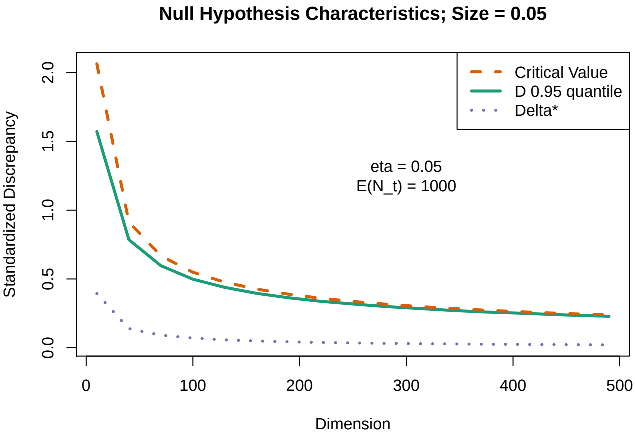
$$\frac{1}{p}\sum_{i=1}^{p}\left|\mu_0^{(k)} - \mu_1^{(k)}\right| \leq p^{-1/2}\|\Sigma_{1,1:p}^{1/2}\|_{op}\sqrt{2\left|\log\frac{\eta}{1-\eta}\right|}.$$

Bound goes as $p^{-1/2}$ for independent, bounded variance case. Converges to zero if $\|\Sigma_{1,1:p}^{1/2}\|_{op}$ is $o(p^{1/2})$, i.e., if effective dimension of $X_{1:p}$ increases with p.

Gaussian Case is Testable

Mean discrepancy bound is explicit and **testable**.

Bound (dotted) and rejection threshold for observed discrepancy (dashed). Details



More General Results

Additional results under distributional assumptions:

- **Gaussian Case**: Bounds on discrepancy in covariance structure.
- **Sub-exponential case**: Mean discrepancy bound random variables. Details

General results for all P_0, P_1:

- For all p, upper bound on **KL divergence** between $P_0(X_{1:p})$ and $P_1(X_{1:p})$ Details
- For all p, lower bound on **test error rate** when discriminating $P_0(X_{1:p})$ from $P_1(X_{1:p})$ Details
- For large p, **no consistent test** of P_0 against P_1
- For large p, **no consistent estimation** of imbalanced parameters $\psi(P_0)$ and $\psi(P_1)$

Way Out? Low-Dimensional Assignment Mechanism

Sufficient condition for strict overlap with respect to any $X_{1:p} \subset (X^{(k)})_{k>0}$.

Assumption (Sufficient Condition for Strict Overlap)

There exists a fixed **balancing variable** B that satisfies

$$X_{1:p} \perp\!\!\!\perp W \mid B \quad \forall X_{1:p} \subset (X^{(k)})_{k>0}, \tag{2}$$

and strict overlap holds with respect to B.

Examples:

- Propensity score is **sparse** (only a function of fixed $X_s \subset (X^{(k)})_{k>0}$).
- Propensity score is a function of **latent class** or **latent factor**.

Way Out? Low-Dimensional Assignment Mechanism

Confounding can be eliminated by some $X_{1:p}$ **only if** it can be eliminated by B.

Proposition

Suppose that B is a balancing variable with respect to covariate sequence $(X^{(k)})_{k>0}$. Then unconfoundendess holds with respect to some covariate set $X_{1:p}$ only if ignorability holds with respect to B.

Tension

- If B is **complex** (e.g., has representation as high-dimensional Gaussian), **overlap** in B is implausible.
- If B is **simple** (e.g., has representation as low-dimensional Gaussian), **ignorability** given B is implausible.

Implications: Regular Semiparametric Estimators (1/2)

Estimate ATE at parametric rates with non-parametric modeling assumptions.

Modular techniques that are ML-compatible. Common threads:

- Estimate $e(X_{1:p})$ and $\mathbb{E}_P[Y \mid W = w, X_{1:p}]$ with predictive models.
- Combine, using efficient influence curve to obtain estimate, perform inference.

With ML/sample splitting in estimation step, applied in high-dimensional settings.

Examples:

- Super Learner + TMLE: Van der Laan and Rose 2011.
- Double/Debiased ML: Chernozhukov et al 2017+.

Implications: Regular Semiparametric Estimators (2/2)

Variance is lower-bounded by **semiparametric efficiency bound**.

$$V^{eff} = \mathbb{E}\left[\frac{\mathrm{Var}(Y(1) \mid X_{1:p})}{e(X_{1:p})} + \frac{\mathrm{Var}(Y(0) \mid X_{1:p})}{1 - e(X_{1:p})} + (\tau(X_{1:p}) - \tau^{ATE})^2\right],$$

where $\tau(X_{1:p})$ is the conditional average treatment effect.

Without strict overlap, variance lower bound is unbounded.

Weak modeling assumptions $\Rightarrow$ strong overlap assumptions.

No causal free lunch in high dimensions.

Ways Forward: Covariate Reduction (1/2)

Overlap assumption can be **relaxed** if covariates are reduced.

If a reduction $d(X_{1:p})$ **discards discriminating information**, but satisfies

$$(Y(0), Y(1)) \perp\!\!\!\perp W \mid d(X_{1:p}),$$

ATE is identified under weaker overlap condition on $d(X_{1:p})$.

Such a $d(X_{1:p})$:

- Cannot be a **balancing score**, i.e., $W \not\perp\!\!\!\perp X_{1:p} \mid d(X_{1:p})$.
- Cannot be characterized by assignment mechanism $P(T \mid W)$ alone.
- Requires information about **outcome** process $P(Y(0), Y(1) \mid W)$.

Ways Forward: Covariate Reduction (2/2)

Example: Generalized prognistic score $r(X_{1:p})$ satisfying

$$(Y(0), Y(1)) \perp\!\!\!\perp X_{1:p} \mid r(X_{1:p})$$

and is a deconfounding score. See approach by Luo et al 2017.

Future Work:

- Process machine learning estimates of $e(X_{1:p})$ and $\mathbb{E}_P[Y \mid W = w, X_{1:p}]$ to estimate $d(X_{1:p})$ that is a function of **both treatment and outcome**. Related to C-TMLE (van der Laan and Gruber 2010).
- Combine **multiple deconfounding scores** $d(X_{1:p})$ to efficiently eliminate nuisance functions, as in regular semiparametric estimation.
- Frame **ignorability as a constraint** in dimension reduction approaches. Three-way relationship; more complex than regression.

Thanks

- Peter Bickel
- Alexander D'Amour
- Peng Ding
- Avi Feller
- Sören Künzel
- Yotam Shem-Tov
- Bin Yu

http://sekhon.berkeley.edu

Superpopulation Setup

Binary treatment W.
Potential outcomes $Y(0), Y(1)$.
Covariates X.

Superpopulation generates triples $((Y(0), Y(1)), W, X) \sim P$. Propensity score: $e(X) := P(W = 1 \mid X)$.

Observe (Y^{obs}, W, X) where $Y^{obs} = (1 - W)Y(0) + WY(1)$.

Estimand is average treatment effect.

$$\tau^{ATE} = E[Y(1) - Y(0)].$$

KL Divergence Bound (1/2)

Strict overlap $\Rightarrow$ P_0 and P_1 cannot be too far apart in terms of **KL divergence**.

Theorem

For $P(W=1)=0.5$, *the strict overlap assumption with bound* η *implies*

$$KL(P_0(X_{1:p}) \| P_1(X_{1:p})) < \left| \log \frac{\eta}{1-\eta} \right|. \tag{3}$$

and vice versa, for P_0 *and* P_1 *switched.*

Follows almost immediately from probability ratio representation of overlap. back

KL Divergence Bound (2/2)

KL divergence accumulates **additively** and **non-decreasingly** in p.

Bound is constant.

Unique discriminating information added by each covariate on average must converge to zero for large p.

Corollary

Let $(X^{(k)})_{k>0}$ be a sequence of covariates, and for each p, let $X_{1:p}$ be a finite subset of $(X^{(k)})_{k>0}$. As p grows large, strict overlap with fixed bound η implies

$$\frac{1}{p}\sum_{k=1}^{p}\mathbb{E}_{P_0} KL(P_0(X^{(k)} \mid X_{1:k-1}) \| P_1(X^{(k)} \mid X_{1:k-1})) \to 0. \tag{4}$$

back

What does an overlap failure look like?

Many opportunities for overlap failure in "big data", particularly when treatment-assigning agents are identifiable.

Example (Deterministic Medical Decision)

Suppose that data are collected where treatment assignment decisions are made by agents using a deterministic rule that varies by every agent.

If the covariate sequence $(X^{(k)})_{k>0}$ contains all of the inputs that go into the decision, and indicators for every agent, then overlap fails.

In **electronic health records**, this can occur when each doctor follows a particular deterministic medical protocol, and the doctor is identified in the data.

Note: Overlap fails even if the protocol is only deterministic for some segment of the population.

Gaussian Test Setting

Suppose $X_{1:p}$ spherical Gaussian under P_0, P_1 with $\sigma = 1$.

Let $\Delta(P_0, P_1)$ be normalized mean discrepancy $p^{1/2}\|\mu_{1,1:p} - \mu_{0,1:p}\|$.
Let $\Delta^*_{p,\eta}$ be upper bound induced by strict overlap with bound η.

Least favorable test is:

$$H_0 : \Delta(P_0, P_1) = \Delta^*_{p,\eta} \quad \text{against} \quad H_A : \Delta(P_0, P_1) > \Delta^*_{p,\eta}.$$

Given sample of N units, define test statistic and its variance:

$$D = p^{1/2}\|\bar{X}^{T=1}_{1:p} - \bar{X}^{T=0}_{1:p}\|; \quad \sigma^2_*(N) = Var_P(D)$$

Under the null, by bound on sub-Gaussian norms given by Hsu et al,

$$P\left(D > \sqrt{\frac{\sigma^2_*(N)}{p}(p + 2\sqrt{pt} + 2t) + {\Delta^*_{p,\eta}}^2\left(1 + 2\left(t/p\right)^{1/2}\right)}\right) \leq \exp(-t). \tag{5}$$

Back

Test Error Lower Bound (1/3)

Strict overlap $\Rightarrow$ no test can discriminate P_0 and P_1 too well.

Let ϕ be a test mapping statistic $S_\phi(X_{1:p})$ to $\{0,1\}$ for hypotheses

$$H_0 : X_{1:p} \sim P_0 \ (\Leftrightarrow W = 0); \quad H_A : X_{1:p} \sim P_1 \ (\Leftrightarrow W = 1).$$

Test error

$$\begin{aligned}\delta_\phi &:= \max\{\text{size}, 1 - \text{power}\} \\ &:= \max\{P(\phi(S_\phi(X_{1:p})) = 1 \mid W = 0), P(\phi(S_\phi(X_{1:p})) = 0 \mid W = 1)\}.\end{aligned}$$

Theorem (Test error lower bound)

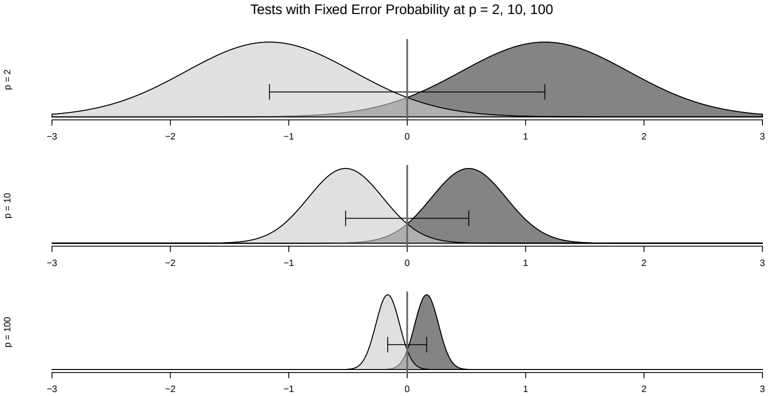
The strict overlap assumption with bound η implies that, for any p, there exists no testing procedure ϕ of $P_0(X_{1:p})$ against $P_1(X_{1:p})$ such that $\delta_\phi < \eta$.

Back

Test Error Lower Bound (2/3)

Test error lower bound fixed for all values p.

Test error is non-increasing in $p \Rightarrow$ tighter constraints on P_0, P_1 for larger p.



Back

Test Error Lower Bound (3/3)

Strict overlap implies that conditions for **consistent tests** of stochastic processes are not satisfied...

Corollary

A test $\phi(S_\phi(X_{1:p}))$ *is* **consistent** *if and only if* $\delta_\phi \to_P 0$ *as p grows large. Asymptotic strict overlap with fixed bound* η *implies that there exists* **no consistent test** *of* P_0 *against* P_1.

... nor are conditions for **consistent estimation** of imbalanced parameters.

Corollary

If P_0 *and* P_1 *differ on a parameter* $\psi(\cdot)$, *asymptotic strict overlap implies that there can exist* **no consistent estimator** *of* $\psi(P_0)$ *or* $\psi(P_1)$ *as p grows large.*

Back

Sub-Exponential Mean Discrepancy Bound (1/2)

Test error bound supports mean discrepancy bound for **multivariate sub-exponential** case.

Theorem (Sub-exponential Mean Distance Bound)

Let $X_{1:p}$ be multivariate sub-exponential with parameters (σ_p^2, b_p) under both P_0 and P_1, as in Gaussian case.

Strict overlap with bound η implies that

$$\|\mu_{0,1:p} - \mu_{1,1:p}\| \leq \begin{cases} \sqrt{8\sigma_p^2 \log \frac{1}{\eta}} & \textit{if } \sigma_p^2/b_p^2 > -2\log\eta \\ 4b_p \log \frac{1}{\eta} & \textit{if } \sigma_p^2/b_p^2 \leq -2\log\eta. \end{cases} \tag{6}$$

Back

Sub-Exponential Mean Discrepancy Bound (2/2)

As in Gaussian case, under some conditions, strict overlap implies **most covariate means are arbitrarily close together** when p is large.

Corollary

In the same setting as Theorem **??**, *strict overlap with bound* η *implies that*

$$\frac{1}{p}\sum_{k=1}^{p}\left|\mu_0^{(k)} - \mu_1^{(k)}\right| \leq \begin{cases} p^{-1/2}\sigma_p\sqrt{8\log\frac{1}{\eta}} & \text{if } \sigma_p^2/b_p^2 > -2\log\eta \\ p^{-1/2}b_p 4\log\frac{1}{\eta} & \text{if } \sigma_p^2/b_p^2 \leq -2\log\eta. \end{cases} \tag{7}$$

Bound goes as $p^{-1/2}$ for the independent, bounded variance case. Converges to zero if $\max\{\sigma_p, b_p\}$ is $o(p^{1/2})$, i.e., effective dimension of $X_{1:p}$ grows with p.

Back

Polar Projection Details

Each point represents a p-dimensional vector; heavy x's represent distribution means.

Coordinates of each point are determined by:

- Normalized distance from mean: $p^{1/2}\|X_{1:p} - \mu_{1:p}\|$.
- Angle (treating mean as origin) in an arbitrary 2-dimensional plane containing the line running between the means of the distributions.

Preserved:

- Distances of points from their mean
- Distance between means of distributions.

Not preserved: Distances between points.

Single Polar Projection | Double Polar Projection | Gaussian Polar Projection

Deconfounding Scores

To maintain identification d must be a **deconfounding score**.

$d(X_{1:p})$ is a **deconfounding score** if and only if ignorability given $X_{1:p}$ implies

$$(Y(0), Y(1)) \perp\!\!\!\perp W \mid d(X_{1:p}).$$

Deconfounding scores include **balancing scores** $b(X_{1:p})$, satisfying:

$$W \perp\!\!\!\perp X_{1:p} \mid b(X_{1:p})$$

and generalized **prognostic scores** $r(X_{1:p})$, satisfying:

$$(Y(0), Y(1)) \perp\!\!\!\perp X_{1:p} \mid r(X_{1:p}).$$

Generalized Identification

Relax overlap requirement with stronger modeling assumptions.

Assumption (Generalized Identification for Estimation)

Given a set of covariates $X_{1:p}$ and a set of functions $\mathcal{D}$ that yield covariate reductions $\mathcal{R}(X_{1:p}) = \{r(X_{1:p}) : d \in \mathcal{D}\}$, for some $d \in \mathcal{D}$

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid r(X_{1:p})$$

and overlap is satisfied for all $r(X_{1:p}) \in \mathcal{D}$.

Key: $\mathcal{D}$ discards information in $X_{1:p}$.

Simpler $\mathcal{D} \Rightarrow$ stronger ignorability assumption, weaker overlap assumption.

$\mathcal{D} := \{$all measureable functions$\}$ recovers standard conditions.

Bias Amplification

If ignorability does not hold, target parameter is not τ^{ATE}

$$\tau_{adj}^{ATE} = \mathbb{E}[\mathbb{E}[Y^{obs} \mid W = 1, X_{1:p}] - \mathbb{E}[Y^{obs} \mid W = 0, X_{1:p}]]$$

Bias has the form

$$\begin{aligned}\tau_{adj}^{ATE} - \tau^{ATE} = \mathbb{E}\Bigg[&(1 - e(X_{1:p}))\frac{Cov(Y(1), W \mid X_{1:p})}{Var(W \mid X_{1:p})} \\ &+ e(X_{1:p})\frac{Cov(Y(0), W \mid X_{1:p})}{Var(W \mid X_{1:p})}\Bigg]\end{aligned}$$

Residual confounding amplified by $Var(W \mid X_{1:p})^{-1} = (e(X_{1:p})(1 - e(X_{1:p})))^{-1}$.

Variance of Conditional Average Treatment Effect

$$\color{blue}{CATE} := \tau(x_i) := \mathbb{E}\Big[D\Big|X = x_i\Big] = \mathbb{E}\Big[Y(t) - Y(c)\Big|X_i = x_i\Big]$$

Decompose the MSE at x_i:

$$\mathbb{E}\left[(D_i - \hat{\tau}_i)^2 | X_i = x_i\right] =$$
$$\underbrace{\color{red}{\mathbb{E}\left[(D_i - \tau(x_i))^2 | X_i = x_i\right]}}_{\color{red}{\text{Approximation Error}}} + \underbrace{\color{blue}{\mathbb{E}\left[(\tau(x_i) - \hat{\tau}_i)^2 | X_i = x_i\right]}}_{\color{blue}{\text{Estimation Error}}}$$

- Since we cannot estimate D_i, we estimate the CATE at x_i
- But the error for the CATE is not the same as the error for the ITE

Supplementary

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 =$$

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 = \sigma^2 = \alpha$$

With one data point?

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 = \sigma^2 = \alpha$$

With one data point?

$$\begin{aligned} E(Y_i - Y_u)^2 &= E(Y_i - \mu + Y_u - \mu)^2 \\ &= E(Y_i - \mu)^2 + E(Y_u - \mu)^2 \\ &= 2\sigma^2 \\ &= 2\alpha \end{aligned}$$

General results for Cover-Hart class, which is a convex cone (Gneiting, 2012)
Back to CATE