Confidence Intervals for Random Forests:
The Jackknife and the Infinitesimal Jackknife

Stefan Wager   swager@stanford.edu
Trevor Hastie   hastie@stanford.edu
Bradley Efron   brad@stat.stanford.edu
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Abstract

We study the variability of predictions made by bagged learners and random forests, and show how to estimate standard errors for these methods. Our work builds on variance estimates for bagging proposed by Efron (1992, 2013) that are based on the jackknife and the infinitesimal jackknife (IJ). In practice, bagged predictors are computed using a finite number B of bootstrap replicates, and working with a large B can be computationally expensive. Direct applications of jackknife and IJ estimators to bagging require $B = \Theta(n^{1.5})$ bootstrap replicates to converge, where n is the size of the training set. We propose improved versions that only require $B = \Theta(n)$ replicates. Moreover, we show that the IJ estimator requires 1.7 times less bootstrap replicates than the jackknife to achieve a given accuracy. Finally, we study the sampling distributions of the jackknife and IJ variance estimates themselves. We illustrate our findings with multiple experiments and simulation studies.

Keywords: bagging, jackknife methods, Monte Carlo noise, variance estimation

1. Introduction

Bagging (Breiman, 1996) is a popular technique for stabilizing statistical learners. Bagging is often conceptualized as a variance reduction technique, and so it is important to understand how the sampling variance of a bagged learner compares to the variance of the original learner. In this paper, we develop and study methods for estimating the variance of bagged predictors and random forests (Breiman, 2001), a popular extension of bagged trees. These variance estimates only require the bootstrap replicates that were used to form the bagged prediction itself, and so can be obtained with moderate computational overhead. The results presented here build on the jackknife-after-bootstrap methodology introduced by Efron (1992) and on the infinitesimal jackknife for bagging (IJ) (Efron, 2013).

Figure 1 shows the results from applying our method to a random forest trained on the "Auto MPG" data set, a regression task where we aim to predict the miles-per-gallon (MPG) gas consumption of an automobile based on 7 features including weight and horsepower. The error bars shown in Figure 1 give an estimate of the sampling variance of the random forest; in other words, they tell us how much the random forest's predictions might change if we
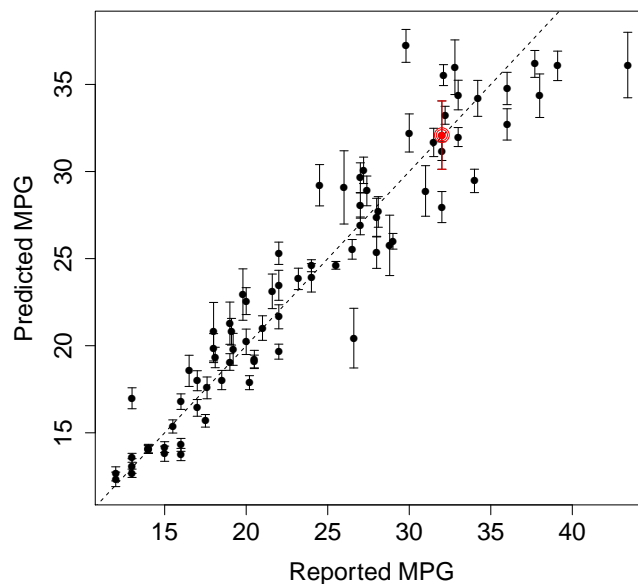
Figure 1: Random forest predictions on the "Auto MPG" data set. The random forest was trained using 314 examples; the graph shows results on a test set of size 78. The error bars are 1 standard error in each direction. Because this is a fairly small data set, we estimated standard errors for the random forest using the averaged estimator from Section 5.2. A more detailed description of the experiment is provided in Appendix C.

trained it on a new training set. The fact that the error bars do not in general cross the prediction-equals-observation diagonal suggests that there is some residual noise in the MPG of a car that cannot be explained by a random forest model based on the available predictor variables.[1]

Figure 1 tells us that the random forest was more confident about some predictions than others. Rather reassuringly, we observe that the random forest was in general less confident about the predictions for which the reported MPG and predicted MPG were very different. There is not a perfect correlation, however, between the error level and the size of the error bars. One of the points, circled in red near (32, 32), appears particularly surprising: the random forest got the prediction almost exactly right, but gave the prediction large error bars of ± 2. This curious datapoint corresponds to the 1982 Dodge Rampage, a two-door Coupe Utility that is a mix between a passenger car and a truck with a cargo tray. Perhaps our random forest had a hard time confidently estimating the mileage of the Rampage

---

1. Our method produces standard error estimates $\hat{\sigma}$ for random forest predictions. We then represent these standard error estimates as Gaussian confidence intervals $\hat{y} \pm z_\alpha \hat{\sigma}$, where z_α is a quantile of the normal distribution.

because it could not quite decide whether to cluster it with cars or with trucks. We present experiments on larger data sets in Section 3.

Estimating the variance of bagged learners based on the preexisting bootstrap replicates can be challenging, as there are two distinct sources of noise. In addition to the sampling noise (i.e., the noise arising from randomness during data collection), we also need to control the Monte Carlo noise arising from the use of a finite number of bootstrap replicates. We study the effects of both sampling noise and Monte Carlo noise.

In our experience, the errors of the jackknife and IJ estimates of variance are often dominated by Monte Carlo effects. Monte Carlo bias can be particularly troublesome: if we are not careful, the jackknife and IJ estimators can conflate Monte Carlo noise with the underlying sampling noise and badly overestimate the sampling variance. We show how to estimate the magnitude of this Monte Carlo bias and develop bias-corrected versions of the jackknife and IJ estimators that outperform the original ones. We also show that the IJ estimate of variance is able to use the preexisting bootstrap replicates more efficiently than the jackknife estimator by having a lower Monte Carlo variance, and needs 1.7 times less bootstrap replicates than the jackknife to achieve a given accuracy.

If we take the number of bootstrap replicates to infinity, Monte Carlo effects disappear and only sampling errors remain. We compare the sampling biases of both the jackknife and IJ rules and present some evidence that, while the jackknife rule has an upward sampling bias and the IJ estimator can have a downward bias, the arithmetic mean of the two variance estimates can be close to unbiased. We also propose a simple method for estimating the sampling variance of the IJ estimator itself.

Our paper is structured as follows. We first present an overview of our main results in Section 2, and apply them to random forest examples in Section 3. We then take a closer look at Monte Carlo effects in Section 4 and analyze the sampling distribution of the limiting IJ and jackknife rules with $B \to \infty$ in Section 5. We spread simulation experiments throughout Sections 4 and 5 to validate our theoretical analysis.

1.1 Related Work

In this paper, we focus on methods based on the jackknife and the infinitesimal jackknife for bagging (Efron, 1992, 2013) that let us estimate standard errors based on the pre-existing bootstrap replicates. Other approaches that rely on forming second-order bootstrap replicates have been studied by Duan (2011) and Sexton and Laake (2009). Directly bootstrapping a random forest is usually not a good idea, as it requires forming a large number of base learners. Sexton and Laake (2009), however, propose a clever work-around to this problem. Their approach, which could have been called a bootstrap of little bags, involves bootstrapping small random forests with around $B = 10$ trees and then applying a bias correction to remove the extra Monte Carlo noise.

There has been considerable interest in studying classes of models for which bagging can achieve meaningful variance reduction, and also in outlining situations where bagging can fail completely (e.g., Skurichina and Duin, 1998; Bühlmann and Yu, 2002; Chen and Hall, 2003; Buja and Stuetzle, 2006; Friedman and Hall, 2007). The problem of producing practical estimates of the sampling variance of bagged predictors, however, appears to have received somewhat less attention in the literature so far.

2. Estimating the Variance of Bagged Predictors

This section presents our main result: estimates of variance for bagged predictors that can be computed from the same bootstrap replicates that give the predictors. Section 3 then applies the result to random forests, which can be analyzed as a special class of bagged predictors.

Suppose that we have training examples $Z_1 = (x_1, y_1)$, ..., $Z_n = (x_n, y_n)$, an input x to a prediction problem, and a base learner $\hat{\theta}(x) = t(x; Z_1, ..., Z_n)$. To make things concrete, the Z_i could be a list of e-mails x_i paired with labels y_i that catalog the e-mails as either spam or non-spam, $t(x; Z_i)$ could be a decision tree trained on these labeled e-mails, and x could be a new e-mail that we seek to classify. The quantity $\hat{\theta}(x)$ would then be the output of the tree predictor on input x.

With bagging, we aim to stabilize the base learner t by resampling the training data. In our case, the bagged version of $\hat{\theta}(x)$ is defined as

$$\hat{\theta}^\infty(x) = \mathbb{E}_*[t(x; Z_1^*, ..., Z_n^*)], \tag{1}$$

where the Z_i^* are drawn independently with replacement from the original data (i.e., they form a bootstrap sample). The expectation $\mathbb{E}_*$ is taken with respect to the bootstrap measure.

The expectation in (1) cannot in general be evaluated exactly, and so we form the bagged estimator by Monte Carlo

$$\hat{\theta}^B(x) = \frac{1}{B} \sum_{b=1}^{B} t_b^*(x), \text{ where } t_b^*(x) = t(x; Z_{b1}^*, ..., Z_{bn}^*) \tag{2}$$

and the Z_{bi}^* are elements in the bth bootstrap sample. As $B \to \infty$, we recover the perfectly bagged estimator $\hat{\theta}^\infty(x)$.

2.1 Basic Variance Estimates

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

$$V(x) = \text{Var}\left[\hat{\theta}^\infty(x)\right].$$

In other words, we ask how much variance $\hat{\theta}^B$ would have once we make B large enough to eliminate the bootstrap effects. We consider two basic estimates of V: The *Infinitesimal Jackknife* estimate (Efron, 2013), which results in the simple expression

$$\widehat{V}_{IJ}^\infty = \sum_{i=1}^{n} \text{Cov}_*[N_i^*, t^*(x)]^2, \tag{3}$$

where $\text{Cov}_*[N_i^*, t^*(x)]$ is the covariance between $t^*(x)$ and the number of times N_i^* the ith training example appears in a bootstrap sample; and the *Jackknife-after-Bootstrap* estimate (Efron, 1992)

$$\widehat{V}_J^\infty = \frac{n-1}{n} \sum_{i=1}^{n} \left(\bar{t}_{(-i)}^*(x) - \bar{t}^*(x)\right)^2, \tag{4}$$

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

The jackknife-after-bootstrap estimate $\widehat{V}_J^\infty$ arises directly by applying the jackknife to the bootstrap distribution. The infinitesimal jackknife (Jaeckel, 1972), also called the non-parametric delta method, is an alternative to the jackknife where, instead of studying the behavior of a statistic when we remove one observation at a time, we look at what happens to the statistic when we individually down-weight each observation by an infinitesimal amount. When the infinitesimal jackknife is available, it sometimes gives more stable predictions than the regular jackknife. Efron (2013) shows how an application of the infinitesimal jackknife principle to the bootstrap distribution leads to the simple estimate $\widehat{V}_{IJ}^\infty$.

2.2 Finite-B Bias

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

$$\widehat{V}_{IJ}^B = \sum_{i=1}^n \widehat{\mathrm{Cov}}_i^2 \text{ with } \widehat{\mathrm{Cov}}_i = \frac{\sum_b (N_{bi}^* - 1)(t_b^*(x) - \bar{t}^*(x))}{B}, \tag{5}$$

and

$$\widehat{V}_J^B = \frac{n-1}{n} \sum_{i=1}^n \hat{\Delta}_i^2, \text{ where } \hat{\Delta}_i = \hat{\theta}_{(-i)}^B(x) - \hat{\theta}^B(x) \tag{6}$$

$$\text{and } \hat{\theta}_{(-i)}^B(x) = \frac{\sum_{\{b\,:\,N_{bi}^* = 0\}} t_b^*(x)}{\left|\{N_{bi}^* = 0\}\right|}.$$

Here, N_{bi}^* indicates the number of times the ith observation appears in the bootstrap sample b.

In our experience, these finite-B estimates of variance are often badly biased upwards if the number of bootstrap samples B is too small. Fortunately, bias-corrected versions are available:

$$\widehat{V}_{IJ-U}^B = \widehat{V}_{IJ}^B - \frac{n}{B^2} \sum_{b=1}^B (t_b^*(x) - \bar{t}^*(x))^2, \text{ and} \tag{7}$$

$$\widehat{V}_{J-U}^B = \widehat{V}_J^B - (e-1) \frac{n}{B^2} \sum_{b=1}^B (t_b^*(x) - \bar{t}^*(x))^2. \tag{8}$$

These bias corrections are derived in Section 4. In many applications, the simple estimators (5) and (6) require $B = \Theta(n^{1.5})$ bootstrap replicates to reduce Monte Carlo noise down to the level of the inherent sampling noise, whereas our bias-corrected versions only require $B = \Theta(n)$ replicates. The bias-corrected jackknife (8) was also discussed by Sexton and Laake (2009).

In Figure 2, we show how $\widehat{V}_{IJ-U}^B$ can be used to accurately estimate the variance of a bagged tree. We compare the true sampling variance of a bagged regression tree with our variance estimate. The underlying signal is a step function with four jumps that are reflected
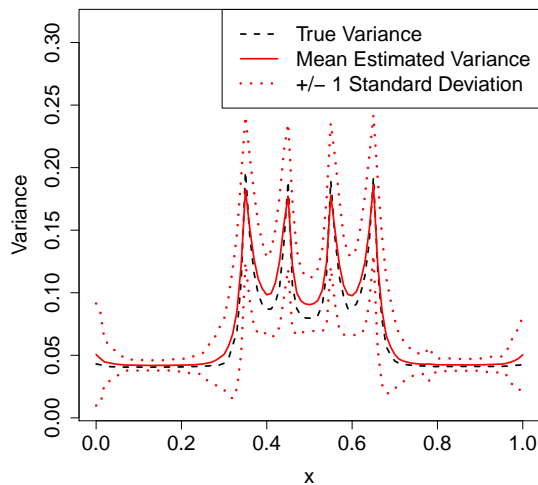
Figure 2: Testing the performance of the bias-corrected infinitesimal jackknife estimate of variance for bagged predictors, as defined in (15), on a bagged regression tree. We compare the true sampling error with the average standard error estimate produced by our method across multiple runs; the dotted lines indicate one-standard-error-wide confidence bands for our standard error estimate.

as spikes in the variance of the bagged tree. On average, our variance estimator accurately identifies the location and magnitude of these spikes.

Figure 3 compares the performance of the four considered variance estimates on a bagged adaptive polynomial regression example described in detail in Section 4.4. We see that the uncorrected estimators $\widehat{V}_J^B$ and $\widehat{V}_{IJ}^B$ are badly biased: the lower whiskers of their boxplots do not even touch the limiting estimate with $B \to \infty$. We also see that that $\widehat{V}_{IJ-U}^B$ has about half the variance of $\widehat{V}_{J-U}^B$. This example highlights the importance of using estimators that use available bootstrap replicates efficiently: with $B = 500$ bootstrap replicates, $\widehat{V}_{IJ-U}^B$ can give us a reasonable estimate of V, whereas $\widehat{V}_J^B$ is quite unstable and biased upwards by a factor 2.

The figure also suggests that the Monte Carlo noise of $\widehat{V}_{IJ}^B$ decays faster (as a function of B) than that of $\widehat{V}_J^B$. This is no accident: as we show in Section 4.2, the infinitesimal jackknife requires 1.7 times less bootstrap replicates than the jackknife to achieve a given level of level of Monte Carlo error.

2.3 Limiting Sampling Distributions

The performance of $\widehat{V}_J^B$ and $\widehat{V}_{IJ}^B$ depends on both sampling noise and Monte Carlo noise. In order for $\widehat{V}_J^B$ (and analogously $\widehat{V}_{IJ}^B$) to be accurate, we need both the sampling error of $\widehat{V}_J^\infty$, namely $\widehat{V}_J^\infty - V$, and the Monte Carlo error $\widehat{V}_J^B - \widehat{V}_J^\infty$ to be small.
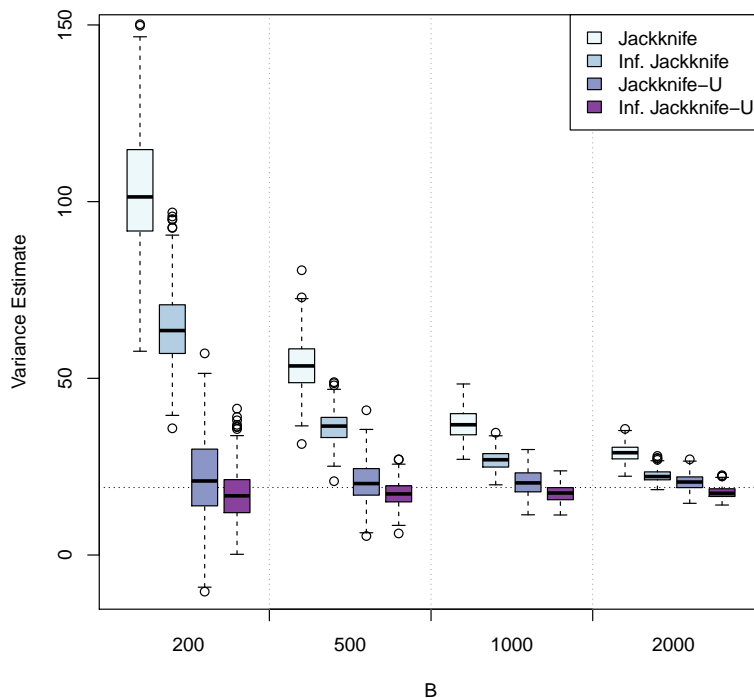
Figure 3: Performance, as a function of B, of the jackknife and IJ estimators and their bias-corrected modifications on the cholesterol data set of Efron and Feldman (1991). The boxplots depict bootstrap realizations of each estimator. The dotted line indicates the mean of all the realizations of the IJ-U and J-U estimators (weighted by B).

It is well known that jackknife estimates of variance are in general biased upwards (Efron and Stein, 1981). This phenomenon also holds for bagging: $\widehat{V}_J^\infty$ is somewhat biased upwards for V. We present some evidence suggesting that $\widehat{V}_{IJ}^\infty$ is biased downwards by a similar amount, and that the arithmetic mean of $\widehat{V}_J^\infty$ and $\widehat{V}_{IJ}^\infty$ is closer to being unbiased for V than either of the two estimators alone.

We also develop a simple estimator for the variance of $\widehat{V}_{IJ}^\infty$ itself:

$$\widehat{\mathrm{Var}\left[\widehat{V}_{IJ}^\infty\right]} = \sum_{i=1}^n \left(C_i^{*,2} - \overline{C_i^{*,2}}\right)^2,$$

where $C_i^* = \mathrm{Cov}_*[N_{bi}^*, t_b^*(x)]$ and $\overline{C_i^{*,2}}$ is the mean of the $C_i^{*,2}$.

3. Random Forest Experiments

Random forests (Breiman, 2001) are a widely used extension of bagged trees. Suppose that we have a tree-structured predictor t and training data $Z_1, ..., Z_n$. Using notation from (2), the bagged version of this tree predictor is

$$\hat{\theta}^B(x) = \frac{1}{B} \sum_{b=1}^{B} t_b^*(x;\, Z_{b1}^*,\, ...,\, Z_{bn}^*).$$

Random forests extend bagged trees by allowing the individual trees t_b^* to depend on an auxiliary noise source ξ_b. The main idea is that the auxiliary noise ξ_b encourages more diversity among the individual trees, and allows for more variance reduction than bagging. Several variants of random forests have been analyzed theoretically by, e.g., Biau et al. (2008), Biau (2012), Lin and Jeon (2006), and Meinshausen (2006).

Standard implementations of random forests use the auxiliary noise ξ_b to randomly restrict the number of variables on which the bootstrapped trees can split at any given training step. At each step, m features are randomly selected from the pool of all p possible features and the tree predictor must then split on one of these m features. If $m = p$ the tree can always split on any feature and the random forest becomes a bagged tree; if $m = 1$, then the tree has no freedom in choosing which feature to split on.

Following Breiman (2001), random forests are usually defined more abstractly for theoretical analysis: any predictor of the form

$$\hat{\theta}^{RF}(x) = \frac{1}{B} \sum_{b=1}^{B} t_b^*(x;\, \xi_b,\, Z_{b1}^*,\, ...,\, Z_{bn}^*) \text{ with } \xi_b \overset{\text{iid}}{\sim} \Xi \tag{9}$$

is called a *random forest*. Various choices of noise distribution Ξ lead to different random forest predictors. In particular, trivial noise sources are allowed and so the class of random forests includes bagged trees as a special case. In this paper, we only consider random forests of type (9) where individual trees are all trained on bootstrap samples of the training data. We note, however, that that variants of random forests that do not use bootstrap noise have also been found to work well (e.g., Dietterich, 2000; Geurts et al., 2006).

All our results about bagged predictors apply directly to random forests. The reason for this is that random forests can also be defined as bagged predictors with different base learners. Suppose that, on each bootstrap replicate, we drew K times from the auxiliary noise distribution Ξ instead of just once. This would give us a predictor of the form

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

Adding the extra draws from Ξ to the random forest does not change the $B \to \infty$ limit of the random forest. If we take $K \to \infty$, we effectively marginalize over the noise from Ξ, and get a predictor

$$\hat{\theta}^{\widetilde{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} \tilde{t}_b^*(x;\, Z_{b1}^*,\, ...,\, Z_{bn}^*), \text{ where}$$

$$\tilde{t}(x;\, Z_1,\, ...,\, Z_n) = \mathbb{E}_{\xi \sim \Xi}\left[t(x;\, \xi,\, Z_1,\, ...,\, Z_n)\right].$$

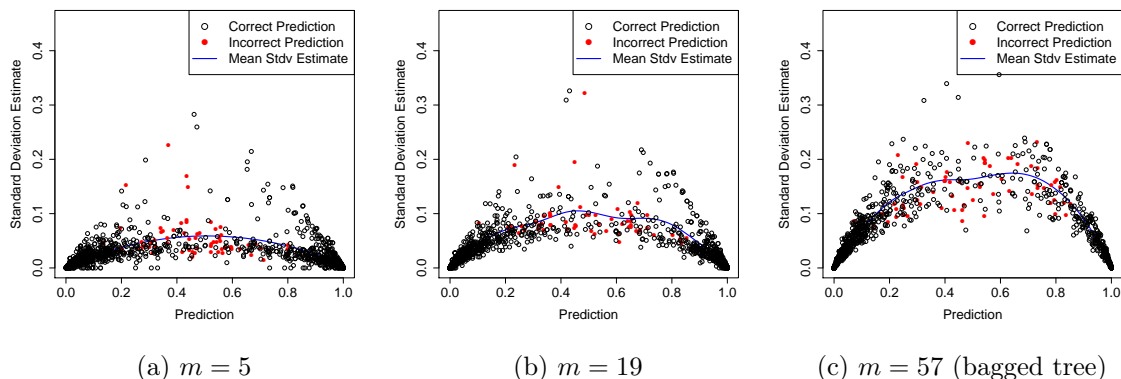(a) $m = 5$ (b) $m = 19$ (c) $m = 57$ (bagged tree)

Figure 4: Standard errors of random forest predictions on the e-mail spam data. The random forests with $m = 5$, 19, and 57 splitting variables were all trained on a train set of size $n = 3,065$; the panels above show class predictions and IJ-U estimates for standard errors on a test set of size 1,536. The solid curves are smoothing splines ($df = 4$) fit through the data (including both correct and incorrect predictions).

In other words, the random forest $\hat{\theta}^{RF}$ as defined in (9) is just a noisy estimate of a bagged predictor with base learner $\tilde{t}$.

It is straight-forward to check that our results about $\widehat{V}_{IJ}^B$ and $\widehat{V}_J^B$ also hold for bagged predictors with randomized base learners. The extra noise from using $t(\cdot; \xi)$ instead of $\tilde{t}(\cdot)$ does not affect the limiting correlations in (3) and (4); meanwhile, the bias corrections from (7) and (8) do not depend on how we produced the t^* and remain valid with random forests. Thus, we can estimate confidence intervals for random forests from N^* and t^* using exactly the same formulas as for bagging.

In the rest of this section, we show how the variance estimates studied in this paper can be used to gain valuable insights in applications of random forests. We use the $\widehat{V}_{IJ-U}^B$ variance estimate (7) to minimize the required computational resources. We implemented the IJ-U estimator for random forests on top of the R package randomForest (Liaw and Wiener, 2002).

3.1 E-mail Spam Example

The e-mail spam data set (spambase) is part of a standard classification task, the goal of which is to distinguish spam e-mail (1) from non-spam (0) using $p = 57$ features. Here, we investigate the performance of random forests on this data set.

We fit the spam data using random forests with $m = 5$, 19 and 57 splitting variables. With $m = 5$, the trees were highly constrained in their choice of splitting variables, while $m = 57$ is just a bagged tree. The three random forests obtained test-set accuracies of 95.1%, 95.2% and 94.7% respectively, and it appears that the $m = 5$ or 19 forests are best. We can use the IJ-U variance formula to gain deeper insight into these numbers, and get a better understanding about what is constraining the accuracy of each predictor.
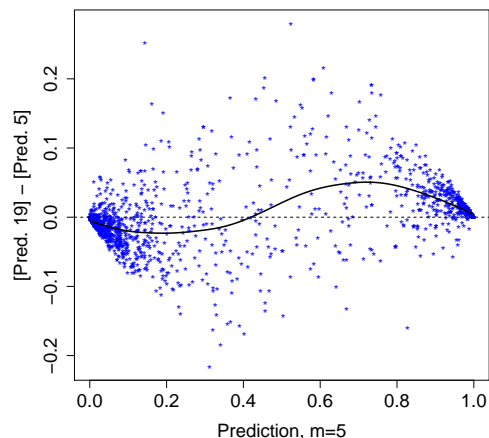
Figure 5: Comparison of the predictions made by the $m = 5$ and $m = 19$ random forests. The stars indicate pairs of test set predictions; the solid line is a smoothing spline ($df = 6$) fit through the data.

In Figure 4, we plot test-set predictions against IJ-U estimates of standard error for all three random forests. The $m = 57$ random forest appears to be quite unstable, in that the estimated errors are high. Because many of its predictions have large standard errors, it is plausible that the predictions made by the random forest could change drastically if we got more training data. Thus, the $m = 57$ forest appears to suffer from overfitting, and the quality of its predictions could improve substantially with more data.

Conversely, predictions made by the $m = 5$ random forest appear to be remarkably stable, and almost all predictions have standard errors that lie below 0.1. This suggests that the $m = 5$ forest may be mostly constrained by bias: if the predictor reports that a certain e-mail is spam with probability 0.5 ± 0.1, then the predictor has effectively abandoned any hope of unambiguously classifying the e-mail. Even if we managed to acquire much more training data, the class prediction for that e-mail would probably not converge to a strong vote for spam or non-spam.

The $m = 19$ forest appears to have balanced the bias-variance trade-off well. We can further corroborate our intuition about the bias problem faced by the $m = 5$ forest by comparing its predictions with those of the $m = 19$ forest. As shown in Figure 5, whenever the $m = 5$ forest made a cautious prediction that an e-mail might be spam (e.g., a prediction of around 0.8), the $m = 19$ forest made the same classification decision but with more confidence (i.e., with a more extreme class probability estimate $\hat{p}$). Similarly, the $m = 19$ forest tended to lower cautious non-spam predictions made by the $m = 5$ forest. In other words, the $m = 5$ forest appears to have often made lukewarm predictions with mid-range values of $\hat{p}$ on e-mails for which there was sufficient information in the data to make confident predictions. This analysis again suggests that the $m = 5$ forest was constrained by bias and was not able to efficiently use all the information present in the data set.

(a) MSE and mean variance
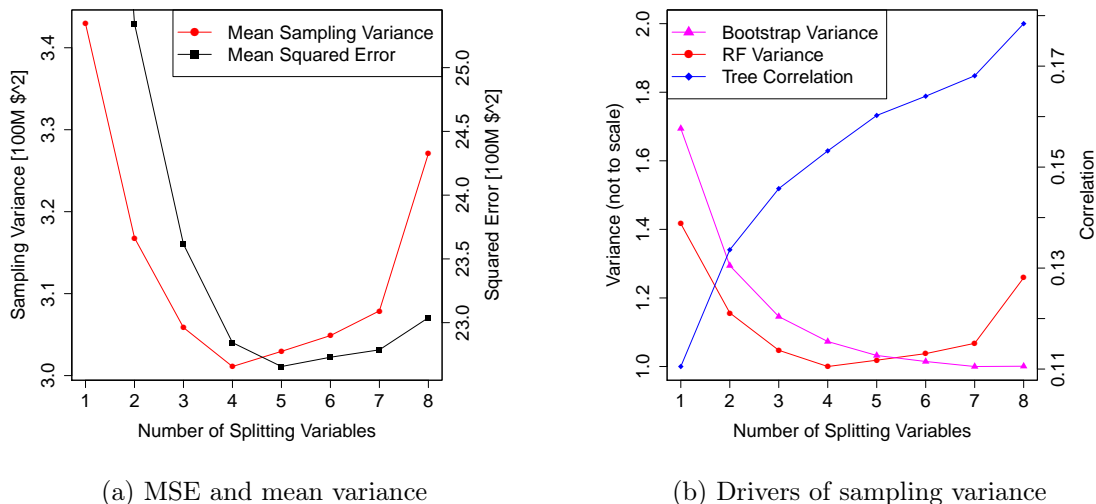
(b) Drivers of sampling variance

Figure 6: Performance of random forests on the California housing data. The left panel plots MSE and mean sampling variance as a function of the number m of splitting variables. The MSE estimate is the out-of bag error, while the mean sampling variance is the average estimate of variance $\widehat{V}_{IJ-U}^{B}$ computed over all training examples. The right panel displays the drivers of sampling variance, namely the variance of the individual bootstrapped trees (bootstrap variance v) and their correlation (tree correlation ρ).

3.2 California Housing Example

In the previous example, we saw that the varying accuracy of random forests with different numbers m of splitting variables primarily reflected a bias-variance trade-off. Random forests with small m had high bias, while those with large m had high variance. This bias-variance trade-off does not, however, underlie all random forests. The California housing data set—a regression task with $n = 20,460$ and $p = 8$—provides a contrasting example.

In Figure 6a, we plot the random forest out-of-bag MSE and IJ-U estimate of average sampling variance across all training examples, with m between 1 and 8. We immediately notice that the sampling variance is not monotone increasing in m. Rather, the sampling variance is high if m is too big or too small, and attains a minimum at $m = 4$. Meanwhile, in terms of MSE, the optimal choice is $m = 5$. Thus, there is no bias-variance trade-off here: picking a value of m around 4 or 5 is optimal both from the MSE minimization and the variance minimization points of view.

We can gain more insight into this phenomenon using ideas going back to Breiman (2001), who showed that the sampling variance of a random forest is governed by two factors: the variance v of the individual bootstrapped trees and their correlation ρ. The variance of the ensemble is then ρv. In Figure 6b, we show how both v and ρ react when we vary m. Trees with large m are fairly correlated, and so the random forest does not get as substantial a variance reduction over the base learner as with a smaller m. With a very small m, however,

the variance v of the individual trees shoots up, and so the decrease in ρ is no longer sufficient to bring down the variance of the whole forest. The increasing ρ-curve and the decreasing v-curve thus jointly produce a U-shaped relationship between m and the variance of the random forest. The $m = 4$ forest achieves a low variance by matching fairly stable base learners with a small correlation ρ.

4. Controlling Monte Carlo Error

In this section, we analyze the behavior of both the IJ and jackknife estimators under Monte Carlo noise. We begin by discussing the Monte Carlo distribution of the infinitesimal jackknife estimate of variance with a finite B; the case of the jackknife-after-bootstrap estimate of variance is similar but more technical and is presented in Appendix A. We show that the jackknife estimator needs 1.7 times more bootstrap replicates than the IJ estimator to control Monte Carlo noise at a given level. We also highlight a bias problem for both estimators, and recommend a bias correction. When there is no risk of ambiguity, we use the short-hand t^* for $t^*(x)$.

4.1 Monte Carlo Error for the IJ Estimator

We first consider the Monte Carlo bias of the infinitesimal jackknife for bagging. Let

$$\widehat{V}_{IJ}^{\infty} = \sum_{i=1}^{n} \mathrm{Cov}_*[N_i^*, \, t^*]^2 \tag{10}$$

be the perfect IJ estimator with $B = \infty$ (Efron, 2013). Then, the Monte Carlo bias of $\widehat{V}_{IJ}^B$ is

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

is the Monte Carlo estimate of the bootstrap covariance. Since t_b^* depends on all n observations, N_{bi}^* and t_b^* can in practice be treated as independent for computing $\mathrm{Var}_*[C_i]$, especially when n is large (see remark below). Thus, as $\mathrm{Var}_*[N_{bi}^*] = 1$, we see that

$$\mathbb{E}_* \left[\widehat{V}_{IJ}^B \right] - \widehat{V}_{IJ}^{\infty} \approx \frac{n\,\hat{v}}{B}, \text{ where } \hat{v} = \frac{1}{B} \sum_{b=1}^{B} (t_b^* - \bar{t}^*)^2. \tag{11}$$

Notice that $\hat{v}$ is the standard bootstrap estimate for the variance of the base learner $\hat{\theta}(x)$. Thus, the bias of $\widehat{V}_{IJ}^B$ grows linearly in the variance of the original estimator that is being bagged.

Meanwhile, by the central limit theorem, C_i converges to a Gaussian random variable as B gets large. Thus, the Monte Carlo asymptotic variance of C_i^2 is approximately $2\,\mathrm{Var}_*[C_i]^2 + 4\,\mathbb{E}_*[C_i]^2\,\mathrm{Var}_*[C_i]$. The C_i can be treated as roughly independent, and so the limiting distribution of the IJ estimate of variance has approximate moments

$$\widehat{V}_{IJ}^B - \widehat{V}_{IJ}^{\infty} \, \dot{\sim} \, \left(\frac{n\,\hat{v}}{B}, \, 2\,\frac{n\,\hat{v}^2}{B^2} + 4\,\frac{\widehat{V}_{IJ}^{\infty}\,\hat{v}}{B} \right). \tag{12}$$

Interestingly, the Monte Carlo mean squared error (MSE) of $\widehat{V}_{IJ}^{B}$ mostly depends on the problem through $\hat{v}$, where $\hat{v}$ is the bootstrap estimate of the variance of the base learner. In other words, the computational difficulty of obtaining confidence intervals for bagged learners depends on the variance of the base learner.

4.1.1 Remark: The IJ Estimator for Sub-bagging

We have focused on the case where each bootstrap replicate contains exactly n samples. However, in some applications, bagging with subsamples of size $m \neq n$ has been found to work well (e.g., Bühlmann and Yu, 2002; Buja and Stuetzle, 2006; Friedman, 2002; Strobl et al., 2007). Our results directly extend to the case where $m \neq n$ samples are drawn with replacement from the original sample. We can check that (10) still holds, but now $\mathrm{Var}\, N_{bi}^{*} = m/n$. Carrying out the same analysis as above, we can establish an analogue to (12):

$$\widehat{V}_{IJ}^{B}(m) - \widehat{V}_{IJ}^{\infty}(m) \dot{\sim} \left(\frac{m\,\hat{v}}{B},\, 2\,\frac{m^2\,\hat{v}^2}{nB^2} + 4\,\frac{m\widehat{V}_{IJ}^{\infty}\,\hat{v}}{nB} \right). \tag{13}$$

For simplicity of exposition, we will restrict our analysis to the case $m = n$ for the rest of this paper.

4.1.2 Remark: Approximate Independence

In the above derivation, we used the approximation

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

We can evaluate the accuracy of this approximation using the formula

$$\mathrm{Var}_* \left[(N_{bi}^* - 1)(t_b^* - \bar{t}^*) \right] - \mathrm{Var}_* \left[N_{bi}^* \right] \mathrm{Var} \left[t_b^* \right]$$
$$= \mathrm{Cov}_* \left[(N_{bi}^* - 1)^2,\, (t_b^* - \bar{t}^*)^2 \right] - \mathrm{Cov}_* \left[(N_{bi}^* - 1),\, (t_b^* - \bar{t}^*) \right]^2.$$

In the case of the sample mean $t(Z_1^*, ..., Z_n^*) = \frac{1}{n} \sum_i Z_i^*$ paired with the Poisson bootstrap, this term reduces to

$$\mathrm{Cov}_* \left[(N_{bi}^* - 1)^2,\, (t_b^* - \bar{t}^*)^2 \right] - \mathrm{Cov}_* \left[(N_{bi}^* - 1),\, (t_b^* - \bar{t}^*) \right]^2 = 2\frac{\left(Z_i - \bar{Z}\right)^2}{n^2},$$

and the correction to (11) would be $2\hat{v}/(nB) \ll n\hat{v}/B$.

4.2 Comparison of Monte Carlo Errors

As shown in Appendix A, the Monte Carlo error for the jackknife-after-bootstrap estimate of variance has approximate moments

$$\widehat{V}_J^{B} - \widehat{V}_J^{\infty} \dot{\sim} \left((e-1)\frac{n\,\hat{v}}{B},\, 2\,(e-1)^2\frac{n\,\hat{v}^2}{B^2} + 4\,(e-1)\frac{\widehat{V}_J^{\infty}\,\hat{v}}{B} \right), \tag{14}$$

where $\widehat{V}_J^{\infty}$ is the jackknife estimate computed with $B = \infty$ bootstrap replicates. The Monte Carlo stability of $\widehat{V}_J^{B}$ again primarily depends on $\hat{v}$.

By comparing (12) with (14), we notice that the IJ estimator makes better use of a finite number B of bootstrap replicates than the jackknife estimator. For a fixed value of B, the Monte Carlo bias of $\widehat{V}_J^B$ is about $e - 1$ or 1.7 times as large as that of $\widehat{V}_{IJ}^B$; the ratio of Monte Carlo variance starts off at 3 for small values of B and decays down to 1.7 as B gets much larger than n. Alternatively, we see that the IJ estimate with B bootstrap replicates has errors on the same scale as the jackknife estimate with $1.7 \cdot B$ replicates.

This suggests that if computational considerations matter and there is a desire to perform as few bootstrap replicates B as possible while controlling Monte Carlo error, the infinitesimal jackknife method may be preferable to the jackknife-after-bootstrap.

4.3 Correcting for Monte Carlo Bias

The Monte Carlo MSEs of $\widehat{V}_{IJ}^B$ and $\widehat{V}_J^B$ are in practice dominated by bias, especially for large n. Typically, we would like to pick B large enough to keep the Monte Carlo MSE on the order of $1/n$. For both (12) and (14), we see that performing $B = \Theta(n)$ bootstrap iterations is enough to control the variance. To reduce the bias to the desired level, namely $\mathcal{O}(n^{-0.5})$, we would need to take $B = \Theta(n^{1.5})$ bootstrap samples.

Although the Monte Carlo bias for both $\widehat{V}_{IJ}^B$ and $\widehat{V}_J^B$ is large, this bias only depends on $\hat{v}$ and so is highly predictable. This suggests a bias-corrected modification of the IJ and jackknife estimators respectively:

$$\widehat{V}_{IJ-U}^B = \widehat{V}_{IJ}^B - \frac{n\,\hat{v}}{B}, \text{ and} \tag{15}$$

$$\widehat{V}_{J-U}^B = \widehat{V}_J^B - (e-1)\,\frac{n\,\hat{v}}{B}. \tag{16}$$

Here $\widehat{V}_{IJ}^B$ and $\widehat{V}_J^B$ are as defined in (5), and $\hat{v}$ is the bootstrap estimate of variance from (11). The letter U stands for unbiased. This transformation effectively removes the Monte Carlo bias in our experiments without noticeably increasing variance. The bias corrected estimates only need $B = \Theta(n)$ bootstrap replicates to control Monte Carlo MSE at level $1/n$.

4.4 A Numerical Example

To validate the observations made in this section, we re-visit the cholesterol data set used by Efron (2013) as a central example in developing the IJ estimate of variance. The data set (introduced by Efron and Feldman, 1991) contains records for $n = 164$ participants in a clinical study, all of whom received a proposed cholesterol-lowering drug. The data contains a measure d of the cholesterol level decrease observed for each subject, as well as a measure c of compliance (i.e. how faithful the subject was in taking the medication). Efron and Feldman originally fit d as a polynomial function of c; the degree of the polynomial was adaptively selected by minimizing Mallows' C_p criterion (1973).

We here follow Efron (2013) and study the bagged adaptive polynomial fit of d against c for predicting the cholesterol decrease of a new subject with a specific compliance level. The degree of the polynomial is selected among integers between 1 and 6 by C_p minimization. Efron (2013) gives a more detailed description of the experiment. We restrict our attention
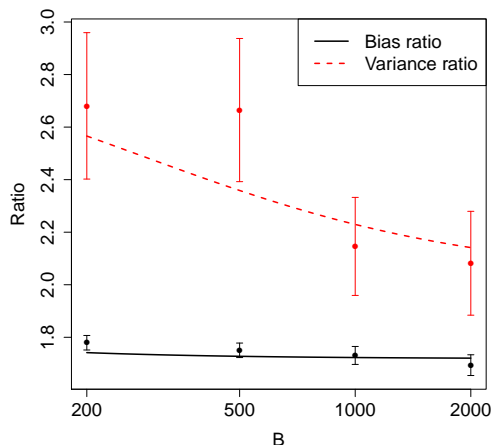
Figure 7: Predicted and actual performance ratios for the uncorrected $\widehat{V}_J^B$ and $\widehat{V}_{IJ}^B$ estimators in the cholesterol compliance example. The plot shows both $\mathrm{Var}[\widehat{V}_J^B]/\mathrm{Var}[\widehat{V}_{IJ}^B]$ and $\mathrm{Bias}[\widehat{V}_J^B]/\mathrm{Bias}[\widehat{V}_{IJ}^B]$. The observations are derived from the data presented in Figure 3; the error bars are one standard deviation in each direction. The solid lines are theoretical predictions obtained from (12) and (14).

to predicting the cholesterol decrease of a new patient with compliance level $c = -2.25$; this corresponds to the patient with the lowest observed compliance level.

In Figure 3, we compare the performance of the variance estimates for bagged predictors studied in this paper. The boxplots depict repeated realizations of the variance estimates with a finite B. We can immediately verify the qualitative insights presented in this section. Both the jackknife and IJ rules are badly biased for small B, and this bias goes away more slowly than the Monte Carlo variance. Moreover, at any given B, the jackknife estimator is noticeably less stable than the IJ estimator.

The J-U and IJ-U estimators appear to fix the bias problem without introducing instability. The J-U estimator has a slightly higher mean than the IJ-U one. As discussed in Section 5.2, this is not surprising, as the limiting $(B \rightarrow \infty)$ jackknife estimator has an upward sampling bias while the limiting IJ estimator can have a downward sampling bias. The fact that the J-U and IJ-U estimators are so close suggests that both methods work well for this problem.

The insights developed here also appear to hold quantitatively. In Figure 7, we compare the ratios of Monte Carlo bias and variance for the jackknife and IJ estimators with theoretical approximations implied by (12) and (14). The theoretical formulas appear to present a credible picture of the relative merits of the jackknife and IJ rules.

5. Sampling Distribution of Variance Estimates

In practice, the $\widehat{V}_{IJ}^B$ and $\widehat{V}_J^B$ estimates are computed with a finite number B of bootstrap replicates. In this section, however, we let B go to infinity, and study the sampling properties of the IJ and jackknife variance estimates in the absence of Monte Carlo errors. In other words, we study the impact of noise in the data itself. Recall that we write $\widehat{V}_{IJ}^\infty$ and $\widehat{V}_J^\infty$ for the limiting estimators with $B = \infty$ bootstrap replicates.

We begin by developing a simple formula for the sampling variance of $\widehat{V}_{IJ}^\infty$ itself. In the process of developing this variance formula, we obtain an ANOVA expansion of $\widehat{V}_{IJ}^\infty$ that we then use in Section 5.2 to compare the sampling biases of the jackknife and infinitesimal jackknife estimators.

5.1 Sampling Variance of the IJ Estimate of Variance

If the data Z_i are independently drawn from a distribution F, then the variance of the IJ estimator is very nearly given by

$$\mathrm{Var}_F\left[\widehat{V}_{IJ}^\infty\right] \approx n \, \mathrm{Var}_F\left[h_F^2(Z)\right], \text{ where} \tag{17}$$

$$h_F(Z) = \mathbb{E}_F\left[\hat{\theta}^\infty | Z_1 = Z\right] - \mathbb{E}_F\left[\hat{\theta}^\infty\right]. \tag{18}$$

This expression suggests a natural plug-in estimator

$$\widehat{\mathrm{Var}[\widehat{V}_{IJ}^B]} = \sum_{i=1}^n \left(C_i^{*,2} - \overline{C_i^{*,2}}\right)^2, \tag{19}$$

where $C_i^* = \mathrm{Cov}_*[N_{bi}^*, t_b^*]$ is a bootstrap estimate for $h_F(Z_i)$ and $\overline{C_i^{*,2}}$ is the mean of the $C_i^{*,2}$. The rest of the notation is as in Section 2.

The relation (17) arises from a general connection between the infinitesimal jackknife and the theory of Hájek projections. The Hájek projection of an estimator is the best approximation to that estimator that only considers first-order effects. In our case, the Hájek projection of $\hat{\theta}^\infty$ is

$$\hat{\theta}_H^\infty = \mathbb{E}_F\left[\hat{\theta}^\infty\right] + \sum_{i=1}^n h_F(Z_i), \tag{20}$$

where $h_F(Z_i)$ is as in (18). The variance of the Hájek projection is $\mathrm{Var}\left[\hat{\theta}_H^\infty\right] = n \, \mathrm{Var}\left[h_F(Z)\right]$.

The key insight behind (17) is that the IJ estimator is effectively trying to estimate the variance of the Hájek projection of $\hat{\theta}^B$, and that

$$\widehat{V}_{IJ}^\infty \approx \sum_{i=1}^n h_F^2(Z_i). \tag{21}$$

The approximation (17) then follows immediately, as the right-hand side of the above expression is a sum of independent random variables. Note that we cannot apply this right-hand side expression directly, as h depends on the unknown underlying distribution F.

The connections between Hájek projections and the infinitesimal jackknife have been understood for a long time. Jaeckel (1972) originally introduced the infinitesimal jackknife

(a) Bagged polynomial fit.
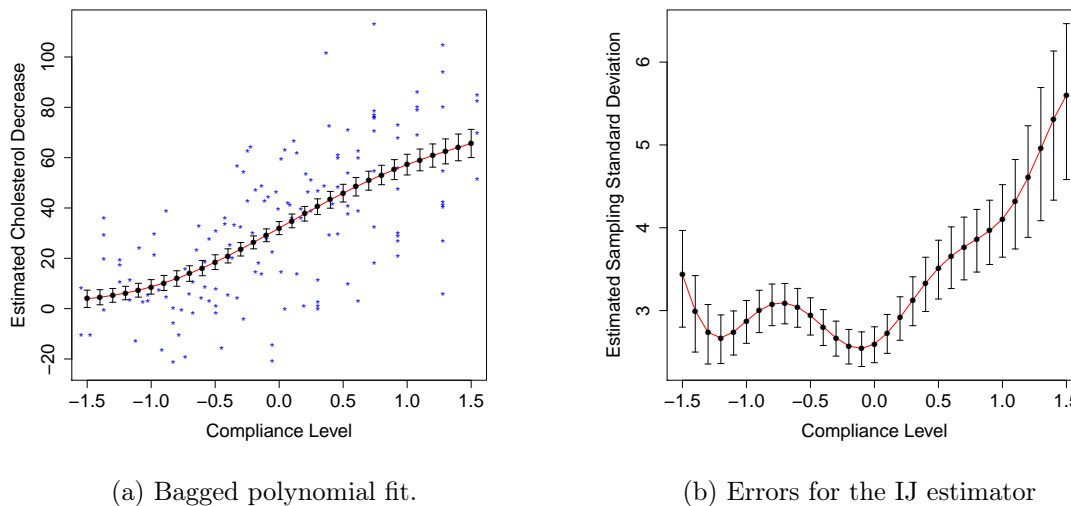
(b) Errors for the IJ estimator

Figure 8: Stability of the IJ estimate of variance on the cholesterol data. The left panel shows the bagged fit to the data, along with error bars generated by the IJ method; the stars denote the data (some data points have x-values that exceed the range of the plot). In the right panel, we use (19) to estimate error bars for the error bars in the first panel. All error bars are one standard deviation in each direction.

as a practical approximation to the first-order variance of an estimator (in our case, the right-hand side of (21)). More recently, Efron (2013) showed that $\widehat{V}_{IJ}^{\infty}$ is equal to the variance of a "bootstrap Hájek projection." In Appendix B, we build on these ideas and show that, in cases where a plug-in approximation is valid, (21) holds very nearly for bagged estimators.

We apply our variance formula to the cholesterol data set of Efron (2013), following the methodology described in Section 4.4. In Figure 8, we use the formula (19) to study the sampling variance of $\widehat{V}_{IJ}^{\infty}$ as a function of the compliance level c. The main message here is rather reassuring: as seen in Figure 8b, the coefficient of variation of $\widehat{V}_{IJ}^{\infty}$ appears to be fairly low, suggesting that the IJ variance estimates can be trusted in this example. Note that, the formula from (19) can require many bootstrap replicates to stabilize and suffers from an upward Monte Carlo bias just like $\widehat{V}_{IJ}^{B}$. We used $B = 100,000$ bootstrap replicates to generate Figure 8.

5.2 Sampling Bias of the Jackknife and IJ Estimators

We can understand the sampling biases of both the jackknife and IJ estimators in the context of the ANOVA decomposition of Efron and Stein (1981). Suppose that we have data $Z_1, ..., Z_n$ drawn independently from a distribution F, and compute our estimate $\hat{\theta}^{\infty}$ based on this data. Then, we can decompose its variance as

$$\text{Var}_F\left[\hat{\theta}^{\infty}\right] = V_1 + V_2 + ... + V_n, \tag{22}$$

where

$$V_1 = n \operatorname{Var}_F \left[\mathbb{E}_F \left[\hat{\theta}^\infty | Z_1 \right] \right]$$

is the variance due to first-order effects, V_2 is the variance due to second-order effects of the form

$$\mathbb{E}_F \left[\hat{\theta}^\infty | Z_1, \, Z_2 \right] - \mathbb{E}_F \left[\hat{\theta}^\infty | Z_1 \right] - \mathbb{E}_F \left[\hat{\theta}^\infty | Z_2 \right] + \mathbb{E}_F \left[\hat{\theta}^\infty \right],$$

and so on. Note that all the terms V_k are non-negative.

Efron and Stein (1981) showed that, under general conditions, the jackknife estimate of variance is biased upwards. In our case, their result implies that the jackknife estimator computed on $n+1$ data points has variance

$$\mathbb{E}_F \left[\widehat{V}_J^\infty \right] = V_1 + 2V_2 + 3V_3 + ... + nV_n. \tag{23}$$

Meanwhile, (21) suggests that

$$\mathbb{E}_F \left[\widehat{V}_{IJ}^\infty \right] \approx V_1. \tag{24}$$

In other words, on average, both the jackknife and IJ estimators get the first-order variance term right. The jackknife estimator then proceeds to double the second-order term, triple the third-order term etc, while the IJ estimator just drops the higher order terms.

By comparing (23) and (24), we see that the upward bias of $\widehat{V}_J^\infty$ and the downward bias of $\widehat{V}_{IJ}^\infty$ partially cancel out. In fact,

$$\mathbb{E}_F \left[\frac{\widehat{V}_J^\infty + \widehat{V}_{IJ}^\infty}{2} \right] \approx V_1 + V_2 + \frac{3}{2}V_3 + ... + \frac{n}{2}V_n, \tag{25}$$

and so the arithmetic mean of $\widehat{V}_J^\infty$ and $\widehat{V}_{IJ}^\infty$ has an upward bias that depends only on third- and higher-order effects. Thus, we might expect that in small-sample situations where $\widehat{V}_J^\infty$ and $\widehat{V}_{IJ}^\infty$ exhibit some bias, the mean of the two estimates may work better than either of them taken individually.

To test this idea, we used both the jackknife and IJ methods to estimate the variance of a bagged tree trained on a sample of size $n = 25$. (See Appendix C for details.) Since the sample size is so small, both the jackknife and IJ estimators exhibit some bias as seen in Figure 9a. However, the mean of the two estimators is nearly unbiased for the true variance of the bagged tree. (It appears that this mean has a very slight upward bias, just as we would expect from (25).)

This issue can arise in real data sets too. When training bagged forward stepwise regression on a prostate cancer data set discussed by Hastie et al. (2009), the jackknife and IJ methods give fairly different estimates of variance: the jackknife estimator converged to 0.093, while the IJ estimator stabilized at 0.067 (Figure 9b). Based on the discussion in this section, it appears that $(0.093 + 0.067)/2 = 0.08$ should be considered a more unbiased estimate of variance than either of the two numbers on their own.

In the more extensive simulations presented in Table 1, averaging $\widehat{V}_{IJ-U}^B$ and $\widehat{V}_{J-U}^B$ is in general less biased than either of the original estimators (although the "AND" experiment seems to provide an exception to this rule, suggesting that most of the bias of $\widehat{V}_{J-U}^B$ for this function is due to higher-order interactions). However, $\widehat{V}_{IJ-U}^B$ has systematically lower

(a) Bagged tree sampling bias.
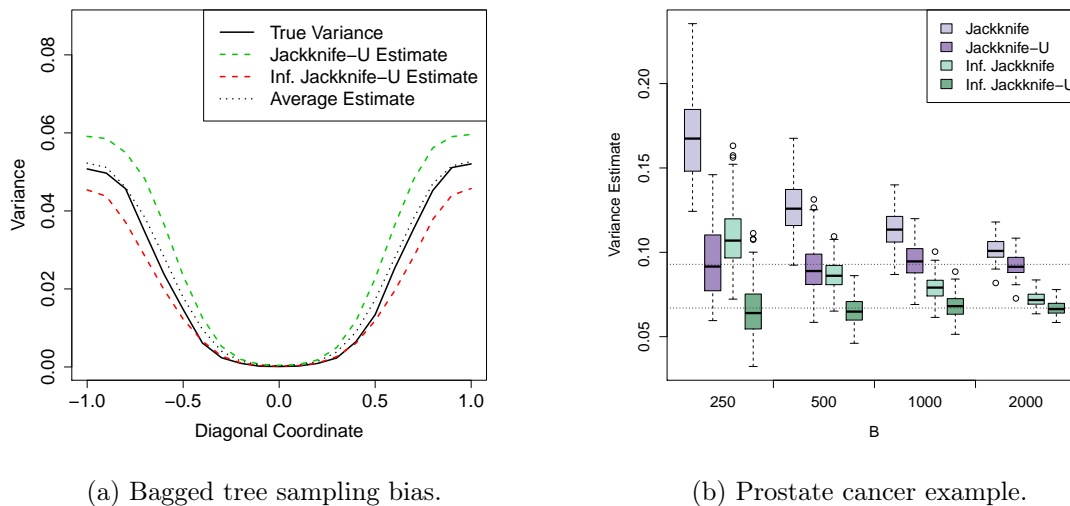
(b) Prostate cancer example.

Figure 9: Sampling bias of the jackknife and IJ rules. In the left panel, we compare the expected values of the jackknife and IJ estimators as well as their mean with the true variance of a bagged tree. In this example, the features take values in $(x_1, x_2) \in [-1, 1]^2$; we depict variance estimates along the diagonal $x_1 = x_2$. The prostate cancer plot can be interpreted in the same way as Figure 3, except that the we now indicate the weighted means of the J-U and IJ-U estimators separately.

variance, which allows it to win in terms of overall mean squared error. Thus, if unbiasedness is important, averaging $\widehat{V}_{IJ-U}^B$ and $\widehat{V}_{J-U}^B$ seems like a promising idea, but $\widehat{V}_{IJ-U}^B$ appears to be the better rule in terms of raw MSE minimization.

Finally, we emphasize that this relative bias result relies on the heuristic relationship (24). While this approximation does not seem problematic for the first-order analysis presented in Section 5.1, we may be concerned that the plug-in argument from Appendix B used to justify it may not give us correct second- and higher-order terms. Thus, although our simulation results seem promising, developing a formal and general understanding of the relative biases of $\widehat{V}_{IJ}^\infty$ and $\widehat{V}_J^\infty$ remains an open topic for follow-up research.

6. Conclusion

In this paper, we studied the jackknife-after-bootstrap and infinitesimal jackknife (IJ) methods (Efron, 1992, 2013) for estimating the variance of bagged predictors. We demonstrated that both estimators suffer from considerable Monte Carlo bias, and we proposed bias-corrected versions of the methods that appear to work well in practice. We also provided a simple formula for the sampling variance of the IJ estimator, and showed that from a sampling bias point of view the arithmetic mean of the jackknife and IJ estimators is often preferable to either of the original methods. Finally, we applied these methods in numerous

Function	n	p	B	ERR	$\widehat{V}_{IJ-U}^{B}$	$\widehat{V}_{J-U}^{B}$	$\frac{1}{2}(\widehat{V}_{IJ-U}^{B} + \widehat{V}_{J-U}^{B})$
Cosine	50	2	200	Bias	$-0.15\,(\pm 0.03)$	$0.14\,(\pm 0.02)$	$\mathbf{-0.01}\,(\pm \mathbf{0.02})$
				Var	$\mathbf{0.08}\,(\pm \mathbf{0.02})$	$0.41\,(\pm 0.13)$	$0.2\,(\pm 0.06)$
				MSE	$\mathbf{0.11}\,(\pm \mathbf{0.03})$	$0.43\,(\pm 0.13)$	$0.2\,(\pm 0.06)$
Cosine	200	2	500	Bias	$-0.05\,(\pm 0.01)$	$0.07\,(\pm 0.01)$	$\mathbf{0.01}\,(\pm \mathbf{0.01})$
				Var	$\mathbf{0.02}\,(\pm \mathbf{0})$	$0.07\,(\pm 0.01)$	$0.04\,(\pm 0.01)$
				MSE	$\mathbf{0.02}\,(\pm \mathbf{0})$	$0.07\,(\pm 0.01)$	$0.04\,(\pm 0.01)$
XOR	50	50	200	Bias	$-0.3\,(\pm 0.03)$	$0.37\,(\pm 0.04)$	$\mathbf{0.03}\,(\pm \mathbf{0.03})$
				Var	$\mathbf{0.48}\,(\pm \mathbf{0.03})$	$1.82\,(\pm 0.12)$	$0.89\,(\pm 0.05)$
				MSE	$\mathbf{0.58}\,(\pm \mathbf{0.03})$	$1.96\,(\pm 0.13)$	$0.89\,(\pm 0.05)$
XOR	200	50	500	Bias	$\mathbf{-0.08}\,(\pm \mathbf{0.02})$	$0.24\,(\pm 0.03)$	$\mathbf{0.08}\,(\pm \mathbf{0.02})$
				Var	$\mathbf{0.26}\,(\pm \mathbf{0.02})$	$0.77\,(\pm 0.04)$	$0.4\,(\pm 0.02)$
				MSE	$\mathbf{0.27}\,(\pm \mathbf{0.01})$	$0.83\,(\pm 0.04)$	$0.41\,(\pm 0.02)$
AND	50	500	200	Bias	$\mathbf{-0.23}\,(\pm \mathbf{0.04})$	$0.65\,(\pm 0.05)$	$\mathbf{0.21}\,(\pm \mathbf{0.04})$
				Var	$\mathbf{1.15}\,(\pm \mathbf{0.05})$	$4.23\,(\pm 0.18)$	$2.05\,(\pm 0.09)$
				MSE	$\mathbf{1.21}\,(\pm \mathbf{0.06})$	$4.64\,(\pm 0.21)$	$2.09\,(\pm 0.09)$
AND	200	500	500	Bias	$\mathbf{-0.04}\,(\pm \mathbf{0.04})$	$0.32\,(\pm 0.04)$	$0.14\,(\pm 0.03)$
				Var	$\mathbf{0.55}\,(\pm \mathbf{0.07})$	$1.71\,(\pm 0.22)$	$0.85\,(\pm 0.11)$
				MSE	$\mathbf{0.57}\,(\pm \mathbf{0.08})$	$1.82\,(\pm 0.24)$	$0.88\,(\pm 0.11)$
Auto	314	7	1000	Bias	$-0.11\,(\pm 0.02)$	$0.23\,(\pm 0.05)$	$\mathbf{0.06}\,(\pm \mathbf{0.03})$
				Var	$\mathbf{0.13}\,(\pm \mathbf{0.04})$	$0.49\,(\pm 0.19)$	$0.27\,(\pm 0.1)$
				MSE	$\mathbf{0.15}\,(\pm \mathbf{0.04})$	$0.58\,(\pm 0.24)$	$0.29\,(\pm 0.11)$

Table 1: Simulation study. We evaluate the mean bias, variance, and MSE of different variance estimates $\widehat{V}$ for random forests. Here, n is the number of test examples used, p is the number of features, and B is the number of trees grown; the numbers in parentheses are 95% confidence errors from sampling. The best methods for each evaluation metric are highlighted in bold. The data-generating functions are described in Appendix C.

experiments, including some random forest examples, and showed how they can be used to gain valuable insights in realistic problems.

Acknowledgments

Appendix A. The Effect of Monte Carlo Noise on the Jackknife Estimator

In this section, we derive expressions for the finite-B Monte Carlo bias and variance of the jackknife-after-bootstrap estimate of variance. Recall from (6) that

$$\widehat{V}_J^B = \frac{n-1}{n} \sum_{i=1}^{n} \hat{\Delta}_i^2, \text{ where } \hat{\Delta}_i = \frac{\sum_{\{b\,:\,N_{bi}^* = 0\}} t_b^*}{|\{N_{bi}^* = 0\}|} - \frac{\sum_b t_b^*}{B}$$

and N_{bi}^* indicates the number of times the ith observation appears in the bootstrap sample b. If $\hat{\Delta}_i$ is not defined because $N_{bi}^* = 0$ for either all or none of the $b = 1, ..., B$, then just set $\hat{\Delta}_i = 0$.

Now $\widehat{V}_J^B$ is the sum of squares of noisy quantities, and so $\widehat{V}_J^B$ will be biased upwards. Specifically,

$$\mathbb{E}_* \left[\widehat{V}_J^B \right] - \widehat{V}_J^\infty = \frac{n-1}{n} \sum_{i=1}^{n} \text{Var}_* \left[\hat{\Delta}_i \right],$$

where $\widehat{V}_J^\infty$ is the jackknife estimate computed with $B = \infty$ bootstrap replicates. For convenience, let

$$B_i = |\{b : N_{bi} = 0\}|,$$

and recall that

$$\text{Var}_* \left[\hat{\Delta}_i \right] = \mathbb{E}_* \left[\text{Var}_* \left[\hat{\Delta}_i | B_i \right] \right] + \text{Var}_* \left[\mathbb{E}_* \left[\hat{\Delta}_i | B_i \right] \right].$$

For all $B_i \neq 0$ or B, the conditional expectation is

$$\mathbb{E}_*[\hat{\Delta}_i | B_i] = \left(1 - \frac{B_i - \mathbb{E}[B_i]}{B} \right) \Delta_i, \text{ where } \Delta_i = \mathbb{E}_* \left[t_b^* | N_{bi}^* = 0 \right] - \mathbb{E}_* \left[t_b^* \right];$$

$\mathbb{E}_*[\hat{\Delta}_i | B_i] = 0$ in the degenerate cases with $B_i \in \{0, B\}$. Thus,

$$\text{Var}_* \left[\mathbb{E}_* \left[\hat{\Delta}_i | B_i \right] \right] = \mathcal{O} \left(\Delta_i^2 / B \right),$$

and so

$$\text{Var}_* \left[\hat{\Delta}_i \right] = \mathbb{E}_* \left[\text{Var}_* \left[\hat{\Delta}_i | B_i \right] \right] + \mathcal{O} \left(\Delta_i^2 / B \right).$$

Meanwhile, for $B_i \notin \{0, B\}$,

$$\text{Var}_* \left[\hat{\Delta}_i | B_i \right] = \frac{1}{B^2} \left(\left(\frac{B}{B_i} - 1 \right)^2 B_i \tilde{v}_i^{(0)} + (B - B_i) \tilde{v}_i^{(+)} \right)$$

$$= \frac{1}{B} \left(\frac{(B - B_i)^2}{B B_i} \tilde{v}_i^{(0)} + \frac{B - B_i}{B} \tilde{v}_i^{(+)} \right)$$

where

$$\tilde{v}_i^{(0)} = \text{Var}_* \left[t_b^* | N_{bi}^* = 0 \right] \text{ and } \tilde{v}_i^{(+)} = \text{Var}_* \left[t_b^* | N_{bi}^* \neq 0 \right].$$

Thus,

$$\text{Var}_* \left[\hat{\Delta}_i \right] = \frac{1}{B} \left(\mathbb{E}_* \left[\frac{(B - B_i)^2}{BB_i} 1_i \right] \tilde{v}_i^{(0)} + \mathbb{E}_* \left[\frac{B - B_i}{B} 1_i \right] \tilde{v}_i^{(+)} \right) + \mathcal{O} \left(\Delta_i^2 / B \right),$$

where $1_i = 1(\{B_i \notin \{0, B\}\})$.

As n and B get large, B_i converges in law to a Gaussian random variable

$$\frac{B_i - Be^{-1}}{\sqrt{B}} \Rightarrow \left(0, \, e^{-1}(1 - e^{-1}) \right)$$

and the above expressions are uniformly integrable. We can verify that

$$\mathbb{E}_* \left[\frac{(B - B_i)^2}{BB_i} 1_i \right] = e - 2 + e^{-1} + \mathcal{O} \left(\frac{1}{B} \right),$$

and

$$\mathbb{E}_* \left[\frac{B - B_i}{B} 1_i \right] = \frac{e - 1}{e} + \mathcal{O} \left(\left(1 - e^{-1} \right)^B \right).$$

Finally, this lets us conclude that

$$\mathbb{E}_* \left[\widehat{V}_J^B \right] - \widehat{V}_J^\infty = \frac{1}{B} \frac{n - 1}{n} \sum_{i=1}^n \left(\left(\frac{(e - 1)^2}{e} \right) \tilde{v}_i^{(0)} + \left(\frac{e - 1}{e} \right) \tilde{v}_i^{(+)} \right) + \mathcal{O} \left(\frac{1}{B} + \frac{n}{B^2} \right),$$

where the error term depends on $\tilde{v}_i^{(0)}$, $\tilde{v}_i^{(+)}$, and $\widehat{V}_J^\infty = (n - 1)/n \sum_{i=1}^n \Delta_i^2$.

We now address Monte Carlo variance. By the central limit theorem, $\hat{\Delta}_i$ converges to a Gaussian random variable as B gets large. Thus, the asymptotic Monte Carlo variance of $\hat{\Delta}_i^2$ is approximately $2 \text{Var}_*[\hat{\Delta}_i]^2 + 4\mathbb{E}_*[\hat{\Delta}_i]^2 \text{Var}_*[\hat{\Delta}_i]$, and so

$$\text{Var}_* \left[\widehat{V}_J^B \right] \approx 2 \left(\frac{1}{B} \frac{n - 1}{n} \right)^2 \sum_{i=1}^n \left(\left(\frac{(e - 1)^2}{e} \right) \tilde{v}_i^{(0)} + \left(\frac{e - 1}{e} \right) \tilde{v}_i^{(+)} \right)^2$$

$$+ 4 \frac{1}{B} \frac{n - 1}{n} \sum_{i=1}^n \Delta_i^2 \left(\left(\frac{(e - 1)^2}{e} \right) \tilde{v}_i^{(0)} + \left(\frac{e - 1}{e} \right) \tilde{v}_i^{(+)} \right).$$

In practice, the terms $\tilde{v}_i^{(0)}$ and $\tilde{v}_i^{(+)}$ can be well approximated by $\hat{v} = \text{Var}_*[t_b^*]$, namely the bootstrap estimate of variance for the base learner. (Note that $\tilde{v}_i^{(0)}$, $\tilde{v}_i^{(+)}$, and $\hat{v}$ can always be inspected on a random forest, so this assumption can be checked in applications.) This lets us considerably simplify our expressions for Monte Carlo bias and variance:

$$\mathbb{E}_* \left[\widehat{V}_J^B \right] - \widehat{V}_J^\infty \approx \frac{n}{B} (e - 1)\hat{v}, \text{ and}$$

$$\text{Var}_* \left[\widehat{V}_J^B \right] \approx 2 \frac{n}{B^2} (e - 1)^2 \hat{v}^2 + 4 \frac{1}{B} (e - 1) \widehat{V}_J^\infty \hat{v}.$$

Appendix B. The IJ estimator and Hájek projections

Up to (27), the derivation below is an alternate presentation of the argument made by Efron (2013) in the proof of his Theorem 1. To establish a connection between the IJ estimate of variance for bagged estimators and the theory of Hájek projections, it is useful to consider $\hat{\theta}^B$ as a functional over distributions. Let G be a probability distribution, and let T be a functional over distributions with the following property:

$$T(G) = \mathbb{E}_G[\tau(Y_1, ..., Y_n)] \text{ for some function } \tau, \tag{26}$$

where the $Y_1, ..., Y_n$ are drawn independently from G. We call functionals T satisfying (26) *averaging*. Clearly, $\hat{\theta}^B$ can be expressed as an averaging functional applied to the empirical distribution $\widehat{F}$ defined by the observations $Z_1, ..., Z_n$.

Suppose that we have an averaging functional T, a sample $Z_1, ..., Z_n$ forming an empirical distribution $\widehat{F}$, and want to study the variance of $T(\widehat{F})$. The infinitesimal jackknife estimate for the variance of $\widehat{T}$ is given by

$$\widehat{V} = \sum_{i=1}^{n} \left(\frac{1}{n} \frac{\partial}{\partial \varepsilon} T\left(\widehat{F}_i(\varepsilon)\right) \right)^2,$$

where $\widehat{F}_i(\varepsilon)$ is the discrete distribution that places weight $1/n + (n-1)/n \cdot \varepsilon$ at Z_i and weight $1/n - \varepsilon/n$ at all the other Z_j.

We can transform samples from $\widehat{F}$ into samples from $\widehat{F}_i(\varepsilon)$ by the following method. Let $Z_1^*, ..., Z_n^*$ be a sample from $\widehat{F}$. Go through the whole sample and, independently for each j, take Z_j^* and with probability ε replace it with Z_i. The sample can now be considered a sample from $\widehat{F}_i(\varepsilon)$.

When $\varepsilon \to 0$, the probability of replacing two of the Z_i^* with this procedure becomes negligible, and we can equivalently transform our sample into a sample from $\widehat{F}_i(\varepsilon)$ by transforming a single random element from $\{Z_j^*\}$ into Z_i with probability $n\varepsilon$. Without loss of generality this element is the first one, and so we conclude that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\mathbb{E}_{\widehat{F}_i(\varepsilon)} [\tau(Z_1^*, ..., Z_n^*)] - \mathbb{E}_{\widehat{F}} [\tau(Z_1^*, ..., Z_n^*)] \right)$$
$$= n \left(\mathbb{E}_{\widehat{F}} \left[\tau(Z_1^*, ..., Z_n^*) \big| Z_1^* = Z_i \right] - \mathbb{E}_{\widehat{F}} [\tau(Z_1^*, ..., Z_n^*)] \right),$$

where τ defines T through (26). Thus,

$$\frac{1}{n} \frac{\partial}{\partial \varepsilon} T(\widehat{F}_i(\varepsilon)) = \mathbb{E}_{\widehat{F}} \left[T \big| Z_1^* = Z_i \right] - \mathbb{E}_{\widehat{F}} [T],$$

and so

$$\widehat{V} = \sum_{i=1}^{n} \left(\mathbb{E}_{\widehat{F}} \left[T \big| Z_1^* = Z_i \right] - \mathbb{E}_{\widehat{F}} [T] \right)^2 \tag{27}$$

$$\approx \sum_{i=1}^{n} \left(\mathbb{E}_F \left[T \big| Z_1^* = Z_i \right] - \mathbb{E}_F [T] \right)^2, \tag{28}$$
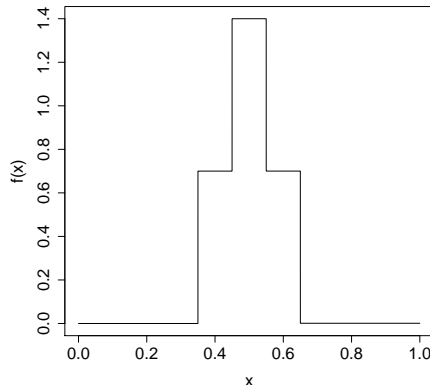
Figure 10: Underlying model for the bagged tree example from Figure 2.

where on the last line we only replaced the empirical approximation $\widehat{F}$ with its true value F. In the case of bagging, this last expression is equivalent to (21).

A crucial step in the above argument is the plug-in approximation (28). If T is just a sum, then the error of (28) is within $\mathcal{O}(1/n)$; presumably, similar statements hold whenever T is sufficiently well-behaved. That being said, it is possible to construct counter-examples where (28) fails; a simple such example is when T counts the number of times Z_1^* is matched in the rest of the training data. Establishing general conditions under which (28) holds is an interesting topic for further research.

Appendix C. Description of Experiments

This section provides a more detailed description of the experiments reported in this paper.

C.1 Auto MPG Example (Figure 1)

The Auto MPG data set, available from the UCI Machine Learning Repository (Bache and Lichman, 2013), is a regression task with 7 features. After discarding examples with missing entries, the data set had 392 rows, which we divided into a test set of size 78 and a train set of size 314. We estimated the variance of the random forest predictions using the $(\widehat{V}_{J-U}^B + \widehat{V}_{IJ-U}^B)/2$ estimator advocated in Section 5.2, with $B = 10,000$ bootstrap replicates.

C.2 Bagged Tree Simulation (Figure 2)

The data for this simulation was drawn from a model $y_i = f(x_i) + \varepsilon_i$, where $x_i \sim U([0, 1])$, $\varepsilon_i \sim \mathcal{N}(0, 1/2^2)$, and $f(x)$ is the step function shown in Figure 10. We modeled the data using 5-leaf regression trees generated using the R package `tree` (Venables and Ripley, 2002); for bagging, we used $B = 10,000$ bootstrap replicates. The reported data is compiled over $1,000$ simulation runs with $n = 500$ data points each.

C.3 Cholesterol Example (Figures 3, 7, and 8)

For the cholesterol data set (Efron and Feldman, 1991), we closely follow the methodology of Efron (2013); see his paper for details. The data set has $n = 164$ subjects and only one predictor.

C.4 E-mail Spam Example (Figures 4 and 5)

The e-mail spam data set (spambase, Bache and Lichman, 2013) is a classification problem with $n = 4,601$ e-mails and $p = 57$ features; the goal is to discern spam from non-spam. We divided the data into train and test sets of size 3,065 and 1,536 respectively. Each of the random forests described in Section 3.1 was fit on the train set using the R package `randomForest` (Liaw and Wiener, 2002) with $B = 40,000$ bootstrap replicates.

C.5 California Housing Example (Figure 6)

The California housing data set (described in Hastie et al., 2009, and available from StatLib) contains aggregated data from $n = 20,460$ neighborhoods. There are $p = 8$ features; the response is the median house value. We fit random forests on this data using the R package `randomForest` (Liaw and Wiener, 2002) with $B = 1,000$ bootstrap replicates.

C.6 Bagged Tree Simulation #2 (Figure 9a)

We drew $n = 25$ points from a model where the x_i are uniformly distributed over a square, i.e., $x_i \sim U([-1, 1]^2)$; the y_i are deterministically given by $y_i = 1(\{||x_i||_2 \geq 1\})$. We fit this data using the R package `tree` (Venables and Ripley, 2002). The bagged predictors were generated using $B = 1,000$ bootstrap replicates. The reported results are based on 2,000 simulation runs.

C.7 Prostate Cancer Example (Figure 9b)

The prostate cancer data (published by Stamey et al., 1989) is described in Section 1 of Hastie et al. (2009). We used forward stepwise regression as implemented by the R function `step` as our base learner. This data set has $n = 97$ subjects and 8 available predictor variables. In figure 9b, we display standard errors for the predicted response of a patient whose features match those of patient #41 in the data set.

C.8 Simulations for Table 1

The data generation functions used in Table 1 are defined as follows. The X_i for $i = 1, ..., p$ are all generated as independent $U([0, 1])$ random variables, and $\varepsilon \sim \mathcal{N}(0, 1)$.

- Cosine: $Y = 3 \cdot \cos(\pi \cdot (X_1 + X_2))$, with $p = 2$.

- XOR: Treating `XOR` as a function with a 0/1 return-value,

$$Y = 5 \cdot [\texttt{XOR}(X_1 > 0.6, X_2 > 0.6) + \texttt{XOR}(X_3 > 0.6, X_4 > 0.6)] + \varepsilon$$

  and $p = 50$.

- AND: With analogous notation,

$$Y = 10 \cdot \mathtt{AND}\,(X_1 > 0.3,\ X_2 > 0.3,\ X_3 > 0.3,\ X_4 > 0.3) + \varepsilon$$

  and $p = 500$.

- Auto: This example is based on a parametric bootstrap built on the same data set as used in Figure 1. We first fit a random forest to the training set, and evaluated the MSE $\hat{\sigma}^2$ on the test set. We then generated new training sets by replacing the labels Y_i from the original training set with $\widehat{Y}_i + \hat{\sigma}\varepsilon$, where $\widehat{Y}_i$ is the original random forest prediction at the ith training example and ε is fresh residual noise.

During the simulation, we first generated a random test set of size 50 (except for the auto example, where we just used the original test set of size 78). Then, while keeping the test set fixed, we generated 100 training sets and produced variance estimates $\widehat{V}$ at each test point. Table 1 reports average performance over the test set.

References

Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(4):1063–1095, 2012.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Andreas Buja and Werner Stuetzle. Observations on bagging. *Statistica Sinica*, 16(2):323, 2006.

Song Xi Chen and Peter Hall. Effects of bagging and bias correction on estimators defined by estimating equations. *Statistica Sinica*, 13(1):97–110, 2003.

Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40 (2):139–157, 2000.

Jiangtao Duan. *Bootstrap-Based Variance Estimators for a Bagging Predictor.* PhD thesis, North Carolina State University, 2011.

Bradley Efron. Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 83–127, 1992.

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, (just-accepted), 2013.

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

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

Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

Jerome H Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683, 2007.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* New York: Springer, 2009.

Louis A Jaeckel. *The Infinitesimal Jackknife.* 1972.

Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Colin L Mallows. Some comments on Cp. *Technometrics*, 15(4):661–675, 1973.

Nicolai Meinshausen. Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999, 2006.

Joseph Sexton and Petter Laake. Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, 53(3):801–811, 2009.

Marina Skurichina and Robert PW Duin. Bagging for linear classifiers. *Pattern Recognition*, 31(7):909–930, 1998.

Thomas A Stamey, John N Kabalin, John E McNeal, Iain M Johnstone, Fuad Freiha, EA Redwine, and N Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. *The Journal of Urology*, 141(5):1076–1083, 1989.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

William N Venables and Brian D Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.