# Random Forests of Interaction Trees for Estimating Individualized Treatment Effects in Randomized Trials

**Xiaogang Su**

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

July 12, 2017 @ NUS IMS Workshop on Precision Medicine

# Outline

- Introduction
  - Rubin's Causal Model
  - Tree Methods
- RFIT
  - SSS for Splitting Data
  - Estimating ITE with RFIT
  - IJ-Based SE Formula
- Simulation Studies
  - RFIT vs. SR
  - Evaluation of SE
- The BCEI Study
  - Other Features of RFIT
- Discussion

# Precision Medicine in General

- Precision medicine aims to optimize the delivery of individualized therapies by integrating comprehensive patient data.
  - Stratified vs. personalized medicine.
  - In terms of sources, data may come from randomized experiments (efficacy) or from observational studies (effectiveness).
  - Two general approaches: estimating the stratified or individualized treatment effects or determining the optimal treatment regime (static or dynamic).
- Our focus is to estimate the individualized (static) treatment effects based on data from randomized trials.
- **Relevant Concepts**: effect moderation or modification, subgroup analysis, qualitative and quantitative treatment-by-covariates interaction, optimal treatment regime, etc.

# Rubin's Causal Model

- 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*: $Y_0(\omega)$ and $Y_1(\omega)$ and the observed outcome $Y(\omega) = \{1 - T(\omega)\}Y_0(\omega) + T(\omega)Y_1(\omega)$. Available data $\{(y_i, t_i, \mathbf{x}_i) = (y(\omega_i), t(\omega_i), \mathbf{x}(\omega_i)) : i = 1, \ldots, n\}$.
- 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*: $E(Y_1) - E(Y_0)$ (ATE).

## Individual Treatment Effects (ITE)

- The "*individual treatment effect*" (ITE) is defined as $E(Y_1 - Y_0|\mathbf{X} = \mathbf{x})$, i.e., the conditional expectation of the difference $Y_1 - Y_0$ 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 to approximate the unit level causal effect.
- One nature approach is to relax conditioning on $\mathbf{X} = \mathbf{x}$ to conditioning on a (data-adaptive) neighborhood of $\mathbf{x}$.

# Recent Statistical Approaches in Precision Medicine

Many recent proposals of statistical approaches in advancing precision medicine (Lipkovich et al., 2017 *SIM*) are available. To name a few,

- Tree-structured methods are dominant:
  - Interaction trees (IT; Su et al., 2009 *JMLR*);
  - Virtual twins (VT; Foster, Taylor, and Ruberg, 2011 *SIM*);
  - SIDES (Lipkovich et al., 2011 *SIM*);
  - Qualitative IT (Dusseldorp and Van Mechelen, 2014 *SIM*);
  - Unbiased (Loh, He, and Man, 2015 *SIM*);
  - Optimal treatment regime (Zhao et al., 2012 *JASA*; Zhang et al., 2012 *Biometrics*; Laber and Zhao, 2015 *Biometrika*; etc.).
- Parametric and semi-parametric (Cai et al., 2011 *Biostat*);
- Bayesian approach (Berger, Wang, and Shen, 2014 *JBS*; Xu et al., 2016 *JASA*);
- LASSO for hierarchical interactions (Bien, Taylor, and Tibshirani, 2013 *Annals* and Tian et al., 2014 *JASA*);
- Logistic-normal mixture model with latent class (Shen and He, 2015 *JASA*).

# Why Tree-Based 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. Among many other merits, tree models
  - Excel at modeling complex (nonlinear) interactions of higher orders (albeit implicitly).
  - Provide a natural way of grouping data with meaningful interpretation.
  - Are capable of modeling high-dimensional data of mixed types (off-the-shelf) with automatic variable selection.

## Random Forests of Interaction Trees (RFIT)

- Tree-structured subgroup analysis (e.g., IT) supply inference on stratified or subpopulation treatment effects. Then we can move backward to ATE by integrating results or move forward to ITE with ensemble learning methods.
- Our present objective focuses on implementation of Random Forests of Interaction Trees (RFIT) for estimating ITE. Our specific contributions include:
  - Explore a new way of splitting data, alternative to greedy search (GS);
  - Standard error formula based on infinitesimal jackknife (IJ)
  - Comparison with separate regression (SR).
  - Implementation of other useful features of RF.

## One Single Split

- Given data $\{(y_i, T_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ obtained from randomized trials, a split is induced by a binary question, e.g., 'if $X_j$ is continuous, is $X_j \leq c$ for a cutoff point $c$?'.
- Every split leads to a $2 \times 2$ table as below:

| | Child Node | |
|---|---|---|
| Trt | Left (L) | Right (R) |
| 1 | $(\bar{y}_{1L}, n_{1L})$ | $(\bar{y}_{1R}, n_{1R})$ |
| 0 | $(\bar{y}_{0L}, n_{0L})$ | $(\bar{y}_{0R}, n_{0R})$ |

- It is natural to consider 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 and $\varepsilon_i \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$.

# The Splitting Statistic

- To assess differential treatment effects between two nodes, it is natural to test for the interaction term, i.e., $H_0: \beta_3 = 0$ where $\beta_3$ corresponds to difference in differences (DID) in econometrics.
- The resultant $t$ or $z$ test is

$$z(X_j; c) = \frac{(\bar{y}_{1L} - \bar{y}_{0L}) - (\bar{y}_{1R} - \bar{y}_{0R})}{\sqrt{\hat{\sigma}^2(1/n_{1L} + 1/n_{0L} + 1/n_{1R} + 1/n_{0R})}},$$

  where $\hat{\sigma}^2$ is the pooled estimator of $\sigma^2$.
- In greedy search (GS), the best split solves $\max_{X_j;\, c} z^2(X_j, c)$, which can be viewed as a two-step search $\max_{X_j} \max_c z^2(X_j, c)$.

# Exhaustive/Greedy Search (GS)

- GS is time consuming, involves discrete optimization with erratic patterns, and suffers from end-cut preferences and variable selection bias problems.
- Su et al. (2016) proposed a smooth sigmoid surrogate (SSS) alternative to GS to amend its deficiencies.
- The main idea of SSS is to replace $\delta_i = 1\{x_i \leq c\}$ in the splitting statistic ($z^2$ here) with a smooth sigmoid function,

$$s_i = \pi\{a \cdot (x_{ij} - c)\}, \text{ with} \pi(x) = \{1 + \exp(-x)\}^{-1}.$$

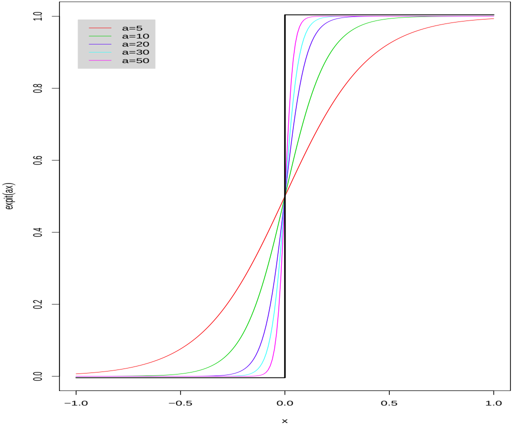


Figure: The expit function $\pi(x) = \{1 + \exp(-a(x - c)\}^{-1}$ with $c = 0$ and different $a$ values.

# Smooth Sigmoid Surrogate (SSS)

- Given predictor $X$, let $\delta_i = 1\{x_i \leq c\}$. First rewrite

$$\left\{ \begin{array}{lcl} n_{1L} & = & \sum_{i=1}^{n} T_i \delta_i \qquad \text{and} \qquad n_{1R} = n_1 - n_{1L} \\ n_{0L} & = & \sum_{i=1}^{n} (1 - T_i) \delta_i \quad \text{and} \qquad n_{0R} = n_0 - n_{0L} \end{array} \right.$$

- Denote the sums

$$\left\{ \begin{array}{lcl} S_{1L} & = & \sum_i y_i T_i \delta_i \qquad \text{and} \qquad S_{1R} = S_1 - S_{1L} \\ S_{0L} & = & \sum_i y_i (1 - T_i) \delta_i \quad \text{and} \qquad S_{0R} = S_0 - S_{0L}. \end{array} \right.$$

- It follows that, for $k = 1, 0$ and $t = \{L, R\}$,

$$\bar{y}_{kt} = S_{kt} / n_{kt} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-4} \left[ \sum_{i=1}^{n} y_i^2 - \sum_{k=0,1} \sum_{t=\{L,R\}} n_{kt} \bar{y}_{kt}^2 \right].$$

# Smooth Sigmoid Surrogate (SSS)

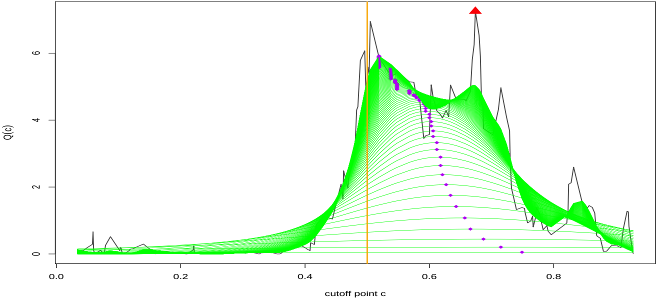
- For each predictor $X_j$, estimate the cutoff $c$ as

$$c_j^{\star} = \arg\max_{c} \tilde{z}^2(X_j; c), \tag{1}$$

  where $\tilde{z}^2(X_j; c)$ denotes the approximated squared $z$ test statistic.
- Solving (1) is a one-dimensional smooth (yet nonconcave) optimization problem.
  - Brent's (1973, *Algorithms for Minimization without Derivatives*) method for 1-D smooth optimization.
  - Strategies for seeking global optimum (unnecessary)

# Smoothed $z^2$ with $a = 1, 2, \ldots, 100$

SSS facilitates a parametric smoothing (with smoothing parameter $a$) to GS by smoothing its generating process.



Data ($n = 100$) were generated from $y = 0.5 + 0.5\,T + 0.5\,z + 0.5\,T \cdot z + \varepsilon$ with $x \sim \text{unif}(0, 1)$, $z = 1\{x \leq .5\}$, and $\varepsilon \sim N(0, 1)$.

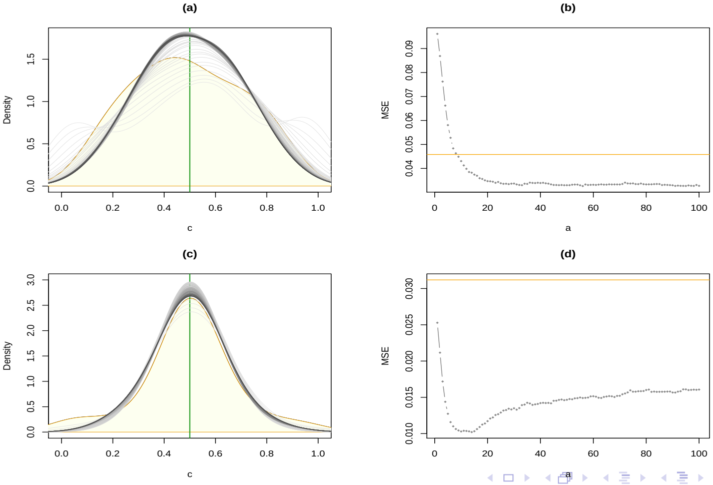
## Comparison of SSS vs. GS

- Data are generated from model

$$y = 0.5 + 0.5\,T + 0.5\,\Delta + 0.5 \cdot T\,\Delta + \varepsilon,$$

  where $\Delta = \Delta(x; c_0) = I(x \geq 0.5)$, $x \sim \text{uniform}[0, 1]$, and $\varepsilon \sim N(0, 1)$.
- Two sample sizes $n = 50$ and $n = 500$
- For SSS, different $a = 1, 2, \ldots, 100$ values are considered.
- For each model configuration, 500 simulation runs are made.
- Performance Criterion $MSE = \sum_{i=1}^{500}(\hat{c}_i - c)^2/500$.

# Empirical density of $\hat{c}$: GS vs. SSS



# Computing Complexity

### Proposition

*Consider the interaction tree setting with one continuous predictor $X$ that has $O(n)$ distinct values. Both GS and SSS are used to find its best cutoff point. In terms of computation complexity, GS is at best $O\{\ln(n)\, n\}$ with the updating scheme and $O(n^2)$ without the updating scheme. Comparatively, SSS is only $O(n)$.*

# Computing Time Comparison

CPU Time (in seconds) averaged over 10 runs. $X$ is generated from a discrete uniform distribution $\{1/K, 2/K, \ldots, K/K\}$, thus $K$ is the number of distinct values of $X$.

| | | $K = 10$ | | $K = 100$ | | $K = 500$ | |
|---|---|---|---|---|---|---|---|
| | | GS | SSS | GS | SSS | GS | SSS |
| $n =$ | 50 | 0.000 | 0.001 | 0.003 | 0.000 | 0.003 | 0.000 |
| | 100 | 0.000 | 0.000 | 0.006 | 0.000 | 0.003 | 0.004 |
| | 500 | 0.002 | 0.000 | 0.012 | 0.003 | 0.047 | 0.000 |
| | 1000 | 0.004 | 0.000 | 0.023 | 0.002 | 0.100 | 0.000 |
| | 2000 | 0.003 | 0.004 | 0.038 | 0.005 | 0.201 | 0.003 |
| | 5000 | 0.008 | 0.002 | 0.094 | 0.002 | 0.462 | 0.001 |
| | 10,000 | 0.017 | 0.005 | 0.182 | 0.005 | 0.899 | 0.010 |

# Variable Selection Bias

- *Variable selection bias* (VSB): A predictor with more values or levels is more likely to be selected as the splitting variable than a predictor with fewer values or levels.
- VSB is deemed inherent in CART with greedy search. Loh (2002; GUIDE) led the efforts in addressing this problem. His approach is to first determine the 'most important' splitting variable (an equally difficult problem) and then find its best cutoff point.
- SSS offers a way of avoiding variable selection bias in GS.

# $\chi^2$ Approximation with Numerically Determined DF

- For each $X_j$, obtain the maximized $\tilde{z}^2(X_j, \hat{c})$ and its distribution can be approximated by a $\chi^2$ distribution with certain df.
- The df is numerically explored with extensive Monte Carlo experiments.
  - If $X_j$ is binary, $df = 1$;
  - If $X_j$ is ordinal, $df \approx 2$ (maximally selected statistic). Minor adjustment for small $K \leq 5$ can be applied, where $K$ is the number of distinct values of $X_j$.
  - If $X_j$ is nominal with $K$ distinct levels, $df \approx 0.0443 + 0.6381K$.
- The best split is selected according to smallest p-value or largest log-worth (defined as $-\log_{10}$ p-value).

# Handling Nominal Covariates

- For a nominal covariate with $K$ distinct levels, there are a total of $2^{K-1} - 1$ ways of bisecting data.
- To speed up, one may first 'ordinal'ize its levels by sorting its levels according to estimated treatment effect at each level and then treat it as if ordinal. However, this 'ordinalization' step introduces over-optimism.
- Exact inference would involve the distribution of order statistics from independent but not identically distributed variables, which involves the concepts of (NP-hard) permanent (see, e.g., Vaughan and Venables, 1972 *JRSSB*) and selection differential (Nagaraja, 1982 *AoS*). However, Nagaraja only considered the IID or balanced case and hence the results cannot be used here.
- From numerical experiences, we notice that the null distribution of $\tilde{z}^2$ resembles $\chi^2$ very much, whose DF varies with $K$ only.

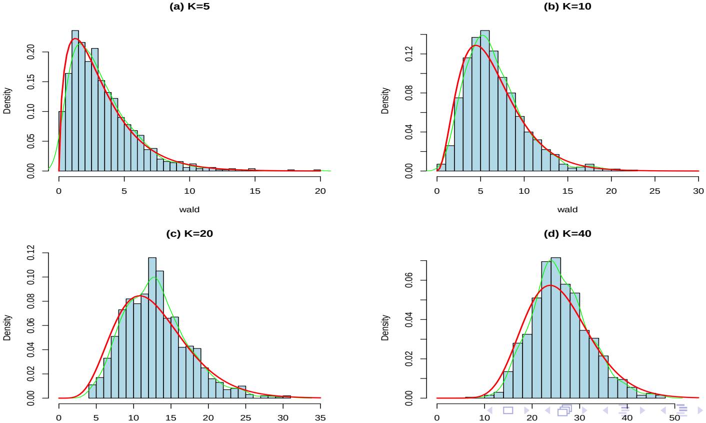
# A Computational Approach to Find df

- Fixed a $K$, we simulate data from the null setting (i.e., $X$ is not an effect-modifier of trt).
- Obtain $\tilde{z}^2(\hat{c})$ at the best cutoff point $\hat{c}$ according to SSS.
- Repeat the experiment for many times to obtain an empirical distribution of $\tilde{z}^2(\hat{c})$.
- Approximate this empirical distribution with a $\chi^2$ distribution. A nonlinear least square problem is involved, presenting a one-D optimization with *df* being the decision variable.
- Repeat the above procedure for different $K$ values and obtain the estimated df of the best approximating $\chi^2$ distribution.
- Plot *df* versus $K$. Fit a model to quantify their relationship. It turns out to be linear. A simple linear regression yields

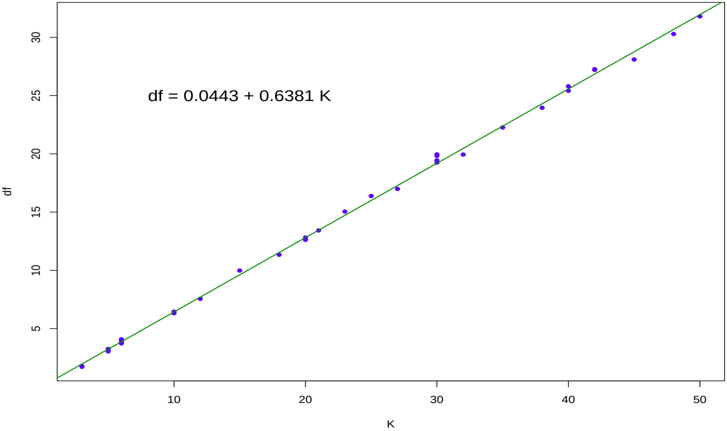
$$df \approx 0.0443 + 0.6381K.$$

# Empirical Distribution of $\tilde{z}^2$

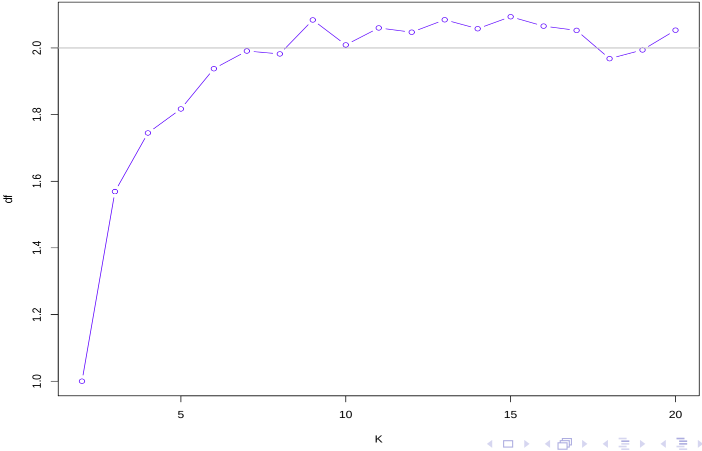
Split on a nominal variable with $K$ distinct levels. The red curve is density of the best $\chi^2$ approximation.



# Approximate DF Formula via Simple Linear Regression



# DF for Ordinal Covariates



# Individualized Treatment Effects (ITE)

- Estimated ITE $\delta = E(Y_1 - Y_0 | \mathbf{X} = \mathbf{x})$ can be useful in various ways,
    - Of key importance in deploying tailored plans in personalized medicine;
    - Affords deeper study of treatment efficacy;
    - Used as a preprocessor in other methods, e.g., virtual twins (VT; Foster, Taylor, and Ruberg, 2011), Zhang et al., (2012; *STAT*), and Laber and Zhao (2015; *Biometrika*).

# A Predictive Modeling Problem with Missing Data

Refer to the following data layout, which presents a missing data problem:

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

# Separate Regression (SR) for Estimating ITE

Regress (RF; super learner, etc.) $Y$ on $\mathbf{x}$ with data in the treated group (trt=1) and then use the fitted model to predict $Y_1$ in the untreated (trt=0) group. Similarly, build another model using data in the untreated group and make prediction for $Y_0$ in the treated 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 | $\mathbf{x}_{n_0+n_1}$ | $y_{n_0+n_1}$ | $Y_{1(n_0+n_1)}$ | $\hat{Y}_{0(n_0+n_1)}$ |

# Random Forests of Interaction Trees (RFIT)

```
Let $\mathcal{L}$ be the training data and $\mathcal{L}'$ be the test data.
Set $m$ and $B$.
For $b = 1, 2, \ldots, B$, do
    ▶ Obtain bootstrap sample $\mathcal{L}_b$.
    ▶ Repeat till a large tree $\mathcal{T}_b$ is obtained.
        ◇ Randomly select $m$ covariates.
        ◇ For $j = 1, 2, \ldots, m$, do
            ○ Apply SSS to find the best binary cut for $X_j$.
            ○ Obtain its associated p-value and log-worth.
        ◇ Bisect $\mathcal{L}_b$ according to the best split of data with maximum
        logworth.
    ▶ Send data $\mathcal{L}'$ down tree $\mathcal{T}_b$.
    ▶ Compute the predicted ITE $\hat{\delta}_{i'b}$ for each $i'$-th individual in $\mathcal{L}'$.
Average over $B$ bootstrap samples $\hat{\delta}_{i'} = \sum_{b=1}^{B} \hat{\delta}_{i'b}/B$.
```

## SE for Ensemble Learners

- The infinitesimal jackknife (IJ), also called nonparametric delta or influence function method, of Efron (2014, *JASA*) provides a general way of obtaining closed-form SE formulas for bootstrap based ensemble learners.
- IJ inspects the effect of an infinitesimal contamination at the *i*-th observation on the estimator.
- Simple to implement and flexible to use for many purposes

## IJ-Based SE Formula

### Proposition

*The IJ estimate of variance of $\hat{\delta}(\mathbf{x})$ is given by*

$$\hat{V} = \sum_{i=1}^{n} \bar{Z}_i^2, \tag{2}$$

*where $\bar{Z}_i = \sum_{b=1}^{B} Z_{bi}/B$ and $Z_{bi} = (N_{bi} - 1)\{\hat{\delta}_b(\mathbf{x}) - \hat{\delta}(\mathbf{x})\}$ with $N_{bi}$ being the number of times that the i-th observation appears in the b-th bootstrap resample. In other words, the quantity $\bar{Z}_i$ is the bootstrap covariance between $N_{bi}$ and $\hat{\delta}_b(\mathbf{x})$.*

# Bias-Corrected SE

## Proposition

*Especially for small or moderate $B$, $\hat{V}$ is biased upwards. A bias-corrected version is given by*

$$\hat{V}_c = \hat{V} - \frac{1}{B^2} \sum_{i=1}^{n} \sum_{b=1}^{B} (Z_{bi} - \bar{Z}_i)^2.$$

*Further assuming approximate independence of $N_{bi}$ and $\hat{\delta}_b(\mathbf{x})$, another computationally easier version is*

$$\hat{V}_c = \hat{V} - \frac{n-1}{B^2} \sum_{b=1}^{B} \{\hat{\delta}_b(\mathbf{x}) - \hat{\delta}(\mathbf{x})\}^2.$$

# Simulation Setting

- First simulate five ($p = 5$) predictors $x_j \sim \text{uniform}[0, 1]$ for $j = 1, \ldots, p$ independently.
- Then we generate $y_0' = \mu_0(\mathbf{x}) + \alpha + \varepsilon_0$ with a nonlinear polynomial
$$\mu_0(\mathbf{x}) = -2 - 2x_1 - 2x_2^2 + 2x_3^3$$
and $\alpha$ and $\varepsilon_0$ being independent from $\mathcal{N}(0, 1)$.
- Next, we generate $y_1' = \mu_1(\mathbf{x}) + \alpha + \varepsilon_1$, where $\mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + \delta(\mathbf{x})$ and $\varepsilon_1 \sim \mathcal{N}(0, 1)$ is independent of both $\alpha$ and $\varepsilon_0$.

## Notes on Simulation Setting

- A random effect term $\alpha$ is introduced to mimic some common characteristics shared by repeated measures $Y_0'$ and $Y_1'$ taken from the same subject.
- The unit-level effect $Y_1' - Y_0'$ equals $\delta(\mathbf{x}) + (\varepsilon_1 - \varepsilon_0)$, where $(\varepsilon_1 - \varepsilon_0)$ represents additional random errors that can not be accounted for by covaraites $\mathbf{x}$.
- The ITE $E(Y_1' - Y_0' | \mathbf{x}) = \delta(\mathbf{x})$.

# Four Models for ITE

- Four models (I)–(IV) are considered for the ITE $\delta(\mathbf{x})$, as given below:

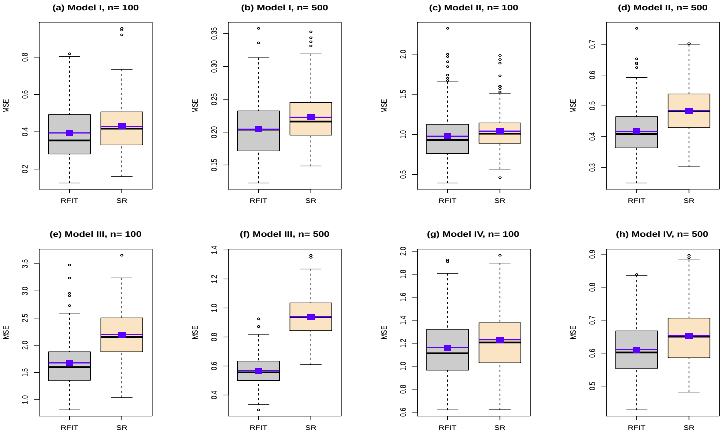
$$\begin{array}{ll} \text{I:} & \delta(\mathbf{x}) = -2 + 2x_1 + 2x_2 \\ \text{II:} & \delta(\mathbf{x}) = -2 + 2\,I(x_1 \leq 0.5) + 2\,I(x_2 \leq 0.5)\,I(x_3 \leq 0.5) \\ \text{III:} & \delta(\mathbf{x}) = -6 + 0.1\exp(4x_1) + 4\exp\{20(x_2 - 0.5)\} + 3x_3 + 2x_4 + x_5 \\ \text{IV:} & \delta(\mathbf{x}) = -10 + 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5. \end{array}$$

- Models III & IV derived from two nonlinear models in Friedman's (1991; *AoS*) MARS paper.

## Other Simulation Settings

- Simulate randomized treatment assignment variable $T$ independently from Bernoulli(0.5) and hence the observed response $y = Ty_1' + (1 - T)y_0'$.
- In order to evaluate their performance, a test sample $\mathcal{D}'$ of size $n' = 2000$ is generated beforehand.
- The mean square error (MSE) $\text{MSE} = \sum_{i=1}^{n'} \{\hat{\delta}(\mathbf{x}_i) - \delta(\mathbf{x}_i)\}^2 / n'$, averaged over simulation runs, is used as performance measure.
- Two sample sizes $n = 100$ and $n = 500$ and a total of 200 simulation runs is used for each model configuration.

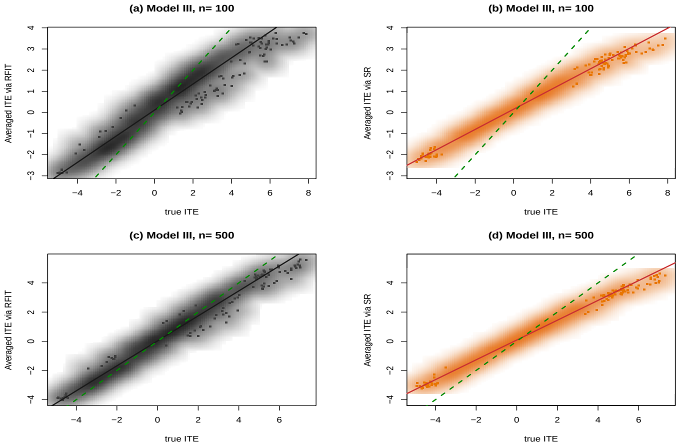
# MSE in Estimating ITE: RFIT vs. GS



# Why RFIT Outperforms SR?

- RFIT has a much easier learning task than SR!
- Suppose that $E(Y_0'|\mathbf{x}) = \mu_0(\mathbf{x})$ and $E(Y_1'|\mathbf{x}) = \mu_1(\mathbf{x}) = \mu_0(\mathbf{x}) + \delta(\mathbf{x})$. SR has to estimate both $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ while RFIT estimates $\delta(\mathbf{x})$ directly.
    - For example, consider the simplest scenario $\mu_0(\mathbf{x}) = \mu_1(\mathbf{x})$ with $\delta(\mathbf{x}) = 0$.
- Like other smoothing procedures, RF has a bias problem (Breiman, 1999). We suspect that SR incurs more bias than RFIT.

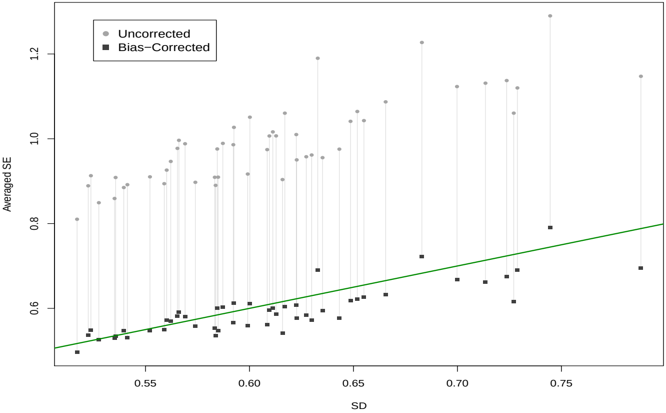
# True vs. Predicted ITE: RFIT and SR



## Simulation Setting

- To investigate the validity of SE, we generated one test data set $\mathcal{D}'$ of size $n' = 50$ from Model III and set it aside.
- 200 Simulation runs are taken. In each, a training data set $\mathcal{D}$ of size $n = 500$ is obtained.
- For each $\mathcal{D}$, $B = 2,000$ bootstrap samples is used to train RFIT and then the trained RFIT is applied to estimate ITE for each observation in $\mathcal{D}'$ together with standard errors.
- At the end of the experiment, we have 200 predicted ITE $\hat{\delta}$ for each observation in $\mathcal{D}'$, together with 200 SEs. Accordingly, we compute the standard deviation (SD) of these ITE estimates $\hat{\delta}$ and average the SE values.
- If the SE formula works well, the averaged SE values should be close to their corresponding SD values.

# SE in Estimating ITE with RFIT



# Observations from Numerical Experiences

- We experimented with other models and similar results were obtained.
- One issue pertains to the number $B$ of bootstrap samples needed. According to Efron (2014), a large $B$, e.g., $B = 2,000$ is needed to guarantee the validity of IJ-based standard errors.
- We experimented with different $B$ values. Generally speaking, ITE estimation stabilizes quickly even with a small $B$, e.g., $B = 100$; however, negative values may frequently occur to the bias-corrected variance estimates when $B$ is small or moderate, e.g., $B = 500$. Thus a large number $B$ of bootstrap samples are needed to have sensible results for the SE formulas.

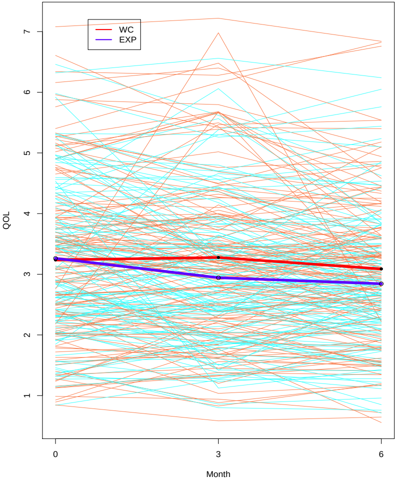
# The BCEI Study

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.

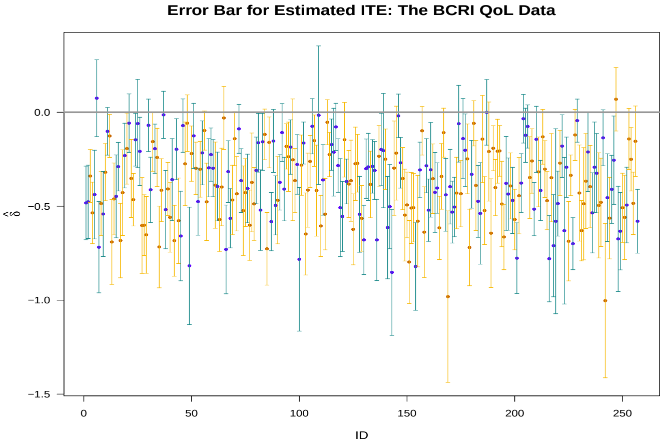
- Founded by NIH (R01) and initialized in 2001;
- 261 BCS's were randomized into the experimental (Exp) and the wait control (WC) groups and followed at baseline, Month 3, and Month 6;
- Four subjects in Exp dropped out and one died in WC during the followup period. 125 in Exp and 131 in WC completed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

# Effectiveness of BCEI on QOL

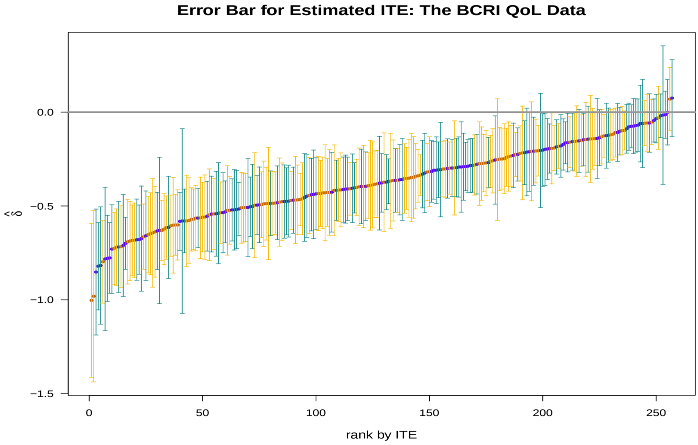
- The outcome variable, Quality of Life (QoL), is obtained from a 50-item instrument with four subdomains: Physical, Psychological, Social, and Spiritual.
- Each item scores on a 0-10 rating scale, *with lower scores indicating better QoL*. The overall QoL score is the grand average.
- The effectiveness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



# ITE with Error Bar by ID



# ITE with Error Bar by Rank



# A Distance Matrix for Differential Treatment Effect

- 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 $i$th 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 $t(i')$.

# Computing the 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

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

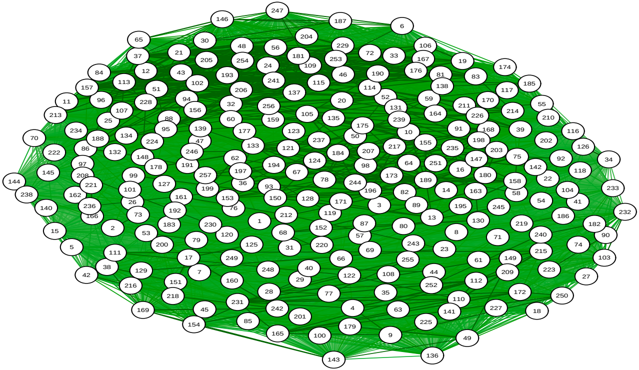
- In ordinary random forests, $\mathbf{B}_b = \mathbf{J} - \mathbf{I}$, where $\mathbf{J}$ is the $q \times q$ matrix of all 1's and $\mathbf{I}$ is the unit matrix. Thus $d_{ii'}^{(b)} = \sum_{t=1}^{q} a_{it} a_{i't} = 1$ if the $i$-th and $i'$-th subjects fall into different terminal nodes; and 0 otherwise.

# Distance Matrix

- Average the distances obtained from $B$ trees: $d_{ii'} = \sum_{b=1}^{B} d_{ii'}^{(b)}/B$. 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.
- The way of constructing the distance matrix $\mathbf{D}$ takes into account the fact that different terminal nodes may show homogeneous treatment effects and is applicable to high-dimensional data with covariates of mixed types.

# A Distance Matrix for Differential Effects of BECI

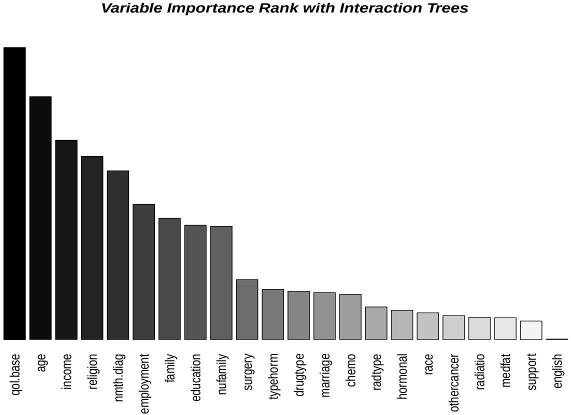
Matrix **D** visualized via a force-directed graph drawing algorithm



# 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 $V_j \leftarrow V_j / B$.
```

# Variable Importance from RFIT: The BCEI QoL Data



## Partial Dependence Plot

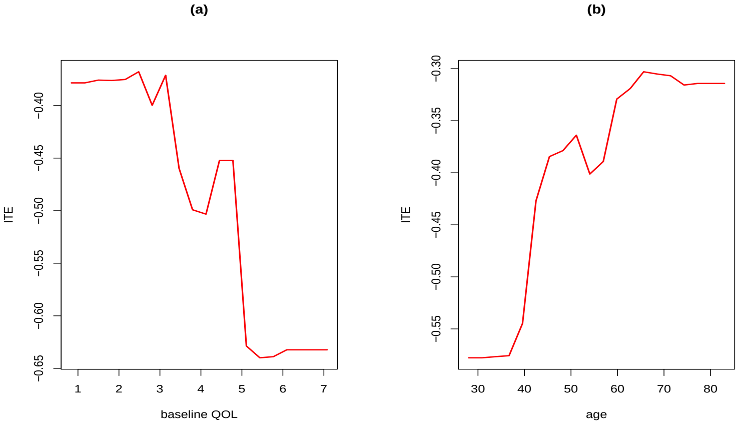
- First proposed by Friedman (1991, *Annals of Statistics*); implemented in R packages `randomForests` and others.
- Can be naturally extended to interaction trees:

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

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

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

# Partial Dependence Plot: the BCEI QoL Data



# Discussion

- SSS yields a superior performance to GS in many aspects and amends its deficiencies. There are interesting issues yet to explore.
- Built on the basis of IT, RFIT outperforms SR in estimating individual treatment effects and offers a number of useful features.
- IJ supplies a closed form SE for ITE estimates from RFIT.
- Future research avenues:
  - Extension to observational data (Su et al., 2012 *JMLR*);
  - More 'honest' estimate of ITE with RFIT;
  - Copula-based predictive modeling of individualized treatment effect;
  - '*Soft sphere tree*' (SST) in $p \gg n$ scenarios.

# References

- Breiman, L. (2001). Random Forests. *Machine Learning*, **45**: 5–32.
- Efron, B. (2014). Estimation and accuracy after model Selection (with discussion). *Journal of the American Statistical Association*, **109**: 991–1007.
- Lipkovich, I., Dmitrienko, A., D'Agostino, R. B. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, **36**: 136–196.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, **100**: 322–331.
- Su, X. G., Kang, J., Fan, J., Levine, R., and Yan, X. (2012). Facilitating Score and Causal Inference Trees for Large Observational Data. *Journal of Machine Learning Research* (*JMLR*), **13**: 2955–2994.
- Su, X. G., Tsai, C.-L., Wang, H., Nickerson, D. and Li, B. (2009). Subgroup Analysis via Recursive Partitioning. *Journal of Machine Learning Research*, **10**: 141–158.
- Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, **15**(1): 1625–1651.

# Thanks! Questions?