Interaction Trees for Exploring Stratified & Individualized Treatment Effects

Xiaogang Su

Department of Mathematical Sciences
University of Texas at El Paso (UTEP)

March 25, 2014 @ NUS - IMS
School and Workshop on Classification and Regression Trees

Outline

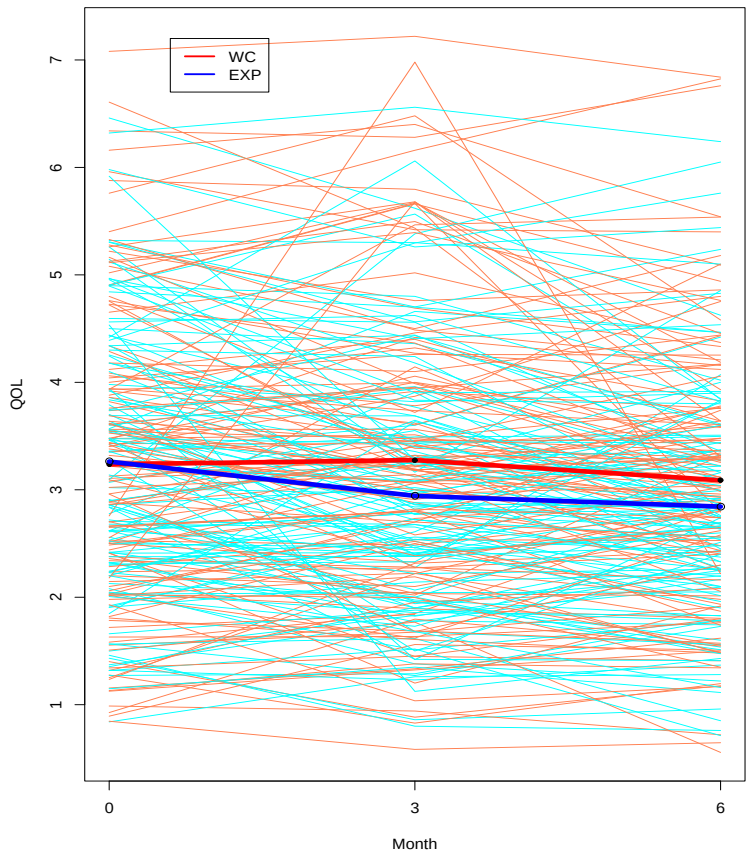
The BCEI Study

The Breast Cancer Education Intervention (BCEI) study (Meneses et al., 2007, ONF) is a randomized controlled longitudinal psycho-educational support intervention trial on quality of life (QoL) targeting women with early-stage breast cancer survivors in the first year of post-treatment survivorship.

- Founded by NIH (R01) and initialized in 2001;
- 261 BCS's were randomized into the experimental (Exp) and the wait control (WC) groups and followed at baseline, Month 3, and Month 6;
- Four subjects in Exp dropped out and one died in WC in the followup period. 125 in Exp and 131 in WC completed the study.

Effectiveness of BCEI on QOL

- The outcome variable, Quality of Life (QoL), is obtained from a 50-item instrument with four subdomains: Physical, Psychological, Social, and Spiritual.
- Each item scores on a 0-10 rating scale, with lower scores indicating better QoL. The overall QoL score is the grand average.
- The effectiveness of BECI is found statistically significant. P-values are $< .0001$ with and without covariate adjustment.



Follow-Up Questions and Studies

- Are there BCS subpopulations where BECI is most (or less) helpful? If so, how are they characterized?
- What variables are effect-modifiers of BECI? Does qualitative interaction possibly exist?
- Given a BCS, what difference could BECI make?
- How do we develop the optimal treatment regimes?
- BCEI is followed by two other R01 projects: Rural Breast Cancer Study (RBCS) and cost-effectiveness analysis (CEA).

For simplicity (independent data and BCEI is most effective at Month 3), all the illustrations in this presentation are based on the change score from Month 3 to Baseline.

Tree-Structured Methods

- The follow-up questions are all related to Stratified & Individualized Treatment Effects, essentially involving treatment-by-covariates interactions.
- **Relevant Concepts**: effect moderation or modification, subgroup analysis, qualitative and quantitative interaction, treatment regime, etc.
- Rubin's causal model (Neyman 1990; Rubin, 1978) provides a fine calibration of causal effects and a general framework for making causal inference.

Potential Outcomes

- *Potential outcomes* $Y_0(\omega)$ and $Y_1(\omega)$ denote the responses that would have been observed if unit ω were assigned to the control group (or the treatment group);
- Either $Y_0(\omega)$ or $Y_1(\omega)$, but not both, can actually be observed depending on the value of $T(\omega)$, an inherent fact called the *fundamental problem of causal inference* (Holland 1986).
- Thus the observed outcome is $Y(\omega) = \{1 - T(\omega)\}Y_0(\omega) + T(\omega)Y_1(\omega)$.
- Available data consist of i.i.d. realizations of $\{Y, T, \mathbf{X}\}$: $\{(y_i, t_i, \mathbf{x}_i) = (y(\omega_i), t(\omega_i), \mathbf{x}(\omega_i)) : i = 1, \ldots, n\}$.

Causal Effect at Different Levels

- Causal inference is concerned with the comparison of the two potential outcomes via the observed data, which can be made at three levels.
  1. *Unit-Level*: $Y_1(\omega) - Y_0(\omega)$.
  2. *Subpopulation-Level*: $\{\omega : \mathbf{X}(\omega) \in A \subset \mathcal{X}\}$:

$$E(Y_1|\mathbf{X} \in A) - E(Y_0|\mathbf{X} \in A).$$

  3. *Population-Level*: $E(Y_1) - E(Y_0)$, called the 'Average Treatment Effect' (ATE).
- These three levels are ordered by decreasing strength. The vast majority of causal inference literature is centered on estimation of ATE.

Individual Treatment Effects (ITE)

- We define "*individual treatment effect*" (ITE) as a conditional expectation $E(Y_1 - Y_0|\mathbf{x})$, given a subject with $\mathbf{X} = \mathbf{x}$.
- ITE is conceptually different from the unit level causal effect $Y_1(\omega) - Y_0(w)$. Strictly speaking, ICE makes conditional causal inference at the subpopulation level $\{\omega : \mathbf{X}(\omega) \in A\}$ with $A = \{\mathbf{x}\}$.
- ITE is the best that one could practically do with available information in order to approximate the unit level causal effect.

Why Tree-Structured Methods?

- A tree model fits piecewise constant models by recursively bisecting the predictor space. It starts simply with a two-sample test statistic but facilitates a comprehensive modeling by recursive partitioning.
- Excels at modeling complex interactions of higher orders (albeit implicitly). Tree models provide a natural way of grouping data.
- Interaction trees (Su et al., 2009 *JMLR*) supplies inference on stratified or subpopulation treatment effects. Then we can move backward to ACE by integrating results or move forward to ICE by ensemble models.

Overview of Interaction Trees (IT) Features

Stratified Treatment Effects

- Single Interaction Tree Analysis
  - Growing a large initial tree; ★
  - Pruning;
  - Tree size selection via validation;
  - Amalgamation;
- Aggregated grouping; ★

Individualized Treatment Effects

- Estimating ITE via Random Forests of Interaction Trees; ★
- Variable importance ranking; ★
- Partial Dependence Plots; ★
- Exploring qualitative interaction;
- Estimating optimal treatment regime.

Stratified Treatment Effects - Subgroup Analysis

Goal: to seek sub-populations that show differential treatment effects

Pros

- Maximum use of available data;
- Deeper insight into the treatment effects;
- Generating new research hypotheses or refining inclusion/exclusion criteria.
- etc.

Cons

- Multiplicity (Type I error);
- Lack of power with reduced sample size (Type II error);
- Complex treatment-by-covariate interactions;
- Pre-planned vs. *post hoc*.
- etc.

The Set-Up

- For simple illustration, consider independent data $\{(y_i, T_i, \mathbf{x}_i) : i = 1, \ldots, n\}$:
  - y_i is the ith response;
  - T_i is the binary treatment indicator: 1- treated and 0 – control;
  - $\mathbf{x}_i \in \mathbb{R}^p$ is a p-dimensional covariate vector.
- Let s denote a split on predictor X_j with cutoff point c, e.g., $X_j \leq c$? if X_j is continuous.
- Each split s induces a 2×2 table as below:

	Child Nodes	
Trt	Left (1)	Right (0)
1	$(\bar{y}_{11},\ n_{11})$	$(\bar{y}_{10},\ n_{10})$
0	$(\bar{y}_{01},\ n_{01})$	$(\bar{y}_{00},\ n_{00})$

The Splitting Statistic

- The t or z test for differential treatment effects between two child nodes amounts to

$$z(X_j; c) = \frac{(\bar{y}_{11} - \bar{y}_{01}) - (\bar{y}_{10} - \bar{y}_{00})}{\sqrt{\hat{\sigma}^2(1/n_{11} + 1/n_{01} + 1/n_{10} + 1/n_{00})}},$$

- Note that $z(X_j; c)$ is the t test for $H_0 : \beta_3 = 0$ in the linear model:

$$y_i = \beta 0 + \beta_1 T_i + \beta_2 \Delta_{ij} + \beta_3 T_i \cdot \Delta_{ij} + \varepsilon_i,$$

where $\Delta_{ij} = 1_{\{x_{ij} \leq c\}}$ is the indicator associated with split s.

Exhaustive/Greedy Search

- The best split $s^\star$ is

$$s^\star = \arg\max_{X_j;\, c} z^2(X_j, c) = \arg\max_{X_j} z^2(X_j, c_j^\star).$$

- It can be viewed as a two-step search.
  1. For each X_j, find its best cutoff point and obtain $z^2(X_j, c_j^\star)$.
  2. Compare $z^2(X_j, c_j^\star)$ across X_j's for $j = 1, \ldots, p$.

Alternative: Approximation via a Sigmoid Function

- Fix X_j. Let $\Delta_{ij} = 1_{\{x_{ij} \leq c\}}$ For $l = 1, 0$ and $t = 1, 0$,

$$
\begin{aligned}
n_{lt} &= \sum_{i=1}^{n} T_i^l (1 - T_i)^{1-l} \Delta_{ij}^t (1 - \Delta_{ij})^{1-t} \\
\bar{y}_{lt} &= \sum_{i} y_i T_i^l (1 - T_i)^{1-l} \Delta_{ij}^t (1 - \Delta_{ij})^{1-t} / n_{lt} \\
\hat{\sigma}^2 &= \sum_{i=1}^{n} y_i^2 - \sum_{k,l=0,1} n_{kl} \bar{y}_{kl}^2
\end{aligned}
$$

- Replace $\Delta_{ij} = 1_{\{x_{ij} \leq c\}}$ with a smooth sigmoid function, e.g., $\pi\{a \cdot (x_{ij} - c)\}$ with $\pi(x) = \{1 + \exp(-x)\}^{-1}$.

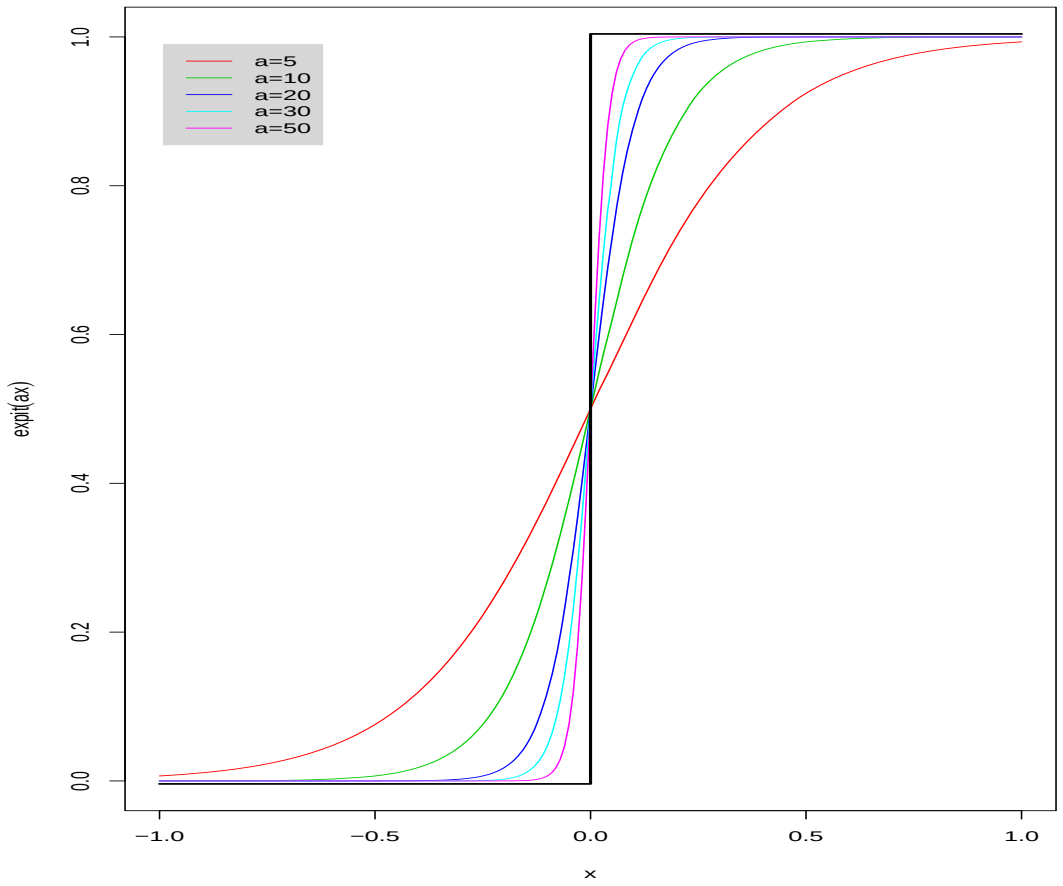


Figure : The expit function $\pi(x) = \{1 + \exp(-a(x - c)\}^{-1}$ with $c = 0$ and different a values.

Smooth Threshold Functions

- With fixed X_j, solving $c_j^\star = \arg\max_c z^2(X_j; c)$ becomes a one-dimensional smooth optimization problem.
- To avoid the end-cut preference problem, optimization with bound constraints, e.g., $(q_{.2}, q_{.98})$, is helpful.
- In some scenarios, the search for $c_j^\star$ can be casted into a model fitting setting and the involved optimization reduces to a separable nonlinear least squares problem.
- Besides improved computational efficiency, this approach can help address the *variable selection bias* problem with recursive partitioning. ♦

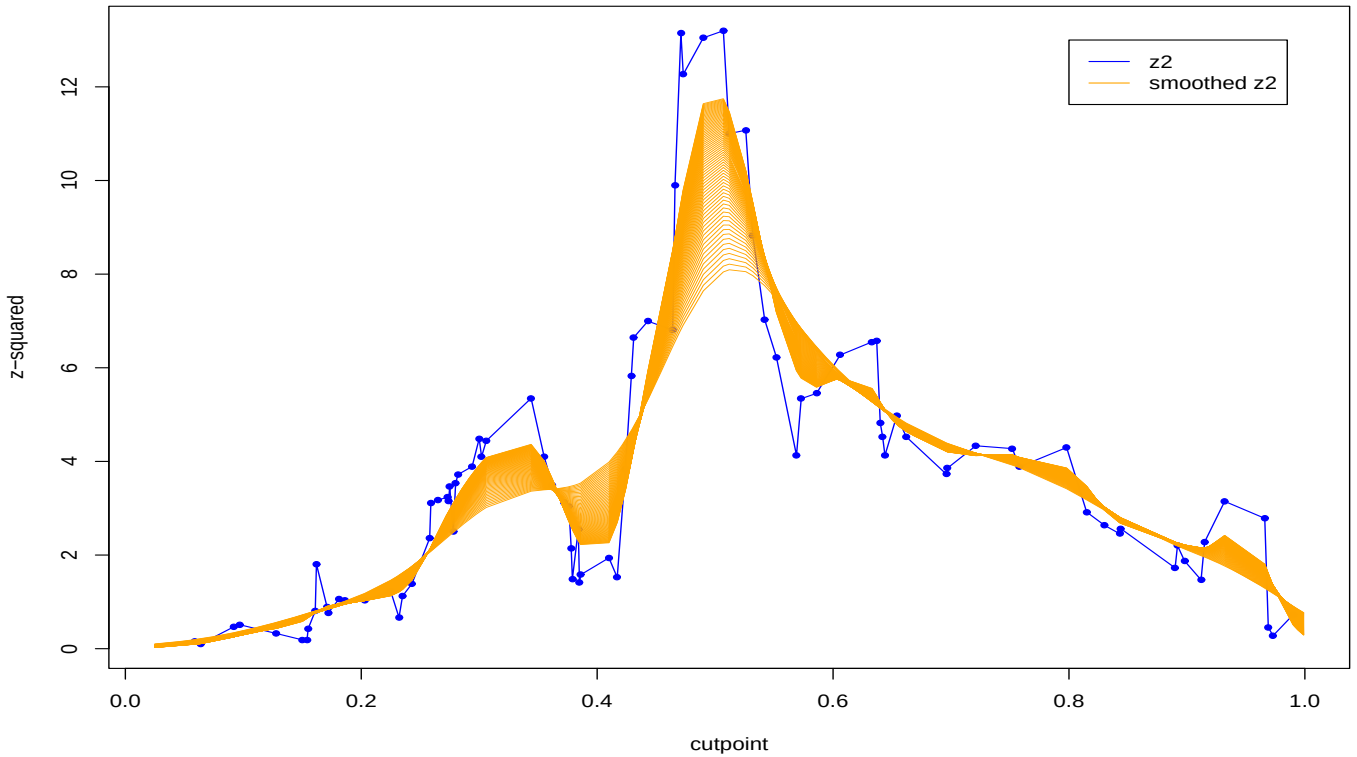


Figure : Smoothed t^2 with $a \in \{30, \ldots, 100\}$, compared to t^2. Data ($n = 100$) were generated from Model: $y = 1 + T + z + T \cdot z + \varepsilon$ with $z = 1\{x \leq .5$ and $\varepsilon \sim N(0, 1)$.

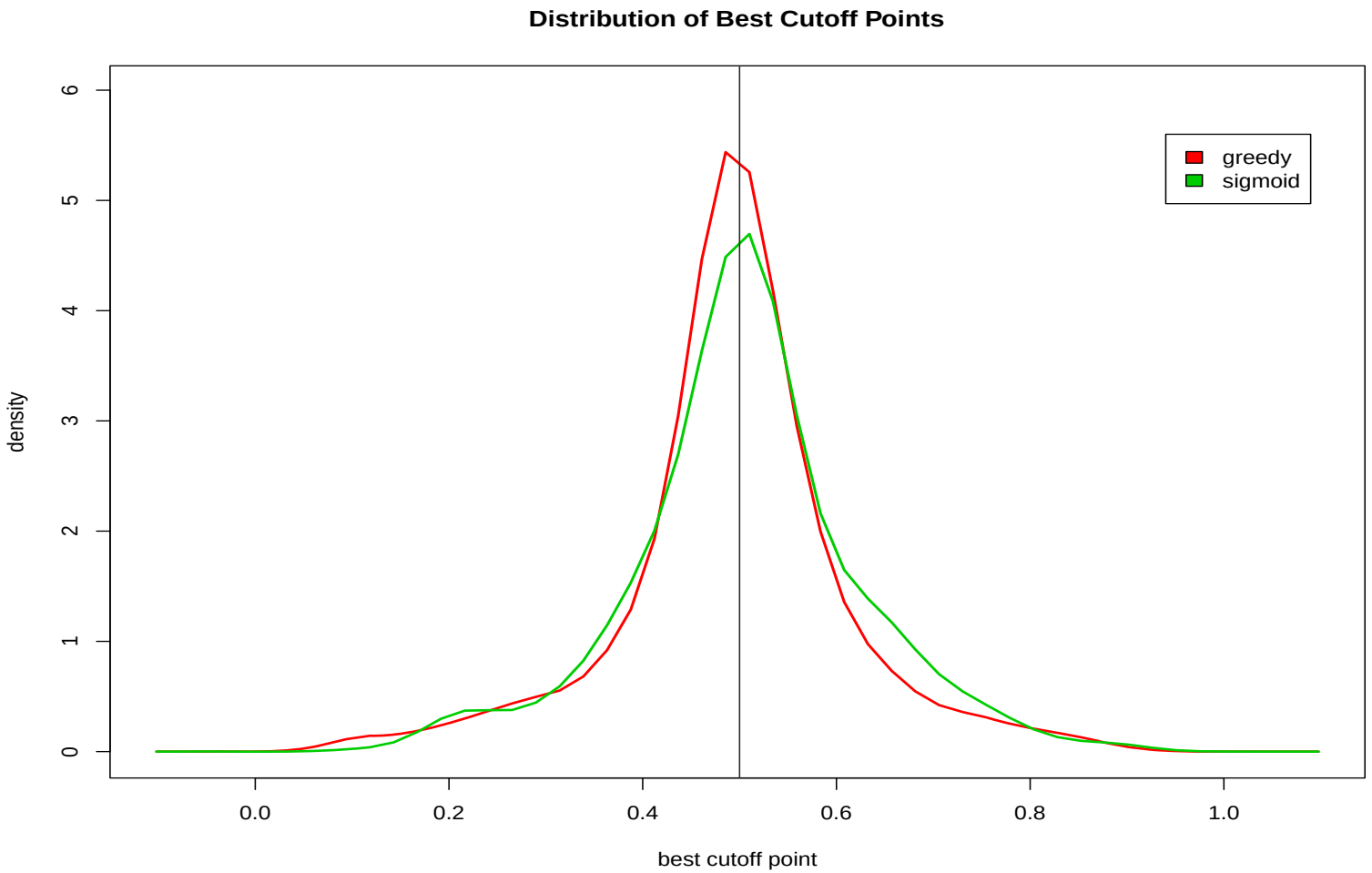
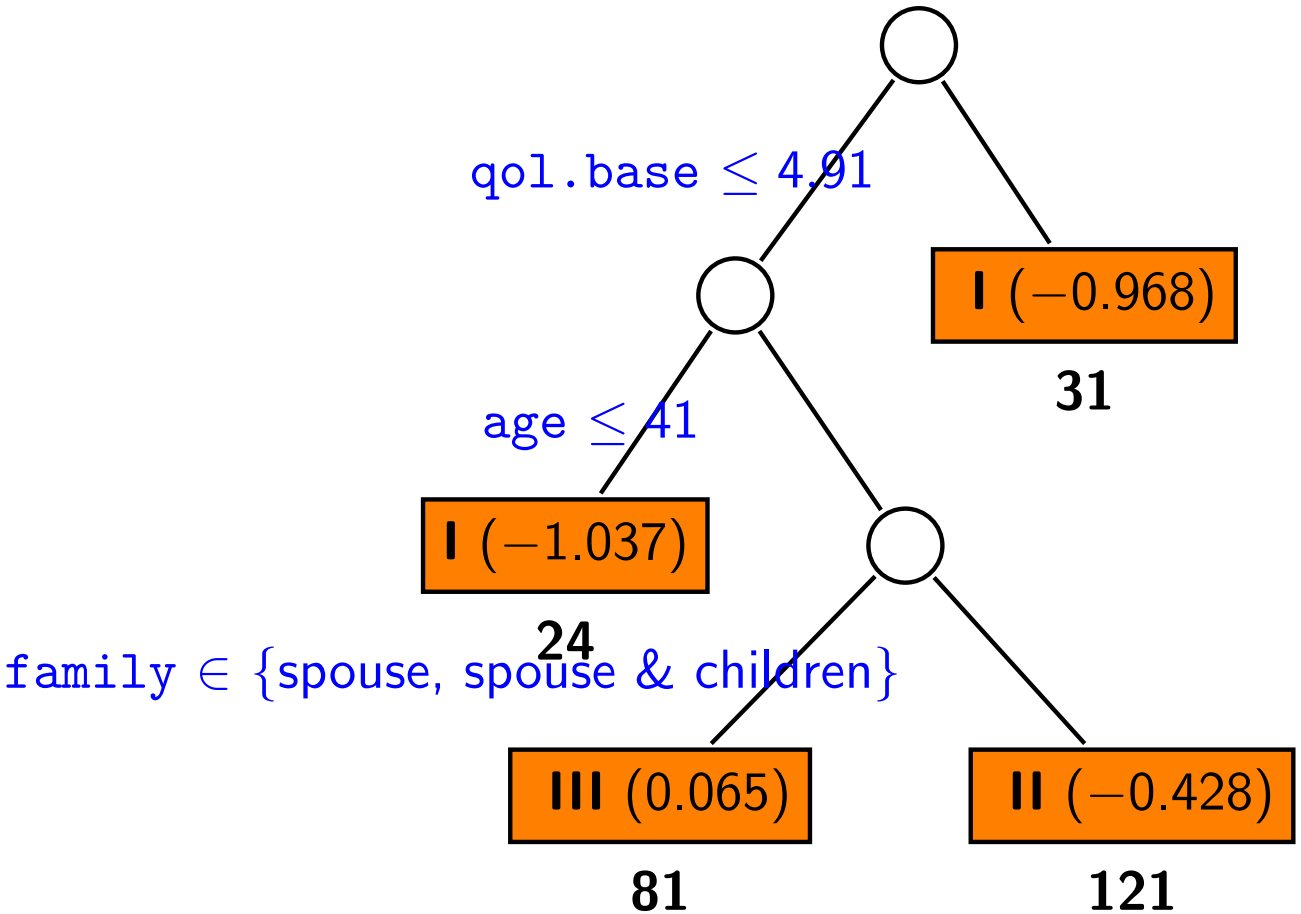


Figure : Distribution of the best cutoff points: 500 simulation runs.

Figure: The IT Structure for BCEI Data.



Aggregated Grouping

- Growing B trees by taking bootstrap samples and apply each tree to the whole data $\mathcal{L}$;
- For each tree $\mathcal{T}_b$, let $t(i)$ denotes the terminal node the ith observation falls into. For any pair of observations (i, i'), define a distance or proximity measure $d_{ii'}^{(b)}$ such that

$$d_{ii'}^{(b)} = \begin{cases} 0 & \text{if } t(i) = t(i'); \\ -\log_{10}(p_{ii'}) & \text{if } t(i) \neq t(i') \end{cases}$$

where $p_{ii'}$ is the p-value from a two-sample statistical test that compares $t(i)$ and $t(i')$.

A Distance Matrix

- Let q be the number of terminal nodes in $\mathcal{T}_b$. Introduce an $n \times q$ (incidence) matrix $\mathbf{A}_b = (a_{it})$ such that $a_{it} = 0$ if observation i falls into terminal node t of $\mathcal{T}_b$. Let $\mathbf{B}_b = (-\log_{10} p_{ii'})$ be the $q \times q$ distance matrix among the q terminal nodes of tree $\mathcal{T}_b$. Then it follows that

$$\mathbf{D}_b = (d_{ii'}^{(b)}) = \mathbf{A}_b \mathbf{B}_b \mathbf{A}_b^t.$$

- In ordinary random forests, $\mathbf{B}_b = \mathbf{J} - \mathbf{I}$, where $\mathbf{J}$ is the $q \times q$ matrix of all 1's and $\mathbf{I}$ is the unit matrix. Thus $d_{ii'}^{(b)} = \sum_{t=1}^{q} a_{it} a_{i't} = 1$ if the i-th and i'-th subjects fall into different terminal nodes; and 0 otherwise.

Forming Groups via Clustering

- Average the distances obtained from B trees: $d_{ii'} = \sum_{b=1}^{B} d_{ii'}^{(b)}$. Then $\mathbf{D} = (d_{ii'})$ is the $n \times n$ distance matrix for all n subjects in terms of heterogeneity of treatment effects.
- Entries in the distance matrix $\mathbf{D}$ measure how two subjects are different in terms of treatment effects.
- Apply clustering to determine the final grouping.

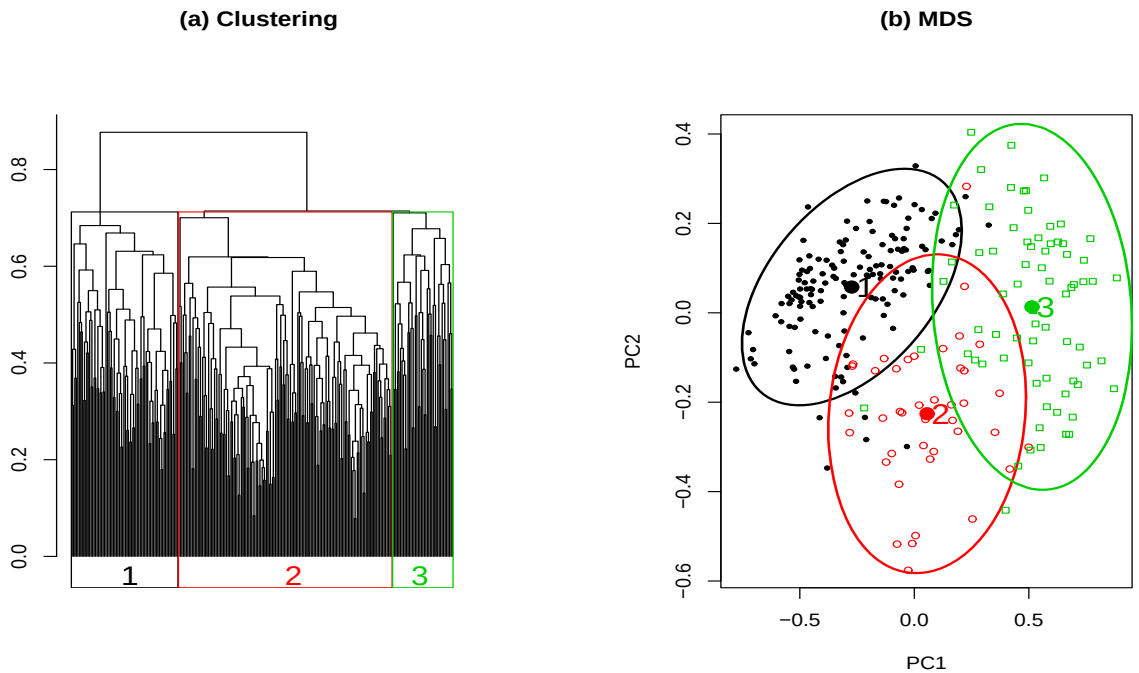


Figure : Aggregated Grouping: Cluster analysis and MDS based on the distance matrix obtained from bagging interaction trees.

Results from Aggregated Grouping

	Exp		WC		Intervention Effect	
Subgroup	Size	Mean	Size	Mean	δ	p-value
I	30	-1.052	42	0.266	-1.318	$< .0001$
II	20	-0.264	21	-0.072	-0.192	.2891
III	75	-0.046	69	-0.038	-0.008	.9424

Estimating ITE

Estimated ITE $\delta = E(Y_1 - Y_0|\mathbf{X} = \mathbf{x})$ can be useful in various ways. Refer to the following data layout for a missing data problem:

id	T	$\mathbf{x}$	y	Y_1	Y_0
1	0	$\mathbf{x}_1$	y_1	·	Y_{01}
2	0	$\mathbf{x}_1$	y_1	·	Y_{01}
	.....				
n_0	0	$\mathbf{x}_{n_0}$	y_{n_0}	·	Y_{0n_0}
n_0+1	1	$\mathbf{x}_{n_0+1}$	y_{n_0+1}	$Y_{1(n_0+1)}$	·
n_0+2	1	$\mathbf{x}_{n_0+2}$	y_{n_0+2}	$Y_{1(n_0+2)}$	·
	.....				
n_0+n_1	1	$\mathbf{x}_{n_0+n_1}$	$y_{n_0+n_1}$	$Y_{1(n_0+n_1)}$	·

Separate Regression for Estimating ITE

Regression Y on $\mathbf{x}$ with data in the treatment group (trt=1) and then use the fitted model to predict Y_1 in the control group; similarly, build another model using data in the control group and make prediction for Y_0 in the treatment group.

id	trt	$\mathbf{x}$	y	Y_1	Y_0
1	0	$\mathbf{x}_1$	y_1	$\hat{Y}_{11}$	Y_{01}
2	0	$\mathbf{x}_1$	y_1	$\hat{Y}_{12}$	Y_{02}
	.....				
n_0	0	$\mathbf{x}_{n_0}$	y_{n_0}	$\hat{Y}_{1n_0}$	Y_{0n_0}
n_0+1	1	$\mathbf{x}_{n_0+1}$	y_{n_0+1}	$Y_{1(n_0+1)}$	$\hat{Y}_{0(n_0+1)}$
n_0+2	1	$\mathbf{x}_{n_0+2}$	y_{n_0+2}	$Y_{1(n_0+2)}$	$\hat{Y}_{0(n_0+2)}$
	.....				
n_0+n_1	1	$\mathbf{x}_{n_0+n_1}$	$y_{n_0+n_1}$	$Y_{1(n_0+n_1)}$	$\hat{Y}_{0(n_0+n_1)}$

Methods for Estimating ITE

- With **Separate Regression (SR)**, there are two ways to compute δ depending on availability of observed response.
  - Method **I**: Given a new subject with **x** only, $\delta = \hat{Y}_1 - \hat{Y}_0$. (Bias Problem)
  - Method **II**: If either Y_1 or Y_0 is available, then $\delta = Y_1 - \hat{Y}_0$ when $T = 1$ and $\delta = \hat{Y}_1 - Y_0$ when $T = 0$. (Large Variance)
- **Random Forests of Interaction Trees (RF-IT)**: Divide data into groups (terminal nodes) where treatment effects are homogeneous and compute $\bar{Y}_{1t} - \bar{Y}_{0t}$ for terminal node t; aggregate results via random forests.

One Simulated Example

- Setting: $\mu_k(\mathbf{x}) = E(Y_k|\mathbf{X} = \mathbf{x})$ and $Y_k = \mu_k(\mathbf{x}) + \varepsilon + \varepsilon_k$ for $k = 0, 1$, where $\varepsilon \sim N(0, 1)$ and $\varepsilon_k \sim N(0, 1)$ independently. ICE $\delta(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$.
- Example: set

$$\mu_0(\mathbf{x}) = 2 + 2\,x_1 + 2\,x_2 + 2\,x_3$$

and

$$\delta(\mathbf{x}) = 0.1\exp(4\,x_1) + 4\,\text{expit}\{20\cdot(x_2 - 0.5)\} + 3\,x_3 + 2\,x_4 + x_5.$$

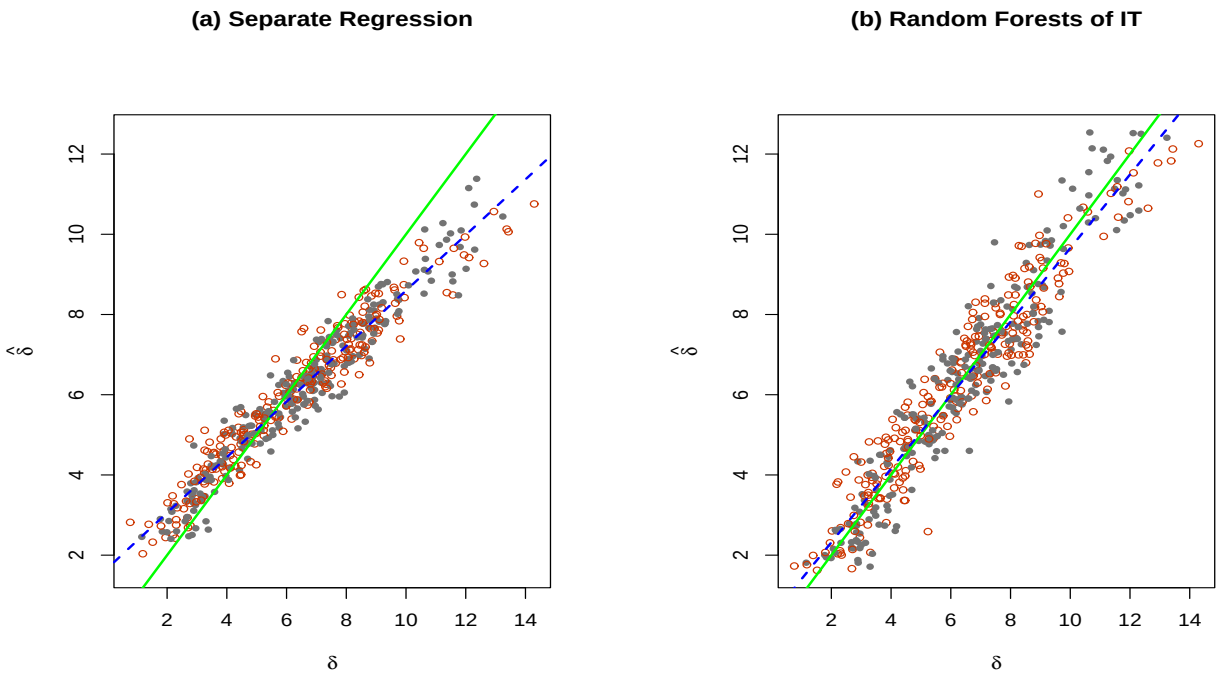


Figure : Comparison of Separate Regression I vs. Random Forests of Interaction Trees in Predicting ITE: A Simulated Example.

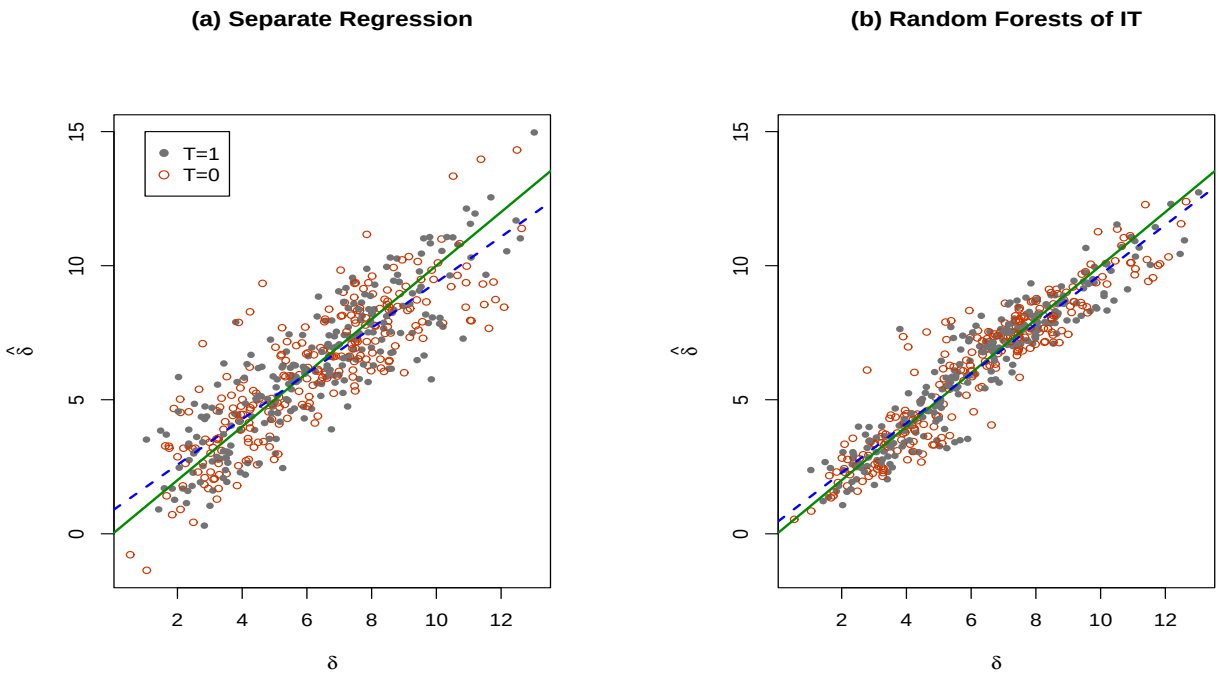


Figure : Comparison of Separate Regression II vs. Random Forests of Interaction Trees in Predicting ITE: A Simulated Example.

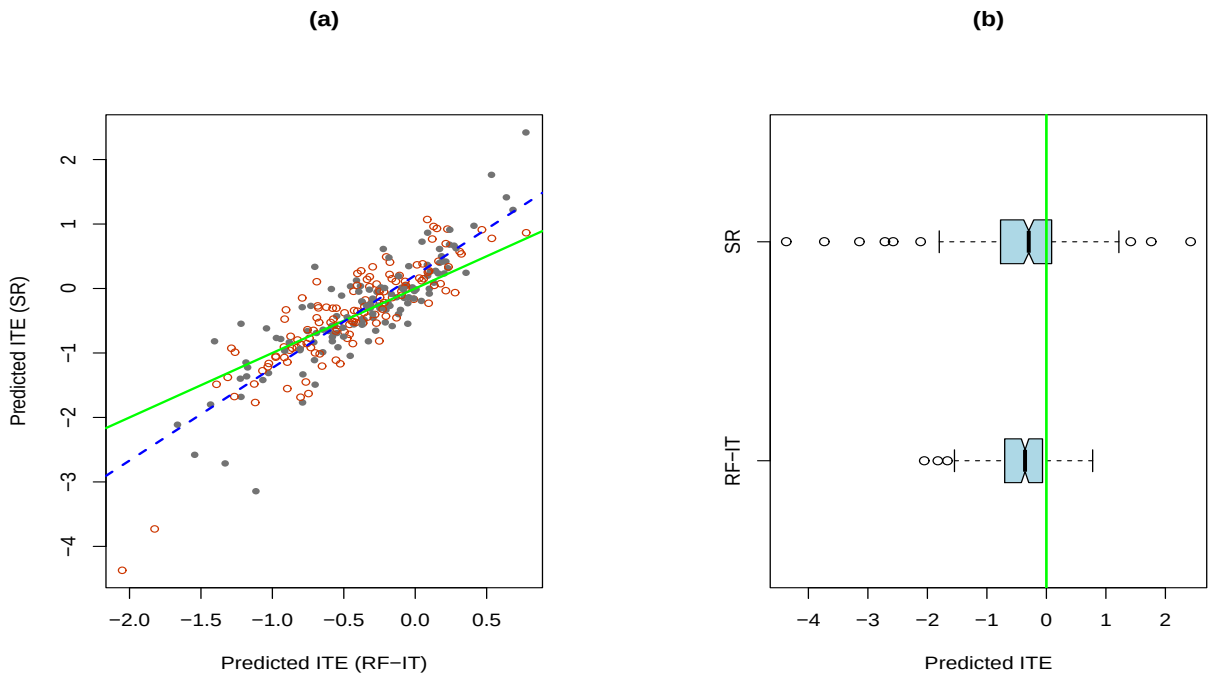


Figure : Comparison of Separate Regression II vs. Random Forests of Interaction Trees in Predicting ITE: The BCEI Data.

Algorithm: Variable Importance

Initialize all V_j's to 0 and Set m.
For $b = 1, 2, \ldots, B$, do

- Obtain bootstrap sample $\mathcal{L}_b$ and the out-of-bag sample $\mathcal{L}_b^{(c)} = \mathcal{L} - \mathcal{L}_b$.
- Based on $\mathcal{L}_b$, grow a large IT tree $\mathcal{T}_b$ by searching over m randomly selected covariates at each split.
- Send $\mathcal{L} - \mathcal{L}_b$ down T_b to compute $G(\mathcal{T}_b)$.
- For each covariate X_j, $j = 1, \ldots, p$, do
  - Permute the values of X_j in $\mathcal{L}_b^{(c)}$;
  - Send the permuted $\mathcal{L}_b^{(c)}$ down $\mathcal{T}_b$ to compute $G_j(\mathcal{T}_b)$.
  - Compute $\Delta V_j = \dfrac{G(\mathcal{T}_b) - G_j(\mathcal{T}_b)}{G(\mathcal{T}_b)}$ if $G(\mathcal{T}_b) > G_j(\mathcal{T}_b)$; and 0 otherwise.
  - Update $V_j \leftarrow V_j + \Delta V_j$.

Average $V_j \leftarrow V_j / B$.

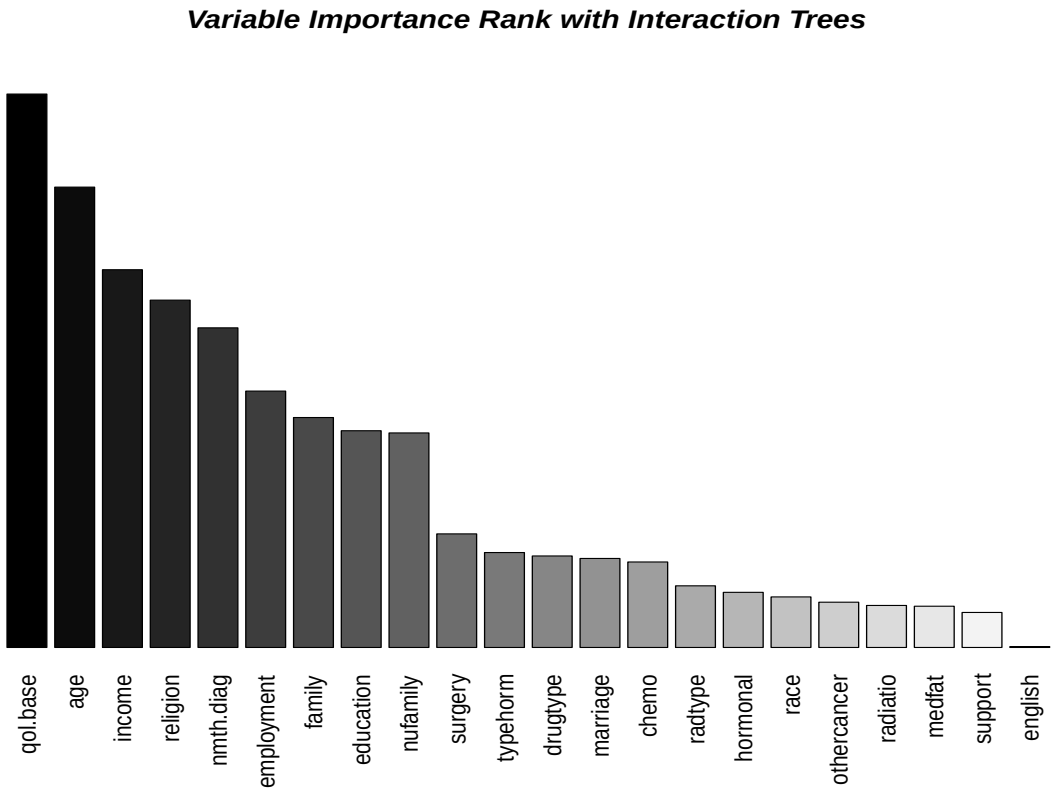


Figure : Variable Importance from Random Forests of Interaction Trees: The BCEI Data.

Partial Dependence Plot

- First proposed by Friedman (1991, *Annals of Statistics*); implemented in R packages `randomForests` and others. Can be naturally extended to interaction trees:

$$f_j(x_j) = E_{\mathbf{x}_{(-j)}} \delta(\mathbf{x}), \text{ for } j = 1, \ldots, p.$$

- To estimate, we compute $\tilde{\delta}(x)$ for a number of values of x and then plot $\tilde{\delta}(x)$ versus x.

$$\begin{aligned} \tilde{\delta}(x_j) &= \frac{1}{n} \sum_{i=1}^{n} \delta(x_j, \mathbf{x}_{i(-j)}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \bar{Y}(x_j, T = 1, \mathbf{x}_{i(-j)}) - \bar{Y}(x_j, T = 0, \mathbf{x}_{i(-j)}) \right\}, \end{aligned}$$

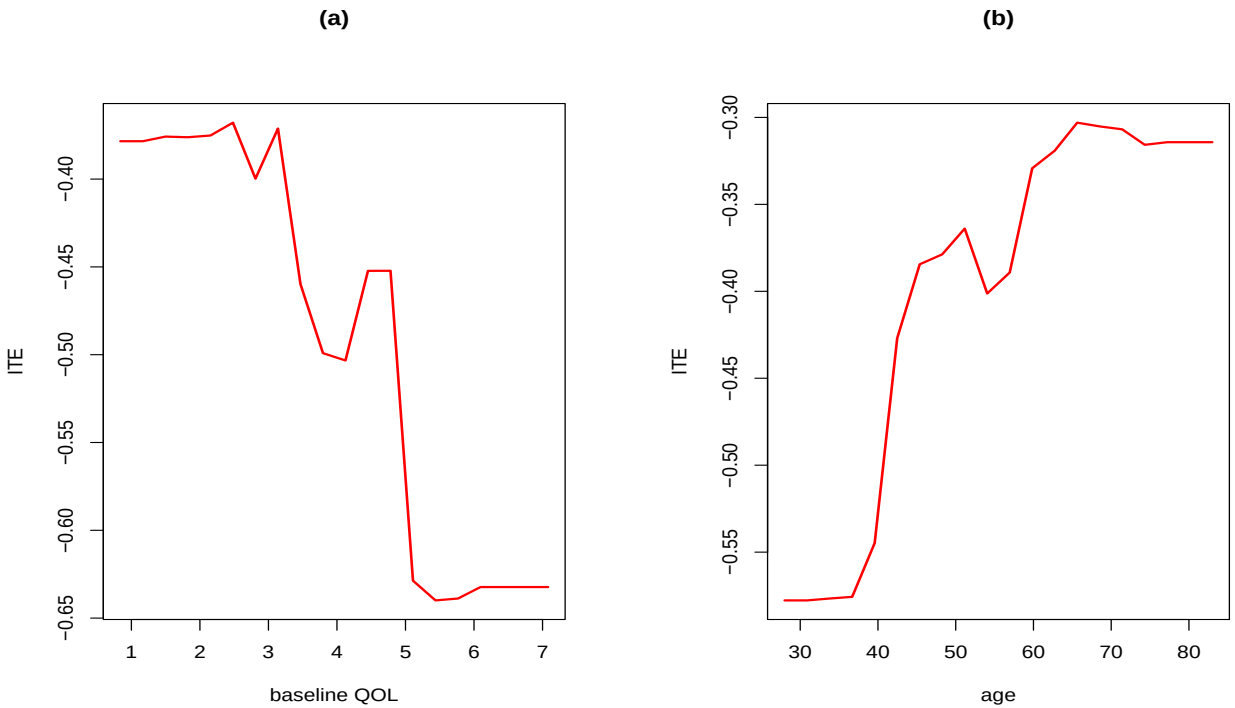


Figure : Partial Dependence Plots from Random Forests of Interaction Trees: The BCEI Data.

Exploring Qualitative Interaction

- Qualitative interaction only possibly exists when ITEs have both positive and negative values.
- Consider a classification problem by setting responses as $1\left\{\hat{\delta}_i \leq 0\right\}$. Run CART analysis and/or random forests.
- With BCEI data, no non-null tree structure was obtained.

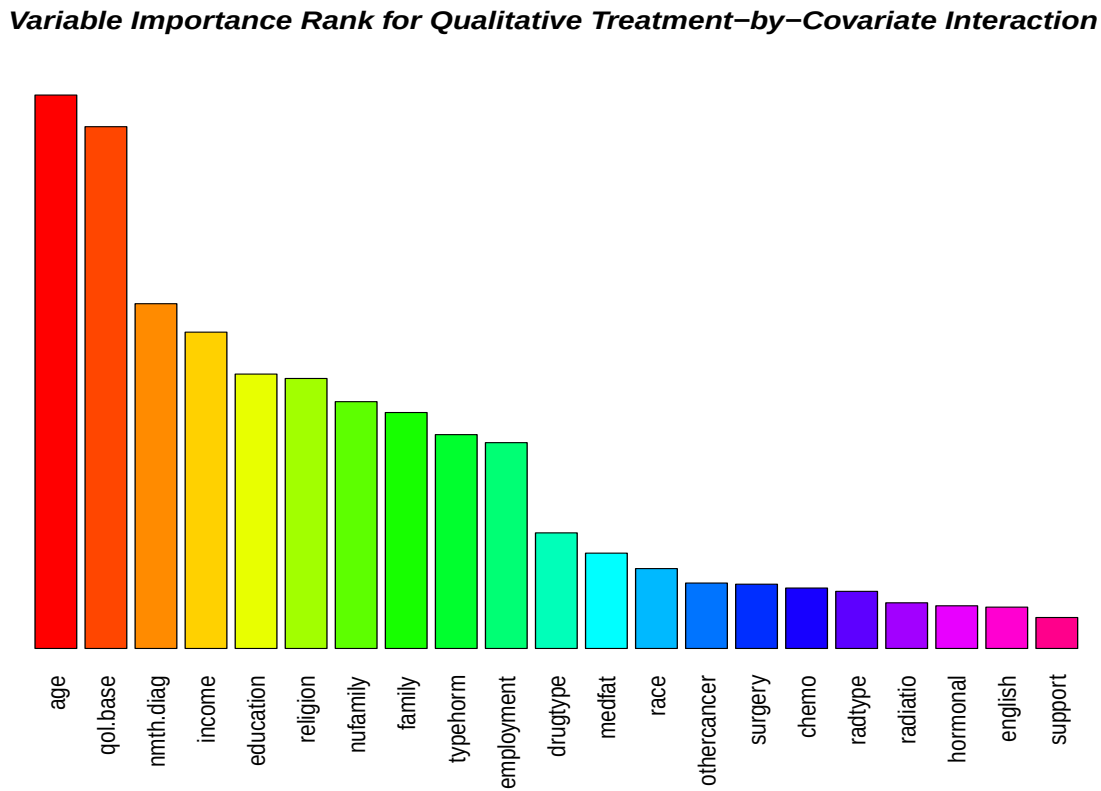


Figure : Variable Importance for Exploring Qualitative Interaction: The BCEI Data.

Optimal Treatment Regime

- Dynamic treatment regimes was first proposed by Murphy (2003, *JRSSB*) who borrowed the idea of system control.
- With independent data, a treatment regime $g(\cdot)$ is a function of $\mathbf{x}$ that maps to the domain of T, i.e., $\{0, 1\}$. The potential outcome with treatment regime g is

$$Y(g) = Y_1 g(\mathbf{x}) + Y_0(1 - g(\mathbf{x})).$$

- The optimal regime (Zhang et al., 2012 *Biometrics*)

$$\begin{aligned} g^\star &= \arg\max_g EY(g) \\ &= \arg\min_g E\left\{|\delta(\mathbf{x})|\left[I(\delta(\mathbf{x}) > 0) - g(\mathbf{x})\right]^2\right\} \end{aligned}$$

- A weighted classification problem.

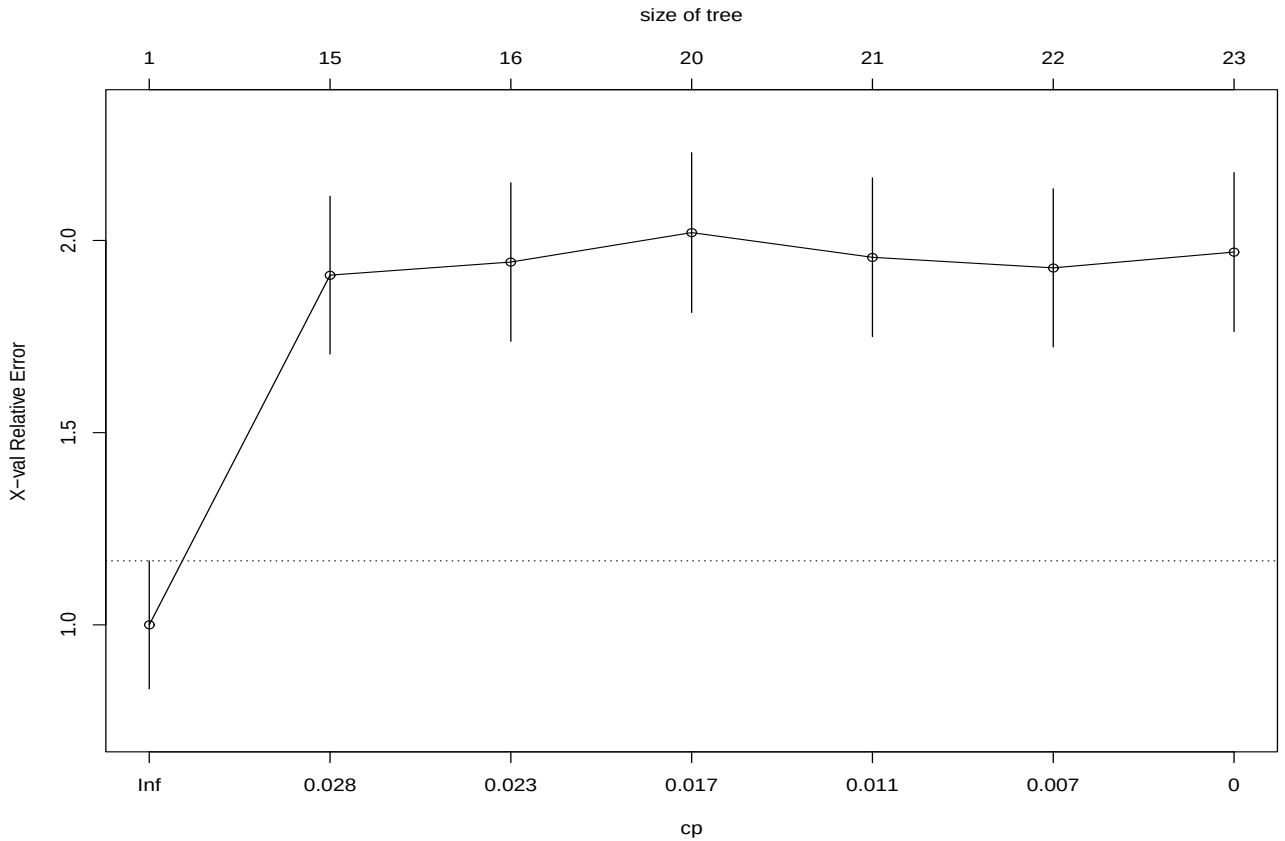


Figure : Tree Selection for Finding Optimal Treatment Regime: The BCEI Data.

Discussion

- Interaction tree facilitates subgroup-level causal inference, which provides the building block for many other features.
- Random forests of IT provides superior performance in estimating individual treatment effects, compared to conventional separate regression.
- Seeking subpopulations with *enhanced* treatment effects? Incorporating *toxicity* or *cost* into the analysis? (Lipkovich et al., 2011 *Statistics in Medicine*)
- How to deal with observational data? (Su et al., 2012 *JMLR*)

Discussion

Thanks!