Semiparametric theory

Jiafeng Chen

August 8, 2020

Econometrics Reading Group

applied talk

A tour of *Semiparametric Theory and Missing Data* [a]

[a] with a little of Ed Kennedy's tutorial and van der Vaart mixed in

Semiparametric models

Observe data $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P_0$. P_0 belongs to a set of distributions $\mathcal{P} = \{P_{\beta,\eta}\}$. The model $\mathcal{P}$ is semiparametric if, generally, β is finite dimensional and η is infinite dimensional.

Example

Suppose $\beta = \mathbb{E}[Z]$ is the parameter of interest. If $\mathcal{P} = \{\mathcal{N}(\beta, 1) : \beta \in \mathbb{R}\}$, the model is parametric. If $\mathcal{P} = \{P : \mathbb{E}_P[Z^2] < \infty\}$ contains all one dimensional distributions with finite second moment, then the model is semiparametric. In this case, we can treat the nuisance parameter as $\eta = \mathcal{L}(Z - \mathbb{E}[Z])$.

Influence functions, RAL estimators

Assume for now that we have a finite-dimensional model $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P_\theta, \theta = (\beta, \eta)$, $\beta \in \mathbb{R}^q, \eta \in \mathbb{R}^r$. Most "reasonable" estimators $\widehat{\beta}_n$ are asymptotically linear:

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Z_i, \theta_0) + o_{P_{\theta_0}}(1)$$

The function $\varphi(\cdot) = \varphi(\cdot, \theta_0)$ is called the influence function of $\widehat{\beta}_n$.

Proposition

If $\widehat{\beta}_n$ is asymptotically linear, its influence function is a.s. unique.

Influence functions, RAL estimators

Definition (Regular estimators)

Consider the local DGP $P_n = P_{\theta_n}$ indexed by the drifting parameter $\theta_n = \theta_0 + h/\sqrt{n}$. Consider $Z_{1n}, \ldots, Z_{nn} \overset{\text{i.i.d.}}{\sim} P_{\theta_n}$. An estimator is regular if the limiting distribution $\sqrt{n}(\widehat{\beta}_n - \beta_n) \overset{\theta_n}{\rightsquigarrow} L_{\theta_0}$ doesn't depend on h.

Example

$Z_1, \ldots, Z_n \sim \mathcal{N}(\beta, 1)$. The sample mean $\widehat{\beta}_n = \frac{1}{n} \sum_i Z_i$ is RAL (regular and asymptotically linear) with influence function $\varphi(Z, \beta) = Z - \beta$

Structure of influence functions (parametric model)

Let $\theta \in \mathbb{R}^p$ be the parameter and let's say we are interested in $\beta(\theta) \in \mathbb{R}^q$ [slightly more general than $\theta = (\beta, \eta)$].

Pathwise derivative representation of the influence function

Theorem (Tsiatis Theorem 3.2; Newey (1994) expression (2.2); Newey (1990) Theorem 2.2)

Let $\Gamma^{q\times p}(\theta) = \frac{\partial \beta}{\partial \theta'}$. Assume Γ exists, full rank, continuous in a neighborhood of θ_0. Let $\widehat{\beta}_n$ be AL with influence function $\varphi(Z)$ s.t. $\mathbb{E}_\theta[\varphi'\varphi]$ exists, continuous in θ in a neighborhood of θ_0. Let $S_\theta(z,\theta) = \frac{\partial}{\partial\theta} p(z,\theta)$ be the score. Then if $\widehat{\beta}_n$ is regular,

$$\mathbb{E}[\varphi(Z)S_\theta'(Z,\theta_0)] = \Gamma(\theta_0)$$

Corollary

If $\theta = (\beta, \eta)$, let $S_\theta = [S_\beta, S_\eta]$, then

$$\mathbb{E}[\varphi(Z)S_\beta'] = I_q$$
$$\mathbb{E}[\varphi(Z)S_\eta'] = 0^{q\times r}$$

Structure of influence functions (parametric model)

Theorem

Let $\Gamma^{q\times p}(\theta) = \frac{\partial \beta}{\partial \theta'}$. Let $\widehat{\beta}_n$ be AL with influence function $\varphi(Z)$. Let $S_\theta(z,\theta) = \frac{\partial}{\partial \theta} p(z,\theta)$ be the score. Then if $\widehat{\beta}_n$ is regular,

$$\mathbb{E}[\varphi(Z) S'_\theta(Z,\theta_0)] = \Gamma(\theta_0)$$

Sketch.

Consider RAL $\widehat{\beta}_n$. Under local sequence $\theta_0 = \theta_n + h/\sqrt{n}$,

$$\begin{aligned}
\overbrace{\sqrt{n}(\widehat{\beta}_n - \beta(\theta_n))}^{\mathcal{N}(0,\mathbb{E}_{\theta_0}[\varphi\varphi'])} &= \sqrt{n}(\widehat{\beta}_n - \beta(\theta_0)) - \sqrt{n}(\beta(\theta_n) - \beta(\theta_0)) \\
&\simeq_{\theta_n} \underbrace{\frac{1}{\sqrt{n}} \sum_i [\varphi(Z,\theta_0) - \mathbb{E}_{\theta_n}[\varphi(Z,\theta_0)]]}_{\mathcal{N}(0,\mathbb{E}_{\theta_0}[\varphi\varphi'])} + \sqrt{n}\mathbb{E}_{\theta_n}[\varphi(Z,\theta_0)] - \sqrt{n}(\beta(\theta_n) - \beta(\theta_0))
\end{aligned}$$

Equate the latter two terms to zero. Differentiate w.r.t. θ to linearize. Get $\mathbb{E}[\varphi S'_\theta]$ from the first, Γ from the second. $\square$

Tangent spaces, Geometry of IF

Let $\theta = [\beta, \eta]$.

The score $S_\theta(Z, \theta_0)$ (which is mean zero $\mathbb{E}_{\theta_0}[S_\theta(Z, \theta_0)] = 0$). Consider $\mathcal{H}$ the Hilbert space formed by all q-dimensional P_0-mean-zero-finite-variance functions, equipped with the covariance inner product.

Definition

Let the tangent space be $\mathcal{T} = \{BS_\theta(Z, \theta_0) : B \in \mathbb{R}^{q \times p}\} \subset \mathcal{H}$. Let the nuisance tangent space be $\Lambda = \{BS_\eta(Z, \theta_0) : B \in \mathbb{R}^{q \times r}\} \subset \mathcal{T} \subset \mathcal{H}$.

Note that an influence function φ necessarily has $\varphi \perp \Lambda$.

Theorem (Converse to Theorem 3.2)

Let $\varphi(Z)$ be such that $\mathbb{E}[\varphi S'_\beta] = I, \mathbb{E}[\varphi S'_\eta] = 0$. The moment condition $m(Z, \beta, \eta) = \varphi(Z) - \mathbb{E}_{\beta,\eta}[\varphi(Z)]$ defines $\widehat{\beta}_n$ that is RAL with influence function $\varphi(Z)$ so long as $\widehat{\eta}_n$ is $\sqrt{n}$-consistent.

Now we can talk about influence functions without talking about RAL estimators. (This is often confusing!)

Geometry of IF

Theorem

The set of all influence functions is the linear variety $\varphi(Z) + \mathcal{T}^{\perp} = \{\varphi(Z) + \phi(Z) : \phi \in \mathcal{T}^{\perp}\}$ where $\varphi(Z)$ is any influence function.

Definition

The variance-minimizing influence function (efficient influence function) is the projection of any influence function onto the tangent space

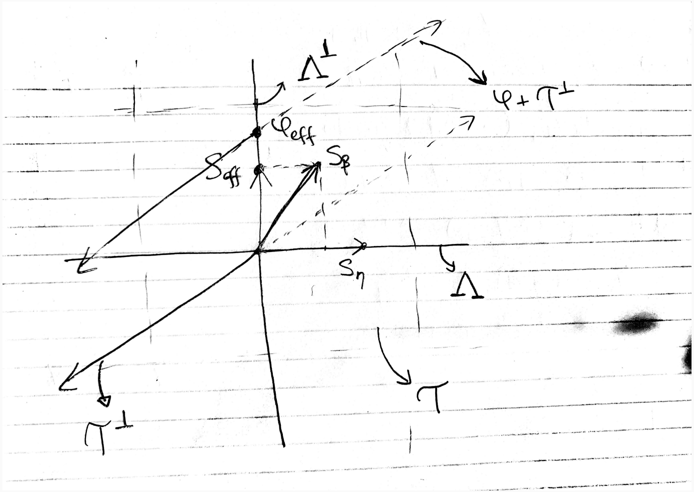
$$\varphi_{\text{eff}}(Z) = \Pi(\varphi(Z)|\mathcal{T}) = \Gamma(\theta_0)\mathit{\Gamma}^{-1}(\theta_0)S_\theta(Z,\theta_0) \overset{\text{subvec}}{=} \mathbb{E}[\varphi S_\theta']\mathbb{E}[S_\theta S_\theta']^{-1}S_\theta = \mathrm{Cov}\cdot\mathrm{Var}^{-1}S_\theta.$$

The efficient score is the residual of the score after projecting onto the nuisance tangent space

$$S_{\text{eff}}(Z,\theta_0) = \Pi(S_\beta|\Lambda^{\perp}) = S_\beta - \mathbb{E}[S_\beta S_\eta']\mathbb{E}[S_\eta S_\eta']^{-1}S_\eta$$

We also have $\varphi_{\text{eff}} = \mathbb{E}[S_{\text{eff}}S_{\text{eff}}']^{-1}S_{\text{eff}}$ under the subvector $\theta = [\beta,\eta]$

Attempt at visualizing tangent spaces and influence functions



Example

Example (Making sure MLEs make sense)

Consider the MLE in $\theta = (\beta, \eta)$. It is known to be efficient. The efficient variance is

$$V^* = \mathbb{E}[S_{\text{eff}} S'_{\text{eff}}]^{-1} = \mathbb{E}[\varphi_{\text{eff}} \varphi'_{\text{eff}}] = [I_\theta^{-1}]_\beta$$

We can calculate the top-left of I_θ^{-1}:

$$V^* = (I_{\beta\beta} - I_{\beta\eta} I_{\eta\eta}^{-1} I'_{\beta\eta})^{-1}.$$

Efficient score is

$$S_{\text{eff}} = S_\beta - I_{\beta\eta} I_{\eta\eta}^{-1} S_\eta$$

and so $\mathbb{E}[S_{\text{eff}} S'_{\text{eff}}] = (I_{\beta\beta} - 2 I_{\beta\eta} I_{\eta\eta}^{-1} I_{\eta\beta} + I_{\beta\eta} I_{\eta\eta}^{-1} I_{\eta\eta} I_{\eta\eta}^{-1} I_{\eta\beta})$

Summary for parametric models

$\theta = [\beta, \eta]$

- Tangent space $\mathcal{T}$ (q-dimensional linear combination of S_θ) lives in the Hilbert space $\mathcal{H}$ of q-dimensional mean-zero finite-variance functions.
- $\mathcal{T} = \Lambda + \mathcal{T}_\beta$ where Λ is the nuisance tangent space
- Influence functions satisfy $\varphi \perp \Lambda, \mathbb{E}[\varphi S_\beta'] = I$. Form linear variety $\varphi + \mathcal{T}^\perp$
- EIF is the projection of any φ on $\mathcal{T}$. The variance of the EIF is the efficiency bound for RAL estimators.
- Efficient score is the projection of S_β on $\Lambda^\perp$ ("S_β that is not explained by S_η")
- EIF is the properly scaled efficient score

Semiparametric models

Now let's make η infinite dimensional. Our model is $\mathcal{P} = \{P_{\beta,\eta} : \beta \in \mathbb{R}^q, \eta \in \Omega\}$. Truth is $P_0 = P_{\beta_0,\eta_0}$

Definition

A parametric submodel $\mathcal{P}_{\beta,\gamma}$ is a parametric model indexed by (β, γ) with γ finite dimensional such that $P_{\beta,\gamma} \in \mathcal{P}$ and $P_0 = P_{\beta_0,\gamma_0}$ for some (β_0, γ_0).

Heuristically, since the semiparametric problem is at least as hard as the parametric problem,

$$V \geq \sup_{\mathcal{P}_{\beta,\gamma}} \left(\mathbb{E} \left[S^{\text{eff}}_{\beta,\gamma} S^{\text{eff}\prime}_{\beta,\gamma} \right] \right)^{-1}$$

Definition (Semiparametric nuisance tangent space)

The semiparametric nuisance tangent space Λ is the mean-square closure of the parametric submodel tangent spaces:

$$\Lambda = \left\{ h \in \mathcal{H} : \exists j = 1, 2, \ldots \text{ s.t. } \lim_{j \to \infty} \left\| h(Z) - B_j^{q \times r_j} S_{\gamma_j}(Z) \right\|^2 = 0 \right\} = \overline{\bigcup_{\mathcal{P}_{\beta,\gamma}} \Lambda_{\beta,\gamma}}$$

Efficient score and the efficiency bound

Definition

Define the efficient score as usual $S_{\text{eff}} = S_\beta - \Pi(S_\beta|\Lambda)$

Theorem

The semiparametric efficiency bound is equal to $\mathbb{E}[S_{\text{eff}} S'_{\text{eff}}]^{-1}$, i.e.

$$\sup_{\mathcal{P}_{\beta,\gamma}} \left(\mathbb{E}\left[S^{\text{eff}}_{\beta,\gamma} S^{\text{eff}\prime}_{\beta,\gamma} \right] \right)^{-1} = \mathbb{E}[S_{\text{eff}} S'_{\text{eff}}]^{-1}$$

Influence functions

Theorem

Any semiparametric RAL estimator for β in $\theta = [\beta, \eta]$ must have an influence function s.t. $\mathbb{E}[\varphi(Z)S'_\beta] = \mathbb{E}[\varphi(Z)S'_{\text{eff}}] = I_q$ and $\Pi(\varphi(Z)|\Lambda) = 0$. The EIF, if it exists, is $\varphi_{\text{eff}} = \mathbb{E}[S_{\text{eff}}S'_{\text{eff}}]^{-1}S_{\text{eff}}$.

If $\beta = \beta(\theta)$, then the more general statement is $\varphi_{\text{eff}} = \Pi(\varphi(Z)|\mathcal{T})$.

Optimal weighting in unconditional GMM

As an example, consider a function $g(z, \beta)$ and the GMM model $Z_i \overset{\text{i.i.d.}}{\sim} P_{\beta,\eta}$ s.t.

$$\mathbb{E}[g(z, \beta_0)] = \int g(z, \beta_0) P_{\beta_0, \eta_0}(z) = 0$$

Consider a parametric submodel indexed by θ. Define $\beta(\theta)$ s.t. $\mathbb{E}_\theta[g(z, \beta(\theta))] = 0$.

For an influence function $\varphi(Z)$, we have that by the pathwise derivative representation

$$\Gamma := \frac{\partial \beta(\theta)}{\partial \theta} = \mathbb{E}[\varphi(Z) S_\theta'].$$

Differentiating $\mathbb{E}_\theta[g(z, \beta(\theta))] = 0$ at θ_0:

$$0 = \frac{\partial}{\partial \theta} \int g p_\theta \, dz = \int \frac{\partial g}{\partial \beta} \frac{\partial \beta}{\partial \theta} p_{\theta_0} \, dz + \int g S_\theta' \, p_{\theta_0} \, dz \implies \mathbb{E}[g S_\theta'] = -\overbrace{\mathbb{E}\left[\frac{\partial g}{\partial \beta}\right]}^{G} \frac{\partial \beta}{\partial \theta}.$$

We know then influence functions look like $-(AG)^{-1} A g$ for a conformable full rank A. Conclude that optimal weighting GMM ($A = G'\Omega^{-1}$) is efficient in the class of RAL estimators in this model.

Side Q: Is there a characterization of the tangent space in the nonlinear GMM model?

Constructing efficient estimators

Assume we know the efficient influence function φ_{eff} or the efficient score S_{eff}.

Natural idea, solve for β in the efficient score equations

$$0 = \frac{1}{n} \sum_{i=1}^{n} S_{\text{eff}}(Z_i, \beta, \widehat{\eta}_n).$$

Here $\widehat{\eta}_n$ is an estimator for η_0, could also concentrate out η and let $\widehat{\eta}_n(\beta)$ solve the score equations for β.

Theorem (van der Vaart Theorem 25.54)

Suppose $\mathbb{E}_{\widehat{\beta}_n, \eta_0}[S_{\text{eff}}(Z, \widehat{\beta}_n, \widehat{\eta}_n)] = o_p(1/\sqrt{n} + ||\widehat{\beta}_n - \beta_0||)$ and $\mathbb{E}_{\beta_0, \eta_0}||\widehat{S}_{\text{eff}} - S_{\text{eff}}||^2 = o_p(1)$. Assume there is a Donsker class that contains $S_{\text{eff}}(\cdot, \tilde{\beta}, \tilde{\eta})$. Under additional regularity conditions, $\widehat{\beta}_n$ is efficient.

Remark

Efficient score equation does not necessarily deliver efficient estimators (if $\widehat{\eta}_n$ is bad). Non-efficient score equations can deliver efficient estimators (if $\widehat{\eta}_n$ is the right amount of bad).

Constructing efficient estimators

In the simpler case $S_{\text{eff}}(Z_i, \beta, \eta) \propto m(Z_i, \eta) - \beta$, we would set naturally that $\widehat{\beta}_n = \frac{1}{n} \sum_{i=1}^n m(Z_i, \widehat{\eta}_n) = \mathbb{P}_n m(\cdot, \widehat{\eta}_n)$.

$$
\begin{aligned}
\widehat{\beta}_n - \beta &= \mathbb{P}_n m(\cdot, \widehat{\eta}_n) - \mathbb{P} m(\cdot, \eta) \\
&= \mathbb{P}_n[m(\cdot, \widehat{\eta}_n) - m(\cdot, \eta)] + \underbrace{[\mathbb{P}_n - \mathbb{P}] m(\cdot, \eta)}_{\xrightarrow{d} \mathcal{N}(0, V)/\sqrt{n}} \\
&= [\mathbb{P}_n - \mathbb{P}][m(\cdot, \widehat{\eta}_n) - m(\cdot, \eta)] + \mathbb{P}[m(\cdot, \widehat{\eta}_n) - m(\cdot, \eta)] + [\mathbb{P}_n - \mathbb{P}] m(\cdot, \eta)
\end{aligned}
$$

If $\mathbb{P}[m(\cdot, \widehat{\eta}_n) - m(\cdot, \eta)]^2 = o_p(1)$ and $\{m(\cdot, \eta) : \eta\}$ is a Donsker class, then the first term is $o_p(1/\sqrt{n})$. If we are lucky, the second term is also $o_p(1/\sqrt{n})$.

If $m(\cdot, \cdot) - \beta$ is not the efficient IF/score, but the second term doesn't vanish, it is still possible for $\widehat{\beta}_n$ to be efficient (All non-doubly-robust ATE estimators have this feature)

Constructing efficient estimators ("synthesis")

A general strategy is to study the pathwise derivative representation of IFs $\frac{\partial \beta}{\partial \theta} = \mathbb{E}[\varphi S_\theta']$. Consider the von Mises expansion

$$\beta(Q) - \beta(P) = \int \varphi(Q)\, d(Q - P) + R_2(Q, P) = -\int \varphi(Q)\, dP + R_2(Q, P).$$

A plug-in estimator $\beta(\widehat{P})$ will have bias approximately $-\mathbb{E}_P[\varphi(\widehat{P})]$.

Natural idea (one-step correction, closely connected to efficient score equations)

$$\widehat{\beta}_n := \beta(\widehat{P}) + \mathbb{P}_n\varphi(\widehat{P}) = \beta + [\mathbb{P}_n - \mathbb{P}][\varphi(\widehat{P}) - \varphi(P)] + R_2(\widehat{P}, P) + (\mathbb{P}_n - \mathbb{P})\varphi(P)$$

Classical paradigm: use empirical process theory to argue that first remainder term is small, check that second remainder is small.

Double/debiased machine learning: sometimes give up efficiency, ensure that "$\frac{\partial \varphi}{\partial \eta} = 0$" (Neyman orthogonality). Use orthogonality and sample splitting to kill the first and second remainder.

Another natural idea (targeted maximum likelihood, van der Laan): construct $\widehat{P}^*$ such that $\beta(\widehat{P}^*) \approx \beta(\widehat{P}) + \mathbb{P}_n\varphi(\widehat{P})$.

Semiparametrics in causal inference

~~Semiparametrics~~ in causal inference

Machine learning in causal inference

Semiparametric models and efficiency

A semiparametric model is a set of distributions $\mathcal{P} = \{P_{\beta,\eta} : \beta \in \mathbb{R}^q, \eta \in H\}$ indexed by (β, η) where η is infinite dimensional (e.g. a function, a distribution, etc.). Observe data $Z_i \sim P_{\beta_0,\eta_0}$

Most models in economics are semiparametric!

- ✓ Interested in a finite-dimensional parameter (a treatment effect, an elasticity, a marginal effect)
- ✓ Cannot write down a parametric likelihood (GMM, linear regression in non-Gaussian models) [if we were willing to we would MLE/parametric Bayes everything]

Most nice estimators for β are asymptotically linear

$$\sqrt{n}(\widehat{\beta}_n - \beta) \simeq \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(Z_i) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\varphi\varphi'])$$

and regular (“smooth” in the data)

Efficiency bound

If $\widehat{\beta}_n$ is a nice estimator, it is asymptotically normal with variance V. The Cramer–Rao lower bound says in a parametric model,

$$V \succeq V_{CRLB} = I_\beta^{-1}$$

Every semiparametric model contains lots of parametric submodels $\mathcal{P}_{\beta,\gamma} = \{P_{\beta,\gamma} : \gamma \in \mathbb{R}^r\}$. We ought to have for every parametric submodel,

$$V \succeq V_{\beta,\gamma}^{CRLB}$$

since the semiparametric problem is harder than the parametric problem. The semiparametric efficiency bound is

$$V^* = \sup_{\mathcal{P}_{\beta,\gamma}} V_{\beta,\gamma}^{CRLB} \preceq V$$

For a given problem, what is the bound? Is it achievable? How to achieve it?

Brief review of identification

Consider binary treatment $W_i \in \{0, 1\}$, potential outcomes $Y_i(0), Y_i(1)$, covariates $X_i \in \mathbb{R}^p$. Assume

1. Strong ignorability / Selection-on-observables $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i$
2. (SUTVA/“consistency”) $Y_i = Y_i^{obs} = W_i Y_i(1) + (1 - W_i) Y_i(0)$
3. Overlap $0 < \epsilon \leq e(X_i) \leq 1 - \epsilon < 1$, where $e(X_i) = \mathbb{P}(W_i = 1 | X_i)$ is the propensity score.

Let $\mu_w(X_i) = \mathbb{E}[Y_i(w) \mid X_i] = \mathbb{E}[Y_i \mid W_i = w, X_i]$ by strong ignorability.

Then the ATE

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)] = \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)}\right]$$

Motivates two plug-in estimators.

We have so far not assumed anything about the functional form of $\mathbb{E}[Y_i(w) \mid X_i]$ or $e(X_i)$.

Desiderata

- Assume the identification problem is solved and we buy all the assumptions
  - “This has made a lot of people very angry and been widely regarded as a bad move”
- Assume i.i.d. sampling from an infinite superpopulation
- Propensity score $e(x)$ may be known or unknown. If we designed an experiment, then it is known
- Generalizes to discrete W_i easily, with continuous W_i things are much more involved.

Nonparametric estimation when everything is discrete

When X_i has finite support, WLOG $X_i \in \{1, \dots, K\}$, everything is easy and we just tabulate:

$$\widehat{\mu}_w(k) = \frac{1}{n_{w,k}} \sum_{i:X_i=k,\, W_i=w} Y_i \qquad \widehat{e}(k) = \frac{1}{n_k} \sum_{i:X_i=k} W_i$$

Natural estimators "outcome modeling / imputation" and inverse propensity score weighting:

$$\widehat{\tau}_{OM} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) \qquad \widehat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{W_i Y_i}{\widehat{e}(X_i)} - \frac{(1-W_i)\, Y_i}{1-\widehat{e}(X_i)}$$

Saturated demeaned regression: coefficient on W in

```
lm(demean_y ~ 1 + w + demean_factor(x) + w * demean_factor(x))
```

Careful: Neither `lm(y ~ 1 + w + x)` nor `lm(y ~ 1 + w + x + w * x)` works!

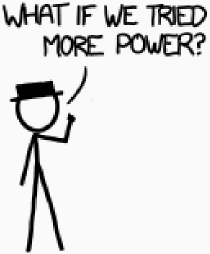
Proposition (One estimator to rule them all)

$\widehat{\tau}_{OM} = \widehat{\tau}_{IPW} = \widehat{\tau}_{REG}$ are numerically equivalent. They are all efficient.

$\widehat{\tau}_{OM}$ is the MLE under normality $(Y_i(w) \mid X_i = k \sim \mathcal{N}(\mu_w(k), \sigma_w^2(k)))$.

$$V_{OM} \geq V_{\text{semiparametric bound}} \geq V_{\text{parametric bound}} = V_{OM}$$

That was easy



https://what-if.xkcd.com/13/

Things are hard if X_i is continuous

- Clearly, the "but I controlled for X"-regression `lm(y ~ w + x + w * x)` is not consistent (worse still, `lm(y ~ w + x)`)
  - Not consistent even if Y, X are already demeaned.
  - Still consistent if we assume complete randomization: $(Y(1), Y(0)) \perp\!\!\!\perp W$
- Outcome modeling with nonparametric estimator $\widehat{\mu}_w(x) \approx \mathbb{E}[Y_i \mid W_i = w, X_i = x]$ and form

$$\widehat{\tau}_{OM} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_1(X_i) - \mu_0(X_i) + \text{Approximation Error}$$

  would be consistent, but it's tricky to establish normality or analytic standard errors.
- It is easy to see that the approximation error has terms like

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(X_i) - \mu_1(X_i) = \overbrace{\mathbb{E}_X[\widehat{\mu}_1(X) - \mu(X)]}^{???} + \overbrace{\left(\frac{1}{n}\sum_{i=1}^{n} (\widehat{\mu}_1 - \mu_1) - \mathbb{E}_X(\widehat{\mu}_1 - \mu_1)\right)}^{\text{hopefully small } (o(1/\sqrt{n}))}$$

- The integrated RMSE $\left(\mathbb{E}_X(\widehat{\mu}_1 - \mu_1)^2\right)^{1/2} \gtrsim n^{-s/(2s+d)} > 1/\sqrt{n}$, so a naive bound would fail

Efficiency of ATE estimators

Theorem (Hahn, 1998)

The semiparametric efficiency bound for the ATE is

$$V^* = \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + [\mu_1(X_i) - \mu_0(X_i) - \tau]^2\right]$$

which is not affected by knowledge of the propensity score.

If we *knew* the propensity score, and form the oracle IPW estimator

$$\widehat{\tau}^*_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}$$

we would find that its variance $V^*_{IPW} > V^*$.

An easy way to remember the the bound is that it is the variance of the oracle augmented IPW (AIPW) estimator

$$\widehat{\tau}^*_{AIPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1-W_i)(Y_i - \mu_0(X_i))}{1-e(X_i)} + \mu_1(X_i) - \mu_0(X_i)$$

Efficient estimators

Theorem (Imbens, Newey, Ridder (2007); Hahn (1998); Hirano, Imbens, Ridder (2003) ...)

Under various tuning parameter choices for estimating μ_w, e with series, the following estimators are first-order equivalent (differ by $o_p(1/\sqrt{n})$) and efficient

$$\widehat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{W_i}{\widehat{e}(X_i)} - \frac{1-W_i}{1-\widehat{e}(X_i)} \right) \qquad \text{(also self-normalized version)}$$

$$\widehat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{W_i(Y_i-\widehat{\mu}_1)}{\widehat{e}(X_i)} - \frac{(1-W_i)(Y_i-\widehat{\mu}_0)}{1-\widehat{e}(X_i)} \right) + \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)$$

$$\widehat{\tau}_{OM} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)$$

$$\widehat{\tau}_{Hahn} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)$$

where $\tilde{\mu}_1(X_i) = \frac{\widehat{\mathbb{E}}[Y_i W_i | X = X_i]}{\widehat{e}(X_i)}$

Under the hood

We can write the ATE problem as a moment problem:

$$\mathbb{E}[m(X_i, W_i, Y_i; \mu_0, \mu_1, e) - \tau] = 0 \quad \text{rewrite to} \quad \mathbb{E}[g(Z_i, \tau, \eta)] = 0.$$

Our ATE estimates set $\frac{1}{n}\sum_{i=1}^{n} g(Z_i, \tau, \widehat{\eta}) = 0$. If we go through the two-step estimation argument, we would find that

$$\widehat{\tau} \simeq \frac{1}{n}\sum_{i=1}^{n} -G^{-1}\left[g(Z_i, \beta_0, \eta_0) + \alpha(Z_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[m(X_i, W_i, Y_i; \mu_0, \mu_1, e) + \alpha(Z_i) - \tau\right]$$

where $\alpha(Z_i)$ is an adjustment term that accounts for the first-step estimation of η and $G = \frac{\partial}{\partial \tau}\mathbb{E}[g(Z_i, \tau, \eta_0)] = -1$.

Newey (1994) makes this rigorous.

AIPW / Doubly-robust estimation

The AIPW representation is $\tau = \mathbb{E}[\phi(Z)]$ where

$$\phi(Z, \mu_1, \mu_0, e) = \mu_1(X) - \mu_0(X) + \frac{W(Y - \mu_1(X))}{e(X)} - \frac{(1 - W)(Y - \mu_0(X))}{1 - e(X)}$$

$$= \frac{WY}{e(X)} - \frac{(1 - W)Y}{1 - e(X)} - \frac{\mu_1(W - e(X))}{e(X)} + \frac{\mu_0(W - e(X))}{1 - e(X)}$$

Using the “wrong” moment conditions gets an adjustment term that is just right!

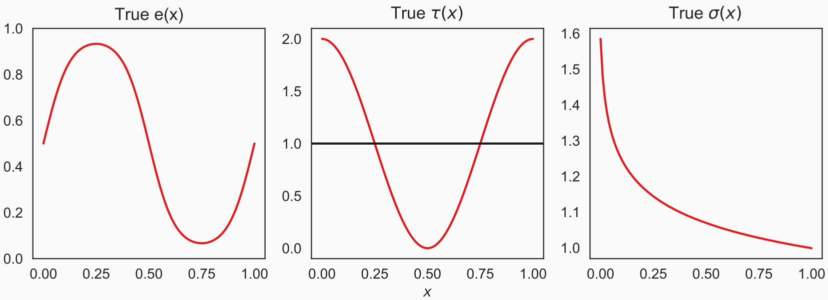
In the IPW case, the adjustment term correlates with the oracle IPW term that decreases overall variance

Remark (Editorializing)

People often motivate doubly-robust / AIPW as having two shots at getting things right…

The doubly-robust form is a natural representation of the ATE

A Monte Carlo



Scalar $X \sim \text{Unif.}$

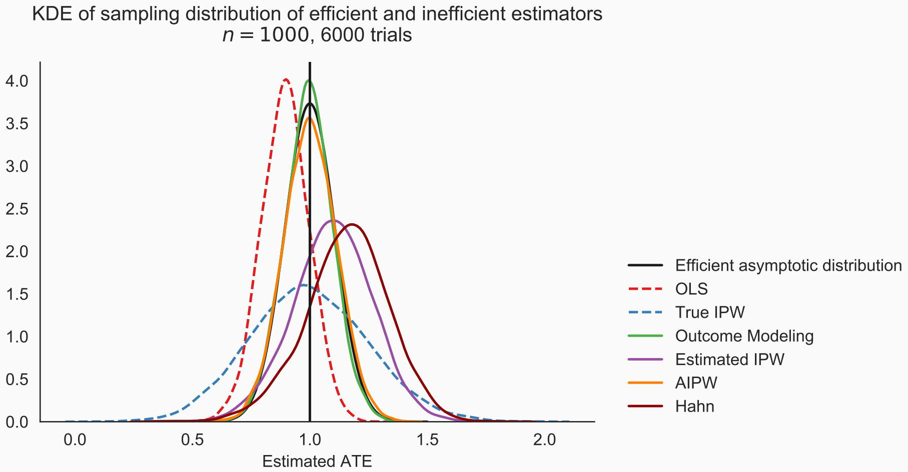
$\mathbb{E}[Y(1) - Y(0)|X] = \tau(X).$

$\mathbb{V}(Y(w)|X) = \sigma^2(X)$

$\tau = 1$

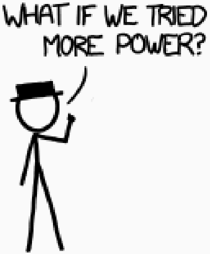
$e(x) \in [0.05, 0.95]$

A Monte Carlo



Intuition: OM does better when $\mu_w(x)$ is "smooth." IPW does better when $e(x)$ is "smooth." AIPW adapts.

That was not as easy



https://what-if.xkcd.com/13/

I promised you machine learning

- Observation: fitting $\mu_1, \mu_0, e(\cdot)$ are essentially prediction problems
- Machine learning is (reductively) a new generation of nonparametric estimators that seem to do well in some empirical applications
- Series estimators perform poorly when $\dim(X) \geq 20$, but something like random forest might work quite well
- Problem: ML methods are very black box, hard to analyze its theoretical properties; classical tools break down
- Proposed solution (more-or-less DML): Have algorithms that work relying only on "weak,"high-level conditions on prediction quality

$$\text{“} \int (\widehat{\mu} - \mu)^2 \, dP(x) = o_p(n^{-1/4}) \text{”}$$

  Use "orthogonality" and sample-splitting to kill terms
- Warning: overlap becomes a very strong condition in high dimensions

DML-AIPW

```
data1, data2 = randomly_split_data_in_half(data)

# Train predictor on one half of the data
mu1_hat_1 = machine_learn("y ~ x", data=data2.where(w == 1))
mu0_hat_1 = machine_learn("y ~ x", data=data2.where(w == 0))
e_hat_1 = machine_learn("w ~ x", data=data2)

mu1_hat_2 = machine_learn("y ~ x", data=data1.where(w == 1))
mu0_hat_2 = machine_learn("y ~ x", data=data1.where(w == 0))
e_hat_2 = machine_learn("w ~ x", data=data1)

# Compute fitted values on the other half
e_hat = [e_hat_1(data1), e_hat2(data2)]
mu1_hat = [mu1_hat_1(data1), mu1_hat_2(data2)]
mu0_hat = [mu0_hat_1(data1), mu0_hat_2(data2)]
```

$$\widehat{\tau}_{DMLAIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i(Y_i - \widehat{\mu}_{1i})}{\widehat{e}_i} - \frac{(1 - W_i)(Y_i - \widehat{\mu}_{0i})}{1 - \widehat{e}_i}\right) + \widehat{\mu}_{1i} - \widehat{\mu}_{0i}$$

Theoretical properties of DML-AIPW

Theorem (Stefan Wager's Stat 361; Chernozhukov et al., 2018)

Let $\widehat{\tau}^*$ be the oracle AIPW estimator, which we know is efficient. Let $\widehat{\mu}_w, \widehat{e}$ be the machine learning output (which are random). Assume

1. Overlap
2. Uniform consistency

$$\sup_x |\widehat{\mu}_w(x) - \mu_w(x)|, \quad \sup_x |\widehat{e}(x) - e(x)| \xrightarrow{p} 0$$

3. Risk decay (more-or-less checkable!)

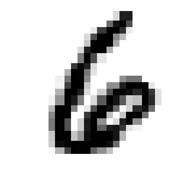
$$\mathbb{E}\left[(\widehat{\mu}_w(x) - \mu_w(x))^2\right] \mathbb{E}\left[(\widehat{e}(x) - e(x))^2\right] = o_p(1/n)$$

Then

$$\sqrt{n}\left(\widehat{\tau}_{DMLAIPW} - \widehat{\tau}^*\right) \xrightarrow{p} 0$$

Monte Carlo, but more power

$Y(1),\ Y(0) \perp\!\!\!\perp W \mid X$, but X can be an image

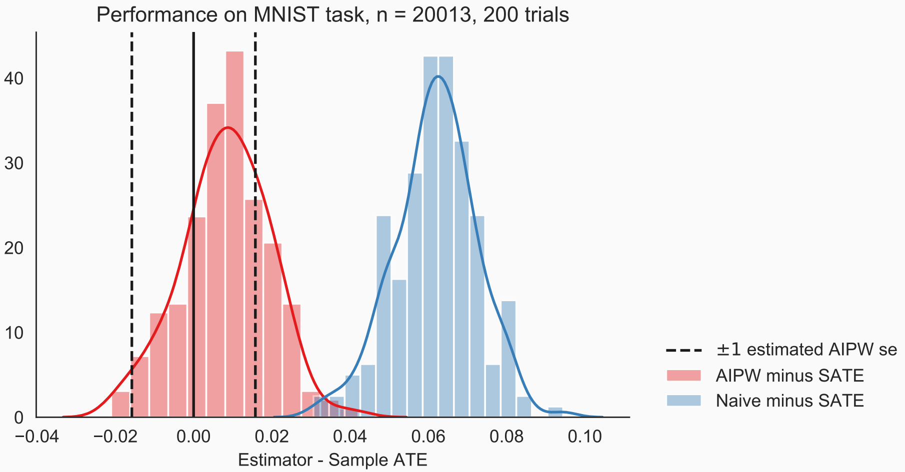


Represented by $[0,1]^{28\times 28} = [0,1]^{784}$

Say $e(X) =$ (digit that X represents) $\times\, 0.1 +$ (mean pixel color)

Only take $4, 5, 6$ to make things simple

Ta-da



Machine learning via $784 \times 20 \times 1$ ReLU networks + tuning. Training takes 3 seconds on my laptop.