Machine Learning and Causal Inference: Applications to Advertising Effectiveness

Professor Susan Athey
Stanford University

August 2017

See papers on arxiv.org: Athey and Imbens "Recursive Partitioning" (PNAS, 2016); Wager and Athey "Random Forests" (JASA, forthcoming); Athey, Imbens, and Wager "Residual Balancing" (2016); Athey, Tibshirani, and Wager "Generalized Random Forests" (2016); Athey and Wager "Optimal Policy" (2017)

ML for Policy and Scientific Understanding

Causal inference for policy and scientific understanding

- Predict hotel prices versus estimate impact of price change
- Identify units at risk versus those with high treatment effect
- See Athey (*Science*, 2017)

Machine learning helps us build more granular statistical models.

- Improves performance of traditional causal inference estimation methods
  - Better use of observables to control for confounders
  - More granual counterfactual predictions about what would have happened to individuals in the absence of the treatment
- Better understanding of how and why a policy works, for continued improvement
- Personalized evaluation of how a policy works
- Personalized policy assignment rules

Correlation v. Causality in Advertising Measurement

Did the advertisement *cause* the sale, or would the sale have happened anyway?

- e.g. Blake, Nosko, &Tadelis, 2016: eBay's search campaigns had -63% return rather than a $+1500\%$ return
- Best solution: A/B tests or bandits
- But what if you don't have enough data to do A/B tests, or have difficulty implementing them properly?
- What is the best we can do with observational data?

A Practical Solution:

- Use best-available observational methods
- Validate observational methods by comparing to experiments
- Use *supplementary analysis* to validate approaches
- Two examples today: techniques based on unconfoundedness, and instrumental variables

Improving Existing Approaches

Approaches to drawing causal inference in observational studies are well-established, but suffer from non-robustness

- Results sensitive to functional form
- Argue about validity based on reasonableness

Recent developments

- Use ML techniques to have data-driven approach to select functional form, modifying techniques to optimize for causal inference v. prediction
  - Avoid "regularization-induced bias"
  - Personalized estimates
  - Centered confidence intervals, asymptotic normality, efficiency
- Validate using many small natural experiments, comparison to A/B tests, evaluation of policy implementation
- *Supplementary Analysis*: Systematic approaches to assessing validity (see, e.g., Athey and Imbens (2015, 2017); Athey, Imbens, Pham, Wager (2017)

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Holland, 1986, Imbens and Rubin, 2015, Rosenbaum and Rubin, 1983, Rubin, 1974), we posit the existence of quantities $Y_i^{(0)}$ and $Y_i^{(1)}$.

- These correspond to the response we **would have measured** given that the i-th subject received treatment $(W_i = 1)$ or no treatment $(W_i = 0)$.
- **NB:** We only get to see $Y_i = Y_i^{(W_i)}$

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

- Define the **average treatment effect (ATE)**, the **average treatment effect on the treated (ATT)**

$$\tau = \tau^{\text{ATE}} = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]; \tau^{\text{ATT}} = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,\middle|\, W_i = 1\right];$$

- and, the **conditional average treatment effect (CATE)**

$$\tau\left(x\right) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,\middle|\, X = x\right].$$

The potential outcomes framework

My stomach hurts
Maybe it was the bad tuna I had for lunch
Too bad I don't have a proper control group, now I'll never know
freshspectrum.com

The potential outcomes framework

If we make no further assumptions, it is not possible to estimate ATE, ATT, CATE, and related quantities.

- This is a failure of identification (infinite sample size), not a small sample issue. Unobserved confounders correlated with both the treatment and the outcome make it impossible to separate correlation from causality.
- One way out is to assume that we have measured enough features to achieve **unconfoundedness** (Rosenbaum and Rubin, 1983)

$$\left\{Y_i^{(0)},\, Y_i^{(1)}\right\} \perp\!\!\!\perp W_i \;\;\big|\;\; X_i.$$

- When this assumption + OVERLAP ($e(x) \in (0,1)$) holds, causal effects are identified and can be estimated.

Estimation Methods

The following methods are efficient when the number of covariates is fixed:

- Propensity score weighting: compare treated and control outcomes, weighted by inverse of probability of being treated
- "Direct" model of the outcome (model of $\mathbb{E}\left[Y_i \,\middle|\, X_i, W_i\right]$), e.g. using regression
- Propensity-score weighted regression of Y on X, W (doubly robust)

The choice among these methods is widely studied:

- Other popular methods include matching, propensity score matching, propensity score blocking, which are not efficient but often do better in practice.
- Note: Hirano, Imbens, Ridder (2003) establish that more efficient to weight by estimated propensity score than actual.
- Which one is better? It depends.

Advertising Effectiveness: Specifics

What is the problem you are trying to solve for separating correlation from causality?

- Users saw the ad because of an action they took indicating interest in the product
- Users saw the ad becaues they were active on the internet that day, and they are more likely to spend money when active ("Activity Bias", e.g. Lewis et al)

What do you need to do to solve this problem?

- Adjust for confounders – compare two individuals for whom the ad assignment was as good as random (allocation due to budget constraints, randomness in ad assignment)
- Observables that indicate user interest, broadly and at this moment
- In online setting, set of websites you have visited recently is related to user interest as well as cookies for targeting
- Problem: there are lots of websites to potentially visit!

Regression Case

Suppose that conditional mean function is given by

$$\mu(w, x) = \beta^{(w)} \cdot x.$$

If we estimate using OLS, then we can estimate the ATT as

$$\widehat{\mathrm{ATT}} = \bar{Y}^{(1)} - \bar{X}^{(1)} \cdot \hat{\beta}^{(0)}$$

Note that OLS is unbiased and efficient, so the above quantity converges to the true values at rate $\sqrt{n}$:

$$\bar{X} \cdot \hat{\beta}^{(0)} - \mu_x^{(1)} \cdot \beta^{(0)} = O_p\Big(\frac{1}{\sqrt{n}}\Big)$$

High-Dimensional Analogs??

Obvious possibility: substitute in the lasso (or ridge, or elastic net) for OLS. But bias is a big problem.

With lasso, for each component j:

$$\hat{\beta}_j^{(w)} - \beta_j^{(w)} = O_p\Big(\sqrt{\frac{\log(p)}{n}}\Big)$$

This adds up across all dimensions, so that we can only guarantee for the ATT:

$$\widehat{\text{ATT}} - \text{ATT} = O_p\Big(\sqrt{\frac{\log(p)}{n}} \|\bar{X}_1 - \bar{X}_0\|_\infty \cdot \|\beta^{(0)}\|_0\Big)$$

Improving the Properties of ATE Estimation in High Dimensions: A "Double-Selection" Method

Belloni, Chernozukov, and Hansen (2013) observe that confounders might be important if they have a large effect on outcomes OR a large effect on treatment assignment. Propose:

- Run LASSO of W on X. Select variables with non-zero coefficients at a selected λ (e.g. cross-validation).
- Run a LASSO of Y on X. Select variables with non-zero coefficients at a selected λ (may be different than first λ).
- Run a OLS of Y on W and the union of selected variables. (Not as good at purely predicting Y as using only second set.)

Result: under "approximate sparsity" of BOTH propensity and outcome models, and constant treatment effects, estimated ATE is asymptotically normal and estimation is efficient.

Doubly Robust Methods

With small data, a “doubly robust” estimator (though not the typical one, where typically people use inverse propensity score weighted regression) is (with $\hat{\gamma}_i = \frac{1}{\hat{e}(X_i)}$):

$$\hat{\mu}_1^0 = \bar{X}_1 \cdot \hat{\beta}^{(0)} + \hat{\mathbb{E}}_{i:W_i=0} \hat{\gamma}_i \big(Y_i - X_i \hat{\beta}^{(0)} \big)$$

To see why, note that the term in parentheses goes to 0 if we estimate $\beta^{(0)}$ well, while to show that we get the right answer if we estimate the propensity score well, we rearrange the expression to be

$$\hat{\mu}_1^0 = \big(\bar{X}_1 - \hat{\mathbb{E}}_{i:W_i=0} (\hat{\gamma}_i X_i) \big) \hat{\beta}^{(0)} + \hat{\mathbb{E}}_{i:W_i=0} \hat{\gamma}_i Y_i$$

The first term has expectation 0, and the second term gives the relevant counterfactual, if the propensity score is well-estimated.

Doubly Robust Methods: A High-Dimensional Analog?

$$\hat{\mu}_1^0 = \bar{X}_1 \cdot \hat{\beta}^{(0)} + \hat{\mathbb{E}}_{i:W_i=0}\hat{\gamma}_i\big(Y_i - X_i\hat{\beta}^{(0)}\big)$$

How does this relate to the truth?

$$\hat{\mu}_1^0 - \mu_1^0 = \bar{X}_1 \cdot (\hat{\beta}^{(0)} - \beta^{(0)}) + \hat{\mathbb{E}}_{i:W_i=0}\hat{\gamma}_i\big(\epsilon_i + X_i\beta^{(0)} - X_i\hat{\beta}^{(0)}\big)$$

$$= \big(\bar{X}_1 - \hat{\gamma}'\bar{X}_0\big) \cdot (\hat{\beta}^{(0)} - \beta^{(0)}) + \hat{\mathbb{E}}_{i:W_i=0}\hat{\gamma}_i\epsilon_i$$

With high dimensions, we could try to estimate $\hat{\beta}$ and the propensity score with LASSO or post-LASSO rather than OLS. However, this may not be good enough. It is also not clear how to get good estimates of the inverse propensity score weights γ_i, in particular if we don't want to assume that the propensity model is sparse (e.g. if the treatment assignment is a complicated function of confounders).

An Efficient Approach with Non-Sparse Propensity

The solution proposed in Athey, Imbens and Wager (2016) for attacking the gap

$$\hat{\mu}_1^0 - \mu_1^0 = \left(\bar{X}_1 - \hat{\gamma}'\bar{X}_0\right) \cdot (\hat{\beta}^{(0)} - \beta^{(0)}) + \hat{\mathbb{E}}_{i:W_i=0}\hat{\gamma}_i\epsilon_i$$

is to bound 1st term by selecting γ_i's using brute force. In particular:

$$\hat{\gamma} = \text{argmin}_{\gamma}\zeta \cdot \|\bar{X}_1 - \gamma'\bar{X}_0\|_\infty + (1-\zeta)\|\gamma\|_2^2$$

The parameter ζ is a tuning parameter; the paper shows that ζ exists such that the γ's exist to tightly bound the first term above. With overlap, we can make $\|\bar{X}_1 - \gamma'\bar{X}_0\|_\infty$ be $O(\sqrt{\frac{log(p)}{n}})$.
Result: If the outcome model is sparse, estimate β using LASSO yielding bias of second term $O_p\left(k\sqrt{\frac{\log(p)}{n}}\right)$, so the bias term is $O(k\frac{log(p)}{n})$, so for k small enough, the last term involving $\hat{\gamma}_i\epsilon_i$ dominates, and ATE estimator is $O(\frac{1}{\sqrt{n}})$.

Summarizing the Approximate Residual Balancing Method of Athey, Imbens, Wager (2016)

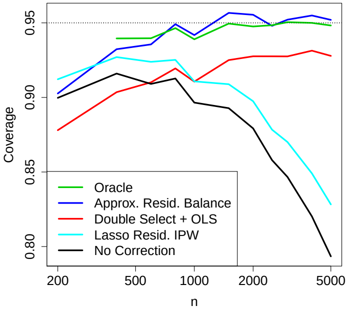
- Estimate lasso (or elastic net) of Y on X in control group.
- Find “approximately balancing” weights that make the control group look like the treatment group in terms of covariates, while attending to the sum of squares of the weights. With many covariates, balance is not exact.
- Adjust the lasso prediction of the counterfactual outcome for the treatment group (if it had been control) using approximately balancing weights to take a weighted average of the residuals from the lasso model.

Main result: if the model relating outcomes to covariates is linear and sparse, and there is overlap, then this procedure achieves the semi-parametric efficiency bound. No other method is known to do this for non-sparse propensity models.
Simulations show that it performs much better than alternatives when propensity is not sparse.

Estimating the Effect of a Welfare-to-Work Program

Data from the California GAIN Program, as in Hotz et al. (2006).

- Program separately randomized in: Riverside, Alameda, Los Angeles, San Diego.
- Outcome: mean earnings over next 3 years.
- We hide county information. Seek to compensate with $p = 93$ controls.
- Full dataset has $n = 19170$.



Supplementary Analysis

To establish validity of models:

- Show multiple methods based on different types of assumptions
- Assess how challenging the problem is:
  - Illustrate overlap through plotting propensity scores for each group
  - Plot the “bias function” (Athey, Imbens, Pham and Wager, 2017)
- “Placebo tests”
- Others depending on setting; see Athey and Imbens (2017)

ML Methods for Causal Inference: Treatment Effect Heterogeneity

- ML methods perform well in practice, but many do not have well established statistical properties
- Unlike prediction, ground truth for causal parameters not directly observed
- Need valid confidence intervals for many applications (AB testing, drug trials); challenges include adaptive model selection and multiple testing
- Different possible questions of interest, e.g.:
  - Identifying subgroups (Athey and Imbens, 2016)
  - Testing for heterogeneity across all covariates (List, Shaikh, and Xu, 2016)
  - Robustness to model specification (Athey and Imbens, 2015)
  - **Personalized estimates** e.g. Wager and Athey, forthcoming; Athey, Tibshirani, and Wager, 2017; Taddy et al 2014
  - **Personalized policy estimation**: ML literature on contextual bandits (e.g. John Langford et al)); Joachins et al; Athey and Wager, 2017

ML Methods for Causal Inference: More general models

- Much recent literature bringing ML methods to causal inference focuses on single binary treatment in environment with unconfoundedness
- Economic models often have more complex estimation approaches
- Athey, Tibshirani and Wager (2016): tackle general GMM case, and establish asymptotic normality
  - Quantile regression
  - Consumer choice
  - Euler equations

  Instrumental Variables
  Panel Regression
  Survival Analysis
- See also "Deep IV" (Taddy, Lewis, Hartford, and Leyton-Brown, 2017)

Heterogeneous Treatment Effects: Related Literature

- Zeilis et al (2008): model trees.
- Imai and Ratkovic (2013) analyze treatment effect heterogeneity with LASSO
- Targeted ML (van der Laan, 2006) can be used as a semi-parametric approach to estimating treatment effect heterogeneity
- Literature on policy estimation, policy learning, and contextual bandits: ML: Langford et al, Swaminathan and Joachims; econometrics: Kitagawa and Tetenov (2015)
- See Wager and Athey (JASA, forthcoming) and Athey and Imbens (PNAS, 2016) for add'l references on treatment effect heterogeneity; see Athey and Wager (2017) for additional references on policy estimation and evaluation

Baseline method: *k*-NN matching

Consider the *k*-**NN matching** estimator for $\tau(x)$:

$$\hat{\tau}(x) = \frac{1}{k} \sum_{\mathcal{S}_1(x)} Y_i - \frac{1}{k} \sum_{\mathcal{S}_0(x)} Y_i,$$

where $\mathcal{S}_{0/1}(x)$ is the set of *k*-nearest cases/controls to x. This is consistent given **unconfoundedness** and regularity conditions.

- **Pro:** Transparent asymptotics and good, robust performance when p is small.
- **Con:** Acute curse of dimensionality.

NB: Kernels have similar qualitative issues as *k*-NN.

Adaptive nearest neighbor matching

Random forests are a a popular heuristic for adaptive nearest neighbors estimation introduced by Breiman (2001).

- **Pro:** Excellent empirical track record.
- **Con:** Often used as a black box, without statistical discussion.

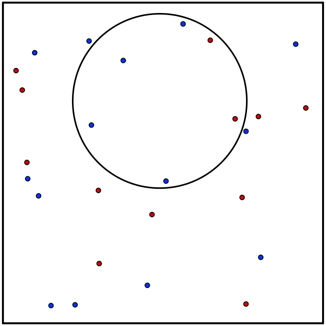
There has been considerable interest in using forest-like methods for treatment effect estimation, but without formal theory.

- Green and Kern (2012) and Hill (2011) have considered using **Bayesian forest algorithms** (BART, Chipman et al., 2010).
- Several authors have also studied related **tree-based methods**: Athey and Imbens (2016), Su et al. (2009), Taddy et al. (2014), Wang and Rudin (2015), Zeilis et al. (2008), ...

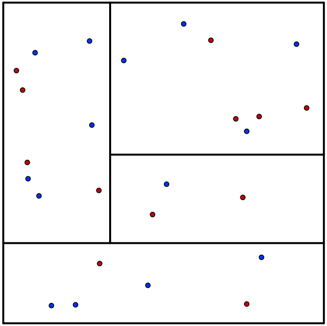
Wager and Athey (JASA, forthcoming) provide the first formal results allowing random forest to be used for provably valid **asymptotic inference**.

Making *k*-NN matching adaptive

Athey and Imbens (2016) introduce **causal tree**: defines neighborhoods for matching based on **recursive partitioning** (Breiman, Friedman, Olshen, and Stone, 1984), advocate sample splitting (w/ modified splitting rule) to get assumption-free confidence intervals for treatment effects in each leaf.



Euclidean neighborhood,
for *k*-NN matching.



Tree-based neighborhood.

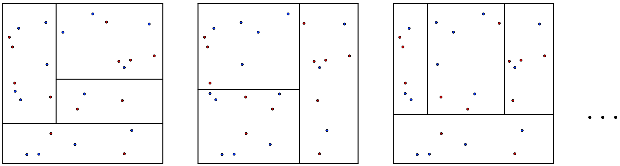
From trees to random forests (Breiman, 2001)

Suppose we have a training set $\{(X_i,\ Y_i,\ W_i)\}_{i=1}^n$, a test point x, and a tree predictor

$$\hat{\tau}\left(x\right) = T\left(x;\ \{(X_i,\ Y_i,\ W_i)\}_{i=1}^n\right).$$

Random forest idea: build and average many different trees T^*:

$$\hat{\tau}\left(x\right) = \frac{1}{B}\sum_{b=1}^{B} T_b^*\left(x;\ \{(X_i,\ Y_i,\ W_i)\}_{i=1}^n\right).$$



From trees to random forests (Breiman, 2001)

Suppose we have a training set $\{(X_i,\, Y_i,\, W_i)\}_{i=1}^n$, a test point x, and a tree predictor

$$\hat{\tau}\left(x\right) = T\left(x;\, \{(X_i,\, Y_i,\, W_i)\}_{i=1}^n\right).$$

Random forest idea: build and average many different trees T^*:

$$\hat{\tau}\left(x\right) = \frac{1}{B}\sum_{b=1}^{B} T_b^*\left(x;\, \{(X_i,\, Y_i,\, W_i)\}_{i=1}^n\right).$$

We turn T into T^* by:

- Bagging / subsampling the training set (Breiman, 1996); this helps smooth over discontinuities (Bühlmann and Yu, 2002).
- Selecting the splitting variable at each step from m out of p randomly drawn features (Amit and Geman, 1997).

Statistical inference with regression forests

Honest trees do not use the same data to select partition (splits) and make predictions. Ex: Split-sample trees, propensity trees.

Theorem. (Wager and Athey, 2015) Regression forests are asymptotically **Gaussian and centered**,

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0, 1), \quad \sigma_n^2(x) \to_p 0,$$

given the following assumptions (+ technical conditions):

1. **Honesty.** Individual trees are honest.
2. **Subsampling.** Individual trees are built on random subsamples of size $s \asymp n^{\beta}$, where $\beta_{\min} < \beta < 1$.
3. **Continuous features.** The features X_i have a density that is bounded away from 0 and ∞.
4. **Lipschitz response.** The conditional mean function $\mu(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right]$ is Lipschitz continuous.

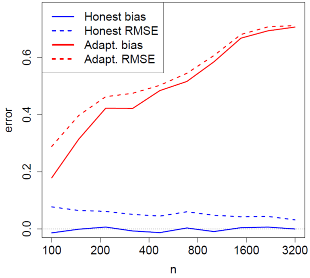
Valid Confidence Intervals

Athey and Imbens (2016), Wager and Athey (2015) highlight the perils of adaptive estimation for confidence intervals, tradeoff between MSE and coverage for trees but not forests.

Single Tree

	Ratio of infeasible MSE: Adaptive to honest†					
TOT-A/TOT-H		1.021		0.754		0.717
F-A/F-H		0.491		0.985		0.993
T-A/T-H		0.935		0.841		0.918
CT-A/CT-H		0.929		0.851		0.785
	Coverage of 90% confidence intervals – adaptive					
TOT-A	0.82	0.85	0.78	0.81	0.69	0.74
F-A	0.89	0.89	0.83	0.84	0.82	0.82
TS-A	0.84	0.84	0.78	0.82	0.75	0.75
CT-A	0.83	0.84	0.78	0.82	0.76	0.79
	Coverage of 90% confidence intervals – honest					
TOT-H	0.90	0.90	0.90	0.89	0.89	0.90
F-H	0.90	0.90	0.90	0.90	0.90	0.90
TS-H	0.90	0.90	0.91	0.91	0.89	0.90
CT-H	0.89	0.90	0.90	0.90	0.89	0.90

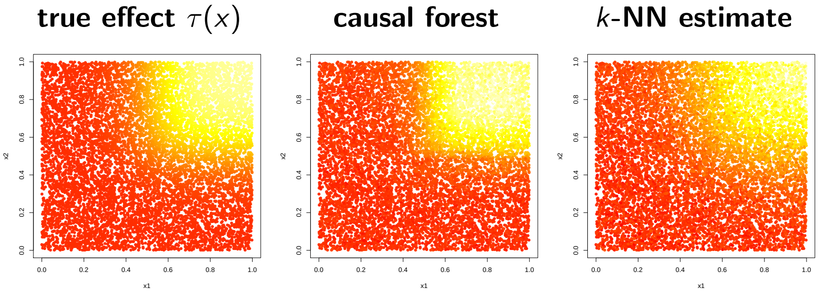
Forests



Causal forest example

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 6$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.



true effect $\tau(x)$

causal forest

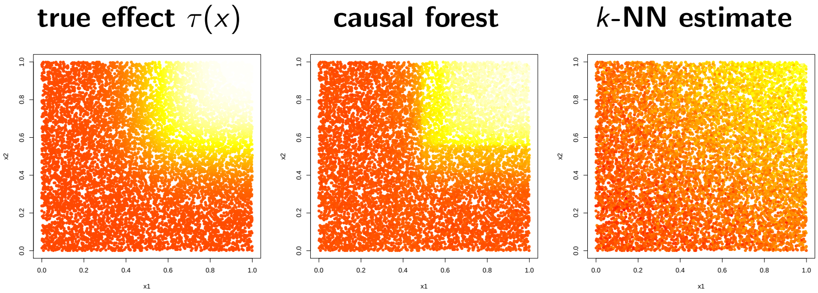
***k*-NN estimate**

Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015) available at `github: susanathey/causalTree`

Causal forest example

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 20$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.



true effect $\tau(x)$

causal forest

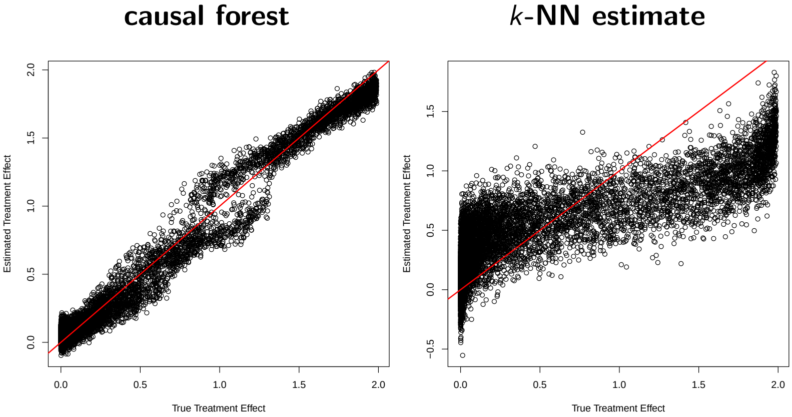
k-NN estimate

Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015) available at `github`: `susanathey/causalTree`

Causal forest example

The causal forest dominates k-NN for both bias and variance.
With $p = 20$, the relative mean-squared error (MSE) for τ is

$$\frac{\text{MSE for } k\text{-NN (tuned on test set)}}{\text{MSE for forest (heuristically tuned)}} = 19.2.$$



causal forest

k-NN estimate

For $p = 6$, the corresponding MSE ratio for τ is 2.2.

Applications of Instrumental Variables in Ad Effectiveness

- Intent-To-Treat v. Treatment on the Treated: Assignment to target group is the intent to treat, but not all targeted users are reached. Instrumental variables gives you the effect of the treatment on the treated.
- "Viewability" and related approaches: Some users did not see the ad because it was too low on the page or didn't render for other reasons
- A/B tests can be used as instruments, e.g. in search advertising, many A/B tests affect ad ranking; for e-commerce websites, A/B tests may affect the prominence of offers and "house" ads

Idea of IV: Use only the part of the variation in the treatment that is explained by the instrument, where the instrument is truly random

Use of ML: Control for confounders; personalized effects

Solving estimating equations with random forests

We have $i = 1, \ldots, n$ i.i.d. samples, each of which has an **observable** quantity O_i, and a set of **auxiliary covariates** X_i.

Examples:

- Non-parametric regression: $O_i = \{Y_i\}$.
- Treatment effect estimation: $O_i = \{Y_i, W_i\}$.
- Instrumental variables regression: $O_i = \{Y_i, W_i, Z_i\}$.

Our **parameter of interest**, $\theta(x)$, is characterized by an estimating equation:

$$\mathbb{E}\left[\psi_{\theta(x),\, \nu(x)}\left(O_i\right) \,\middle|\, X_i = x\right] = 0 \text{ for all } x \in \mathcal{X},$$

where $\nu(x)$ is an optional **nuisance parameter**.

The GMM Setup: Examples

Our parameter of interest, $\theta(x)$, is characterized by

$$\mathbb{E}\left[\psi_{\theta(x),\,\nu(x)}\left(O_i\right) \,\middle|\, X_i = x\right] = 0 \;\text{ for all }\; x \in \mathcal{X},$$

where $\nu(x)$ is an optional **nuisance parameter**.

- **IV regression**, with treatment assignment W and instrument Z. We care about the treatment effect $\tau(x)$:

$$\psi_{\tau(x),\,\mu(x)} = \begin{pmatrix} Z_i\left(Y_i - W_i\,\tau(x) - \mu(x)\right) \\ Y_i - W_i\,\tau(x) - \mu(x) \end{pmatrix}.$$

Solving heterogeneous estimating equations

The classical approach is to rely on **local solutions** (Fan and Gijbels, 1996; Hastie and Tibshirani, 1990; Loader, 1999).

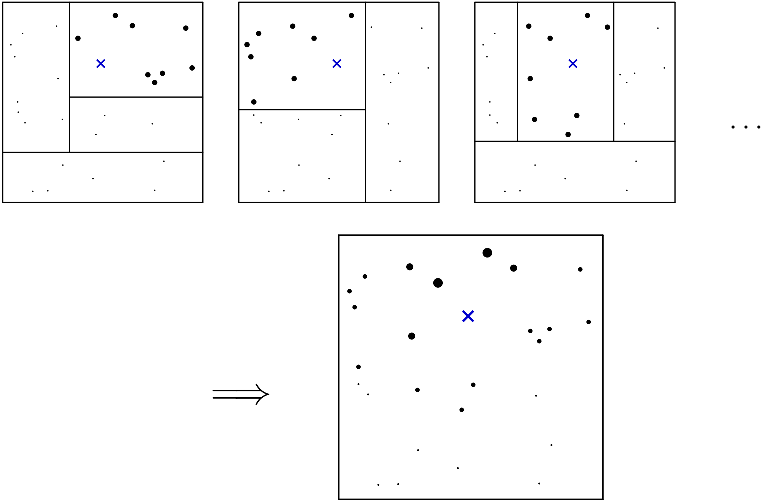
$$\sum_{i=1}^{n} \alpha(x;\, X_i)\; \psi_{\hat{\theta}(x),\, \hat{\nu}(x)}(O_i) = 0,$$

where the weights $\alpha(x;\, X_i)$ are obtained from, e.g., a **kernel**.

We use random forests to get good **data-adaptive** weights. Has potential to be help mitigate the **curse of dimensionality**.

- Building many trees with small leaves, then solving the estimating equation in each leaf, and finally **averaging the results** is a bad idea. Quantile and IV regression are badly **biased** in very small samples.
- Using RF as an "adaptive kernel" protects against this effect.

The random forest kernel



Forests induce a kernel via **averaging tree-based neighborhoods**.
This idea was used by Meinshausen (2006) for quantile regression.

Forests for GMM Parameter Heterogeneity

- Local GMM/ML uses kernel weighting to estimate personalized model for each individual, weighting nearby observations more.
  - Problem: curse of dimensionality
- We propose forest methods to determine what dimensions matter for "nearby" metric, reducing curse of dimensionality.
  - Estimate model for each point using "forest-based" weights: the fraction of trees in which an observation appears in the same leaf as the target
- We derive splitting rules optimized for objective
- Computational trick:
  - Use approximation to gradient (based on parent node parameters) to construct pseudo-outcomes
  - Then apply a splitting rule inspired by regression trees to these pseudo-outcomes
- Asymptotic normality: more subtle than regression/causal forests, but ultimately similar arguments apply

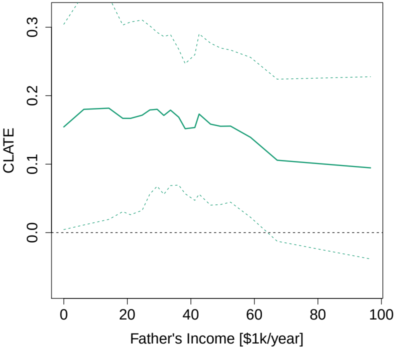
Empirical Application: Family Size

Angrist and Evans (1998) study the effect of family size on women's labor market outcomes. Understanding heterogeneity can guide policy.

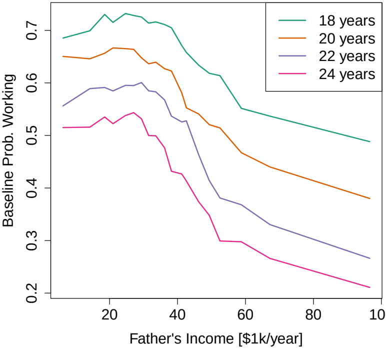
- Outcomes: participation, female income, hours worked, etc.
- Treatment: more than two kids
- Instrument: first two kids same sex
- First stage effect of same sex on more than two kids: .06
- Reduced form effect of same sex on probability of work, income: .008, $132
- LATE estimates of effect of kids on probability of work, income: .133, $2200

Treatment Effects: Magnitude of Decline

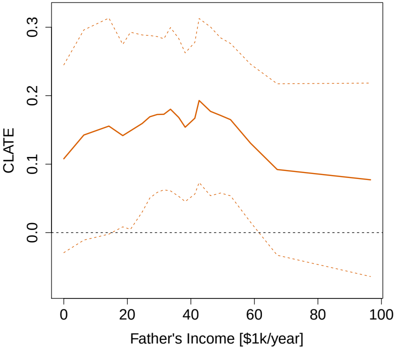
Effect on Participation



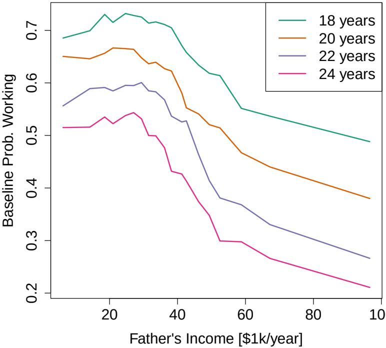
Baseline Probability of Working



Treatment Effects: Magnitude of Decline



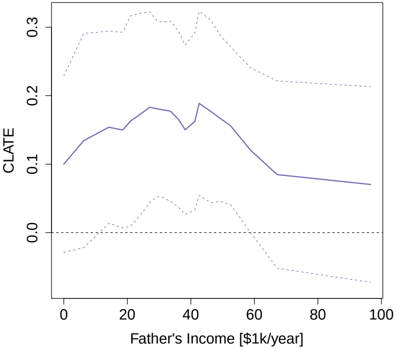
Effect on Participation



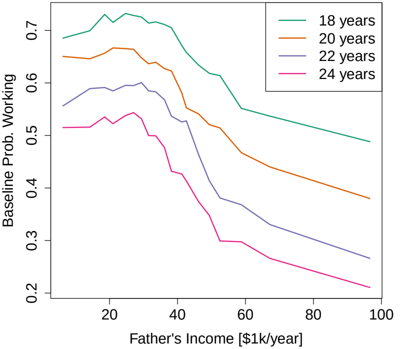
Baseline Probability of Working

Treatment Effects: Magnitude of Decline

Effect on Participation

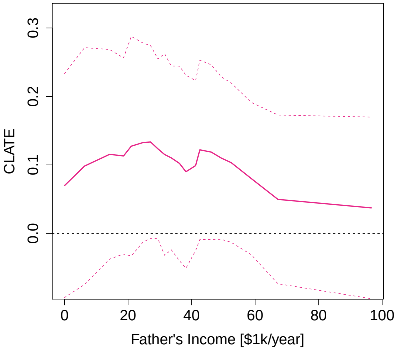


Baseline Probability of Working

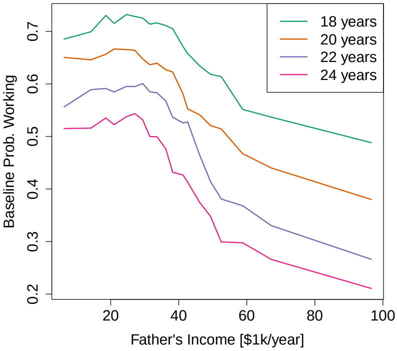


Treatment Effects: Magnitude of Decline

Effect on Participation



Baseline Probability of Working



Asymptotic normality of GRFs

Theorem. (Athey, Tibshirani and Wager, 2016) Given regularity of both the estimating equation and the data-generating distribution, GRFs are **consistent** and **asymptotically normal**:

$$\frac{\hat{\theta}_n(x) - \theta(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}\left(0,\, 1\right), \quad \sigma_n^2 \to 0.$$

Proof sketch.

- Influence functions: Hampel (1974); also parallels to use in Newey (1994).
- Influence function heuristic motivates approximating GRFs with regression forests applied to (infeasible) pseudo-outcomes

$$\tilde{\theta}_i = \theta - \xi^\top A^{-1}\ \psi_{\theta_P,\, \nu_P}\left(O_i\right)$$

- Analyze the approximating regression forests using Wager and Athey (2015)
- Use coupling result to derive conclusions about GRFs.

Pre-computing nuisance parameters

- For the IV case, can improve performance by residualizing the variables prior to constructing moments using leave-one-out estimators
    - $\tilde{Y}_i = Y_i - \hat{y}^{(-i)}(X_i)$, $\tilde{W}_i = Y_i - \hat{w}^{(-i)}(X_i)$, $\tilde{Z}_i = Z_i - \hat{z}^{(-i)}(X_i)$
- Moment conditions constructed using residualized variables have the same solution, but are not sensitive to errors in estimating nuisance parameters
- See Chernozhukov et al (2017) who argue for using Neyman-orthogonal moments in estimating structural parameters θ, e.g. the ATE, and Wager et al (2016)

Conclusions for Heterogeneous Parameter Estimation

- Local ML/GMM using kernel-based methods useful for allowing flexible estimation of heterogeneity in parameter estimates, but limited by curse of dimensionality
- Replacing kernel weighting function with forest-based function makes use of one of best-performing ML methods for non-parametric estimation without sacrificing asymptotic normality
- Proposed method solves computational/engineering challenges and guards against overfitting/instability
- Honesty (sample splitting) important

Efficient Policy Estimation

- **Learning optimal policy** assignment and **estimating treatment effect heterogeneity** closely related
- ML literature proposed variety of methods (Langford et al; Swaminathan and Joachims; in econometrics, Kitagawa and Tetenov)
- Estimating the value of a personalized policy closely related to **estimating average treatment effect** (comparing treat all policy to treat none policy)
- Lots of econometric theory about how to estimate average treatment effects efficiently (achieve semi-parametric efficiency bound)
- Athey and Wager (2017): prove that bounds on regret (gap between optimal policy and estimated policy) can be tightened using an algorithm consistent with econometric theory
  - Theory provides guidance for algorithm choice

Setup and Approach

Setup

- Policy $\pi : \mathcal{X} \to \{\pm 1\}$
- Given a class of policies Π, the optimal policy π^* and the regret $R(\pi)$ of any other policy are respectively defined as

$$\pi^* = \text{argmax}_{\pi \in \Pi} \left\{ \mathbb{E}\left[Y_i\left(\pi\left(X_i\right)\right)\right]\right\} \tag{1}$$

$$R(\pi) = \mathbb{E}\left[Y_i\left(\pi^*\left(X_i\right)\right)\right] - \mathbb{E}\left[Y_i\left(\pi\left(X_i\right)\right)\right]. \tag{2}$$

- Goal: estimate a policy π that minimizes regret $R(\pi)$.

Approach: Estimate $Q(\pi)$, choose policy to minimize $\hat{Q}(\pi)$:

$$Q(\pi) = \mathbb{E}\left[Y_i\left(\pi\left(X_i\right)\right)\right] - \frac{1}{2}\mathbb{E}\left[Y_i(-1) + Y_i(+1)\right] \tag{3}$$

$$\hat{\pi} = \text{argmax}_{\pi \in \Pi} \left\{ \widehat{Q}(\pi) \right\}, \tag{4}$$

Main Result

- Let $V(\pi)$ denote the semiparametrically efficient variance for estimating $Q(\pi)$.
- Let $V_* := V(\pi^*)$ denote the semiparametrically efficient variance for evaluating π^*
- Let $V_{\max}$ denote a sharp bound on the worst case efficient variance $\sup_\pi V(\pi)$ for any policy π. Results:
- Given policy class Π with VC dimension $VC(\Pi)$, proposed learning rule yields policy $\hat{\pi}$ with regret bounded by

$$R(\hat{\pi}) = \mathcal{O}_P\left(\sqrt{V_* \log\left(\frac{V_{\max}}{V_*}\right)\frac{VC(\Pi)}{n}}\right). \qquad (5)$$

- We also develop regret bounds for non-parametric policy classes Π with a bounded entropy integral, such as finite-depth decision trees.

Policy

- Estimate $Q(\pi)$ as

$$\widehat{Q}_{DML}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \pi\left(X_i\right) \widehat{\Gamma}_i \tag{6}$$

$$\widehat{\Gamma}_i := \hat{\mu}_{+1}^{(-k(i))}(X_i) - \hat{\mu}_{-1}^{(-k(i))}(X_i) + W_i \, \frac{Y_i - \hat{\mu}_{W_i}^{(-k(i))}(X_i)}{\hat{e}_{W_i}^{(-k(i))}\left(X_i\right)}, \tag{7}$$

- $k(i) \in \{1, ..., K\}$ denotes the fold containing the i-th conservation.
- Estimated propensity score for treatment W_i: $\hat{e}_{W_i}^{(-k(i))}\left(X_i\right)$
- Estimate this using a classifier with labels $\text{sign}(\widehat{\Gamma}_i)$ and weights $|\widehat{\Gamma}_i|$

Conclusions

Contributions from causal inference and econometrics literature:

- Identification and estimation of causal effects
- Classical theory to yield asymptotically normal and centered confidence intervals
- Semiparametric efficiency theory

Contributions from ML:

- Practical, high performance algorithms for personalized prediction and policy estimation

Putting them together:

- Practical, high performance algorithms
- Causal effects with valid confidence intervals
- Consistent with insights from efficiency theory