Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests

Lucas Mentch   LKM54@CORNELL.EDU

Giles Hooker   GILES.HOOKER@CORNELL.EDU

Department of Statistical Science
Cornell University
Ithaca, NY 14850, USA

Editor: Bin Yu

Abstract

This work develops formal statistical inference procedures for predictions generated by supervised learning ensembles. Ensemble methods based on bootstrapping, such as bagging and random forests, have improved the predictive accuracy of individual trees, but fail to provide a framework in which distributional results can be easily determined. Instead of aggregating full bootstrap samples, we consider predicting by averaging over trees built on subsamples of the training set and demonstrate that the resulting estimator takes the form of a U-statistic. As such, predictions for individual feature vectors are asymptotically normal, allowing for confidence intervals to accompany predictions. In practice, a subset of subsamples is used for computational speed; here our estimators take the form of incomplete U-statistics and equivalent results are derived. We further demonstrate that this setup provides a framework for testing the significance of features. Moreover, the internal estimation method we develop allows us to estimate the variance parameters and perform these inference procedures at no additional computational cost. Simulations and illustrations on a real data set are provided.

Keywords: trees, u-statistics, bagging, subbagging, random forests

1. Introduction

This paper develops tools for performing formal statistical inference for predictions generated by a broad class of methods developed under the algorithmic framework of data analysis. In particular, we focus on ensemble methods—combinations of many individual, frequently tree-based, prediction functions—which have played an important role. We present a variant of bagging and random forests, both initially introduced by Breiman (1996, 2001b), in which base learners are built on randomly chosen subsamples of the training data and the final prediction is taken as the average over the individual outputs. We demonstrate that this fits into the statistical framework of U-statistics, which were shown to have minimum variance by Halmos (1946) and later demonstrated to be asymptotically normal by Hoeffding (1948). This allows us to demonstrate that under weak regularity conditions, predictions generated by these subsample ensemble methods are asymptotically normal. We also provide a method to consistently estimate the variance in the limiting distribution without increasing the computational cost so that we may produce confidence intervals and formally test feature significance in practice. Though not the focus of this

paper, it is worth noting that this subbagging procedure—suggested by Andonova et al. (2002) for use in model selection—was shown by Zaman and Hirose (2009) to outperform traditional bagging in many situations.

We consider a general supervised learning framework in which an outcome $Y \in \mathbb{R}$ is predicted as a function of d features $\boldsymbol{X} = (X_1, ..., X_d)$ by the function $\mathbb{E}[Y|\boldsymbol{X}] = F(\boldsymbol{X})$. We also allow binary classification so long as the model predicts the probability of success, as opposed to a majority vote, so that the prediction remains real valued. Additionally, we assume a training set $\{(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)\}$ consisting of n independent examples from the process that is used to produce the prediction function $\hat{F}$. Throughout the remainder of this paper, we implicitly assume that the dimension of the feature space d remains fixed, though nothing in the theory provided prohibits a growing number of features so long as our other explicit conditions on the statistical behavior of trees are met.

Statistical inference proceeds by asking the counterfactual question, "What would our results look like if we regenerated these data." That is, if a new training set was generated and we reproduced $\hat{F}$, how different might we expect the predictions to be? To illustrate, consider the hypothesis that the feature X_1 does not contribute to the outcome at any point in the feature space:

$$H_0 : \exists\, F_1 \ \ s.t. \ \ F(x_1, ..., x_d) = F_1(x_2, ..., x_d) \quad \forall\, (x_1, ..., x_d) \in \mathcal{X}$$

A formal statistical test begins by calculating a test statistic $t_0 = t((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ and asks, "If our data was generated according to H_0 and we generated a new training set and recalculated t, what is the probability that this new statistic would be larger than t_0?" That is, we are interested in estimating $P(t > t_0 | H_0)$. In most fields, a probability of less than 0.05 is considered sufficient evidence to reject the assumption that the data was generated according to H_0. Of course, a 0.05 chance can be obtained by many methods (tossing a biassed coin, for example) so we also seek a statistic t such that when H_0 is false, we are likely to reject. This probability of correctly rejecting H_0 is known as the power of the test, with more powerful tests clearly being more useful.

Here we propose to conduct the above test by comparing predictions generated by $\hat{F}$ and $\hat{F}_1$. Before doing so however, we consider the simpler hypothesis involving the value of a prediction:

$$H_0' : F(x_1, \ldots, x_p) = f_0.$$

Though often of less scientific importance, hypotheses of this form allow us to generate confidence intervals for predictions. These intervals are defined to be those values of f_0 for which we do not have enough evidence to reject H_0'. In practice, we choose f_0 to be the prediction $\hat{F}(x_1, ..., x_p)$ generated by the ensemble method in order to provide a formalized notion of plausible values of the prediction, which is, of course, of genuine interest. Our results begin here because the statistical machinery we develop will provide a distribution for the values of the prediction. This allows us to address H_0', after which we can combine these tests to address hypotheses like H_0.

Although this form of statistical analysis is ubiquitous in scientific literature, it is worthwhile contrasting this form of analysis with an alternative based on probably approximately correct (PAC) theory, as developed by Vapnik and Chervonenkis (1971) and Valiant (1984).

PAC theory provides a uniform bound on the difference between the true error and observed training error of a particular estimator, also referred to as a hypothesis. In this framework, $err(F)$ is some error of the function F which is estimated by $\widehat{err}(F)$ based on the data. A bound is then found for $P(\sup_{F \in \mathcal{F}} |\widehat{err}(F) - err(F)| > \epsilon)$ where $\mathcal{F}$ is some class of functions that includes $\hat{F}$. Since this bound is uniform over $\mathcal{F}$, it applies to $\hat{F}$ and we might think of comparing $\widehat{err}(\hat{F})$ with $\widehat{err}(\hat{F}_1)$ using such bounds. While appealing, these bounds provide the accuracy of our estimate of $err(\hat{F})$ but do not account for how the true $err(\hat{F})$ might change when $\hat{F}$ is reproduced with new training data. The uniformity of these bounds could be used to account for the uncertainty in $\hat{F}$ if it is chosen to minimize $\widehat{err}(F)$ over $\mathcal{F}$, but this is not always the case, for example, when using tree-based methods. We also expect the same uniformity to make PAC bounds conservative, thereby resulting in tests with lower power than those we develop.

Our analysis relies on the structure of subsample-based ensemble methods, specifically making use of classic U-statistic theory. These estimators have a long history (see, for example, original work by Kendall (1938) or Wilcoxon (1945), or Lee (1990) which has a modern overview), frequently focussed on rank-based non-parametric tests, and have been shown to have an asymptotically normal sampling distribution by Hoeffding (1948). Our application to subsample ensembles requires the extension of these results to some new cases as well as methods to estimate the asymptotic variance, both of which we provide.

U-statistics have traditionally been employed in the context of statistical parameter estimation. From this classical statistical perspective, we treat ensemble-tree methods like bagging and random forests as estimators and thus the limiting distributions and inference procedures we develop are with respect to the expected prediction generated by the ensemble. That is, given a particular prediction point $\boldsymbol{x}^*$, our limiting normal distributions are centered at the expected ensemble-based prediction at $\boldsymbol{x}^*$ and not necessarily $F(\boldsymbol{x}^*)$. Such forms of distributional analysis are common in other nonparametric regression settings—see Eubank (1999) Section 4.8, for example. More details on appropriate interpretations of the results are provided throughout the paper, in particular in Section 4.1.

In order to claim that the inference procedures proposed here are asymptotically valid for $F(\boldsymbol{x}^*)$, the ensemble must consistently predict $F(\boldsymbol{x}^*)$ at a rate of $\sqrt{n}$ or faster. Though this is the case for many classical estimators, establishing fast uniform rates of convergence for tree-based ensembles has proven extremely difficult. Breiman et al. (1984) discuss consistency of general partition-type models in the final chapter of their seminal book in the context of both classification and regression. Biau et al. (2008) restrict their attention to classification, but prove consistency of certain idealized bagging and random forest estimators, provided the individual trees are consistent. This paper also discusses a more general version of bagging, where the samples used to construct individual base learners may be proper subsamples of the training set taken with replacement as opposed to full bootstrap samples, so as to include the subbagging approach. Biau (2012) further examines the consistency of random forests and investigates their behavior in the presence of a sparse feature space. Recently, Denil et al. (2013) proved consistency for a mathematically tractable variant of random forests and in some cases, achieved empirical performance on par with the original random forest procedure suggested by Breiman. Zhu et al. (2015) prove consistency for their Reinforcement Learning Trees, where embedded random forests are used to decide splitting variables, and achieve significant improvements in empirical MSE for some data

sets. However, no rates of convergence have been developed that could be applied to analyze the ensemble methods we consider here.

Beyond these consistency efforts, mathematical analyses of ensemble learners has been somewhat limited. Sexton and Laake (2009) propose estimating the standard error of bagged trees and random forests using jackknife and bootstrap estimators. Recently, Wager et al. (2014) proposed applying the jackknife and infinitesimal jackknife procedures introduced by Efron (2014) for estimating standard errors in random forest predictions. Chipman et al. (2010) have received significant attention for developing *BART*, a Bayesian "sum-of-trees" statistical model for the underlying regression function that allows for pointwise posterior inference throughout the feature space as well as estimates for individual feature effects. Recently, Bleich et al. (2014) extended the BART approach by suggesting a permutation-based approach for determining feature relevance and by introducing a procedure to allow variable importance information to be reflected in the prior.

The layout of this paper is as follows: we demonstrate in Section 2 that ensemble methods based on subsampling can be viewed as U-statistics. In Section 3 we provide consistent estimators of the limiting variance parameters so that inference may be carried out in practice. Inference procedures, including a test of significance for features, are discussed in Section 4. Simulations illustrating the limiting distributions and inference procedures are provided in Section 5 and the inference procedures are applied to a real data set provided by Cornell University's Lab of Ornithology in Section 6.

2. Ensemble Methods as U-statistics

We begin by introducing the subbagging and subsampled random forest procedures that result in estimators in the form of U-statistics. In both cases, we provide an algorithm to make the procedure explicit.

2.1 Subbagging

We begin with a brief introduction to U-statistics; see Lee (1990) for a more thorough treatment. Let $Z_1, ..., Z_n \overset{iid}{\sim} F_{Z,\theta}$ where θ is the parameter of interest and suppose that there exists an unbiased estimator h of θ that is a function of $k \leq n$ arguments. Then we can write

$$\theta = \mathbb{E}h(Z_1, ..., Z_k)$$

and without loss of generality, we may further assume that h is permutation symmetric in its arguments since any given h may be replaced by an equivalent permutation symmetric version. The minimum variance unbiased estimator for θ is given by

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

where the sum is taken over all $\binom{n}{k}$ subsamples of size k and is referred to as a U-statistic with kernel h of rank k. When both the kernel and rank remain fixed, Hoeffding (1948) showed that these statistics are asymptotically normal with limiting variance $\frac{k^2}{n}\zeta_{1,k}$ where

$$\zeta_{1,k} = cov(h(Z_1, ..., Z_k), h(Z_1, Z_2^{'}, ..., Z_k^{'})) \tag{2}$$

and $Z_2^{'}, ..., Z_k^{'} \overset{iid}{\sim} F_{Z,\theta}$. The 1 in the subscript comes from the fact that there is 1 example in common between the two subsamples. In general, $\zeta_{c,k}$ denotes a covariance in the form of (2) with c examples in common.

Given infinite computing power and a consistent estimate of $\zeta_{1,k}$, Hoeffding's original result is enough to produce a subbagging procedure with asymptotically normal predictions. Suppose that as our training set, we observe $Z_1 = (\boldsymbol{X}_1, Y_1), ..., Z_n = (\boldsymbol{X}_n, Y_n) \overset{iid}{\sim} F_{\boldsymbol{X},Y}$ where $\boldsymbol{X} = (X_1, ..., X_d)$ is a vector of features and $Y \in \mathbb{R}$ is the response. Fix $k \le n$ and let $(\boldsymbol{X}_{i_1}, Y_{i_1}), ..., (\boldsymbol{X}_{i_k}, Y_{i_k})$ be a subsample of the training set. Given a feature vector $\boldsymbol{x}^* \in \mathcal{X}$ where we are interested in making a prediction, we can write the prediction at $\boldsymbol{x}^*$ generated by a tree that was built using the subsample $(\boldsymbol{X}_{i_1}, Y_{i_1}), ..., (\boldsymbol{X}_{i_k}, Y_{i_k})$ as a function $T_{\boldsymbol{x}^*}$ from $(\mathcal{X} \times \mathbb{R}) \times \cdots \times (\mathcal{X} \times \mathbb{R})$ to $\mathbb{R}$. Taking all $\binom{n}{k}$ subsamples, building a tree and predicting at $\boldsymbol{x}^*$ with each, we can write our final subbagged prediction at $\boldsymbol{x}^*$ as

$$b_n(\boldsymbol{x}^*) = \frac{1}{\binom{n}{k}} \sum_{(i)} T_{\boldsymbol{x}^*}((\boldsymbol{X}_{i_1}, Y_{i_1}), ..., (\boldsymbol{X}_{i_k}, Y_{i_k})). \tag{3}$$

by averaging the $\binom{n}{k}$ tree-based predictions. Treating each ordered pair as one of k inputs into the function $T_{\boldsymbol{x}^*}$, the estimator in (3) is in the form of a U-statistic since tree-based estimators produce the same predictions independent of the order of the training data. Thus, provided the distribution of predictions at $\boldsymbol{x}^*$ has a finite second moment and $\zeta_{1,k} > 0$, the distribution of subbagged predictions at $\boldsymbol{x}^*$ is asymptotically normal. Note that in this context, $\zeta_{1,k}$ is the covariance between predictions at $\boldsymbol{x}^*$ generated by trees trained on data sets with 1 sample in common.

Of course, building $\binom{n}{k}$ trees is compuationally infeasible for even moderately sized training sets and an obvious substantial improvement in computationally efficiency can be achieved by building and averaging over only $m_n < \binom{n}{k}$ trees. In this case, the estimator in (3), appropriately scaled, is called an *incomplete* U-statistic. When the m_n subsamples are selected uniformly at random with replacement from the $\binom{n}{k}$ possibilities, the resulting incomplete U-statistic remains asymptotically normal; see Janson (1984) or Lee (1990) page 200 for details.

Though more computationally efficient, there remains a major shortcomming with this approach: the number of samples used to build each tree, k, remains fixed as $n \to \infty$. We would instead like k to grow with n so that trees can be grown to a greater depth, thereby presumably producing more accurate predictions. Incorporating this, our estimator becomes

$$b_{n,k_n,m_n}(\boldsymbol{x}^*) = \frac{1}{m_n} \sum_{(i)} T_{\boldsymbol{x}^*, k_n}((\boldsymbol{X}_{i_1}, Y_{i_1}), ..., (\boldsymbol{X}_{i_{k_n}}, Y_{i_{k_n}})). \tag{4}$$

Statistics of this form were discussed by Frees (1989) and called *Infinite Order* U-statistics (IOUS) in the complete case, when $m_n = \binom{n}{k_n}$, and *resampled* statistics in the incomplete case. Specifically, Frees considers the situation where, given an *i.i.d* sample $Z_1, Z_2, ...$ and kernel h_{k_n}, $\lim_{n \to \infty} h_{k_n}(Z_{i_1}, ..., Z_{i_{k_n}}) = h(Z_1, Z_2, ...)$ and $\theta = \mathbb{E}h(Z_1, Z_2, ...)$

and goes on to develop sufficient conditions for consistency and asymptotic normality whenever m_n grows faster than n. In contrast, the theorem below introduces a central limit theorem for estimators of the same form as in (4) but with respect to their individual means $\mathbb{E}b_{n,k_n,m_n}(\boldsymbol{x}^*)$ and covers all possible growth rates of m_n with respect to n. In this context, only minimal regularity conditions are required for asymptotic normality. We begin with an assumption on the distribution of estimates for the general U-statistic case.

Condition 1: Let $Z_1, Z_2, ... \overset{iid}{\sim} F_Z$ with $\theta_{k_n} = \mathbb{E}h_{k_n}(Z_1, ..., Z_{k_n})$ and define $h_{1,k_n}(z) = \mathbb{E}h_{k_n}(z, Z_2, ..., Z_{k_n}) - \theta_{k_n}$. Then for all $\delta > 0$,

$$\lim_{n \to \infty} \frac{1}{\zeta_{1,k_n}} \int_{|h_{1,k_n}(Z_1)| \geq \delta\sqrt{n\zeta_{1,k_n}}} h_{1,k_n}^2(Z_1)dP = 0.$$

This condition serves to control the tail behavior of the predictions and allows us to satisfy the Lindeberg condition needed to obtain part *(i)* of Theorem 1 below.

Theorem 1 Let $Z_1, Z_2, ... \overset{iid}{\sim} F_Z$ and let U_{n,k_n,m_n} be an incomplete, infinite order U-statistic with kernel h_{k_n} that satisfies Condition 1. Let $\theta_{k_n} = \mathbb{E}h_{k_n}(Z_1, ..., Z_{k_n})$ such that $\mathbb{E}h_{k_n}^2(Z_1, ..., Z_{k_n}) \leq C < \infty$ for all n and some constant C, and let $\lim \frac{n}{m_n} = \alpha$. Then as long as $\lim \frac{k_n}{\sqrt{n}} = 0$ and $\lim \zeta_{1,k_n} \neq 0$,

(i) if $\alpha = 0$, then $\frac{\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{k_n^2 \zeta_{1,k_n}}} \overset{d}{\to} \mathcal{N}(0,1)$.

(ii) if $0 < \alpha < \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\frac{k_n^2}{\alpha}\zeta_{1,k_n} + \zeta_{k_n,k_n}}} \overset{d}{\to} \mathcal{N}(0,1)$.

(iii) if $\alpha = \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\zeta_{k_n,k_n}}} \overset{d}{\to} \mathcal{N}(0,1)$.

Condition 1, though necessary for the general U-statistic setting, is a bit obscure. However, in our regression context, when the regression function is bounded and the errors have exponential tails, a more intuitive Lipschitz-type condition given in Proposition 1 is sufficient. Though stronger than necessary, this alternative condition allows us to satisfy the Lindeberg condition and is reasonable to expect of any supervised learning method.

Proposition 1: For a bounded regression function F, if there exists a constant c such that for all $k_n \geq 1$,

$$\Big|h((\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}), (\boldsymbol{X}_{k_n+1}, Y_{k_n+1})) - h((\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}), (\boldsymbol{X}_{k_n+1}, Y_{k_n+1}^*))\Big|$$
$$\leq c\big|Y_{k_n+1} - Y_{k_n+1}^*\big|$$

where $Y_{k_n+1} = F(\boldsymbol{X}_{k_n+1}) + \epsilon_{k_n+1}$, $Y_{k_n+1}^* = F(\boldsymbol{X}_{k_n+1}) + \epsilon_{k_n+1}^*$, and where ϵ_{k_n+1} and $\epsilon_{k_n+1}^*$ are i.i.d. with exponential tails, then Condition 1 is satisfied.

A number of important aspects of these results are worth pointing out. First, note from Theorem 1 that the trees are built with subsamples that are approximately square root of the size of the full training set. This condition is not necessary for the proof, but ensures that the variance of the U-statistic in part (i) converges to 0 as is typically the case in central limit theorems. By maintaining this relatively small subsample size, we can build many more trees and maintain a procedure that is computationally equivalent to traditional bagging based on full bootstrap samples. Also note that no particular assumptions are placed on the dimension d of the feature space; the number of features may grow with n so long as the stated conditions remain satisfied.

The final condition of Theorem 1, that $\lim \zeta_{1,k_n} \neq 0$, though not explicitly controllable, should be easily satisfied in many cases. As an example, suppose that the terminal node size is bounded by $\mathcal{T}$ so that trees built with larger training sets are grown to greater depths. Then if the form of the response is $Y = F(\boldsymbol{X}) + \epsilon$ where ϵ has variance σ^2, ζ_{1,k_n} will be bounded below by $\sigma^2/\mathcal{T}$. Finally, note that the assumption of exponential tails on the distribution of regression errors in Proposition 1 is stronger than necessary. Indeed, so long as $k_n = o(\sqrt{n})$, we need only insist that $nP\left(|\epsilon| > \sqrt{n}\right) \to 0$.

The proofs of Theorem and Proposition 1 are provided in Appendix A. The subbagging algorithm that produces asymptotically normal predictions at each point in the feature space is provided in Algorithm 1.

Algorithm 1 Subbagging

Load training set
Select size of subsamples k_n and number of subsamples m_n
for i in 1 to m_n **do**
    Take subsample of size k_n from training set
    Build tree using subsample
    Use tree to predict at $\boldsymbol{x}^*$
end for
Average the m_n predictions to get final estimate $b_{n,k_n,m_n}(\boldsymbol{x}^*)$

Note that this procedure is precisely the original bagging algorithm suggested by Breiman, but with proper subsamples used to build trees instead of full bootstrap samples. In Section 3, we provide consistent estimators for the limiting variance parameters in Theorem 1 so that we may carry out inference in practice.

We would also like to acknowledge similar work currently in progress by Wager (2014). Wager builds upon the potential nearest neighbor framework introduced by Lin and Jeon (2006) and seeks to provide a limiting distribution for the case where many trees are used in the ensemble, roughly corresponding to our result (i) in Theorems 1 and 2. The author considers only an idealized class of trees based on the assumptions in Meinshausen (2006) as well as additional *honesty* and *regularity* conditions that allow k_n to grow at a faster rate, and demonstrates that when many Monte Carlo samples are employed, the infinitesimal jackknife estimator of variance is consistent and predictions are asymptotically normal. This estimator has roughly the same computational complexity as those we propose in Section 3 and should scale well subject to some additional bookkeeping. In contrast, the theory we provide here takes into account all possible rates of Monte Carlo sampling via the

three cases discussed in Theorems 1 and 2 and we provide a consistent means for estimating each corresponding variance.

2.2 Random Forests

The distributional results described above for subbagging do not insist on a particular tree building method. So long as the trees generate predictions that satisfy minimal regularity conditions, the experimenter is free to use whichever building method is preferred. The subbagging procedure does, however, require that each tree in the ensemble is built according to the same method.

This insistence on a uniform, non-randomized building method is in contrast with random forests. The original random forests procedure suggested by Breiman (2001b) dictates that at each node in each tree, the split may occur on only a randomly selected subset of features. Thus, we may think of each tree in a random forest as having an additional randomization parameter ω that determines the eligible features that may potentially be split at each node. In a general U-statistic context, we can write this *random kernel* U-statistic as

$$U_{\omega;n,k_n,m_n} = \frac{1}{m_n} \sum_{(i)} h_{k_n}^{(\omega_i)}(Z_{i_1}, ..., Z_{i_{k_n}}) \tag{5}$$

so that we can write a random forest estimator as

$$r_{n,k_n,m_n}(\boldsymbol{x}^*) = \frac{1}{m_n} \sum_{(i)} T_{\boldsymbol{x}^*,k_n}^{(\omega_i)}((\boldsymbol{X}_{i_1}, Y_{i_1}), ..., (\boldsymbol{X}_{i_{k_n}}, Y_{i_{k_n}})).$$

Due to this additional randomness, random forests and random kernel U-statistics in general do not fit within the framework developed in the previous section so we develop new theory for this expanded class. Suppose $\omega_1, ..., \omega_{m_n} \overset{iid}{\sim} F_\omega$ and that these randomization parameters are selected independently of the original sample $Z_1, ..., Z_n$. Consider the statistic

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

so that $U_{\omega;n,k_n,m_n}^* = \mathbb{E}_\omega U_{\omega;n,k_n,m_n}$. Taking the expectation with respect to ω, the kernel becomes fixed and hence $U_{\omega;n,k_n,m_n}^*$ conforms to the non-random kernel U-statistic theory. Thus, $U_{\omega;n,k_n,m_n}^*$ is asymptotically normal in both the complete and incomplete cases, as well as in the complete and incomplete infinite order cases, by Theorem 1. Given this asymptotic normality of $U_{\omega;n,k_n,m_n}^*$, in order to retain asymptotic normality of the corresponding random kernel version, we need only show that

$$\sqrt{n}(U_{\omega;n,k_n,m_n}^* - U_{\omega;n,k_n,m_n}) \overset{P}{\to} 0.$$

We make use of this idea in the proof of the following theorem, which is provided in Appendix A.

Theorem 2 *Let $U_{\omega;n,k_n,m_n}$ be a random kernel U-statistic of the form defined in equation (5) such that $U^*_{\omega;n,k_n,m_n}$ satisfies Condition 1 and suppose that $\mathbb{E}h^2_{k_n}(Z_1, ..., Z_{k_n}) < \infty$ for all n, $\lim \frac{k_n}{\sqrt{n}} = 0$, and $\lim \frac{n}{m_n} = \alpha$. Then, letting β index the subsamples, so long as $\lim \zeta_{1,k_n} \neq 0$ and*

$$\lim_{n \to \infty} \mathbb{E}\left(h^{(\omega)}_{k_n}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h^{(\omega)}_{k_n}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2 \neq \infty,$$

$U_{\omega;n,k_n,m_n}$ is asymptotically normal and the limiting distributions are the same as those provided in Theorem 1.

Note that the variance parameters ζ_{1,k_n} and ζ_{k_n,k_n} in the context of random kernel U-statistics are still defined as the covariance between estimates generated by the (now random) kernels. Thus, in the specific context of random forests, these variance parameters correspond to the covariance between predictions generated by trees, but each tree is built according to its own randomization parameter ω and this covariance is taken over ω as well. The final condition of Theorem 2 that

$$\lim_{n \to \infty} \mathbb{E}\left(h^{(\omega)}_{k_n}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h^{(\omega)}_{k_n}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2 \neq \infty,$$

simply ensures that the randomization parameter ω does not continually pull predictions from the same subsample further apart as $n \to \infty$. This condition is satisfied, for example, if the response Y is bounded and should also be easily satisfied for any reasonable implementation of random forests.

The subsampled random forest algorithm that produces asymptotically normal predictions is provided in Algorithm 2. As with subbagging, this subsampled random forest algorithm is exactly a random forest with subsamples used to build trees instead of full bootstrap samples.

Algorithm 2 Subsampled Random Forest

    Load training set
    Select size of subsamples k_n and number of subsamples m_n
    for i in 1 to m_n **do**
        Select subsample of size k_n from training set
        Build tree based on randomization parameter ω_i
        Use this tree to predict at $\boldsymbol{x}^*$
    end for
    Average the m_n predictions to obtain final estimate $r_{n,k_n,m_n}(\boldsymbol{x}^*)$

Another random-forest-type estimator based on a crossed design that results in an infinite order *generalized* U-statistic is provided in Appendix B. However, the above formulation most resembles Breiman's original procedure and is more computationally feasible than the method mentioned in Appendix B, so we consider only this random kernel version of random forests in the simulations and other work that follows.

3. Estimating the Variance

The limiting distributions provided in Theorem 1 depend on the unknown mean parameter $\theta_{k_n} = \mathbb{E}U_{n,k_n,m_n}$ as well as the unknown variance parameters ζ_{1,k_n} and ζ_{k_n,k_n}. In order for us to be able to use these distributions for statistical inference in practice, we must establish consistent estimators of these parameters. It is obvious that we can use the sample mean—i.e. the prediction from our ensemble—as a consistent estimate of θ_{k_n}, but determining an appropriate variance estimate is less straightforward.

In equation (2) of the previous section, we defined ζ_{c,k_n} as the covariance between two instances of the kernel with c shared arguments, so the sample covariance between predictions may serve as a consistent estimator for both ζ_{1,k_n} and ζ_{k_n,k_n}. However, in practice we find that this often results in estimates close to 0, which may then lead to an overall negative variance estimate.

It is not difficult to show - see Lee (1990) page 11 for details - that an equivalent expression for ζ_{c,k_n} is given by

$$\zeta_{c,k_n} = var\Big(\mathbb{E}\big(h_{k_n}(Z_1, ..., Z_{k_n}) \mid Z_1 = z_1, ..., Z_c = z_c\big)\Big).$$

To estimate ζ_{c,k_n} for our tree-based ensembles, we begin by selecting c observations $\tilde{z}_1, ..., \tilde{z}_c$, which we refer to as initial fixed points, from the training set. We then select several subsamples of size k_n from the training set, each of which must include $\tilde{z}_1, ..., \tilde{z}_c$, build a tree with each subsample, and record the mean of the predictions at x^*. Let n_{MC} (MC for "Monte Carlo") denote the number of subsamples drawn so that this average is taken over n_{MC} predictions. We then repeat the process for $n_{\tilde{z}}$ initial sets of fixed points and take our final estimate of ζ_{c,k_n} as the variance over the $n_{\tilde{z}}$ final averages, yielding the estimator

$$\hat{\zeta}_{c,k_n} = var\left(\frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} T_{x^*,k_n}(\mathcal{S}_{\tilde{z}^{(1)},i}) , ..., \frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} T_{x^*,k_n}(\mathcal{S}_{\tilde{z}^{(n_{\tilde{z}})},i}) \right)$$

where $\tilde{z}^{(j)}$ denotes the jth set of initial fixed points and $\mathcal{S}_{\tilde{z}^{(j)},i}$ denotes the ith subsample that includes $\tilde{z}^{(j)}$ (which is used here as shorthand for the argument to the tree function T_{x^*,k_n}). Now, since we assume that the orginal data in the training set is i.i.d., the random variables $\frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} T_{x^*,k_n}(\mathcal{S}_{\tilde{z}^{(1)},i})$ are also i.i.d. and since the sample variance is a U-statistic, $\hat{\zeta}_{c,k_n}$ is a consistent estimator. The algorithm for calculating $\hat{\zeta}_{1,k_n}$ is provided in Algorithm 3. Note that when $c = k_n$, each of the subsamples is identical so we need only use $n_{MC} = 1$ which simplifies the estimation procedure for ζ_{k_n,k_n}. The procedure for calculating $\hat{\zeta}_{k_n,k_n}$ is provided in Algorithm 4.

Choosing the values of $n_{\tilde{z}}$ and n_{MC} will depend on the situation. The number of iterations required to accurately estimate the variance depends on a number of factors, including the tree building method and true underlying regression function. Of course, ideally these estimation parameters should be chosen as large as is computationally feasible. In our simulations, we find that in most cases, only a relatively small number of initial fixed point sets are needed, but many more Monte Carlo samples are often necessary for accurate estimation. In most cases, we used an n_{MC} of at least 500. Recall that because our trees are built with small subsamples, we can build correspondingly more trees at the same computational cost.

Algorithm 3 ζ_{1,k_n} Estimation Procedure

---

    for i in 1 to $n_{\tilde{z}}$ **do**
        Select initial fixed point $\tilde{z}^{(i)}$
        for j in 1 to n_{MC} **do**
            Select subsample $\mathcal{S}_{\tilde{z}^{(i)},j}$ of size k_n from training set that includes $\tilde{z}^{(i)}$
            Build tree using subsample $\mathcal{S}_{\tilde{z}^{(i)},j}$
            Use tree to predict at $\boldsymbol{x}^*$
        end for
        Record average of the n_{MC} predictions
    end for
    Compute the variance of the $n_{\tilde{z}}$ averages

Algorithm 4 ζ_{k_n,k_n} Estimation Procedure

---

    for i in 1 to $n_{\tilde{z}}$ **do**
        Select subsample of size k_n from training set
        Build tree using subsample this subsample
        Use tree to predict at $\boldsymbol{x}^*$
    end for
    Compute the variance of the $n_{\tilde{z}}$ predictions

3.1 Internal vs. External Estimation

The algorithms for producing the subbagged or subsampled random forest predictions as well as the above algorithms for estimating the variance parameters are all that is needed to perform statistical inference. We can begin with Algorithm 1 or 2 to generate the predictions, followed by Algorithms 3 and 4 to estimate the variance parameters ζ_{1,k_n} and ζ_{k_n,k_n}. This procedure of running these 3 algorithms seperately is what we will refer to as the *external* variance estimation method, since the the variance parameters are estimated outside of the orginal ensemble. By contrast, we could instead generate the predictions and estimate the variance parameters in one procedure by taking the mean and variance of the predictions generated by the trees used to estimate ζ_{1,k_n}. Algorithm 5 outlines the steps in this *internal* variance estimation method.

This internal variance estimation method is more computationally efficient and has the added benefit of producing variance estimates by simply changing the way in which the subsamples are selected. This means that we are able to obtain all parameter estimates we need to conduct inference at no greater computational cost than building the original ensemble. Although Theorems 1 and 2 dictate that the subsamples used in the ensemble be selected uniformly at random, we find that the additional correlation introduced by selecting the subsamples in this way and using the same subsamples to estimate all parameters does not affect the limiting distribution.

Algorithm 5 Internal Variance Estimation Method

---

    for i in 1 to $n_{\tilde{z}}$ **do**

       Select initial fixed point $\tilde{z}^{(i)}$

       for j in 1 to n_{MC} **do**

          Select subsample $\mathcal{S}_{\tilde{z}^{(i)},j}$ of size k_n from training set that includes $\tilde{z}^{(i)}$

          Build tree using subsample $\mathcal{S}_{\tilde{z}^{(i)},j}$

          Use tree to predict at $\boldsymbol{x}^*$ and record prediction

       end for

       Record average of the n_{MC} predictions

    end for

    Compute the variance of the $n_{\tilde{z}}$ averages to estimate ζ_{1,k_n}

    Compute the variance of all predictions to estimate ζ_{k_n,k_n}

    Compute the mean of all predictions to estimate θ_{k_n}

---

4. Inference Procedures

In this section, we describe the inference procedures that may be carried out after performing the estimation procedures.

4.1 Confidence Intervals

In Section 2, we showed that predictions from subbagging and subsampled random forests are asymptotically normal and in Section 3 we provided consistent estimators for the parameters in the limiting normal distributions. Thus, given a training set, we can estimate the approximate distribution of predictions at any given feature vector of interest $\boldsymbol{x}^*$. To produce a confidence interval for predictions at $\boldsymbol{x}^*$, we need only estimate the variance parameters and take quantiles from the appropriate limiting distribution. Formally, our confidence interval is $[LB, UB]$ where the lower and upper bounds, LB and UB, are the $\alpha/2$ and $1 - \alpha/2$ quantiles respectively of the normal distribution with mean $\hat{\theta}_{k_n}$ and variance $\frac{k_n^2}{\hat{\alpha}}\hat{\zeta}_{1,k_n} + \hat{\zeta}_{k_n,k_n}$ where $\hat{\zeta}_{1,k_n}$ and $\hat{\zeta}_{k_n,k_n}$ are the variance estimates and $\hat{\alpha} = n/m_n$. This limiting distribution is that given in result (ii) of Theorem 1 which is the distribution we recommend using in practice.

As mentioned in the introduction, these confidence intervals can also be used to address hypotheses of the form

$$
\begin{aligned}
H_0 &: \ \theta_{k_n} = c \\
H_1 &: \ \theta_{k_n} \neq c.
\end{aligned}
$$

Formally, we can define the test statistic

$$
t = \frac{\hat{\theta}_{k_n} - c}{sd(\hat{\theta}_{k_n})}
$$

and reject H_0 if $|t|$ is greater than the the $1 - \frac{\alpha}{2}$ quantile of the standard normal. This corresponds to a test with type 1 error rate α so that $P[\text{reject } H_0 \mid H_0 \text{ true}] = P[\ |t| >$

$1 - \frac{\alpha}{2} \mid \theta_{k_n} = c] = \alpha$. However, this testing procedure is equivalent to simply checking whether c is within the calculated confidence interval: if c is in the confidence interval, then we fail to reject this hypothesis that the true mean prediction is equal to c, otherwise we reject.

Finally, recall that these confidence intervals are for the expected prediction θ_{k_n} and not necessarily for the true value of the underlying regression function $\theta = F(\boldsymbol{x}^*)$. If the tree building method employed is consistent so that $\theta_{k_n} \xrightarrow{P} \theta$, then as the sample size increases, the tree should be (on average) producing more accurate predictions, but in order to claim that our confidence intervals are asymptotically valid for θ, we need for this convergence to occur at rate of $\sqrt{n}$ or faster. However, in general, the rate of convergence will depend on not only the tree-building method and true underlying regression function, but also on the location of the prediction point within the feature space.

Note that Theorems 1 and 2 apply not only to tree-based ensembles, but to any estimator that can be written in the form of an infinite order U-statistic, as in equation (4). Some of these ensembles may be straightforward to analyze, but for others, such as random forest predictions near the edge of the feature space, it may be difficult to establish a universal rate of consistency. However, even when $\sqrt{n}$-consistency cannot be guaranteed, these intervals still provide valuable information not currently available with existing tools. In these cases, the confidence interval provides a reasonable range of values for where the prediction might fall if the ensemble was recomputed using a new training set; areas of the feature space where confidence intervals are relatively large indicate regions where the ensemble is particularly unstable.

Compare this, for example, to the standard approach of withholding some (usually small) portion of the training set and comparing predictions made at these hold-out points to the corresponding known responses. Such an approach provides some information as to the *accuracy* of the learner at specific locations throughout the feature space, but says nothing about the *stability* of these predictions. Thus, instead of relying only on measures of overall goodness-of-fit such as MSE or SSE, these intervals allow users to investigate prediction variability at particular points or regions and in this sense, can be seen as a measure of how much the accuracy of predictions at that point is due to chance.

4.2 Tests of Significance

The limiting distributions developed in Theorems 1 and 2 also allow us a way to test the significance of features. In many situations, data are recorded for a large number of features but a sparse true regression structure is suspected. Suppose that the training set consists of d features, $X_1, ..., X_d$ and consider a reduced set $\boldsymbol{X}^{(R)} \subset \{X_1, ..., X_d\}$. Let $\boldsymbol{x}_{\text{TEST}} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ be a set of feature vectors where we are interested in making predictions. Also, let g denote the function that maps feature vectors to their corresponding true mean prediction and let $g^{(R)}$ denote the same type of function that maps from the reduced feature space. That is, for a particular prediction point of interest $\boldsymbol{x}^*$, $g(\boldsymbol{x}^*)$ is the true mean prediction θ_{k_n} generated by trees built using the full feature space, and $g^{(R)}(\boldsymbol{x}^*)$ is the true mean prediction $\theta_{k_n}^{(R)}$ generated by trees that are only permitted to utilize features in the reduced set. Then, for each test point $\boldsymbol{x}_i \in \boldsymbol{x}_{\text{TEST}}$, we would like to know whether

$g(\boldsymbol{x}_i) = g^{(R)}(\boldsymbol{x}_i)$ so that we can determine the predictive influence of features not in $\boldsymbol{X}^{(R)}$. More formally, we would like to test the hypothesis

$$H_0 : g(\boldsymbol{x}_i) = g^{(R)}(\boldsymbol{x}_i) \ \forall \ \boldsymbol{x}_i \in \boldsymbol{x}_{\text{TEST}} \tag{6}$$
$$H_1 : g(\boldsymbol{x}_i) \neq g^{(R)}(\boldsymbol{x}_i) \ \text{ for some } \boldsymbol{x}_i \in \boldsymbol{x}_{\text{TEST}}.$$

Rejecting this null hypothesis means that a feature not in the reduced feature space $\boldsymbol{X}^{(R)}$ makes a significant contribution to the prediction at at least one of the test points.

To perform this test with a training set of size n, we take m_n subsamples, each of size k_n, and build a tree with each subsample. Denote these subsamples $S_1, ..., S_{m_n}$ and for a given feature vector $\boldsymbol{x}_i$, let $\hat{g}(\boldsymbol{x}_i)$ denote the average over the predictions at $\boldsymbol{x}_i$ generated from the m_n trees. Then, using the same subsamples $S_1, ..., S_{m_n}$, again build a tree with each, but using only those features in $\boldsymbol{X}^{(R)}$, and let $\hat{g}^{(R)}(\boldsymbol{x}_i)$ be the average prediction at $\boldsymbol{x}_i$ generated by these trees. Finally, define the difference function

$$\hat{D}(\boldsymbol{x}_i) = \hat{g}(\boldsymbol{x}_i) - \hat{g}^{(R)}(\boldsymbol{x}_i)$$

as the difference between the two ensemble predictions. Note that we can write

$$
\begin{aligned}
\hat{D}(\boldsymbol{x}_i) &= \hat{g}(\boldsymbol{x}_i) - \hat{g}^{(R)}(\boldsymbol{x}_i) \\
&= \frac{1}{m_n} \sum_{(j)} T_{\boldsymbol{x}_i, k_n}(S_j) - \frac{1}{m_n} \sum_{(j)} T^{(R)}_{\boldsymbol{x}_i, k_n}(S_j) \\
&= \frac{1}{m_n} \sum_{(j)} \left(T_{\boldsymbol{x}_i, k_n}(S_j) - T^{(R)}_{\boldsymbol{x}_i, k_n}(S_j) \right)
\end{aligned}
$$

so that this difference function is a U-statistic. Thus, if we have only a single test point of interest, $\hat{D}$ is asymptotically normal, so $\hat{D}^2$ is asymptotically χ_1^2 and we can use $\hat{D}^2$ as a test statistic.

However, it is more often the case that we have several test points of interest. In this case, define $\hat{\mathbb{D}}$ to be the vector of observed differences in the predictions

$$\hat{\mathbb{D}} = \left(\hat{D}(\boldsymbol{x}_1), ..., \hat{D}(\boldsymbol{x}_N) \right)$$

so that, provided a joint distribution exists with respect to Lebesgue measure, $\hat{\mathbb{D}}$ has a multivariate normal distribution with mean vector

$$\boldsymbol{\mu} = \left(g(\boldsymbol{x}_1) - g^{(R)}(\boldsymbol{x}_1), ..., g(\boldsymbol{x}_N) - g^{(R)}(\boldsymbol{x}_N) \right)^T$$

which we estimate with

$$\hat{\boldsymbol{\mu}} = \hat{\mathbb{D}}^T$$

as well as a covariance matrix Σ. This covariance matrix has parameters Σ_{1, k_n} and Σ_{k_n, k_n}, the multivariate analogues of ζ_{1, k_n} and ζ_{k_n, k_n}. Consistent estimators for these multivariate

parameters can be obtained by simply replacing the variance calculation in Algorithms 3 and 4 with a covariance. For clarity, the procedure for obtaining $\hat{\Sigma}_{1,k_n}$ is provided in Algorithm 6 in Appendix C.

Finally, combining these predictions to form a consistent estimator $\hat{\Sigma}$ we have that

$$\hat{\boldsymbol{\mu}}^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}} \sim \chi^2_N$$

under H_0. Thus, in order to test the hypothesis in (6), we compare the test statistic $\hat{\boldsymbol{\mu}}^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}$ to the $1 - \alpha$ quantile of the χ^2_N distribution to produce a test with type 1 error rate α. If our test statistic is larger than this critical value, we reject the null hypothesis.

4.3 Further Testing Procedures

This setup, though straightforward, may not always definitively decide the significance of features. In some cases, even randomly generated features that are unrelated to the response can be reported significant. Depending on the building method, tree-based algorithms may take advantage of additional randomness in features even when the particular values of those features do not directly contribute to the response. For this reason, we also recommend repeating the testing procedure by comparing predictions generated using the full data set to predictions generated by a data set with randomly generated values—commonly obtained by permuting the values in the training set—for the features not in the reduced feature set to test hypostheses of the form

$$H_0 : g(\boldsymbol{x}_i) = g^{(RAND)}(\boldsymbol{x}_i) \ \ \forall \, \boldsymbol{x}_i \in \boldsymbol{x}_{\text{TEST}}$$
$$H_1 : g(\boldsymbol{x}_i) \neq g^{(RAND)}(\boldsymbol{x}_i) \ \text{ for some } \boldsymbol{x}_i \in \boldsymbol{x}_{\text{TEST}}.$$

The testing procedure remains exactly the same except that to calculate the second set of trees, we simply substitute the reduced training set for a training set with the same number of features, but with randomized values taking the place of the original values for the additional features. Rejecting this null hypothesis allows us to conclude that not only do the additional features not in the reduced training set make a significant contribution to the predictions, but that the contribution is significantly more than could be obtained simply by adding additional randomness.

There are also two additional tests that may be performed. First, we can test whether predictions generated by a training set with randomized values for the additional features are significantly different from predictions generated by the reduced feature set. If a significant difference is found, then the trees in the ensemble are making use of the additional randomness or possibly an accidental structure in the randomized features. As a final check, we can compare predictions generated by two training sets, each with randomized values for the features not in the reduced set. In the unlikely event that a significant difference is found between these predictions, it is again likely due to an accidental structure in the randomized values. Both of these tests can be performed in exactly the same fashion by substituting the appropriate training sets.
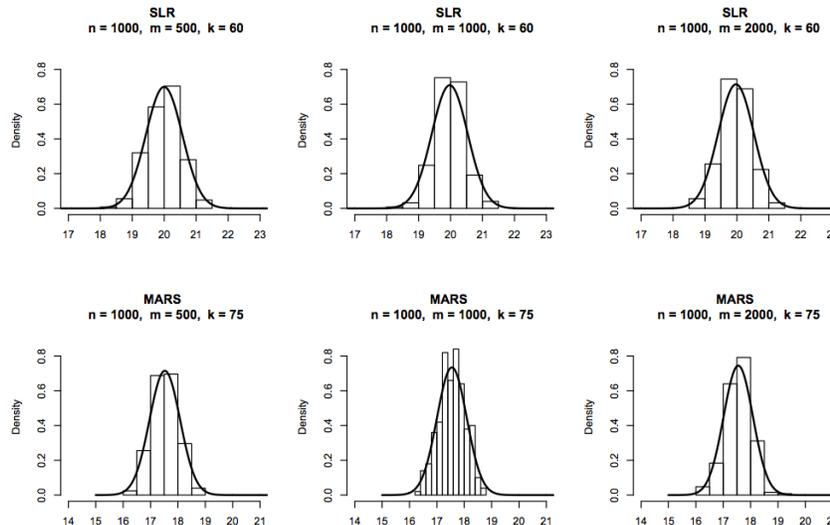
Figure 1: Histograms of subbagged predictions at $x_1 = 10$ in the SLR case (top row) and at $x_1 = \cdots = x_5 = 0.5$ in the MARS case (bottom row). The total sample size, number of subsamples, and size of each subsample are denoted by n, m, and k, respectively in the plot titles.

5. Simulations

We present here a small simulation study in order to illustrate the limiting distributions derived in Section 2 and also to demonstrate the inference procedures proposed in the previous section. We consider two different underlying regression functions:

1. $g(x_1) = 2x_1$; $\quad \mathcal{X} = [0, 20]$

2. $g(\boldsymbol{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.05)^2 + 10x_4 + 5x_5$; $\quad \mathcal{X} = [0, 1]^5$

The first function corresponds to simple linear regression (SLR) and was chosen for simplicity and ease of visualization. The second was initially considered by Friedman (1991) in development of the Multivariate Adaptive Regression Spline (MARS) procedure and was recently investigated by Biau (2012). In each case, features were selected uniformly at random from the feature spaces and responses were sampled from $g(\boldsymbol{x}) + \epsilon$, where $\epsilon \overset{iid}{\sim} \mathcal{N}(0, 10)$, to form the training sets.

5.1 Limiting Distributions

We begin by illustrating the distributions of subbagged predictions. In the SLR case, predictions were made at $x_1 = 10$ and in the MARS case, predictions were made at $x_1 = \cdots = x_5 = 0.5$. The histograms of subbagged predictions are shown in Figure 1. Each histogram is comprised of 250 simulations.

For each histogram, the size of the training set n, number of subsamples m, and size of each subsample k, is provided in the title. Each tree in the ensembles was built using the

16
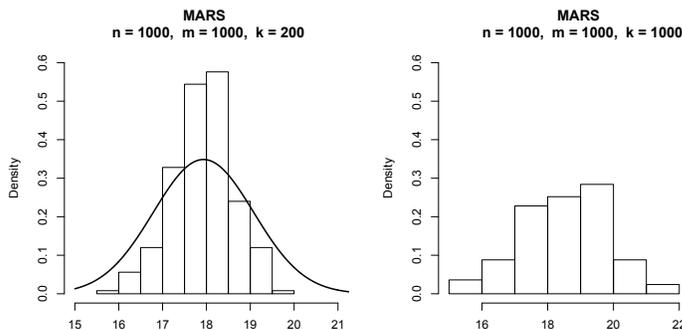
Figure 2: Histograms of subbagged predictions with larger subsample size k and full bootstrap samples. Predictions are made at $x_1 = \cdots = x_5 = 0.5$.

`rpart` function in R, with the additional restriction that at least 3 observations per node were needed in order for the algorithm to consider splitting on that node. Overlaying each histogram is the density obtained by estimating the parameters in the limiting distribution. In each case, we take the limiting distribution to be that given in result *(ii)* of Theorem 1; namely that the predictions are normally distributed with mean $\mathbb{E}b_{n,k,m}(\boldsymbol{x}^*)$ and variance $\frac{1}{\alpha}\frac{k^2}{m}\zeta_{1,k} + \frac{1}{m}\zeta_{k,k}$.

The mean $\mathbb{E}b_{n,k,m}(\boldsymbol{x}^*) = \theta_k$ was estimated as the empirical mean across the 250 subbagged predictions. To estimate $\zeta_{k,k}$, 5000 new subsamples of size k were selected and with each subsample, a tree was built and used to predict at $x_1 = 10$ and $\hat{\zeta}_{k,k}$ was taken as the empirical variance between these predictions. To estimate $\zeta_{1,k}$, we follow the procedure in Algorithm 3 with $n_{\tilde{\boldsymbol{z}}} = 50$ and $n_{MC} = 1000$ in the SLR cases and with $n_{\tilde{\boldsymbol{z}}} = 250$ and $n_{MC} = 1000$ in the MARS cases. Note that since we are only interested in verifying the distributions of predictions, the variance parameters are estimated only once for each case and not for each ensemble.

It is worth noting that the same variance estimation procedure with $n_{\tilde{\boldsymbol{z}}} = 250$ and $n_{MC} = 250$ lead to an overestimate of the variance, so we reiterate that using a large n_{MC} seems to provide better results, even when $n_{\tilde{\boldsymbol{z}}}$ is relatively small. In each case, we use $\frac{n}{m}$ as a plug-in estimate for $\alpha = \lim \frac{n}{m}$. We also repeated this procedure and generated the distribution of predictions according to the internal variance estimation method described in Algorithm 5. Details and histograms are provided in Appendix D. These distributions appear to be the same as when the subsamples are selected uniformly at random, as in the external variance estimation method.

Note that the distributional results in Theorem 1 require $\lim \frac{k}{\sqrt{n}} = 0$, so in practice, the subsample size k should be small relative to n. In the above simulations, we choose k slightly larger than $\sqrt{n}$ and the distributions appear normally distributed with the correct limiting distribution. However, though this restriction on the growth rate of the subsample size is sufficient for asymptotic normality, it is perhaps not necessary. In our simulations, we found that ensembles built with larger k are still approximately normal, but begin to

17
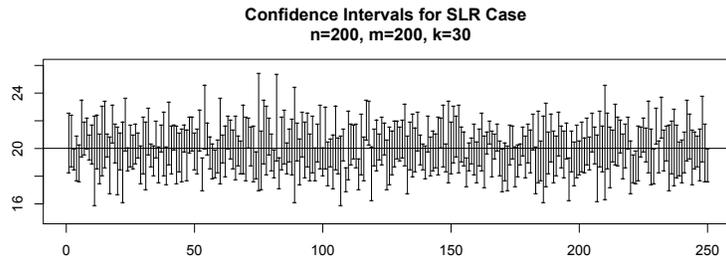
Figure 3: Confidence Intervals for subbagged predictions.

look increasingly further from normal as k increases. The histograms in Figure 2 show the distribution of subbagged predictions in the MARS case with $n = m = 1000$ and $k = 200$ and also with $n = m = k = 1000$ so that we are using full bootstrap samples to build the ensembles in the latter case. The parameters in the limiting distribution are estimated in exactly the same manner as with the smaller k for the case where $k = 200$. In the bootstrap case, we cannot follow our subbagging procedure exactly since the bootstrap samples used to build each tree in the ensemble must be taken with replacement, so we do not attempt to estimate the variance. These distributions look less normal and we begin to overestimate the variance in the case where $k = 200$.

5.2 Confidence Intervals

We move now to building confidence intervals for predictions and examine their coverage probabilities. We begin with the SLR case, with $n = 200$, $m = 200$, and $k = 30$ and as above, predict at $x_1 = 10$. To build the confidence intervals, we generate 250 data sets and with each data set, we produce a subbagged ensemble, estimate the parameters in the limiting distribution, and take the 0.025 and 0.975 quantiles from the estimated limiting normal distribution to form an approximate 95% confidence interval. The mean of this limiting normal θ_k was estimated as the mean of the predictions generated by the ensemble. The variance parameter $\zeta_{k,k}$ was estimated by drawing 500 new subsamples, not necessarily used in the ensemble, and calculating the variance between predictions generated by the resulting 500 trees and $\zeta_{1,k}$ was estimated externally using $n_{\bar{z}} = 50$ and $n_{MC} = 250$.

In order to assess the coverage probability of our confidence intervals, we first need to estimate the true mean prediction θ_k at $x_1 = 10$ that would be generated by this subbagging ensemble. To estimate this true mean, we built 1000 subbagged ensembles and took the mean prediction generated by these ensembles, which we found to be 20.02 - very close to the true underlying value of 20. In this case, we found a coverage probability of 0.912, which means that 228 of our 250 confidence intervals contained our estimate of the true mean prediction. These confidence intervals are shown in Figure 3. The horizontal line in the plot is at 20.02 and represents our estimate of the true expected prediction.

This same procedure was repeated for the SLR case with $n = m = 1000$ and $k = 60$, the MARS case with $n = m = 500$ and $k = 50$, and the MARS case with $n = m = 1000$ and $k = 75$. In each case, we produced 250 confidence intervals and the coverage probabilities

are shown in Table 1. The parameters in the limiting distributions were estimated externally in exactly the same fashion, using $n_{\tilde{z}} = 50$ and $n_{MC} = 250$ to estimate $\zeta_{1,k}$. These slightly higher coverage probabilities mean that we are overestimating the variance, which is likely due to smaller values of the estimation parameters $n_{\tilde{z}}$ and n_{MC} being used to estimate $\zeta_{1,k}$.

Underlying Function	n	k	θ_k	Coverage Probability	
				External Variance Est.	Internal Variance Est.
SLR	200	30	19.94	0.912	0.912
SLR	1000	60	19.99	0.956	0.936
MARS	500	50	17.43	0.980	0.984
MARS	1000	75	17.56	0.996	0.996

Table 1: Coverage probabilities

We also repeated this procedure for generating confidence intervals using the internal variance estimation method. The resulting coverage probabilities are remarkably similar to the external variance estimation method and are shown in Table 1. These ensembles were built using $n_{\tilde{z}} = 50$ and $n_{MC} = 250$.

5.3 Hypothesis Testing

We now explore the hypothesis testing procedure for assessing feature significance. We focus on the MARS case, where our training set now consists of 6 features $X_1, ..., X_6$, but the response Y depends only on the first 5. The values of the additional feature X_6 are sampled uniformly at random from the interval $[0, 1]$ and independently of the first 5 features.

We begin by looking at the distribution of test statistics when the test set consists 41 equally spaced points between 0 and 1. That is, the first test point is $x_1 = \cdots = x_6 = 0$, the second is $x_1 = \cdots = x_6 = 0.025$, and so on so that the last test point is $x_1 = \cdots = x_6 = 1$. For this test, we are interested in looking at the difference between trees built using all features and those built using only the first 5 so that in the notation in Section 4.2, $\boldsymbol{X}^{(R)} = \{X_1, ..., X_5\}$. We ran 250 simulations with $n = 1000$, $m = 1000$, and $k = 75$ using a test set consisting of all 41 test points, the 20 central-most points, and the 5 central-most points. The parameter $\Sigma_{1,k}$ was estimated externally using $n_{\tilde{z}} = 100$ and $n_{MC} = 5000$ and $\Sigma_{k,k}$ was estimated by taking the covariance of the difference in predictions generated by 5000 trees. These covariance parameters are estimated only once instead of within each ensemble since we are only interested in the distribution of test statistics. Histograms of the resulting test statistics along with an overlay of the estimated χ^2 densities are shown in the top row of Figure 4.

We repeated this procedure, this time with a test set consisting of points in the center of the hypercube so we are not predicting close to any edge of the feature space. The value of each feature in the test set was selected uniformly at random from $[0.25, 0.75]$. A total of 41 such test points were generated, and histograms of the 250 resulting test statistics were produced in the cases where we use 5 of these test points, 20 test points, and all 41 test points. These histograms and estimated χ^2 densities are shown in the bottom row of
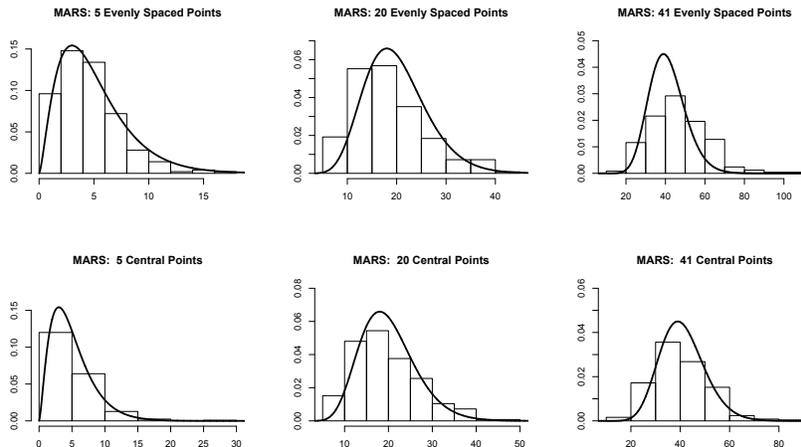
Figure 4: Histograms of simulated test statistics with estimated χ^2 overlay. The top row of histograms involve test points equally spaced between 0 and 1 and the bottom row corresponds to points randomly selected from the interior of the feature space.

Figure 4. Note that the bottom row appears to be a better fit and thus there appears to be some bias occuring when test points are selected near the edges of the feature space.

To check the alpha level of the test—the probability of incorrectly rejecting the null hypothesis—we simulated 250 new training sets and used the test set consisting of 41 randomly selected central points. For each training set, we built full and reduced subbagged estimates, allowed and not allowed to utilize X_6 respectively, estimated the parameters, and performed the hypothesis test. For each ensemble, the variance parameter $\Sigma_{1,k}$ was estimated externally using $n_{\tilde{z}} = 50$ and $n_{MC} = 1000$ and $\Sigma_{k,k}$ was estimated externally on an independently generated set of trees.

In this setup, none of the 250 simulations resulted in rejecting the null hypothesis, so our empirical alpha level was 0. A histogram of the resulting test statistics is shown on the left of Figure 5; the critical value, the 0.95 quantile of the χ^2_{41}, is 56.942. Thus, as with the confidence intervals, we are being conservative. Recall that our confidence interval simulations with $n = 1000$, $m = 1000$, and $k = 75$ predicting at $\boldsymbol{x} = (0.5, ..., 0.5)$ captured our true value 99.6% of the time, so this estimate of 0, though conservative, is not necessarily unexpected. We also repeated this procedure using an internal variance estimate with $n_{\tilde{z}} = 50$ and $n_{MC} = 1000$ and found an alpha level of 0.14. The histogram of test statistics resulting from the internal variance estimation method is shown on the right in Figure 5. Here we see that the correlation introduced by not taking an i.i.d. selection of subsamples to build the ensemble may be slightly inflating the test statistics.

5.4 Random Forests

Thus far, our simulations have dealt only with subbagged ensembles, but recall that Theorem 2 established the asymptotic normality for predictions generated by random forests as

20
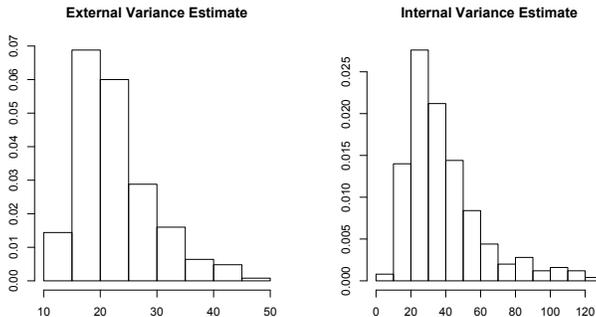
Figure 5: Histograms of simulated test statistics
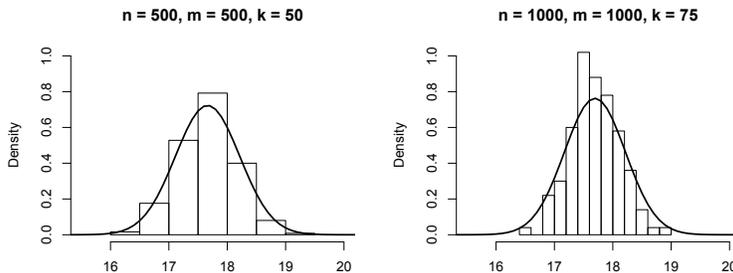
Figure 6: Histograms of random forest predictions at $x_1 = \cdots = x_5 = 0.5$ with estimated normal density overlaid.

well. The histograms in Figure 6 show the distribution of predictions generated by subsampled random forests at $x_1 = \cdots = x_5 = 0.5$ when the true underlying function is the MARS function. These trees were grown using the `randomForest` function in `R` with the `ntree` argument set to 1. At each node in each tree, 3 of the 5 features $X_1, ..., X_5$ were selected at random as candidates for splits and we insisted on at least 2 observations in each terminal node. The histograms show the empirical distribution of 250 subsampled random forest predictions and the overlaid density is the limiting normal $\mathcal{N}(\mathbb{E}r_{n,k,m}(\boldsymbol{x}^*), \frac{1}{\alpha}\frac{k^2}{m}\zeta_{1,k} + \frac{1}{m}\zeta_{k,k})$ with the variance parameters estimated externally. Our estimate of the mean of this distribution was taken as the empirical mean of the 250 predictions. Our estimate of $\zeta_{k,k}$ was taken as the empirical variance of predictions across 5000 new trees and the estimate for $\zeta_{1,k}$ was calculated with $n_{\tilde{z}} = 250$ and $n_{MC} = 2000$.

6. Real Data

We now apply our inference methods to a real data set provided by the Lab of Ornithology at Cornell University. The data is part of the ongoing *eBird* citizen science project described in Sullivan et al. (2009). This project is hosted by Cornell's Lab of Ornithology and relies
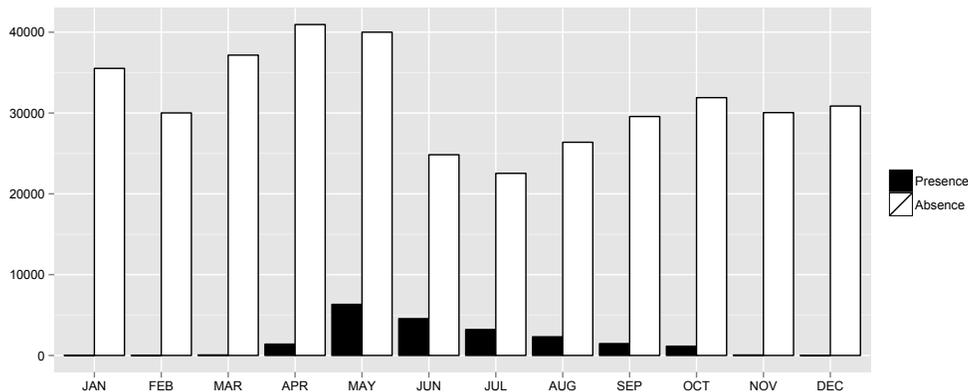
Figure 7: Monthly counts of Indigo Bunting observations.

on citizens, referred to as *birders*, to submit reports of bird observations. Location, bird species observed and not observed, effort level, and number of birds of each species observed are just a few of the variables participants are asked to provide. In addition to the data contained in these reports, landcover characteristics as reported in the 2006 United States National Land Cover Database are also available so that information about the local terrain may be used to help predict species abundance.

For our analysis, we restrict our attention to observations (and non-observations) of the Indigo Bunting species. For the first part of our analysis, we further restrict our attention to observations made during the year 2010. A little more than 400,000 reports of either presence or absence of Indigo Buntings were recorded during 2010 and the data set consists of 23 features. Like many species, the abundance of Indigo Buntings is known to fluctuate throughout the year, so we have two primary goals: (1) to produce confidence intervals for monthly abundance and (2) to show that the feature 'month' is significant for predicting abundance.

A presence/absence plot of Indigo Buntings by month is shown in Figure 7. A few features of this plot are worth pointing out. Most obviously, there are many more absence observations each month than presence observations. This makes sense because each time a birder submits a report, they note when Indigo Buntings are not present. Next, we see that this species is only observed during the warmer months, so month seems highly significant for predicting abundance. Finally, we see that all months have a large number of reports, so we need not worry about underreporting issues throughout the year.

First we produce the confidence intervals. The goal is to get an idea of the monthly abundance, which we can think of as 'probability of observation', so our test points will be 12 vectors, one for each month. For the values of the other features, we will use the average of the values recorded for that feature, or, in the case of categorical features, use the most popular category. Some features, such as elevation, have missing values which are not included in calculating the averages. We also removed the *day* feature from the training set, since day of the year is able to capture any effect of month. Since the training set consists of approximately 400,000 observations, we use a subsample size $k = 650$, slightly more than
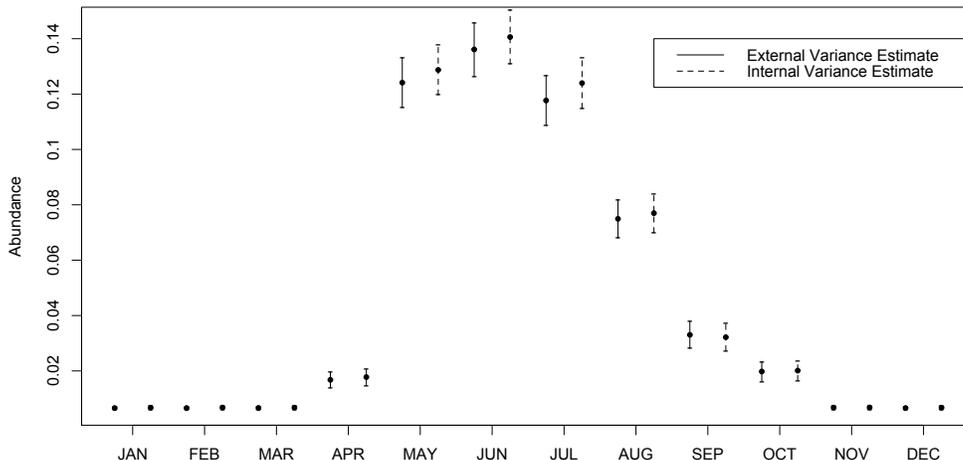
Figure 8: Monthly confidence intervals for Indigo Bunting abundance.

the square root of the training set size. We build a total of $m = 5000$ trees and take the variance of these predictions at each point to be our estimates of $\zeta_{k,k}$. We use an external estimate with $n_{\tilde{z}} = 250$ and $n_{MC} = 5000$ to estimate $\zeta_{1,k}$. For each tree built, we require a minimum of 20 observations to consider splitting an internal node. We also repeat this procedure using an internal estimate of variance with $n_{\tilde{z}} = 250$ and $n_{MC} = 5000$.

The confidence intervals are shown in Figure 8. Note that the pattern seen in Figure 7 is mirrored in the confidence interval plot: the confidence intervals are higher during months where more positive observations are recorded. It is also interesting to note that the width of the confidence intervals is larger during months of higher abundance. Observe in Figure 7 that even for months with many positive observations reported, many more negative observations are also reported. Thus, for these months there are likely a number of trees in the ensemble with nearly all positive or negative observations so we expect a higher variance in predictions. For months when very few positive observations are made, nearly all observations in the terminal node will be abscence observations, thus resulting in a very small variance and much narrower confidence intervals. Based on a visual inspection of the confidence intervals, there appears to be a clear significant difference in abundance between certain months, but we need to account for correlations in our predictions so we also conduct hypothesis tests.

We conduct these formal tests for the significance of month, following the procedure in Section 4.2. To perform this test, we randomly selected 20 points from the training set as the test set and calculated the test statistic based on an internal variance estimate with $n_{\tilde{z}} = 250$ and $n_{MC} = 5000$. We calculated a test statistic of 4233.10 and a critical value, the 0.95 quantile of the χ^2_{20}, of only 31.41 which yields a p-value of approximately 0, so it seems that month is highly significant for predicting abundance. However, when we generated random values for month and repeated the testing procedure, we calculated a

test statistic of 58.02 which, though significantly smaller than the test statistic calculated on the original training set, is still significant. To ensure that the randomized months we selected did not add any accidental structure, we compared predictions generated using this training set to those generated by another training set, also with randomized months. Here we find a test statistic of only 2.36 and thus there is no significant difference between these predictions, so the trees are simply taking advantage of additional randomness. Finally, we test for a difference in predictions generated by the original training set and those generated by the training set with random values of month. In this case, we calculated a test statistic of 2336.14 which is highly significant so we can conclude that month is significant for predicting abundance and the contribution to the prediction is significantly more than would be expected by simply adding a random feature to the model.

This significant effect of month comes as no surprise as Indigo Buntings are known to be a migratory species. However, many scientists also believe that migrations may change from year to year and thus year may also be significant for predicting abundance. For this test, we used the full training data set consisting of approximately 1 million observations from 2004 to 2010 and as a test set, randomly selected 20 observations from the training set. Given this larger training set, we increased our subsample size to $k = 1000$ and our Monte Carlo sample size to $n_{MC} = 8000$ and again performed the tests using the internal variance estimation method. In the first setup where we test predictions from the full training set against predictions generated from the training set without year, we find a test statistic of 94.43 which means that year is significant for making predictions as it is larger than the critical value of 31.41. However, as was the case in testing the significance of month, we find that a randomly generated year feature is also significant with a test statistic of 52.51. Following in the same manner as above, we compare predictions from two training sets, each with a randomized year feature, and we find no significant difference in these predictions with a test statistic of only 4.70. Finally, we test for a difference in predictions between the full training set and a data set with random year and find that there is a significant difference in these predictions with a test statistic of 109.72. Thus, as was the case with month, we can conclude that year is significant for predicting abundance and the contribution to the prediction is significantly more than would be expected by simply adding a random feature to the model.

7. Discussion

This work demonstrates that formal statistical inference procedures are possible within the context of supervised ensemble learning, even when individual base learners are difficult to analyze mathematically. Demonstrating that ensembles built with subsamples can be viewed as U-statistics allows us to calculate limiting distributions for predictions and consistent estimation procedures for the parameters allow us to compute confidence intervals for predictions and formally test the significance of features. Moreover, using the internal variance estimation procedure, we are able to do so at no additional computational cost. In his controversial paper, Breiman (2001a) contrasted traditional statistical data analysis models that emphasize interpretation with the modern algorithmic approach of machine learning where the primary goal is prediction and among the differences invoked was the statisti-

cian's concern for formalized inference. Our hope is that this work be seen as something of a bridge between Breiman's two cultures.

The distributions and procedures we discuss apply to a very general class of supervised learning algorithms. We focus on bagging and random forests with tree-based base learners due to their popularity, but any supervised ensemble method that satisfies the conditions in Theorems 1 and 2 will generate predictions that have these limiting distributions and the inference procedures can be carried out in the same way. By the same reasoning, our procedures also make no restrictions on the tree-building method.

There are also some small modifications that can be made to our procedure which would still allow for an asymptotically normal limiting distribution. First, we assumed that our subsamples were selected with replacement so that in theory, the same subsample could be selected multiple times. This choice was based primarily on the fact that ensuring the exact subsample was not taken twice would typically require extra computational work. However, even for relatively small data sets, the probability of selecting the same subsample more than once is very small, and not surprisingly, the limiting distributions remain identical when the subsamples are taken without replacement. Furthermore, we also did not allow repeat observations within subsamples, which made the resulting estimator a U-statistic. Had we selected the subsamples themselves with replacement, our estimator would be a V-statistic—a closely related class of estimators introduced by von Mises (1947)—and similar theory could be developed. It is also worth pointing out that in practice, we always selected the same number of subsamples m_n and subsample size k_n for each prediction point $\boldsymbol{x}^*$ but in theory, these could be chosen differently for each point of interest and the same can be said of the estimation parameters $n_{\tilde{\boldsymbol{z}}}$ and n_{MC}. However, choosing different values of these parameters for different prediction points involves significantly more bookkeeping and we advise against it in practice.

Our approach also raises some issues. Perhaps most obviously, the parameter in our inference procedures is the expected prediction generated by trees built in the prescribed fashion, as opposed to the true value of the underlying regression function, $F(\boldsymbol{x}^*)$. Though some tree-building methods have been shown to be consistent, the bias introduced may not be negligible so we have to be careful about interpreting our results. It may be possible to employ a residual bootstrap to try and correct for bias and we plan to explore this in future work. Another open question that we hope to address in future work is how to select the test points when testing feature significance. In the eBird application, we randomly selected points from the training set, but it would be beneficial to investigate optimal strategies for selecting both the number and location of test points. Finally, the distributional results we provide could potentially allow us to test more complex hypotheses about the structure of the underlying regression function, for example, the interaction between covariates.

Acknowledgments

Appendix A.

We present here the proofs of the theorems provided in Section 2. We begin with a lemma from Lee (1990).

Lemma 3 *(Lee 1990, Lemma A, page 201) Let $a_1, a_2, ...$ 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 the random variables $M_1, ..., M_n$ have a multinomial distribution, $multinomial(m_n; \frac{1}{m_n}, ..., \frac{1}{m_n})$. Then as $m_n, n \to \infty$, the limiting distribution of*

$$m_n^{-1/2} \sum_{i=1}^{n} a_i(M_i - m_n/n)$$

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

Additionally, it will be useful to have the limiting distribution of complete infinite order U-statistics, which we provide in the lemma below.

Lemma 4 *Let $Z_1, Z_2, ... \overset{iid}{\sim} F_Z$ and let U_{n,k_n} be a complete, infinite order U-statistic with kernel h_{k_n} satisfying Condition 1 and $\theta_{k_n} = \mathbb{E}h_{k_n}(Z_1, ..., Z_{k_n})$ such that $\mathbb{E}h_{k_n}^2(Z_1, ..., Z_{k_n}) \leq C < \infty$ for all n and some constant C and $\lim \frac{k_n}{\sqrt{n}} = 0$. Then*

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

The proof of Lemma 3 is provided in Lee (1990) page 201. The proof of Lemma 4 follows in exactly the same fashion as the proof of result *(i)* in Theorem 1 below. We take advantage of these lemmas in the following proofs.

Theorem 1 *Let $Z_1, Z_2, ... \overset{iid}{\sim} F_Z$ and let U_{n,k_n,m_n} be an incomplete, infinite order U-statistic with kernel h_{k_n} that satisfies Condition 1. Let $\theta_{k_n} = \mathbb{E}h_{k_n}(Z_1, ..., Z_{k_n})$ such that $\mathbb{E}h_{k_n}^2(Z_1, ..., Z_{k_n}) \leq C < \infty$ for all n and some constant C, and let $\lim \frac{n}{m_n} = \alpha$. Then as long as $\lim \frac{k_n}{\sqrt{n}} = 0$ and $\lim \zeta_{1,k_n} \neq 0$,*

(i) *if $\alpha = 0$, then $\frac{\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{k_n^2 \zeta_{1,k_n}}} \overset{d}{\to} \mathcal{N}(0, 1)$.*

(ii) *if $0 < \alpha < \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\frac{k_n^2}{\alpha} \zeta_{1,k_n} + \zeta_{k_n,k_n}}} \overset{d}{\to} \mathcal{N}(0, 1)$.*

(iii) *if $\alpha = \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\zeta_{k_n,k_n}}} \overset{d}{\to} \mathcal{N}(0, 1)$.*

Proof:

(i) Suppose first that $\alpha = 0$. In the interest of clarity, we follow the Hájek projection method discussed in Van der Vaart (2000) chapters 11 and 12. Define the Hájek projection of $U_{n,k_n,m_n} - \theta_{k_n}$ as

$$\hat{U}_{n,k_n,m_n} = \sum_{i=1}^{n} \mathbb{E}(U_{n,k_n,m_n} - \theta_{k_n} \mid Z_i) - (n-1)\mathbb{E}(U_{n,k_n,m_n} - \theta_{k_n})$$

$$= \sum_{i=1}^{n} \mathbb{E}(U_{n,k_n,m_n} - \theta_{k_n} \mid Z_i)$$

so that for each term in the sum, we have

$$\mathbb{E}(U_{n,k_n,m_n} - \theta_{k_n} \mid Z_i) = \mathbb{E}\left(\frac{1}{m_n} \sum_{\beta} h_{k_n}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \theta_{k_n} \;\middle|\; Z_i \right)$$

$$= \frac{1}{m_n} \sum_{\beta} \mathbb{E}\left(h_{k_n}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \theta_{k_n} \mid Z_i \right) \tag{7}$$

where, in keeping with the notation in Van der Vaart (2000), we let β index the subsamples. Define $h_{1,k_n}(x) = \mathbb{E}h_{k_n}(x, Z_2, ..., Z_{k_n}) - \theta_{k_n}$ and let W be the number of subsamples that contain i. Since we assume that the subsamples are selected uniformly at random with replacement,

$$W \sim Binom\left(m_n, \frac{\binom{n-1}{k_n-1}}{\binom{n}{k}} \right)$$

so we can rewrite (7) as

$$\frac{1}{m_n} \sum_{\beta} \mathbb{E}\left(\mathbb{E}(h_{k_n}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \theta_{k_n} | Z_i) \mid W \right)$$

$$= \frac{1}{m_n} \mathbb{E}(W h_{1,k_n}(Z_i))$$

$$= \frac{1}{m_n} \left(m_n \left(\frac{\binom{n-1}{k_n-1}}{\binom{n}{k}} \right) h_{1,k_n}(Z_i) \right)$$

$$= \frac{k_n}{n} h_{1,k_n}(Z_i)$$

so that taking the sum yields

$$\hat{U}_{n,k_n,m_n} = \frac{k_n}{n} \sum_{i=1}^{n} h_{1,k_n}(Z_i).$$

Now we establish the asymptotic normality of $\hat{U}_{n,k_n,m_n}$. Define the triangular array

$$k_{n_1}h_{1,k_{n_1}}(Z_1), \ \dots \ , k_{n_1}h_{1,k_{n_1}}(Z_{n_1})$$
$$k_{n_1+1}h_{1,k_{n_1+1}}(Z_1), \quad \dots \quad \dots \quad , k_{n_1+1}h_{1,k_{n_1+1}}(Z_{n_1+1})$$
$$. \qquad\qquad\qquad\qquad .$$
$$. \qquad\qquad\qquad\qquad\qquad .$$
$$. \qquad\qquad\qquad\qquad\qquad\quad .$$
$$k_{n_1+j}h_{1,k_{n_1+j}}(Z_1), \quad \dots \quad \dots \qquad \dots \quad \dots \qquad \dots \quad \dots \quad , k_{n_1+j}h_{1,k_{n_1+j}}(Z_{n_1+j})$$

so that for each variable in the array, we have

$$\mathbb{E}(k_n h_{1,k_n}(Z_i)) = k_n(\theta_{k_n} - \theta_{k_n}) = 0$$

and

$$var(k_n h_{1,k_n}(Z_i)) = k_n^2 var(h_{1,k_n}(Z_i)) = k_n^2 \zeta_{1,k_n}$$

and thus the row-wise sum of the variances is

$$s_n^2 = \sum_{i=1}^{n} var(k_n h_{1,k_n}(Z_i)) = n k_n^2 \zeta_{1,k_n}.$$

For $\delta > 0$, the Lindeberg condition is given by

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{nk_n^2\zeta_{1,k_n}} \int_{|k_n h_{1,k_n}(Z_i)| \geq \delta k_n \sqrt{n\zeta_{1,k_n}}} k_n^2 h_{1,k_n}^2(Z_i)dP$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{n\zeta_{1,k_n}} \int_{|h_{1,k_n}(Z_i)| \geq \delta \sqrt{n\zeta_{1,k_n}}} h_{1,k_n}^2(Z_i)dP$$

$$\leq \lim_{n\to\infty} \max_{1\leq i\leq n} \frac{1}{\zeta_{1,k_n}} \int_{|h_{1,k_n}(Z_i)| \geq \delta \sqrt{n\zeta_{1,k_n}}} h_{1,k_n}^2(Z_i)dP$$

$$= \lim_{n\to\infty} \frac{1}{\zeta_{1,k_n}} \int_{|h_{1,k_n}(Z_1)| \geq \delta \sqrt{n\zeta_{1,k_n}}} h_{1,k_n}^2(Z_1)dP \qquad (8)$$

$$= 0$$

by Condition 1, and thus the Lindeberg condition is satisfied. Thus, by the Lindeberg-Feller central limit theorem,

$$\frac{\sum_j k_n h_{1,k_n}(Z_j)}{s_n} \xrightarrow{d} \mathcal{N}(0,1)$$

or, rewriting,

$$\frac{\sqrt{n}\hat{U}_{n,k_n,m_n}}{\sqrt{k_n^2\zeta_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Now, we need to compare the limiting variance ratio of U_{n,k_n,m_n} and its Hájek projection $\hat{U}_{n,k_n,m_n}$. For incomplete U-statistics of fixed rank, Blom (1976) showed that the variance of the incomplete U-statistic U_m consisting of m subsamples selected uniformly at random with replacement is given by

$$\frac{\zeta_k}{m_n} + \left(1 - \frac{1}{m_n}\right) var(U)$$

where U is the complete U-statistic analogue. Extending this result to our situation where k and m may both depend on n, we have

$$var(U_{n,k_n,m_n}) = \frac{\zeta_{k_n,k_n}}{m_n} + \left(1 - \frac{1}{m_n}\right) var(U_{n,k_n})$$

where the variance of the complete U-statistic U_{n,k_n} is given by

$$\sum_{c=1}^{k_n} \frac{k_n!^2}{c!(k_n - c)!^2} \frac{(n - k_n)(n - k_n - 1)\cdots(n - 2k_n + c + 1)}{n(n-1)\cdots(n - k_n + 1)} \zeta_{c,k_n}.$$

The details of this calculation are described in Van der Vaart (2000) page 163. Thus, looking at the limit of the variance ratio, we have

$$\lim_{n\to\infty} \left(\frac{var(U_{n,k_n,m_n})}{var(\hat{U}_{n,k_n,m_n})}\right) = \lim_{n\to\infty} \left(\frac{\frac{\zeta_{k_n,k_n}}{m_n} + \left(1 - \frac{1}{m_n}\right) var(U_{n,k_n})}{\frac{k_n^2}{n}\zeta_{1,k_n}}\right)$$

$$= \lim_{n\to\infty} \left(\frac{n\,\zeta_{k_n,k_n}}{m_n\,k_n^2\,\zeta_{1,k_n}}\right) + \lim_{n\to\infty}\left(1 - \frac{1}{m_n}\right)\lim_{n\to\infty}\left(\frac{n\,var(U_{n,k_n})}{k_n^2\,\zeta_{1,k_n}}\right)$$

$$= 0 + \lim_{n\to\infty}\left(\frac{n\,var(U_{n,k_n})}{k_n^2\,\zeta_{1,k_n}}\right)$$

$$= \lim_{n\to\infty}\left(\frac{\sum_{c=1}^{k_n}\frac{k_n!^2}{c!(k_n-c)!^2}\frac{(n-k_n)(n-k_n-1)\cdots(n-2k_n+c+1)}{(n-1)\cdots(n-k_n+1)}\zeta_{c,k_n}}{k_n^2\zeta_{1,k_n}}\right)$$

$$= \lim_{n\to\infty}\left(\frac{\frac{k_n!^2}{(k_n-1)!^2}\frac{(n-k_n)(n-k_n-1)\cdots(n-2k_n+2)}{(n-1)\cdots(n-k_n+1)}\zeta_{1,k_n}}{k_n^2\zeta_{1,k_n}}\right)$$

$$= \lim_{n\to\infty}\left(\frac{(n - k_n)(n - k_n - 1)\cdots(n - 2k_n + 2)}{(n-1)\cdots(n - k_n + 1)}\right)$$

$$= 1.$$

Note that in the second line, $\lim\left(\frac{n\,\zeta_{k_n,k_n}}{m_n\,k_n^2\,\zeta_{1,k_n}}\right) = 0$ since $\frac{n}{m_n} \to 0$, $\zeta_{1,k_n} \not\to 0$ by assumption, and $\zeta_{k_n,k_n} \not\to \infty$ since $\mathbb{E}h_{k_n}^2(Z_1,...,Z_{k_n})$ is bounded. Finally, by Slutsky's Theorem and Theorem 11.2 in Van der Vaart (2000), we have

$$\frac{\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{k_n^2\zeta_{1,k_n}}} = \frac{\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n} - \hat{U}_{n,k_n,m_n} + \hat{U}_{n,k_n,m_n})}{\sqrt{k_n^2\zeta_{1,k_n}}}$$

$$= \frac{\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n} - \hat{U}_{n,k_n,m_n})}{\sqrt{k_n^2 \zeta_{1,k_n}}} + \frac{\sqrt{n}\hat{U}_{n,k_n,m_n}}{\sqrt{k_n^2 \zeta_{1,k_n}}}$$

where $\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n} - \hat{U}_{n,k_n,m_n})/\sqrt{k_n^2 \zeta_{1,k_n}} \xrightarrow{P} 0$ and

$$\frac{\sqrt{n}\hat{U}_{n,k_n,m_n}}{\sqrt{k_n^2 \zeta_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0,1)$$

so that

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

as desired. $\square$

(ii) & (iii) Now suppose $\alpha > 0$. We follow the proof technique in Lee (1990) page 200 which is based on the work of Janson (1984). Let $\mathcal{S}_{(n,k_n)} = \{S_i : i = 1, ..., \binom{n}{k_n}\}$ denote the set of all possible subsamples of size k_n. In the following work, we use the notation (n, k_n) in place of $\binom{n}{k_n}$ in subscripts and summation notation. Consider the random vector $M_{n,k_n} = (M_{S_1}, ..., M_{S_{(n,k_n)}})$, where the ith element denotes the number of times the ith subsample appears in U_{n,k_n,m_n}. Since the subsamples are selected uniformly at random with replacement, $M_{n,k_n} \sim multinomial(m_n; \frac{1}{\binom{n}{k_n}}, ..., \frac{1}{\binom{n}{k_n}})$. Let ϕ_{n,k_n,m_n} be the characteristic function of $\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})$ and let ϕ denote the limiting characteristic function of $\sqrt{n}(U_{n,k_n} - \theta_{k_n})$ where U_{n,k_n} is the corresponding complete U-statistic. Additionally, let $\phi_{n,k_n,m_n}^{(M)}$ be the characteristic function of the random variable

$$m_n^{-1/2} \sum_{i=1}^{(n,k)} \left(M_{S_i} - \frac{m_n}{\binom{n}{k_n}}\right)\left(h_{k_n}(S_i) - \theta_{k_n}\right)\Big| Z_1, ..., Z_n.$$

Then we have

$$\phi_{n,k_n,m_n}(t) = \mathbb{E}\left(exp\left[it\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})\right]\right)$$

$$= \mathbb{E}\left(exp\left[itm_n^{-1/2}\left(\sum_{i=1}^{(n,k_n)} M_{S_i}(h_{k_n}(S_i) - \theta_{k_n})\right)\right]\right)$$

$$= \mathbb{E}\left(\mathbb{E}\left(exp\left[itm_n^{-1/2}\left(\sum_{i=1}^{(n,k_n)} M_{S_i}(h_{k_n}(S_i) - \theta_{k_n})\right)\right]\Big| Z_1, ..., Z_n\right)\right)$$

$$= \mathbb{E}\left(\mathbb{E}\left(exp\left[itm_n^{-1/2}\left(\sum_{i=1}^{(n,k_n)} \left(M_{S_i} + \frac{m_n}{\binom{n}{k_n}} - \frac{m_n}{\binom{n}{k_n}}\right)\right.\right.\right.\right.$$
$$\left.\left.\left.\left. \times \left(h_{k_n}(S_i) - \theta_{k_n}\right)\right)\right]\Big| Z_1, ..., Z_n\right)\right)$$

$$
= \mathbb{E}\Bigg(\mathbb{E}\Bigg(exp\Bigg[itm_n^{-1/2} \Bigg(\sum_{i=1}^{(n,k_n)} \frac{m_n}{\binom{n}{k_n}}(h_{k_n}(S_i) - \theta_{k_n}) \Bigg) \Bigg]
$$

$$
\times exp\Bigg[itm_n^{-1/2} \Bigg(\sum_{i=1}^{(n,k)} \Bigg(M_{S_i} - \frac{m_n}{\binom{n}{k_n}} \Bigg) (h_{k_n}(S_i) - \theta_{k_n}) \Bigg) \Bigg] \Bigg| Z_1, ..., Z_n \Bigg) \Bigg)
$$

$$
= \mathbb{E}\Bigg(exp\Bigg[itm_n^{-1/2} \Bigg(\sum_{i=1}^{(n,k_n)} \frac{m_n}{\binom{n}{k_n}}(h_{k_n}(S_i) - \theta_{k_n}) \Bigg) \Bigg]
$$

$$
\times \mathbb{E}\Bigg(exp\Bigg[itm_n^{-1/2} \Bigg(\sum_{i=1}^{(n,k_n)} \Bigg(M_{S_i} - \frac{m_n}{\binom{n}{k_n}} \Bigg) (h_{k_n}(S_i) - \theta_{k_n}) \Bigg) \Bigg] \Bigg| Z_1, ..., Z_n \Bigg) \Bigg)
$$

$$
= \mathbb{E}\Bigg(exp\Big[it\sqrt{m_n}U_{n,k_n} \Big] \; \phi_{n,k_n,m_n}^{(M)}(t) \Bigg)
$$

and now, taking limits, we have

$$
\lim_{n\to\infty} \phi_{n,k_n,m_n}(t) = \lim_{n\to\infty} \mathbb{E}\Bigg(exp\Big[it\sqrt{m_n}U_{n,k_n} \Big] \; \phi_{n,k_n,m_n}^{(M)}(t) \Bigg)
$$

$$
= \mathbb{E}\Bigg(\lim_{n\to\infty} exp\Big[it\sqrt{m_n}U_{n,k_n} \Big] \; \lim_{n\to\infty} \phi_{n,k_n,m_n}^{(M)}(t) \Bigg)
$$

so that by the preceeding lemmas,

$$
\lim_{n\to\infty} \phi_{n,k_n,m_n}(t) = \lim_{n\to\infty} \mathbb{E}\Bigg(exp\Big[it\sqrt{m_n}U_{n,k_n} \Big] \Bigg) exp\Big[-t^2\zeta_{k_n,k_n}/2 \Big]
$$

$$
= \lim_{n\to\infty} \mathbb{E}\Bigg(exp\Big[it\Big(\frac{\sqrt{m_n}}{\sqrt{n}}\Big)\sqrt{n}U_{n,k_n} \Big] \Bigg) exp\Big[-t^2\zeta_{k_n,k_n}/2 \Big]
$$

$$
= \phi(\alpha^{-1/2}t)exp\Big[-t^2\zeta_{k_n,k_n}/2 \Big]
$$

$$
= exp\Big[-(t\alpha^{-1/2})^2 k_n^2\zeta_{1,k_n}/2 \Big] exp\Big[-t^2\zeta_{k_n,k_n}/2 \Big]
$$

$$
= exp\Big[-t^2\Big(\frac{k_n^2}{\alpha}\zeta_{1,k_n} + \zeta_{k_n,k_n}\Big)/2 \Big]
$$

which is the characteristic function of a Normal distribution with mean 0 and variance $\frac{k_n^2}{\alpha}\zeta_{1,k_n} + \zeta_{k_n,k_n}$. Note that when $\alpha = \infty$, the first term in the variance is 0, so the limiting variance reduces to ζ_{k_n,k_n}, as desired. ∎

Proposition 1: *For a bounded regression function F, if there exists a constant c such that for all $k_n \geq 1$,*

$$\left|h((\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}), (\boldsymbol{X}_{k_n+1}, Y_{k_n+1})) - h((\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}), (\boldsymbol{X}_{k_n+1}, Y_{k_n+1}^*))\right|$$
$$\leq c\left|Y_{k_n+1} - Y_{k_n+1}^*\right|$$

where $Y_{k_n+1} = F(\boldsymbol{X}_{k_n+1}) + \epsilon_{k_n+1}$, $Y_{k_n+1}^* = F(\boldsymbol{X}_{k_n+1}) + \epsilon_{k_n+1}^*$, and where ϵ_{k_n+1} and $\epsilon_{k_n+1}^*$ are i.i.d. with exponential tails, then Condition 1 is satisfied.

Proof. First consider the particular $Y_j^* = F(\boldsymbol{X}_j)$. Since F is bounded, we can define

$$\sup_{\boldsymbol{X}_j \in \mathcal{X}} \left|F(\boldsymbol{X}_j)\right| \leq M < \infty$$

and since tree-based predictions cannot fall outside the range of responses in the training set, $\left|h((\boldsymbol{X}_1, Y_1^*), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}^*))\right| \leq M$ for all possible $\boldsymbol{X}_1, ..., \boldsymbol{X}_{k_n}$. Furthermore, by the Lipschitz condition,

$$\left|h((\boldsymbol{X}_1, Y_1^*), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}^*)) - h((\boldsymbol{X}_1, Y_1^*), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}))\right| \leq c\sum_{j=2}^{k_n} \left|\epsilon_j\right|$$

and thus, applying Jensen's Inequality, we have

$$\sup_{\boldsymbol{X}_1 \in \mathcal{X}} \left|h_{1,k_n}((\boldsymbol{X}_1, Y_1^*))\right|$$

$$= \sup_{\boldsymbol{X}_1 \in \mathcal{X}} \left| \int h((\boldsymbol{X}_1, Y_1^*), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_{k_n}, Y_{k_n})) - h((\boldsymbol{X}_1, Y_1^*), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}^*)) \right.$$

$$\left. + h((\boldsymbol{X}_1, Y_1^*), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}^*))dP - \theta_{k_n} \right|$$

$$\leq \sup_{\boldsymbol{X}_1 \in \mathcal{X}, ..., \boldsymbol{X}_{k_n} \in \mathcal{X}} \int \left| h((\boldsymbol{X}_1, Y_1^*), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_{k_n}, Y_{k_n})) - h((\boldsymbol{X}_1, Y_1^*), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}^*))\right| dP$$

$$+ \sup_{\boldsymbol{X}_1 \in \mathcal{X}} h((\boldsymbol{X}_1, Y_1^*), ..., (\boldsymbol{X}_{k_n}, Y_{k_n}^*)) - \theta_{k_n}$$

$$\leq ck_n\mathbb{E}\left|\epsilon_1\right| + M - \theta_{k_n}.$$

Now, define the set of interest in the Lindeberg condition as A_n so that we may write

$$A_n = \left\{ \left| h_{1,k_n} \big((\boldsymbol{X}_1, Y_1) \big) \right| \geq \delta \sqrt{n \zeta_{1,k_n}} \right\}$$

$$= \left\{ \left| h_{1,k_n} \big((\boldsymbol{X}_1, Y_1) \big) - h_{1,k_n} \big((\boldsymbol{X}_1, Y_1^*) \big) + h_{1,k_n} \big((\boldsymbol{X}_1, Y_1^*) \big) \right| \geq \delta \sqrt{n \zeta_{1,k_n}} \right\}$$

$$\subseteq \left\{ \left| h_{1,k_n} \big((\boldsymbol{X}_1, Y_1) \big) - h_{1,k_n} \big((\boldsymbol{X}_1, Y_1^*) \big) \right| \geq \delta \sqrt{n \zeta_{1,k_n}} + \left| h_{1,k_n} \big((\boldsymbol{X}_1, Y_1^*) \big) \right| \right\}$$

$$\subseteq \left\{ \left| \epsilon_1 \right| \geq \frac{1}{c} \left(\delta \sqrt{n \zeta_{1,k_n}} + M - \theta_{k_n} \right) + k_n \mathbb{E} \left| \epsilon_1 \right| \right\}$$

$$=: A_n^*$$

Finally, continuing from equation (8) of Theorem 1,

$$\lim_{n \to \infty} \frac{1}{\zeta_{1,k_n}} \int_{A_n} h_{1,k_n}^2 \big((\boldsymbol{X}_1, Y_1) \big) dP$$

$$= \lim_{n \to \infty} \frac{1}{\zeta_{1,k_n}} \int_{A_n} \left(h_{1,k_n} \big((\boldsymbol{X}_1, Y_1) \big) - h_{1,k_n} \big((\boldsymbol{X}_1, Y_1^*) \big) + h_{1,k_n} \big((\boldsymbol{X}_1, Y_1^*) \big) \right)^2 dP$$

$$\leq \lim_{n \to \infty} \frac{2}{\zeta_{1,k_n}} \int_{A_n} \left(h_{1,k_n} \big((\boldsymbol{X}_1, Y_1) \big) - h_{1,k_n} \big((\boldsymbol{X}_1, Y_1^*) \big) \right)^2 dP$$

$$+ \lim_{n \to \infty} \frac{2}{\zeta_{1,k_n}} \int_{A_n} h_{1,k_n}^2 \big((\boldsymbol{X}_1, Y_1^*) \big) dP$$

$$\leq \lim_{n \to \infty} \frac{2}{\zeta_{1,k_n}} \int_{A_n^*} c^2 \epsilon_1^2 dP + \lim_{n \to \infty} \frac{2}{\zeta_{1,k_n}} P(A_n^*) \left(c k_n \mathbb{E} \left| \epsilon_1 \right| + M - \theta_{k_n} \right)^2$$

$$= \lim_{n \to \infty} \frac{2}{\zeta_{1,k_n}} P \left[\left| \epsilon_1 \right| \geq \frac{1}{c} \left(\sqrt{n \zeta_{1,k_n}} + M - \theta_{k_n} \right) + k \mathbb{E} \left| \epsilon_1 \right| \right]$$

$$\times \left(c k_n \mathbb{E} \left| \epsilon_1 \right| + M - \theta_{k_n} \right)^2$$

$$= 0$$

as desired, so long as $nP \left(\left| \epsilon_1 \right| > \sqrt{n} \right) \to 0$, which is the case with exponential tails, and $k_n = o(\sqrt{n})$. ∎

Theorem 2 *Let $U_{\omega; n, k_n, m_n}$ be a random kernel U-statistic of the form defined in equation (5) such that $U_{\omega; n, k_n, m_n}^*$ satisfies Condition 1 and suppose that $\mathbb{E} h_{k_n}^2 (Z_1, ..., Z_{k_n}) < \infty$ for all n, $\lim \frac{k_n}{\sqrt{n}} = 0$, and $\lim \frac{n}{m_n} = \alpha$. Then, letting β index the subsamples, so long as $\lim \zeta_{1,k_n} \neq 0$ and*

$$\lim_{n \to \infty} \mathbb{E} \left(h_{k_n}^{(\omega)} (Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)} (Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2 \neq \infty,$$

$U_{\omega; n, k_n, m_n}$ is asymptotically normal and the limiting distributions are the same as those provided in Theorem 1.

Proof. We begin with the case where $\alpha = 0$ and we make use of this result in the proof of the case where $\alpha > 0$. As in Section 2.2, define $U^*_{\omega;n,k_n,m_n} = \mathbb{E}_\omega U_{\omega;n,k_n,m_n}$. We have

$$\mathbb{E}(U_{\omega;n,k_n,m_n} - U^*_{\omega;n,k_n,m_n})^2$$

$$= \mathbb{E}\left[\left(\frac{1}{m_n}\sum_\beta h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}}) - \mathbb{E}_\omega\left(\frac{1}{m_n}\sum_\beta h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}})\right)\right)^2\right]$$

$$= \mathbb{E}\left[\frac{1}{m_n^2}\left(\sum_\beta h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}}) - \mathbb{E}_\omega\left(\sum_\beta h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}})\right)\right)^2\right]$$

$$= \mathbb{E}\frac{1}{m_n^2}\left(\sum_\beta h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}}) - \left(\sum_\beta \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}})\right)\right)^2$$

$$= \mathbb{E}\left[\frac{1}{m_n^2}\sum_\beta \left(h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1},...,Z_{\beta_{k_n}})\right)^2\right]$$

$$+ \mathbb{E}\left[\frac{1}{m_n^2}\sum_{\beta_i \neq \beta_j} \left(h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}})\right)\right.$$

$$\left. \times \left(h_{k_n}^{(\omega)}(Z_{\beta_{j_1}},...,Z_{\beta_{j_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{j_1}},...,Z_{\beta_{j_{k_n}}})\right)\right]$$

We focus now on the second term, involving the cross terms with different subsamples and randomization parameters. Splitting apart the expectation and moving the expectation with respect to ω inside, we can write the second term as

$$\mathbb{E}\left[\frac{1}{m_n^2}\sum_{\beta_i \neq \beta_j} \left(h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}})\right)\right.$$

$$\left. \times \left(h_{k_n}^{(\omega)}(Z_{\beta_{j_1}},...,Z_{\beta_{j_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{j_1}},...,Z_{\beta_{j_{k_n}}})\right)\right]$$

$$= \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{m_n^2}\sum_{\beta_i \neq \beta_j} \mathbb{E}_\omega\left(h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}})\right)\right.$$

$$\left. \times \mathbb{E}_\omega\left(h_{k_n}^{(\omega)}(Z_{\beta_{j_1}},...,Z_{\beta_{j_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{j_1}},...,Z_{\beta_{j_{k_n}}})\right)\right]$$

$$= \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{m_n^2}\sum_{\beta_i \neq \beta_j} \left(\mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{i_1}},...,Z_{\beta_{i_{k_n}}})\right)\right.$$

$$\times \left(\mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{j_1}}, ..., Z_{\beta_{j_{k_n}}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_{j_1}}, ..., Z_{\beta_{j_{k_n}}}) \right) \Bigg]$$

$$= \mathbb{E}_{\boldsymbol{X}} \left[\frac{1}{m_n^2} \sum_{\beta_i \neq \beta_j} 0 \times 0 \right]$$

$$= 0$$

and thus we need only investigate the first term. We have

$$\mathbb{E} \left[\frac{1}{m_n^2} \sum_\beta \left(h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2 \right]$$

$$= \frac{1}{m_n^2} \sum_\beta \mathbb{E} \left(h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2$$

$$= \frac{1}{m_n^2} m_n \mathbb{E} \left(h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2$$

$$= \frac{1}{m_n} \mathbb{E} \left(h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2 .$$

Putting all of this together, we have

$$\lim_{n \to \infty} \mathbb{E} \left(\frac{\sqrt{n} \left(U_{\omega;n,k_n,m_n} - U^*_{\omega;n,k_n,m_n} \right)}{\sqrt{k_n^2 \zeta_{1,k_n}}} \right)^2$$

$$= \lim_{n \to \infty} \frac{n}{m_n} \frac{1}{k_n^2 \zeta_{1,k_n}} \mathbb{E} \left(h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2$$

$$= 0$$

so long as $k_n^2 \zeta_{1,k_n} \not\to 0$ and $\lim_{n\to\infty} \mathbb{E} \left(h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) - \mathbb{E}_\omega h_{k_n}^{(\omega)}(Z_{\beta_1}, ..., Z_{\beta_{k_n}}) \right)^2 \neq \infty$, as desired. $\square$

Now we handle the case where $\alpha > 0$. First note that when $k_n = 1$ for all n, this reduces to simply averaging over an *i.i.d.* sample and thus asymptotic normality can be obtained via the classic central limit theorem so assume that eventually $k_n > 1$. The remaining steps in this proof are nearly identical to the proof of results *(ii)* and *(iii)* of Theorem 1. Again, let $\mathcal{S}_{(n,k_n)} = \{S_i : i = 1, ..., \binom{n}{k_n}\}$ denote the set of all possible subsamples of size k_n and let $M_{n,k_n} = (M_{S_1}, ..., M_{S_{(n,k_n)}})$ denote the random vector that counts the number of times each subsample appears so that $M_{n,k_n} \sim multinomial(m_n; \frac{1}{\binom{n}{k_n}}, ..., \frac{1}{\binom{n}{k_n}})$ since we assume the subsamples are selected uniformly at random with replacement. Define

$\theta^*_{k_n} = \mathbb{E}U^*_{\omega;n,k_n,m_n}$, let ϕ_{n,k_n,m_n} be the characteristic function of $\sqrt{m_n}(U_{\omega;n,k_n,m_n} - \theta^*_{k_n})$, and let ϕ denote the limiting characteristic function of $\sqrt{n}(U^*_{\omega;n,k_n,m_n} - \theta^*_{k_n})$. Finally, let $\phi^{(M)}_{n,k_n,m_n}$ be the characteristic function of the random variable

$$m_n^{-1/2} \sum_{i=1}^{(n,k)} \left(M_{S_i} - \frac{m_n}{\binom{n}{k_n}} \right) \left(h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n} \right) \Big| Z_1, ..., Z_n, \omega.$$

Then we have

$$
\begin{aligned}
\phi_{n,k_n,m_n}(t) &= \mathbb{E}\left(exp\left[it\sqrt{m_n}(U_{\omega;n,k_n,m_n} - \theta^*_{k_n}) \right] \right) \\
&= \mathbb{E}\left(exp\left[itm_n^{-1/2} \left(\sum_{i=1}^{\binom{n,k_n}{}} M_{S_i}(h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n}) \right) \right] \right) \\
&= \mathbb{E}\left(\mathbb{E}\left(exp\left[itm_n^{-1/2} \left(\sum_{i=1}^{\binom{n,k_n}{}} M_{S_i}(h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n}) \right) \right] \Big| Z_1, ..., Z_n, \omega \right) \right) \\
&= \mathbb{E}\Bigg(\mathbb{E}\bigg(exp\bigg[itm_n^{-1/2} \bigg(\sum_{i=1}^{\binom{n,k_n}{}} \left(M_{S_i} + \frac{m_n}{\binom{n}{k_n}} - \frac{m_n}{\binom{n}{k_n}} \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \times \left(h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n} \right) \bigg) \bigg] \bigg| Z_1, ..., Z_n, \omega \bigg) \Bigg) \\
&= \mathbb{E}\Bigg(\mathbb{E}\bigg(exp\bigg[itm_n^{-1/2} \left(\sum_{i=1}^{\binom{n,k_n}{}} \frac{m_n}{\binom{n}{k_n}} (h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n}) \right) \bigg] \\
&\qquad \times exp\bigg[itm_n^{-1/2} \left(\sum_{i=1}^{(n,k)} \left(M_{S_i} - \frac{m_n}{\binom{n}{k_n}} \right) (h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n}) \right) \bigg] \bigg| Z_1, ..., Z_n, \omega \bigg) \Bigg) \\
&= \mathbb{E}\Bigg(exp\bigg[itm_n^{-1/2} \left(\sum_{i=1}^{\binom{n,k_n}{}} \frac{m_n}{\binom{n}{k_n}} (h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n}) \right) \bigg] \\
&\qquad \times \mathbb{E}\bigg(exp\bigg[itm_n^{-1/2} \left(\sum_{i=1}^{\binom{n,k_n}{}} \left(M_{S_i} - \frac{m_n}{\binom{n}{k_n}} \right) (h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n}) \right) \bigg] \bigg| Z_1, ..., Z_n, \omega \bigg) \Bigg) \\
&= \mathbb{E}\Bigg(exp\bigg[itm_n^{-1/2} \bigg(\sum_{i=1}^{\binom{n,k_n}{}} \frac{m_n}{\binom{n}{k_n}} \Big(h_{k_n}^{(\omega_i)}(S_i) - \mathbb{E}_\omega h_{k_n}^{(\omega_i)}(S_i) \\
&\qquad\qquad\qquad\qquad\qquad + \mathbb{E}_\omega h_{k_n}^{(\omega_i)}(S_i) - \theta^*_{k_n} \Big) \bigg) \bigg] \; \phi^{(M)}_{n,k_n,m_n}(t) \Bigg) \\
&= \mathbb{E}\Bigg(exp\bigg[it\sqrt{m_n}\bigg(\frac{1}{\binom{n}{k_n}} \sum_{i=1}^{\binom{n,k_n}{}} \Big(h_{k_n}^{(\omega_i)}(S_i) - \mathbb{E}_\omega h_{k_n}^{(\omega_i)}(S_i) \Big)
\end{aligned}
$$

$$+ \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{(n,k_n)} \left(\mathbb{E}_\omega h_{k_n}^{(\omega_i)}(S_i) - \theta_{k_n}^* \right) \Big) \Big] \phi_{n,k_n,m_n}^{(M)}(t) \Bigg).$$

Now, note that since we are in the case where $\alpha > 0$, $m_n = O(n) \ll (n, k_n)$ and thus, by the previous result in the case where $\alpha = 0$, the first term converges to 0 and we have

$$\lim_{n\to\infty} \phi_{n,k_n,m_n}(t) = \mathbb{E}\Bigg(\lim_{n\to\infty} exp\Big[it\sqrt{m_n} \, \frac{1}{\binom{n}{k_n}} \sum_{i=1}^{(n,k_n)} \left(\mathbb{E}_\omega h_{k_n}^{(\omega_i)}(S_i) - \theta_{k_n}^* \right) \Big] \lim_{n\to\infty} \phi_{n,k_n,m_n}^{(M)}(t) \Bigg)$$

$$= \mathbb{E}\Bigg(\lim_{n\to\infty} exp\Big[it\sqrt{m_n} \, U_{\omega;n,k_n,m_n}^* \Big] \lim_{n\to\infty} \phi_{n,k_n,m_n}^{(M)}(t) \Bigg)$$

so that by exactly the same arguments as in the proof of Theorem 1

$$\lim_{n\to\infty} \phi_{n,k_n,m_n}(t) = exp\Big[-t^2\Big(\frac{k_n^2}{\alpha}\zeta_{1,k_n}^* + \zeta_{k_n,k_n}^*\Big)/2 \Big]$$

which is the characteristic function of a Normal distribution with mean 0 and variance $\frac{k_n^2}{\alpha}\zeta_{1,k_n}^* + \zeta_{k_n,k_n}^*$. Further, when $\alpha = \infty$, the variance reduces to ζ_{k_n,k_n}^*, as desired. ∎

Appendix B.

Crossed Design Random Forests

Typically, each tree in a random forest is built according to an independently selected randomization parameter ω. Alternatively, for each ω, we could choose to build an entire set of trees, one for each of the m_n subsamples, so that if Ω_n randomization parameters are used, a total of $\Omega_n \times m_n$ trees are built. We could then write the prediction at x^* generated by this alternative random forest estimator as

$$\frac{1}{\Omega_n}\frac{1}{m_n} \sum_{(i)} \sum_{(j)} T_{x^*,k_n}((\boldsymbol{X}_{i_1},Y_{i_1}),...,(\boldsymbol{X}_{i_{k_n}},Y_{i_{k_n}}); \omega_j) \qquad (9)$$

where here, we explicitly treat ω as an input to the function. Statistics of the form in (9) are referred to as *two-sample* or *generalized* U-statistics and similar results regarding asymptotic normality have been established for fixed-rank kernels; see Lee (1990) or Van der Vaart (2000) for details.

We mention this as a possible alternative only because the resulting estimator takes the well established form of a generalized U-statistic. Readers familiar with U-statistics may be more comfortable with this approach than with the random kernel approach that more closely resembles the type of random forests used in practice. However, since this formulation strays from Breiman's original procedure and is far more computationally intensive, we consider only the random kernel version described in Section 2.2.

37

Appendix C.

The algorithm for estimating Σ_{1,k_n} as needed for the hypothesis testing procedure is given below.

Algorithm 6 Σ_{1,k_n} Estimation Procedure

---
  for i in 1 to $n_{\tilde{z}}$ **do**
    Select initial fixed point $\tilde{z}^{(i)}$
    for j in 1 to n_{MC} **do**
      Select subsample $\mathcal{S}_{\tilde{z}^{(i)},j}$ of size k_n from training set that includes $\tilde{z}^{(i)}$
      Build *full* tree using subsample $\mathcal{S}_{\tilde{z}^{(i)},j}$
      Build *reduced* tree using subsample $\mathcal{S}_{\tilde{z}^{(i)},j}$ utilizing only reduced feature space
      Use both full tree and reduced tree to predict at each test point
      Record difference in predictions
    end for
    Record average of the n_{MC} differences in predictions
  end for
  Compute the covariance of the $n_{\tilde{z}}$ averages

Appendix D.

Distribution of Subbagged Predictions using an Internal Variance Estimate

Here we examine the distribution of predictions when the ensemble is built according to the internal variance estimation method described in Algorithm 5. Since we are only interested in the distribution of predictions, we omit the steps in Algorithm 5 for estimating the variance parameters and use the same estimates as in the external case above. For both the SLR case with $n = 1000, k = 60$ and the MARS case with $n = 1000, k = 75$, we use $n_{\tilde{z}} = 50$ and $n_{MC} = 250$ to produce a total of $m = 12500$ predictions in each ensemble. A total of 250 ensembles were built and the resulting histograms with estimated normal densities overlaid are shown in Figure 9 below. We see that these distributions appear to be the same as when the subsamples are selected uniformly at random, as in the external variance estimation method.
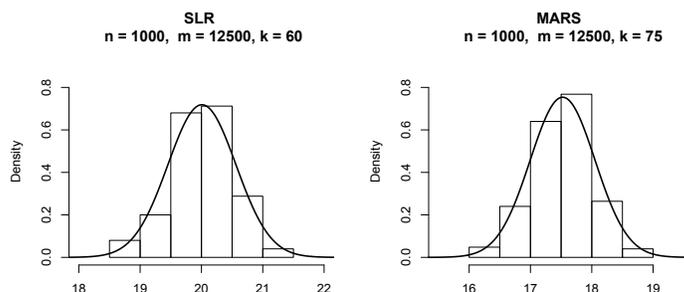
Figure 9: Histograms of subbagged predictions with using an internal estimate of variance. Predictions are made at $x_1 = 10$ in the SLR case and at $x_1 = \cdots = x_5 = 0.5$ in the MARS case.

References

Savina Andonova, Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. A simple algorithm for learning stable machines. In *ECAI*, pages 513–517, 2002.

Gérard Biau. Analysis of a Random Forests Model. *The Journal of Machine Learning Research*, 98888:1063–1095, 2012.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of Random Forests and Other Averaging Classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

Justin Bleich, Adam Kapelner, Edward I George, Shane T Jensen, et al. Variable selection for bart: An application to gene regulation. *The Annals of Applied Statistics*, 8(3): 1750–1781, 2014.

Gunnar Blom. Some Properties of Incomplete U-Statistics. *Biometrika*, 63(3):573–580, 1976.

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

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001a.

Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001b.

Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1st edition, 1984.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Misha Denil, David Matheson, and Nando de Freitas. Narrowing the gap: Random forests in theory and in practice. arXiv:1310.1415v1 [stat.ML], October 2013.

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

Randall L Eubank. *Nonparametric regression and spline smoothing.* CRC press, 1999.

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.

Paul R Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43, 1946.

Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Svante Janson. The Asymptotic Distributions of Incomplete u-Statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(4):495–505, 1984.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Alan J Lee. *U-Statistics: Theory and Practice.* CRC Press, New York, 1990.

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

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

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

Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280, 1971.

R von Mises. On the Asymptotic Distribution of Differentiable Statistical Functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

Stefan Wager. Asymptotic Theory for Random Forests. arXiv:1405.0352 [math.ST], May 2014.

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

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6): 80–83, 1945.

Faisal Zaman and Hideo Hirose. Effect of subsampling rate on subbagging and related ensembles of stable classifiers. In *Pattern Recognition and Machine Intelligence*, pages 44–49. Springer, 2009.

Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, (just-accepted), 2015.