Summary and discussion of: "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife"

Journal club report*

Zhipeng LIANG

Contents

1 Summary

1.1 Bagging and Random Forest

Suppose that we have training examples $Z_1 = (x_1, y_1), ..., Z_n = (x_n, y_n)$, the dataset $\mathcal{D} = \{Z_i\}_{i=1}^n$, an input x to a prediction problem, and a base learner $\hat{\theta} = t(x; Z_1, ..., Z_n)$.

For bagging method, we stabilize the base learner t by resampling the training data. The bagging can be viewed as the solution to

$$\hat{\theta}^B(x) = \frac{1}{B}\sum_{b=1}^{B} t_b^*(x), \text{ where } t_b^*(x) = t\left(x; Z_{b1}^*, \ldots, Z_{bn}^*\right), \tag{1}$$

where the Z_{bi}^* are the i-th elements in the b-th bootstrap sample.

*The Latex document and the reproduced code can be found in https://github.com/liangzp/MATH-5472.

Moreover, the random forest constructs trees base on not only bootstrap samples but also random feature selection. Thus it can be viewed as the solution to the following problem

$$\hat{\theta}^{RF}(x) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{K}\sum_{k=1}^{K} t_b^*\left(x;\xi_{kb},Z_{b1}^*,\ldots,Z_{bn}^*\right) \text{ with } \xi_{kb} \overset{\text{iid}}{\sim} \Xi, \tag{2}$$

where ξ_{kb} can be viewed as the modeling for the feature selection noise.

To be specific, we introduce the generalized bagged version of $\hat{\theta}^{\infty}(x)$ is defined as

$$\hat{\theta}^{\infty}(x) = \mathbb{E}_{Z_i\sim P_*,\xi\sim\Xi}\left[t\left(x;\xi,Z_1^*,\ldots,Z_n^*\right)\right], \tag{3}$$

where P^* is the bootstrap distribution and Ξ is the auxiliary noise distribution. Then (1) can be viewed as a Monte Carlo approximation to the bootstrap distribution P^* with trivial distribution Ξ while random forest is a Monte Carlo approximation to both the bootstrap distribution P^* and nontrivial auxiliary noise distribution.

First, to understand the bootstrap distribution, note that the empirical distribution is

$$P_\delta(x,y) = \frac{1}{n}\sum_{i=1}^{n}\delta\left(x=x_i,y=y_i\right),$$

where $\delta(\cdot)$ is the Dirac mass centered at (x_i,y_i). Then the bootstrap distribution is

$$P_*(x,y) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n} k\delta\left(x=x_i,y=y_i\right)\frac{1}{k!}\exp^{-1}.$$

The derivation is from the approximation of the poisson distribution with parameter $\lambda = 1$ Poisson(1) to the multinomial distribution Multi$(n,\frac{1}{n})$ when $n\to\infty$ since we use Poisson bootstrap here.

1.2 IJ and J estimator

The goal of this paper is to study the sampling variance of bagged learners

$$V(x) = \mathrm{Var}(\hat{\theta}^{\infty}(x)).$$

This paper consider two basic estimates of $V(\cdot)$:

1. The Infinitesimal Jackknife estimate

$$\widehat{V}_{IJ}^{\infty} = \sum_{i=1}^{n}\mathrm{Cov}_*\left[N_i^*,t^*(x)\right]^2$$

where $\mathrm{Cov}[N_i^*,t(x)]$ is the covariance between $t^*(x)$ and the number of times N_i^* the i-th training example appears in a bootstrap sample. This is the direct application of the Theorem 1 in [1].

2. The Jacknife-after-Bootstrap estimate

$$\widehat{V}_J^{\infty} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\bar{t}^*_{(-i)}(x) - \bar{t}^*(x) \right)^2$$

where $\bar{t}^*_{(-i)}(x)$ is the average of $t^*(x)$ over all the bootstrap samples not containing the i-th example and $\bar{t}^*(x)$ is the mean of all the $t^*(x)$.

In practice, we can only ever work with a finite number B of bootstrap replicates. The natural Monte Carlo approximations to the estimators introduced above are

$$\widehat{V}_{IJ}^{B} = \sum_{i=1}^{n} \widehat{\mathrm{Cov}}_i^{\,2} \text{ with } \widehat{\mathrm{Cov}}_i = \frac{\sum_b \left(N^*_{bi} - 1\right) \left(t^*_b(x) - \bar{t}^*(x)\right)}{B}, \tag{4}$$

and

$$\widehat{V}_{J}^{B} = \frac{n-1}{n} \sum_{i=1}^{n} \hat{\Delta}_i^2,$$

where

$$\hat{\Delta}_i = \hat{\theta}^B_{(-i)}(x) - \hat{\theta}^B(x)$$

and

$$\hat{\theta}^B_{(-i)}(x) = \frac{\sum_{\{b: N^*_{bi}=0\}} t^*_b(x)}{\left|\{N^*_{bi} = 0\}\right|},$$

where N^*_{bi} indicates the number of times the i-th observation appears in the bootstrap sample b.

However, these finite-B estimates of variance are often badly biased upwards if the number of bootstrap samples B is too small. To correct the bias, we investigate the bias

$$\mathbb{E}_*\left[\widehat{V}_{IJ}^{B}\right] - \widehat{V}_{IJ}^{\infty} = \sum_{i=1}^{n} \mathrm{Var}_*\left[C_i\right], \text{ where } C_i = \frac{\sum_b \left(N^*_{bi} - 1\right)\left(t^*_b - \bar{t}^*\right)}{B}.$$

where $\bar{t}^*(x) = \frac{1}{b}\sum_{j=1}^{b} t^*_b(x)$.

They use the approximation

$$\mathrm{Var}_*\left[\left(N^*_{bi} - 1\right)\left(t^*_b - \bar{t}^*\right)\right] \approx \mathrm{Var}_*\left[N^*_{bi}\right] \mathrm{Var}_*\left[t^*_b\right]$$

which holds when $t(x; Z^*_1, Z^*_2, \cdots, Z^*_n) = \frac{1}{n}\sum_{i=1}^{n} Z^*_i$ as the mean estimator and Poisson bootstrap. To be specific, the approximation error is dominated by the approximated term.

Thus this bias can be approximated as

$$\mathbb{E}_*\left[\widehat{V}_{IJ}^{B}\right] - \widehat{V}_{IJ}^{\infty} \approx \frac{n}{B^2} \sum_{b=1}^{B} \left(t^*_b(x) - \bar{t}^*(x)\right)^2.$$

Thus we have the bias-correction version

$$\widehat{V}_{IJ-U}^{B} = \widehat{V}_{IJ}^{B} - \frac{n}{B^2}\sum_{b=1}^{B}\left(t_b^*(x) - \bar{t}^*(x)\right)^2, \tag{5}$$

and

$$\widehat{V}_{J-U}^{B} = \widehat{V}_{J}^{B} - (e-1)\frac{n}{B^2}\sum_{b=1}^{B}\left(t_b^*(x) - \bar{t}^*(x)\right)^2. \tag{6}$$

1.3 Upward Bias and Downward Bias

Suppose that we have data $Z_1, \cdots, Z_n$ drawn independently from a distribution F, and compute the estimate $\hat{\theta}^{\infty}$ from the data as in 3. Then using the ANOVA decomposition [4] we know

$$\mathrm{Var}_F\left[\hat{\theta}^{\infty}\right] = V_1 + V_2 + \ldots + V_n, \tag{7}$$

where V_k are all non-negative.

[4] also show that, under general condition, the Jacknife estimator of variance is biased upward. To be specific, the Jackknife estimator for the $n+1$-th data point has the variance

$$\mathbb{E}_F\left[\widehat{V}_J^{\infty}\right] = V_1 + 2V_2 + 3V_3 + \ldots + nV_n. \tag{8}$$

However, as pointed out by [5, 1], $\hat{V}_{IJ}^{\infty}$ is equivalent to the vaience of a "bootstrap Hjek projection", i.e.,

$$\widehat{V}_{IJ}^{\infty} \approx \sum_{i=1}^{n} h_F^2\left(Z_i\right) \tag{9}$$

where $h_F(Z) = \mathbb{E}_F\left[\hat{\theta}^{\infty} \mid Z_1 = Z\right] - \mathbb{E}_F\left[\hat{\theta}^{\infty}\right]$. (9) suggests that

$$\mathbb{E}_F\left[\widehat{V}_{IJ}^{\infty}\right] \approx V_1. \tag{10}$$

Compare (7), (8) and (10) we can conclude that Jackknife estimator has a upward bias by double the second-order term, triple the third-term and so on while Infinitesimal Jackknife estimator has a downward bias by dropping higher-order terms. Thus this paper suggests to use the the arithmetic mean of $\widehat{V}_J^{\infty}$ and $\widehat{V}_{IJ}^{\infty}$.

2 Result and Discussion

2.1 Empirical Bayes Calibration

My implementation is based on Stefan Wager's randomForestCI, ranger's implementation of IJ and sklearn's randomForestCI. All the figures can be reproduced using the code in my Github repository. Stefan Wager's code is for R while sklearn's is based on python. However, neither of them implemented Jacknife-after-Bootstrap estimator.

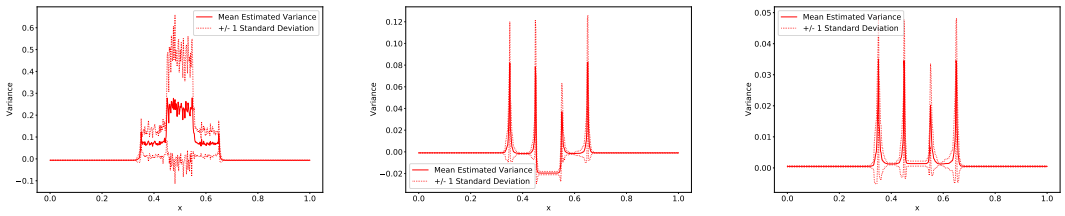
2.2 Jacknife-after-Bootstrap v.s. Infinitesimal Jackknife

In the first experiment I testing the performance of my version of the bias-corrected infinitesimal jackknife estimate of variance for bagged predictors, defined in 5, on the simulation data. The experiment setup is the same as described in the paper while here I only outline some points. The data is generaed via $y = f(x) + \epsilon$ where ϵ is a Gaussian noise and

$$f(x) = \begin{cases} 14 & 0.45 \leq x \leq 0.55 \\ 7 & 0.35 \leq x \leq 0.65 \\ 0 & 0 \leq x \leq 1. \end{cases} \tag{11}$$

There is a jump of level 7 at four discontinuous points, 0.35, 0.45, 0.55 and 0.65. Thus the ideal variance estimator should capture them. As we can see, Figure 1c shows that the infinitesimal jackknife estimator of the bagging predictors does achieve it. However, in Figure 1a my version of Jacknife-after-Bootstrap estimator although react to the discontinuous points but fail to capture the true variance. Then I increase the bagging size to 50000 and we can see from Figure 1b the Jacknife-after-Bootstrap estimator finally capture the true variance, which implies the Jacknife-after-Bootstrap estimator is more vulunerble to the Monte Carlo error.

Before we start to present other experiments, note that both in the originl code by Stefan Wager and sklearn's implementation, the bias-corrected infinitesimal jackknife estimate in (5) is not directly used. Instead, Wager conducted empirical Bayes calibration for the infinitesimal jackknife estimate using the g-modeling in [2]. Without such a empirical Bayes calibration, most of the elements in the infinitesimal jackknife estimate will be negative in my implementation.



(a) (b) (c)

Figure 1: The left figure is Jacknife-after-Bootstrap (Bagging size = 10000). The middle one is Jacknife-after-Bootstrap (Bagging size = 50000) and the right figure is Infinitesimal Jackknife.

2.3 Advantage of Bias-correction

In this experiment, I test the advantage of bias-correction version against the vallina one. I test their performance on the cholesterol dataset [3]. I closely follow the experiment setup in [1]. For completeness I outline some key ingredients here. I use the bagging polyregression and the choice of degree follows C_p criterior [6], i.e., for polyregression regression with degree

m, the C_p value is defined as

$$C_p(m) = \left\| \boldsymbol{y} - X_m \hat{\beta}_m \right\|^2 + 2\sigma^2 m,$$

where X_m is the design matrix with polynomial factors whose degree at most m and

$$\hat{\beta}_m = \underset{\hat{\beta}}{\operatorname{argmin}} \left\| \boldsymbol{y} - X_m \hat{\beta} \right\|^2.$$

Follows the setup in [1], we use the value $\sigma = 22$ in this experiments. We use the best degree for each bootstrap sample for the final regression model and apply the Infinitesimal Jackknife 4 and bias-correction version 5. As we can see in Figure 2, the bias-correction version (on the right column) is significantly lower than the biased one. Moreover, as the bootstrap sample sizes increase, the Monte Carlo error decreases which leads to the closer performance between biased IJ and bias-correction one.

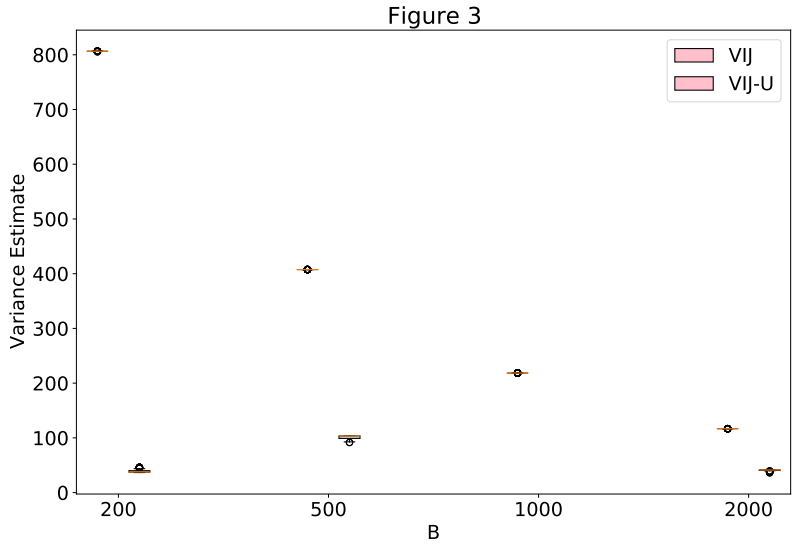


Figure 2: Boxplot of the Variance Estimate of Infinitesimal Jackknife and Bias-corrected Version

2.4 Application in Random Forest

In this section we aim to gain some qualitative insight from the Infinitesimal jackknife estimator on the random forest. From Figure 3, 4 and 5 we plot test-set predictions against IJ-U estimates of standard error for all three random forests with maximum number of features as 5, 19 and 57 (which reduce to the bagging version). The sample variance increases with the maximum number of select features. Thus the $m = 57$ forest tends to suffer overfitting under this limited sample size, which hinder it from achieving the best test

error amongs them. In contrast, the $m = 5$ forest has small variance across all samples, which may implies it suffies from bias due to the limited capacity of utiliziing features. From Figure 5 we plotthe prediction probability of $m = 5$ forest versus the difference between prediction probability of $m = 19$ against $m = 5$. To visulize the main trend from the scatters, we interpolate the scatters with a polynomial curve with degree 5 (red line). The red line remains negative when prediction is less than $1/2$ and remains positive when prediction is greater than $1/2$, which implies that the $m = 19$ forest has more confidence along the correct direction, and further verify the insight gained from the Infinitesimal Jackknife estimator.

2.5 Empirical Bayes Calibration

In this section I conduct experiments to show the necessarity of the empirical Bayes calibration. First I follows the similar setup in 2.2 and plot the variance estimate before Figure 6a and after the empirical Bayes calibration Figure 6b. In this experiment I decrease the number of bagging from 10000 which is 20 times the sample size, i.e., $\Theta(n^{1.5})$ as recommended in this paper to 500 which is $\Theta(n)$ where n is the sample size. Insufficient bagging number would incur unignorable Monte Carlo error. However, 6a validate the performance of the inifinite Jackknife even before the calibration whose elements are all non-negative and properly capture the jump points of the underlying function. Instead, after the calibration, the inifinite Jackknife has non-zero variance estimate even on the non-jump points and has a upward bias between 0.45 to 0.55 in Figure 6b. I susspect that the noise variance is too small to incur significant Monte Carlo noise and ruin the bias-corrected inifinite Jackknife estimator. I increase the variance from $\frac{1}{2^2}$ to 2^2. As shown by Figure 6c and Figure 6d, the variance is significant to ruin bias-corrected inifinite Jackknife estimator but cannot render it to have negative solution. Calibrated inifinite Jackknife also fail to capture the variance in this experiment. In conclusion, empirical Bayes calibration has no advantage in my simple simulation study.

Next I turn to the real-dataset experiment using the same setup in Section 2.4. Figure 7a is the results before calibration while Figure 7b is after calibration and both of them are tested using random forest with maximum number of feature as 5. It is easy to find that the variance estimate before calibration has a lot of negative number, which are invalid. Instead, after calibration, all the variance estimate becomes positive. I also test them under $m = 19$ and $m = 57$. Although the negative number decrease in larger m, there is still some negative variance estimate before the calibration. Note that now the bagging number $B = \Theta(n)$ which is also the recommended magniture for the Infinitesimal Jackknife, thus the calibration is necessary in this dataset.

Finally we summarize the empirical Bayes calibration algorithm used by Prof. Wager.

The remaining point in this algorithm is the negative likelihood function. Roughly speaking, the negative likelihood approximate the posterior distribution with the discrete distribution generated via $g_r(\hat{\boldsymbol{\eta}})$. Another point is they use mask in their algorithm to get rid of the negative points in the approximation to the posterior distribution.

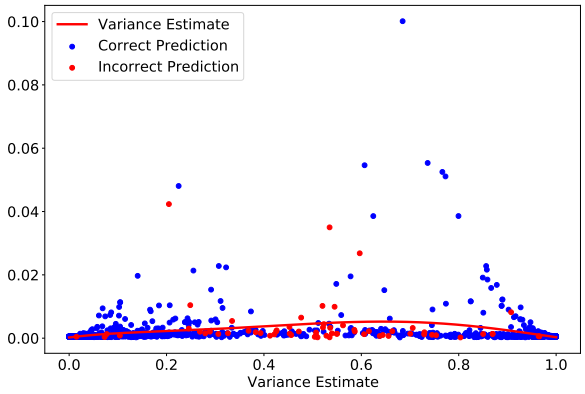


Figure 3: Infinitesimal Jackknife ($m = 5$)

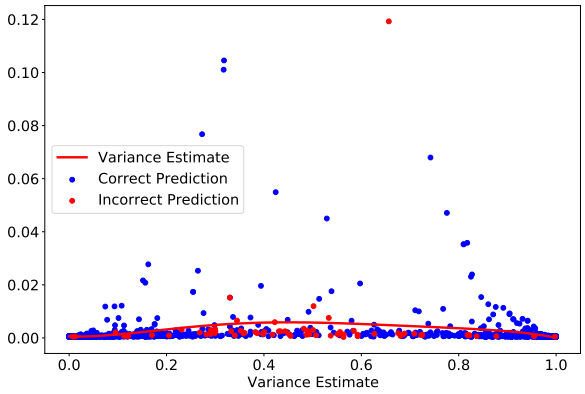


Figure 4: Infinitesimal Jackknife ($m = 19$)

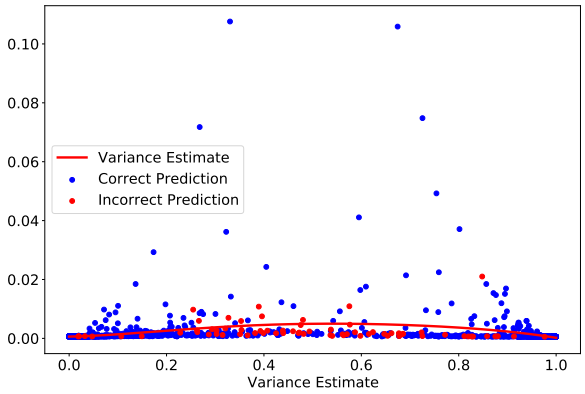
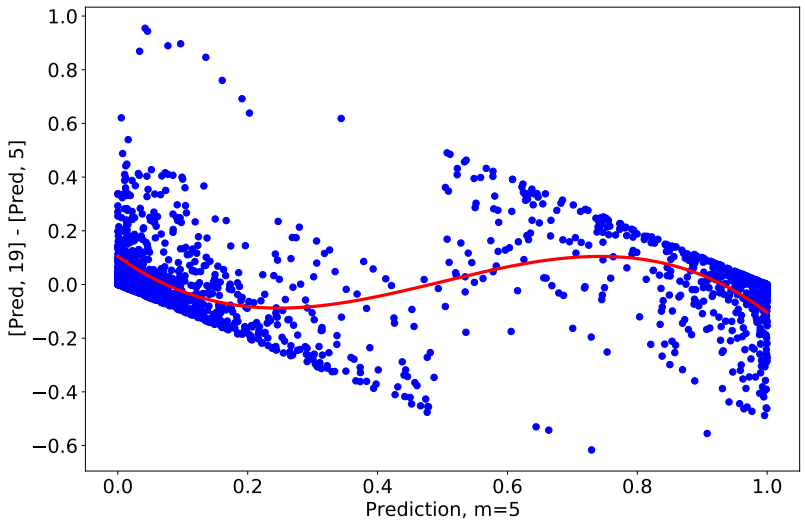


Figure 5: Infinitesimal Jackknife ($m = 57$)



Algorithm 1 Algorithm

Input: to calibrate vector $\boldsymbol{x}$, degree for modeling Gaussian prior p, number of bins b and noise level σ^2.

Output: calibrated vector $\boldsymbol{x}^{\mathrm{c}}$.

$x_{\min} = \min_i X_i - 2 * \mathrm{std}(\boldsymbol{X})$, $x_{\max} = \max_i X_i + 2 * \mathrm{std}(\boldsymbol{X})$

$\boldsymbol{x}_{\mathrm{vals}}$ be a set of size b and takes values uniformly between $x_{\min}$ and $x_{\max}$.

$\boldsymbol{k}$ be the pdf of standard normal distribution evaluated on x_{vals}.

$\boldsymbol{X}(m)$ be a matrix of size $n \times p$ with the ij-element as X_i to the power of j, i.e., X_i^j.

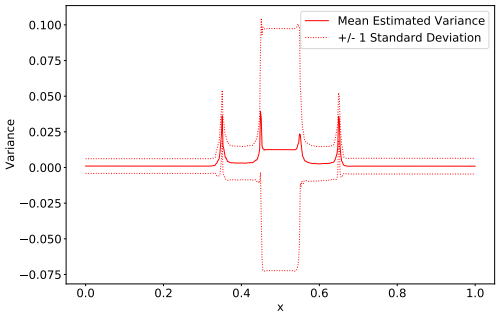
Minimizer a negative likelihood function to derive $\hat{\boldsymbol{\eta}}$ and $g_r(\hat{\boldsymbol{\eta}}) = \exp(\boldsymbol{X}(m)\hat{\boldsymbol{\eta}}) = \exp(\sum_{i=1}^{p} \boldsymbol{X}(m)[:, i]\hat{\eta}_i)$.

$g(\hat{\boldsymbol{\eta}}) = (1-\epsilon)\frac{g_r(\hat{\boldsymbol{\eta}})}{\sum_i (g_r(\hat{\boldsymbol{\eta}}))_i} + \epsilon \mathrm{Uniform}[x_{\min}, x_{\max}]$.

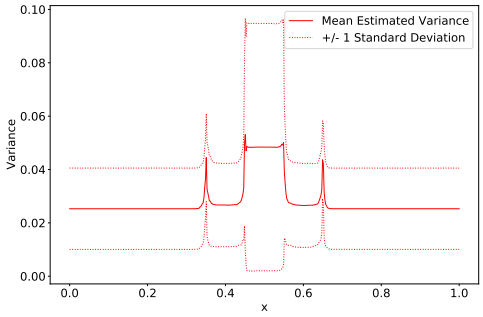
Derive the posterior expectation $\boldsymbol{x}^c = \mathbb{E}_{g(\hat{\boldsymbol{\eta}})}[\boldsymbol{x}]$.

3 Conclusion

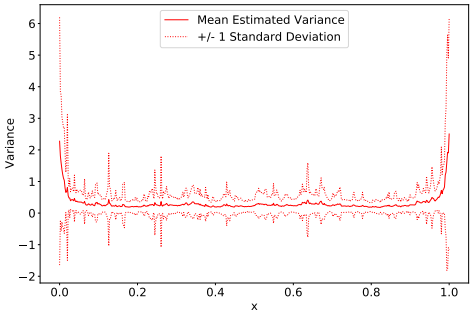
This paper analyze Infinitesimal Jackknife and Jackknife-after-Bootstrap estimator for constructing the confidence interval in bagging and random forest. It analyze the Monto Carlo error and propose a bias-correction version for both of them. Finally it recommends to use the the arithmetic mean of the jackknife and IJ estimators from the analysis of direction of the sampling bias. Note that the empirical Bayes calibration is essential for the algorithm to work in reality.



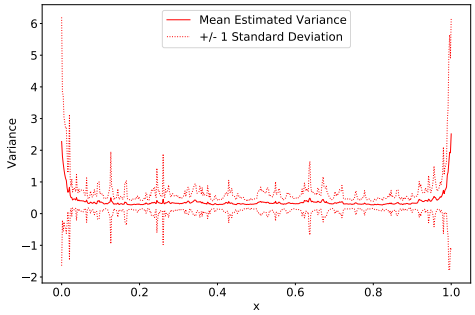
(a) Before Calibration ($\sigma^2 = 1/2^2$)



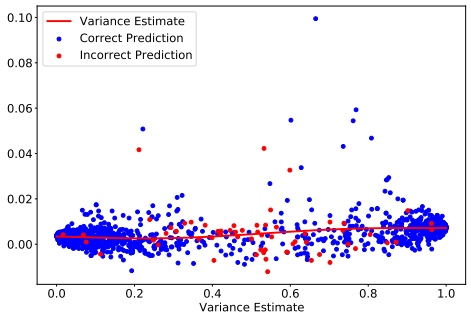
(b) After Calibration ($\sigma^2 = 1/2^2$)



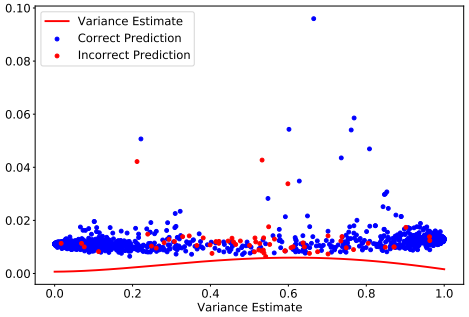
(c) Before Calibration ($\sigma^2 = 2^2$)



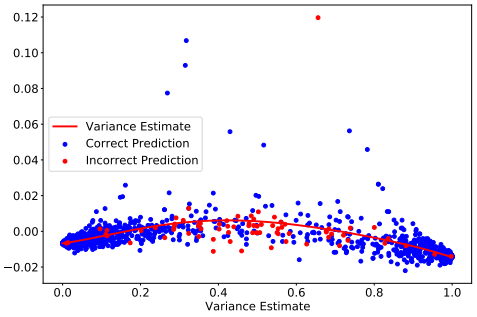
(d) After Calibration ($\sigma^2 = 2^2$)



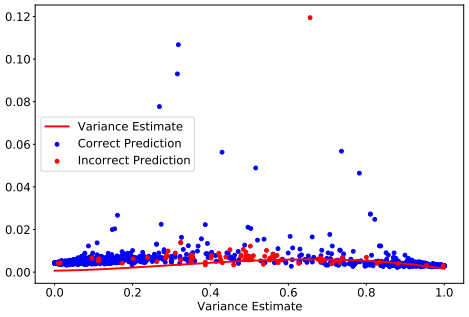
(a) Before Calibration and $m = 5$



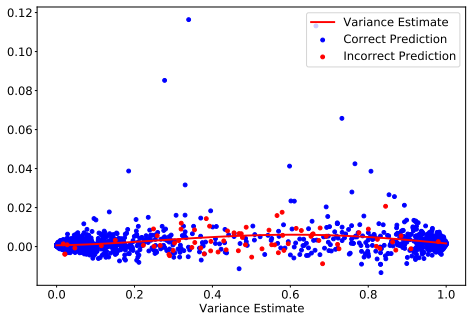
(b) After Calibration and $m = 5$



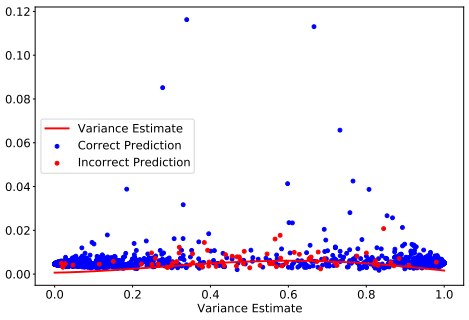
(c) Before Calibration and $m = 19$



(d) After Calibration and $m = 19$



(e) Before Calibration and $m = 57$



(f) After Calibration and $m = 57$

References

[1] Bradley Efron. *Model Selection Estimation and Bootstrap Smoothing.* Division of Biostatistics, Stanford University, 2012.

[2] Bradley Efron. Two modeling strategies for empirical bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(2):285, 2014.

[3] Bradley Efron and David Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991.

[4] Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

[5] Louis A Jaeckel. The infinitesimal jackknife. 1972.

[6] Colin L Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.