

Special Session on Random Forest and Ensembles

Maite Termenon¹

¹Computational Intelligence Group

2012 January 27

Article to Present



Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/cstda



Estimating residual variance in random forest regression

Guillermo Mendez^a, Sharon Lohr^{b,*}

^a 2124 E. Fremont Drive, Tempe AZ 85282, United States

^b School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804, United States

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions

Problem Definition

- $T = \{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$ of independent observations that follow a nonparametric regression model:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \quad (1)$$

where \mathbf{x}_i is a p -vector of covariates, f is the unknown mean function, ε_i are i.i.d random variables from a continuous symmetric distribution \mathcal{G} with mean 0 and variance σ^2 .

Mean Squared Prediction Error

- RF provides a predictor $\hat{f}(x_0)$ of a new observation with covariates x_0 .
- The mean squared prediction error (MSPE) for $\hat{f}(x_0)$ provides a measure of accuracy of the predicted value.
- MSPE includes:
 - squared bias for prediction
 - variance
- MSPE overestimates the residual variance, σ^2 , for moderate sample sizes.

Why is important the residual variance?

- An estimator of σ^2 is needed to determine which explanatory variables have significant relationships with the response, and to assess model fit.
- But an estimate of this residual variability in the data is not easily available using RF.

Proposal

- They proposed two estimators for residual variance:
 - A **residual-based estimator** of σ^2 using the RF algorithm.
 - A **difference-based estimator** of σ^2 using the RF algorithm.

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions

MSPE

- To estimate σ^2 , we need to separate the residual error from error in estimating f .
- The mean squared prediction error (MSPE) for a prediction $\hat{f}(\mathbf{x}_0)$ is:

$$\begin{aligned}\text{MSPE}[\hat{f}(\mathbf{x}_0)] &= E_Y \left[\left(Y - \hat{f}(\mathbf{x}_0) \right)^2 \mid \mathbf{x}_0 \right] \\ &= \sigma^2 + \text{Bias}^2 \left[\hat{f}(\mathbf{x}_0) \right] + \text{Var} \left[\hat{f}(\mathbf{x}_0) \right] \quad (3)\end{aligned}$$

- An unbiased estimator of MSPE will be positively biased when used as an estimator of σ^2 because the MSPE incorporates the error due to estimating $f(\mathbf{x}_0)$ as well as the pure error σ^2 .

MSPE in RF

- RF reduce the variance of a prediction by constructing K randomized trees and averaging over the multiple tree predictions.
- The bias term, however, is not reduced.

OOB observations

- The out-of-bag (OOB) observations are those that were not used to construct the tree.
- We denote these observations by T_k^{OOB} .
- Suppose there are K_i trees that do not use observation i during their construction.
- Averaging predictions at \mathbf{x}_i over these trees, RF OOB prediction is obtained:

$$\hat{f}^{OOB}(\mathbf{x}_i) = K_i^{-1} \sum_{k=1}^{K_i} \hat{f}_k(\mathbf{x}_i) I[(y_i, \mathbf{x}_i) \in T_k^{OOB}]$$

where $I(\cdot)$ is the indicator function.

Integrated MSPE

- Breiman estimated the integrated MSPE:

$$\text{MSPE}[\hat{f}] = \sigma^2 + E_X \left\{ \text{Bias}^2[\hat{f}(\mathbf{x}_0)] + \text{Var}[\hat{f}(\mathbf{x}_0)] \right\}.$$

- Using:

$$\text{mspe}[\hat{f}] = n^{-1} \sum_{i=1}^n [y_i - \hat{f}^{\text{OOB}}(\mathbf{x}_i)]^2 \quad (5)$$

Integrated MSPE

- Since the observations used in $\hat{f}^{OOB}(\mathbf{x}_i)$ are independent of y_i ,

$$\begin{aligned} E[\text{mspe}(\hat{f})] &= \sigma^2 + n^{-1} \sum_{i=1}^n E[(\hat{f}^{OOB}(\mathbf{x}_i) - f(\mathbf{x}_i))^2] \\ &= \sigma^2 + n^{-1} \sum_{i=1}^n \text{Bias}^2[\hat{f}^{OOB}(\mathbf{x}_i)] + n^{-1} \sum_{i=1}^n \text{Var}[\hat{f}^{OOB}(\mathbf{x}_i)] \end{aligned} \quad (6)$$

- This estimator works well for estimating the mean squared error when predicting new observations, but is biased when estimating σ^2 .
- The value of $E[\text{mspe}(\hat{f})] - \sigma^2$ is dominated by the squared bias term because large (small) values of $f(\mathbf{x}_i)$ tend to have negative (positive) bias since the prediction is a weighted average of the other y_j ($i \neq j$) values.

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation**
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions

Naive estimator

- We define a naive estimator of σ^2 using (5):

$$\hat{\sigma}_{NV}^2 := \text{mspe}[\hat{f}] = n^{-1} \sum_{i=1}^n [y_i - \hat{f}^{OOB}(\mathbf{x}_i)]^2 \quad (7)$$

- It is consistent for estimating σ^2 if $n^{-1} \sum_{i=1}^n [\hat{f}^{OOB}(\mathbf{x}_i) - f(\mathbf{x}_i)]^2$ converges in probability to 0, which requires mean squared error consistency of the estimated function.
- Although $\hat{\sigma}_{NV}^2$ will in general be consistent, for moderate values of n the bias of $\hat{\sigma}_{NV}^2$ from (6) may be large.

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation**
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions

Bias-corrected Estimator

- We subtract an estimator of the bias from σ_{NV}^2 to derive a bias-corrected residual variance estimator.
- Once a random forest is constructed, the prediction at a new point x_0 is a weighted average of the y values in T .
- The i th OOB prediction is a weighted average of the other $y_j (j \neq i)$ values in the training data but the bootstrapping makes the weighting system more complex.

OOB prediction for k th tree

- K_i = number of trees constructed without using observation i .
- $\tau_k(\mathbf{x}_i)$ = terminal node that contains \mathbf{x}_i for the k th tree ($k = 1, \dots, K_i$)
- b_{jk} = number of times observation j is selected in the k th bootstrap sample.
- OOB prediction for the k th tree: $\hat{f}_k^{OOB}(\mathbf{x}_i) = \sum_{j \neq i} w_{ijk} y_j$, where:

$$w_{ijk} = \begin{cases} b_{jk}/n(\tau_k(\mathbf{x}_j)) & \text{if } \mathbf{x}_i \text{ falls in } \tau_k(\mathbf{x}_j) \\ 0 & \text{otherwise} \end{cases}$$

- $n(\tau(\mathbf{x}_j))$ = number of observations in $\tau_k(\mathbf{x}_j)$.

OOB prediction for kth tree

- If an observation is bootstrapped more than once for a given tree, then all replicates will end up in the same terminal node.
- So, $\sum_{j \neq i} w_{ijk} = 1$ for each tree and the RF OOB prediction of observation i :

$$\hat{f}^{OOB}(\mathbf{x}_i) = K_i^{-1} \sum_{k=1}^{K_i} \hat{f}_k^{OOB}(\mathbf{x}_i) = \sum_{j \neq i} \bar{w}_{ij} y_j \quad (8)$$

- where $\bar{w}_{ij} = K_i^{-1} \sum_{k=1}^{K_i} w_{ijk}$

Bias

- We use the expression in (8) to estimate the last two terms in (6).
- The bias of σ_{NV}^2 as an estimator of σ^2 is given in (6):

$$R(\hat{f}) = n^{-1} \sum_{i=1}^n E[(\hat{f}^{OOB}(\mathbf{x}_i) - f(\mathbf{x}_i))^2] \quad (9)$$

- The bias $R(\hat{f})$ may be estimated using a full bootstrap, but procedure is computationally intensive.

Bootstrap-related Procedures

- We propose two bootstrap-related procedures that are less computationally intensive to estimate $R(\hat{f})$ directly, motivated by the fact that different sets of observations will be OOB for bootstrapped samples even if the same trees are used.

Parametric Bootstrap

Procedure 1. Parametric bootstrap

1. Calculate OOB predictions $\{\hat{f}^{OOB}(\mathbf{x}_i)\}$, $i = 1, \dots, n$, as in (8).
2. Generate B bootstrap samples $\{y_i^b\}$, $i = 1, \dots, n$, according to model (1) with parameters $\{\hat{f}^{OOB}(\mathbf{x}_i)\}$ and $\hat{\sigma}_{NV}^2$ given by (7).
3. Recompute $\{\hat{f}^{OOB,b}(\mathbf{x}_i)\}$, $i = 1, \dots, n$, using (8) and $\{y_i^b\}$ where the weights remain constant (no more trees are constructed).
4. Estimate $R(\hat{f})$ by

$$\hat{R}_P[\hat{f}] = (nB)^{-1} \sum_{b=1}^B \sum_{i=1}^n \left(\hat{f}^{OOB,b}(\mathbf{x}_i) - \hat{f}^{OOB}(\mathbf{x}_i) \right)^2. \quad (10)$$

Nonparametric Bootstrap

Procedure 2. Nonparametric bootstrap

1. Calculate OOB predictions $\{\hat{f}^{OOB}(\mathbf{x}_i)\}$, $i = 1, \dots, n$, as in (8) and then calculate $e_i = r_i - \bar{r}$, $i = 1, \dots, n$, where $r_i = y_i - \hat{f}^{OOB}(\mathbf{x}_i)$ and $\bar{r} = n^{-1} \sum_i r_i$.
2. Generate B bootstrap samples $\{y_i^b\}$, $i = 1, \dots, n$, according to

$$y_i^b = \hat{f}^{OOB}(\mathbf{x}_i) + e_i^b,$$

where e_i^b has been sampled, with replacement, from e_1, \dots, e_n .

3. Recompute $\{\hat{f}^{OOB,b}(\mathbf{x}_i)\}$, $i = 1, \dots, n$, using (8) and $\{y_i^b\}$ where the weights remain constant (no more trees are constructed).
4. Estimate $R(\hat{f})$ by

$$\hat{R}_N[\hat{f}] = (nB)^{-1} \sum_{b=1}^B \sum_{i=1}^n \left(\hat{f}^{OOB,b}(\mathbf{x}_i) - \hat{f}^{OOB}(\mathbf{x}_i) \right)^2. \quad (11)$$

Residual Variance Estimation

- Afterward, we can estimate σ^2 by using either bias-corrected estimator:

$$\hat{\sigma}_{BCP}^2 = \hat{\sigma}_{NV}^2 - \hat{R}_P[\hat{f}]$$

$$\hat{\sigma}_{BCN}^2 = \hat{\sigma}_{NV}^2 - \hat{R}_N[\hat{f}] \quad (12), (13)$$

- where σ_{NV}^2 is given by (7).
- Both estimators are consistent when σ_{NV}^2 is consistent since the bias correction tends to 0 as $n \rightarrow \infty$.

Negative Bias-corrected Estimators

- Bias-corrected estimators can, theoretically, be negative, especially for small values of σ^2 and small n .
- We recommend replacing such a value by zero.
- Negative values of these estimators may be a sign that $R(\hat{f})$ is relatively larger than the residual variance.

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation**
 - Bias-corrected estimator
 - Proximity measures estimator**
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions

Proximity measure

- Distance between two points \mathbf{x} and \mathbf{z} can be measured by the proportion of times they both land in the same terminal node among a set of trees.
- The terminal node of a tree represents a hyperrectangle in \mathfrak{R}^p , where p is the number of predictors.
- Trees grow until a specified terminal node size is reached, so the volumes of the hyperrectangles shrink as the number of observations increases.
- We use squared differences of the y values from pairs of observations that fall in the same terminal node to estimate σ^2 since the corresponding values, $f(\mathbf{x})$ and $f(\mathbf{z})$, must also be close together for smooth functions.

Proximity Measures estimator

- We propose a variance estimator that does not require the data to be sorted in any way.
 - using differences of responses with nonzero proximity measures to estimate σ^2 .
 - weighting the contribution of each pair of observations by the value of their proximity measure.

OOB proximity measure

- Defining proximity indicator between OOB observations \mathbf{x}_i and \mathbf{x}_j for a given tree k as $I(\mathbf{x}_i \in \tau_k(\mathbf{x}_j))$ where $\tau_k(\mathbf{x})$ is the terminal node of tree k that contains \mathbf{x} .
- OOB proximity measure for observations i and j , and tree k :

$$Q_k(\mathbf{x}_i, \mathbf{x}_j) = I(\mathbf{x}_i \in \tau_k(\mathbf{x}_j)) / \sum_{h=1}^K \sum_{l \neq m} I(\mathbf{x}_l \in \tau_h(\mathbf{x}_m)) \quad (14)$$

where k is the number of tree in the forest.

Estimator using Proximity Measures

- The estimator of σ^2 we propose:

$$\begin{aligned}\hat{\sigma}_{PROX}^2 &= \frac{1}{2} \sum_{k=1}^K \sum_{i \neq j} Q_k(\mathbf{x}_i, \mathbf{x}_j) (y_i - y_j)^2 \\ &= \frac{1}{2} \sum_{i \neq j} Q(\mathbf{x}_i, \mathbf{x}_j) (y_i - y_j)^2,\end{aligned}\tag{15)(16}$$

where $Q(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^K Q_k(\mathbf{x}_i, \mathbf{x}_j)$.

- Weights $Q(\mathbf{x}_i, \mathbf{x}_j)$ are symmetric, non-negative, and sum to one:

$$Q(\mathbf{x}_i, \mathbf{x}_j) = Q(\mathbf{x}_j, \mathbf{x}_i); \quad Q(\mathbf{x}_i, \mathbf{x}_j) \geq 0; \quad \sum_{i \neq j} Q(\mathbf{x}_i, \mathbf{x}_j) = 1$$

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results**
- 5 Variance of male and female test scores
- 6 Conclusions

Simulation Study

- We generated 1000 simulated data sets (replicates).
- For each replicate, 300 trees were constructed.
- We simulated the data sets under model (1) with $\mathbf{x} = (X_1, \dots, X_p)'$, where each X_j was distributed identically and independently uniform in $[0, 1]$.

Simulation Study

- We computed the empirical bias and empirical mean squared error (MSE) of each estimator:

$$\text{Bias}[\hat{\sigma}^2] = 1000^{-1} \sum_{r=1}^{1000} \hat{\sigma}_r^2 - \sigma^2,$$

$$\text{MSE}[\hat{\sigma}^2] = 1000^{-1} \sum_{r=1}^{1000} (\hat{\sigma}_r^2 - \sigma^2)^2,$$

where r is the replicate index and $\hat{\sigma}_r^2$ is the residual estimator using the r th simulated data set.

Factorial Design

- We performed a factorial design with the following factors:

- function, with number of covariates p :

$$f_1(\mathbf{x}) = 2x_1 + x_2^2 + x_3^2 + 2x_4^3, \quad p = 4;$$

$$f_2(\mathbf{x}) = \sin(2\pi x_1) + \cos(\pi x_2) + \sin(\pi x_3) + \cos(2\pi x_4), \quad p = 4;$$

$$f_3(\mathbf{x}) = -4x_1 + 3x_2 - 2.5x_3 + 1.5x_4 - x_5 + 0.5x_6, \quad p = 6;$$

$$f_4(\mathbf{x}) = 5 + \begin{cases} -4x_3 & \text{if } x_1 \geq 0.5 \\ -4x_4 + 3x_5 & \text{if } x_1 < 0.5 \text{ and } x_2 \geq 0.5 \\ 3x_5 - 2x_6 & \text{otherwise,} \end{cases} \quad p = 6.$$

- Number of potential covariates included in data set: p or $p + 6$. If $p + 6$ covariates are used, the 6 extra covariates are noise variables unrelated to the response.
- σ^2 : 2 or 4.
- Error distribution: $N(0, 2)$ or $t(4)$ for $\sigma^2 = 2$; $N(0, 4)$ or $t(8/3)$ for $\sigma^2 = 4$
- Terminal node size: $n_t = 1$ or 5.
- Sample size: $n = 250, 500, 1000$.

Parameters

- For Procedures 1 and 2:
 - Number of bootstraps replicates: $B = 100$.
 - Procedure 1: Normal distribution to simulate the error terms.
- All simulations were done in R version 2.6.1
- Randomized trees were constructed using the randomForest package (Liaw and Wiener, 2002).

Estimators used in Experiment

- **NV**: Naive Estimator.
- **BCP**: Bias Corrected Parametric bootstrap Estimator (Procedure1).
- **BCN**: Bias Corrected Nonparametric bootstrap Estimator (Procedure2).
- **PROX**: Proximity Measure Estimator.
- Results for $\sigma^2 = 4$ are similar to $\sigma^2 = 2$ and are not shown.

Results - Empirical Biases (I)

Table 1
 Empirical biases when $\sigma^2 = 2$.

| Error distribution | Variance estimator | $n = 250$ | | $n = 500$ | | $n = 1000$ | |
|---|--------------------|-----------|-----------|-----------|-----------|------------|-----------|
| | | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ |
| <i>$f = f_1$, no noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.239 | 0.246 | 0.179 | 0.213 | 0.163 | 0.175 |
| $N(0, 2)$ | BCP | 0.001 | -0.071 | -0.022 | -0.065 | -0.009 | -0.056 |
| $N(0, 2)$ | BCN | 0.001 | -0.071 | -0.022 | -0.065 | -0.009 | -0.056 |
| $N(0, 2)$ | PROX | 0.115 | 0.207 | 0.064 | 0.178 | 0.043 | 0.140 |
| $t(4)$ | NV | 0.246 | 0.249 | 0.194 | 0.184 | 0.155 | 0.172 |
| $t(4)$ | BCP | 0.011 | -0.057 | -0.005 | -0.082 | -0.012 | -0.064 |
| $t(4)$ | BCN | 0.011 | -0.057 | -0.005 | -0.083 | -0.012 | -0.064 |
| $t(4)$ | PROX | 0.095 | 0.237 | 0.042 | 0.153 | 0.002 | 0.147 |
| <i>$f = f_1$ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.255 | 0.231 | 0.188 | 0.205 | 0.142 | 0.144 |
| $N(0, 2)$ | BCP | 0.050 | 0.020 | 0.024 | 0.025 | 0.006 | -0.006 |
| $N(0, 2)$ | BCN | 0.050 | 0.020 | 0.024 | 0.025 | 0.006 | -0.006 |
| $N(0, 2)$ | PROX | 0.223 | 0.325 | 0.164 | 0.284 | 0.118 | 0.226 |
| $t(4)$ | NV | 0.219 | 0.290 | 0.200 | 0.163 | 0.178 | 0.143 |
| $t(4)$ | BCP | 0.013 | 0.061 | 0.031 | -0.022 | 0.038 | -0.011 |
| $t(4)$ | BCN | 0.013 | 0.061 | 0.031 | -0.022 | 0.037 | -0.011 |
| $t(4)$ | PROX | 0.158 | 0.413 | 0.130 | 0.299 | 0.105 | 0.273 |
| <i>$f = f_2$, no noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.386 | 0.402 | 0.306 | 0.318 | 0.237 | 0.252 |
| $N(0, 2)$ | BCP | 0.088 | 0.025 | 0.055 | -0.009 | 0.019 | -0.022 |
| $N(0, 2)$ | BCN | 0.088 | 0.025 | 0.055 | -0.009 | 0.019 | -0.022 |
| $N(0, 2)$ | PROX | 0.407 | 0.469 | 0.296 | 0.402 | 0.230 | 0.322 |
| $t(4)$ | NV | 0.400 | 0.404 | 0.309 | 0.305 | 0.227 | 0.241 |
| $t(4)$ | BCP | 0.105 | 0.039 | 0.062 | -0.019 | 0.014 | -0.029 |
| $t(4)$ | BCN | 0.105 | 0.039 | 0.062 | -0.019 | 0.014 | -0.029 |
| $t(4)$ | PROX | 0.411 | 0.496 | 0.291 | 0.399 | 0.191 | 0.327 |

Results - Empirical Biases (II)

Table 2
 Empirical biases when $\sigma^2 = 2$.

| Error distribution | Variance estimator | $n = 250$ | | $n = 500$ | | $n = 1000$ | |
|--|--------------------|-----------|-----------|-----------|-----------|------------|-----------|
| | | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ |
| <i>f = f₂ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.535 | 0.533 | 0.396 | 0.373 | 0.289 | 0.294 |
| $N(0, 2)$ | BCP | 0.279 | 0.274 | 0.189 | 0.143 | 0.111 | 0.097 |
| $N(0, 2)$ | BCN | 0.279 | 0.274 | 0.189 | 0.143 | 0.111 | 0.097 |
| $N(0, 2)$ | PROX | 0.639 | 0.721 | 0.470 | 0.585 | 0.365 | 0.501 |
| $t(4)$ | NV | 0.534 | 0.525 | 0.390 | 0.394 | 0.304 | 0.267 |
| $t(4)$ | BCP | 0.274 | 0.243 | 0.180 | 0.154 | 0.122 | 0.068 |
| $t(4)$ | BCN | 0.274 | 0.243 | 0.180 | 0.154 | 0.122 | 0.068 |
| $t(4)$ | PROX | 0.590 | 0.802 | 0.427 | 0.668 | 0.344 | 0.514 |
| <i>f = f₃ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.255 | 0.231 | 0.188 | 0.205 | 0.142 | 0.144 |
| $N(0, 2)$ | BCP | 0.050 | 0.020 | 0.024 | 0.025 | 0.006 | -0.006 |
| $N(0, 2)$ | BCN | 0.050 | 0.020 | 0.024 | 0.025 | 0.006 | -0.006 |
| $N(0, 2)$ | PROX | 0.223 | 0.325 | 0.164 | 0.284 | 0.118 | 0.226 |
| $t(4)$ | NV | 0.219 | 0.290 | 0.200 | 0.163 | 0.134 | 0.143 |
| $t(4)$ | BCP | 0.013 | 0.061 | 0.031 | -0.022 | -0.004 | -0.011 |
| $t(4)$ | BCN | 0.013 | 0.061 | 0.031 | -0.022 | -0.004 | -0.011 |
| $t(4)$ | PROX | 0.158 | 0.413 | 0.130 | 0.299 | 0.068 | 0.273 |
| <i>f = f₄ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.535 | 0.533 | 0.396 | 0.373 | 0.289 | 0.294 |
| $N(0, 2)$ | BCP | 0.279 | 0.274 | 0.189 | 0.143 | 0.111 | 0.097 |
| $N(0, 2)$ | BCN | 0.279 | 0.274 | 0.189 | 0.143 | 0.111 | 0.097 |
| $N(0, 2)$ | PROX | 0.639 | 0.721 | 0.470 | 0.585 | 0.365 | 0.501 |
| $t(4)$ | NV | 0.534 | 0.525 | 0.390 | 0.394 | 0.304 | 0.267 |
| $t(4)$ | BCP | 0.274 | 0.243 | 0.180 | 0.154 | 0.122 | 0.068 |
| $t(4)$ | BCN | 0.274 | 0.243 | 0.180 | 0.154 | 0.122 | 0.068 |
| $t(4)$ | PROX | 0.590 | 0.802 | 0.427 | 0.668 | 0.344 | 0.514 |

Results - Empirical MSEs (I)

Table 3

Empirical MSEs when $\sigma^2 = 2$.

| Error distribution | Variance estimator | $n = 250$ | | $n = 500$ | | $n = 1000$ | |
|--|--------------------|-----------|-----------|-----------|-----------|------------|-----------|
| | | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ |
| <i>f = f₁, no noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.102 | 0.107 | 0.055 | 0.069 | 0.037 | 0.042 |
| $N(0, 2)$ | BCP | 0.037 | 0.040 | 0.020 | 0.023 | 0.009 | 0.012 |
| $N(0, 2)$ | BCN | 0.037 | 0.040 | 0.020 | 0.023 | 0.009 | 0.012 |
| $N(0, 2)$ | PROX | 0.047 | 0.080 | 0.021 | 0.051 | 0.010 | 0.029 |
| $t(4)$ | NV | 0.342 | 0.278 | 0.207 | 0.164 | 0.100 | 0.117 |
| $t(4)$ | BCP | 0.237 | 0.172 | 0.147 | 0.110 | 0.067 | 0.076 |
| $t(4)$ | BCN | 0.236 | 0.172 | 0.146 | 0.110 | 0.067 | 0.076 |
| $t(4)$ | PROX | 0.201 | 0.250 | 0.119 | 0.138 | 0.054 | 0.099 |
| <i>f = f₁ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.109 | 0.097 | 0.057 | 0.063 | 0.030 | 0.030 |
| $N(0, 2)$ | BCP | 0.040 | 0.038 | 0.019 | 0.019 | 0.009 | 0.009 |
| $N(0, 2)$ | BCN | 0.040 | 0.038 | 0.019 | 0.019 | 0.009 | 0.009 |
| $N(0, 2)$ | PROX | 0.085 | 0.148 | 0.045 | 0.101 | 0.022 | 0.061 |
| $t(4)$ | NV | 0.283 | 0.601 | 0.299 | 0.135 | 0.076 | 0.101 |
| $t(4)$ | BCP | 0.201 | 0.436 | 0.227 | 0.093 | 0.052 | 0.070 |
| $t(4)$ | BCN | 0.201 | 0.435 | 0.226 | 0.093 | 0.052 | 0.070 |
| $t(4)$ | PROX | 0.207 | 0.715 | 0.180 | 0.209 | 0.050 | 0.165 |
| <i>f = f₂, no noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.201 | 0.211 | 0.116 | 0.127 | 0.068 | 0.074 |
| $N(0, 2)$ | BCP | 0.051 | 0.037 | 0.022 | 0.020 | 0.010 | 0.009 |
| $N(0, 2)$ | BCN | 0.051 | 0.037 | 0.022 | 0.020 | 0.010 | 0.009 |
| $N(0, 2)$ | PROX | 0.207 | 0.261 | 0.105 | 0.183 | 0.062 | 0.113 |
| $t(4)$ | NV | 0.419 | 0.413 | 0.267 | 0.236 | 0.121 | 0.115 |
| $t(4)$ | BCP | 0.227 | 0.194 | 0.150 | 0.113 | 0.061 | 0.047 |
| $t(4)$ | BCN | 0.227 | 0.193 | 0.150 | 0.113 | 0.061 | 0.047 |
| $t(4)$ | PROX | 0.365 | 0.466 | 0.210 | 0.289 | 0.088 | 0.158 |

Results - Empirical MSEs (II)

Table 4
 Empirical MSEs when $\sigma^2 = 2$.

| Error distribution | Variance estimator | $n = 250$ | | $n = 500$ | | $n = 1000$ | |
|--|--------------------|-----------|-----------|-----------|-----------|------------|-----------|
| | | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ | $n_t = 5$ | $n_t = 1$ |
| <i>f = f₂ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.332 | 0.336 | 0.182 | 0.160 | 0.094 | 0.097 |
| $N(0, 2)$ | BCP | 0.117 | 0.119 | 0.057 | 0.039 | 0.021 | 0.019 |
| $N(0, 2)$ | BCN | 0.117 | 0.119 | 0.057 | 0.039 | 0.021 | 0.019 |
| $N(0, 2)$ | PROX | 0.452 | 0.571 | 0.244 | 0.364 | 0.143 | 0.262 |
| $t(4)$ | NV | 0.610 | 0.482 | 0.311 | 0.309 | 0.176 | 0.155 |
| $t(4)$ | BCP | 0.351 | 0.228 | 0.170 | 0.155 | 0.090 | 0.078 |
| $t(4)$ | BCN | 0.351 | 0.228 | 0.170 | 0.155 | 0.090 | 0.077 |
| $t(4)$ | PROX | 0.614 | 0.866 | 0.315 | 0.616 | 0.188 | 0.357 |
| <i>f = f₃ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.109 | 0.097 | 0.057 | 0.063 | 0.030 | 0.030 |
| $N(0, 2)$ | BCP | 0.040 | 0.038 | 0.019 | 0.019 | 0.009 | 0.009 |
| $N(0, 2)$ | BCN | 0.040 | 0.038 | 0.019 | 0.019 | 0.009 | 0.009 |
| $N(0, 2)$ | PROX | 0.085 | 0.148 | 0.045 | 0.101 | 0.022 | 0.061 |
| $t(4)$ | NV | 0.283 | 0.601 | 0.299 | 0.135 | 0.076 | 0.101 |
| $t(4)$ | BCP | 0.201 | 0.436 | 0.227 | 0.093 | 0.052 | 0.070 |
| $t(4)$ | BCN | 0.201 | 0.435 | 0.226 | 0.093 | 0.052 | 0.070 |
| $t(4)$ | PROX | 0.207 | 0.715 | 0.180 | 0.209 | 0.050 | 0.165 |
| <i>f = f₄ with 6 noise covariates</i> | | | | | | | |
| $N(0, 2)$ | NV | 0.332 | 0.336 | 0.182 | 0.160 | 0.094 | 0.097 |
| $N(0, 2)$ | BCP | 0.117 | 0.119 | 0.057 | 0.039 | 0.021 | 0.019 |
| $N(0, 2)$ | BCN | 0.117 | 0.119 | 0.057 | 0.039 | 0.021 | 0.019 |
| $N(0, 2)$ | PROX | 0.452 | 0.571 | 0.244 | 0.364 | 0.143 | 0.262 |
| $t(4)$ | NV | 0.610 | 0.482 | 0.311 | 0.309 | 0.176 | 0.155 |
| $t(4)$ | BCP | 0.351 | 0.228 | 0.170 | 0.155 | 0.090 | 0.078 |
| $t(4)$ | BCN | 0.351 | 0.228 | 0.170 | 0.155 | 0.090 | 0.077 |
| $t(4)$ | PROX | 0.614 | 0.866 | 0.315 | 0.616 | 0.188 | 0.357 |

Results for $n = 3000$

Table 5

Empirical biases and MSEs when $\sigma^2 = 2$, $n = 3000$, and $n_t = 5$.

| Error distribution | Variance estimator | Bias | MSE |
|--|--------------------|-------|-------|
| <i>f</i> = <i>f</i> ₃ with 6 noise covariates | | | |
| <i>N</i> (0, 2) | NV | 0.080 | 0.010 |
| <i>N</i> (0, 2) | BCP | 0.004 | 0.004 |
| <i>N</i> (0, 2) | BCN | 0.004 | 0.004 |
| <i>N</i> (0, 2) | PROX | 0.061 | 0.006 |
| <i>t</i> (4) | NV | 0.071 | 0.011 |
| <i>t</i> (4) | BCP | 0.002 | 0.020 |
| <i>t</i> (4) | BCN | 0.002 | 0.020 |
| <i>t</i> (4) | PROX | 0.032 | 0.006 |

Comments on Results

- Parametric and nonparametric bootstrap procedures perform similarly, and both have smaller MSEs than the naive estimator.
- The full bootstrap did not perform better than the bootstraps in Procedures 1 and 2.
 - Full bootstrap may slightly overcorrect for the bias because new trees are grown for each iteration.
- The empirical variances of $\hat{R}(\hat{f})$ were very small relative to the mean squared errors in Tables 3 and 4, and decreased with increasing n .
- All the proposed estimators tend to perform better when the error terms are normally distributed than when the errors follow a t distribution.

More Comments on Results

- Proximity estimator performs better when the node size is set to the default value of five compared to one.
- Bias-corrected estimators performed better than the proximity estimator when the generated errors are normally distributed.
- For sample sizes less than 1000, the proximity estimator generally performed worse than the bootstrap bias-corrected estimators.
- With sample size of 3000, the proximity estimator performs very well, with small bias and MSE even though the simulated data included six noise variables.

How to Solve Negative Effect

- The presence of the six additional noise variables has a negative effect on the performance of the estimators.
- To improve the performance of the variance estimators is to adopt a two-stage procedure.
 - Implement random forest on the full data set.
 - Rerun the algorithm after omitting covariates with variable importance scores near zero

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions

A Real Study

- We use our models to estimate residual variability of 2873 male and 2720 female students on the 2007 Arizona Instrument to Measure Standards (AIMS).
- Our data set consists of the scale scores for 10th-grade students taking the mathematics test in 11 schools from 2 school districts.
- Covariates: School, District, Ethnicity, Age, primary Language, Startsch, Startdist, Numyrsch and Accom.
- Numyrsch value is missing for 38 students; we treated the missing values as a separate category for prediction.

Data Statistics Analysis

- 19 male and 14 female students achieved a perfect score on the exam.
 - female students have mean 721.8 and variance 1866.6
 - male students have mean 721.4 and variance 2124.6
- The male/female variance ratio is 1.138
- F test has p-value 0.0006 and 95% confidence interval [1.057, 1.226].
- F test is sensitive to the assumption of normality, so we also performed Levene's test and Shoemaker's modification 1 of the F test. Each had p-value < 0.003 .

Histograms of male and female scores

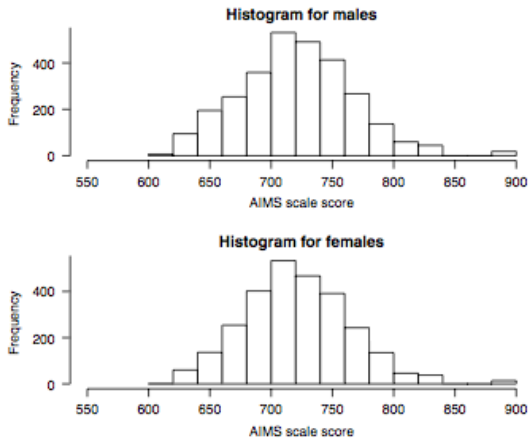


Fig. 3. Histograms of male and female scale scores on AIMS test.

Methodology

- We first fit a single RF model on all the data, using 5 observations per terminal node.
- The variable importance scores are shown in Fig. 4.
 - Gender has the lowest variable importance score.
- To be able to perform an F test for equality of the residual variances for males and females, we fit separate RF models for male and female students.
 - The separate RF models have similar predictions to the model fit with all the data,
 - but use independent sets of data.

Variable importance scores

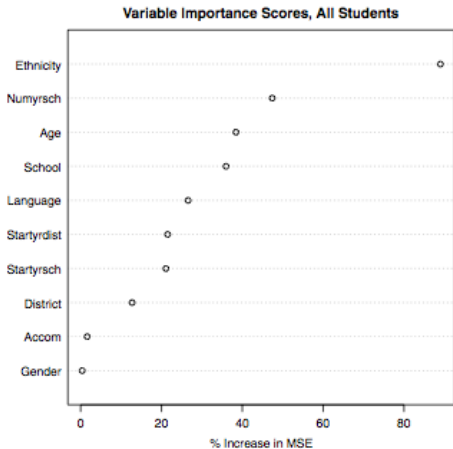


Fig. 4. Variable importance scores for covariates used in random forest predictions.

Degrees of Freedom

- Vector of predicted values for the data set: $\hat{\mathbf{f}} = \mathbf{W}\mathbf{y}$, where the (i, j) element of \mathbf{W} is \bar{w}_{ij} .
- As predictions are done using only OOB observations, the diagonal elements of \mathbf{W} are 0.
 - $tr(\mathbf{W}) = 0$, used to estimate model degrees of freedom in linear smoothers.
- We estimate the residual degrees of freedom by $n - tr(\mathbf{W}\mathbf{W}')$, obtaining 2816 residual df for the males and 2666 residual df for the females.

Estimates of Residual Variability

Table 6

Estimates of residual variance for male and female students.

| | $\hat{\sigma}_{NV}^2$ | $\hat{\sigma}_{BCP}^2$ | $\hat{\sigma}_{BCN}^2$ | $\hat{\sigma}_{PROX}^2$ |
|------------------|-----------------------|------------------------|------------------------|-------------------------|
| Male | 1649.5 | 1570.8 | 1571.3 | 1589.6 |
| Female | 1426.5 | 1359.4 | 1359.3 | 1492.8 |
| Ratio | 1.16 | 1.16 | 1.16 | 1.06 |
| 95% CI for ratio | [1.07, 1.25] | [1.07, 1.25] | [1.07, 1.25] | [0.99, 1.15] |

Comments on Results

- Bootstrap bias corrections both reduced the estimated residual variance by about the same amount.
- Proximity estimate was less than the naive estimate for males, but greater than the naive estimate for females.
- For these data, we think that the bias-corrected estimates are more reliable.
- We find that male students exhibit greater residual and raw variability for this data set.

More Comments on Results

- These data are not a representative sample of Arizona population.
 - Our results apply only to two school districts.
- Indeed, the estimated male/female residual variance ratio differs for the two districts studied:
- In addition, we had only limited covariate information, mostly demographic.
- We would expect greater reduction in variability, and perhaps different estimated ratios, with richer covariate information.
 - Estimators of residual variance using RF present a new tool that can account for complex interactions.

Outline

- 1 Motivation
- 2 Mean squared prediction error for random forest
- 3 Residual variance estimation
 - Bias-corrected estimator
 - Proximity measures estimator
- 4 Simulation results
- 5 Variance of male and female test scores
- 6 Conclusions


Summary

- We proposed three estimators of residual variance using the popular random forest algorithm:
 - Two Residual-based estimators that subtract an estimator of average bias from the estimator of prediction error which uses the OOB portion of data.
 - A Difference-based estimator which uses OOB proximity measures as weights.

Conclusions

- Methods proposed would also work for robust estimates of variability.
- These estimators are promising for statistical inference with RF models.
- They are easy to compute and implement even when the data are high dimensional.
- They may contain both continuous and categorical covariates.

References I

-  [Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News: The Newsletter of the R Project 2, pp. 18–22. Available online at <http://cran.r-project.org/doc/Rnews/>.