NONPARAMETRIC VARIABLE SELECTION BASED ON IMPORTANCE SCORES:

THE EARTH AND RANDOM FOREST VARIABLE SELECTION ALGORITHMS

KJELL DOKSUM
JOINT WITH SHIJIE TANG AND KAM TSUI

Department of Statistics
University of Wisconsin-Madison

MARCH, 2008

1. EARTH IMPORTANCE SCORES.
   VARIABLE SELECTION.
2. RANDOM FOREST IMPORTANCE SCORES.
   RANDOM FOREST VARIABLE SELECTION (RFVS)
3. COMPARISON OF EARTH AND RFVS.
4. MORE COMPARISONS.
   C_p, AIC, SBC, LASSO, GUIDE and MARS.
5. EARTH AND RFVS PRE-SCREENING TO IMPROVE MARS, GUIDE, AND RANDOM FOREST PREDICTIONS.
6. EARTH THRESHOLD SELECTION.
7. ASYMPTOTIC CONSISTENCY OF EARTH.

EARTH IMPORTANCE SCORES. VARIABLE SELECTION

DESIGNING AN EXPERIMENT IS LIKE GAMBLING WITH THE DEVIL: ONLY A RANDOM STRATEGY CAN DEFEAT ALL HIS BETTING SYSTEMS. (R.A. Fisher)

NONPARAMETRIC VARIABLE SELECTION: RANDOMIZATION WORKS WELL.

Y$=$ RESPONSE; $X_1, \cdots, X_d$$=$ COVARIATES. d IS LARGE.
VARIABLE SELECTION: KEEP THE IMPORTANT X_j's.
WRITE $\mathbf{X} = (X_1, \cdots, X_d)$,
$\mathbf{X}_{-j} = (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_d) \equiv \mathbf{X} - X_j$.

EARTH STRATEGY:
IF GIVEN $\mathbf{X}_{-j}$, X_j IS INDEPENDENT OF Y, THEN X_j IS NOT IMPORTANT AND SHOULD BE DROPPED.

SELECT X_j IF IT HAS A LARGE NONPARAMETRIC IMPORTANCE SCORE.

IMPORTANCE SCORE: CONDITIONALLY GIVEN $\mathbf{X}_{-j}$, COMPUTE THE NONPARAMETRIC REGRESSION OF Y ON X_j.

HOW TO DO THIS? OBSERVE i.i.d. $(\mathbf{X}^{(i)}, Y^{(i)})$, $i = 1, \cdots, n$. SET $\mathbf{X}_{-j}^{(i)} = \mathbf{X}^{(i)} - X_j^{(i)}$.

1) SELECT $\mathbf{X}^{(k)}$ AT RANDOM FROM $\{\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(n)}\}$

2) CONSTRUCT A TUBE IN R^{d-1} CENTERED AT $\mathbf{X}_{-j}^{(k)} = \mathbf{X}^{(k)} - X_j^{(k)}$

$$T_k = (Tube)_k = \{\mathbf{x}_{-j} \in R^{d-1} : |\mathbf{x}_{-j} - \mathbf{X}_{-j}^{(k)}| \le \delta\}$$

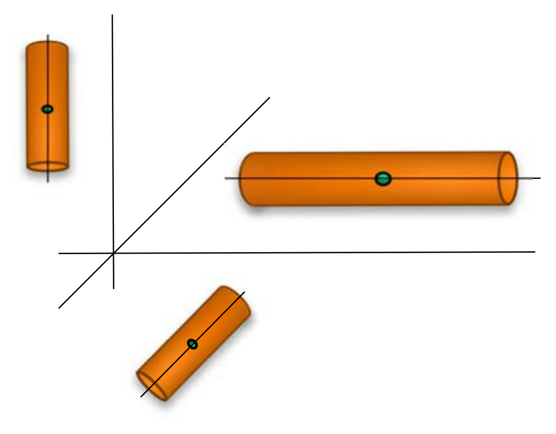


Figure: Random Tubes (WITH HELP FROM Wei-Yin Loh)

EARTH IMPORTANCE SCORES. VARIABLE SELECTION

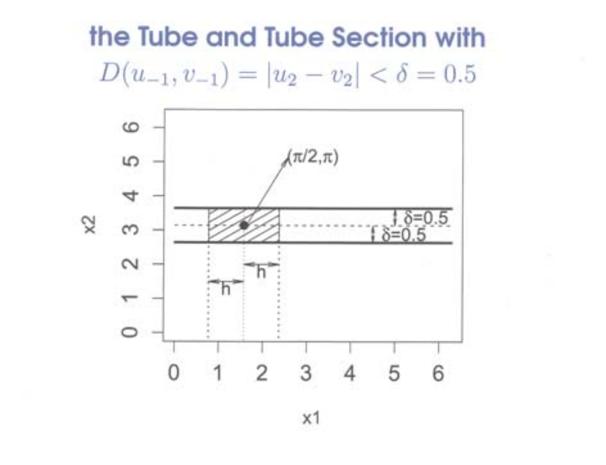


Figure: TUBE SELECTION FOR X_1 IMPORTANCE SCORE

3) COMPUTE A LOCALLY LINEAR t-STATISTIC $t_j^{(k)}(h)$ over the section $[X_j^{(k)} - h, X_j^{(k)} + h]$ in R.
4) Find $t_j^{(k)} \equiv \max_h |t_j^{(k)}(h)| =$ IMPORTANCE SCORE FOR VARIABLE X_j FOR TUBE T_k.
5) COMPUTE THE INITIAL IMPORTANCE SCORE FOR VARIABLE X_j AS $t_j = M^{-1} \sum_{k=1}^{M} t_j^{(k)}$.
6) USING 5), BOOST WEAK VARIABLES BY ADJUSTING THE METRIC $|\cdot|$ ADAPTIVELY.

NOTE: THE BANDWIDTH h IS SELECTED TO MAXIMIZE AN ESTIMATE OF THE EFFICACY OF THE LOCAL t-STATISTIC. EFFICACY IS A PROXY FOR POWER. WE WANT TO MAXIMIZE THE PROBABILITY OF SELECTING A RELEVANT VARIABLE.

EFFICACY IS VERY DIFFERENT FROM MEAN SQUARED ERROR. SMALL BIAS IS NEEDED ONLY WHEN X_j IS IRRELEVANT.

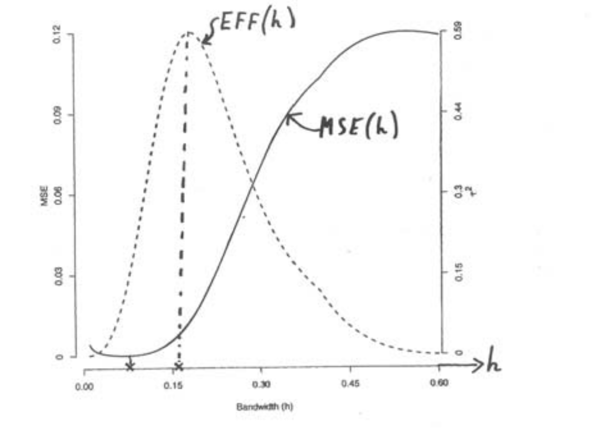


Figure: BUMP MODEL, $x_0 = 0.4$, n=1000, $\sigma^2 = 0.05$.

EARTH THRESHOLD: PERMUTE THE Y's INSIDE THE TUBES AT RANDOM, THEN COMPUTE

$t_j^* \equiv$ AVERAGE OF THE IMPORTANCE SCORES FOR THE "PERMUTED TUBES".

SELECT X_j IF $t_j \geq ct_j^*$. HERE THE THRESHOLD CONSTANTS IS SELECTED USING TRAINING AND TEST SETS, OR USING ASYMPTOTICS.

WE WANT (AS SAMPLE SIZE $n \to \infty$)

PROB(WRONGLY SELECT $\{X_j\}$) $\to 0$ and

PROB(WRONGLY DELETE $\{X_k\}$) $\to 0$.

RANDOM FOREST IMPORTANCE SCORES

RFVS: RANDOM FOREST VARIABLE SELECTION.
CART: CLASSIFICATION AND REGRESSION TREES.
CART CONSTRUCTS A NONPARAMETRIC REGRESSION MODEL FIT $\hat{\mu}(\mathbf{X})$.
SET $\mu(\mathbf{X}) = E(Y|\mathbf{X})$,
$\mu(\mathbf{X}_{-j}) \equiv E(Y|\mathbf{X}_{-j}) = \mu(\mathbf{X}_j^*)$,
WHERE $\mathbf{X}_j^*$ IS $\mathbf{X}$ WITH X_j REPLACED BY X_j^* INDEPENDENT OF $(\mathbf{X}_{-j}, Y)$.

RFVS CRITERIA:

$$\begin{aligned}\Delta_j &= E[Y - \mu(\mathbf{X}_{-j})]^2 - E[Y - \mu(\mathbf{X})]^2 \\ &= E[\mu(\mathbf{X}) - \mu(\mathbf{X}_{-j})]^2\end{aligned}$$

IF X_j IS INDEPENDENT OF Y GIVEN $\mathbf{X}_{-j}$, THEN $\Delta_j = 0$.

THE ESTIMATE OF Δ_j IS

$$\hat{\Delta}_j = n^{-1}\{\sum[Y^{(i)} - \hat{\mu}(\mathbf{X}_{-j}^{(i)})]^2 - \sum[Y^{(i)} - \hat{\mu}(\mathbf{X}^{(i)})]^2\}$$

HERE $\hat{\mu}(\mathbf{X}_{-j}^{(i)})$ AND $\hat{\mu}(\mathbf{X}^{(i)})$ ARE ESTIMATES OF $E(Y|\mathbf{X}_{-j} \in C_{-j})$ AND $E(Y|\mathbf{X}_{-j} \in C_{-j}, X_j \in C_j)$. IF X_j AND $\mathbf{X}_{-j}$ ARE DEPENDENT, THEN $\hat{\Delta}_j$ CAN BE LARGE EVEN IF, GIVEN $\mathbf{X}_{-j}$, X_j IS INDEPENDENT OF Y.

NEXT GENERATE TREES AT RANDOM (A BREIMAN BOOTSTRAP), GET $\hat{\Delta}_j^{(1)}, \cdots, \hat{\Delta}_j^{(k)}$. THE RF IMPORTANCE SCORE IS

$t_j^{RF} = \hat{\Delta}_j^{(\cdot)} / SE(\hat{\Delta}_j^{(\cdot)})$

RANDOM FOREST VARIABLE SELECTION: KEEP THE VARIABLES WITH LARGE IMPORTANCE SCORES t_j^{RF}.

COMPARE WITH SHIBATA(1981) WHO KEEPS THE MODEL(VARIABLES) WITH SMALLEST VALUES OF

$$[n + 2N(m)]n^{-1} \sum_{i=1}^{n} [Y^{(i)} - \hat{\beta}_{(m)}^{T}(\mathbf{X}^{(i)})]^2$$

WHERE $N(m) =$ NUMBER OF NON-ZERO β'S IN MODEL m.

THRESHOLD FOR RFVS:
EXPAND THE DESIGN MATRIX
$\mathbf{X}^* = (\mathbf{X}, X^*_{d+1}, \cdots, X^*_{d+r})$ WHERE THESE X^*_j ARE NOISE VARIABLES INDEPENDENT OF $(\mathbf{X}, Y)$.

SELECT THE VARIABLE X_j IF $t_j > ct'$.
WHERE t' IS THE AVERAGE OF THE IMPORTANCE SCORES FOR $X^*_{d+1}, \cdots, X^*_{d+r}$.

SUGGESTED:
$r = 30, c = 2, X^*_{d+1}, \cdots, X^*_{d+r} \sim \mathit{UNIFORM}(0, 1)$.

COMPARISON OF EARTH AND RFVS.

EARTH DIVIDES BY THE STANDARD ERROR IN EACH RANDOM TUBE. RFVS DIVIDES BY THE STANDARD ERROR ACROSS RANDOM TREES.

EARTH DEALS WITH CONFOUNDING BY CONDITIONING ON $\mathbf{X}_{-j}$, THEN USES SIMPLE (ONE X) NP REGRESSION. RFVS DEALS WITH CONFOUNDING BY CONDITIONING ON $\mathbf{X}$ AND $\mathbf{X}_j^*$. RFVS IS BASED ON DOING MULTIPLE REGRESSION (d X's) TWICE.

LOOKING AHEAD: RFVS DOES VERY WELL FOR INDEPENDENT $X_1, \cdots, X_d$. FOR STRONGLY DEPENDENT X's, NOT SO MUCH.

MORE COMPARISONS

EARTH AND RFVS AGAINST MARS, GUIDE ($\approx$ *CART*), C_p, AIC, SBC, LASSO.

1ST MONTE CARLO MODEL:
LINEAR REGRESSION. HOW MUCH DO NP METHODS EARTH, RFVS, MARS, GUIDE LOSE COMPARED TO PARAMETRIC METHODS C_p, AIC, SBC, LASSO?

PARAMETRIC MODEL.

MODEL (1):
$X_1, X_2, \cdots, X_{20}$ i.i.d. UNIFORM [0,1], $\epsilon \sim N(0,1)$.

$$Y = 1.25X_1 + X_2 + 0.75X_3 + 0.5X_4 + 0.25X_5 + \epsilon.$$

$X_6, X_7, \cdots, X_{20}$ ARE INDEPENDENT OF Y.

MORE COMPARISONS

		X_1	X_2	X_3	X_4	X_5	$X_i, i \geq 6$	AllX_i	
	C_p	100	100	99.4	49.0	0.6	0	92.5	
PAR METHODS	AIC	100	100	100	100	81.4	1.58	91.5	
	SBC	100	100	100	96.0	39.4	0.08	96.4	BEST OVERALL
	Lasso	100	100	100	100	88.2	5.29	73.0	
	GUIDE	100	100	100	99.4	64.2	0.67	94.8	
NP METHODS	MARS	99.2	99.2	99.2	98.4	93.8	5.31	72.9	
	RFVS	100	100	100	67.8	25.2	0.85	90.4	
	EARTH	100	99.4	96.6	49.0	14.2	0.79	89.0	

Table: Columns $1, \cdots, 5$ gives the percentage of simulation trials where variables $X_1, \cdots, X_5$ were selected for Model (1). Column 6 gives the average number of irrelevant variables per simulation falsely identified as relevant variable. Column 7 gives the percentage of correct identifications for 20 variables.

NONLINEAR (in X_2, X_3, X_4, X_5) MODEL.

MODEL (2):
$X_1, X_2, \cdots, X_{20}$ i.i.d UNIFORM(0,1), $\epsilon \sim N(0,1)$.

$$Y = X_1 + 5sin(2\pi X_2 + 2\pi X_3) + 8(X_4 - 0.5)^2 + e^{X_5} + \epsilon.$$

$X_6, \cdots, X_{20}$ ARE IRRELEVANT.

	X_1	X_2	X_3	X_4	X_5	$X_i, i \geq 6$	AllX_i
C_p	74.0	1.2	2.2	1.4	99.0	0.22	82.8
AIC	83.6	16.4	16.2	14.6	99.0	1.43	79.3
SBC	43.4	0.6	0.4	2.2	81.6	0.02	81.3
Lasso	86.4	18.2	20.6	19.6	99.8	3.10	71.7
GUIDE	80.2	57	55.6	57.4	98.2	4.69	69.0
MARS	77.0	83.2	73.4	80.4	80.6	9.77	45.9
RFVS	28.4	100	100	90.4	87.2	0.91	90.8
EARTH	39.8	99.6	99.4	82.4	81.4	1.06	89.8

Table: Columns $1, \cdots, 5$ gives the percentage of simulation trials where variables $X_1, \cdots, X_5$ were correctly selected for Model (2). Column 6 gives the average number of irrelevant variables per simulation falsely identified as relevant variable. Column 7 gives the percentage of correct identifications for 20 variables.

MODEL (3):
$Y = X_1 + 5\sin(2\pi X_2 + 2\pi X_3) + 8(X_4 - 0.5)^2 + e^{X_5} + \epsilon.$

GIVEN X_4, X_6 AND Y ARE INDEPENDENT.
WITHOUT X_4, X_6, AND Y ARE DEPENDENT
BECAUSE $CORR(X_4, X_6) \cong 0.9$.

THE SUM OF THE X_4 AND X_6 PERCENTAGES SHOULD BE 100.

$X_8, \cdots, X_{20}$ ARE IRRELEVANT FOR Y.

EARTH IMPORTANCE SCORES. VARIABLE SELECTION

	$X_1^{(5)}$	X_2	X_3	$X_4^{(6,7)}$	$X_5^{(1)}$	$X_6^{(4)}$	$X_7^{(4)}$	$X_i, i \geq 8$	$X_4 + X_6$	
C_p	24.6	3.4	4.8	2.2	85.2	6.0	5.4	0.68	8.2	
AIC	79.8	17.4	16.4	22.8	97.6	19.2	14.8	2.24	42.0	PAR METHODS
SBC	34.4	1.8	1.4	1.4	87.2	1.8	0.6	0.11	3.2	
Lasso	88.4	20.8	21.2	15.2	100	12.2	21.0	2.45	27.4	
GUIDE	78.6	51.8	50.0	59.4	97.2	41.8	30.2	4.14	101.2	
MARS	75	77.4	79.4	74.0	75.4	74.6	71.8	8.15	148.6	NP METHODS
RFVS	78.2	100	100	91.0	96.8	66.6	8.2	0.75	157.6	
EARTH	76.2	97.8	97.8	76.6	91.8	49.2	8.2	1.41	125.8	

Table: Columns $1, \cdots, 7$ gives the percentage of simulation trials where variables $X_1, \cdots, X_7$ were selected for Model (3). Column 8 gives the average number of irrelevant variables per simulation falsely identified as relevant variable. The superscript indicates what variables(s) the variable is associated with. Column 9 = $(\#X_4$ included $+ \#X_6$ included$)$ / $\#$trials.

EARTH AND RFVS PRE-SCREENING TO IMPROVE MARS, GUIDE, AND RANDOM FOREST PREDICTIONS.

MODELS AND CRITERIA FROM FRIEDMAN (1991) MARS PAPER.

CRITERIA = ISE $\cong$ PREDICTION ERROR STANDARDIZED TO BE COMPARABLE ACROSS MODELS.

$ISE(\bar{Y}) = 1$ ALWAYS.

Simulation experiment (Model and Criteria from MARS paper)

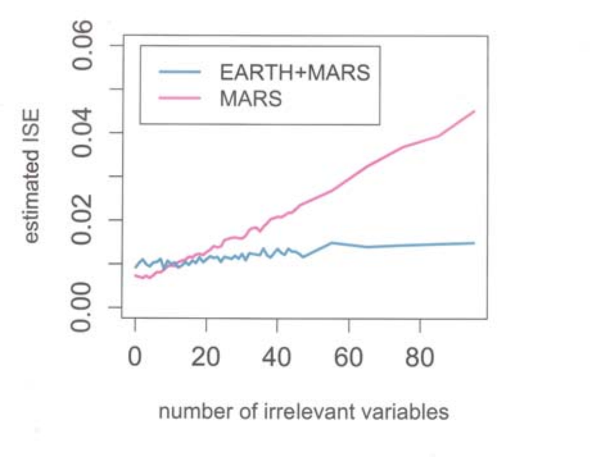
- $Y^{(i)} = \mu(\mathbf{X}^{(i)}) + \epsilon_i, \quad i = 1, \cdots, n,$
  $\mathbf{X}^{(i)} \in (0,1)^d, \quad \epsilon_i \sim N(0,1).$
- $\mathbb{D} = \text{DATA} = \{(\mathbf{X}^{(i)}, Y^{(i)}), i = 1, \cdots, n\}$, simulated from the model.
- $\hat{\mu}(\mathbf{x})$ is the estimator of $\mu(\mathbf{x})$ and predictor of Y based on $\mathbb{D}$.
- Evaluate the algorithm: use ISE = scaled Integrated Squared Error for this model based on Monte Carlo.

In MARS paper, $d = 10$ and $n = 50$, or 100, or 200.

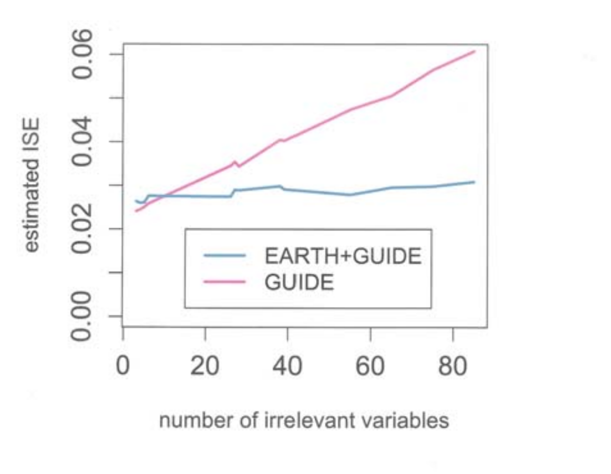
Mean Function 1 (MARS)

$$\mu(\mathbf{x}) = 0.1e^{4x_1} + \frac{4}{1+e^{-20(x_2-1/2)}} + 3x_3 + 2x_4 + x_5$$

$n = 1000$, $\widehat{\text{ISE}}$ vs. number of irrelevant variables are shown in next two slides.


EARTH+MARS
MARS
estimated ISE
0.00
0.02
0.04
0.06
0
20
40
60
80
number of irrelevant variables

EARTH AND RFVS PRE-SCREENING



EARTH AND RFVS PRE-SCREENING

Mean Function 2 (MARS)

$$\mu(\mathbf{x}) = 10sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5.$$

$n = 1000$, $\widehat{\text{ISE}}$ vs. number of irrelevant variables are shown in next two slides (breakdowns: ($\widehat{\text{ISE}} > 1$) for MARS omitted from the graph).

Table 2: Number of breakdowns among 100 trials.

d	30	40	50	60	70	80	90	100
MARS	1	1	0	2	2	5	6	7
EARTH+MARS	0	0	0	0	0	0	0	0

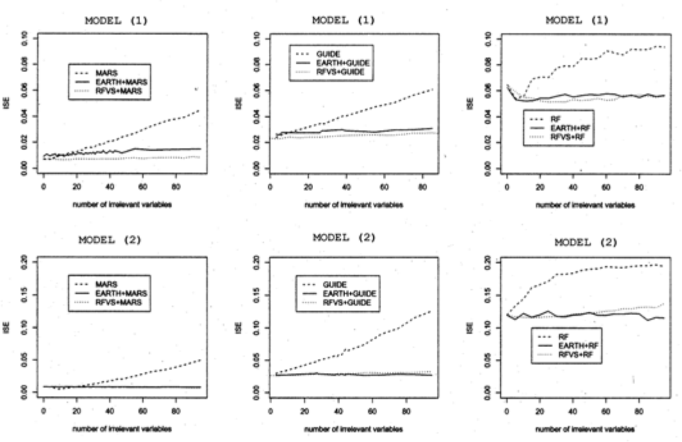


Figure: ISE's of MARS, GUIDE and Random Forest, and ISE's of these procedures preceded by EARTH and preceded by RFVS.

EARTH THRESHOLD SELECTION BASED ON TRAINING AND TEST SETS.

1) COMPUTE p-VALUES BASED ON IMPORTANCE SCORES AND NULL IMPORTANCE SCORES FOR A TRAINING SET.

2) USE DIFFERENT p-VALUE THRESHOLDS STARTING AT 0.01 TO GET SELECTED SETS OF X's. FOR EACH SET, CHECK PREDICTION ACCURACY USING A TEST SET. CHOOSE THE p-VALUE THRESHOLD WITH THE BEST PREDICTION ACCURACY.

l	5	6	7	8	9	13	14	15	16	17	18	19	20
p-values	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.35	0.1	0.03	0.00	0.40
l	21	22	23	24	25	26	27	28	29	30	31	32	33
p-values	0.27	0.20	0.14	0.15	0.24	0.16	0.17	0.15	0.16	0.16	0.02	0.09	0.06

Table: EARTH p-values for the 26 variables X_l, $l = \{5, 6, \cdots, 33\}$ using 5000 observations in a training set. "pole" data: www.liacc.up.pt/ ltorgo/Regression/DataSets.html

threshold(p_0)	selected variables	MARS	GUIDE	RF
0.01	5,6,7,8,9,13,14,19	361.92	97.73	54.26
0.02	5,6,7,8,9,13,14,15,19	361.92	89.81	39.55
0.03	5,6,7,8,9,13,14,15,19,31	361.92	86.37	41.02
0.06	5,6,7,8,9,13,14,15,18,19,31	341.10	85.05	41.64
0.09	5,6,7,8,9,13,14,15,18,19,31,33	322.66	85.14	36.02
0.1	5,6,7,8,9,13,14,15,18,19,31,32,33	311.42	92.78	36.88
0.14	5,6,7,8,9,13,14,15,17,18,19,31,32,33	304.33	90.06	38.65
...	...	...	...	...
(0.4, 1]	all	263.26	99.00	36.40

Table: Mean square prediction errors using variables selected by EARTH using p-value threshold p_0 for a test set with 10000 observation.

ASYMPTOTIC CONSISTENCY OF EARTH

RETT: REGRESSION TUBE t-STATISTIC SELECTION. FIX BANDWIDTH h. NO ADAPTION. M=1. ONLY ONE TUBE FOR EACH X_j.
AS $n \to \infty, d \to \infty, m = \#$ OBSERVATIONS IN THE SECTION FOR VARIABLES X_j.

ASSUME $m/n \to \lambda, 0 < \lambda \leq 1$. SET $t_j = \sqrt{m}\hat{\beta}_j/(s_{0j}/s_{x_j})$.
THEN $t_j/\sqrt{m} \to_P \beta_j/(\sigma_{0j}/\sigma_{x_j}) \equiv \tau_j$ as $n \to \infty$.

DIMENSION REDUCTION RULE:
KEEP X_j IF $|t_j| > c$,
DROP X_j IFF $|t_j| \leq c$. WLOG:

$$
\begin{aligned}
\tau_j &\neq 0, j = 1, \cdots, d_1; \\
&= 0, j = d_1 + 1, \cdots, d; \; d_0 = d - d_1;
\end{aligned}
$$

DEF: CONSISTENCY $\Leftrightarrow$ P(CORRECT DECISION MADE FOR ALL X_j) $\rightarrow 1$ as $n \rightarrow \infty$.
CONSISTENCY $\Leftrightarrow$
$P(\min_{j \leq d_1} |t_j| \geq c \textit{ and } \max_{j > d_1} |t_j| < c) \rightarrow 1.$

THEOREM: RETT IS CONSISTENT PROVIDED $min|\tau_j| = m^{-r} b_m$, WHERE $0 < r < 1/3$, $b_m \rightarrow \infty$, AND $\log(d_0) = o(m^r)$, $\log(d_1) = o(m^{1/2-r} b_m)$ AND $c = O(m^{r/2})$.

NOTE: THE "ALTERNATIVE" $|\tau_j|$ CAN TEND TO ZERO, BUT SLOWLY. d_0 AND d_1 CAN TEND TO INFINITY. THE THRESHOLD c TENDS TO INFINITY SLOWLY.

PROOF: SHOW
$P(\max_{j>d_1}|t_j| > c) \to 0$ AND
$P(\min_{j\le d_1}|t_j| \le c) \to 0$.

$P(\max_{j>d_1}|t_j| > c) \le \sum_{d_1+1}^{d} P(|t_j| > c)$.
Peter Hall 92: UNDER MOMENT CONDITIONS,
$P(|t_j| > c) \le \frac{2}{\sqrt{2\pi}} e^{-c^2/2}(1 + A(1+c)^3 m^{-1/2} a^3) \equiv \mathcal{H}$,
WHEN $c \le m^{1/6}/a$, UNDER OUR ASSUMPTION,
$P(\max_{j>d_1}|t_j| > c) \le d_0 \mathcal{H} \to 0$.

ALSO,
$P(\min_{j\le d_1}|t_j| \le c) \to 0$.