Statistica Sinica Preprint No: SS-2022-0131	
Title	Bias, Consistency, and Alternative Perspectives of the Infinitesimal Jackknife
Manuscript ID	SS-2022-0131
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0131
Complete List of Authors	Wei Peng, Lucas Mentch and Len Stefanski
Corresponding Authors	Lucas Mentch
E-mails	lkm31@pitt.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

Statistica Sinica

Bias, Consistency, and Alternative Perspectives of the Infinitesimal Jackknife

Wei Peng*, Lucas Mentch* and Len Stefanski

University of Pittsburgh and North Carolina State University*

Though introduced nearly 50 years ago, the infinitesimal jackknife (IJ) remains a popular modern tool for quantifying predictive uncertainty in complex estimation settings. In particular, when supervised learning ensembles are constructed via bootstrap samples, recent work demonstrated that the IJ estimate of variance is particularly convenient and useful. However, despite the algebraic simplicity of its final form, its derivation is rather complex. As a result, studies clarifying the intuition behind the estimator or rigorously investigating its properties have been severely lacking. This work aims to take a step forward on both fronts. We demonstrate that surprisingly, the exact form of the IJ estimator can be obtained via a straightforward linear regression of the individual bootstrap estimates on their respective weights or via the classical jackknife. The latter realization allows us to formally investigate the bias of the IJ variance estimator and better characterize the settings in which its use is appropriate. Finally, we extend these results to the case of U-statistics where base models are constructed via subsampling rather than bootstrapping and provide a consistent estimate of the resulting variance.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

1. Introduction

Given a sample $X_1, ..., X_n \sim \mathrm{P}$, a parameter of interest θ, and an estimator $\hat{\theta} = s(X_1, ..., X_n)$, it is often of interest to estimate $\mathrm{Var}(\hat{\theta})$. Given data $\boldsymbol{x} = (x_1, ..., x_n)$ and an estimate $\hat{\theta} = s(\boldsymbol{x})$, to provide a bootstrap estimate of variance, we draw B (re)samples of size n with replacement to form bootstrap samples $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_B^*$ from which we calculate bootstrap estimates $\hat{\theta}_1, ..., \hat{\theta}_B$. The nonparametric bootstrap variance estimate of $\hat{\theta}$ is then taken as the empirical variance of $\hat{\theta}_1, ..., \hat{\theta}_B$ (Efron, 1979, 2014). Within this context, given the necessity of calculating $\hat{\theta}_1, ..., \hat{\theta}_B$, it is natural to consider the estimator $\tilde{\theta}_B = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b$ as a "bootstrap smoothed" or "bagged" alternative of $\hat{\theta}$ (Efron and Tibshirani, 1994; Breiman, 1996).

The standard bootstrap approach to assess the variability of $\tilde{\theta}_B$ is computationally burdensome, requiring bootstrap replicates of not only the original data, but of the bootstrap samples as well. This double bootstrap (Beran, 1988), is especially costly whenever the original statistic s is computationally costly. A variety of approaches have been suggested to reduce the computational burden of these sorts of problems. Some of these (White, 2000; Davidson and MacKinnon, 2000, 2002, 2007; Giacomini et al., 2013; Chang and Hall, 2015) employ what is now referred to as the fast double bootstrap whereby only a single second-level bootstrap sample is collected.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

Sexton and Laake (2009) propose an alternative nonparametric means by which $\mathrm{Var}(\tilde{\theta}_B)$ may be estimated, suggesting also that the number of second-level bootstrap replicates B' may be small. In lieu of full bootstrap samples, subsampling, or m-out-of-n bootstrap sampling, was proposed by Politis and Romano (1994) and Bickel et al. (1997). More recently, Sengupta et al. (2016) proposed a combination of these approaches, first subsampling and then employing a single second-level resample. Similarly, Kleiner et al. (2014) proposed the bag of little bootstraps which involves splitting the original dataset into a number of subsamples and then taking bootstrap samples on each subset.

Though the above approaches can reduce the computational complexity, each nonetheless involves further resampling. Recently, Efron (2014) alleviated this issue by developing an algebraically compact, closed-form estimator for the variance of a bagged estimate. Instead of additional resampling, Efron's proposal required only additional bookkeeping to recall which samples in the original data appeared how many times in each bootstrap sample. This development has proved especially helpful for estimating the variance in predictions generated via supervised learning ensembles like random forests (Wager et al., 2014; Wager and Athey, 2018; Zhou et al., 2021).

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

Though its form is algebraically simple, Efron's variance estimator can appear somewhat mysterious. Its development comes from an application of the original theory for the infinitesimal jackknife (IJ) involving functional derivatives. Likely as a result, studies investigating its statistical properties as well as the contexts in which the estimator would be appropriate are lacking. Efron, for example, notes that the appropriateness of his nonparametric delta method (IJ) approach follows from the fact that the bagged estimates represent a more smooth function of the data. Thus, while clearly a significant result, these estimates would not necessarily apply in other resampling schemes without smoothness guarantees.

Here we take a step forward both in better understanding the intuition behind this important estimator as well as in understanding its statistical properties. In addition to the IJ derivation utilized by Efron, we consider two alternative approaches that are more straightforward and easily motivated. The first exploits the important fact that conditional on the observed data, the bagged estimate depends only on the resampling weights. We consider a linear approximation to this function of bootstrap weights (i.e. standard linear regression) and demonstrate that this approach exactly reproduces the infinitesimal jackknife results given in Efron (2014) whenever all bootstrap samples are employed. As an additional benefit, this

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

setup motivates a more general procedure for estimating the variance of any resampled estimate.

In addition to the linear regression and IJ approaches, we also consider a classical jackknife motivation and again demonstrate its equivalence in the full bootstrap context. This alternative representation of the estimator allows us to explore its asymptotic properties and in particular, the bias. While the variance estimators motivated by the jackknife, infinitesimal jackknife, and linear regression approaches are shown to be identical when all bootstrap samples are used, they differ in practical settings when only a randomly selected subsample are employed, suggesting different bias corrections that might be imposed.

Finally, we derive the form of the IJ variance estimate in the U-statistic regime and discover that the variance estimators commonly employed in practice are actually a sort of "pseudo" infinitesimal jackknife in that they are a linear approximation but differ from the form derived directly from the definition. We then investigate the properties of the pseudo infinitesimal jackknife and provide a consistent estimate of the variance of (generalized) U-statistics.

The remainder this paper proceeds as follows. In Section 2, we provide the background and historical motivation for the IJ method. In Section 3,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

in addition to Efron's recent derivation in Efron (2014), we provide three alternative approaches for obtaining variance estimates for any bootstrap smoothed statistic, examine their bias, and demonstrate their equivalence when all bootstrap samples are employed. Finally, in Section 4, we derive the IJ estimator for the variance of (generalized) U-statistics and discuss its consistency. These results provide formal guarantees for the validity of confidence intervals constructed on such estimators in settings where not all subsamples are employed, and therefore apply directly to the kinds of modern supervised learning ensembles like random forests commonly employed in practice.

2. Background of the Infinitesimal Jackknife (IJ)

Let $\mathcal{D}_n$ denote a sample of observed values from real-valued random variables $X_1, \ldots, X_n$ that are i.i.d. from a distribution P. In practice, we are often interested in estimating statistical functionals – functions of the underlying distribution P, often estimated via the empirical distribution $\mathbb{P}_n$. Assume that s is permutation symmetric in these n arguments and denote this statistic as $s(X_1, \ldots, X_n) = f(\mathbb{P}_n)$. These "functions of functions" were first introduced by Volterra (1887) and today are a familiar topic of advanced analysis. Any statistic that treats the samples equivalently can also be viewed as a function of $\mathbb{P}_n$, albeit without always necessarily having an ex-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

plicit form of f. We can further extend the domain of f to any non-negative functions on $X_1, \ldots, X_n$ by defining $f(\mathbb{P}_n) = f(c \cdot \mathbb{P}_n)$, for any $c > 0$.

A common task, especially in today's big data era, is to find an appropriate and feasible means of estimating the variance of $f(\mathbb{P}_n)$. Historically, there have been three primary methods: the infinitesimal jackknife (Miller, 1974), influence curves (Hampel, 1974; Huber et al., 1972), and the delta method (Efron, 1982). Though each method was motivated differently, Efron (1981) pointed out that the three methods are identical. We thus refer to the common estimator as IJ, which is defined as

$$\mathrm{IJ} = \frac{1}{n^2} \sum_i \mathrm{D}_i^2, \tag{2.1}$$

where

$$\mathrm{D}_i = \lim_{\epsilon \to 0} \frac{f((1-\epsilon)\mathbb{P}_n + \epsilon \delta_{X_i}) - f(\mathbb{P}_n)}{\epsilon} \tag{2.2}$$

and δ_x is the Dirac delta function.

We now briefly review the original derivation of the IJ, following closely to the original constructions given by Mises (1947) and Jaeckel (1972). Let $\mathcal{P}$ be the set of all linear combinations of P and an arbitrary finite number of the δ_x measures. Let $\mathcal{P}^+$ be the set of positive measures in $\mathcal{P}$, not including the zero measures and assume f is defined for the probability measures in $\mathcal{P}^+$. As above, extend f to all of $\mathcal{P}^+$ by letting $f(c \cdot \mathrm{Q}) = f(\mathrm{Q})$ for all

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

$c > 0$. Note that $\mathcal{P}^+$ is convex and includes $\mathbb{P}_n$. We say f is differentiable at G in $\mathcal{P}^+$, if there exists a function $f'(\mathrm{G}, x)$, defined at all x in $\mathbb{R}$, with the following property:

Definition 1 (Jaeckel (1972))**.** Let H be any member of $\mathcal{P}$ such that $\mathrm{G}+t\mathrm{H}$ is in $\mathcal{P}^+$ for all t in some interval $0 \le t \le t_{\mathrm{H}}$, $t_{\mathrm{H}} > 0$, so that $f(\mathrm{G} + t\mathrm{H})$ is defined for t in this interval. Then for any such H, $f'(\mathrm{G}, x)$ satisfies

$$\frac{df(\mathrm{G} + t\mathrm{H})}{dt}|_{t=0} := \lim_{t\to 0} \frac{f(\mathrm{G} + t\mathrm{H}) - f(\mathrm{G})}{t} = \int f'(\mathrm{G}, x)\, d\mathrm{H}(x). \tag{2.3}$$

If $\mathrm{H} = \mathrm{G}$, we see that $\int f'(\mathrm{G}, x)\, d\mathrm{G}(x) = 0$ since $f(c\mathrm{G}) = f(\mathrm{G})$. On the other hand, if $\mathrm{H} = \delta_x - \mathrm{G}$, we find

$$\lim_{t\to 0} \frac{f((1-t)\mathrm{G} + t\delta_x) - f(\mathrm{G})}{t} = \int f'(\mathrm{G}, x)\, d(\delta_x - \mathrm{G})(x) = f'(\mathrm{G}, x). \tag{2.4}$$

Indeed, Hampel (1974) defined $f'(\mathrm{G}, x)$ by (2.4) and called it the "influence curve", since it reflects the influence of f by adding a small mass on G at x. Additionally, the derivative of $f(\mathrm{G} + t\mathrm{H})$ at arbitrary t_0 with $0 < t_0 < t_{\mathrm{H}}$ is given by

$$\frac{df(\mathrm{G} + t\mathrm{H})}{dt}|_{t=t_0} = \frac{df(\mathrm{G} + t_0\mathrm{H} + u\mathrm{H})}{du}|_{u=0} = \int f'(\mathrm{G} + t_0\mathrm{H}, x)\, d\mathrm{H}(x). \tag{2.5}$$

Now assume that f is differentiable (in the sense defined above) at all G in some convex neighbor of P in $\mathcal{P}^+$ such that $\mathbb{P}_n$ lies in the neighborhood

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

with probability approaching one. We now describe the motivation for addressing when the IJ estimator might provide a sensible estimate of the variance of $f(\mathbb{P}_n)$. Parameterizing the segment from P to $\mathbb{P}_n$ by $\mathrm{P}(t) = \mathrm{P} + t(\mathbb{P}_n - \mathrm{P})$ for $0 \le t \le 1$, we have that if $\mathbb{P}_n$ lies in the neighborhood of P, then

$$\begin{aligned} f(\mathbb{P}_n) - f(\mathrm{P}) &= f(\mathrm{P}(1)) - f(\mathrm{P}(0)) \\ &= \frac{df(\mathrm{P}(c))}{dt}|_{t=c} \\ &= \int f'(\mathrm{P}(c))\, d(\mathbb{P}_n - \mathrm{P}) \quad \text{(by (2.5))} \end{aligned} \tag{2.6}$$

for some c in $[0, 1]$, where the second equality above is due to the mean value theorem. Now, for large n, $\mathbb{P}_n$ is near P and one would expect that $f'(\mathrm{P}(c))$ is close to $f'(\mathrm{P})$ in such a way that

$$\begin{aligned} f(\mathbb{P}_n) - f(\mathrm{P}) &= \int f'(\mathrm{P}, x)\, d(\mathbb{P}_n - \mathrm{P})(x) + o_p\left(1/\sqrt{n}\right) \\ &= \frac{1}{n}\sum_i f'(\mathrm{P}, X_i) + o_p\left(1/\sqrt{n}\right) \end{aligned} \tag{2.7}$$

where the last equality is due to the fact that $\int f'(\mathrm{P}, x)\, d\mathrm{P}(x) = 0$. We hope that $f(\mathbb{P}_n) - f(\mathrm{P})$ is dominated by the first term, so that the remainder is $o_p\left(1/\sqrt{n}\right)$. Since this first term is a sum of i.i.d. random variables, $\sqrt{n}(f(\mathbb{P}_n) - f(\mathrm{P}))$ is asymptotic normal with mean 0 and variance $\mathrm{V} = \int [f'(\mathrm{P}, x)]^2\, d\mathrm{P}(x)$. Since P is unknown and $f'(\mathrm{P}, x)$ depends on both f and P, we generally will not know $f'(\mathrm{P}, x)$ in advance, and so we could

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

estimate V by

$$\mathrm{V} = \int [f'(\mathrm{P}, x)]^2 \, d\mathrm{P}(x) \approx \int f'^2(\mathbb{P}_n, x) \, d\mathbb{P}_n(x) = \frac{1}{n} \sum_i [f'(\mathbb{P}_n, X_i)]^2. \quad (2.8)$$

Then $\mathrm{Var}(f(\mathbb{P}_n))$ can be estimated by $n^{-2} \sum [f'(\mathbb{P}_n, X_i)]^2$, which corresponds to the definition of the IJ estimator given in (2.1) with $\mathrm{D}_i = f(\mathbb{P}_n, X_i)$. In summary, to obtain the final estimate of the variance of $f(\mathbb{P}_n)$, we need to introduce two steps of approximation. First, in (2.7), we approximate $f(\mathbb{P}_n)$ with a linear statistic at $\mathbb{P}_n$. Then, in (2.8), we approximate P with $\mathbb{P}_n$ and $f'(\mathrm{P})$ with $f'(\mathbb{P}_n)$. Thus, in evaluating the quality of the IJ estimator, we must determine (i) whether f is close to a linear statistic and (ii) whether $f'(\mathbb{P}_n)$ is close to $f'(\mathrm{P})$. The following sections provide in-depth investigations into these issues for popular types of statistics formed by resampling.

3. Infinitesimal Jackknife for Bootstrap ($\mathrm{IJ_B}$)

We now focus on the bootstrap setting where the infinitesimal jackknife has seen the most success as a method for estimating variance. Suppose that $s(X_1, ..., X_n)$ is statistic, not necessarily a function of $\mathbb{P}_n$, and we take all possible bootstrap samples $(X_1^*, ..., X_n^*)$, plug each into s to obtain a corresponding bootstrap replication s^*, and then take the average. We call the new statistic the bootstrap smoothed (bagged) alternative of s

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

and denote it as $\mathbb{E}_*[s^*]$, where $\mathbb{E}_*[\cdot]$ denotes the expectation taken over the bootstrap sampling procedure conditional on the data. Note that $\mathbb{E}_*[s^*]$ is now a function of $\mathbb{P}_n$. The dependence of f on $\mathbb{P}_n$ can be expressed explicitly as

$$f(\mathbb{P}_n) = \int \cdots \int s \, d\mathbb{P}_n \times \cdots \times d\mathbb{P}_n = \int s \, [d(\mathbb{P}_n)]^n. \tag{3.1}$$

According to Eq. (2.4), we have

$$\begin{aligned} f'(\mathbb{P}_n, x) &= n \left(\int s(x, x_2, \ldots, x_n) \, d\mathbb{P}_n(x_2) \times \cdots \times d\mathbb{P}_n(x_n) - f(\mathbb{P}_n) \right) \\ &= n \left(\int s(x, \cdots) \, [d\mathbb{P}_n]^{n-1} - \int s \, [d(\mathbb{P}_n)]^n \right) \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} f'(\mathrm{P}, x) &= n \left(\int s(x, x_2, \ldots, x_n) \, d\mathrm{P}(x_2) \times \cdots \times d\mathrm{P}(x_n) - f(\mathrm{P}) \right) \\ &= n \left(\int s(x, \ldots) \, [d\mathrm{P}]^{n-1} - \int s \, [d\mathrm{P}]^n \right). \end{aligned} \tag{3.3}$$

By the Glivenko-Cantelli theorem, $\sup_x |\mathbb{P}_n(x) - \mathrm{P}(x)| \xrightarrow{a.s.} 0$. Though the distance between $\mathbb{P}_n$ and P is small as n increases, $f'(\mathbb{P}_n, x)$ does not necessarily converge to $f'(\mathrm{P}, x)$ since $\mathbb{P}_n$ appears as the differential $[d\mathbb{P}_n]^n$ a total of n times rather than a constant number of times.

3.1 Three Approaches for Variance Estimation

We turn now to the question of how to derive an estimator for $\mathrm{Var}(\mathbb{E}_*[s^*])$. In the following subsections, we lay out three different approaches, including

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

the Efron's infinitesimal jackknife formulation and ultimately demonstrate that all three are equivalent when all bootstrap samples are utilized.

Method 1: The Infinitesimal Jackknife Approach:

Since $\mathbb{E}_*[s^*]$ can be viewed as a function of $\mathbb{P}_n$, estimating $\mathrm{Var}(\mathbb{E}_*[s^*]) = \mathrm{Var}(f(\mathbb{P}_n))$ is a standard problem for the infinitesimal jackknife method and we denote the estimator $\mathrm{IJ_B}$. From the definition of $\mathbb{E}_*[s^*]$, we have

$$\begin{aligned}
&f((1-\epsilon)\mathbb{P}_n + \epsilon\delta_{X_i}) \\
&= n^{-n} \sum \frac{s(X_1^*, \ldots, X_n^*) n!}{(w_1^*!)(w_2^*!)\ldots(w_n^*!)} \left[(1-\epsilon)^{\sum_{k\neq i} w_k^*} (1+(n-1)\epsilon)^{w_i^*}\right] \\
&= n^{-n} \sum \frac{s(X_1^*, \ldots, X_n^*) n!}{(w_1^*!)(w_2^*!)\ldots(w_n^*!)} \left[1 + n\epsilon(w_i^* - 1)\right] + o(\epsilon^2) \\
&= f(\mathbb{P}_n) + \epsilon n \mathrm{Cov}_*(s^*, w_i^*) + o(\epsilon^2),
\end{aligned}$$

where $w_i^* = \#\{j : X_j^* = X_i\}$. By Eq. (2.2), $\mathrm{D}_i = n\mathrm{Cov}_*(s^*, w_i^*) = n\mathrm{Cov}_i$ and thus

$$\mathrm{IJ_B} = \sum_i \mathrm{Cov}_i^2 = \sum_i \mathrm{Cov}_*^2(s^*, w_i^*). \tag{3.4}$$

The estimator in (3.4) was first derived by Efron (2014) as a straightforward application of (2.1) and (2.2). Nonetheless, Cov_i^2 may seem a bit abstract and this work did not discuss how well the estimator might be expected to perform in various settings. However, if we go back to the original idea of infinitesimal jackknife given in Section 2, we can see that Cov_i^2 is identical to $f'(\mathbb{P}_n, X_i)$ in (3.2), which simply comes from the approxima-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

tions of $\frac{1}{n}\sum_i f'(\mathrm{P}, X_i)$ to $f(\mathbb{P}_n)$ and $f'(\mathbb{P}_n, x)$ to $f'(\mathrm{P}, x)$, and thus, how well the estimator performs in practice depends entirely on the accuracy of the approximations.

Method 2: The Jackknife Approach:

We now propose an estimator for $\mathrm{Var}(\mathbb{E}_*[s^*])$ motivated by the classical jackknife procedure. Let $\mathcal{D}_n[i]$ denote the dataset remaining after deletion of the ith sample,

$$\mathcal{D}_n[i] = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

and let $t(\mathcal{D}_n[i])$, assumed to be well defined, denote the corresponding deleted point value of the statistic of interest, called the ith jackknife replicate. Letting t denote $\mathbb{E}_*[s^*]$, the jackknife estimate of variance is then defined as

$$\frac{n-1}{n}\sum(t(\mathcal{D}_n[i]) - \bar{t})^2 \tag{3.5}$$

where $\bar{t}$ is the mean of $t(\mathcal{D}_n[1]), \ldots, t(\mathcal{D}_n[n])$. This estimate in (3.5) is referred to as the Jackknife-After-Bootstrap (JAB) introduced by Efron (1992). These n jackknife replicates $t(\mathcal{D}_n[1]), \ldots, t(\mathcal{D}_n[n])$ can be thought of as approximations to $t(\mathcal{D}_n)$ and thus it is therefore natural to use their sample variance to estimate the variance of $t(\mathcal{D}_n)$, scaled by a factor of n

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

to account for their correlation. Note that

$$t(\mathcal{D}_n[i]) = \int s_{n-1}(x_1, x_{i-1}, x_{i+1}, \ldots, x_n)\, d\mathbb{P}_n(x_1) \cdots \mathbb{P}_n(x_{i-1}) \times \mathbb{P}_n(x_{i+1}) \cdots \mathbb{P}_n(x_n)$$

where $s_{n-1}(\cdot)$ is the version of s with $n-1$ arguments. Now suppose that we consider fixing the ith position rather than deleting the ith sample. We would thus replace $t(\mathcal{D}_n[i])$ by $t_{(i,j)}$ where

$$t_{(i,j)} = \int s(x_1, x_2, \ldots, x_{i-1}, X_j, x_{i+1}, \ldots, x_n)$$
$$d\mathbb{P}_n(x_1) \cdots \mathbb{P}_n(x_{i-1}) \times \mathbb{P}_n(x_{i+1}) \cdots \mathbb{P}_n(x_n).$$

Since s is permutation symmetric, $t_{(i,j)}$ is independent of i and so $t_{(1,j)} = t_{(i,j)}$ for all i. In the end, we still obtain n values $t_{(1,1)}, \ldots, t_{(1,n)}$ just as with the traditional JAB. We can therefore make use of those values and consider

$$\mathrm{JK}_{\mathrm{B}}^{\sharp} = \sum_j (t_{(1,j)} - \bar{t})^2 = \sum_j (e_j - s_0)^2 \tag{3.6}$$

as an estimate of the variance of $\mathbb{E}_*[s^*]$, where $\bar{t}$ is the mean of $t_{(1,1)}, \ldots, t_{(1,n)}$. Note that the goal here is to follow in the footsteps of the classical idea of the jackknife, but with a twist in the terms of the JAB to arrive at an estimator for $\mathrm{Var}(\mathbb{E}_*[s^*])$. An example may help in understanding the connection between JAB and JK_{B}. Let s_{n-1} denote a decision tree built with $n-1$ i.i.d. samples and similarly define s_n to be a decision tree built with n i.i.d. samples. Now, $t(\mathcal{D}_n[j])$ will be the average of all decision trees constructed

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

with $n-1$ samples randomly selected from $(X_1, \ldots, X_{j-1}, X_j, \ldots, X_n)$ (with replacement) and $t_{(1,j)}$ will be the average of all decision trees constructed with n samples, with the first sample held fixed at X_j and with the rest randomly selected from $(X_1, \ldots, X_n)$ (with replacement). The first type of bagging keeps X_j out of picture, giving it 0 probability to be selected into training and thus giving more weights to other samples, while the second gives X_j more weight and thereby reduces those of others. Both can be viewed as a variant of the version of bagging utilizing all bootstrap samples from $X_1, \ldots, X_n$, the variance of which is what we are seeking to estimate.

In the interest of completeness, we will later compare the JAB with the three other proposed methods in a small simulation study.

Method 3: The OLS Linear Regression Approach:

Recall that in a standard bootstrap procedure, B resamples of size n are sampled uniformly at random (with replacement) from the rows of $\mathcal{D}_n$. Each observation in the original dataset receives equal weight and so we can write that weight vector as $\mathbf{w}^* \sim \text{Multinomial}(\frac{1}{n}, \ldots, \frac{1}{n})$. Now realize that conditional on the original data $\mathcal{D}_n$, each bootstrap estimate s^* is a function of only those (empirical) weights $(w_1^*, \ldots, w_n^*)$ corresponding to the observations actually selected in the bootstrap sample. Consider the linear space spanned by $\mathbf{w}^* = (w_1^*, \ldots, w_n^*)$ and denote the l^* as the projection of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

$s(\mathbf{w}^*) - \mathbb{E}_*[s^*]$ onto that linear space.

We can use $\mathrm{Var}_*(l^*)$ as an estimate of $\mathrm{Var}(\mathbb{E}_*[s^*])$ which corresponds exactly to the setup where the relationship between the bootstrap estimates and observation weights is estimated via ordinary least squares linear regression.

Remark 1. In the sections that follow, we rely more heavily on Methods 1 and 2 for assessing the properties of the infinitesimal jackknife variance estimator. Nonetheless, the OLS approach in Method 3 reveals helpful and important insight into how the infinitesimal jackknife can be viewed. An explicit walkthrough of how the linear regression setup can be used to derive the infinitesimal jackknife estimator is provided in Appendix A.

3.2 Comparing the Three Approaches

We now examine how the three estimators derived above compare to each other. The first result below gives that the three estimators are identical whenever all bootstrap samples are used.

Theorem 1. *Suppose that we have data $\mathcal{D}_n$ and a statistic s. Let $(X_1^*, \ldots, X_n^*)$ be a general bootstrap sample of $\mathcal{D}_n$, $s^* = s(X_1^*, \ldots, X_n^*)$ and $w_j^* = \#\{i : X_i^* = X_j\}$, then (1) $\mathbb{E}_*[s^* w_j^*] = e_j$, (2) $l^* = \sum_j w_j^* \beta_j$, where $\beta_j = (e_j - s_0)$, and (3) $\mathrm{Var}_*(l^*) = \mathrm{JK}_{\mathrm{B}}^\sharp = \mathrm{IJ}_{\mathrm{B}}$ where $e_j = \mathbb{E}_*[s^* \mid X_1^* = X_j]$ and $s_0 = \mathbb{E}_*[s^*]$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

The proof of Theorem 1 can be found in Appendix B. Consider the more practical setting in which we draw only B bootstrap samples $(X^*_{b1}, \ldots, X^*_{bn})$ and calculate $s^*_b = s(X^*_{b1}, \ldots, X^*_{bn})$ for each $b = 1, \ldots, B$. And consider the bagged estimate $\overline{s^*} = \frac{1}{B}\sum_{b=1}^{B} s^*_b$. By the law of total variance, we have

$$\begin{aligned} \mathrm{Var}(\overline{s^*}) &= \mathrm{Var}(\mathbb{E}[\overline{s^*} \mid \mathcal{D}_n]) + \mathbb{E}[\mathrm{Var}(\overline{s^*} \mid \mathcal{D}_n)] \\ &= \mathrm{Var}(\mathbb{E}_*[s^*]) + \frac{1}{B}\mathbb{E}[\mathrm{Var}_*(s^*)]. \end{aligned} \tag{3.7}$$

The first term in Eq. (3.7) is dominant and so to provide a good estimate of $\mathrm{Var}(\overline{s^*})$, we must provide a good estimate of $\mathrm{Var}(\mathbb{E}_*[s^*])$. Since we do not employ all bootstrap samples, we cannot use $\mathrm{IJ_B}$ directly but we can use the B bootstrap replications to provide an estimate. First, a natural estimate of Cov_j is simply $\widehat{\mathrm{Cov}}_j$, the sample covariance of $(s^*_1, \ldots, s^*_B)$ and $(w^*_{1j}, \ldots, w^*_{Bj})$. Next, as for $e_j - s_0$, we can estimate s_0 with $\sum_{b=1}^{B} s^*_b/B$, and since e_j the expected value of s^* given $X^*_i = X_j$ for any $i = 1, \ldots, n$, a natural estimate would be the weighted average of the mean of s^*_b where $X^*_i = X_j$. Here, the weights are the proportion of times when $X^*_i = X_j$ across the B bootstrap samples. A straightforward calculation gives that

$$\widehat{e_j} = \sum_{b=1}^{B} \frac{w^*_{bj}}{\sum w^*_{bj}} s^*_b \quad \text{and} \quad \widehat{e_j} - \widehat{s_0} = \sum_{b=1}^{B} \left(\frac{w^*_{bj}}{\sum w^*_{bj}} - \frac{1}{B} \right) s^*_b. \tag{3.8}$$

Finally, the natural estimate of $\mathrm{Var}_*(l^*)$ is given by $\widehat{\mathrm{Var}}(\hat{l})$, where $\hat{l} = (\hat{l}_1, \ldots, \hat{l}_B)$ is the projection of $(s^*_1 - \bar{s^*}, \ldots, s^*_B - \bar{s^*})$ onto the linear space

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

spanned by $(w^*_{1j}, \ldots, w^*_{Bj})$ for $j = 1, \ldots, n$ which can be readily computed via least squares and $\widehat{\mathrm{Var}}(\hat{l}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{l}_b)^2$.

Putting all of this together, we have the following three limited number of bootstrap replications (i.e. not all bootstrap samples employed) variance estimators corresponding to the infinitesimal jackknife, jackknife, and OLS linear regression methods:

$$\begin{aligned}
\hat{\sigma}^2_{\mathrm{IJ}} &:= \widehat{\mathrm{IJ}}_{\mathrm{B}} = \sum_j \widehat{\mathrm{Cov}}_j^{\,2} = \sum_j \left[\frac{1}{B-1}\sum_b (s^*_b - \bar{s^*})(w^*_{bj} - \bar{w^*_j})\right]^2 \\
\hat{\sigma}^2_{\mathrm{JK}} &:= \widehat{\mathrm{JK}^{\sharp}_{\mathrm{B}}} = \sum_j (\widehat{e_j} - \widehat{s_0})^2 = \sum_j \left[\sum_b \left(\frac{w^*_{bj}}{\sum w^*_{bj}} - \frac{1}{B}\right) s^*_b\right]^2 \\
\hat{\sigma}^2_{\mathrm{OLS}} &:= \widehat{\mathrm{Var}}(\hat{l}) = \frac{1}{B-1}\sum_b (\hat{l}_b)^2.
\end{aligned} \tag{3.9}$$

Elaborating on how these methods relate to each other, note that rewriting $\hat{\sigma}^2_{\mathrm{IJ}}$ and $\hat{\sigma}^2_{\mathrm{JK}}$ as

$$\hat{\sigma}^2_{\mathrm{IJ}} = \sum_j \left[\sum_b \frac{w^*_{bj}}{B-1}(s^*_b - \bar{s^*})\right]^2, \quad \hat{\sigma}^2_{\mathrm{JK}} = \sum_j \left[\sum_b \frac{w^*_{bj}}{\sum_b w^*_{bj}}(s^*_b - \bar{s^*})\right]^2 \tag{3.10}$$

respectively, we can see that $\hat{\sigma}^2_{\mathrm{JK}}$ merely replaces $\sum_{b=1}^{B} w^*_{bj}$ with $B-1$. $\sum_{b=1}^{B} w^*_{bj}$ is actually close to $B-1$ since its expectation is B and standard deviation $\sqrt{B(1-1/n)}$. For $\hat{\sigma}^2_{\mathrm{IJ}}$ and $\hat{\sigma}^2_{\mathrm{OLS}}$, let $\mathbf{s} = (s^*_1 - \bar{s^*}..., s^*_B - \bar{s^*})^T$ and let $\mathbf{w}^*_j = (w^*_{1j}, \ldots, w^*_{Bj})^T$ for $j = 1, \ldots, n$. Further, denote the matrix of $[\mathbf{w}^*_1 - \bar{w^*_1}\mathbf{1}_B, \ldots, \mathbf{w}^*_n - \bar{w^*_n}\mathbf{1}_B)]$ as $\mathbf{W}$ and let $\mathbf{U}\Sigma\mathbf{V}^T$ be the singular value

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

decomposition (SVD) of $\mathbf{W}/\sqrt{B-1}$. Then

$$\begin{aligned}\hat{\sigma}^2_{\text{IJ}} &= \frac{1}{(B-1)}\mathbf{s}^T\frac{\mathbf{W}\mathbf{W}^T}{B-1}\mathbf{s} \\ &= \frac{1}{(B-1)}\mathbf{s}^T\mathbf{U}\cdot\Sigma^2\cdot\mathbf{U}^T\mathbf{s}.\end{aligned} \tag{3.11}$$

Since $n^{-1}\sum_j \mathbf{w}^*_j = \mathbf{1}_B$, the column space of $[\mathbf{w}^*_1,\ldots,\mathbf{w}^*_n]$ is the same as that of $[\mathbf{w}^*_1-\bar{w}^*_1\mathbf{1}_B,\ldots,\mathbf{w}^*_n-\bar{w}^*_n\mathbf{1}_B,\mathbf{1}_B]$, the projection of $\mathbf{s}$ onto the column space of $[\mathbf{w}^*_1,\ldots,\mathbf{w}^*_n]$ is the same as that onto $[\mathbf{w}^*_1-\bar{w}^*_1\mathbf{1}_B,\ldots,\mathbf{w}^*_n-\bar{w}^*_n\mathbf{1}_B,\mathbf{1}_B]$. Furthermore, since $\mathbf{1}^T_B\cdot\mathbf{s}=0$, the projection will be the same as the projection onto $[\mathbf{w}^*_1-\bar{w}^*_1\mathbf{1}_B,\ldots,\mathbf{w}^*_n-\bar{w}^*_n\mathbf{1}_B]=\mathbf{W}$. As the rank of $\mathbf{W}$ is $n-1$ and $\mathbf{U}$ forms an orthonormal basis of the column space of $\mathbf{W}$, we have

$$\hat{\sigma}^2_{\text{OLS}} = \frac{1}{(B-1)}\mathbf{s}^T\mathbf{U}\mathbf{U}^T\mathbf{s} = \frac{1}{(B-1)}\mathbf{s}^T\mathbf{U}\cdot\mathbf{I}_{n-1}\cdot\mathbf{U}^T\mathbf{s}. \tag{3.12}$$

Therefore,

$$\begin{aligned}\left|\hat{\sigma}^2_{\text{IJ}}-\hat{\sigma}^2_{\text{OLS}}\right| &= \left|\frac{1}{(B-1)}\mathbf{s}^T\mathbf{U}\left(\Sigma^2-\mathbf{I}_{n-1}\right)\mathbf{U}^T\mathbf{s}\right| \\ &= \left|\frac{1}{(B-1)}\text{trace}\left(\mathbf{s}^T\mathbf{U}\mathbf{U}^T\mathbf{s}\left(\Sigma^2-\mathbf{I}_{n-1}\right)\right)\right| \\ &\le \max_i\left(\Sigma^2-\mathbf{I}_{n-1}\right)\frac{1}{B-1}||\mathbf{s}||^2 \\ &= \max_i\left|\Sigma^2-\mathbf{I}_{n-1}\right|\widehat{\text{Var}}(s^*).\end{aligned} \tag{3.13}$$

To see why Σ^2 is close to $\mathbf{I}_{n-1}$, note that Σ^2 is a diagonal matrix of rank $n-1$ with the diagonal entries being the non-zero eigenvalues of $(B-1)^{-1}\mathbf{W}\mathbf{W}^T$, which are equal to the non-zero eigenvalues of $(B-1)^{-1}\mathbf{W}^T\mathbf{W}$. Holding

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

n fixed, as B gets large, the non-zero eigenvalues will approximate those of $\mathbb{E}_*[(B-1)^{-1}\mathbf{W}^T\mathbf{W}] = (\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)$, which are simply $n-1$ ones.

We stress here that our goal is not to contrast these three estimators in order to suggest an optimal form for use in practice, but rather to demonstrate how these estimators which are already used in practice are connected in theory. While these alternative derivations and estimators are helpful for providing insight into the infinitesimal jackknife in general, $\hat{\sigma}^2_{\mathrm{IJ}} = \widehat{\mathrm{IJ}}_{\mathrm{B}}$ seems to be the estimator that has garnered the most practical interest. In the following subsections, we thus focus on assessing its properties. A brief simulation study comparing the estimates obtained from the three different methods laid out above as well as the Jackknife-After-Bootstrap (JAB) is presented in Appendix E.

3.3 The Bias of $\widehat{\mathrm{IJ}}_{\mathrm{B}}$

We now begin to analyze the properties of $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ as an estimator. As discussed above in reference to (3.7), in order to understand how well $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ estimates $\mathrm{Var}(\overline{s^*})$, we need only understand how well it estimates $\mathrm{Var}(\mathbb{E}_*[s^*])$. Here we consider both the Monte Carlo bias and sampling bias of $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ where the sampling bias is considered with respect to variation in the data, whereas the Monte Carlo bias is considered with respect to bootstrap process con-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

ditional on the data. We combine these two sources of bias with

$$\begin{aligned}\mathbb{E}[\mathbb{E}_*[\widehat{\mathrm{IJ}}_{\mathrm{B}}]] - \mathrm{Var}(\mathbb{E}_*[s^*]) &= \mathbb{E}[\mathbb{E}_*[\widehat{\mathrm{IJ}}_{\mathrm{B}}] - \mathrm{IJ}_{\mathrm{B}}]] + \mathbb{E}[\mathrm{IJ}_{\mathrm{B}}] - \mathrm{Var}(\mathbb{E}_*[s^*]) \\ &\propto \underbrace{\mathbb{E}_*[\widehat{\mathrm{IJ}}_{\mathrm{B}}] - \mathrm{IJ}_{\mathrm{B}}}_{\textbf{Monte Carlo Bias}} + \underbrace{\mathbb{E}[\mathrm{IJ}_{\mathrm{B}}] - \mathrm{Var}(\mathbb{E}_*[s^*])}_{\textbf{Sampling Bias}}.\end{aligned} \tag{3.14}$$

and consider their impact separately in the following subsections. Note that here we only discuss about the bias of Monte Carlo thereby give no consideration to the variability of $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ arising from the Monte Carlo choice of the bootstrap replications.

Monte Carlo Bias: We first consider the Monte Carlo bias of $\widehat{\mathrm{IJ}}_{\mathrm{B}}$, which is relatively straightforward. Note that

$$\begin{aligned}\mathbb{E}_*[\widehat{\mathrm{IJ}}_{\mathrm{B}}] - \mathrm{IJ}_{\mathrm{B}} &= \sum_j \mathbb{E}_*[\widehat{\mathrm{Cov}}_j^2] - \mathrm{Cov}_j^2 = \sum_j \mathbb{E}_*[\widehat{\mathrm{Cov}}_j^2] - \mathbb{E}_*^2[\widehat{\mathrm{Cov}}_j] \\ &= \sum_j \mathrm{Var}_*(\widehat{\mathrm{Cov}}_j)\end{aligned} \tag{3.15}$$

and some calculation gives that

$$\begin{aligned}\mathrm{Var}_*(\widehat{\mathrm{Cov}}_j) &= -\frac{B-2}{B(B-1)}\mathrm{Cov}_*^2(s^*, w_j^*) + \frac{\mathrm{Var}_*(s^*)\mathrm{Var}_*(w_j^*)}{B(B-1)} \\ &\quad + \frac{\mathbb{E}_*[(s^* - \mathbb{E}_*[s^*])^2(w_j^* - \mathbb{E}_*[w_j^*])]^2}{B} \\ &:= \mathrm{I} + \mathrm{II}\end{aligned}$$

where

$$\begin{aligned}\mathrm{I} &= \frac{1}{B}\mathrm{Var}_*\left((s^* - \mathbb{E}_*[s^*])(w_j^* - \mathbb{E}_*[w_j^*])\right), \\ \mathrm{II} &= \frac{1}{B(B-1)}\left[\mathrm{Var}_*(s^*)\mathrm{Var}_*(w_j^*) + \mathrm{Cov}_*^2(s^*, w_j^*)\right].\end{aligned} \tag{3.16}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

The first term I is the dominant term and is $\mathcal{O}(1/B)$. Essentially, we see here that using $\widehat{\mathrm{Cov}}_j^2$ to estimate Cov_j^2 is analogous to using $\bar{X}^2$ to estimate $\mathbb{E}^2[X]$ for some random variable X, which is biased as $\mathbb{E}[\bar{X}^2]-\mathbb{E}^2[X] = \mathrm{Var}(X)/B$. Indeed, $\mathrm{Var}(X)/B$ may not be negligible, especially when B is small and the coefficient of variation of X is large. The variance estimator defined below offers a bias correction for $\widehat{\mathrm{IJ}}_{\mathrm{B}}$.

Definition 2. A Monte Carlo bias corrected version of $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ is given by

$$\widehat{\mathrm{IJ}}_{\mathrm{B}}^{mc} = \widehat{\mathrm{IJ}}_{\mathrm{B}} - \frac{1}{B}\sum_j \widehat{\mathrm{Var}}\left((s^* - \bar{s^*})(w_j^* - \bar{w_j^*})\right), \tag{3.17}$$

where $\widehat{\mathrm{Var}}$ denotes sample variance.

The bias correction term above is a sum over n terms. Then if B is small, the bias correction term will be significant. In recent work, Wager et al. (2014) proposed the Monte Carlo bias corrected estimator

$$\widehat{\mathrm{IJ}}_{\mathrm{B}}^{whe} = \widehat{\mathrm{IJ}}_{\mathrm{B}} - \frac{n}{B}\widehat{\mathrm{Var}}(s^*). \tag{3.18}$$

We see that if $\mathrm{Var}_*((s^*-\mathbb{E}_*[s^*])(w_j^*-\mathbb{E}_*[w_j^*])$ is close to $\mathrm{Var}_*(s^*-\mathbb{E}_*[s^*])\mathrm{Var}_*(w_j^*-\mathbb{E}^*[w_j^*]) = (1-\frac{1}{n})\mathrm{Var}_*(s^*-\mathbb{E}_*[s^*])$, then (3.18) is close to (3.17). This happens when n is sufficiently large that w_j does not heavily impact the values of s^*, so that w_j^* is nearly independent of s^* in calculating $\mathrm{Var}_*((s^*-\mathbb{E}_*[s^*])(w_j^*-\mathbb{E}_*[w_j^*])$. Simulations provided in the appendix demonstrate

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

this point. For small values of n, (3.17) is usually more natural and accurate.

Sampling Bias: Appendix B provides explicit calculations to examine how IJ_{B} behaves on some simple familiar examples. In particular, we look at two linear statistics, the sample mean and variance, and show, not surprisingly that IJ_{B} is asymptotically unbiased. On the other hand, for statistics like the sample maximum that are far from linear, the estimators perform quite poorly even when large numbers of bootstrap samples are employed. The following proposition provides an equivalence condition under which IJ_{B} is asymptotically unbiased and suggests that $\mathbb{E}_*[s^*]$ needs to be asymptotically linear. Unfortunately, this condition is not practically verifiable but gives some insight of how hard it is for IJ_{B} to be accurate in estimating the variance of $\mathbb{E}_*[s^*]$, and it naturally leads to the discussion of U-statistics, an alternative version of bagging, in the next section.

Proposition 1. *Let $\mathbb{E}_*[s^*]$ be the bootstrap smoothed alternative of s, then*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\mathrm{Var}_*(l^*)]}{\mathrm{Var}(\mathbb{E}_*[s^*])} = 1 \iff \lim_{n\to\infty} n(1-\rho) = 1 \tag{3.19}$$

where ρ is the correlation between e_1 and e_2 and $e_i = \mathbb{E}_*[s^* \mid X_1^* = X_i]$.

To help provide some intuition, consider the case where $s = \bar{X}$, which is linear. Here, we have $\rho = \frac{n^2-1}{n^2+n-1} = 1 - \frac{1}{n} + O(\frac{1}{n^2})$. From Proposition 1,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

we can see that $1-\rho = \frac{1}{n} + o(\frac{1}{n})$ is the condition required and a statistic as smoothed as sample mean just results in $1-\rho = \frac{1}{n} + O(\frac{1}{n^2})$. In this sense, it would be reasonable to say that the condition generally fairly restrictive, requiring $\mathbb{E}^*[s^*]$ to be close to linear.

In general, the replicates of bootstrap samples defies understanding the bias and other statistical properties of IJ_B comprehensively. To understand even $\mathbb{E}_*[s^*]$ alone requires an understanding of s^* for each bootstrap sample. Our main point here is thus that IJ_B, the infinitesimal jackknife for bootstrap, may only work well in very limited scenarios where the bootstrap statistic is asymptotic linear. The bootstrap statistic $\mathbb{E}_*[s^*]$ may have comparatively small variance, but is not necessarily much more linear than the original statistic s. We thus recommend taking care when employing IJ_B. On the contrary, U-statistics, as averages across subsamples, not only improve the stability, but also must be more linear than s. This second fact makes the infinitesimal jackknife consistent for variance estimation, under relatively more mild conditions.

4. Pseudo Infinitesimal Jackknife for U-statistics (ps-IJ_U)

While the infinitesimal jackknife is naturally appealing for estimating the variance of bagged statistics, the current resurgence in interest is due in large part to its potential for quantifying the uncertainty of complex su-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

pervised machine learning ensembles often formed by subsampling. Immediately following Efron (2014), Wager et al. (2014) showed that the same infinitesimal jackknife approach could be used to generate confidence intervals for random forests. In recent years, a number of works have expanded on this idea (Mentch and Hooker, 2016; Wager and Athey, 2018; Peng et al., 2022). Among the key breakthroughs was the realization that predictions generated by averaging across ensembles of subsampled regression estimates could be seen as akin to classical U-statistics. We thus now explore how the infinitesimal jackknife can be extended to the case of subsampling without replacement. In this case, $\mathbb{E}_*[s^*]$ is a U-statistic, which is often more convenient for theoretical analysis and also more likely to be close to linear. Here s is a (permutation-symmetric) function of k i.i.d. random variables and the U-statistic can be written as $\mathrm{U} = \binom{n}{k}^{-1} \sum_{(n,k)} s(X_{i_1}, \ldots, X_{i_k})$ where the sum is taken over all $\binom{n}{k}$ subsamples of size k. In the classical language of U-statistics, we would refer to s as the kernel and k as the order or rank of the kernel. The U-statistics we discuss throughout the remainder of this paper are assumed non-degenerate.

4.1 The Pseudo Infinitesimal Jackknife for U-statistics (ps-IJ$_{\mathrm{U}}$)

In recent work, Wager and Athey (2018) investigated the consistency and asymptotic normality of random forests where the individual trees were

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

constructed with subsamples of the original data. As part of this work, the authors proposed another estimate of variance wherein the format of the IJ for bootstrap was simply copied over to this subsampling regime to arrive at

$$\sum_j \mathrm{Cov}_*^2(s^*, w_j^*), \tag{4.1}$$

where $*$ refers to the subsampling procedure. We refer to this estimator as ps-IJ$_{\mathrm{U}}$, since it is not derived from the definition of the infinitesimal jackknife. The alternative estimator derived directly from the definition in (2.1) is denoted by IJ$_{\mathrm{U}}$. Generally speaking, IJ$_{\mathrm{U}}$ and ps-IJ$_{\mathrm{U}}$ are quite close to each other, though ps-IJ$_{\mathrm{U}}$ has a slighter simpler expression. Further discussion and comparisons are given in Appendix C.

It is, however, possible to provide a more rigorous motivation for ps-IJ$_{\mathrm{U}}$ that follows from the spirit of the infinitesimal jackknife if not the exact formulation of it. Recall from Section 2 (see (2.7)), we assume that $f(\mathbb{P}_n) - f(\mathrm{P})$ can be written as $\frac{1}{n}\sum_i f'(\mathrm{P}, X_i) + o_p(1/\sqrt{n})$, where the dominant term is a sum of i.i.d. random variables and we estimate the variance of $f'(\mathrm{P}, X_i)$ by $\frac{1}{n}\sum_i f'^2(\mathbb{P}_n, X_i)$. Now suppose that we write $f(\mathbb{P}_n) - f(\mathrm{P})$ as $\sum_i g(X_i) + o_p(1/\sqrt{n})$, where $g(X_i)$ is not necessarily $\frac{1}{n}f'(\mathrm{P}, X_i)$. From classical U-statistic theory, there is a natural candidate for $g(X_i)$: the Hájek projection - $\mathbb{E}[f(\mathbb{P}_n) - f(\mathrm{P})|X_i] = \frac{k}{n}\mathbb{E}[s - \mathbb{E}[s]|X_i]$. Since $\mathrm{Var}\left(\sum g(X_i)\right) = \frac{k^2}{n}V_1$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

$V_1 = \text{Var}(\mathbb{E}[s|X_1])$, we need only to propose a reasonable estimate for V_1 and we can then use $\frac{k^2}{n}\hat{V}_1$ as an estimate of the variance of the U-statistic. Since $V_1 = \mathbb{E}[\mathbb{E}[s|X_1] - \mathbb{E}[s]]^2$, a natural candidate of $\hat{V}_1$ would be

$$\frac{1}{n}\sum_j (\mathbb{E}_*[s^*|X_1^* = X_j] - \mathbb{E}_*[s^*])^2 = \frac{1}{n}\sum_j (e_j - s_0)^2. \tag{4.2}$$

As it turns out, (4.1) is the same as (4.2).

Proposition 2. *Let $\mathcal{D}_n^* = (X_1^*, \ldots, X_k^*)$ denote a subsample of size k from the original data $\mathcal{D}_n$ and define $w_j^* = \mathbf{1}_{X_j \in \mathcal{D}_n^*}$. Then*

$$\text{Cov}_*(s^*, w_j^*) = \frac{k}{n}(e_j - s_0),$$

where $*$ refers to the subsampling procedure, $e_j = \mathbb{E}_*[s^*|X_1^* = X_j]$ and $s_0 = \mathbb{E}_*[s^*]$. Thus,

$$\text{ps-IJ}_{\text{U}} = \sum_j \text{Cov}_*^2(s^*, w_j^*) = \left(\frac{\binom{k}{1}}{\binom{n}{1}}\right)^2 \sum_j (e_j - s_0)^2. \tag{4.3}$$

To understand the bias of ps-IJ$_{\text{U}}$, the H-decomposition is quite useful. To set this up, we first need to introduce following notation for kernels $s^1, \ldots, s^k$ of degrees $1, \ldots, k$. These kernels are defined recursively as

$$s^1(x_1) = s_1(x_1)$$

$$s^c(x_1, \ldots, x_c) = s_c(x_1, x_2, ..., x_c) - \sum_{j=1}^{c} \sum_{i_1,\ldots,i_j \in \{1,\ldots,c\}} s^j(x_{i_1}, \ldots, x_{i_j})$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

where $s_c(x_1,\ldots,x_c) = \mathbb{E}[s(x_1,\ldots,x_c,X_{c+1},\ldots,X_k)]-\mathbb{E}[s]$. Let $V_j = \mathrm{Var}(s^j)$ for $j = 1,\ldots,k$. Then $\mathrm{Var}(\mathrm{U}) = \sum_{j=1}^{k}\binom{k}{j}^2\binom{n}{j}^{-1}V_j$, and $\mathbb{E}[\mathrm{IJ}_{\mathrm{U}}]$ can also be written as a linear combination of those V_j. In particular, we have the following theorem.

Theorem 2. *The pseudo-IJ estimator of the variance of a U-statistic is defined as*

$$\text{ps-IJ}_{\mathrm{U}} = \frac{k^2}{n^2}\sum_{j=1}^{n}[e_j - s_0]^2 \tag{4.4}$$

where $e_j = \mathbb{E}_*[s^*|X_1^* = X_j]$ *and* $s_0 = \mathbb{E}_*[s^*]$. *Then*

$$\mathbb{E}[\text{ps-IJ}_{\mathrm{U}}] = \sum_{j=1}^{k} r_j \binom{k}{j}^2\binom{n}{j}^{-1}V_j, \tag{4.5}$$

where

$$r_j = \left(\frac{n-k}{n}\right)^2 \frac{j}{1-j/n}, \quad \textit{for } i = 1,\ldots,k. \tag{4.6}$$

Note that although our goal here is to use $\frac{k^2}{n}\hat{V}_1$ to estimate $\frac{k^2}{n}V_1$, $\mathbb{E}[\frac{k^2}{n}\hat{V}_1] = \mathbb{E}[\text{ps-IJ}_{\mathrm{U}}]$ involves higher order terms of $V_2,\ldots,V_k$. Though not ideal, it is unavoidable since we do not have new data generated from the underlying distribution. If we simply multiply ps-IJ$_{\mathrm{U}}$ by $(\frac{n}{n-k})^2\frac{n-1}{n}$ as proposed in Wager and Athey (2018), then only the first term is unbiased, but it doubles the quadratic term, triples the cubic term etc. (a similar phenomenon was discovered by Efron (1981) for the jackknife variance estimator). This explains why this estimation procedure is inflated in practice.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

In many applications, k is not small and so the higher order terms of Var(U) are not negligible and the effect of r_j cannot be ignored. Indeed, as alluded to at the beginning of this section, in many modern machine learning applications like random forests, k corresponds to the number subsamples utilized in the construction of each base learner and so k is generally best chosen to be as large as possible.

4.2 The Consistency of ps-IJ$_{\mathrm{U}}$ and Its Derivatives

In this subsection, we investigate the properties and consistency of ps-IJ$_{\mathrm{U}}$, beginning with the idea of generalized U-statistics, recently defined in Peng et al. (2022).

Definition 3 (Generalized U-statistic (Peng et al., 2022))**.** Suppose $X_1, \ldots, X_n$ are i.i.d. samples from some distribution P and let s denote a (possibly randomized) real-valued function that is permutation symmetric in its $k \leq n$ arguments. A generalized U-statistic with kernel s of order (rank) k refers to any estimator of the form

$$\mathrm{U}_{n,k,N,\omega} = \frac{1}{\hat{N}} \sum_{(n,k)} \rho s(X_{i_1}, \ldots, X_{i_k}; \omega) \tag{4.7}$$

where ω denotes i.i.d. randomness, independent of the original data. The ρ denotes i.i.d. Bernoulli random variables determining which subsamples are selected where $\mathbb{P}(\rho = 1) = N/\binom{n}{k}$ and $\hat{N}$ corresponds to the sum of the ρ.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

When $N = \binom{n}{k}$, the estimator in (4.7) is a generalized complete U-statistic and is denoted as $\text{U}_{n,k,\omega}$. When $N < \binom{n}{k}$, these estimators are generalized incomplete U-statistics. Note that each collection of subsamples is paired with an individual ρ sampled i.i.d., but we use a single ρ here for notational convenience.

Generalized U-statistics are essentially incomplete U-statistics with potentially extra randomness and where the order of the kernel may grow with n. Random forests, for example, are generalized U-statistics in which the additional randomness ω determines which features are eligible for splitting at each node in each tree. In recent work, Peng et al. (2022) proved that if $\frac{k}{n}(\zeta_k/k\zeta_{1,\omega} - 1) \to 0$, then the complete generalized U-statistic $\text{U}_{n,k,\omega}$ is asymptotically normal with variance $\frac{k^2}{n}\zeta_{1,\omega}$, where $\zeta_k = \text{Var}(s)$, the variance of the kernel, and $\zeta_{1,\omega} = V_1 = \text{Var}(\mathbb{E}[s|X_1])$, the covariance of two kernels with only one sample in common. Let $e_i^\omega = \binom{n-1}{k-1}^{-1} \sum s(X_i, \ldots; \omega)$ and let $s_0^\omega = \binom{n}{k}^{-1} \sum s(\ldots; \omega)$. Like ρ, each collection of subsamples is paired with an i.i.d. ω. Then the corresponding ps-IJ$_{\text{U}}$ for generalized U-statistics is defined as

$$\text{ps-IJ}_{\text{U}}^\omega = \frac{k^2}{n^2} \sum [e_i^\omega - s_0^\omega]^2. \tag{4.8}$$

The following theorem gives that the same conditions used by Peng et al. (2022) to establish asymptotic normality are sufficient to establish the con-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

sistency of ps-IJ$_{\mathrm{U}}^{\omega}$. In other words, if the (generalized) U-statistic is nearly linear, then ps-IJ$_{\mathrm{U}}^{\omega}$/Var(U$_{n,k,\omega}$) converges to one in probability.

Theorem 3. *Let $X_1, \ldots, X_n$ be i.i.d. from* P *and* U$_{n,k,\omega}$ *be a generalized complete U-statistic with kernel $s(X_1, \ldots, X_k; \omega)$. Let $\theta = \mathbb{E}[s]$, $\zeta_{1,\omega} = \mathrm{Var}(\mathbb{E}[s|X_1])$ and $\zeta_k = \mathrm{Var}(s)$. Then if $\frac{k}{n}\left(\frac{\zeta_k}{k\zeta_{1,\omega}} - 1\right) \to 0$,*

$$\text{ps-IJ}_{\mathrm{U}}^{\omega}/\mathrm{Var}(\mathrm{U}_{n,k,\omega}) \xrightarrow{p} 1. \tag{4.9}$$

Corollary 1. *If the conditions in Theorem 3 are met, then*

$$\begin{aligned} \mathrm{U}_{n,k,\omega} \pm z_{\alpha/2}\frac{n}{n-k}\sqrt{\text{ps-IJ}_{\mathrm{U}}^{\omega}} &= \mathrm{U}_{n,k,\omega} \pm z_{\alpha/2}\frac{n}{n-k}\sqrt{\sum \mathrm{Cov}_*^{\omega}(s^*, w_i^*)^2} \\ &= \mathrm{U}_{n,k,\omega} \pm z_{\alpha/2}\frac{k}{n-k}\sqrt{\sum (e_i^{\omega} - s_0^{\omega})^2} \end{aligned} \tag{4.10}$$

provides an asymptotically valid confidence interval for θ with confidence level $1-\alpha$. Note that $\mathrm{Cov}_*^{\omega}(s^*, w_i^*)$ can be defined in the same fashion as (4.8) to include the extra randomness and equals $\frac{k}{n}(e_i^{\omega} - s_0^{\omega})$.

The $\frac{n}{n-k}$ appearing in (4.10) corrects for finite sample bias. Note that Theorem 3.5 in Wager and Athey (2018) can be viewed as a special case of Theorem 3, where the kernel is taken as $\mathbb{E}_{\omega}[s(X_1, ..., X_k; \omega)]$. However, in the random forest setting, calculating such a statistic involves building all possible randomized trees on a collection of subsamples, a task which is generally computationally impossible in practical settings. Indeed, because of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

their inherent computational burden in calculating $\binom{n}{k}$ base estimates, the complete forms of these estimators are almost never utilized in practice. Fortunately, asymptotic normality for incomplete generalized U-statistics was also established in Peng et al. (2022). Thus, to establish asymptotically valid confidence intervals for these incomplete counterparts, we need only establish a consistent means of estimating the asymptotic variance of $\mathrm{U}_{n,k,N,\omega}$.

Let

$$\widehat{\text{ps-IJ}}_{\mathrm{U}}^{\omega} = \frac{k^2}{n^2}\sum_i [\hat{e}_i^{\omega} - \hat{s}_0^{\omega}]^2, \tag{4.11}$$

where

$$\hat{s}_0^{\omega} = \frac{1}{N}\sum s(...;\omega), \quad \hat{e}_i^{\omega} = \frac{n}{Nk}\sum s(X_i, ...;\omega). \tag{4.12}$$

Here, $\sum s(...;\omega)$ denotes the sum of all kernels that make up the incomplete U-statistic, whereas $\sum s(X_i, ...;\omega)$ denotes the sum of all kernels that make up the incomplete U-statistic and include X_i in their respective subsamples. The following theorem shows that we can obtain a consistent estimate of $\zeta_{1,\omega}$ so long as N is large enough to ensure that $\frac{n}{Nk\zeta_{1,\omega}} \to 0$. Importantly, this means that N need not be on the order of $\binom{n}{k}$.

Theorem 4. *Let $X_1, \ldots, X_n$ be i.i.d. from* P *and* $\widehat{\text{ps-IJ}}_{\mathrm{U}}^{\omega}$ *be as (4.11). Let* $\zeta_{1,\omega} = \mathrm{Var}(\mathbb{E}[s|X_1])$ *and* $\zeta_k = \mathrm{Var}(s)$. *Assume* $\mathbb{E}[s]$ *and* ζ_k *are bounded.*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

Then if $\frac{k}{n}\left(\frac{\zeta_k}{k\zeta_{1,\omega}} - 1\right) \to 0$ *and* $\frac{n}{Nk\zeta_{1,\omega}} \to 0$, *we have*

$$\widehat{\text{ps-IJ}}_{\text{U}}^{\omega} \Big/ \frac{k^2}{n}\zeta_{1,\omega} \xrightarrow{p} 1. \tag{4.13}$$

Remark 2. Consider the case that $\zeta_k/k\zeta_{1,\omega} \leq c_1$ and $k\zeta_k \geq c_2$ for some constants c_1 and c_2. These conditions simply imply that the variance of the kernel vanishes at a rate of at most $1/k$ and the linear term of the variance is not negligible. Theorem 4 states that we need only have $n \gg k$ and $N \gg nk$ in order to ensure that $\widehat{\text{ps-IJ}}_{\text{U}}^{\omega}/\frac{k^2}{n}\zeta_{1,\omega} \xrightarrow{p} 1$.

Theorem 4 provides a consistent estimate of the variance of predictions generated by subsampled random forests, where the number of trees in the random forest need not be $\binom{n}{k}$ and the trees themselves may be randomized. Note, of course, that the conditions in Theorem 4 are not guaranteed to hold generally and would need to be verified for whatever trees or other base learners are employed by the forest in practice.

To verify the conditions in Theorem 4, it is important to understand how $\zeta_{1,\omega}$ behaves as $k, n \to \infty$. Although, it is difficult in general to quantify $\zeta_{1,\omega}$ for the kernels based on adaptive nearest neighbor methods (most notably, decision trees), recently, work by Wager and Athey (2018) and Peng et al. (2022) proved that for "double sample trees", a special type of tree that uses one half of the subsample to build the tree structure and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

the other to form the prediction, $\zeta_{1,\omega}$ is well behaved and vanishes at a rate of $k^{-(1+\epsilon)}$, $\forall \epsilon > 0$. Additionally, Peng et al. (2022) provides empirical evidence in a small simulation that this term actually behaves well for trees built according to the original CART criterion as well.

It is also worth noting in (4.12) that the sum on left is taken over $\hat{N}$ terms whereas the right is taken over $\hat{N}_i$ terms. Note that $N = \mathbb{E}[\hat{N}]$ and $\frac{Nk}{n} = \mathbb{E}[\hat{N}_i]$ simply ease of proof and could be replaced by their empirical values $\hat{N}$ and $\hat{N}_i$, which are more natural and could result in more accurate estimation as suggested by the simulations in Appendix F. Applying the replacement, we have

$$\begin{aligned}
\frac{k^2}{n^2}\sum_i [\hat{e}_i^{\omega} - \hat{s}_0^{\omega}]^2 &= \sum_i \frac{k^2}{n^2}[\hat{e}_i^{\omega} - \hat{s}_0^{\omega}]^2 \\
&\xrightarrow{\text{replace}} \sum \frac{\hat{N}_i^2}{\hat{N}^2}\left[\frac{1}{\hat{N}_i}\sum s(X_i, \ldots; \omega) - \frac{1}{\hat{N}}\sum s(\ldots;\omega)\right]^2 \\
&= \sum_i \left[\frac{1}{\hat{N}}\sum s(\ldots;\omega) w_i^* - \frac{1}{\hat{N}}\sum s(\ldots;\omega)\frac{\hat{N}_i}{\hat{N}}\right]^2 \\
&= \sum_i \widehat{\text{Cov}}^2(s^*, w_i^*).
\end{aligned} \tag{4.14}$$

In the recent literature on random forests discussed above, $\sum_i \widehat{\text{Cov}}^2(s^*, w_i^*)$ is often the quantity used to estimate the variance of predictions in practice.

Recent work has established that $\text{U}_{n,k,N,\omega}$ has different asymptotic distributions depending on the number of subsamples N that are employed

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

(Peng et al., 2022). When $N \ll n/k$, $\mathrm{U}_{n,k,N,\omega} \sim \mathcal{N}(0, \zeta_k/N)$, when $N = O(n/k)$, $\mathrm{U}_{n,k,N,\omega} \sim \mathcal{N}(0, \frac{k^2}{n}\zeta_{1,\omega} + \frac{\zeta_k}{N})$, and when $N \gg n/k$ and $\mathcal{N}(0, \frac{k^2}{n}\zeta_{1,\omega})$. Thankfully, regardless of the setting, there are only two variance parameters that may need to be estimated: ζ_k and $\zeta_{1,\omega}$. We can estimate ζ_k simply by calculating sample variance of base learners built on non-overlapping sub-samples. Based on the above arguments, $\zeta_{1,\omega}$ can be estimated by $\widehat{\text{ps-IJ}}_{\mathrm{U}}^{\omega}$. Therefore, it is guaranteed that $\widehat{\text{ps-IJ}}_{\mathrm{U}}^{\omega} \Big/ \frac{k^2}{n}\zeta_{1,\omega} \xrightarrow{p} 1$ and $\widehat{\zeta}_k \Big/ \zeta_k \xrightarrow{p} 1$. Thus, Theorem 4 provides a means by which we can consistently estimate the variance of the (asymptotic) normal distribution established for random forests in Peng et al. (2022). This result together with Peng et al. (2022) thus provides a more complete picture of the distributional results for random forests, resting upon the same assumptions as on the decision trees.

Interestingly, for a random forest built with $N = O(n)$ decision trees, we will not be able to estimate its variance consistently by using only the trees contained in the random forest; this requires $\gg \frac{n}{k\zeta_{1,\omega}}$ decision trees. Because $k\zeta_{1,\omega} \leq \zeta_k$ and ζ_k usually tends toward 0 as k grows, $k\zeta_{1,\omega}$ thereby vanishes. This implies that the number of trees required for consistent variance estimation would be $\gg n$. These results shed light on the intuition established throughout the machine learning literature that it is always significantly more computationally intensive to estimate the variance of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

ensembles than to obtain the ensemble itself. A brief simulation study comparing the variance estimates is provided in Appendix F.

5. Discussion

The work above provides an in-depth examination of the infinitesimal jackknife estimate of variance for resampled statistics. We provided alternative perspectives on the estimator, most notably demonstrating its equivalence to an OLS regression of the bootstrap estimates on their respective sampling weights. Ultimately we derived three alternative estimators under the bootstrap regime and demonstrated their equivalence when all bootstrap samples are employed. We also examined both the Monte Carlo and sampling bias of the IJ estimator in the bootstrap setting and proposed a novel bias-corrected estimator, providing conditions under which it is asymptotically unbiased. In the latter portion of the work, we examine how these preliminary results translate outside the bootstrap setup by looking instead at subsampling. Here the statistics resemble a U-statistic and we derived corresponding results for generalized U-statistics, a new tool for analyzing modern learning ensembles like random forests. We also provided a formal motivation for the pseudo IJ often employed in practice and establish its consistency under similar linearity conditions. Importantly, we further established consistency for the finite sample (incomplete) versions of these

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

estimators so that one needn't utilize all possible subsamples in order for the empirical ps-IJ$_{\mathrm{U}}$ to form a consistent estimator.

Finally, recall from the previous section we assume the statistics are approximately linear so that when we write $\mathrm{Var(U)} = \sum_{j=1}^{k} \binom{k}{j}^2 \binom{n}{j}^{-1} V_j$, the variance is dominated by the first order term k^2/nV_1 and we propose an estimate of V_1 accordingly. One may wonder whether, if the remaining terms are not negligible, an improved estimate can be provided by also including estimates for V_j for $j = 2, \ldots, k$. While such estimates can be provided, establishing the superiority of the resulting estimator is a far more in depth undertaking that we reserve for future work. Further detailed discussion is provided in Appendix D.

Supplementary Material

Online supplementary material is provided, which contains an extended discussion and walk-through of the connection between OLS linear regression and the infinitesimal jackknife. Example calculations, simulations, and proofs are also provided.

References

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association 83*(403), 687–697.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

REFERENCES

Bickel, P. J., F. Götze, and W. R. van Zwet (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica 7*, 1–31.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Chang, J. and P. Hall (2015). Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika 102*(1), 203–214.

Davidson, R. and J. MacKinnon (2000). Improving the reliability of bootstrap tests. *Queens University Working paper no. 995*.

Davidson, R. and J. G. MacKinnon (2002). Fast double bootstrap tests of nonnested linear regression models. *Econometric Reviews 21*(4), 419–429.

Davidson, R. and J. G. MacKinnon (2007). Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics & Data Analysis 51*(7), 3259–3281.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*, 1–26.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika 68*(3), 589–599.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, Volume 38. Siam.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society: Series B (Methodological) 54*(1), 83–111.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association 109*(507), 991–1007.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

REFERENCES

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Giacomini, R., D. N. Politis, and H. White (2013). A warp-speed method for conducting monte carlo experiments involving bootstrap estimators. *Econometric theory 29*(3), 567.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association 69*(346), 383–393.

Huber, P. J. et al. (1972). The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics 43*(4), 1041–1067.

Jaeckel, L. A. (1972). *The infinitesimal jackknife*. Bell Telephone Laboratories.

Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(4), 795–816.

Mentch, L. and G. Hooker (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research 17*(1), 841–881.

Miller, R. G. (1974). The jackknife–a review. *Biometrika 61*(1), 1–15.

Mises, R. v. (1947). On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics 18*(3), 309–348.

Peng, W., T. Coleman, and L. Mentch (2022). Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics 16*(1), 232 – 292.

Politis, D. N. and J. P. Romano (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 2031–2050.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0131

Sengupta, S., S. Volgushev, and X. Shao (2016). A subsampled double bootstrap for massive data. *Journal of the American Statistical Association 111*(515), 1222–1232.

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

Volterra, V. (1887). *Sopra le funzioni che dipendono da altre funzioni*. Tip. della R. Accademia dei Lincei.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wager, S., T. Hastie, and B. Efron (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research 15*, 1625–1651.

White, H. (2000). A reality check for data snooping. *Econometrica 68*(5), 1097–1126.

Zhou, Z., L. Mentch, and G. Hooker (2021). V-statistics and variance estimation. *The Journal of Machine Learning Research 22*(1), 13112–13159.