Journal of Machine Learning Research 22 (2021) 1-48 Submitted 5/21; Revised 9/21; Published 10/21

V-statistics and Variance Estimation

Zhengze Zhou ZZ433@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA

Lucas Mentch LKM31@PITT.EDU
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260, USA

Giles Hooker GHOOKER@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Editor: Genevera Allen

Abstract

As machine learning procedures become an increasingly popular modeling option among applied researchers, there has been a corresponding interest in developing valid tools for understanding their statistical properties and uncertainty. Tree-based ensembles like random forests remain one such popular option for which several important theoretical advances have been made in recent years by drawing upon a connection between their natural subsampled structure and the classical theory of U-statistics. Unfortunately, the procedures for estimating predictive variance resulting from these studies are plagued by severe bias and extreme computational overhead. Here, we argue that the root of these problems lies in the use of subsampling without replacement and that with-replacement subsamples, resulting in V-statistics, substantially alleviates these problems.

We develop a general framework for analyzing the asymptotic behavior of V-statistics, demonstrating asymptotic normality under precise regularity conditions and establishing previously unreported connections to U-statistics. Importantly, these findings allow us to produce a natural and efficient means of estimating the variance of a conditional expectation, a problem of wide interest across multiple scientific domains that also lies at the heart of uncertainty quantification for supervised learning ensembles.

Keywords: V-statistics, U-statistics, Variance Estimation, Uncertainty Quantification, Supervised Ensembles

1. Introduction

There has been considerable interest in recent years in developing inferential procedures for random forests and related methods (Mentch and Hooker, 2016; Wager and Athey, 2018; Cui et al., 2017). In most cases, these procedures exploit the underlying subsampling structure to represent predictions from the resulting ensemble in terms of U-statistics. This representation then allows central limit theorems to be derived by extending those classical results (Hoeffding, 1948). While these advances have led to promising new inferential methodology (Mentch and Hooker, 2016; Hooker and Mentch, 2018; Zhou et al., 2018; Peng et al., 2019; Athey et al., 2019), each necessitates explicit variance estimation of the random forests predictions.

The primary component of this variance can be expressed as the variance of a conditional expectation. Estimation of parameters of this form has been explored for some time in the broader academic literature as it arises in a number of practical applications (Zouaoui and Wilson, 2003; Staum, 2009). Mentch and Hooker (2016); Hooker and Mentch (2018) exploit this representation explicitly, considering a structured subsampling approach that produces a nested Monte Carlo estimate for the variance of random forest predictions. Years earlier, Sun et al. (2011) investigated a similar estimator in an operations research context, using failure time stability as a motivating example. Goda (2017) recently derived a generalized version of this estimator and demonstrated that estimation could be done in a non-nested fashion. Wang and Lindsay (2014) were explicitly interested in estimating the variance of classical U-statistics and introduced a "partition resampling" scheme that was shown to be the best unbiased estimator as long as the kernel degree k satisfies $k \leq n/2$. In the context of ensemble variance estimation, Sexton and Laake (2009) proposed a bootstrap-of-little-bags approach and Wager et al. (2014); Wager and Athey (2018) utilize an extension of the Infinitesimal Jackknife estimator put forth in Efron (2014).

Unfortunately, all of these estimation procedures exhibit considerable upward bias unless the size of the ensemble is much larger than that required to construct a typical, predictively-accurate random forest. Such biased estimates generally result in conservative confidence intervals and significant decreases in the power of resulting hypothesis tests (Mentch and Hooker, 2016; Zhou et al., 2018). Methods proposed to correct this bias also frequently result in negative variance estimates, leading to *ad hoc* workarounds such as a pseudo-Bayes approach to enable inference (Athey et al., 2019).

In this paper, we argue that the difficulty of correcting the bias in these estimates results from the U-statistic construction itself that employs subsamples *without replacement*. This structure results in negative correlation between the trees in the ensemble that induces undesired bias in the bias correction. By simply employing subsampling *with* replacement, the entire ensemble can be represented as an expectation over the empirical distribution of the data and we show that this framework provides considerably more reliable inference than in the case of U-statistics as well as reducing the sensitivity of predictive performance to subsample size. Buja and Stuetzle (2006) studied some equivalence for bagging on resampling with or without replacement along with the effect of kernel size on bias and variance.

Formally, methods based around subsampling with replacement fall under the framework of V-statistics. These have been treated by either demonstrating their asymptotic equivalence to U-statistics, or via a series expansions (von Mises, 1947). However, the equivalence hold only when the size of the V-statistic kernel grows more slowly than $n^{1/4}$ – meaning that the number of observations given to each tree grows very slowly – and series expansions are complicated by the non-differential greedy tree-building process. Instead, we show that a V-statistic of any order can be exactly represented as a U-statistic but employing a different kernel, allowing us to invoke the central limit theorems already derived for infinite-order U-statistics, although some variation must be made to account for the use of random sampling for subsets.

Importantly, the general framework we develop allows us to present a unified theory for the general problem of variance estimation from which we can bridge the gap between the numerous approaches discussed above. In particular, we propose *balanced* method (BM) for variance estimation and show that it enjoys lower bias than the alternatives given ensembles of equal size. Further, we establish a close connection between the BM and the Infinitesimal Jackknife (IJ) and prove their equivalence under a natural condition. To estimate variance in the limiting distribution in finite sample case, we develop a bias-corrected version of BM through an ANOVA-like framework (Sun et al., 2011). The new estimator is shown to produce much more reliable results with a moderate number of base learners such as would be incurred, for example, by utilizing a traditional bootstrap approach.

Through the remainder of this paper, Section 2 and 3 review U- and V-statistics, their use in the theory of machine learning methods and derive their equivalence, although with modifications to the expression for variance in the incomplete case. In Section 4, we then examine variance estimates where we derive a more efficient estimate version of the estimate in Mentch and Hooker (2016) and show that this is equivalent to the Infinitesimal Jackknife if every observation is used the same number of times in the ensemble. This framework allows us to explicitly derive bias corrections in Section 5 and we show that while the natural estimate of the bias over-corrects the variance for the case of U-statistics, it is unbiased in the case of V-statistics. Section 6 extends the asymptotic results to randomized ensembles. Empirical studies in Section 7 corroborate the developed asymptotic theories and the effectiveness of variance estimation procedure; it also suggest that the change from without replacement to with replacement subsampling does not have a consistent effect on predictive performance but that the latter makes that performance less sensitive to subsample size, bringing the method closer to the original bootstrap sampling proposed in Breiman (2001). These results suggest that subsampling with replacement should be considered the appropriate default for ensemble methods.

Proofs of all results, additional experiment studies and further discussions are collected in the Appendix. Code accompanying this paper can be found at `https://github.com/ZhengzeZhou/V-statistics-and-Variance-Estimation`.

1.1 A Motivating Example

We will illustrate the practical importance of our contribution through a motivating example using Boston Housing Data[1]. The data set contains 506 samples with 13 features and the target is the median value of owner-occupied homes in $1000's in the area of Boston Mass. To simulate practical use cases, we leave 20% of the data as test set and train on the remaining samples. A random sample is selected from the test set and a 95% confidence interval for predictions are calculated by the methods described later in this paper.

Previous work suggests building ensembles by sampling without replacement (U-statistics), and estimating predictive variance by Infinitesimal Jackknife (IJ). We build the a random forests with subsample size 100 on the training data and calculate the intervals on the selected test sample, while varying the number of base learners $B = 100, 500, 1000, 5000, 10000$. The results are depicted by dashed lines in Figure 1 . Here the confidence interval of predictions are calculated as $predicted\_value \pm 1.96 \times estimated\_standard\_deviation$. The width

[1] `https://archive.ics.uci.edu/ml/machine-learning-databases/housing/`

of the intervals decrease as B becomes larger and small values of B results in conservative estimates. The ideal number of B is usually prohibitively large in practice ($B > 5000$ in this example) in order to get accurate variance estimation. This demonstrates the insufficiency of existing work in conducting inference for ensembles, despite solid theoretical properties of U-statistics.

This paper proposes the use of V-statistics by sampling with replacement along with a bias-corrected variance estimator (solid lines in Figure 1) . We can see that the proposed bias corrected estimates yield more accurate confidence intervals with moderate size of B. We would like to emphasize that the contributions of our work are twofold: in additional to theoretical advancements in analyzing asymptotics of V-statistics, we also provide a general framework for efficient variance estimation, which is of great significance for practical applications.

We note here that this interval is based on a central limit theorem that is centered on the expectation of the random forests, rather than using a target conditional mean $\mathbb{E}(Y|X = x)$. The consistency of such intervals will then depend on the size of the bias associated with the particular tree-building method used in the random forests; see (Wager and Athey, 2018; Scornet et al., 2015) for examples.

Our focus in this paper is on the structure of the ensemble rather than its constituent and we do not address confidence interval consistency here. Alternatively, Zhou and Hooker (2018) multiplies the calculated standard error by $\sqrt{2}$ to produce "reproduction intervals", giving the range in which an independently-generated random forests would fall and thus a notion of stability. For the sake of simplicity, we have used standard interval calculations below.

2. Related Work on U-statistics

We first give a brief introduction on the notion of U-statistics, and then illustrate how it can be utilized in the analysis of ensemble models.

Assume that we have a training set $\mathcal{D} = \{Z_1, \ldots, Z_n\}$ of i.i.d. observations of the form $Z_i = (\mathbf{X}_i, Y_i)$ drawn from an underlying distribution F_Z, where $\mathbf{X} = (X_1, \ldots, X_p) \in \mathcal{X}$ are p covariates. We want to estimate a parameter of interest θ. Suppose there exists an unbiased estimator h of θ that is a function of $k \leq n$ arguments (we call h a kernel of size or degree k) so that

$$\theta = \mathbb{E}h\left(Z_1, \ldots, Z_k\right)$$

and without loss of generality, assume that h is permutation symmetric in its arguments and $\mathbb{E}h^2\left(Z_1, \ldots, Z_k\right) < \infty$. Then the minimum variance unbiased estimator for θ is given by

$$U_n = \frac{1}{\binom{n}{k}} \sum_{i} h\left(Z_{i_1}, \ldots, Z_{i_k}\right) \tag{1}$$

where $\{Z_{i_1}, \ldots, Z_{i_k}\}$ consists of k distinct elements from the original sample $\{Z_1, \ldots, Z_n\}$ and the sum is taken over all $\binom{n}{k}$ subsamples of size k. The estimator in (1) is referred to as a *complete* U-statistic with kernel h of degree k.

There are some natural extensions of (1). To produce more predicive ensembles, we would like k to grow with n so the kernel will have access to more information from the

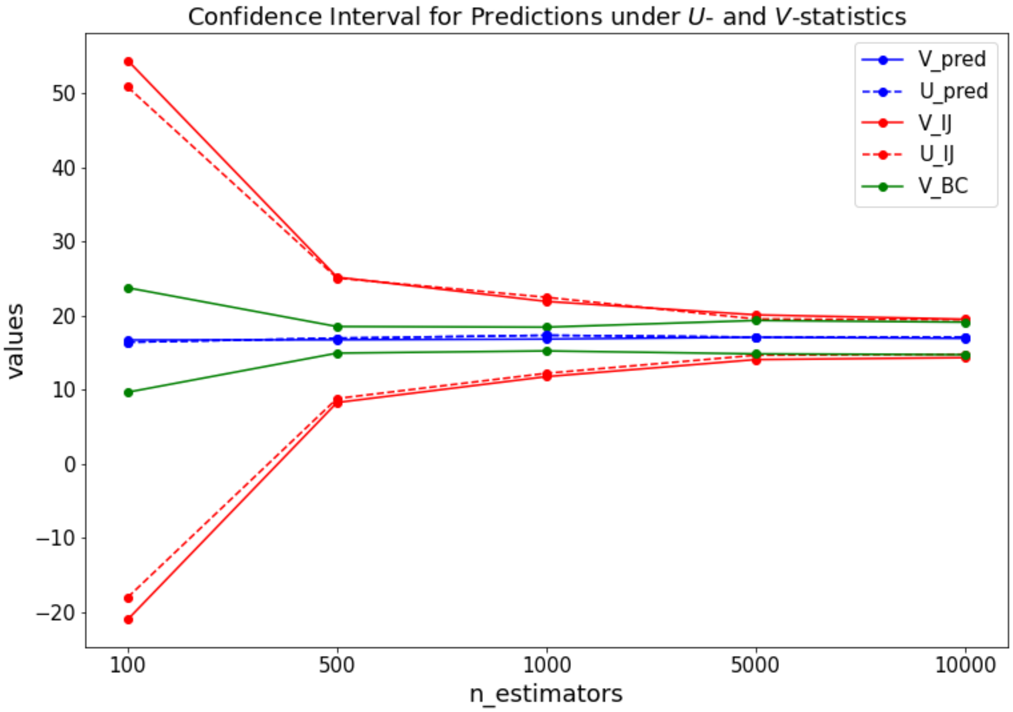


Figure 1: Confidence intervals for predictions based on U- and V-statistics. Dashed and solid lines are for U- and V-statistics respectively. The x-axis denotes the number of base learners used in an ensemble. Prediction values are drawn in blue, while IJ estimates of 95% confidence intervals are drawn in red. Bias corrected estimator for V-statistics is shown in green.

data. This results in a kernel that varies with n, and an *Infinite Order* U-statistic (IOUS; Frees, 1989)

$$U_{n,k_n} = \frac{1}{\binom{n}{k_n}} \sum_i h_{k_n}\left(Z_{i_1}, \ldots, Z_{i_{k_n}}\right). \tag{2}$$

Further, evaluating all $\binom{n}{k_n}$ kernels is computationally infeasible for even moderately sized n or k_n and thus an estimate can be achieved by averaging over only $B_n < \binom{n}{k_n}$ subsamples. Incorporating this, the estimator becomes an *Incomplete Infinite Order* U-statistic

$$U_{n,k_n,B_n} = \frac{1}{B_n} \sum_i h_{k_n}\left(Z_{i_1}, \ldots, Z_{i_{k_n}}\right). \tag{3}$$

In (2) and (3) we use subscripts to denote that values of k and B may depend on n, and the degree of kernel h is k_n.

U-statistics of form (1) were first studied in Halmos (1946) and Hoeffding (1948), where the latter also shows that these statistics are asymptotically normally distributed. Sen (1992) provides a review of Hoeffding's seminal paper and outlines the importance of U-statistics in modern statistical theory. A comprehensive treatment of the classical U-statistics results can be found in Lee (1990) and Serfling (2009). Certain basic properties, such as almost sure consistency and asymptotic normality, are proved to hold in the case of (2) and (3) in Frees (1989). The connection between U-statistics and ensemble methods had not been observed until very recently in the work of Mentch and Hooker (2016) and Wager and Athey (2018).

For simplicity we will focus on the regression setting, where predictions are assumed to be continuous. This can also incorporate binary classification as long as the model predicts the probability by averaging outputs instead of predicting a label obtained from a majority vote. We are interested in estimating the conditional mean function at a test point x

$$\mu(x) = \mathbb{E}(Y|X=x).$$

Given a base learner h, ensemble methods generate resamples $R_1, \ldots, R_B$ of the original data, apply h to each resample, and produce final point estimates by averaging over those generated by each model, yielding estimates of the form

$$\frac{1}{B}\sum_{i=1}^{B} h\left(x; \omega_i, R_i\right).$$

Here, the ω_i denotes an auxiliary randomization parameter as used in randomized ensembles like random forests, but which may be dropped for simpler (non-randomized) estimation procedures like bagging[2]; Peng et al. (2019) refers to these estimators as generalized U-statistics. When all instances of the randomness are considered, note that the kernel again becomes nonrandom (we can write the kernel as $E_{\omega_i} h\left(x; \omega_i, R_i\right)$), as in Wager and Athey (2018) where the authors assume B is large enough for Monte Carlo effects not to matter.

[2] In some papers, ω_i incorporates drawing the subsamples in bagging and doing both resampling and random feature selection in random forests. Here ω has a different meaning as we write out the resamples explicitly.

The conventional procedure in random forests is to take $R_1, \ldots, R_B$ to be bootstrap samples, which turns out very difficult to analyze statistically. Mentch and Hooker (2016) propose the following procedure to construct an ensemble. Given a training set $\mathcal{D}$ of size n, an ensemble consisting of B_n base learners is constructed using subsamples of size k_n

$$U_{n,k_n,B_n}(x) = \frac{1}{B_n} \sum_{i=1}^{B_n} h_{k_n}\left(x; Z^*_{i_1}, \ldots, Z^*_{i_{k_n}}\right) \tag{4}$$

where $\{Z^*_{i_1}, \ldots, Z^*_{i_{k_n}}\}$ is drawn *without* replacement from $\{Z_1, \ldots, Z_n\}$. This fits into the statistical framework of U-statistics and asymptotic normality can be demonstrated under some regularity conditions (see Peng et al. (2019) for refined results). In particular, the explicit expression for the variance of predictions at any given point can be written in closed-form

$$\frac{k_n^2}{n}\zeta_{1,k_n} + \frac{1}{B_n}\zeta_{k_n,k_n}. \tag{5}$$

For a given c, $1 \leq c \leq k_n$, the variance parameters are defined as

$$\zeta_{c,k_n} = \text{cov}\left(h_{k_n}(Z_1, \ldots, Z_{k_n}), h_{k_n}(Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{k_n})\right) \tag{6}$$

where $Z'_{c+1}, \ldots, Z'_{k_n}$ are i.i.d. copies from the same distribution F_Z and independent of the original data $Z_1, \ldots, Z_n$. For notational simplicity, we drop the test point x in (6).

Within these results, as in ours below, the asymptotic distribution is centered at $\theta_k = \mathbb{E}h_{k_n}(Z_1, \ldots, Z_{k_n})$ instead of the true conditional mean $\mathbb{E}(Y|X = x)$. As noted above, this means that any inferential statements must, in general, be made about the sampling structure of the ensemble rather than the underlying data generating process. A careful analysis of specific choices of the base learner h and the relationship between covariates X and response Y are central in achieving consistent predictions and is not the focus of this paper. Some work along these lines includes Wager and Athey (2018) which focus on particular tree-building methods, and Scornet et al. (2015) which demonstrate the $\mathbb{L}^2$ consistency for random forests when the underlying response corresponds to an additive regression model. Also note that the asymptotic normality result in Wager and Athey (2018) can be viewed as a special case of (5). Here, the authors assume that ensemble size B is large enough for Monte Carlo effects not to matter, in which case (5) reduces to $\frac{k_n^2}{n}\zeta_{1,k_n}$.

3. V-statistics

V-statistics are closely related to U-statistics except that the data used in each kernel is sampled *with* replacement. Similar to (1), a *complete* V-statistic with kernel h of degree k is defined as

$$V_n = n^{-k} \sum_{i_i=1}^{n} \ldots \sum_{i_k=1}^{n} h_k\left(Z_{i_1}, \ldots, Z_{i_k}\right) \tag{7}$$

where $\{Z_{i_1}, \ldots, Z_{i_k}\}$ consists of k elements from $\{Z_1, \ldots, Z_n\}$ and the sum is taken over all n^k subsamples of size k. An *Infinite Order* V-statistic (IOVS) is defined analogously to (2)

$$V_{n,k_n} = n^{-k_n} \sum_{i_i=1}^{n} \dots \sum_{i_{k_n}=1}^{n} h_{k_n}\left(Z_{i_1}, \dots, Z_{i_{k_n}}\right). \tag{8}$$

Asymptotic equivalence between V- and U-statistics for fixed kernel degree is a well studied topic (Lee, 1990). Previous work has also shown that the equivalence hold when the size of V-statistic kernel grows more slowly than $n^{\frac{1}{4}}$ (Shieh, 1994). We provide a rigorous analysis of this argument in Appendix A. However, this growth rate of kernel size is fairly restrictive: the number of observations given to each base learner grows very slowly, which can harm predictive performance. In the following, we show that a V-statistic of any order can be exactly represented as a U-statistic but employing a different kernel, thus enabling us to discard the restriction to attain a more general asymptotic results of V-statistics.

3.1 Representation As U-statistics

This section develops a broader connection between V- and U-statistics to show that the former automatically achieve almost all the properties of the latter. A complete, infinite order V-statistic V_{n,k_n} with kernel h_{k_n} can be written as a corresponding U-statistic but with a more complicated kernel derived from $h_{k_n}(\cdot)$.

Let Ω denote the set $\{1, 2, \dots, n\}$. We use $B^{k_n}(\Omega)$ to denote all size k_n permutations of Ω with replacement, and let $S^{k_n}(\Omega)$ denote subsamples of size k_n without replacement so that $|B^{k_n}(\Omega)| = n^{k_n}$ and $|S^{k_n}(\Omega)| = \binom{n}{k_n}$. We can write V_{n,k_n} as

$$V_{n,k_n} = n^{-k_n} \sum_{b \in B^{k_n}(\Omega)} h_{k_n}\left(Z_b\right)$$

where b has k_n elements and Z_b are those Z's with index in b.

Equivalently, V_{n,k_n} can be expressed as

$$V_{n,k_n} = n^{-k_n} \sum_{s \in S^{k_n}(\Omega)} \sum_{b \in B^{k_n}(s)} \omega_b h_{k_n}\left(Z_b\right)$$

where ω_b is the weight associated with each evaluation of h_{k_n} to account for the multiplicity in sampling the same b from $B^{k_n}(s)$ for different s. For $b = \{i_1, i_2, \dots, i_{k_n}\}$, we use $u(b) \in \{1, 2, \dots, k_n\}$ to denote the number of unique elements in b and we have

$$\omega_b = \frac{1}{\binom{n-u(b)}{k_n-u(b)}}.$$

We can thus express V_{n,k_n} as a U-statistic

$$V_{n,k_n} = \frac{1}{\binom{n}{k_n}} \sum_{s \in S^{k_n}(\Omega)} h^*_{k_n}\left(Z_s\right) \tag{9}$$

where the kernel $h^*_{k_n}$ is defined as

$$h^*_{k_n}(Z_s) = \frac{\binom{n}{k_n}}{n^{k_n}} \sum_{b \in B^{k_n}(s)} \omega_b h_{k_n}\left(Z_b\right).$$

Here ω_b is defined as before, and

$$\sum_{b \in B^{k_n}(s)} \omega_b = \frac{n^{k_n}}{\binom{n}{k_n}}.$$

A general result for the aymptotics of V-statistics is stated in the following theorem, allowing us to remove the restriction $k_n = o(n^{\frac{1}{4}})$.

First we define a class to which the kernel function h_{k_n} belongs:

$$\mathcal{H} = \left\{ h : \sup_{k_n} \mathbb{E} h^2_{k_n} \left(Z_{i_1}, \ldots, Z_{i_{k_n}} \right) < \infty \right\}$$

where $(i_1, \ldots, i_{k_n})$ are chosen from $\{1, \ldots, k_n\}$ with replacement.

Theorem 1 *Let $Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} F_Z$ and let V_{n,k_n,B_n} be an incomplete, infinite order V-statistic with kernel h_{k_n} such that $h_{k_n} \in \mathcal{H}$. Let $\theta^*_{k_n} = \mathbb{E} h^*_{k_n}$. Then under the assumption that $\lim \frac{\zeta^*_{k_n,k_n}}{n\zeta^*_{1,k_n}} \to 0$, we have*

$$\frac{\left(V_{n,k_n,B_n} - \theta^*_{k_n} \right)}{\sqrt{\frac{k_n^2}{n} \zeta^*_{1,k_n} + \frac{1}{B_n} \zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

In the complete case where $B_n = n^{k_n}$, we have

$$\frac{\left(V_{n,k_n} - \theta^*_{k_n} \right)}{\sqrt{\frac{k_n^2}{n} \zeta^*_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Here, the variance parameter ζ^*_{1,k_n} is defined as in Equation (6) by replacing kernel h_{k_n} with $h^*_{k_n}$

$$\zeta^*_{1,k_n} = cov\left(h^*_{k_n}(Z_1, \ldots, Z_{k_n}), h^*_{k_n}\left(Z_1, Z_2^{'}, \ldots, Z_{k_n}^{'} \right) \right)$$

and $\zeta_{k_n,k_n} = var(h_{k_n}(Z_1, \ldots, Z_{k_n}))$ is still the variance across individual kernels h_{k_n}.

As above, we note that our asymptotic distribution is centered on$\theta^*_{k_n} = \mathbb{E} h^*_{k_n}$. This theorem provides a more general result for the asymptotics of V-statistics. It is essentially a reduction to U-statistics by constructing a new kernel representation. The variance expression $\frac{k_n^2}{n} \zeta^*_{1,k_n} + \frac{1}{B_n} \zeta_{k_n,k_n}$ again can be viewed as two parts: the first part $\frac{k_n^2}{n} \zeta^*_{1,k_n}$ comes from the complete case; the second part $\frac{1}{B_n} \zeta_{k_n,k_n}$ is the additional Monte Carlo variance introduced due to incomplete case, which is why ζ_{k_n,k_n} only involves the original kernel h_{k_n} instead of the composite kernel $h^*_{k_n}$. Peng et al. (2019) provides a unified analysis of these two components by incorporating the choice of subsamples into the randomization parameters of the generalized U-statistics. This strategy is not available in our case, as detailed in Section 6.

Unlike Theorem 10 (see Appendix A) where the expected value of V- and U-statistics are the same asymptotically when $k_n = o(n^{\frac{1}{4}})$, in the more general case here one may not have the expected value of the new kernel $h^*_{k_n}$ equals that of h_{k_n}. The central limit theorem centers at the expectation of the statistics of interest. How to quantify the bias of the predictions is out of the scope of this paper; see Wager and Athey (2018) for a careful treatment of decision tree based ensembles.

The introduction of new kernel $h^*_{k_n}$ facilitates theoretical analysis, but it brings challenges in estimating variance component ζ^*_{1,k_n} directly: it is not feasible to calculate $h^*_{k_n}(Z_s)$ for any $s \in S^{k_n}(\Omega)$. We will see in Section 4 that as a general variance estimation method, the Infinitesimal Jackknife (IJ) can be applied to the original kernel function h_{k_n}. And based on Theorem 3, Balanced Variance Estimation Method is equivalently valid without resorting to evaluating $h^*_{k_n}(Z_s)$ directly.

4. Variance Estimation

This section addresses how to estimate variance in the limiting distribution. Mentch and Hooker (2016) propose Internal Variance Estimation Method (IM) based on a two-level sampling procedure. Inspired from this, we design the Balanced Variance Estimation Method (BM) which is shown to have lower bias compared to IM. Unlike IM and BM, the Infinitesimal Jackknife (IJ) employed in Wager and Athey (2018) does not depend on an explicit expression for the variance term. All methods presented can apply to both U- and V-statistics, though they exhibit different performances when sampling without or with replacement, especially in terms of bias. For notational simplicity, during the development of BM and IM we will use h_{k_n} to denote the kernel function and ζ_{c,k_n} to denote variance parameters. For general V-statistics as characterized by Theorem 1, one can substitute them for $h^*_{k_n}$ and ζ^*_{c,k_n} whenever needed.

IM and BM operate by directly estimating ζ_{1,k_n} and ζ_{k_n,k_n} as defined in (6). Notice that $\zeta_{k_n,k_n} = \mathrm{var}(h_{k_n}(Z_1,\ldots,Z_{k_n}))$, which can be simply estimated as the variance across all base learners. The estimation for ζ_{1,k_n} is much more involved. The sample covariance between predictions may serve as a consistent estimator, but in practice it is numerically unstable and often results in negative variance estimates (Mentch and Hooker, 2016). Thus we work with the equivalent expression for ζ_{1,k_n} (Lee, 1990)

$$\zeta_{1,k_n} = \mathrm{var}\left(\mathbb{E}\left(h_{k_n}(Z_1,\ldots,Z_{k_n})\,|Z_1 = z_1\right)\right). \tag{10}$$

Expressions of the form from (10) belong to an important theme in statistics: estimating the variance of a conditional expectation. It is usually related to uncertainty quantification and has been studied intensively in a number of fields (Zouaoui and Wilson, 2003; Staum, 2009). For a more detailed review, we refer readers to Sun et al. (2011).

In what follows, assume we have data $\mathcal{D} = \{Z_1,\ldots,Z_n\}$ of i.i.d. observations of the form $Z_i = (\mathbf{X}_i, Y_i)$, and a kernel function $h_{k_n}(Z_1,\ldots,Z_{k_n})$. For simplicity, we suppress notations by dropping the test point x in the kernel expression.

4.1 Internal Variance Estimation Method

IM was first proposed in Mentch and Hooker (2016) wherein the estimates are obtained as a result of restructuring the ensemble building procedure. It can be viewed as a nested two-level Monte Carlo, where we need to choose n_{OUT} and n_{IN} for the number of outer and inner iterations respectively. See Algorithm 1 for details.

Algorithm 1 Internal Variance Estimation Method

```
for $i$ in 1 to $n_{\text{OUT}}$ do
    Select initial fixed point $\tilde{\mathbf{z}}^{(i)}$
    for $j$ in 1 to $n_{\text{IN}}$ do
        Select subsample $\mathcal{S}_{\tilde{\mathbf{z}}^{(i)},j}$ of size $k_n$ from training set that includes $\tilde{\mathbf{z}}^{(i)}$
        Build base learner and evaluate $h_{k_n}(\mathcal{S}_{\tilde{\mathbf{z}}^{(i)},j})$
    end for
    Record average of the $n_{\text{IN}}$ predictions
end for
Compute the variance of the $n_{\text{OUT}}$ averages to estimate $\zeta_{1,k_n}$
Compute the variance of all predictions to estimate $\zeta_{k_n,k_n}$
Compute the mean of all predictions to obtain final ensemble prediction
```

We use the shorthand $h_{i,j}$ to denote $h_{k_n}(\mathcal{S}_{\tilde{\mathbf{z}}^{(i)},j})$. The average across inner level is calculated as $\bar{h}_i = \frac{1}{n_{\text{IN}}}\sum_{j=1}^{n_{\text{IN}}} h_{i,j}$. Further we use $\bar{h} = \frac{1}{n_{\text{OUT}}}\sum_{i=1}^{n_{\text{OUT}}} \bar{h}_i$ to denote the average across outer level i. Then the estimates for ζ_{1,k_n} and ζ_{k_n,k_n} can be expressed as

$$\hat{\zeta}_{1,k_n}^{\text{IM}} = \frac{1}{n_{\text{OUT}}-1}\sum_{i=1}^{n_{\text{OUT}}}\left(\bar{h}_i - \bar{h}\right)^2$$

and

$$\hat{\zeta}_{k_n,k_n}^{\text{IM}} = \frac{1}{n_{\text{IN}} \times n_{\text{OUT}} - 1}\sum_{i=1}^{n_{\text{OUT}}}\sum_{j=1}^{n_{\text{IN}}}\left(h_{i,j} - \bar{h}\right)^2 .$$

4.2 Balanced Variance Estimation Method

As will be shown in Figure 2 in Section 5.1, the estimator for $\hat{\zeta}_{1,k_n}^{\text{IM}}$ given by IM is severely biased upwards when n_{IN} and n_{OUT} are not sufficiently large ($B_n = n_{\text{IN}} \times n_{\text{OUT}}$). IM is not optimal in the sense that it does not utilize all the information in the ensemble. In particular, $h_{i,j}$ is only used once in the outer iteration i when conditioned on $\tilde{\mathbf{z}}^{(i)}$. Ideally we could also utilize $h_{i,j}$ by conditioning on the remaining $k_n - 1$ inputs. Further, we need to choose two hyperparameters n_{OUT} and n_{IN} instead of fixing the number of base learners B_n. It is not clear what combination will yield optimal performance under the same computational budget this trade-off will likely differ depending on whether we wish to optimize predictive performance or variance estimation.

To address these issues, we design the Balanced Variance Estimation Method (Algorithm 2). In the following, we use h_b to represent $h_{k_n}(\mathcal{S}_b)$ if there is no ambiguity. Let $N_{i,b}$ denote the number of times the i^{th} training sample appears in subsample $\mathcal{S}_b$. In the case of U-statistics where we sample without replacement, $N_{i,b} \in \{0,1\}$. For V-statistics $N_{i,b}$ can be larger than 1 due to sampling with replacement. Summing over b gives $N_i = \sum_{b=1}^{B_n} N_{i,b}$ and the averaged version $\bar{N}_i = \frac{N_i}{B_n}$. For $1 \le i \le n$, define

$$m_i = \sum_{b=1}^{B_n} \omega_{i,b} h_b$$

Algorithm 2 Balanced Variance Estimation Method

```
for $b$ in 1 to $B_n$ do
    Select subsample $\mathcal{S}_b$ of size $k_n$ from training set $\mathcal{D}$ of size $n$.
    Build base learner and evaluate $h_{k_n}(\mathcal{S}_b)$
end for
for $i$ in 1 to $n$ do
    Calculate $m_i$ as the average of $h_{k_n}(\mathcal{S}_b)$ where the $i^{th}$ training sample appears in $\mathcal{S}_b$, weighted by the number of appearance.
end for
Compute the variance of $m_i$ to estimate $\zeta_{1,k_n}$
Compute the variance of all predictions to estimate $\zeta_{k_n,k_n}$
Compute the mean of all predictions to obtain final ensemble prediction
```

where $\omega_{i,b} = \frac{N_{i,b}}{N_i}$. Further define the average of m_i as $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ and the overall average of h_b as $\bar{h} = \frac{1}{B_n}\sum_{b=1}^{B_n} h_b$. The estimates for ζ_{1,k_n} and ζ_{k_n,k_n} can be written as

$$\hat{\zeta}^{\mathrm{BM}}_{1,k_n} = \frac{1}{n-1}\sum_{i=1}^{n}\left(m_i - \bar{m}\right)^2,$$

and

$$\hat{\zeta}^{\mathrm{BM}}_{k_n,k_n} = \frac{1}{B_n-1}\sum_{b=1}^{B_n}\left(h_b - \bar{h}\right)^2.$$

4.3 Infinitesimal Jackknife

The Infinitesimal Jackknife (IJ) was first studied by Jaeckel (1972) as an extension for the jackknife to estimate variance. The basic idea of the jackknife is to omit one observation and recompute the estimate using the remaining samples. Alternatively, if we assign a weight to each observation, omitting one is equivalent to setting the corresponding weight to zero. More generally, we can give each observation a weight slightly less than one every time. IJ is the limiting case as this deficiency in the weight approaches zero. Efron (1982) provided a more detailed treatment of these resampling plans. More recently, IJ was found to be a powerful tool for estimating standard errors in bagging (Efron, 2014). Wager et al. (2014) and Wager and Athey (2018) applied IJ in the context of random forests.

In our setting, the Infinitesimal Jackknife estimate of variance can be expressed as

$$\hat{V}_{\mathrm{IJ}} = \sum_{i=1}^{n} \mathrm{cov}^2\left(N_{i,b}, h_b\right)$$

where $\mathrm{cov}(N_{i,b}, h_b) = \frac{\sum_{b=1}^{B_n}(N_{i,b}-\bar{N}_i)(h_b-\bar{h})}{B_n}$ and $\bar{N}_i$ is defined in Section 4.2. As a general variance estimation method for ensembles, IJ is applied upon the original kernel function (Efron, 2014). Consistency results of IJ for U-statistics typed ensembles were developed in Wager and Athey (2018); Ghosal and Hooker (2020).

IJ does not rely on an explicit expression of the variance term and is targeted at estimating the limiting variance assuming B_n is sufficiently large. By applying IJ, we are essentially estimating $\frac{k_n^2}{n}\zeta_{1,k_n}$ (as in Theorem 10) or $\frac{k_n^2}{n}\zeta^*_{1,k_n}$ (as in Theorem 1). For general V-statistics, it is not practical to use BM on the composite kernel $h^*_{k_n}$ to get an estimate for ζ^*_{1,k_n} directly, and we can instead use IJ on the original kernel h_{k_n} to get variance estimates.

A direct connection exists between BM and IJ, which we will show below.

Definition 2 *Balanced Subsample Structure*

We call a subsample structure balanced if $B_n \times k_n$ is a multiple of n, and each training sample appears exactly $r_n = \frac{B_n \times k_n}{n}$ times.

For U-statistics, this structure implies that each training observation appears in exactly r_n base learners. For V-statistics, each sample is required to occur r_n times but may be used in fewer than r_n base learners since the sampling is done with replacement.

Theorem 3 *If we have balanced subsample structure, the Balanced Variance Estimation Method and the Infinitesimal Jackknife estimator satisfy*

$$\frac{k_n^2}{n}\hat{\zeta}^{BM}_{1,k_n} = \frac{n}{n-1}\hat{V}_{IJ}.$$

Remark 4 *The scaling factor $\frac{n}{n-1}$ is a result of how we calculate the empirical variance. If instead we define $\hat{\zeta}^{BM}_{1,k_n} = \frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2$, then the two estimators are equal: $\frac{k_n^2}{n}\hat{\zeta}^{BM}_{1,k_n} = \hat{V}_{IJ}$. Theorem 3 also enables us to apply BM on the original kernel h_{k_n} to get valid variance estimates in general V-statistics without resorting to the calculation of composite kernel $h^*_{k_n}$, which is infeasible in practice. Further implications of this result are discussed in Appendix I.*

5. Bias Corrections for Variance Estimates

We have so far presented three variance estimation methods (IM, BM and IJ) to estimate the variance of predictions given by an ensemble, being either U-statistics or V-statistics. Although both IM and BM are targeted at the specific variance expression (Equation (10)), IJ is a more general procedure. Applying IM or BM for the general V-statistics (Theorem 1) is infeasible since it involves the evaluation of the complex kernel $h^*_{k_n}$, while IJ can be naturally applied on the original kernel h_{k_n} (Efron, 2014). By the connection of BM and IJ in Theorem 3, BM applying on the original kernel also yields valid estimates. As a result, although we need the new kernel $h^*_{k_n}$ for theoretical analysis, variance estimates can be achieved without explicitly evaluation. In this section, we will simply refer to the variance components of both $U-$ and $V-$statistics as ζ_{1,k_n} and ζ_{k_n,k_n}.

5.1 Bias in Variance Estimation

As briefly mentioned in Section 1, all existing variance estimation methods exhibit severe bias when the number of base learners is not sufficiently large. We now conduct a simple simulation to demonstrate the extent of this bias. Suppose $X \sim 20 \times \text{unif}(0,1)$ and $Y = 2X + \mathcal{N}(0,1)$. An ensemble of decision trees is built to predict Y from X, and we calculate the variance of the prediction at $x = 10$ using IM, BM and IJ. In our simulation, we fix the number of training observations $n = 500$ and kernel size $k_n = 100$. The number of base learners B_n is varied among $100, 1000$ and 10000.

Figure 2 shows the result for both U- and V-statistics. Notice that although larger B_n indicates lower variance (see Equation (5)), the effect of this is negligible as the dominating component of the variance is $\frac{k_n^2}{n}\zeta_{1,k_n}$ in our case (compare Figure 6 with 7 in Appendix C). In order to provide a fair comparison, the variance shown in the figure is for $\frac{k_n^2}{n}\zeta_{1,k_n} + \frac{1}{1000}\zeta_{k_n,k_n}$ for each value of B_n. Thus, different values of B_n only have effect on the estimation for ζ_{1,k_n} and ζ_{k_n,k_n}. Appendix G provides at a closer look at the relationship between two variance components as B_n varies.

We can easily observe that all three methods (IM, BM, IJ) badly overestimate the variance (notice the log scale on y-axis). The bias mainly arises from an overestimation of ζ_{1,k_n} (see Figure 6 and 7 in Appendix C). However, BM and IJ are better than IM since they utilize more information. The plot also corroborates Theorem 3: BM and IJ are exactly the same up to a scaling factor.

In Appendix C, we include additional simulation results on the effect of different kernel size k_n. It is worth noting that the pattern of bias is consistent among V-statistics: an overestimation of ζ_{1,k_n} leads to severe bias which diminishes as the number of base learners increases. This effect exists in U-statistics as well, as the estimated variance decreases as B_n increases. However for U-statistics, the variance estimates tend to underestimate when large kernels are used (Figure 4b and 5b in Appendix C). This is partly caused by the fact that the sampling scheme with U-statistics is not equivalent to sampling from the empirical distribution, especially when the kernel size is large. Both perspectives on variance estimation, either on estimating the variance of conditional expectation or resorting to Infinitesimal Jackknife, are based on the idea of using the empirical distribution of the data to approximate the true underlying distribution. U-statistics, which operate by sampling without replacement, are not equivalent to sampling from empirical distribution, thus resulting in the underestimation phenomenon. In Wager and Athey (2018), the authors use a correction factor $\frac{n(n-1)}{(n-k_n)^2}$ as an empirical adjustment for this effect. Figure 8 in Appendix F shows the result when this correction is applied (denoted by corrected-IJ). Empirically the correction mitigates the underestimation bias, and exhibits a similar pattern as V-statistics: a bias due to overestimation of ζ_{1,k_n}.

5.2 A Bias-corrected Estimator

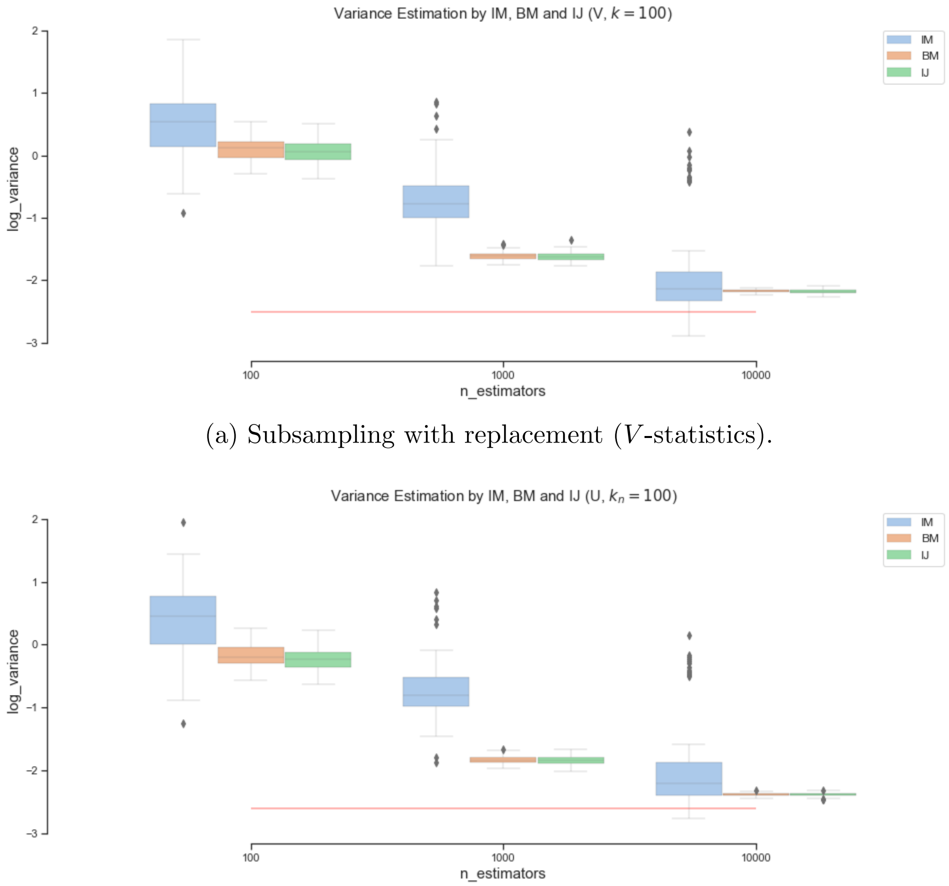
In this section, we present a bias-corrected estimator for ζ_{1,k_n} under the framework of V-statistics. We use an ANOVA-like estimation of variance components similar to Sun et al. (2011). Derivations are collected in Appendix D.

Following notation form Section 4.2, define

$$SS_\tau = \sum_{i=1}^{n} N_i \left(m_i - \bar{m}\right)^2$$

and

$$SS_\epsilon = \sum_{i=1}^{n} \sum_{b=1}^{B_n} N_{i,b} \left(h_b - m_i\right)^2.$$



(a) Subsampling with replacement (V-statistics).

(b) Subsampling without replacement (U-statistics).

Figure 2: Variance estimation by three different methods: Internal Variance Estimation Method (IM), Balanced Variance Estimation Method (BM) and Infinitesimal Jackknife (IJ). The red line denotes true (log) variance obtained by generating data, training the ensemble 100 times and calculating the empirical variance of predictions. For IM, we choose $n_{\text{OUT}} = (10, 20, 50)$ for n_estimators $= (100, 1000, 10000)$ respectively.

A bias-corrected estimate is given by

$$\hat{\zeta}_{1,k_n} = \frac{SS_\tau - (n-1)\,\hat{\sigma}_\epsilon^2}{C - \sum_{i=1}^n N_i^2/C}$$

where $C = \sum_{i=1}^n N_i = B_n k_n$ and $\hat{\sigma}_\epsilon^2 = \frac{SS_\epsilon}{C-n}$.

As a special case for the *Balanced Subsample Structure*, we have $N_1 = N_2 = \ldots = N_i = r_n$, then

$$\hat{\sigma}_\epsilon^2 = \frac{SS_\epsilon}{C-n} = \frac{1}{n\,(r_n - 1)} \sum_{i=1}^n \sum_{b=1}^{B_n} N_{i,b}\,(h_b - m_i)^2$$

and

$$\hat{\zeta}_{1,k_n} = \frac{1}{n-1} \sum_{i=1}^n (m_i - \bar{m})^2 - \frac{1}{r_n}\hat{\sigma}_\epsilon^2. \tag{11}$$

The calculation for $\hat{\zeta}_{1,k_n}$ in (11) may seem complicated at first. In Appendix E, we show that under Balanced Subsample Structure we have $\hat{\zeta}_{1,k_n} \approx \frac{1}{n-1}\sum_{i=1}^n (m_i - \bar{m})^2 - \frac{1}{B_n}\frac{n}{k_n}\hat{\zeta}^{\mathrm{BM}}_{k_n,k_n}$, which is simply the original BM estimator minus a correction term calculated from $\hat{\zeta}^{\mathrm{BM}}_{k_n,k_n}$. This indicates that one can actually calculate the bias-corrected estimator without any extra computational effort. In this case, the estimate for the limiting variance $\frac{k_n^2}{n}\zeta_{1,k_n} + \frac{1}{B_n}\zeta_{k_n,k_n}$ is $\frac{k_n^2}{n(n-1)}\sum_{i=1}^n (m_i - \bar{m})^2 - \frac{k_n - 1}{B_n}\hat{\zeta}^{\mathrm{BM}}_{k_n,k_n}$, where it's clear that the incomplete part is negligible relative to the bias correction term we need to apply.

Figure 3 shows simulation results for this bias-corrected estimator (we call it corrected-V in the remaining part of this paper) compared with BM and IJ under the framework of V-statistics. Here we no longer display IM since it is systematically worse. We can see that even with only 100 base learners, the bias-corrected estimator achieves relatively accurate estimation of the variance. The bias-corrected term may introduce some instability when B_n is very small, but for a moderate size B_n it has much lower bias compared to BM and IJ.

It is worth pointing out that this bias correction method does not work for U-statistics, for the same reason mentioned before: the sampling schema is not equivalent to sampling from empirical distribution. Figure 9 in Appendix F shows that the bias-corrected estimator over-corrects the variance for U-statistics. In Athey et al. (2019), the authors developed a method called the bootstrap of little bags to estimate variance based on the work of Sexton and Laake (2009). They also encountered the challenge of negative variance when B_n is small. In their software, an improper uniform prior over $[0, \infty)$ was employed to help mitigate this issue. We conjecture that the phenomenon also stems from the mechanism of sampling without replacement. In Appendix F, we present an empirically accurate correction to U-statistics as well.

We briefly discuss computational costs to end this section. The majority of computational efforts are spent at building the ensembles, which is scaled with the total number of base learners B_n. The time needed for variance calculation using either IM, BM or IJ along with the bias correction are only marginal. That being said, in many cases to get accurate variance estimates a prohibitively large B_n is needed. The bias correction method we developed in this section can reduce B_n to a moderate size, and thus improves computational efficiency for statistical inference.

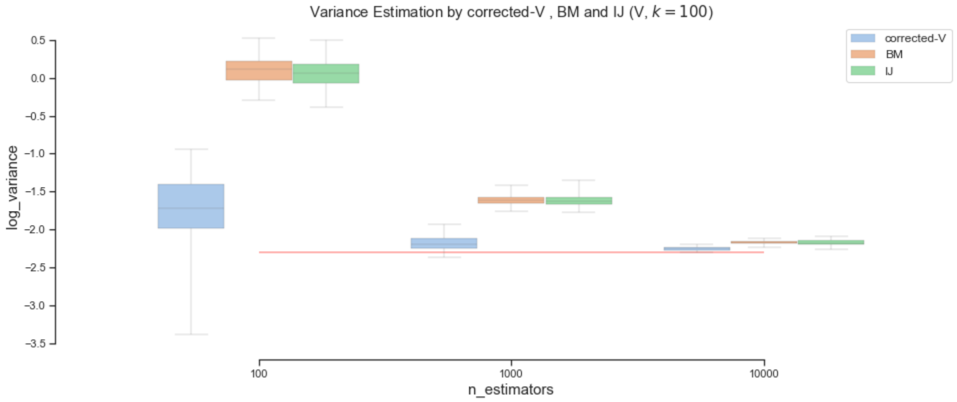


Figure 3: Variance estimation by three different methods: corrected-V, BM and IJ. The red line denotes true (log) variance obtained by generating data, training the ensemble 100 times and calculating the empirical variance of predictions.

6. Randomized Ensembles

As briefly mentioned before, randomized ensembles are widely used in practice. A general principle to achieve good performance in ensembles is to make individual learners both accurate and diverse (Zhou, 2012). To increase diversity, randomization is added to each base learner. For example in random forests (Breiman, 2001), each split is chosen from a randomly selected subset of all possible features.

Similar to Peng et al. (2019), we define the notion of a *generalized* complete V-statistic

$$V_{n,k_n,\omega} = n^{-k_n} \sum_{i_1=1}^{n} \dots \sum_{i_{k_n}=1}^{n} h_{k_n}\left(Z_{i_1}, \dots, Z_{i_{k_n}}; \omega\right). \tag{12}$$

Note that for each kernel h_{k_n} we consider an i.i.d. sample of random ω_i but the subscript is dropped for notational convenience.

Similarly define the generalized incomplete statistic by

$$V_{n,k_n,B_n,\omega} = \frac{1}{B_n} \sum_{i} h_{k_n}\left(Z_{i_1}, \dots, Z_{i_{k_n}}; \omega\right). \tag{13}$$

Following the same idea developed in Mentch and Hooker (2016) and Wager and Athey (2018), consider the expected version of (12)

$$V^*_{n,k_n,\omega} = \mathbb{E}_\omega V_{n,k_n,\omega} = n^{-k_n} \sum_{i_1=1}^{n} \dots \sum_{i_{k_n}=1}^{n} E_\omega h_{k_n}\left(Z_{i_1}, \dots, Z_{i_{k_n}}; \omega\right) \tag{14}$$

where the expectation is taken over the randomization parameter ω. In this case, $V^*_{n,k_n,\omega}$ can be viewed as a non-randomized V-statistic with kernel $h^E_{k_n} = \mathbb{E}_\omega h_{k_n}$ where Theorem 1 applies. We state this result formally in the following corollary.

Corollary 5 *Let* $Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} F_Z$ *and let* $V_{n,k_n,\omega}$ *be a generalized complete* V*-statistic defined in (12) and the corresponding expected version* $V^*_{n,k_n,\omega}$ *in (14). Under the same conditions as Theorem 1, we have*

$$\frac{\left(V^*_{n,k_n,\omega} - \theta^*_{k_n}\right)}{\sqrt{\frac{k_n^2}{n}\zeta^*_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

where all parameters $\theta^*_{k_n}, \zeta^*_{1,k_n}, \zeta^*_{k_n,k_n}$ *are defined using new non-randomized kernel* $h^E_{k_n}$ *instead of* h_{k_n}*.*

Given this, in order to retain the asymptotic normality of the corresponding randomized case (13), there are two steps: first we show that $\frac{V_{n,k_n,\omega} - V^*_{n,k_n,\omega}}{\mathrm{Var}(V^*_{n,k_n,\omega})} \xrightarrow{P} 0$ and thus $V_{n,k_n,\omega}$ has the same asymptotic distribution as $V^*_{n,k_n,\omega}$. Then the asymptotics of $V_{n,k_n,B_n,\omega}$ can be derived from that of $V_{n,k_n,\omega}$.

Theorem 6 *Let* $V_{n,k_n,B_n,\omega}$ *be a generalized incomplete* V*-statistic of the form defined in (13). Further assume the corresponding statistic* $V^*_{n,k_n,\omega}$ *in (14) satisfies Corollary 5 and* $\lim_{n\to\infty} k_n^2 \zeta^*_{1,k_n} > 0$*. Then as long as*

$$\sup_{k_n} \mathbb{E}\left(h_{k_n}\left(Z_{i_1}, \ldots, Z_{i_{k_n}}; \omega\right) - \mathbb{E}_\omega h_{k_n}\left(Z_{i_1}, \ldots, Z_{i_{k_n}}; \omega\right)\right) < \infty$$

where $(i_1, \ldots, i_{k_n})$ *are chosen from* $\{1, \ldots, k_n\}$ *with replacement, we have*

$$\frac{\left(V_{n,k_n,B_n,\omega} - \theta^*_{k_n}\right)}{\sqrt{\frac{k_n^2}{n}\zeta^*_{1,k_n} + \frac{1}{B_n}\zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Here, $\zeta_{k_n,k_n} = var(h_{k_n}(Z_1, \ldots, Z_{k_n}, \omega))$ *is the variance across individual randomized kernels, and all parameters* $\theta^*_{k_n}, \zeta^*_{1,k_n}, \zeta^*_{k_n,k_n}$ *are defined using new kernel* $h^E_{k_n}$ *instead of* h_{k_n} *as in Corollary 5.*

The analysis for randomized ensembles in Peng et al. (2019) directly treats the randomized kernel, rather than first establishing a result for the expectation over ω. This is easier in the framework of U-statistics; in the case of V-statistics, the new kernel $h^*_{k_n}$ constructed as in Section 3.1 no longer has independent randomization parameters, since different $h^*_{k_n}$ might share the same kernel h_{k_n}.

The condition $\lim_{n\to\infty} k_n^2 \zeta^*_{1,k_n} > 0$ required here has also appeared in Lemma 4.1 of Song et al. (2019). We believe it is generally satisfied with many base learners including trees, see Peng et al. (2019) for an in-depth analysis for the behavior of ζ_{1,k_n}.

7. Empirical Studies

Here, we conduct two suites of experiments. All simulations are implemented in Python. For building random forests, we apply *RandomForestClassifier* and *RandomForestRegressor* from *scikit-learn* Friedman (1991). Unless otherwise noted, default parameter values are used.

7.1 Predictive Performance

In this section, we evaluate the predictive performance for different sampling strategies. In particular, we focus on the scenario of U-statistics (sample without replacement) and V-statistics (sample with replacement), and varying subsample size (proportion of the size of training data to be $0.2, 0.4, 0.6, 0.8, 1.0$). We address the following two questions empirically:

1. Should we subsample with or without replacement in terms of prediction performance?

2. What is the best subsample size?

There are six datasets taken from UCI Machine learning Repository (see Appendix K for details) and we also include a regression function (denoted by MARS) which was initially considered by Friedman (1991) for multivariate adaptive regression splines, and has since been used as a benchmark in many random forests publications.

Each model is built using 100 trees and to full depth until a leaf is pure or contain fewer than 2 data points. 20% of samples are left as test set. For classification, $\sqrt{p}$ of features are considered when searching for best splits, and $\lfloor p/3 \rfloor$ for regression. Table 7.1 summarizes our results. The first three datasets are regression tasks for which we report root-mean-squared error and the remaining four are classification with accuracy given by correct classification percentage. We repeat the process 20 times and denote standard error in parenthesis. Top performance entries are marked in bold separately for sampling with vs. without replacement and an asterisk indicates the best performance across all scenarios.

We make two observations here. First, for both sampling with replacement or without, there is a best subsample size for prediction, though the proportion varies across different datasets. Accuracy decreases as the subsample size moves away from the ideal proportion. It is worth noting that performance discrepancy, which is defined to be the maximum performance difference across five subsample size settings, is generally larger in the case of U-statistics than V-statistics. For example, in diabetes dataset, there is a 4.1357 RMSE difference in U-statistics scenario compared to 1.03 in V-statistics. Similarly for retinopathy dataset, the accuracy discrepancy is 2.79% versus 1.26%. It may suggest that sampling with replacement is more robust to changes in subsample sizes; possibly as a result of combining trees built on different numbers of unique data points.

On the other hand, we did not see an obvious performance gap between two sampling techniques. The best result can be generated either by sampling with replacement or without depending on the specific data at hand. In practice, one will need to use cross validation to choose the best sampling strategy and subsample size, if predictive performance is the primary concern.

Dataset	n	p	U-statistics (sample without replacement)					V-statistics (sample with replacement)				
			0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
boston	506	13	3.9696 (0.1112)	3.5754 (0. 0788)	3.3225 (0.1131)	3.2103 (0.1050)	**3.0588*** (0.0787)	4.0358 (0.0656)	3.6799 (0.0758)	3.4705 (0.0936)	3.3738 (0.1057)	**3.2944** (0.0887)
diabetes	442	10	**54.6137** (0.6936)	54.6289 (0.5164)	55.5140 (0.6097)	56.9017 (0.5681)	58.7494 (0.8187)	**54.3830*** (0.7100)	54.4282 (0.6347)	54.5220 (0.7739)	55.0782 (0.5642)	55.4130 (0.8877)
MARS	500	5	3.2106 (0.0648)	2.8850 (0.0980)	2.7211 (0.0845)	2.6329 (0.0482)	**2.5465*** (0.0484)	3.2908 (0.1059)	2.9706 (0.0644)	2.8201 (0.0683)	2.7498 (0.0591)	**2.6965** (0.0425)
iris	150	4	**0.9667*** (0.0000)	**0.9667*** (0.0000)	0.9633 (0.0100)	0.9500 (0.0167)	0.9333 (0.0000)	0.9650 (0.0073)	**0.9667*** (0.0000)	0.9617 (0.0119)	**0.9667*** (0.0000)	0.9633 (0.0100)
digits	1797	64	0.9507 (0.0053)	0.9660 (0.0034)	0.9688 (0.0043)	**0.9735*** (0.0048)	0.9731 (0.0045)	0.9481 (0.0054)	0.9588 (0.0058)	0.9651 (0.0049)	0.9681 (0.0048)	**0.9689** (0.0036)
retinopathy	1151	19	**0.6935** (0.0156)	0.6825 (0.0112)	0.6766 (0.0114)	0.6708 (0.0082)	0.6656 (0.0125)	0.6892 (0.0135)	**0.6946*** (0.0126)	0.6868 (0.0125)	0.6846 (0.0109)	0.6820 (0.0117)
breast_cancer	569	30	**0.9772*** (0.0070)	0.9746 (0.0038)	0.9750 (0.0064)	0.9741 (0.0019)	0.9745 (0.0026)	0.9741 (0.0071)	0.9737 (0.0039)	0.9746 (0.0038)	**0.9759** (0.0047)	0.9754 (0.0035)

Table 1: Predictive performance on seven datasets under different sampling strategies.

7.2 Asymptotic Normality and Variance Estimation

In this section, we illustrate empirically the asymptotic normality property and variance estimation algorithms for V-statistics. We will first utilize the MARS function (Friedman, 1991; Mentch and Hooker, 2016) such that we have access to the underlying data generating distribution: $\mathrm{y} = \mathrm{f(x)} = 10\sin(\pi \mathrm{x}_1\mathrm{x}_2) + 20(\mathrm{x}_3 - 0.05)^2 + 10\mathrm{x}_4 + 5\mathrm{x}_5 + \epsilon$, where $\mathcal{X} \sim U([0,1]^5)$ and $\epsilon \sim \mathcal{N}(0,1)$.

Our simulation runs for 500 iterations. In each iteration we generate $n = 500$ training observations and train random forests with subsample size $k = 100, 250, 500$ and the number of trees $B = 500, 1000, 2500, 5000$. We make evaluation on three test points: $p_1 = [0.5, 0.5, 0.5, 0.5, 0.5]$ and p_2, p_3 are randomly drawn from $U([0,1]^5)$.

For each k and B, let $\hat{f}_{i,j}$ denote the prediction at the ith data point and jth iteration ($i = 1, 2, 3$ and $j = 1, \ldots, 500$). Similarly $\hat{V}_{u,ij}$ and $\hat{V}_{c,ij}$ are the variance estimates for IJ and corrected-V respectively. The following three metrics are reported:

- Normality: We test the normality of predictions $\hat{f}_{i,j}$ for $j = 1, 2, \ldots 500$ based on D'Agostino (1971) and D'Agostino and Pearson (1973) which combine skewness and kurtosis, and is implemented by *scipy.stats.normaltest*[3] in Python. In Table 2, we report test statistics and corresponding p-values (in parenthesis). We can see from the p-values reported that normality for predictions generally hold, even for large subsample size. See Appendix H for a larger scale of experiments on asymptotic normality for ensembles.

- Variance ratio: The estimated variance for each setting is given by $\bar{\hat{V}}_{u,i} = \frac{1}{500}\sum_{t=1}^{500} \hat{V}_{u,ij}$ and $\bar{\hat{V}}_{c,i} = \frac{1}{500}\sum_{t=1}^{500} \hat{V}_{c,ij}$. And true variance $V(\hat{f}_i)$ is approximated by the empirical variance of $\hat{f}_{i,j}$ for $j = 1, 2, \ldots 500$. Note that in practice we cannot calculate the true asymptotic variance, but the between-simulation variance can serve as a good approximation. The variance ratio is defined as $\frac{\bar{\hat{V}}_{u,i}}{V(\hat{f}_i)}$ and $\frac{\bar{\hat{V}}_{c,i}}{V(\hat{f}_i)}$, where a value close to 1 is ideal. We can see similar patterns across three tables. The original version of IJ produces highly biased variance estimates, where the bias diminishes as the number of trees B becomes larger. The bias-corrected version successfully alleviates the issue. For $k = 100$, it starts to produce reasonable estimates for 1000 trees, and the variance ratios are close to one for larger B values. We can also see that it becomes harder to estimate the variance as the subsample sizes grow.

- Coverage probability: constructing 95% confidence intervals by $\hat{f}_{i,j} \pm 1.96 \times \sqrt{\hat{V}_{u,ij}}$ or $\hat{f}_{i,j} \pm 1.96 \times \sqrt{\hat{V}_{c,ij}}$ for $j = 1, 2, \ldots 500$ in each setting and we can calculate a coverage probability by checking whether the expected prediction value (approximated by $\bar{\hat{f}}_i = \frac{1}{500}\sum_{t=1}^{500} \hat{f}_{i,j}$) falls into this interval. (Note that this does not assess coverage of an underlying $\theta^*_{k_n} = \mathbb{E}h^*_{k_n}$.) This is strongly related to our results for variance ratios. A larger variance ratios will produce conservative intervals, thus generating higher coverage probability. The bias-corrected algorithm produces coverage probability close to 0.95 with reasonable number of base learners.

[3] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html`

		p_1		p_2		p_3	
		original	corrected	original	corrected	original	corrected
B = 500	normality	3.5285 (0.1713)		0.2274 (0.8925)		0.6600 (0.7189)	
	var_ratio	7.6985	1.4962	8.3871	1.5648	5.7719	1.2725
	coverage	100.0	97.2	100.0	96.4	100.0	95.2
B = 1000	normality	1.2341 (0.5395)		1.1715 (0.5567)		1.2146 (0.5448)	
	var_ratio	4.9540	1.3839	4.9290	1.3116	3.8187	1.2048
	coverage	100.0	96.4	100.0	96.6	99.8	96.0
B = 2500	normality	1.9708 (0.3733)		2.8710 (0.2380)		0.1502 (0.9276)	
	var_ratio	2.3068	1.0328	2.5635	1.1058	2.1548	1.0770
	coverage	99.2	94.2	99.8	94.6	99.4	94.4
B = 5000	normality	3.2266 (0.1992)		0.8750 (0.6456)		2.7195 (0.2567)	
	var_ratio	1.7073	1.0305	1.7748	1.0460	1.5446	1.0138
	coverage	98.0	95.0	99.2	94.8	97.3	93.0

(a) k = 100

		p_1		p_2		p_3	
		original	corrected	original	corrected	original	corrected
B = 500	normality	0.0309 (0.9846)		2.9390 (0.2300)		1.1261 (0.5695)	
	var_ratio	9.2069	2.6342	9.1624	2.5651	7.0298	2.1752
	coverage	100.0	99.2	100.0	99.0	100.0	98.6
B = 1000	normality	0.6423 (0.7253)		0.4018 (0.8180)		1.1071 (0.5749)	
	var_ratio	5.7815	2.0194	5.1873	1.7729	4.6717	1.7826
	coverage	100.0	99.4	100.0	98.4	100.0	97.8
B = 2500	normality	8.7518 (0.0126)		3.4691 (0.1765)		0.9235 (0.6302)	
	var_ratio	2.6779	1.3200	2.5242	1.2031	2.1364	1.1433
	coverage	99.8	95.0	99.4	95.4	98.8	94.8
B = 5000	normality	2.0317 (0.3621)		2.2734 (0.3209)		0.7299 (0.6942)	
	var_ratio	1.8688	1.1504	1.8303	1.1103	1.5598	1.0553
	coverage	98.8	95.0	99.0	94.2	96.8	92.0

(b) k = 250

		p_1		p_2		p_3	
		original	corrected	original	corrected	original	corrected
B = 500	normality	0.5147 (0.7731)		2.1454 (0.3421)		2.3230 (0.3130)	
	var_ratio	10.5311	4.4362	11.3409	4.7405	7.3644	3.1930
	coverage	100.0	99.8	100.0	100.0	100.0	99.4
B = 1000	normality	8.0832 (0.0176)		0.5806 (0.7481)		0.3763 (0.8285)	
	var_ratio	6.2113	2.8874	6.4252	2.9916	4.5268	2.2424
	coverage	100.0	99.8	100.0	99.4	100.0	99.4
B = 2500	normality	1.0393 (0.5947)		2.6963 (0.2579)		0.7730 (0.6794)	
	var_ratio	2.8223	1.5995	3.3763	1.9137	2.4909	1.5236
	coverage	99.8	96.2	99.4	97.8	99.0	96.4
B = 5000	normality	2.5268 (0.2827)		1.2847 (0.5361)		0.0623 (0.9693)	
	var_ratio	2.0838	1.405	1.9944	1.3492	1.6583	1.2001
	coverage	98.4	94.8	98.2	95.8	96.4	93.6

(c) k = 500

Table 2: Asymptotic normality and variance estimation results for MARS function.

		$k = 100$		$k = 250$		$k = 500$	
		original	corrected	original	corrected	original	corrected
B = 500	normality	0.10		0.00		0.15	
	var_ratio	12.34(3.64)	1.41(0.36)	13.81(3.41)	2.32(0.53)	20.73(8.75)	5.08(1.98)
	coverage	100.0(0.0)	94.33(0.04)	100.0(0.0)	98.3(0.02)	100.0(0.0)	99.7(0.01)
B = 1000	normality	0.15		0.00		0.05	
	var_ratio	7.12(2.07)	1.25(0.34)	7.62(2.10)	1.63(0.39)	11.07(4.87)	3.08(1.23)
	coverage	100.0(0.0)	95.0(0.05)	100.0(0.0)	96.9(0.03)	100.0(0.0)	98.7(0.02))
B = 2500	normality	0.10		0.05		0.05	
	var_ratio	3.49(0.93)	1.10(0.28)	3.67(0.90)	1.21(0.25)	5.00(2.10)	1.78(0.06)
	coverage	100.0(0.0)	94.7(0.05)	100.0(0.0)	94.1(0.04)	99.6(0.02)	96.6(0.03)
B = 5000	normality	0.10		0.00		0.05	
	var_ratio	2.32(0.63)	1.08(0.28)	2.31(0.54)	1.06(0.22)	2.93(1.08)	1.34(0.39)
	coverage	99.2(0.02)	94.0(0.05)	99.3(0.01)	91.7(0.05)	98.7(0.03)	93.3(0.02)

Table 3: Asymptotic normality and variance estimation results for Protein Tertiary Structure across 20 test samples.

In order to see how our results work in real world settings, we pick a relatively large scale dataset: Physicochemical Properties of Protein Tertiary Structure Data Set[4]. The data set contains 45730 samples with 9 covariates and the target is the size of the residue.

To simulate the situation where one can attain alternative training data drawn from the same data generating distribution to quantify sampling uncertainty, we randomly select 45000 samples and partition them to 45 sets with 1000 in each. These 45 sample sets act as independent draws from the unknown data generating distribution. As in the previous experiment, we run for 45 iterations with subsample size $k = 100, 250, 500$ and the number of trees $B = 500, 1000, 2500, 5000$. We treat the average of the resulting 45 random forests as the target for inference.

We evaluate normality, variance ratio and coverage probability across 20 randomly selected test points. In Table 3 we report the average rejection percentage for the normality test, average variance ratio and average coverage probability. Numbers in the parentheses denote standard deviations. We observe similar patterns as in the MARS setting.

8. Conclusion

In this paper, we present a framework for analyzing the asymptotics of V-statistics where the kernel size k_n grows with the number of samples n. It is shown that a central limit theorem can be established similar to the work in Mentch and Hooker (2016), Wager and Athey (2018) and Peng et al. (2019), which focus on the case of U-statistics. The result brings new insight into the analysis of ensemble methods.

We also provide unified treatment of variance estimation in both U- and V-statistics. We observe that existing methods for estimating the limiting variance exhibit severe bias and would require a prohibitively large number of base learners to achieve accurate results, hindering any practical applications such as constructing confidence intervals or conducting

[4] `https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure`

hypothesis tests. To this end, we propose a new method called Balanced Variance Estimation Method (BM), and carefully analyze its connection to other methods. In particular, we demonstrate an equivalence between BM and Infinitesimal Jackknife. Additionally, a bias correction method is developed which is shown to produce more accurate variance estimation with a moderate size of base learners.

Practically, we would suggest sampling with replacement in building ensembles since the bias correction for V-statistics is theoretically sound and much less involved. What's more, it appears that the asymptotic normality for V-statistics holds across a broarder spectrum compared to its U-statistics counterpart (Appendix H). We speculate that the relative insensitivity of V-statistics performance to subsample size results from averaging over trees with different numbers of unique samples. More generally, we might speculate that our equivalent kernel representation in fact admits an approximation in terms of U-statistics of somewhat lower orders. For example, the weight given to the value of the kernel with k_n unique data points is asymptotically negligible. This analysis may lead to better understanding of these properties and the rates at which V-statistics subsamples can be allowed to grow.

From another theoretical point of view, the analysis we provide here is essentially a reduction to U-statistics. We will further explore whether other approaches like Taylor expansion using differential methods Serfling (2009) could be applied to attain similar results. It would also be valuable to see if the results presented in this paper could be extended to high dimensional cases.

Acknowledgments

All three authors were partially supported by NSF grant DMS-1712554. LM was partially funded by NSF grant DMS-1712041.

Appendix A. Equivalence between V- and U-statistics

We first show that the asymptotic behavior of V_{n,k_n} is the same as that of U_{n,k_n}, provided $k_n = o(n^{\frac{1}{4}})$. The following important lemma relates V_{n,k_n} to a family of U-statistics, which is a simple extension from Theorem 1 in (Lee, 1990, p.183) to the case where the kernel size k_n is changing with n.

Lemma 7 *Let V_{n,k_n} be a complete, infinite order V-statistic based on a permutation symmetric kernel h_{k_n} of degree k_n as defined in (8). Then we may write*

$$V_{n,k_n} = n^{-k_n} \sum_{j=1}^{k_n} j! S_{k_n}^{(j)} \binom{n}{j} U_n^{(j)}$$

where $U_n^{(j)}$ is a U-statistic of degree j. The kernel $\phi_{(j)}$ of $U_n^{(j)}$ is given by

$$\phi_{(j)}\left(z_1, \ldots, z_j\right) = \left(j! S_{k_n}^{(j)}\right)^{-1} \sum\nolimits_{(j)}^{*} h_{k_n}\left(z_{i_1}, \ldots, z_{i_{k_n}}\right)$$

where the sum $\sum_{(j)}^{*}$ is taken over all k_n-tuples $(i_1, \ldots, i_{k_n})$ formed from $\{1, 2, \ldots, j\}$ having exactly j indices distinct, and the quantities $S_{k_n}^{(j)}$ are Stirling numbers of the second kind Rennie and Dobson (1969).

Intuitively, as n grows, if k_n grows slowly enough, V_{n,k_n} should behave like U_{n,k_n}, as the difference brought by sampling *with* or *without* replacement becomes negligible. Theorem 8 extends a result in Shieh (1994) which makes this argument rigorous.

Theorem 8 *Suppose $h_{k_n} \in \mathcal{H}$, $k_n = o(n^{\frac{1}{4}})$ and $\lim_{n\to\infty} Var(\sqrt{n}U_{n,k_n}) > 0$. Then V_{n,k_n} and U_{n,k_n} have the same asymptotic distribution.*

Remark 9 *This theorem only states that asymptotically V_{n,k_n} and U_{n,k_n} are indistinguishable. The assumption that $\lim_{n\to\infty} Var(\sqrt{n}U_{n,k_n}) > 0$ simply indicates that the rate of convergence for U_{n,k_n} is $\sqrt{n}$. Theorem 8 may possibly hold under other regimes, such as with degenerate kernels, where the convergence rate is not $\sqrt{n}$, but this is out of the scope of this paper.*

As in Equation (3), by averaging only $B_n < n^{k_n}$ set of subsamples we have an *incomplete, infinite order* V-statistic

$$V_{n,k_n,B_n} = \frac{1}{B_n} \sum_{i} h_{k_n}\left(Z_{i_1}, \ldots, Z_{i_{k_n}}\right) \tag{15}$$

where $\{Z_{i_1}, \ldots, Z_{i_k}\}$ is again drawn *with* replacement from $\{Z_1, \ldots, Z_n\}$. Under some regularity conditions, similar asymptotic results as Theorem 1 in Mentch and Hooker (2016); Theorem 1 in Peng et al. (2019) can be shown.

Theorem 10 *Let* $Z_1, Z_2, \dots, Z_n \overset{iid}{\sim} F_Z$ *and let* V_{n,k_n,B_n} *be an incomplete, infinite order* V*-statistic with kernel* h_{k_n}*. Let* $\theta_{k_n} = \mathbb{E}h_{k_n}(Z_1, \dots, Z_{k_n})$ *such that* $h_{k_n} \in \mathcal{H}$*. Then under the assumptions that* $k_n = o(n^{\frac{1}{4}})$*,* $\lim_{n\to\infty} k_n^2 \zeta_{1,k_n} > 0$ *and* $\lim_{n\to\infty} \frac{\zeta_{k_n,k_n}}{n\zeta_{1,k_n}} \to 0$*, we have*

$$\frac{(V_{n,k_n,B_n} - \theta_{k_n})}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n} + \frac{1}{B_n}\zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

In the complete case where $B_n = n^{k_n}$*, we have*

$$\frac{(V_{n,k_n} - \theta_{k_n})}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Note that the first two assumptions $k_n = o(n^{\frac{1}{4}})$ and $\lim_{n\to\infty} k_n^2 \zeta_{1,k_n} > 0$ ensure that we can apply Theorem 8. The proof requires an additional lemma and is collected together in Appendix B.2.

Appendix B. Proofs

B.1 Proof of Theorem 8

Proof By Slutsky's theorem, we only need to show $\frac{(V_{n,k_n} - U_{n,k_n})}{\sqrt{\mathrm{Var}(U_{n,k_n})}} \xrightarrow{p} 0$. Since we assume $\lim_{n\to\infty} \mathrm{Var}(\sqrt{n}U_{n,k_n}) > 0$, it suffices to prove $\sqrt{n}(V_{n,k_n} - U_{n,k_n}) \xrightarrow{p} 0$. We seek to prove L^1 convergence, which implies convergence in probability. According to Lemma 7, V_{n,k_n} could be written as

$$\begin{aligned} V_{n,k_n} &= \sum_{j=1}^{k_n} \frac{j! S_{k_n}^{(j)} \binom{n}{j}}{n^{k_n}} U_n^{(j)} \\ &= \frac{k_n! S_{k_n}^{(k_n)} \binom{n}{k_n}}{n^{k_n}} U_{n,k_n} + \sum_{j=1}^{k_n-1} \frac{j! S_{k_n}^{(j)} \binom{n}{j}}{n^{k_n}} U_n^{(j)} \\ &= \frac{n(n-1)\dots(n-k_n+1)}{n^{k_n}} U_{n,k_n} + \sum_{j=1}^{k_n-1} \frac{j! S_{k_n}^{(j)} \binom{n}{j}}{n^{k_n}} U_n^{(j)}. \end{aligned}$$

Since we assume the second moment of kernel h is bounded, $U_n^{(j)}$ could also be bounded by a constant $C < \infty$. We have

$$\begin{aligned} \mathbb{E}\left|\sqrt{n}\left(V_{n,k_n} - U_{n,k_n}\right)\right| &= \sqrt{n}\mathbb{E}\left|\left(\frac{n(n-1)\dots(n-k_n+1)}{n^{k_n}} - 1\right)U_{n,k_n} + \sum_{j=1}^{k_n-1} \frac{j! S_{k_n}^{(j)} \binom{n}{j}}{n^{k_n}} U_n^{(j)}\right| \\ &\le \sqrt{n}\mathbb{E}\left|\left(\frac{n(n-1)\dots(n-k_n+1)}{n^{k_n}} - 1\right)U_{n,k_n}\right| + \mathbb{E}\left|\sum_{j=1}^{k_n-1} \frac{j! S_{k_n}^{(j)} \binom{n}{j}}{n^{k_n}} U_n^{(j)}\right| \\ &\le C\left(\left|\sqrt{n}\left(\frac{n(n-1)\dots(n-k_n+1)}{n^{k_n}} - 1\right)\right| + \sqrt{n}\sum_{j=1}^{k_n-1} \frac{j! S_{k_n}^{(j)} \binom{n}{j}}{n^{k_n}}\right). \end{aligned}$$

First it's easy to see

$$
\begin{aligned}
&\sqrt{n}\left(\frac{n(n-1)\dots(n-k_n+1)}{n^{k_n}}-1\right)\\
\sim&\sqrt{n}\left(1-\frac{\sum_{i=1}^{k_n-1}}{n}-1\right)\\
\sim&\frac{k_n^2}{\sqrt{n}}\to 0
\end{aligned}
$$

as $n\to\infty$ when $k_n=o(n^{\frac{1}{4}})$. An upper bound for $S_{k_n}^{(j)}$ is provided in Rennie and Dobson (1969)

$$S_{k_n}^{(j)}\leq\frac{1}{2}\binom{k_n}{j}j^{k_n-j}.$$

Thus,

$$
\begin{aligned}
\frac{j!S_{k_n}^{(j)}\binom{n}{j}}{n^{k_n}}&=\frac{j!\binom{n}{j}}{n^j}\frac{S_{k_n}^{(j)}}{n^{k_n-j}}\\
&\leq\frac{1}{2}\frac{j!\binom{n}{j}}{n^j}\frac{\binom{k_n}{j}j^{k_n-j}}{n^{k_n-j}}\\
&\leq\frac{1}{2}\frac{j!\binom{n}{j}}{n^j}\frac{k_n^{k_n-j}k_n^{k_n-j}}{n^{k_n-j}}\\
&=\frac{1}{2}\frac{j!\binom{n}{j}}{n^j}\left(\frac{k_n^2}{n}\right)^{k_n-j}\\
&\leq\left(\frac{k_n^2}{n}\right)^{k_n-j}.
\end{aligned}
$$

Let $a_n=\frac{k_n^2}{\sqrt{n}}$, and we know $a_n\to 0$. Taking the sum yields

$$
\begin{aligned}
\sqrt{n}\sum_{j=1}^{k_n-1}\frac{j!S_{k_n}^{(j)}\binom{n}{j}}{n^{k_n}}&\leq\sqrt{n}\sum_{j=1}^{k_n-1}\left(\frac{k_n^2}{n}\right)^{k_n-j}\\
&\leq\sum_{j=1}^{k_n-1}\left(\frac{k_n^2}{\sqrt{n}}\right)^{k_n-j}\\
&=\sum_{j=1}^{k_n-1}a_n^{k_n-j}\\
&\leq\frac{a_n}{1-a_n}\\
&\to 0.
\end{aligned}
$$

We could conclude that $\mathbb{E}|\sqrt{n}(V_{n,k_n}-U_{n,k_n})|\to 0$. ■

B.2 Proof of Theorem 10

Since $k_n = o(n^{\frac{1}{4}})$ and $\lim_{n\to\infty} k_n^2 \zeta_{1,k_n} > 0$, the complete case follows directly from Theorem 8 and Theorem 1 in Peng et al. (2019). We will need the following lemma for the incomplete case.

Lemma 11 *Let $a_1, a_2, \dots$ be a sequence of constants such that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n a_i = 0$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n a_i^2 = \sigma^2$ and let random variables $M_1, \dots, M_n$ have a multinomial distribution, multinomial$(B_n; \frac{1}{n}, \dots, \frac{1}{n})$. Then as $B_n, n \to \infty$, the limiting distribution of*

$$B_n^{-\frac{1}{2}} \sum_{i=1}^n a_i \left(M_i - \frac{B_n}{n}\right)$$

is $\mathcal{N}(0, \sigma^2)$.

Proof The characteristic function of $(M_1, \dots, M_n)$ is $(\frac{1}{n}e^{it_1} + \dots + \frac{1}{n}e^{it_n})^{B_n}$ since it's multinomial$(B_n; \frac{1}{n}, \dots, \frac{1}{n})$. Thus the characteristic function of $B_n^{-\frac{1}{2}}\sum_{i=1}^n a_i \left(M_i - \frac{B_n}{n}\right)$ is given by

$$\begin{aligned}
\mathbb{E}\left(e^{itB_n^{-\frac{1}{2}}\sum_{i=1}^n a_i\left(M_i - \frac{B_n}{n}\right)}\right) &= e^{-itB_n^{-\frac{1}{2}}\frac{B_n}{n}\sum_{i=1}^n a_i}\mathbb{E}\left(e^{itB_n^{-\frac{1}{2}}\sum_{i=1}^n a_i M_i}\right) \\
&= e^{-it\bar{a}_n B_n^{\frac{1}{2}}}\left(\frac{1}{n}e^{ita_1 B_n^{-\frac{1}{2}}} + \dots + \frac{1}{n}e^{ita_n B_n^{-\frac{1}{2}}}\right)^{B_n} \\
&= e^{-it\bar{a}_n B_n^{\frac{1}{2}}}\left(\frac{1}{n}\left(n + itB_n^{-\frac{1}{2}}\sum_{i=1}^n a_i + \frac{1}{2}\left(itB_n^{-\frac{1}{2}}\right)^2\sum_{i=1}^n a_i^2 + \dots\right)\right)^{B_n} \\
&= e^{-it\bar{a}_n B_n^{\frac{1}{2}}}\left(1 + itB_n^{-\frac{1}{2}}\bar{a}_n + \frac{1}{2}\sigma_n^2\left(itB_n^{-\frac{1}{2}}\right)^2 + o\left(B_n^{-1}\right)\right)^{B_n}
\end{aligned}$$

where $\bar{a}_n = \frac{1}{n}\sum_{i=1}^n a_i$ and $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n a_i^2$. Taking the logarithm gives

$$\begin{aligned}
\log\mathbb{E}\left(e^{itB_n^{-\frac{1}{2}}\sum_{i=1}^n a_i\left(M_i - \frac{B_n}{n}\right)}\right) &= -it\bar{a}_n B_n^{\frac{1}{2}} + B_n\log\left(1 + itB_n^{-\frac{1}{2}}\bar{a}_n + \frac{1}{2}\sigma_n^2\left(itB_n^{-\frac{1}{2}}\right)^2 + o\left(B_n^{-1}\right)\right) \\
&= -it\bar{a}_n B_n^{\frac{1}{2}} + B_n\left(itB_n^{-\frac{1}{2}}\bar{a}_n + \frac{1}{2}\left(\sigma_n^2 - \bar{a}_n^2\right)\left(itB_n^{-\frac{1}{2}}\right)^2\right) + o\left(1\right) \\
&= -\frac{1}{2}\left(\sigma_n^2 - \bar{a}_n^2\right)t^2 + o\left(1\right).
\end{aligned}$$

Since we assume tht $\bar{a_n} \to 0$ and $\sigma_n^2 \to \sigma^2$, the above quantity converges to $-\frac{1}{2}\sigma^2 t^2$, which is the logarithm of the characteristic function of $\mathcal{N}(0, \sigma^2)$. ■

Now we could prove the major part of Theorem 10.

Proof Without loss of generality we will assume $\theta_{k_n} = 0$. Suppose $(M_1, \dots, M_{n^{k_n}})$ have a multinomial distribution, multinomial $(B_n; \frac{1}{n^{k_n}}, \dots, \frac{1}{n^{k_n}})$. We could rewrite V_{n,k_n,B_n} as

$$\begin{aligned}
V_{n,k_n,B_n} &= \frac{1}{B_n}\sum_{i} h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}}\right)\\
&= \frac{1}{B_n}\sum_{i=1}^{n^{k_n}} M_i h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}}\right)\\
&= \frac{1}{B_n}\sum_{i=1}^{n^{k_n}} \left(M_i - \frac{B_n}{n^{k_n}} + \frac{B_n}{n^{k_n}}\right) h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}}\right)\\
&= \frac{1}{B_n}\sum_{i=1}^{n^{k_n}} \left(M_i - \frac{B_n}{n^{k_n}}\right) h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}}\right) + \frac{1}{n^{k_n}}\sum_{i=1}^{n^{k_n}} h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}}\right)\\
&= \frac{1}{B_n}\sum_{i=1}^{n^{k_n}} \left(M_i - \frac{B_n}{n^{k_n}}\right) h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}}\right) + V_{n,k_n}.
\end{aligned}$$

To show $\frac{V_{n,k_n,B_n}}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0,1)$, it is equivalent to prove

$$\lim_{n\to\infty}\mathbb{E}\left[\exp\left(it\frac{V_{n,k_n,B_n}}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)\right] = \exp\left(-\frac{1}{2}t^2\right).$$

From the above decomposition of V_{n,k_n,B_n}, we have

$$\begin{aligned}
&\lim_{n\to\infty}\mathbb{E}\left[\exp\left(it\frac{V_{n,k_n,B_n}}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)\right]\\
=&\lim_{n\to\infty}\mathbb{E}\left[\exp\left(it\frac{(\frac{1}{B_n}\sum_{i=1}^{n^{k_n}}(M_i-\frac{B_n}{n^{k_n}})h_{k_n}(Z_{i_1},\ldots,Z_{i_{k_n}})+V_{n,k_n})}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)\right]\\
=&\lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}\left[\exp\left(it\frac{(\frac{1}{B_n}\sum_{i=1}^{n^{k_n}}(M_i-\frac{B_n}{n^{k_n}})h_{k_n}(Z_{i_1},\ldots,Z_{i_{k_n}})+V_{n,k_n})}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)|Z_1,\ldots,Z_n\right]\right]\\
=&\lim_{n\to\infty}\mathbb{E}\left[\exp\left(it\frac{V_{n,k_n}}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)\mathbb{E}\left[\exp\left(it\frac{\frac{1}{B_n}\sum_{i=1}^{n^{k_n}}(M_i-\frac{B_n}{n^{k_n}})h_{k_n}(Z_{i_1},\ldots,Z_{i_{k_n}})}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)|Z_1,\ldots,Z_n\right]\right]
\end{aligned}$$

Since $\frac{V_{n,k_n}}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0,1)$ and by Lemma 11,

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[\exp\left(it\frac{V_{n,k_n,B_n}}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)\right] \\
=&\lim_{n\to\infty} \mathbb{E}\left[\exp\left(it\frac{V_{n,k_n}}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)\mathbb{E}\left[\exp\left(it\frac{\frac{1}{B_n}\sum_{i=1}^{n^{k_n}}(M_i-\frac{B_n}{n^{k_n}})h_{k_n}(Z_{i_1},\dots,Z_{i_{k_n}})}{\sqrt{\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}}}\right)|Z_1,\dots,Z_n\right]\right] \\
=&\exp\left(-\frac{\frac{k_n^2}{n}\zeta_{1,k_n}}{2\left(\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}\right)}t^2\right)\exp\left(-\frac{\frac{1}{B_n}\zeta_{k_n,k_n}}{2\left(\frac{k_n^2}{n}\zeta_{1,k_n}+\frac{1}{B_n}\zeta_{k_n,k_n}\right)}t^2\right) \\
=&\exp\left(-\frac{1}{2}t^2\right).
\end{aligned}
$$

■

B.3 Proof of Theorem 1

Proof By Equation (9), the complete case follows directly from Theorem 1 in Peng et al. (2019). The incomplete case follows exactly the same proof as Theorem 10. ■

B.4 Proof of Theorem 3

Proof

In the case of *Balanced Subsample Structure* where $r_n = \frac{B_n \times k_n}{n}$, we have $\bar{m} = \bar{h}$ and $N_i = r_n$ for all i.

First we can rewrite $\hat{\zeta}_{1,k_n}^{\text{BM}}$ as

$$
\begin{aligned}
\hat{\zeta}_{1,k_n}^{\text{BM}} &= \frac{1}{n-1}\sum_{i=1}^{n}(m_i-\bar{m})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{b=1}^{B_n}\omega_{i,b}h_b-\bar{h}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{b=1}^{B_n}\frac{N_{i,b}}{N_i}h_b-\bar{h}\right)^2 \\
&= \frac{1}{n-1}\frac{1}{r_n^2}\sum_{i=1}^{n}\left(\sum_{b=1}^{B_n}N_{i,b}\left(h_b-\bar{h}\right)\right)^2.
\end{aligned}
$$

Then for $\hat{V}_{\text{IJ}} = \sum_{i=1}^{n} \text{cov}^2(N_{i,b}, h_b)$, we look at each individual term

$$\begin{aligned}
\text{cov}\left(N_{i,b}, h_b\right) &= \frac{\sum_{b=1}^{B_n}(N_{i,b} - \bar{N}_i)(h_b - \bar{h})}{B_n} \\
&= \frac{1}{B_n}\left(\sum_{b,Z_i \in b}\left(N_{i,b} - \bar{N}_i\right)\left(h_b - \bar{b}\right) + \sum_{b,Z_i \notin b}\left(N_{i,b} - \bar{N}_i\right)\left(h_b - \bar{b}\right)\right) \\
&= \frac{1}{B_n}\left(\sum_{b,Z_i \in b}\left(N_{i,b} - \frac{k_n}{n}\right)\left(h_b - \bar{h}\right) + \sum_{b,Z_i \notin b}\left(0 - \frac{k_n}{n}\right)\left(h_b - \bar{h}\right)\right) \\
&= \frac{1}{B_n}\left(\sum_{b,Z_i \in b} N_{i,b}\left(h_b - \bar{h}\right) - \frac{k_n}{n}\sum_{b}\left(h_b - \bar{h}\right)\right) \\
&= \frac{1}{B_n}\sum_{b,Z_i \in b} N_{i,b}\left(h_b - \bar{h}\right) \\
&= \frac{1}{B_n}\sum_{b=1}^{B_n} N_{i,b}\left(h_b - \bar{h}\right)
\end{aligned}$$

where Z_i denotes the i^{th} training sample.

Combining two previous identities

$$\begin{aligned}
\hat{V}_{\text{IJ}} &= \sum_{i=1}^{n} \text{cov}^2(N_{i,b}, h_b) \\
&= \frac{1}{B_n^2}\sum_{i=1}^{n}\left(\sum_{b=1}^{B_n} N_{i,b}(h_b - \bar{b})\right)^2 \\
&= \frac{(n-1)r_n^2}{B_n^2}\hat{\zeta}_{1,k_n}^{\text{BM}} \\
&= \frac{n-1}{n}\frac{k_n^2}{n}\hat{\zeta}_{1,k_n}^{\text{BM}}
\end{aligned}$$

as claimed.

■

B.5 Proof of Theorem 6

Proof The assumption $\lim_{n\to\infty} k_n^2 \zeta_{1,k_n}^* > 0$ implies $\lim_{n\to\infty} \text{Var}(\sqrt{n} V_{n,k_n,B_n,\omega}^*) > 0$. We first show the complete case. Similar to the proof of Theorem 2 in Mentch and Hooker (2016), we have

$$\mathbb{E}(V_{n,k_n,\omega} - V_{n,k_n,\omega}^*)^2 = \frac{1}{(n^{k_n})^2}\mathbb{E}\sum_{i}\left(h_{k_n}\left(Z_{i_1}, \ldots, Z_{i_{k_n}}; \omega\right) - \mathbb{E}_\omega h_{k_n}\left(Z_{i_1}, \ldots, Z_{i_{k_n}}; \omega\right)\right)^2.$$

Thus,

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left(\frac{\sqrt{n}(V_{n,k_n,\omega} - V^*_{n,k_n,\omega})}{\sqrt{\mathrm{Var}(\sqrt{n}V^*_{n,k_n,\omega})}}\right)^2 \\
=&\lim_{n\to\infty} \mathbb{E}\frac{n}{n^{k_n}}\frac{1}{\mathrm{Var}(\sqrt{n}V^*_{n,k_n,\omega})}\frac{1}{n^{k_n}}\sum_i \mathbb{E}\left(h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}};\omega\right) - \mathbb{E}_\omega h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}};\omega\right)\right)^2 \\
=&0
\end{aligned}
$$

since $\sup_{k_n}\mathbb{E}\left(h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}};\omega\right) - \mathbb{E}_\omega h_{k_n}\left(Z_{i_1},\ldots,Z_{i_{k_n}};\omega\right)\right) < \infty$ and $\lim_{n\to\infty}\mathrm{Var}(\sqrt{n}V^*_{n,k_n,\omega}) > 0$.

The incomplete case follows exactly as in Mentch and Hooker (2016). ■

Appendix C. Additional Simulation Results

Simulations in this section are based on a simple setting where $X \sim 20 \times \mathrm{unif}(0,1)$ and $Y = 2X + \mathcal{N}(0,1)$. The number of training observations $n = 500$. The model is an ensemble of decision trees.

Figure 4 and 5 displays variance estimation by IM, BM and IJ for kernel size $k_n = 250$ and 400.

Figure 6 and Figure 7 shows the estimated values for each variance components ζ_{1,k_n} and ζ_{k_n,k_n} for $k_n = 100$. Since IJ does not target at ζ_{1,k_n} directly, we rescaled the estimated by a factor $\frac{k_n^2}{n}$ according to Theorem 3. The estimators for ζ_{k_n,k_n} for the three methods shown are essentially the same as they are all calculating the variance across all base learners' predictions.

Appendix D. Derivations on Bias-corrected Estimator

Our goal is to provide an estimator of ζ_{1,k_n} based on expression given in (10)

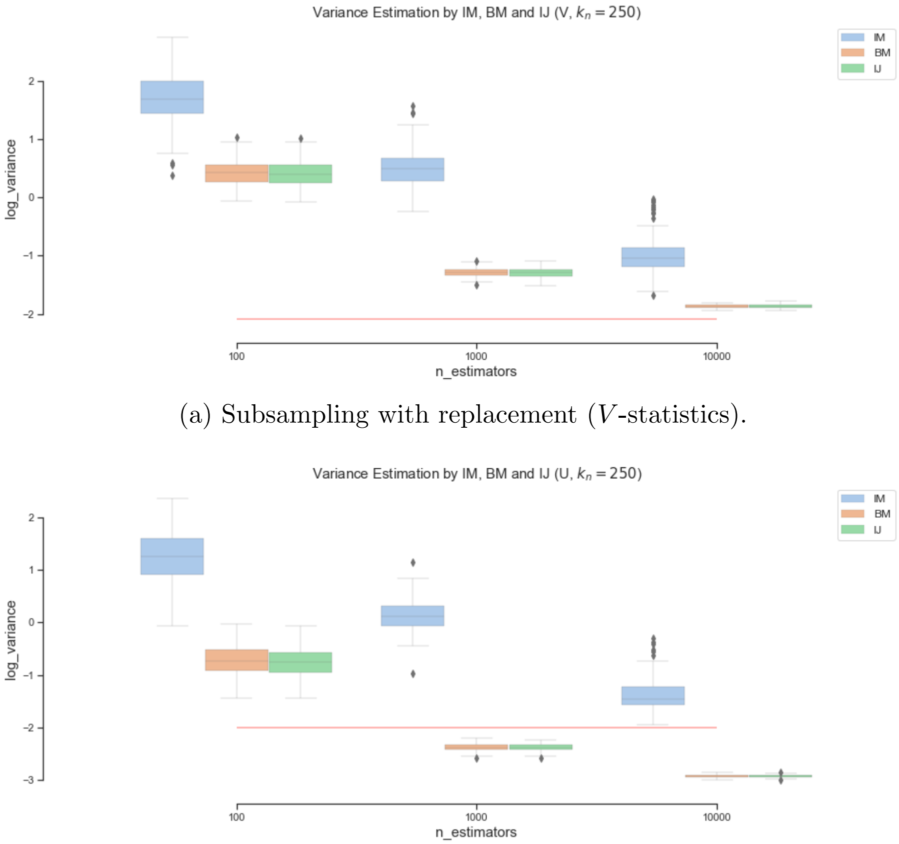
$$\zeta_{1,k_n} = \mathrm{var}\left(\mathbb{E}\left(h_{k_n}\left(Z_1,\ldots,Z_{k_n}\right)|Z_1 = z_1\right)\right).$$

To simplify notations, we introduce a more general mathematical representation. Consider a random variable X and its conditional distribution given a random variable Z, denoted by $M = \mathbb{E}(X|Z)$. We want to estimate the variance of $\sigma_M^2 = \mathrm{Var}(M)$. Use F_Z and $F_{X|Z}$ to denote the distribution for Z and the conditional distribution X given Z respectively.

Consider the following sampling framework: for $k = 1,\ldots,K$:

1. Sample Z_k randomly from F_Z.

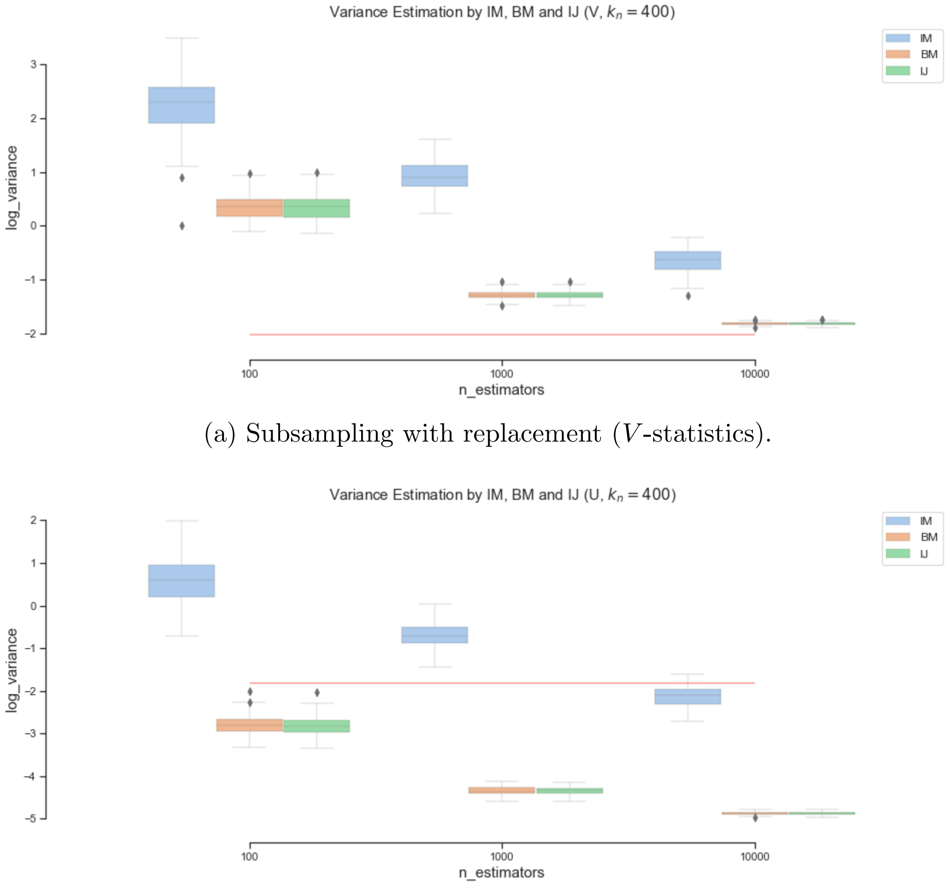
2. For $j = 1,\ldots,n_k$: Sample X_{kj} randomly from $F_{X|Z=Z_k}$.



(a) Subsampling with replacement (V-statistics).

(b) Subsampling without replacement (U-statistics).

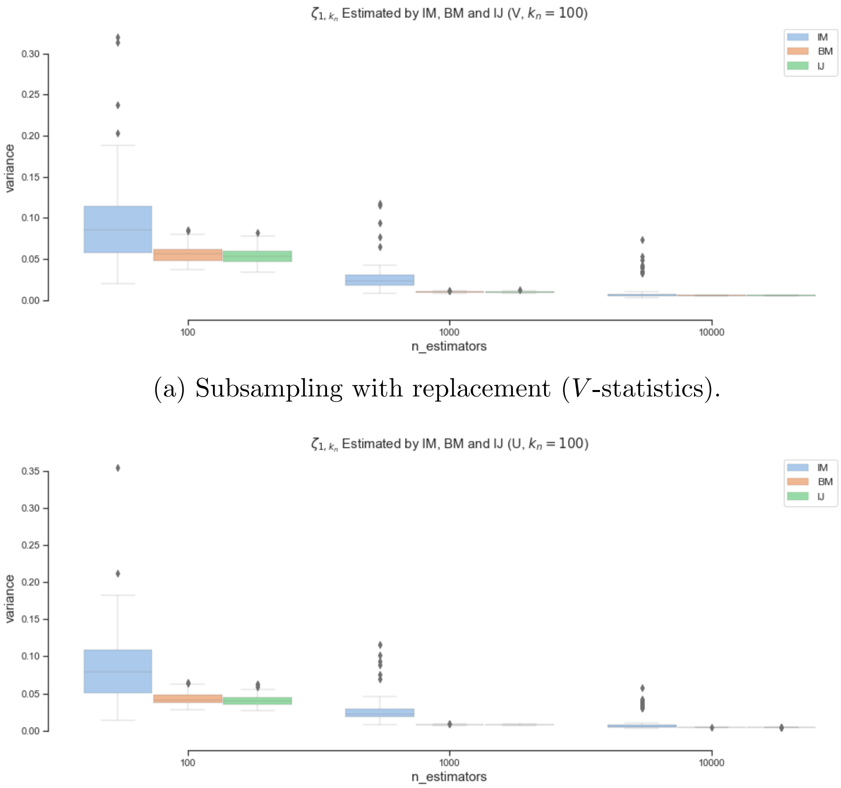
Figure 4: Variance estimation by three different methods: IM, BM and IJ. The kernel size $k_n = 250$. The variance shown is for prediction at test point $x = 10$. The red line denotes true (log) variance obtained by generating data, training the ensemble 100 times and calculating the empirical variance of predictions.



(a) Subsampling with replacement (V-statistics).

(b) Subsampling without replacement (U-statistics).

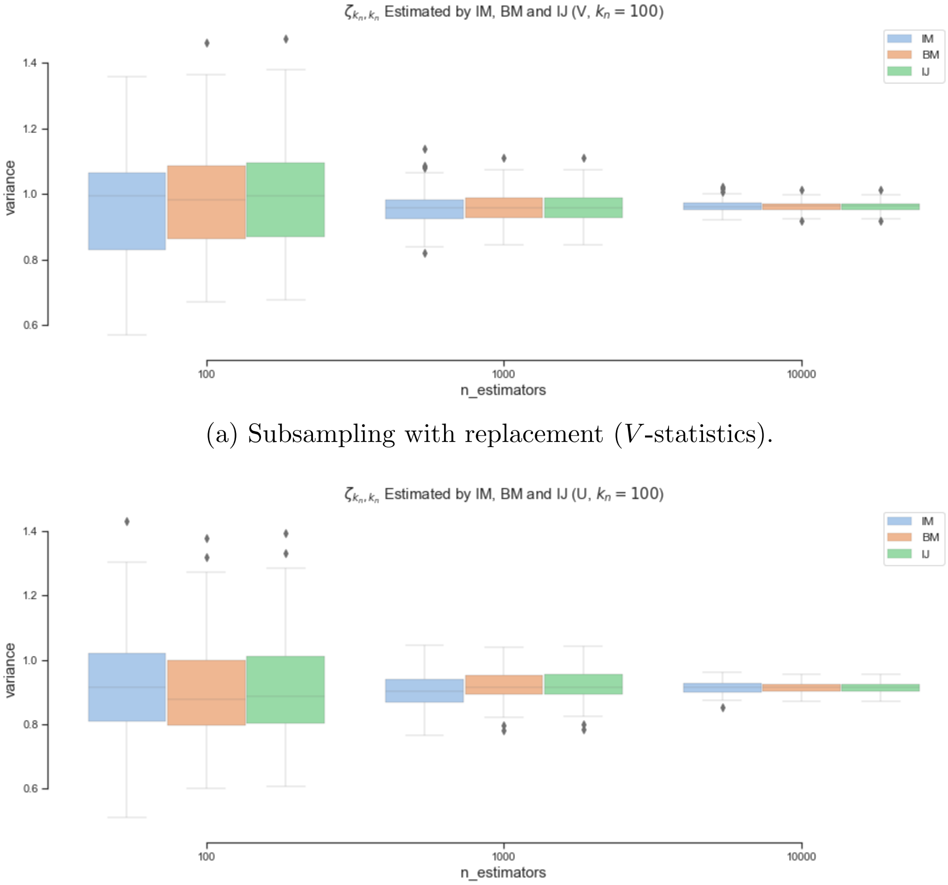
Figure 5: Variance estimation by three different methods: IM, BM and IJ. The kernel size $k_n = 400$. The variance shown is for prediction at test point $x = 10$. The red line denotes true (log) variance obtained by generating data, training the ensemble 100 times and calculating the empirical variance of predictions.



(a) Subsampling with replacement (V-statistics).

(b) Subsampling without replacement (U-statistics).

Figure 6: ζ_{1,k_n} estimated by three different methods: IM, BM and IJ. The kernel size $k_n = 100$. The variance shown is for prediction at test point $x = 10$.



(a) Subsampling with replacement (V-statistics).

(b) Subsampling without replacement (U-statistics).

Figure 7: ζ_{k_n,k_n} estimated by three different methods: IM, BM and IJ. The kernel size $k_n = 100$. The variance shown is for prediction at test point $x = 10$.

We'll use the collections of samples X_{kj} $(k = 1, \ldots, K, j = 1, \ldots, n_k)$ to provide an estimator for σ_M. Define $C = \sum_{k=1}^{K} n_k$, $\sigma_\epsilon^2 = \mathbb{E}(\mathrm{Var}(X|Z))$ and the following two sum of squares

$$SS_\tau = \sum_{k=1}^{K} n_k (\bar{X}_k - \bar{\bar{X}})^2,$$

$$SS_\epsilon = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (X_{kj} - \bar{X}_k)^2$$

where $\bar{\bar{X}} = \frac{1}{C} \sum_{k=1}^{K} n_k \bar{X}_k$, $\bar{X}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} X_{kj}$. Following the calculations in Sun et al. (2011), we have

$$\mathbb{E}(SS_\tau) = \left(C - \sum_{i=1}^{K} n_i^2 / C \right) \sigma_M^2 + (K-1)\, \sigma_\epsilon^2,$$

$$\mathbb{E}(SS_\epsilon) = (C - K)\sigma_\epsilon^2.$$

Thus we can get the estimator for σ_M^2 as

$$\hat{\sigma}_M^2 = \frac{SS_\tau - (K-1)\hat{\sigma}_\epsilon^2}{C - \sum_{i=1}^{K} n_i^2 / C}$$

where

$$\hat{\sigma}_\epsilon^2 = \frac{SS_\epsilon}{C - K}.$$

The unbiasedness of these estimators is shown by Searle et al. (2009). By setting $Z = Z_1$ and $X = h_{k_n}(Z_1, \ldots, Z_{k_n})$ gives the estimator presented in Section 5.2.

Appendix E. An Alternative Version of Bias Correction

Our subsampling methods choose each data point with equal probability and we thus expect to obtain an approximately balanced subsample. For simplicity, our derivation assumes this structure holds exactly.

Recall that we have $N_1 = N_2 = \ldots = N_i = r_n$, then

$$\hat{\sigma}_\epsilon^2 = \frac{SS_\epsilon}{C - n} = \frac{1}{n(r_n - 1)} \sum_{i=1}^{n} \sum_{b=1}^{B_n} N_{i,b} (h_b - m_i)^2,$$

and

$$\hat{\zeta}_{1,k_n} = \frac{1}{n-1} \sum_{i=1}^{n} (m_i - \bar{m})^2 - \frac{1}{r_n} \hat{\sigma}_\epsilon^2.$$

We could rewrite $\hat{\sigma}_\epsilon^2$ as

$$\begin{aligned}
\hat{\sigma}_\epsilon^2 &= \frac{1}{n(r_n-1)} \sum_{i=1}^{n} \sum_{b=1}^{B_n} N_{i,b}(h_b - m_i)^2 \\
&= \frac{1}{n(r_n-1)} \sum_{i=1}^{n} \sum_{b=1}^{B_n} N_{i,b}(h_b - \bar{h} + \bar{h} - m_i)^2 \\
&= \frac{1}{n(r_n-1)} \sum_{i=1}^{n} \sum_{b=1}^{B_n} N_{i,b}(h_b - \bar{h})^2 + \frac{1}{n(r_n-1)} \sum_{i=1}^{n} \sum_{b=1}^{B_n} N_{i,b}(\bar{h} - m_i)^2 \\
&\quad + \frac{2}{n(r_n-1)} \sum_{i=1}^{n} \sum_{b=1}^{B_n} N_{i,b}(h_b - \bar{h})(\bar{h} - m_i) \\
&= \frac{1}{n(r_n-1)} \sum_{b=1}^{B_n} (h_b - \bar{h})^2 \sum_{i=1}^{n} N_{i,b} + \frac{1}{n(r_n-1)} \sum_{i=1}^{n} (\bar{h} - m_i)^2 \sum_{b=1}^{B_n} N_{i,b} \\
&\quad + \frac{2}{n(r_n-1)} \sum_{i=1}^{n} (\bar{h} - m_i) \sum_{b=1}^{B_n} N_{i,b}(h_b - \bar{h}) \\
&= \frac{1}{n(r_n-1)} \sum_{b=1}^{B_n} (h_b - \bar{h})^2 k_n + \frac{1}{n(r_n-1)} \sum_{i=1}^{n} (\bar{h} - m_i)^2 r_n \\
&\quad + \frac{2}{n(r_n-1)} \sum_{i=1}^{n} (\bar{h} - m_i) r_n (m_i - \bar{h}) \\
&= \frac{k_n}{n(r_n-1)} \sum_{b=1}^{B_n} (h_b - \bar{h})^2 - \frac{r_n}{n(r_n-1)} \sum_{i=1}^{n} (\bar{h} - m_i)^2 \\
&= \frac{k_n}{n(r_n-1)} \sum_{b=1}^{B_n} (h_b - \bar{h})^2 - \frac{r_n}{n(r_n-1)} \sum_{i=1}^{n} (m_i - \bar{m})^2.
\end{aligned}$$

Plug this into the expression for $\hat{\zeta}_{1,k_n}$, we have

$$
\begin{aligned}
\hat{\zeta}_{1,k_n} &= \frac{1}{n-1}\sum_{i=1}^{n}(m_i-\bar{m})^2 - \frac{1}{r_n}\hat{\sigma}_\epsilon^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(m_i-\bar{m})^2 - \frac{1}{r_n}\left(\frac{k_n}{n(r_n-1)}\sum_{b=1}^{B_n}\left(h_b-\bar{h}\right)^2 - \frac{r_n}{n(r_n-1)}\sum_{i=1}^{n}(m_i-\bar{m})^2\right) \\
&= \left(\frac{1}{n-1}-\frac{1}{n(r_n-1)}\right)\sum_{i=1}^{n}(m_i-\bar{m})^2 - \frac{k_n}{r_n n(r_n-1)}\sum_{b=1}^{B_n}\left(h_b-\bar{h}\right)^2 \\
&= \left(\frac{1}{n-1}-\frac{1}{n(r_n-1)}\right)\sum_{i=1}^{n}(m_i-\bar{m})^2 - \frac{k_n}{r_n n(r_n-1)}\left(B_n-1\right)\hat{\zeta}_{k_n,k_n}^{\mathrm{BM}} \\
&\approx \frac{1}{n-1}\sum_{i=1}^{n}(m_i-\bar{m})^2 - \frac{1}{B_n}\frac{n}{k_n}\hat{\zeta}_{k_n,k_n}^{\mathrm{BM}}.
\end{aligned}
$$

The approximation in the last line holds as long as r_n grows with n, which is a reasonable assumption in most cases.

Appendix F. Bias Correction for U-statistics

Simulations in this section are based on a simple setting where $X \sim 20 \times \mathrm{unif}(0,1)$ and $Y = 2X + \mathcal{N}(0,1)$. The number of training observations $n = 500$. The model is an ensemble of decision trees built under the framework of U-statistics: each tree is constructed using subsamples *without* replacement.

Figure 8 shows the result for U-statistics by employing the correction by Wager and Athey (2018).

Figure 9 shows the result of corrected-V developed in Section 5.2 applied to U-statistics.

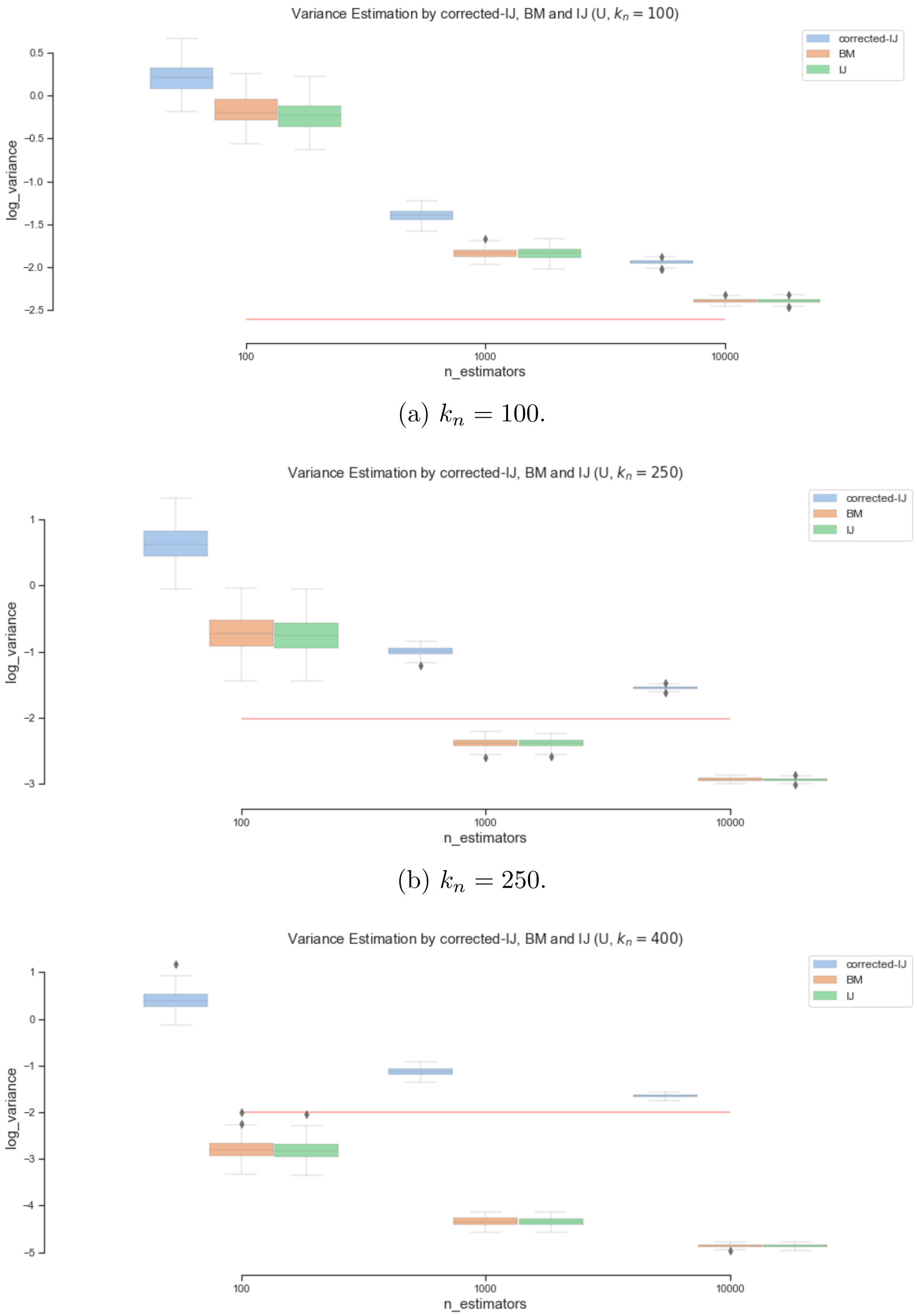
For simplicity, we will use the simpler but approximate variance estimation described in Appendix E

$$
\hat{\zeta}_{1,k_n} = \frac{1}{n-1}\sum_{i=1}^{n}(m_i-\bar{m})^2 - \frac{1}{B_n}\frac{n}{k_n}\hat{\zeta}_{k_n,k_n}^{\mathrm{BM}}.
$$

We find that if we scale the correction term by $\frac{n-k_n}{n}$, and include the correction term in Wager and Athey (2018), it works for U-statistics empirically. The estimator for ζ_{1,k_n} for U-statistics is

$$
\hat{\zeta}_{1,k_n}^{U} = \frac{n(n-1)}{(n-k_n)^2}\left(\frac{1}{n-1}\sum_{i=1}^{n}(m_i-\bar{m})^2 - \frac{1}{B_n}\frac{n-k_n}{k_n}\hat{\zeta}_{k_n,k_n}^{\mathrm{BM}}\right).
$$

The blue bars (denoted by corrected-U) in Figure 10 shows the result. We can see that by combining both correction terms, the estimator yields stable and accurate variance estimation. How to theoretically analyze bias correction for U-statistics remains a promising future endeavor.



(a) $k_n = 100$.

(b) $k_n = 250$.

(c) $k_n = 400$.

Figure 8: Variance Estimation by three different methods: corrected-IJ, BM and IJ. The kernel size $k_n = 100, 250, 400$. The variance shown is for prediction at test point $x = 10$. The red line denotes true (log) variance obtained by generating data, training the ensemble 100 times and calculating the empirical variance of predictions.

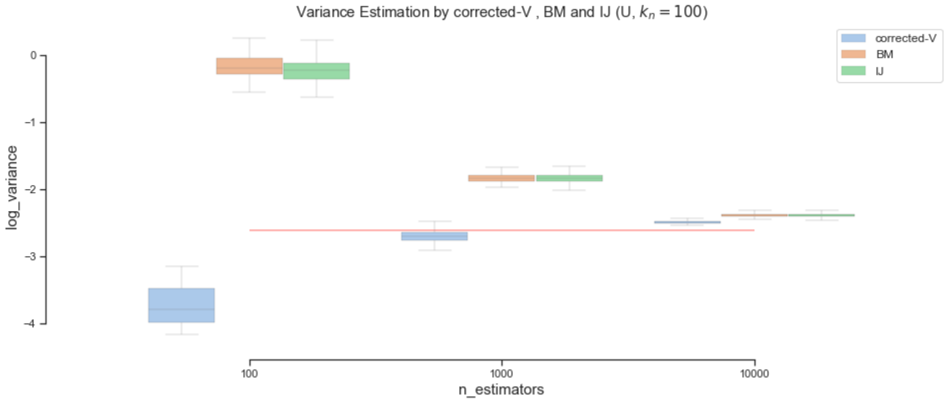


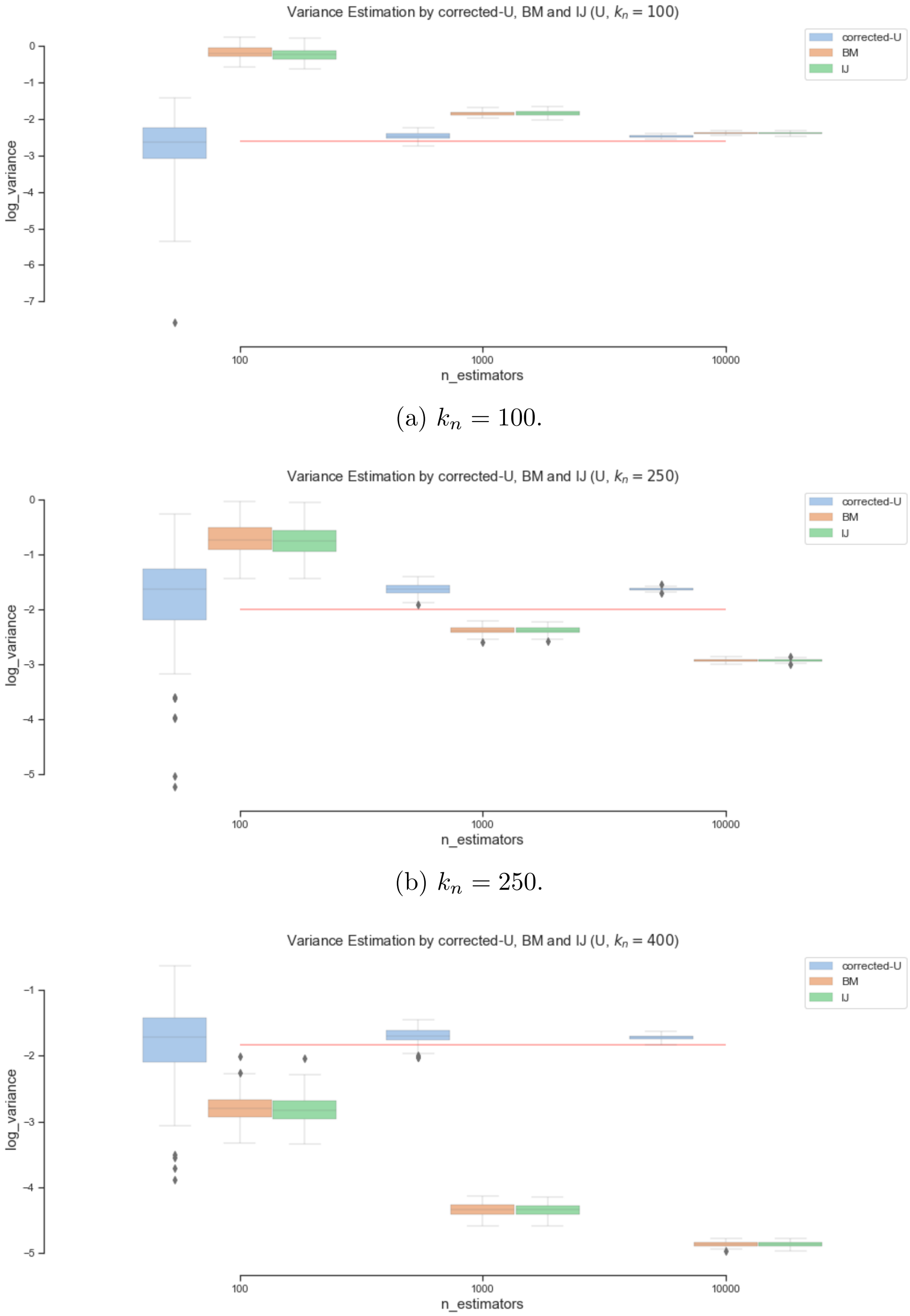
Figure 9: Variance estimation by three different methods: corrected-V, BM and IJ. The variance shown is for prediction at test point $x = 10$. The red line denotes true (log) variance obtained by generating data, training the ensemble 100 times and calculating the empirical variance of predictions.

Appendix G. A Closer Look at Variance Components

A major difference between our work and Wager and Athey (2018) is that we also take into account the effect of Monte Carlo effect brought by the number of base learners B_n. Thus our variance has two components, where the first part $\frac{k_n^2}{n}\zeta_{1,k_n}$ corresponds to the complete case and the second part $\frac{1}{B_n}\zeta_{k_n,k_n}$ is the additional Monte Carlo variance introduced due to the incomplete case.

One can imagine that for smaller B_n, the second part of Monte Carlo variance might be much larger than the first part, while as B_n gets larger the effect diminishes and $\frac{k_n^2}{n}\zeta_{1,k_n}$ becomes the dominating one. We conduct an experiment to visualize this transition. As before, let $X \sim 20 \times \text{unif}(0,1)$ and $Y = 2X + \mathcal{N}(0,1)$. The model is an ensemble of decision trees built under the framework of V-statistics. We fix the number of training observations $n = 1000$ and kernel size $k_n = 10$ and build the ensembles with $B_n = 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000$. For each B_n, 100 models are built and we calculate empirical variance of the predictions at test point $x = 10$. This procedure is repeated 10 times and we report the average of empirical variance.

Figure 11 shows four lines: the empirical variance; the two variance components $\frac{k_n^2}{n}\zeta_{1,k_n}$ and $\frac{1}{B_n}\zeta_{k_n,k_n}$; the total estimated variance which is simply the sum $\frac{k_n^2}{n}\zeta_{1,k_n} + \frac{1}{B_n}\zeta_{k_n,k_n}$. We estimate ζ_{1,k_n} and ζ_{k_n,k_n} using an ensemble of size $B_n = 1000$. The dotted black line aligns well with the black line, which indicates that our variance estimates give accurate results. For small $B_n = 100$, each observation is expected to only appear once in the ensemble (since $r_n = \frac{k_n \times B_n}{n} = 1$), and as a result base learners will be approximately independent. In this case, the variance of the ensemble prediction should mainly come from $\frac{1}{B_n}\zeta_{k_n,k_n}$. When B_n grows, dependence between some base learners kicks in and the effect of $\frac{k_n^2}{n}\zeta_{1,k_n}$ gradually becomes the dominating part as $\frac{1}{B_n}\zeta_{k_n,k_n}$ decreases. This transition is depicted in Figure



(a) $k_n = 100$.

(b) $k_n = 250$.

(c) $k_n = 400$.

Figure 10: Variance Estimation by three different methods: corrected-U, BM and IJ. The kernel size $k_n = 100, 250, 400$. The variance shown is for prediction at test point $x = 10$. The red line denotes true (log) variance obtained by generating data, training the ensemble 100 times and calculating the empirical variance of predictions.

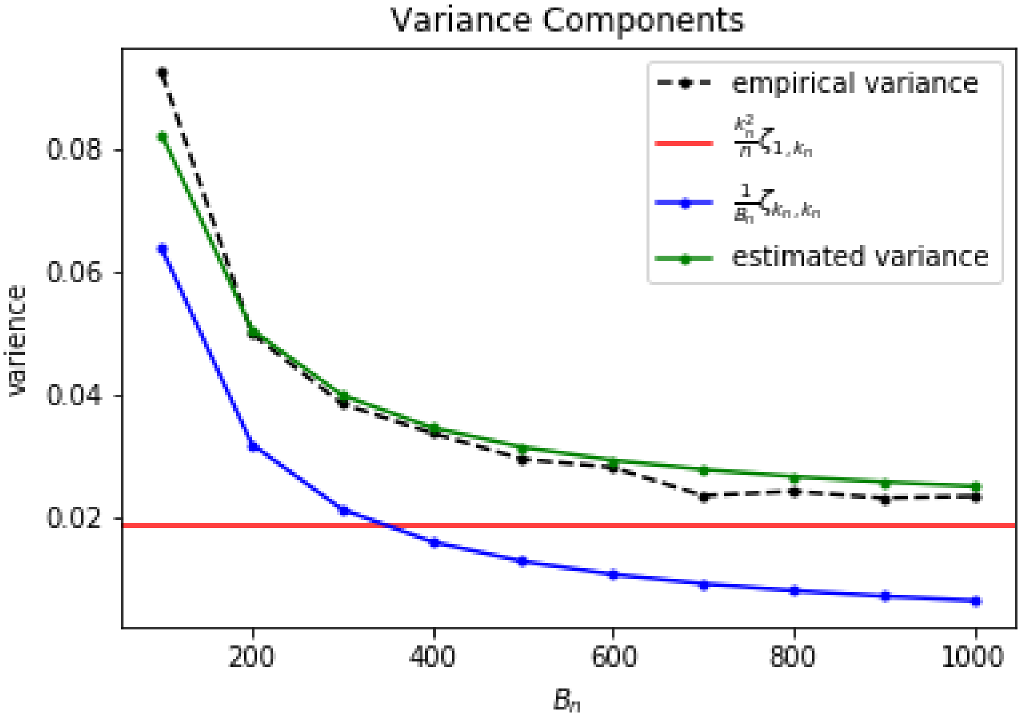


Figure 11: Variance components for different B_n. The number of training observations $n = 1000$ and kernel size $k_n = 10$. The variance shown is for prediction at test point $x = 10$. Four lines shown are: empirical variance, two variance components ($\frac{k_n^2}{n}\zeta_{1,k_n}$ and $\frac{1}{B_n}\zeta_{k_n,k_n}$) and their sum as estimated variance.

11. Note that in practice the additional Monte Carlo variance introduced due to incomplete case is usually negligible as we would choose larger k_n and B_n.

Appendix H. Additional Simulation Results on Normality for Ensembles

The experiment is conducted under the same setting of Section 7.2. Let $y = f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.05)^2 + 10x_4 + 5x_5 + \epsilon$, where $\mathcal{X} \sim U([0,1]^5)$ and $\epsilon \sim \mathcal{N}(0,1)$.

We fix $B = 500$ and vary subsample size $k = 100, 200, 300, 400, 500$. We randomly general 100 test points from $U([0,1]^5)$. For each test point, we conduct 100 simulations where in each iteration we generate $n = 500$ training observations to fit a random forests and record its prediction. Finally a normality test is conducted on 100 predictions of the same test point. The experiment is repeated for both V- and U-statistics. Table 4 shows the percentage of p-values falls below 0.05 (a normal hypothesis is rejected). We can see that in our setting normality for predictions generally hold for V-statistics across all kernel size, while for U-statistics it starts to break down for large kernel size.

	V-statistics	U-statistics
100	0.1	0.07
200	0.06	0.05
300	0.06	0.04
400	0.05	0.11
500	0.07	0.18

Table 4: Normality Test for Ensembles.

Appendix I. Variance Estimation for V-statistics and Its Implications

We prove the theoretical asymptotics for general V-statistics utilizing a composite kernel $h^*_{k_n}$, which is infeasible to evaluate in practice. Thus it remains a challenge to quantify its variance directly which involves ζ^*_{1,k_n}, and we use Infinitesimal Jackknife as a workaround. IJ was initially developed for computing standard errors and confidence interval in bagging (Efron, 2014). As a general tool, IJ does not rely on any specific variance expressions and is applied upon the original kernel h_{k_n} instead of $h^*_{k_n}$. Consistency of IJ under the framework of U-statistics was proved in Wager and Athey (2018); Ghosal and Hooker (2020).

The result in Theorem 3 establishes a connection between BM and IJ. Thus applying BM on the original kernel h_{k_n} of a V-statistic should yield valid variance estimates. This is how we calculate limiting variance in all of the empirical studies. Although we do not theoretically prove IJ is consistent for V-statistics, the empirical results in Table 2 show a promising sign. If this is the case, then we should have $\zeta_{1,k_n} = \zeta^*_{1,k_n}$ for a general V-statistics under the conditions of Theorem 1. We leave this conjecture as a future work.

Appendix J. U-statistics Results for Section 7.2

This section shows the results for U-statistics under the same setting of Section 7.2.

Table 5 presents the normality test, variance ratio and coverage for the same three test points as in Table 2 using the original BM variance estimation method. This further supports our previous finding that for U-statistics a prohibitive number of base learner is needed for valid inference; while for larger kernel sizes, normality starts to break down and as a result the estimated variance is no longer valid.

Corresponding U-statistics result for Protein Tertiary Structure data is shown in Table 6.

Appendix K. Datasets Information

Six of the seven datasets in Table 7.1 are taken from UCI Machine Learning Repository[5]:

- boston: `https://archive.ics.uci.edu/ml/machine-learning-databases/housing/`. The dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. This is a regression task to predict median value of owner-occupied homes in $1000's.

[5] `https://archive.ics.uci.edu/ml/index.php`

		p_1	p_2	p_3
		original	original	original
B = 500	normality	2.4464 (0.2943)	1.7276 (0.4216)	3.9023 (0.1421)
	var_ratio	5.6619	6.1710	4.1774
	coverage	100.0	100.0	99.8
B = 1000	normality	4.0015 (0.1352)	0.7224 (0.6968)	0.0999 (0.9513)
	var_ratio	3.5290	3.6977	2.7600
	coverage	99.8	100.0	99.8
B = 2500	normality	1.6279 (0.4431)	4.2958 (0.1167)	0.4680 (0.7914)
	var_ratio	1.6725	1.8111	1.6239
	coverage	98.2	99.0	97.2
B = 5000	normality	3.8553(0.1455)	0.7108 (0.7009)	2.1749 (0.3371)
	var_ratio	1.2396	1.3060	1.1624
	coverage	96.6	98.2	94.6

(a) k = 100

		p_1	p_2	p_3
		original	original	original
B = 500	normality	3.8301 (0.1473)	3.0423 (0.2185)	1.3942 (0.4980)
	var_ratio	3.2972	2.9889	2.3566
	coverage	99.6	100.0	99.0
B = 1000	normality	5.9749 (0.0504)	0.3142 (0.8546)	0.0418 (0.9793)
	var_ratio	1.9543	1.7152	1.6739
	coverage	98.4	98.6	97.2
B = 2500	normality	7.4631 (0.0239)	2.0683 (0.0.3555)	2.3659 (0.3064)
	var_ratio	0.9110	0.8726	0.7628
	coverage	91.8	92.0	89.6
B = 5000	normality	6.2611 (0.0437)	1.4629 (0.4812)	1.9362 (0.3798)
	var_ratio	0.5918	0.6795	0.5756
	coverage	85.4	87.6	83.4

(b) k = 250

		p_1	p_2	p_3
		original	original	original
B = 500	normality	28.9413 (5.19e-07)	1.7923 (0.4081)	7.0993 (0.0287)
	var_ratio	12.27e-05	17.39e-05	16.13e-05
	coverage	1.4	1.4	0.4
B = 1000	normality	11.3388 (0.0034)	3.8287 (0.1474)	8.5205 (0.0141)
	var_ratio	9.03e-05	9.63e-05	8.48e-05
	coverage	1.2	0.6	1.4
B = 2500	normality	2.9161 (0.2327)	0.3811 (0.8265)	13.3368 (0.0013)
	var_ratio	3.17e-05	3.63e-05	3.48e-05
	coverage	0.6	0.6	0.6
B = 5000	normality	5.2564 (0.7222)	4.0334 (0.1331)	10.9238 (0.0042)
	var_ratio	1.63e-05	1.91e-05	1.65e-05
	coverage	0.4	0.4	0.6

(c) k = 500

Table 5: Asymptotic normality and variance estimation results for MARS function: U-statistics.

		$k = 100$	$k = 250$	$k = 500$
		original	original	original
B = 500	normality	0.10	0.15	0.10
	var_ratio	10.77(2.65)	13.20(5.70)	6.88(1.73)
	coverage	100.0(0.0)	100.0(0.0)	99.88(0.00)
B = 1000	normality	0.15	0.15	0.15
	var_ratio	6.03(1.49)	6.95(2.46)	3.62(1.01)
	coverage	100.0(0.0)	100.0(0.0)	99.33(0.01)
B = 2500	normality	0.10	0.20	0.15
	var_ratio	3.07(0.77)	3.29(1.20)	1.67(0.40)
	coverage	99.33(0.01)	99.7(0.01)	96.8(0.03)
B = 5000	normality	0.20	0.15	0.15
	var_ratio	2.00(0.46)	2.05(0.72)	1.01(0.24)
	coverage	98.22(0.02)	97.67(0.03)	91.0(0.05)

Table 6: Asymptotic normality and variance estimation results for Protein Tertiary Structure across 20 test samples: U-statistics.

- diabetes: `https://archive.ics.uci.edu/ml/datasets/diabetes`. The attributes are diabetes patient records and the target is an integer between 25 and 346. We simply cast it as a regression problem.

- iris: `https://archive.ics.uci.edu/ml/datasets/Iris`. This is a classification problem. The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

- digits: `https://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits`. A classification task to predict integers from 0 to 9 with 64 attributes.

- retinopathy: `https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set`. This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not.

- breast_cancer: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`. This is a classification task to predict whether the diagnosis is malignant or benign based on features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

References

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. The Annals of Statistics, 47(2):1148–1178, 2019.

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

Andreas Buja and Werner Stuetzle. Observations on bagging. Statistica Sinica, pages 323–351, 2006.

Yifan Cui, Ruoqing Zhu, Mai Zhou, and Michael Kosorok. Consistency of survival tree and forest models: splitting bias and correction. arXiv preprint arXiv:1707.09631, 2017.

Ralph D'Agostino and Egon S Pearson. Tests for departure from normality. empirical results for the distributions of b_2 and $\sqrt{b}_1$. Biometrika, 60(3):613–622, 1973.

Ralph B D'Agostino. An omnibus test of normality for moderate and large size samples. Biometrika, 58(2):341–348, 1971.

Bradley Efron. The jackknife, the bootstrap and other resampling plans. SIAM, 1982.

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

Edward W Frees. Infinite order u-statistics. Scandinavian Journal of Statistics, pages 29–45, 1989.

Jerome H Friedman. Multivariate adaptive regression splines. The annals of statistics, pages 1–67, 1991.

Indrayudh Ghosal and Giles Hooker. Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. Journal of Computational and Graphical Statistics, pages 1–10, 2020.

Takashi Goda. Computing the variance of a conditional expectation via non-nested monte carlo. Operations Research Letters, 45(1):63–67, 2017.

Paul R Halmos. The theory of unbiased estimation. The Annals of Mathematical Statistics, pages 34–43, 1946.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics, pages 293–325, 1948.

Giles Hooker and Lucas Mentch. Bootstrap bias corrections for ensemble methods. Statistics and Computing, 28(1):77–86, 2018.

Louis A Jaeckel. The infinitesimal jackknife. Bell Telephone Laboratories, 1972.

Justin Lee. U-statistics: Theory and Practice. Citeseer, 1990.

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

Wei Peng, Tim Coleman, and Lucas Mentch. Asymptotic distributions and rates of convergence for random forests and other resampled ensemble learners. arXiv preprint arXiv:1905.10651, 2019.

Basil Cameron Rennie and Annette Jane Dobson. On stirling numbers of the second kind. Journal of Combinatorial Theory, 7(2):116–121, 1969.

Erwan Scornet, Gérard Biau, Jean-Philippe Vert, et al. Consistency of random forests. The Annals of Statistics, 43(4):1716–1741, 2015.

Shayle R Searle, George Casella, and Charles E McCulloch. Variance components, volume 391. John Wiley & Sons, 2009.

PK Sen. Introduction to hoeffding (1948) a class of statistics with asymptotically normal distribution. In Breakthroughs in statistics, pages 299–307. Springer, 1992.

Robert J Serfling. Approximation theorems of mathematical statistics, volume 162. John Wiley & Sons, 2009.

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

Grace S Shieh. Infinite order v-statistics. Statistics & Probability Letters, 20(1):75–80, 1994.

Yanglei Song, Xiaohui Chen, Kengo Kato, et al. Approximating high-dimensional infinite-order u-statistics: Statistical and computational guarantees. Electronic Journal of Statistics, 13(2):4794–4848, 2019.

Jeremy Staum. Monte carlo computation in finance. In Monte Carlo and Quasi-Monte Carlo Methods 2008, pages 19–42. Springer, 2009.

Yunpeng Sun, Daniel W Apley, and Jeremy Staum. Efficient nested simulation for estimating the variance of a conditional expectation. Operations research, 59(4):998–1007, 2011.

Ralph von Mises. On the asymptotic distribution of differentiable statistical functions. The annals of mathematical statistics, 18(3):309–348, 1947.

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

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. The Journal of Machine Learning Research, 15(1):1625–1651, 2014.

Qing Wang and Bruce Lindsay. Variance estimation of a general u-statistic with application to cross-validation. Statistica Sinica, pages 1117–1141, 2014.

Yichen Zhou and Giles Hooker. Boulevard: Regularized stochastic gradient boosted trees and their limiting distribution. arXiv preprint arXiv:1806.09762, 2018.

Yichen Zhou, Zhengze Zhou, and Giles Hooker. Approximation trees: Statistical stability in model distillation. arXiv preprint arXiv:1808.07573, 2018.

Zhi-Hua Zhou. Ensemble methods: foundations and algorithms. CRC press, 2012.

Faker Zouaoui and James R Wilson. Accounting for parameter uncertainty in simulation input modeling. Iie Transactions, 35(9):781–792, 2003.