Generalized Random Forests[1]

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

[1]This section is based on [Athey et al., 2019]

Outline

Background on Random Forests

- Introduced by Breiman (2001) as an ensemble learning method.
- Constructs a multitude of decision trees for various tasks:
  - Classification: Outputs the mode of the classes.
  - Regression: Output the mean/average prediction.
- Addresses overfitting, a common issue with single decision trees.
- Widely used for nonparametric conditional mean estimation:

$$\mu(x) = \mathbb{E}[Y_i \mid X_i = x] \tag{1}$$

- Supported by theoretical results on consistency and asymptotic properties.

Limitations of Standard Random Forests

- Traditional RF mainly for conditional mean estimation
- Struggles with complex statistical tasks beyond means
- Introduction to local moment conditions:

$$\mathbb{E}[\psi_{\theta(x),\nu(x)}(O_i) \mid X_i = x] = 0 \quad \forall x \in \mathcal{X} \tag{2}$$

- Need for a generalized approach

Generalized Random Forests Overview

- **Algorithm Essentials:**
  - Combines sample splitting and deep trees for robust estimation.
  - Applies local weighting for problem-specific analysis.
- **Extending Capabilities:**
  - Broadens RF applicability to a variety of statistical problems.
  - Integrates new methodologies and formalizes asymptotic behavior.
- **Advanced Weighting Techniques:**
  - Moves from averaging to adaptive weighting for improved accuracy.
  - Aligns with local likelihood concepts for better generalization.
- **Tailored Adaptations:**
  - Adapts neighborhood functions to the nature of the data.
  - Using custom rules for more effective partitioning.
- **Efficient Splitting Framework:**
  - Innovates splitting rules for specific challenges.
  - Incorporates gradient methods and optimized software for performance.

Applications of GRFs

- Causal Inference: Estimation of heterogeneous treatment effects.
- Quantile Regression: Estimating conditional quantiles of a response variable.
- Policy Learning: Discovering optimal policy rules.

Software and Implementation

- Open-source software packages are available for implementing GRFs.
- R package ‘grf‘ for Generalized Random Forests.
- Python implementations are also available.

Generalized Random Forests

- Generalized Random Forests (GRFs) extend traditional Random Forests (RFs).
- Prediction at a point x is not merely an average but uses adaptive weighting.
- GRFs are designed to estimate any quantity $\theta(x)$.
- They handle a broader range of statistical settings and provide formal asymptotic results.

References: Breiman (2001), Amit and Geman (1997), Breiman (1996), Dietterich (2000), Ho (1998), Breiman et al. (1984), Hastie, Tibshirani, and Friedman (2009)

Forest-Based Local Estimation

- Forest-based local estimation uses n i.i.d samples with observables O_i and covariates X_i.
- The goal is to estimate solutions to:

$$E[\psi_{\theta(x),\nu(x)}(O_i)|X_i = x] = 0$$

- $\theta(x)$ is our parameter of interest, and $\nu(x)$ is a nuisance parameter.

Estimation Equation

- Weights $\alpha_i(x)$ define the relevance of training examples to $\theta(\cdot)$ at x.
- The empirical estimation equation is:

$$(\hat{\theta}(x), \hat{\nu}(x)) \in \arg\min_{\theta,\nu} \left\{ \left\| \sum_{i=1}^{n} \alpha_i(x) \psi_{\theta,\nu}(O_i) \right\|_2 \right\}$$

- When the expression has a unique root:

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

References: Fan, Farmen and Gijbels (1998), Newey (1994a), Staniswalis (1989), Stone (1977), Tibshirani and Hastie (1987)

Adaptive Weights and Forest Structure

- GRFs learn adaptive weights $\alpha_i(x)$ by averaging over trees.
- For a tree b, $L_b(x)$ is the leaf containing x:

$$\alpha_{bi}(x) = \frac{1(\{X_i \in L_b(x)\})}{|L_b(x)|}, \quad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x)$$

- These weights sum to 1 and define an adaptive neighborhood.

References: Hothorn et al. (2004), Meinshausen (2006)

Equivalence to Regression Trees

- For regression trees estimating $\mu(x) = E[Y_i|X_i = x]$:
- The GRF estimate is equivalent to averaging tree predictions:

$$\sum_{i=1}^{n} \frac{1}{B} \left(\sum_{b=1}^{B} \alpha_{bi}(x)(Y_i - \hat{\mu}(x)) \right) = 0$$

- $\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b(x)$
- Where $\hat{\mu}_b(x)$ is the prediction of a single tree.

References: Breiman (2001)

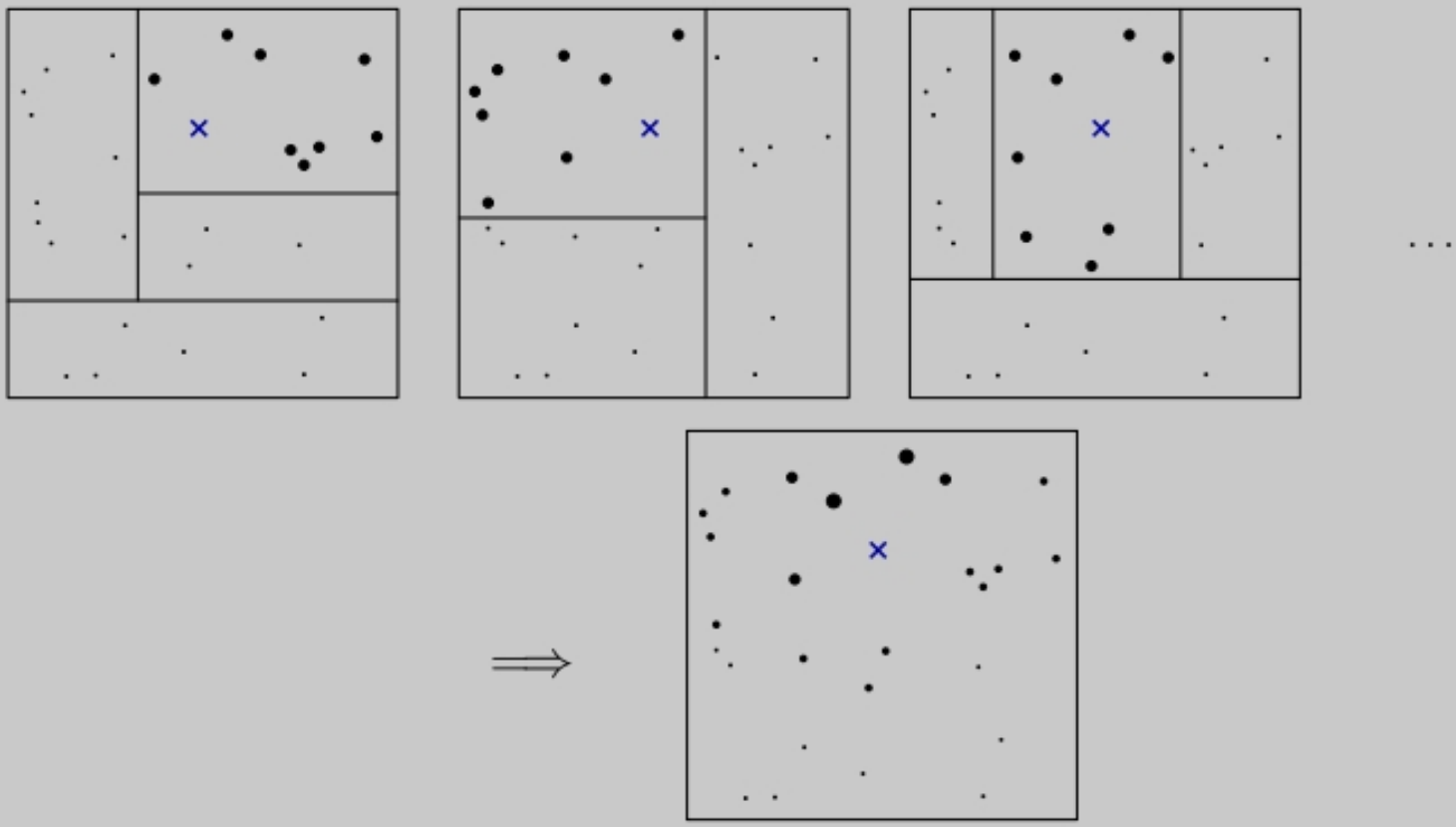

FIG. 1. *Illustration of the random forest weighting function. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Then the forest averages all these tree-based weightings, and effectively measures how often each training example falls into the same leaf as x.*

Introduction to the Splitting Problem

- Goal: Construct trees that induce weights $\alpha_i(x)$ for accurate $\theta(x)$ estimates.
- Key Difference: Use of recursive partitioning on subsamples to generate $\alpha_i(x)$.
- Objective: Mimic Breiman's algorithm, focusing on heterogeneity in $\theta(x)$.

Greedy Splitting Strategy

- Splits are selected greedily to maximize immediate improvement in tree fit.
- Starting with a parent node $P \subseteq X$ and given a sample of data J, we solve:

$$(\hat{\theta}_P, \hat{\nu}_P)(J) \in \arg\min_{\theta,\nu} \left\{ \left\| \sum_{i \in J, X_i \in P} \psi_{\theta,\nu}(O_i) \right\|_2 \right\}.$$

Splitting to Minimize Error I

- Aim: Improve the accuracy of our θ-estimates by partitioning P into C_1, C_2.
- We seek to minimize:

$$\text{err}(C_1, C_2) = \sum_{j=1,2} \mathbb{P}[X \in C_j | X \in P] \; \mathbb{E}[(\hat{\theta}_{C_j}(J) - \theta(X))^2 | X \in C_j],$$

where $\hat{\theta}_{C_j}(J)$ are estimates fit over children C_j.

Splitting to Minimize Error II

- Standard **regression tree methods** (e.g., **CART** by Breiman et al., 1984) minimize **in-sample error**, using plug-in estimators.
- For binary treatment effects, **Athey and Imbens (2016)** propose **unbiased estimates** for $err(C_1, C_2)$ with **overfitting penalties**, building on **Mallows (1973)**.
- Where $\theta(x)$ is defined by a **moment condition**, direct loss minimization is infeasible, precluding unbiased estimates of $err(C_1, C_2)$.
- An **abstract characterization** addresses this issue when direct minimization isn't possible.

Challenge of Direct Loss Minimization

- Direct loss minimization is not viable when $\theta(x)$ is identified through moment conditions.
- To address this, we consider Proposition 1, which leads to a more abstract criterion.
- The criterion $\Delta(C_1, C_2)$ is defined as:

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

where n_{C_j} is the number of observations in child C_j.

Combining Δ-Criterion and Implications

- Proposition 1 suggests maximizing $\Delta(C_1, C_2)$ for better splits, enhancing the heterogeneity of in-sample θ-estimates.
- This criterion is designed to address bias from sampling variance, similar to Athey and Imbens (2016), and aims to stabilize tree construction.
- By integrating this error term into the splitting strategy, we refine traditional methods, systematically increasing θ-estimate heterogeneity.
- As a result, the Δ-criterion leads to a more robust model with improved estimation accuracy.

Challenges in Optimization

Optimizing the criterion:

- Directly optimizing the criterion $\Delta(C1, C2)$ by solving for $\hat{\theta}_{C1}$ and $\hat{\theta}_{C2}$ can be computationally expensive.
- To alleviate computational costs, an approximate criterion $\bar{\Delta}(C1, C2)$ is used.

Gradient-Based Approximations

- Gradient-Based Estimates: Approximations $\tilde{\theta}_C$ for true values $\hat{\theta}_C$ leverage gradient information.
- Consistent Estimation: A_P consistently estimates the gradient of the expected ψ-function.
- Approximation Formula:

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

- Differentiable ψ-Function:

$$A_P = \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i:X_i \in P\}} \nabla \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \tag{3}$$

- The term $\xi^\top A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i)$ represents the influence function for the ith observation in computing $\hat{\theta}_P$.
- Nondifferentiable ψ: Special consideration is needed, such as with quantile regression.

Recursive Partitioning Algorithm Steps

- **Labeling Step:**
  - Compute estimates $\hat{\theta}_P$, $\hat{\nu}_P$, and the derivative matrix A_P^{-1} using parent data.
  - Obtain pseudo-outcomes ρ_i using the formula:

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

- **Regression Step:**
  - Perform a standard CART regression to split on ρ_i.
  - Split parent P into two axis-aligned children $C1$ and $C2$ to maximize the criterion:

$$\Delta(C1, C2) = \sum_{j=1}^{2} \left(\frac{1}{|\{i : X_i \in Cj\}|} \sum_{\{i : X_i \in Cj\}} \rho_i \right)^2$$

- After the regression split, relabel observations in each child node by solving the estimating equation and proceed recursively.

Efficiency of Gradient-Based Tree Splitting

- **Consistent Performance**
  - More consistent than direct optimization at each split.
  - Relies on gradient-based estimates.
- **Computational Efficiency**
  - **Split-Selection**: Dominant computational step in tree growth.
  - **Labeling Step**: Minimal impact, performed once per node.
  - **Regression Criterion**: Enables single data pass evaluation.
- **Complexity Management**
  - Streamlines split process, reducing optimization complexity.
  - Enhances overall algorithm efficiency.
- **Theoretical Foundations and Connections**
  - Basis for algorithms like gradient boosting and recursive partitioning.
  - Analogous to change-point detection in statistical models.
  - Aligns with moment-based change-point methods in literature.

Theoretical Justification

- The approximation error of the criterion $\bar{\Delta}(C1, C2)$ as opposed to the exact criterion is within an acceptable tolerance.
- This suggests that using the gradient-based approach for splitting does not introduce significant inefficiency.
- Proposition 2 ensures that under certain conditions, the approximate and exact criteria are equivalent up to a small order probability term.

Introduction to Asymptotic Analysis

- We aim to establish asymptotic Gaussianity of $\hat{\theta}(x)$ for GRFs.
- Focus on providing tools for statistical inference about $\theta(x)$.
- Covariate space $X = [0,1]^p$ and parameter space $(\theta, \nu) \in B \subset \mathbb{R}^k$.
- Features X have a density bounded away from 0 and ∞ (weak dependence).
- Primary Goal: Formal characterization underpinning theoretical results.

Key Assumptions

- Assumption 1: Lipschitz x-signal for $M_{\theta,\nu}(x)$.
- Assumption 2: Smooth identification with twice continuously differentiable M-function.
- Assumption 3: Lipschitz (θ, ν)-variogram for score functions $\psi_{\theta,\nu}$.
- Assumption 4: Regularity of ψ-functions, allowing empirical processes to be Donsker.
- Assumption 5: Existence of solutions to the estimating equation.
- Assumption 6: Convexity of the score function and strong convexity of expected score.

Tree Specifications and Asymptotic Gaussianity

- Trees must be honest, regular, and symmetric for theoretical results.
- Minimum split probability condition with a random number of variables considered.
- Specification 1 requires trees to:
  - Make balanced splits with a minimum fraction ω.
  - Randomize splits with a minimum probability π.
  - Use subsampling size s such that $s/n \to 0$ and $s \to \infty$.
- Under these settings, $\hat{\theta}(x)$ is consistent and asymptotically Gaussian.
- Applicability to various cases like instrumental variables regression and quantile regression.

Least Squares Regression in GRFs

Consider the least squares regression setting where $\psi_\theta(Y_i) = Y_i - \theta$.
Assumptions 26 are satisfied:

- Assumption 2: Variance $V = 1$.
- Assumption 3: Variogram $\gamma(\theta, \theta') = 0$.
- Assumption 4: ψ is Lipschitz.
- Assumption 6: Score function $\psi_\theta(y) = -\frac{d}{d\theta}\frac{(y-\theta)^2}{2}$.

Assumption 1 requires that the conditional mean function $E[Y_i|X_i = x]$ is Lipschitz in x, a common regression forest assumption.

Example 2: Quantile Regression

- **Definition**: Quantile regression characterizes the relationship between variables by estimating quantiles of the conditional distribution.
  - $\psi_\theta(Y_i) = q - \mathbb{1}\{Y_i \leq \theta\}$
  - $M_\theta(x) = q - F_x(\theta)$, where $F_x(\cdot)$ is the CDF of $Y_i|X_i = x$
- **Assumptions for Quantile Regression**:
  - Assumption 1: Conditional exceedance probabilities $P[Y_i > y|X_i = x]$ are Lipschitz-continuous in x.
  - Assumption 2: $f_x(y)$ has a continuously bounded first derivative, not zero at quantile $y = F_x^{-1}(q)$.
  - Assumption 3: $f_x(y)$ is uniformly bounded from above.
  - Assumption 4: Monotonicity of ψ and univariate $O_i = Y_i$.
  - Assumption 5: Directly satisfied by the nature of quantile regression.
  - Assumption 6: Negative subgradient of V-shaped function due to $f_x(\theta) > 0$.
- **Statistical Inference**: These assumptions allow for the use of quantile regression in statistical inference, providing a robust alternative to mean regression.

Application to Quantile Regression

- Quantile regression forests (QRFs) address nonparametric quantile regression.
- Pioneered by Meinshausen (2006), who proposed a consistent forest-based algorithm.
- QRFs use random forest weights for solving estimating equations.
- Unlike Meinshausen, we introduce a quantile-specific splitting rule.
- Our method emphasizes heterogeneity in conditional quantiles.

Estimating Equations and Splitting Rules

- The qth quantile $\theta_q(x)$ of $Y|X = x$ is identified via:

$$\psi_\theta(Y_i) = q1\{Y_i > \theta\} - (1-q)1\{Y_i \leq \theta\}$$

- Our splitting scheme adapts to this moment function.
- Pseudo-outcomes ρ_i guide the tree to split based on the qth quantile.
- Gradient-based trees separate observations around the parent node's quantile.

Comparative Evaluation

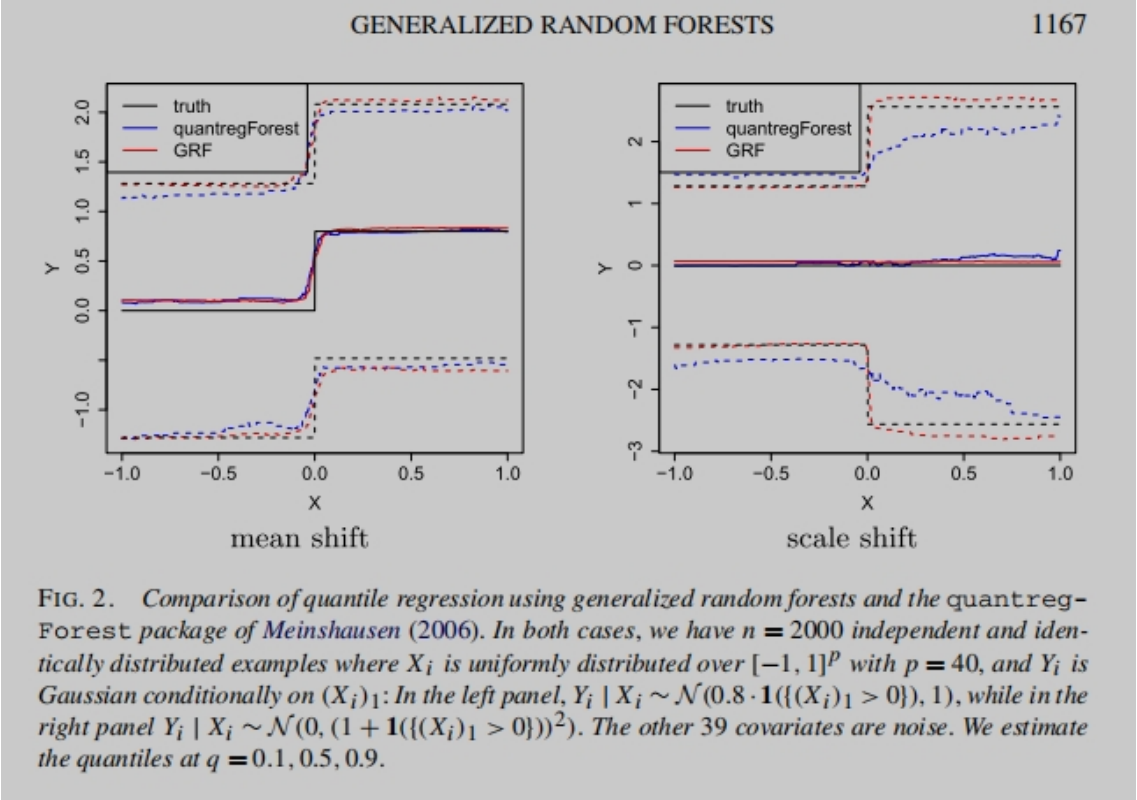

FIG. 2. *Comparison of quantile regression using generalized random forests and the* `quantregForest` *package of Meinshausen* (2006). *In both cases, we have* $n = 2000$ *independent and identically distributed examples where* X_i *is uniformly distributed over* $[-1, 1]^p$ *with* $p = 40$, *and* Y_i *is Gaussian conditionally on* $(X_i)_1$: *In the left panel,* $Y_i \mid X_i \sim \mathcal{N}(0.8 \cdot \mathbf{1}(\{(X_i)_1 > 0\}), 1)$, *while in the right panel* $Y_i \mid X_i \sim \mathcal{N}(0, (1 + \mathbf{1}(\{(X_i)_1 > 0\}))^2)$. *The other* 39 *covariates are noise. We estimate the quantiles at* $q = 0.1, 0.5, 0.9$.

Figure: Figure 2: Comparison of QRF methods. Left: Mean shift detection. Right: Scale shift detection.

- Both methods capture mean shifts in conditional distributions.
- Our method also detects scale shifts, unlike plain CART regression splits.
- Our QRFs are robust to changes in conditional quantile distributions.

Assessing Performance Differences

- Our method shows smoother sample paths, attributed to honesty as per Section 2.4.
- Without honesty, our QRFs still identify jumps but with oscillations similar to Meinshausen's.
- Our QRFs focus on quantile shifts, offering a nuanced view compared to regression splits.
- Generalized random forests demonstrate sensitivity to quantile-specific dynamics.

References I


Athey, S., Tibshirani, J., and Wager, S. (2019).
Generalized random forests.