Heterogeneous Treatment Effects and Parameter Estimation with Causal Forests and Gradient Forests

Susan Athey
Stanford University

Machine Learning and Econometrics

See Wager and Athey (forthcoming, JASA)
and Athey, Tibshirani and Wager
https://arxiv.org/abs/1610.01271

Treatment Effect Heterogeneity

Heterogeneous Treatment Effects

- Insight about mechansims
- Designing policies, selecting groups for application/eligibility
- Personalized policies

Literature: Many Covariates

- See Wager and Athey (2015) and Athey and Imbens (2016) for ML-based analyses and many references on treatment effect heterogeneity
- 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

ML Methods for Causal Inference: Treatment Effect Heterogeneity

- ML methods perform well in practice, but many do not have well established statistical properties (see Chen and White (1999) for early analysis of neural nets)
- 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** (Wager and Athey, 2015; Taddy et al 2014; others)

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 ($W_i = 0$).

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, 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** $W_i \in \{0, 1\}$.

Goal is to estimate the **conditional average treatment effect**

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

NB: In experiments, we only get to see $Y_i = Y_i^{(W_i)}$.

The potential outcomes framework

If we make no further assumptions, estimating $\tau(x)$ is not possible.

- Literature often assumes **unconfoundedness** (Rosenbaum and Rubin, 1983)

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp\!\!\!\perp W_i \mid X_i.$$

- When this assumption holds, methods based on matching or propensity score estimation are usually consistent.

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, even when $p = 20$ and $n = 20k$.

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.

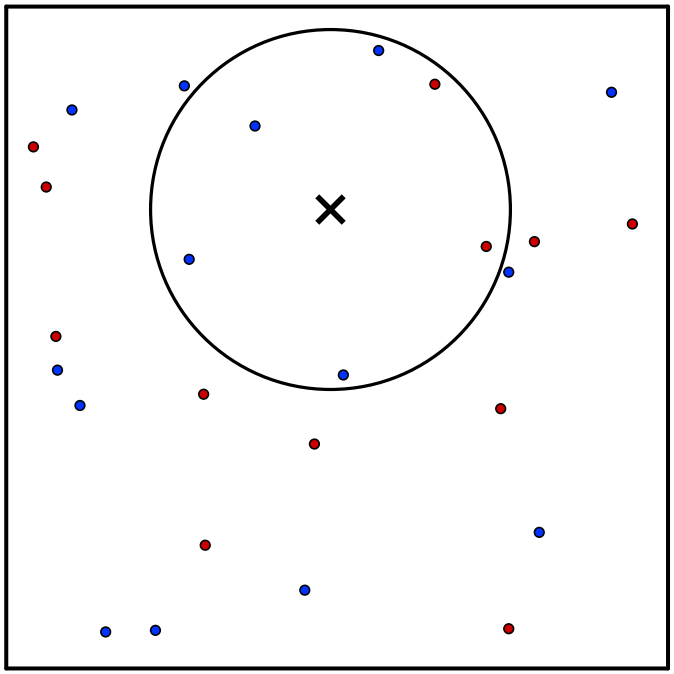
There has been considerable interest in using forest-like methods for treatment effect estimation, but without formal theory.

- 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), ...

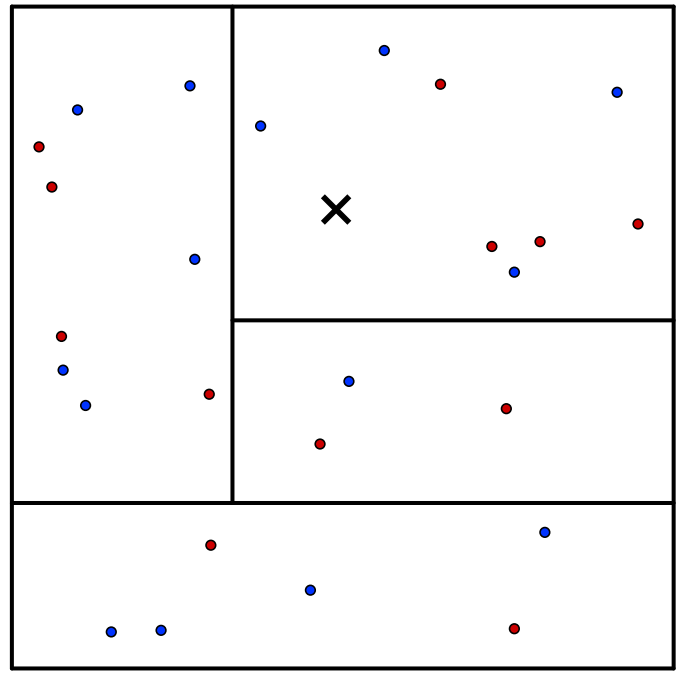
Wager and Athey (2015) provide the first formal results allowing random forest to be used for provably valid **asymptotic inference**.

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.

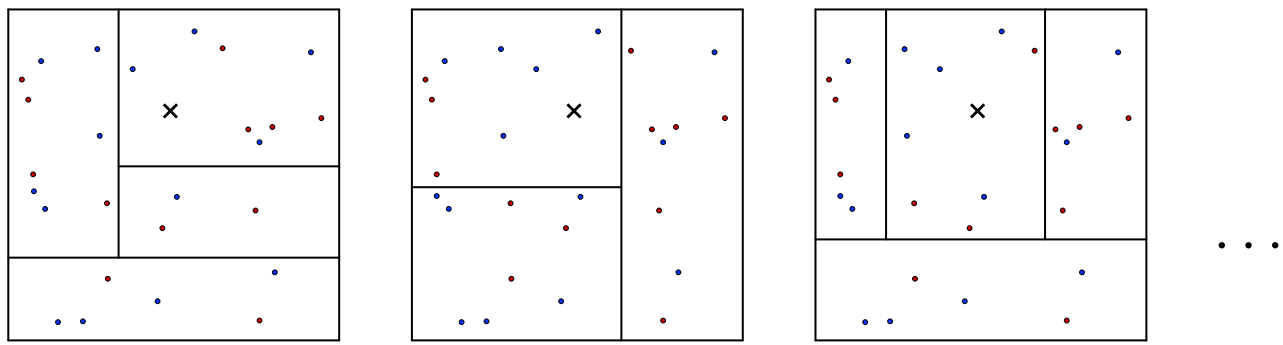
From trees to random forests (Breiman, 2001)

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

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

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

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



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

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

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

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

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 $\beta_{\text{min}} < \beta < 1$.
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.

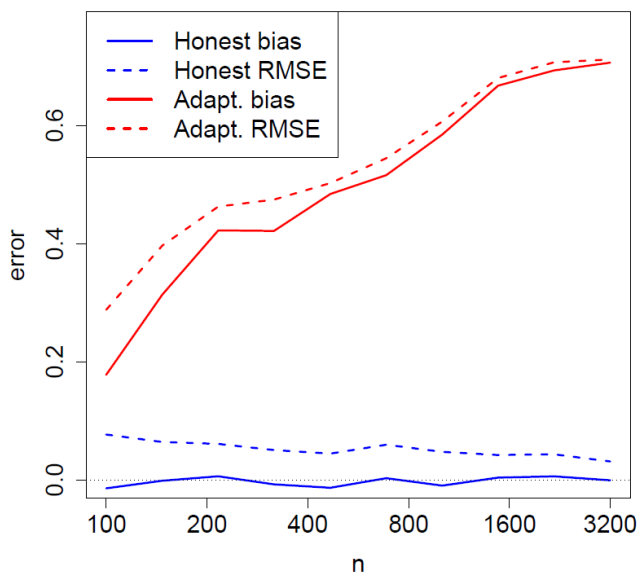
Valid Confidence Intervals

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



Proof idea

Use the shorthand $Z_i = (X_i,\ Y_i)$ for training examples.

- The regression forest prediction is

$$\hat{\mu} := \hat{\mu}\left(Z_1,\ ...,\ Z_n\right).$$

- The **Hájek projection** of the regression forest is

$$\mathring{\mu} = \mathbb{E}\left[\hat{\mu}\right] + \sum_{i=1}^{n} \left(\mathbb{E}\left[\hat{\mu} \,\middle|\, Z_i\right] - \mathbb{E}\left[\hat{\mu}\right]\right).$$

- Classical statistics (Hoeffding, Hájek) tells us that

$$\mathrm{Var}\left[\mathring{\mu}\right] \leq \mathrm{Var}\left[\hat{\mu}\right], \text{ and that } \lim_{n\to\infty} \mathrm{Var}\left[\mathring{\mu}\right] / \mathrm{Var}\left[\hat{\mu}\right] = 1$$

  implies **asymptotic normality**.

Proof idea

Now, let $\hat{\mu}_b^*(x)$ denote the estimate for $\hat{\mu}(x)$ given by a single regression tree, and let $\mathring{\mu}_b^*$ be its Hájek projection,

- Using the **adaptive nearest neighbors** framework of Lin and Jeon (2006), we show that

$$\text{Var}\left[\mathring{\mu}_b^*\right] \,/\, \text{Var}\left[\hat{\mu}_b^*\right] \gtrsim \log^{-p}(s).$$

- As a consequence of the **ANOVA decomposition** of Efron and Stein (1981), the full forest gets

$$\frac{\text{Var}\left[\mathring{\mu}\right]}{\text{Var}\left[\hat{\mu}\right]} \geq 1 - \frac{s}{n} \frac{\text{Var}\left[\hat{\mu}_b^*\right]}{\text{Var}\left[\mathring{\mu}_b^*\right]},$$

  thus yielding the **asymptotic normality** result for $s \asymp n^\beta$ for any $0 < \beta < 1$.

- For **centering**, we bound the bias by requiring $\beta > \beta_{\text{min}}$.

Variance estimation for regression forests

We estimate the variance of the regression forest using the **infinitesimal jackknife for random forests** (Wager, Hastie, and Efron, 2014). For each of the $b = 1, ..., B$ trees comprising the forest, define

- The estimated response as $\hat{\mu}_b^*(x)$, and
- The number of times the i-th observation was used as N_{bi}^*.

Then, defining Cov_* as the covariance taken with respect to all the trees comprising the forest, we set

$$\hat{\sigma}^2 = \frac{n-1}{n} \left(\frac{n}{n-s} \right)^2 \sum_{i=1}^{n} \text{Cov}_* \left[\hat{\mu}_b^*(x),\, N_{bi}^* \right]^2 .$$

Theorem. (Wager and Athey, 2015) Given the same conditions as used for asymptotic normality, the infinitesimal jackknife for regression forests is consistent:

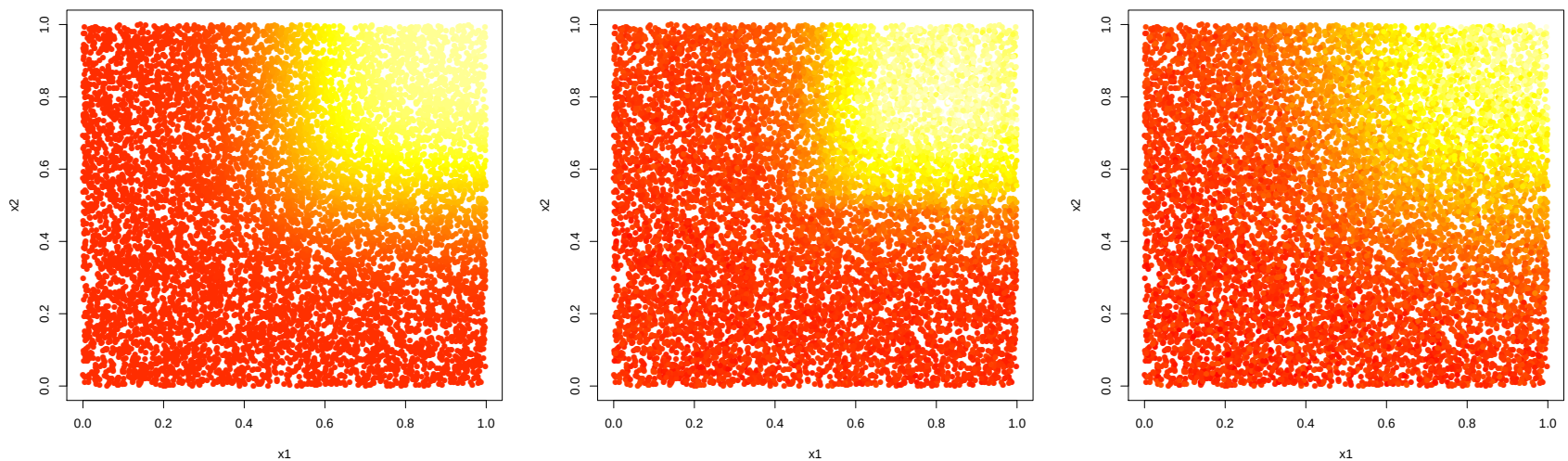
$$\hat{\sigma}_n^2(x) \,/\, \sigma_n^2(x) \to_p 1.$$

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**



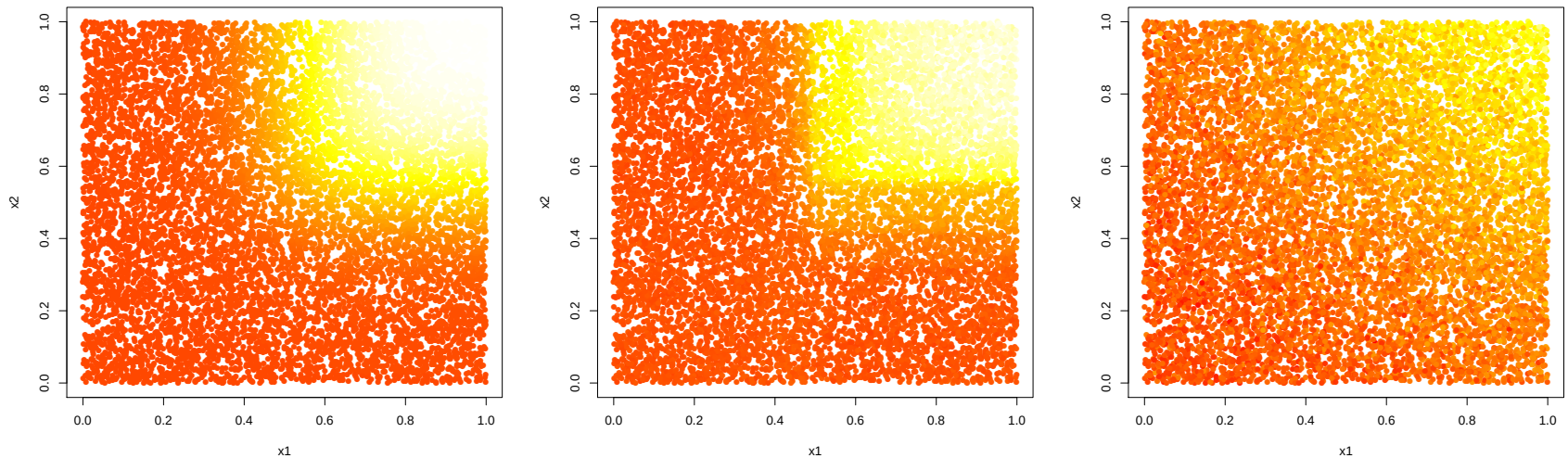
Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015) available at `github: susanathey/causalTree`

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`

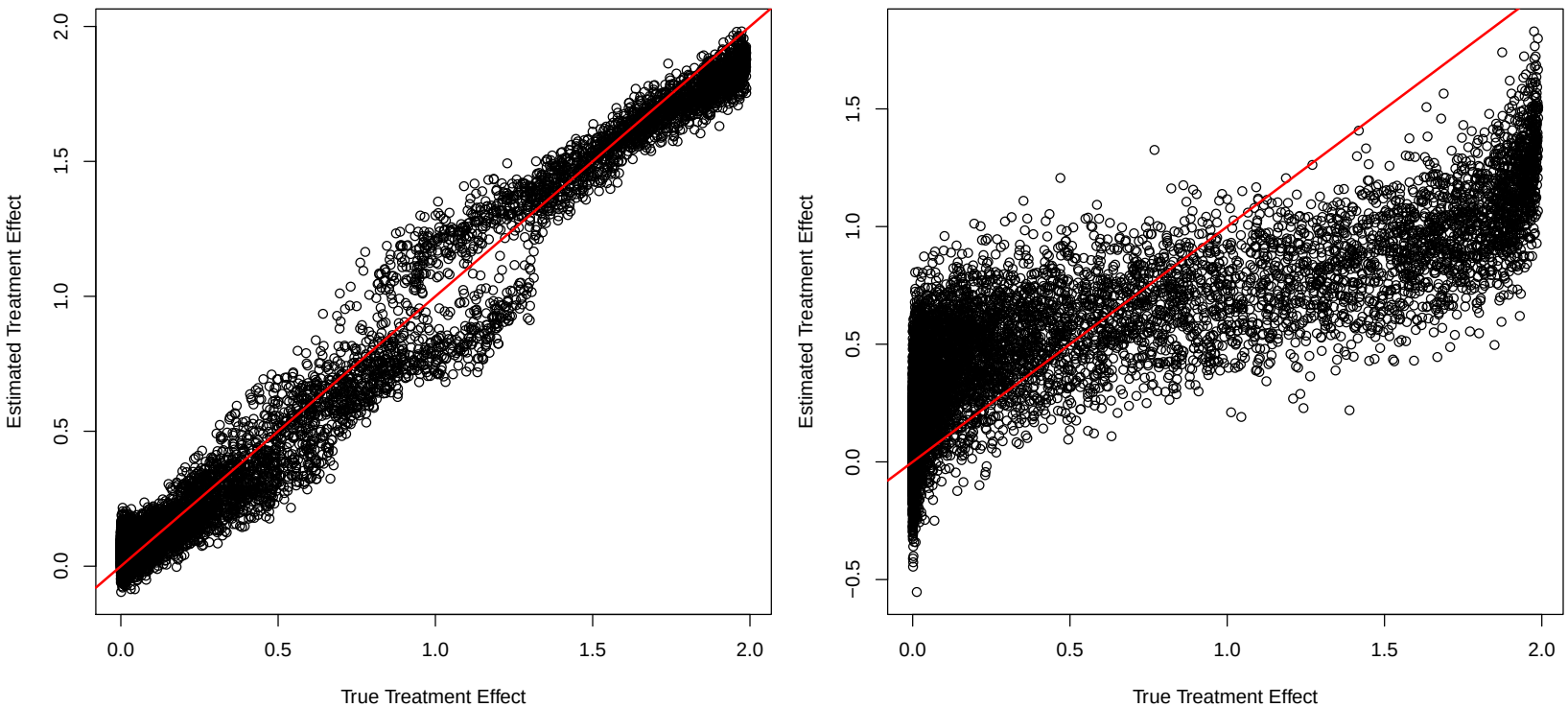
Causal forest example

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.

Application: General Social Survey

The General Social Survey is an extensive survey, collected since 1972, that seeks to measure demographics, political views, social attitudes, etc. of the U.S. population.

Of particular interest to us is a **randomized experiment**, for which we have data between 1986 and 2010.

- **Question A:** Are we spending too much, too little, or about the right amount on **welfare**?
- **Question B:** Are we spending too much, too little, or about the right amount on **assistance to the poor**?

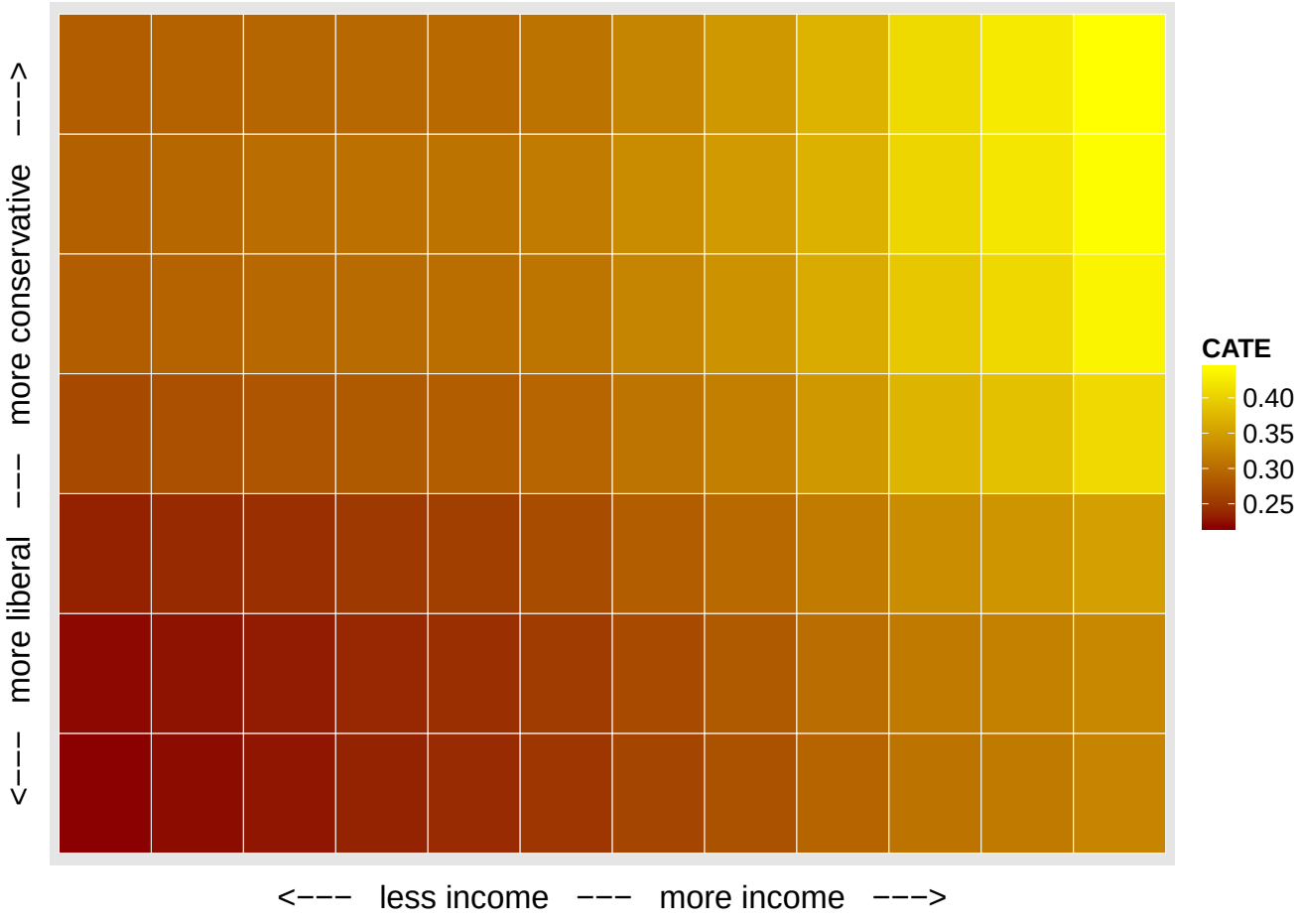
Treatment effect: how much less likely are people to answer **too much** to question B than to question A.

- We want to understand how the treatment effect depends on **covariates**: political views, income, age, hours worked, ...

NB: This dataset has also been analyzed by Green and Kern (2012) using Bayesian additive regression trees (Chipman, George, and McCulloch, 2010).

Application: General Social Survey

A causal forest analysis uncovers **strong treatment heterogeneity** ($n = 28,686$, $p = 12$).



ML Methods for Causal Inference: More general models

- Much recent literature bringing ML methods to causal inference focus 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:
  - Quantile regression
  - Instrumental Variables
  - Panel regression
  - Consumer choice
  - Euler equations
  - Survival analysis

ML Methods for Causal Inference: More general models

Average Treatment Effects with IV or Unconfoundedness

- In a series of influential papers, Belloni, Chernozhukov, Hansen, et al generalized LASSO methods to average treatment effect estimation through instrumental variables models, unconfoundedness, and also moment-based methods
- See also Athey, Imbens and Wager (2016) combine regularized regression and high-dimensional covariate balancing for average treatment effect estimation; and references therein on more recent papers on ATE estimation in high dimensions

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 to construct pseudo-outcomes
  - Then apply a splitting rule inspired by regression trees to these pseudo-outcomes

Related Work

(Semi-parametric) local maximum likelihood/GMM

- Local ML (Hastie and Tibshirani, 1987) weights nearby observations; e.g. local linear regression. See Loader, C. (1999); also Hastie and Tibshirani (1990) on GAM; see also Newey (1994)
- Lewbel (2006) asymptotic prop of kernel-based local GMM
- Other approaches include Sieve: Chen (2007) reviews

Score-based test statistics for parameter heterogeneity

- Andrews (1993), Hansen (1992), and many others, e.g. structural breaks, using scores of estimating equations
- Zeiles et al (2008) apply this literature to split points, when estimating models in the leaves of a single tree.

Splitting rules

- CART: MSE of predictions for regression, Gini impurity for classification, survival (see Bouhamad et al (2011))
- Statistical tests, multiple testing corrections: Su et al (2009)
- Causal trees/forests: adaptive v. honest est. (Athey and Imbens, 2016); propensity forests (Wager and Athey, 2015)

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:

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

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

The GMM Setup: Examples

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

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

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

- **Quantile regression**, where $\theta(x) = F_x^{-1}(q)$ for $q \in (0,\, 1)$:

$$\psi_{\theta(x)}(Y_i) = q\,\mathbf{1}(\{Y_i > \theta(x)\}) - (1-q)\,\mathbf{1}(\{Y_i \leq \theta(x)\})$$

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

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

Solving heterogeneous estimating equations

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

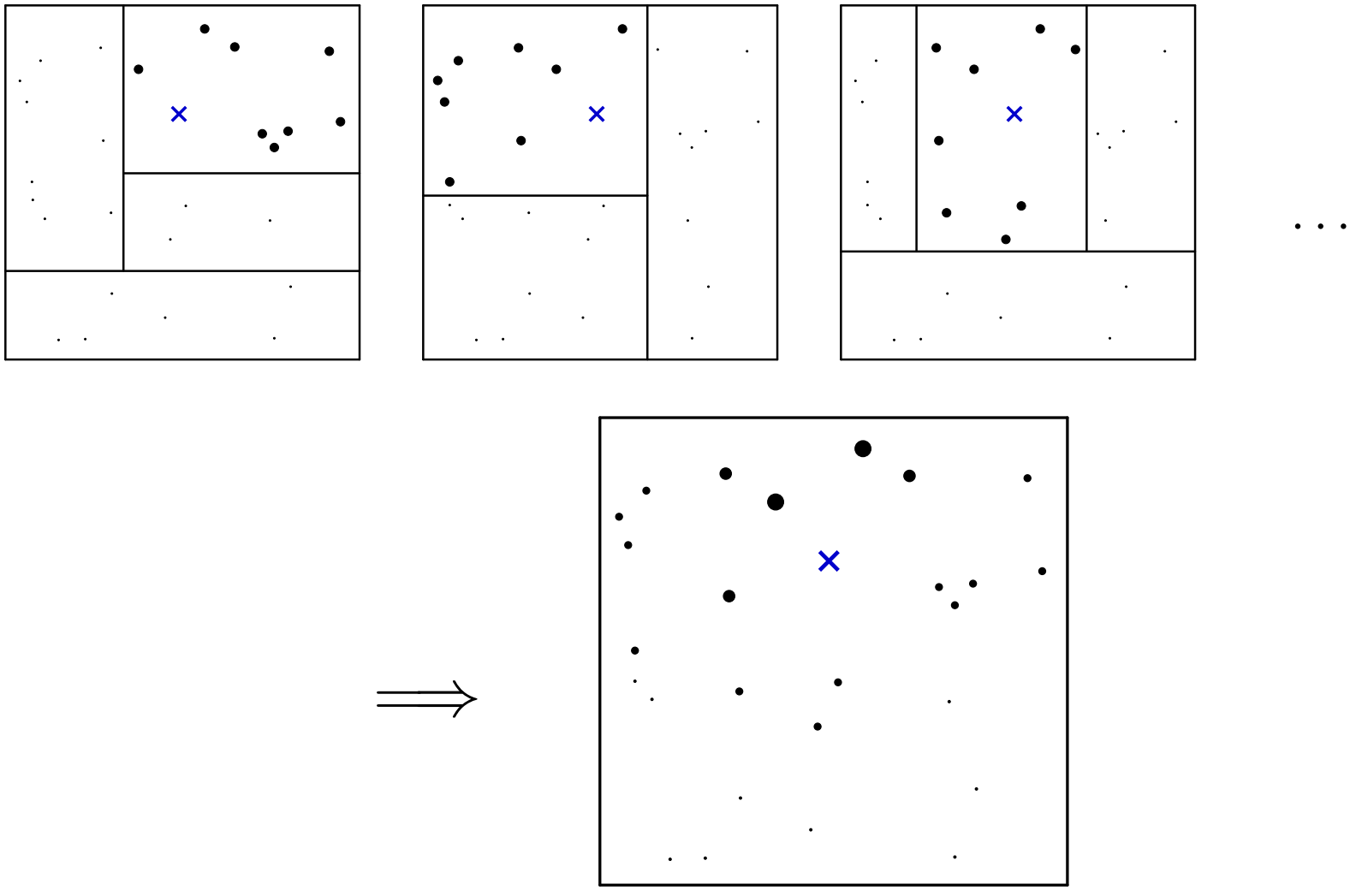
$$\sum_{i=1}^{n} \alpha(x;\, X_i)\; \psi_{\hat{\theta}(x),\, \hat{\nu}(x)}\left(O_i\right) = 0,$$

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.

Solving estimating equations with random forests

We want to use an estimator of the form

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

where the weights $\alpha(x;\, X_i)$ are from a random forest.

Key Challenges:

- How do we grow trees that yield an **expressive** yet **stable** neighborhood function $\alpha(\cdot;\, X_i)$?
- We do not have access to **"prediction error"** for $\theta(x)$, so how should we **optimize splitting**?
- How should we account for **nuisance parameters**?
- Split evaluation rules need to be **computationally efficient**, as they will be run many times for each split in each tree.

Step #1: Conceptual motivation

Following CART (Breiman et al., 1984), we use **greedy splits**. Each split directly seeks to improve the fit as much as possible.

- For regression trees, in large samples, the **best split** is that which **increases the heterogeneity** of the predictions the most.
- The same fact also holds **locally** for estimating equations.

We split a parent node P into two children C_1 and C_2. In **large samples** and with **no computational constraints**, we would like to maximize

$$\Delta\left(C_1,\, C_2\right) = n_{C_1} n_{C_2} \left(\hat{\theta}_{C_1} - \hat{\theta}_{C_2}\right)^2,$$

where $\hat{\theta}_{C_1}$, $\hat{\theta}_{C_2}$ **solve the estimating equation in the children**.

Step #2: Practical realization

Computationally, solving the estimating equation in each possible child to get $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ can be **prohibitively expensive**.

To avoid this problem, we use a **gradient-based approximation**. The same idea underlies gradient boosting (Friedman, 2001).

$$\hat{\theta}_C \approx \tilde{\theta}_C := \hat{\theta}_P - \frac{1}{|\{i : X_i \in C\}|} \sum_{\{i:X_i \in C\}} \xi^\top A_P^{-1} \, \psi_{\hat{\theta}_P, \hat{\nu}_p}(O_i),$$

$$A_P = \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i:X_i \in P\}} \nabla \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i),$$

where $\hat{\theta}_P$ and $\hat{\nu}_P$ are obtained by solving the estimating equation once in the parent node, and ξ is a vector that picks out the θ-coordinate from the (θ, ν) vector.

Step #2: Practical realization

In practice, this idea leads to a **split-relabel** algorithm:

1. **Relabel step:** Start by computing pseudo-outcomes

$$\tilde{\theta}_i = -\xi^\top A_P^{-1} \, \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \in \mathbb{R}.$$

2. **Split step:** Apply a CART-style regression split to the $\widetilde{Y}_i$.

This procedure has several advantages, including the following:

- **Computationally**, the most demanding part of growing a tree is in scanning over all possible splits. Here, we reduce to a regression split that can be efficiently implemented.
- **Statistically**, we only have to solve the estimating equation once. This reduces the risk of hitting a numerically unstable leaf—which can be a risk with methods like IV.
- From an **engineering** perspective, we can write a single, optimized split-step algorithm, and then use it everywhere.

Step #3: Variance correction

Conceptually, we saw that—in large samples—we want splits that maximize the heterogeneity of the $\hat{\theta}(X_i)$. In small samples, we need to account for **sampling variance**.

We need to penalize for the following two sources of variance.

- Our **plug-in estimates** for the heterogeneity of $\hat{\theta}(X_i)$ will be **overly optimistic** about the large-sample parameter heterogeneity. We need to correct for this kind of over-fitting.
- We **anticipate "honest" estimation**, and want to avoid leaves where the **estimating equation is unstable**. For example, with IV regression, we want to avoid leaves with an unusually weak 1st-stage coefficient.

This is a generalization of the analysis of Athey and Imbens (2016) for treatment effect estimation.

Gradient forests

Our label-and-regress splitting rules can be used to grow an ensemble of trees that yield a forest kernel. We call the resulting procedure a **gradient forest**.

- Regression forests are a special case of gradient forests with a squared-error loss.

Available as an `R`-package, `gradientForest`, built on top of the `ranger` package for random forests (Wright and Ziegler, 2015).

Asymptotic normality of gradient forests

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

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

Proof sketch.

- Influence functions: Hampel (1974); also parallels to use in Newey (1994).
- Influence function heuristic motivates approximating gradient forests with a class of regression forests.
- Analyze the approximating regression forests using Wager and Athey (2015)
- Use coupling result to derive conclusions about gradient forests.

Asymptotic normality of gradient forests: Proof details

- Influence function heuristic motivates approximating gradient forests with a class of regression forests. Start as if we knew true parameter value in calculating influence fn:
  - Let $\tilde{\theta}_i^*(x)$ denote the influence function of the i-th observation with respect to the true parameter value $\theta(x)$: $\widetilde{Y}_i^*(x) = -\xi^\top V(x)^{-1} \psi_{\theta(x),\,\nu(x)}(O_i)$
  - Pseudo-forest predictions: $\tilde{\theta}^*(x) = \theta(x) + \sum_{i=1}^n \alpha_i \tilde{\theta}_i^*(x)$.
- Apply Wager and Athey (2015) to this. Key points: $\tilde{\theta}^*(x)$ is linear function, so we can write it as an average of tree predictions, with trees built on subsamples. Thus it is U-statistic; can use the ANOVA decomposition.
- Coupling result: conclusions about gradient forests.

Suppose that the gradient forest estimator $\hat{\theta}(x)$ is consistent for $\theta(x)$. Then $\hat{\theta}(x)$ and $\tilde{\theta}^*(x)$ are coupled,

$$\tilde{\theta}^*(x) - \hat{\theta}(x) = o_P\left(\left\|\sum_{i=1}^n \alpha_i(x)\,\psi_{\theta(x),\,\nu(x)}\,(O_i)\right\|_2\right). \qquad (1)$$

Simulation example: Quantile regression

In quantile regression, we want to estimate the q-th quantile of the conditional distribution of Y given X, namely $\theta(x) = F_x^{-1}(q)$.

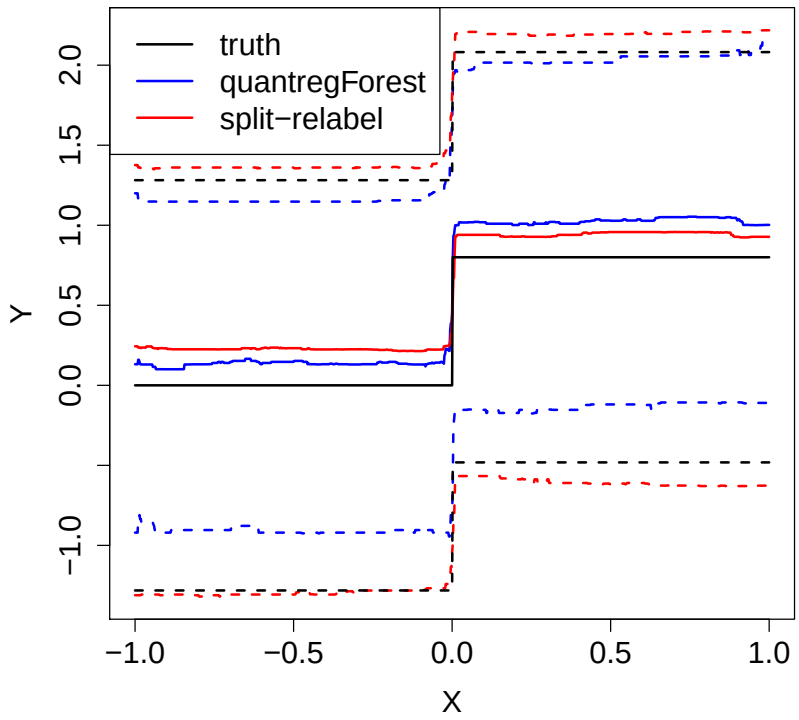
- Meinshausen (2006) used the random forest kernel for quantile regression. However, he used standard **CART regression splitting** instead of a tailored splitting rule.
- In our split-relabel paradigm, **quantile splits** reduce to **classification splits** ($\hat{\theta}_P$ is the q-th quantile of the parent):

$$\widetilde{Y}_i = \mathbf{1}(\{Y_i > \hat{\theta}_P\}).$$

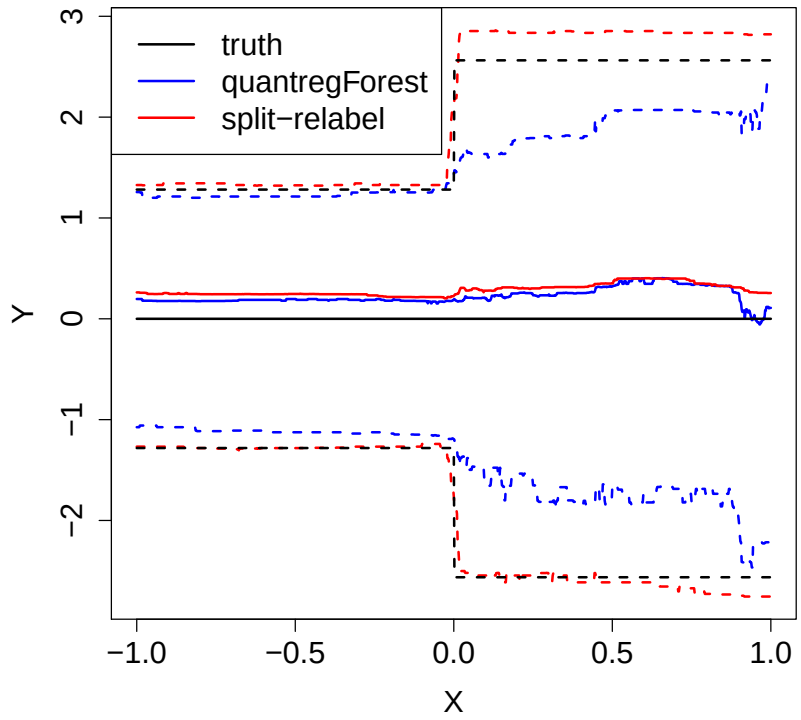
- To estimate **many quantiles**, we do **multi-class** classification.

Simulation example: Quantile regression

Case 1: Mean Shift



Case 2: Scale Shift



The above examples show quantile estimates at $q = 0.1, 0.5, 0.9$, on Gaussian data with $n = 2,000$ and $p = 40$. The package `quantregForest` implements the method of Meinshausen (2006).

Simulation example: Instrumental variables

We want to estimate **heterogeneous treatment effects** with endogenous treatment assignment: Y_i is the treatment, W_i is the treatment assignment, and Z_i is an instrument satisfying:

$$\{Y_i(w)\}_{w \in \mathcal{W}} \perp\!\!\!\perp Z_i \,\big|\, X_i.$$

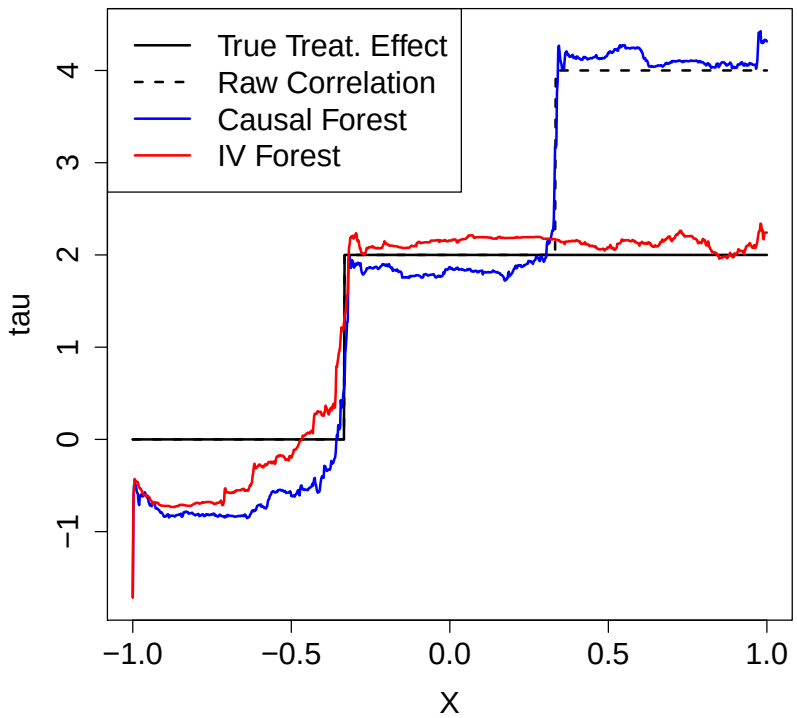
- Our **split-relabel** formalism tells us to use pseudo-outcomes

$$\tilde{\tau}_i = \left(Z_i - \bar{Z}_p\right)\left(\left(Y_i - \overline{Y}_p\right) - \hat{\tau}_P\left(W_i - \overline{W}_p\right)\right),$$

  where $\hat{\tau}_P$ is the IV solution in the parent, and $\overline{Y}_P$, $\overline{W}_p$, $\bar{Z}_p$ are averages over the parent.
- This is just IV regression residuals projected onto the instruments.

Simulation example: Instrumental variables

Using IV forests is important



We have **spurious correlations**:

- OLS for Y on W given X has two jumps, at $X_1 = -1/3$ and at $X_1 = 1/3$.
- The causal effect $\tau(X)$ only has a jump at $X_1 = -1/3$.
- $n = 10,000$, $p = 20$.

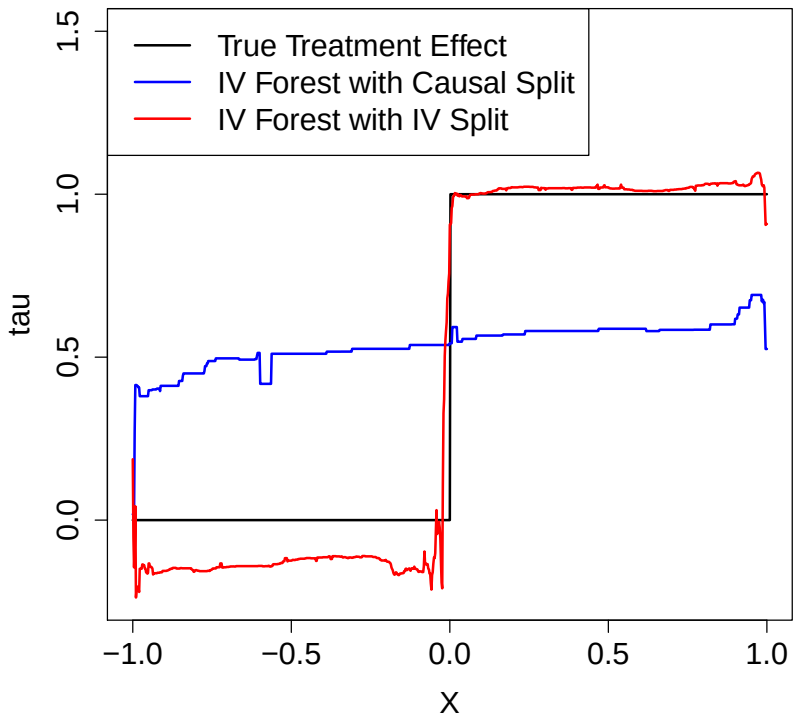
The response function is

$$Y_i = (2W_i - 1)\,\mathbf{1}(\{X_{1,i} > -1/3\}) + (3A - 1.5)\,\mathbf{1}(\{X_{1,i} > 1/3\}) + 2\varepsilon_i.$$

A_i is correlated with W_i.

Simulation example: Instrumental variables

Using IV splits is important



We have **useless correlations**:

- The joint distribution of $(W_i,\ Y_i)$ is independent of the covariates X_i.
- But: the causal effect $\tau(X)$ has a jump at $X_1 = 0$.
- $n = 5,000,\ p = 20$.

The response function is

$$Y_i = 2 \cdot \mathbf{1}(\{X_{1,i} \leq 0\})\, A_i + \mathbf{1}(\{X_{1,i} > 0\})\, W_i + \mathbf{1}(1 + 0.73 \cdot \mathbf{1}(\{X_{1,i} > 0\}))\varepsilon_i.$$

A_i is correlated with W_i.

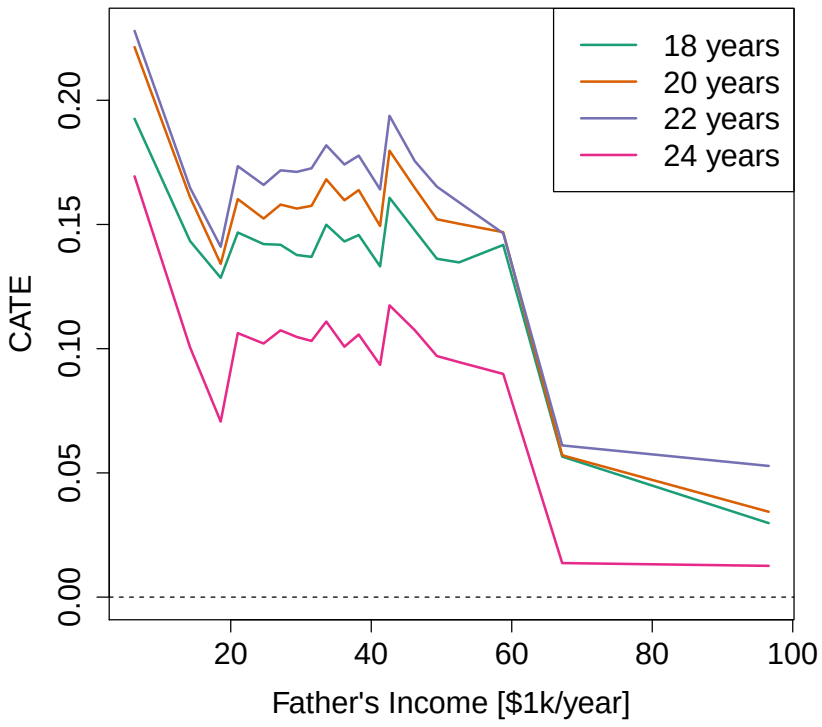
Empirical Application: Family Size

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

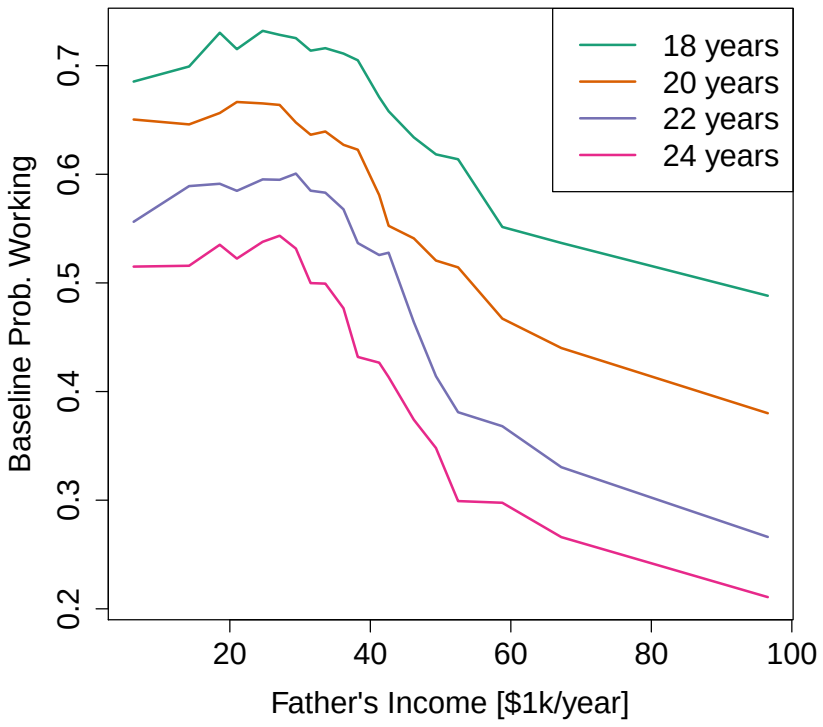
- Outcomes: participation, female income, hours worked, etc.
- Treatment: more than two kids
- Instrument: first two kids same sex
- First stage effect of same sex on more than two kids: .06
- Reduced form effect of same sex on probability of work, income: .008, $132
- LATE estimates of effect of kids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

Treatment Effects: Magnitude of Decline

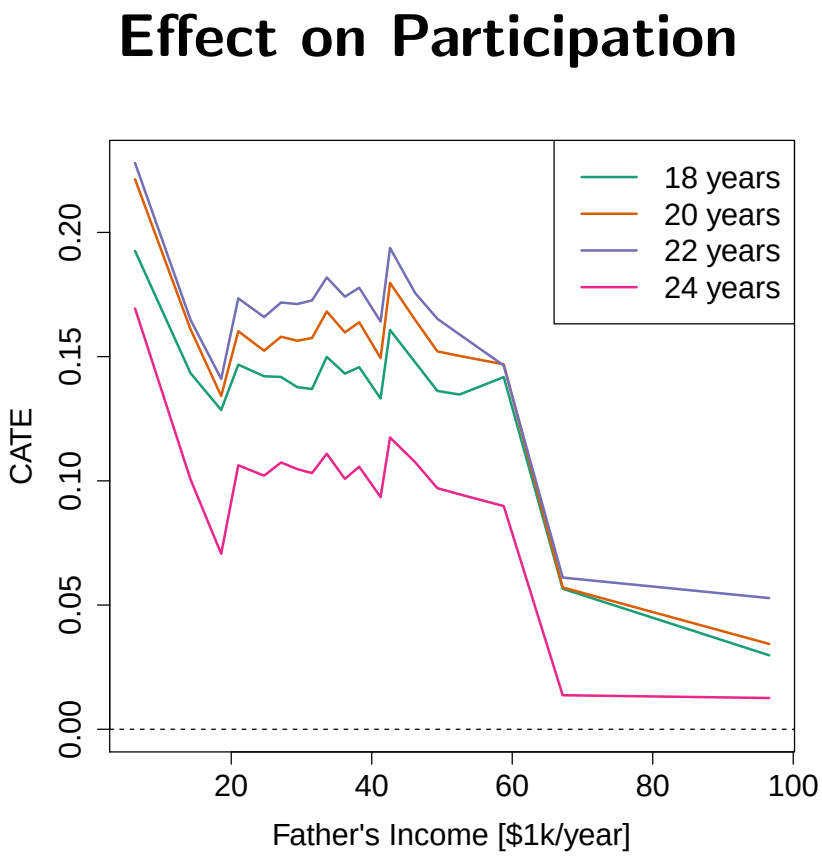
Effect on Participation

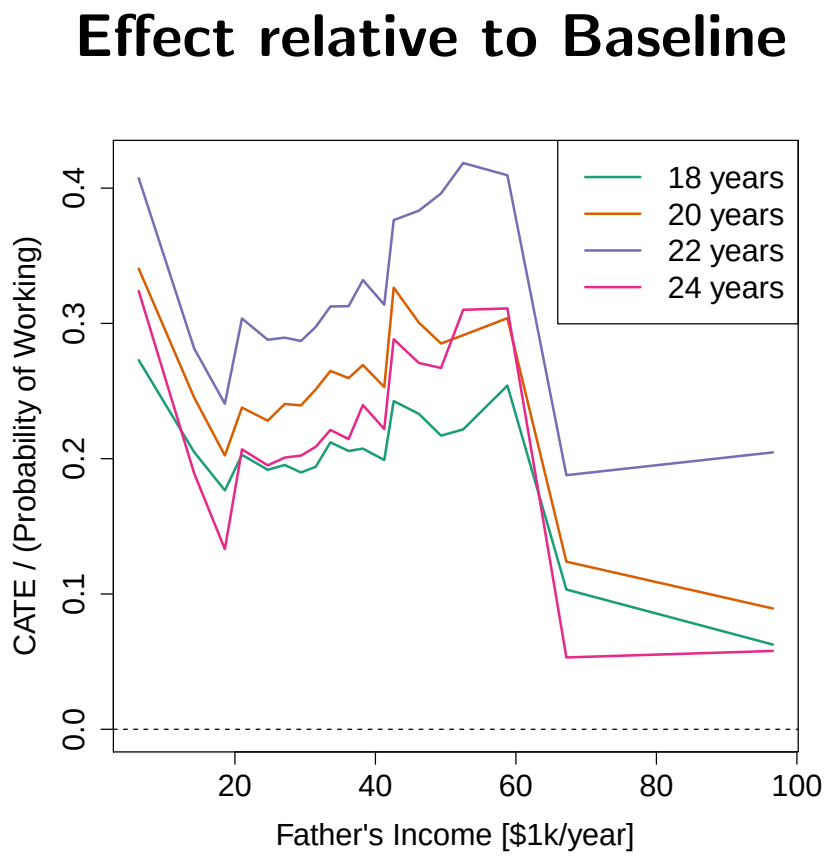


Baseline Probability of Working



Treatment Effects: Magnitude of Decline





Treatment Effects: Magnitude of Decline