Heterogeneous Treatment Effects and Parameter Estimation with Generalized Random Forests

Susan Athey
Stanford University

Machine Learning and Econometrics

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

Heterogeneous Parameter Estimates

- Heterogeneous treatment effects, elasticities, etc.
- Could estimate models within leaves of shallow trees; or regularize interaction effects in models
- Generalized random forests: look for parameter heterogeneity flexibly

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

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

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).

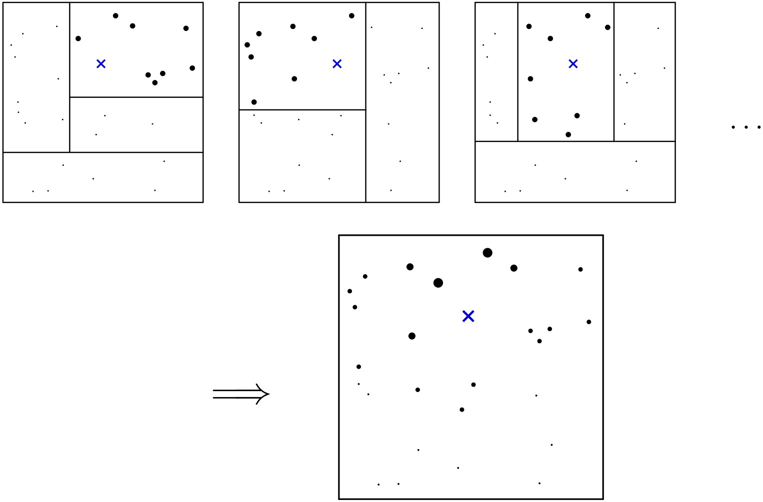
$$\sum_{i=1}^{n} \alpha(x;\, X_i)\; \psi_{\hat{\theta}(x),\, \hat{\nu}(x)}\left(O_i\right) = 0,$$

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

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

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.

Generalized Random 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 **generalized random forest**.

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

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

Asymptotic normality of GRF

Theorem. (Athey, Tibshirani and Wager, 2016) Given regularity of both the estimating equation and the data-generating distribution, generalized random forests 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 generalized random forests with a class of regression forests.
- Analyze the approximating regression forests using Wager and Athey (2015)
- Use coupling result to derive conclusions about GRF.

Asymptotic normality of GRF: Proof details

- Influence function heuristic motivates approximating GRFs 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 $\theta(x)$: $\tilde{\theta}_i^*(x) = -\xi^\top V(x)^{-1} \psi_{\theta(x),\, \nu(x)}(O_i)$
  - 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 GRFs.

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

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

Simulation example: Quantile regression

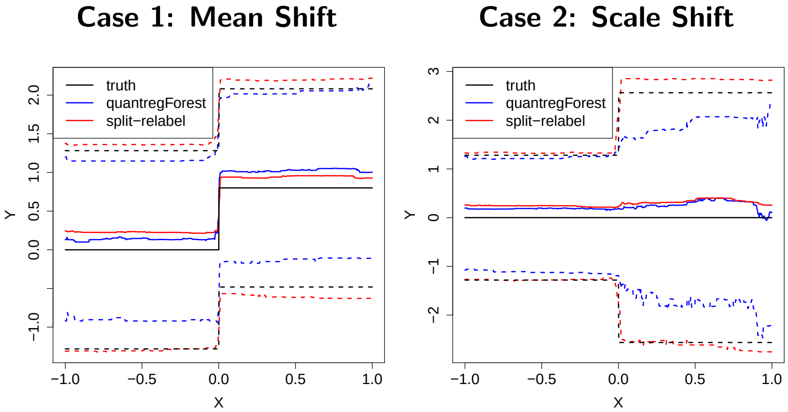
In quantile regression, we want to estimate the q-th quantile of the conditional distribution of Y given X, namely $\theta(x) = F_x^{-1}(q)$.

- 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



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:

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

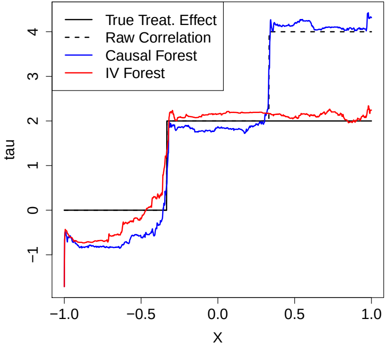
- Our **split-relabel** formalism tells us to use pseudo-outcomes

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

  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 $X_1 = -1/3$.
- $n = 10,000$, $p = 20$.

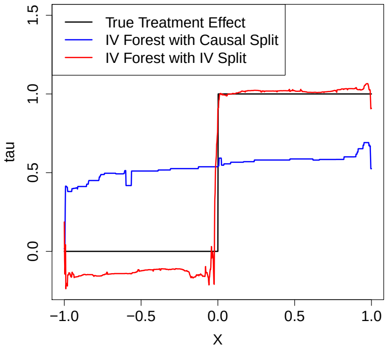
The response function is

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

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

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

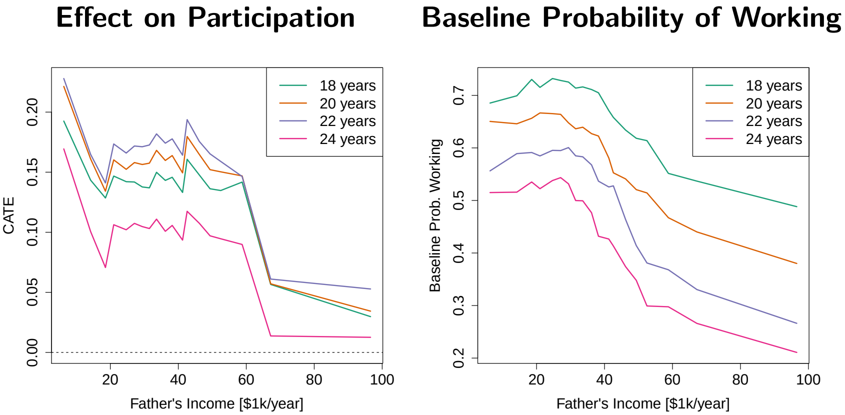
A_i is correlated with W_i.

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.

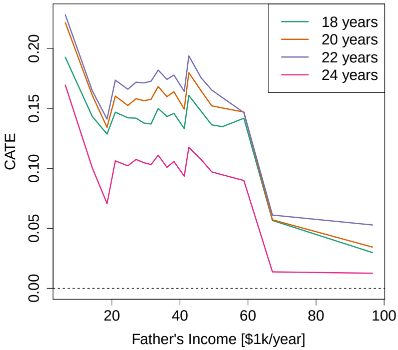
- Outcomes: participation, female income, hours worked, etc.
- Treatment: more than two kids
- Instrument: first two kids same sex
- First stage effect of same sex on more than two kids: .06
- Reduced form effect of same sex on probability of work, income: .008, $132
- LATE estimates of effect of kids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

Treatment Effects: Magnitude of Decline

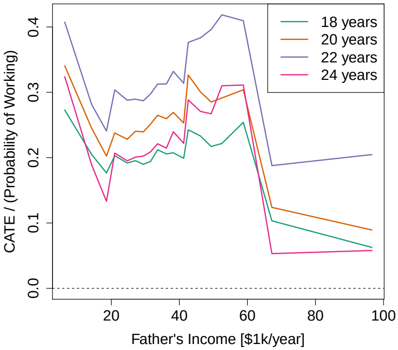


Treatment Effects: Magnitude of Decline

Effect on Participation



Effect relative to Baseline



Treatment Effects: Magnitude of Decline