Scalable and Efficient Hypothesis Testing with Random Forests

Tim Coleman, Wei Peng, & Lucas Mentch

Department of Statistics
University of Pittsburgh

April 17, 2019

Abstract

Throughout the last decade, random forests have established themselves as among the most accurate and popular supervised learning methods. While their black-box nature has made their mathematical analysis difficult, recent work has established important statistical properties like consistency and asymptotic normality by considering subsampling in lieu of bootstrapping. Though such results open the door to traditional inference procedures, all formal methods suggested thus far place severe restrictions on the testing framework and their computational overhead precludes their practical scientific use. Here we propose a permutation-style testing approach to formally assess feature significance. We establish asymptotic validity of the test via exchangeability arguments and show that the test maintains high power with orders of magnitude fewer computations. As importantly, the procedure scales easily to big data settings where large training and testing sets may be employed without the need to construct additional models. Simulations and applications to ecological data where random forests have recently shown promise are provided.

Keywords: Random forests, Permutation tests, Variable importance, Exchangeability

1 Introduction

Advances in computing power and big data collection have produced numerous situations in which complex supervised learning methods can drastically outperform more rigid classical statistical models in terms of predictive accuracy. Despite these advances, many such models and algorithms have remained largely impenetrable to traditional statistical analysis. The random forests algorithm (Breiman, 2001) is among the relatively few procedures that remain very popular with a long history of success for which formal statistical properties have recently begun to be developed. Recent exciting progress describing the consistency and asymptotic distribution of random forest predictions have paved the way for the development of statistical inference. As detailed below, however, methods proposed to this point for assessing variable importance have either been *ad hoc* and susceptible to producing misleading results even in simple settings or have come with severe restrictions on the testing framework while incurring extreme computational overhead. The primary goal of this paper is to formally develop a statistically valid permutation test approach that maintains high power with orders of magnitude fewer required computations that scales naturally and efficiently to large data settings.

Permutation tests have their roots in the work of Fisher (1937) using contingency tables. The canonical permutation test framework relies on an assumption of exchangeability of observations, or at least the asymptotic exchangeability of observations. Under this assumption, generally associated with the null hypothesis, the permutation test is conducted as follows. Given two (independent) iid samples $\boldsymbol{X} = X_1, ..., X_n$ and $\boldsymbol{Y} = Y_1, ..., Y_m$, consider the joined sample, $\boldsymbol{Z} = \boldsymbol{X} \uplus \boldsymbol{Y}$, where $\uplus$ indicates concatenation of datasets and let $\mathcal{G}$ be the group of all permutations of the indices $1, ..., m+n$. Then, consider a statistic of the form $T = r(Z_1, ...Z_{m+n})$, and let T_0 be the statistic calculated on the original data. A p-value for the hypothesis the null hypothesis $H_0 : X \overset{d}{=} Y$ is given by

$$p = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} I(|T_0| > |T(g\boldsymbol{Z})|) := 1 - \hat{J}_n(T_0; \boldsymbol{Z}) + \hat{J}_n(-T_0; \boldsymbol{Z})$$

where $\hat{J}_n(\cdot; \boldsymbol{Z})$ is the permutation distribution function, typically referred to as the conditional distribution. To achieve a test with type I error rate α, we reject H_0 if $p < \alpha$. Pesarin and Salmaso (2010) note that this p-value is conditionally unbiased, i.e. $P(p < \alpha | \boldsymbol{Z}, H_0) \leq \alpha$ and $P(p < \alpha | \boldsymbol{Z}, H_1) \geq \alpha$. However, this procedure is *not* typically unbiased for more general hypotheses, such as $H_0^f : \mathbb{E}(f(X_1)) = \mathbb{E}(f(Y_1))$ for some integrable function $f(\cdot)$. As such, many permutation procedures are heuristics for hypotheses like H_0^f that may provide some practical use and intuition, but without verified statistical validity. A main focus of this paper will be providing rigorous results about the asymptotic validity of the proposed testing procedure for hypotheses like H_0^f, utilizing results for exchangeable random variables and asymptotic results for permutation tests.

1.1 Related Work on Permutation Tests

Modern analysis of permutation tests expands upon the works of Hoeffding (1952) and Lehmann et al. (1949), who demonstrates the convergence of the permutation distribution to the sampling distribution for a wide variety of test statistics. This suggests that permutation tests are asymptotically as efficient as parametric tests. Much modern work has focused on extending permutation tests to situations where the data may not be iid or even exchangeable, e.g. (Romano, 1990). Studentization is typically proposed as a means of forcing the sampling distribution of a statistic to converge to a normal distribution (referred to as the *unconditional* distribution), and then it is shown that the permutation distribution converges to the unconditional distribution. This idea has underpinned results in Neuhaus (1993); Janssen (1997, 2005), who provide various sufficient conditions for the convergence to the unconditional distribution.

Permutation tests are exact tests for the hypotheses of equal distribution, under the assumption of iid sequences, but as noted above, are not necessarily valid for more general hypotheses. Convergence to the unconditional distribution ensures that the permutation distribution can be used for a finite sample exact test of equality of distribution and an asymptotically valid test for more general hypotheses. In this work, we will first focus on the exactness property of theses tests and then prove results regarding the asymptotic validity of our procedure for more general hypotheses. The individual models used in ensemble learners, often referred to as weak or base learners, naturally lend themselves to the permutation framework, by being exchangeable in many practical cases, and behave asymptotically like iid sequences, allowing for more tenable theoretical results.

1.2 Related Work on Random Forests

Decision trees recursively partition the covariate space and generate predictions by fitting some simple model – often an average or majority vote – within each resulting region. Of particular interest are the classical **C**lassification **A**nd **R**egression **T**rees (Breiman et al., 1984). CART procedures often have low bias, but tend to overfit to the training data without significant pruning. Bagging predictors (Breiman, 1996) stabilize the variance of weak learning methods by training many weak learners on resampled data, typically offering some gains in predictive performance. Random forests (Breiman, 2001) augment the bagging procedure by introducing auxiliary randomness in the construction of each weak learner, leading to trees with a lower degree of dependence but higher individual variances. Since their introduction, random forests have sustained a long track-record of empirical success in terms of predictive accuracy; see Fernández-Delgado et al. (2014) for a recent large-scale comparison in which random forests outperform nearly all competitors.

The flexibility offered by random forests and other machine learning methods makes theoretical analysis of the statistical properties of these models challenging. Breiman (2001) proposed out of bag (oob) measures for variable importance as a quick heuristic, though a substantial amount of work since has shown that these importance measures are biased towards inflating the importance of correlated variables (Toloşi and Lengauer, 2011; Nicodemus et al., 2010) or variables with many levels (Strobl et al., 2007). Nonetheless, these importance measures remain the most commonly used tool for assessing variable importance in random forests, largely because of their computational feasibility and availability in software. Remedial measures using statistically valid tree fitting were proposed in Hothorn et al. (2006). These conditional inference trees (ctrees) use permutation tests to evaluate the conditional effect of a particular covariate on the response of the distribution and were shown by Strobl et al. (2007) to remedy many pathologies in oob variable importance measures. Further work (Altmann et al., 2010; Janitza et al., 2016) has developed permutation based importance measures but failed to analyze to their statistical validity. These procedures generally involve training many random forests, following much of the standard bootstrap approach to inference, and as such are often computationally intractable. Ishwaran and Lu (2019) provides confidence intervals for the standard oob measures that are valid whenever the random forest is assumed to be L_2 consistent.

Extensions of random forests beyond classification/regression forests have also been proposed. Meinshausen (2006) used the weights learned in a regression forest to perform quantile regression, and proved pointwise consistency of these estimates under mild assumptions. Ishwaran and Lu (2008) suggested using random forests to perform survival analysis, and whose consistency properties were studied in Cui et al. (2017). More recently, Zhu et al. (2015) proposed using trees in a reinforcement learning framework. Finally, Athey et al. (2016) proposed the unifying framework of generalized random forests, which use random forests weights for general local parameter estimation.

Perhaps the most promising line of work has come in recent years by considering ensembles of trees built with subsamples rather than bootstrap samples. Wager et al. (2014) applied the infinitesimal jackknife variance estimate developed in Efron (2014) to produce closed form variance estimates for random forest predictions. Scornet et al. (2015) provided the first consistency results for Breiman's original random forest procedure for additive regression functions. Mentch and Hooker (2016) derived the closed form asymptotic distribution for random forest predictions whenever the subsample size $k = o(\sqrt{n})$. Most recently, Wager and Athey (2017) proved both consistency and asymptotic normality for subsampled random forests for potentially larger subsamples whenever trees are restricted to being built according to *honesty* and *regularity* conditions and large numbers of trees are constructed relative to n.

The asymptotic normality established in Mentch and Hooker (2016) was obtained by cast-

ing random forests as incomplete, infinite-order U-statistics. In addition to establishing normality and providing the closed form asymptotic variance, the authors also lay out a formal hypothesis testing procedure for evaluating variable importance. This test involves drawing a large number of subsamples and constructing two decision trees with each – one on the original data and one where the effect of some features is muted with respect to the response. The difference in predictions then serves as the U-statistic kernel for which the asymptotic distribution can be estimated. The test rejects the null hypothesis that the features are not important when the estimated distribution is centered at some $c > 0$.

This test, though valid, is quite computationally prohibitive. The hypotheses are presumed to be evaluated at predefined test locations in some test set $\mathcal{T}$ and whenever $|\mathcal{T}| = N_t > 1$, calculating the test statistic involves estimating an $N_t \times N_t$ covariance matrix. While asymptotic results are available for situations in which relatively few trees are built, accurate estimation of the covariance necessitates constructing a very large number of trees and becomes computationally infeasible for more than 20-30 test points. Mentch and Hooker (2017) extends the procedure to tests for additivity and provide an alternative approximate test involving random projections that allows the procedure to scale up slightly but with additional computational overhead. Even employing the potentially more efficient infinitesimal jackknife variance estimate utilized in Wager et al. (2014) and Wager and Athey (2017) requires the number of trees constructed be at least on the order of n to be valid.

In contrast, the method we propose here is almost entirely computationally immune to the number of test points employed. The permutation scheme we employ avoids the need for an explicit covariance estimation and thus does not require a larger number of trees for larger datasets. Instead, our hypothesis tests provide valid p-values for variable importance while maintaining the same order of computational complexity as the original random forest procedure. Put simply, if the size and structure of the available data allows for a random forest model to be constructed, our testing procedure can be readily employed.

The remainder of this paper is laid out as follows. In Section 2, we give an overview of the testing procedure, and further highlight its benefits over existing methods. In Section 3, we present results regarding the statistical properties of the proposed test, namely that it attains validity for the desired hypotheses. This follows from first deriving the unconditional distribution of random forest squared errors and then demonstrating convergence of the permutation distribution to the unconditional distribution. In Section 4, we present simulation studies of the testing procedure for a variety of underlying regression functions. In Section 5, we apply our procedure to multiple ecological datasets where random forests have been successfully employed in applied work. In addition to the main text, all technical proofs are provided in Appendix A, and additional simulations demonstrating the robustness of the proposed procedure are presented in Appendix B.

2 Overview of the Testing Procedure

2.1 Set Up

Here, we detail the general framework considered in the remainder of this paper. Consider a sample $\mathcal{D}_n = \{Z_1, Z_2, ..., Z_n\}$, with $Z_k = (\boldsymbol{X}_k, Y_k)$ consisting of observations on covariates $\boldsymbol{X} = (X_1, ..., X_p) \in \mathcal{X}$ and a response $Y \in \mathcal{Y}$. In this work, it is assumed that $Z_k \overset{iid}{\sim} F$ where F is some distribution with support on $\mathcal{X} \times \mathcal{Y}$. In the regression context, we assume that

$$Y = m(\boldsymbol{x}) + \epsilon$$

where $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$ and ϵ is an independent noise process, typically with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) < \infty$. The goal of random forest is to accurately estimate $m(\boldsymbol{x})$. A subsampled random forest is constructed by drawing subsamples, of size $k_n < n$, from $\mathcal{D}_n$, drawing a randomization parameter ξ from some distribution Ξ, and constructing a randomized decision tree on the subsample. This process is repeated B times and the random forest prediction at some $\boldsymbol{x} \in \mathcal{X}$ is given by

$$RF_{B,k_n}(\boldsymbol{x}) = \frac{1}{B}\sum_{j=1}^{B} T_{j,k_n}(\boldsymbol{x}; \xi_j; \mathcal{D}_n). \tag{2.1}$$

We can similarly evaluate the RF prediction accuracy at a test point (not in $\mathcal{D}_n$), with known location $\boldsymbol{x}$ and response y via its mean squared error

$$MSE_{RF}(\boldsymbol{x}; Y, \mathcal{D}_n) := \left(\frac{1}{B}\sum_{j=1}^{B} T_{k_n,j}(\boldsymbol{x}) - y\right)^2.$$

In practice, there are $\binom{n}{k_n} \times |\Xi|$ possible trees, where $|\Xi|$ denotes cardinality of the set of possible randomization elements. For sufficiently large B, Equation 2.1 can be made arbitrarily close to $\mathbb{E}_\Xi(RF_{B,k_n}(\boldsymbol{x})|\mathcal{D}_n)$, where the expectation is taken over the subsampling and randomization distributions, conditional on the data. Equation 2.1 can be interpreted as a Monte Carlo approximation to the infinite forest. The effect of this approximation is studied in Scornet (2016).

Similarly, we can calculate the MSE of a forest at a collection of test points

$$\mathcal{T} = [(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{N_t}, y_{N_t})]$$

as $MSE_{RF}(\mathcal{T}) = \frac{1}{N_t}\sum_{\ell=1}^{N_t} MSE_{RF}(\boldsymbol{x}_\ell; Y_\ell, \mathcal{D}_n)$. Let RF^π be defined similarly to Equation 2.1, but with $\mathcal{D}_n$ replaced by $\mathcal{D}_n^\pi$, where $\mathcal{D}_n^\pi$ replaces X_j with an alternate copy drawn

independent of Y given the rest of the covariates. We then evaluate whether a variable X_j is important by conducting a test of the following hypotheses

$$\begin{aligned} H_0^j &: \mathbb{E}(MSE_{RF}(\mathcal{T})) = \mathbb{E}(MSE_{RF^\pi}(\mathcal{T})) \\ H_1^j &: \mathbb{E}(MSE_{RF}(\mathcal{T})) < \mathbb{E}(MSE_{RF^\pi}(\mathcal{T})) \end{aligned} \tag{2.2}$$

where the expectation is taken over the training data and the auxillary randomness, but is conditional on $\mathcal{T}$. We call X_j important if we are able to reject H_0^j. This definition of importance is model based and therefore less general than alternative definitions such as that utilized in the recent *knockoff* literature (Barber et al., 2015; Candes et al., 2016), where a variable X_j is deemed unimportant if

$$(Y \perp\!\!\!\perp X_j) \mid \boldsymbol{X}_{-j}.$$

It should be noted that the above is neither necessary nor sufficient for H_0^j. However, in practice, model based inference is the default approach in numerous applications and also the context in which testing with knockoffs is generally carried out. If variable X_j is expensive to measure, or only available for a subset of subjects, it may be practically advantageous to determine if X_j is important for an accurate model. Moreover, for the prediction problem in finite samples, a random forest may provide optimal generalization error, and as such may appear to provide the best available estimate of the conditional mean function for the given data.

2.2 Testing Procedure

Here we present the procedure for comparing the mean squared errors of two random forests. We first provide an overview and heuristic justification for the test, followed by a more technical discussion of its validity.

In general, if two randomized ensemble methods produce predictions that are similarly accurate, then the permutation distribution of discrepancies in accuracy should be centered around 0. In our particular setting for testing feature significance, we compare the accuracy of two random forests built on different data. For a given (original) dataset $\mathcal{D}_n$, we first construct $\mathcal{D}_n^\pi$ in such a way so as to remove any dependence the response may have on these features. However, rather than permuting the data and retraining entire random forests, we first train trees on both $\mathcal{D}_n$ and $\mathcal{D}_n^\pi$ separately, record predictions at the test locations, and then permute the predictions between the forests. The new forests formed at each iteration thus consist of some trees built on the original data and some built with the permuted counterpart so that on average, we should expect the accuracy both forests to be the same. In this light, the testing procedure can be seen as directly analogous to a classic permutation test to evaluate equality in distribution across two groups. In

particular, note that this procedure requires only $2B$ trees regardless of the size of the test set – a substantial computational improvement over existing procedures such as those in Coleman et al. (2017) and Mentch and Hooker (2016) that each require training enormous numbers of trees, with growing numbers needed for larger test sets.

Pseudo-code for the permutation test is provided in Algorithm 1. We use $\oplus$ to denote concatenation of data matrices by column, and $\uplus$ to denote concatenation by row. $\ominus$ denotes removal of columns from a dataframe, and $||\cdot||_2$ is the standard L_2 norm. In order to prevent p-values exactly equal to 0, we add 1 to the numerator and denominator, ensuring that under H_0 the p-values are stochastically larger than uniform random variables. In all, this suffices to make the testing procedure slightly more conservative, but more amenable to potential p-value transforming procedures. Crucially, note that this procedure requires no explicit variance estimation of the N_t predictions made by individual forests, thereby providing a dramatic computational speed-up over existing parametric approaches (Mentch and Hooker, 2016, 2017) that require the estimation of a $N_t \times N_t$ covariance matrix.

Algorithm 1: Permutation test pseudocode for variable importance

Data: Training data $\mathcal{D}_n$ test sample $(\mathcal{T} = [(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{N_t}, y_{N_t})])$, specified feature(s) of interest, $\boldsymbol{X}_S$

Result: p-value, $\tilde{p}$ for importance of $\boldsymbol{X}_S$ at points in $\mathcal{T}_n$

SET number of permutations n_{perm}, subsample size k_n, and $n_{tree} = B$;

DEFINE $\boldsymbol{X}_S^\pi$ by permuting the rows of $\mathcal{D}_n$ and selecting the columns corresponding to $\boldsymbol{X}_S$;

DEFINE $\mathcal{D}_n^\pi = \mathcal{D}_n \ominus \boldsymbol{X}_S \oplus \boldsymbol{X}_S^\pi$;

for i *in* $\{1, ..., B\}$ **do**

  SAMPLE k_n rows from $\mathcal{D}_n$: $\mathcal{D}_i^* = \{Z_{i,1}^*, ..., Z_{i,k_n}^*\}$;

  SAMPLE k_n rows from $\mathcal{D}_n^\pi$: $\mathcal{D}_i^{*\pi} = \{Z_{i,1}^{*\pi}, ..., Z_{i,k_n}^{*\pi}\}$;

  TRAIN trees $T_i(\cdot)$ on $\mathcal{D}_{i,k_n}^*$ and $T_i^\pi(\cdot)$ on $\mathcal{D}_{i,k_n}^{*\pi}$;

  PREDICT at $\mathcal{T}_n$ using T_i, T_i^π, generating $\boldsymbol{T}_i = [T_i(\boldsymbol{x}_1), ..., T_i(\boldsymbol{x}_{N_t})]$ and $\boldsymbol{T}_i^\pi = [T_i^\pi(\boldsymbol{x}_1), ..., T_i^\pi(\boldsymbol{x}_{N_t})]$

end

CALCULATE $MSE_0 = \left|\left|\frac{1}{B}\sum_{i=1}^B \boldsymbol{T}_i - \boldsymbol{y}\right|\right|_2^2$ and $MSE_0^\pi = \left|\left|\frac{1}{B}\sum_{i=1}^B \boldsymbol{T}_i^\pi - \boldsymbol{y}\right|\right|_2^2$;

for j *in* $\{1, ..., n_{perm}\}$ **do**

  SAMPLE $\boldsymbol{T}_{j,1}^*, ..., \boldsymbol{T}_{j,B}^*$ from $\{\boldsymbol{T}_1, ...\boldsymbol{T}_B, \boldsymbol{T}_1^\pi, ..., \boldsymbol{T}_B^\pi\}$ without replacement, call the B remaining trees $\boldsymbol{T}_{j,1}^{*\pi}, ..., \boldsymbol{T}_{j,B}^{*\pi}$;

  CALCULATE $MSE_j^* = \left|\left|\frac{1}{B}\sum_{l=1}^B \boldsymbol{T}_{j,l}^* - \boldsymbol{y}\right|\right|_2^2$ and $MSE_j^{*\pi} = \left|\left|\frac{1}{B}\sum_{l=1}^B \boldsymbol{T}_{j,l}^{*\pi} - \boldsymbol{y}\right|\right|_2^2$

end

CALCULATE $\tilde{p} = \frac{1}{N_0+1}\left[1 + \sum_{j=1}^{N_0} I\big((MSE_0^\pi - MSE_0) > (MSE_j^{*\pi} - MSE_j^*)\big)\right]$

3 Theoretical Justification

We now develop the theoretical backing for the hypothesis testing procedure outlined above. In particular, we establish asymptotic normality for subsampled random forest predictions under mild conditions and extend these results to general functionals of the predictions. As noted above, to use these distributions directly is unwieldly - estimation of the variance parameters often consumes an unrealistic amount of computational power. For our procedure, we need only that the permutation distribution of test statistics approaches the unconditional sampling distribution, which we establish via arguments similar to Chung and Romano (2013). For readability, proofs are reserved for Appendix A.

3.1 Exchangeable Random Variables & Permutation Tests

We begin by recalling the definition of an exchangeable sequence of random variables:
Definition 1. *(Exchangeable Sequences)* A sequence of random variables $X_1, X_2, ...$ is called *exchangeable* if for every finite sub collection indexed by $i_1, ..., i_k$

$$(X_{i_1}, X_{i_2}, ..., X_{i_k}) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, ..., X_{\pi(k)})$$

where $\pi(\cdot)$ is any of the $k!$ permutations of the indices.

The notion of exchangeability can be extended to general probability spaces (where the random quantities may be infinite dimensional) by replacing the distribution function with the probability measure; it can also be equivalently stated as distributional invariance under the elements of the group of permutation matrices. See Aldous (1985) for a thorough review of related properties. An especially important result regarding exchangeable sequences is De Finetti's theorem (De Finetti, 1937), which gives that infinite sequences of exchangeable random variables are a mixture of an iid sequence.

Permutation tests naturally lend themselves to exchangeable data by providing a means of evaluating the fundamental hypothesis that the joint distribution of a collection of random variables is invariant under permutations. They maintain exactness for the null hypothesis whenever $X_i \overset{iid}{\sim} P$ and independently $Y_j \overset{iid}{\sim} Q$ because the joint measure of the data factorizes as

$$\mu(\boldsymbol{X}, \boldsymbol{Y}) = \prod_{i=1}^{n} P(X_i) \prod_{j=1}^{m} Q(Y_j)$$

which is invariant to permutations of observations, and thus exchangeable, if and only if $P = Q$.

Much of the modern work for permutation tests has focused on modifications to account for pathologies due to violations of the exchangeability assumption. Chung and Romano

(2013) propose a studentization of the permutation test statistic when conducting inference a functional of two distributions. Consider, for example, $X_1, ..., X_n \overset{iid}{\sim} F_X = \mathcal{N}(0, 5)$ and independently let $Y_1, ..., Y_m \overset{iid}{\sim} F_Y = \mathcal{N}(0, 1)$. Clearly, median$(F_X)$ = median(F_Y), but the data are no longer exchangeable, and so an unstudentized permutation test of H_0 : median(F_X) = median(F_Y) is no longer valid at a pre-specified level (see Chung and Romano (2013) Section 4 for a demonstration of this phenomenon). However, note that this assumption is violated only because the data are no longer identically distributed. On the other hand, permutation tests can remain valid for data that are correlated but identically distributed so long as the pairwise dependence is constant.

Theorem 1. *Under the conditions outlined above, denote a sequence of (potentially randomized) trees trained on subsamples from* $\mathcal{D}_n$ *as* $\{T_k(\cdot)\}_1^\infty$. *Moreover, consider an independently drawn test point,* $\boldsymbol{Z}^* = (\boldsymbol{X}^*, Y^*) \sim F$. *Then, the residuals, defined as:*

$$r_k = T_k(\boldsymbol{X}) - Y^*$$

form an infinitely exchangeable sequence of random variables.

In the case of a single ensemble of trees in a random forest, for example, exchangeability is readily apparent as the order in which trees are trained has no bearing on their predicted values. Indeed, Theorem 1 can be extended to any bagged learning method where the randomization parameters are iid including bootstrapped random forests or model-averaged neural networks where the networks are generated in an iid fashion, such as with the randomized dropping out of nodes proposed in Krizhevsky et al. (2012).

Two Forest Case

The primary goal of this work is to identify covariates that produce statistically significant improvements in model accuracy. To assess this, we consider building two forests, one on the original dataset $\mathcal{D}_n$ and another on a second dataset $\mathcal{D}_n^\pi$ wherein the covariate(s) of interest $\boldsymbol{X}_S$ are altered so as to mute their potential effect on the response. This muting can be achieved in various ways:

- Outright exclusion: The covariates in $\boldsymbol{X}_S$ are simply removed from the second training dataset.
- Random permutation: Each covariate in $\boldsymbol{X}_S$ is randomly shuffled so that $\boldsymbol{X}_S$ is replaced by some permuted alternative $\boldsymbol{X}_S^\pi$ in the second training dataset.
- Knockoffs: Each covariate in X_i in $\boldsymbol{X}_S$ is replaced by some knockoff alternative X_i^π sampled from the distribution of $X_i|\boldsymbol{X}_{-i}$ so that $\boldsymbol{X}_S$ is replaced by a randomized alternative $\boldsymbol{X}_S^\pi$ in the second training dataset. See Candes et al. (2016) for details.

Consider two samples, $\mathcal{D}_n^1$ and $\mathcal{D}_n^2$ drawn independently from the same population, and a collection of subsampled trees, say $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, trained on each. Then, $\boldsymbol{T}_1 \stackrel{d}{=} \boldsymbol{T}_2$, but the trees are no longer exchangeable, because the conditioning random vector (i.e. the training data) is different and thus in general the within sample correlation between trees will be higher than the between sample correlation. In practice, the trees are *approximately* exchangeable - higher within sample correlations tend to die out when the subsample size grows slower than n; see Lemma 1 below. Provided that the tree distributions converge weakly to some distribution, we can establish the notion of *asymptotic exchangeability*. This idea is well studied, if rarely explicitly mentioned - Romano (1990); Chung and Romano (2013) provide conditions for permuted test statistics to be asymptotically exchangeable. Good (2002) provides further examples of asymptotic exchangeability, particularly in an ANOVA context. Moreover, recall that Breiman's original motivation for adding the additional randomness in random forests was to force the tree predictions to be less correlated, which further dampens the effect of higher within-sample correlation.

In practical terms, this means that replacing the covariates under investigation with a randomized counterpart – either a permutation or knock-off – leads to a procedure with more desirable properties than when variables of interest are simply dropped, creating a lower-dimensional covariate space. In particular, to attain near-exchangeability under the null hypothesis, the individual trees should be constructed in a near-identical fashion. This includes ensuring that the trees are trained on the same dimension feature space, on the same subsample size, and with other various tree-specific parameters controlled.

In terms of the test described in Section 2, we consider an original dataset, $\mathcal{D}_n$ with $n \times p$ design matrix $\boldsymbol{X}$. As above, let $\mathcal{S} \subset \{1, ..., p\}$ and define $\boldsymbol{X}_{\mathcal{S}} = \{X_j : \ j \in \mathcal{S}\}$ and $\boldsymbol{X}_{-\mathcal{S}} = \{X_j : \ j \notin \mathcal{S}\}$ where we take $\boldsymbol{X}_{\mathcal{S}}$ to be the set of covariates we are interested in testing for predictive significance. We then create a randomized version of $\boldsymbol{X}_{\mathcal{S}}$ independent of Y, typically via either random permutation or by generation of knockoffs, which we denote by $\boldsymbol{X}_{\mathcal{S}}^{\pi}$. Note in particular that when the entire joint density $P(\boldsymbol{X})$ of the covariates is known, Algorithm 1 of Candes et al. (2016) can be used to generate the knockoffs that make up $\boldsymbol{X}_{\mathcal{S}}^{\pi}$ which then ensures that $[\boldsymbol{X}_{-\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}}] \stackrel{d}{=} [\boldsymbol{X}_{-\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}}^{\pi}]$.

Note that by construction, $\tilde{\boldsymbol{X}}_{\mathcal{S}} \perp\!\!\!\perp Y | \boldsymbol{X}_{-\mathcal{S}}$ and consequently, if we now replace $\boldsymbol{X}_{\mathcal{S}}$ with $\boldsymbol{X}_{\mathcal{S}}^{\pi}$ in the design matrix to form a new training dataset $\mathcal{D}_n^{\pi}$, then the trees trained on $\mathcal{D}_n^{\pi}$ inherit the conditional independence

$$T(\boldsymbol{x}; \mathcal{D}_n^{\pi}) \perp\!\!\!\perp Y | \ \boldsymbol{X}_{-\mathcal{S}}.$$

Under the assumption (null hypothesis) that $\boldsymbol{X}_{-\mathcal{S}}$ are the only predictively important covariates, we would expect predictions from trees trained on $\mathcal{D}_n$ to have the same distribution as those generated from trees trained on $\mathcal{D}_n^{\pi}$. As such, the trees should be asymptotically exchangeable between forests and it follows that we can test this exchangeability assump-

tion via a permutation test. The mean squared error on the test set can be used in a permutation test of the following hypotheses:

$$\begin{aligned} H_0: &\quad T(\boldsymbol{x}; \mathcal{D}_n) \stackrel{d}{=} T(\boldsymbol{x}; \mathcal{D}_n^\pi) \\ H_1: &\quad T(\boldsymbol{x}; \mathcal{D}_n) \stackrel{d}{\neq} T(\boldsymbol{x}; \mathcal{D}_n^\pi). \end{aligned} \tag{3.1}$$

Note that unlike traditional permutation tests, this particular test does not maintain the exactness property due to the higher within sample dependence of the trees in each training dataset. Even when $T(\boldsymbol{x}; \mathcal{D}_n)$ has the same distribution as $T(\boldsymbol{x}; \mathcal{D}_n^\pi)$, the joint distribution for the two groups of trees may not be invariant under permutations, because the dependency structure is different between groups vs within groups. We can mitigate this by forcing the trees to be more independent by, for example, building with relatively small subsamples. Empirical results presented in later sections, demonstrate that the deviation from exactness is quite small, even for small sample sizes, corroborating the theory that the asymptotic sequence of trees behaves like an iid sequence.

3.2 Unconditional Sampling Distribution of the MSE

Within-forest exchangeability of trees is not sufficient to justify the proposed permutation procedure for testing the null hypothesis $\mathbb{E}(MSE_{RF}(\mathcal{T})) = \mathbb{E}(MSE_{RF^\pi}(\mathcal{T}))$ at the nominal level. Instead, we need to consider comparisons of trees from two forests, and establish conditions to justify exchanging trees between forests. An important step in this direction is to establish the existence of a limiting sequence of sub-bagged trees that behave like an iid sequence as such a sequence is much more amenable to theoretical analysis.

Condition 1. *There exists a random function* T_∞ *such that* $\lim_{n\to\infty} T_{k_n} \stackrel{d}{=} T_\infty$

For now we leave this as an assumption; in future sections we provide sufficient conditions for this to be satisfied. Note that this condition is very similar in spirit to Assumption 15.7.1 in Lehmann and Romano (2006), which is fundamental to the validity subsampling based intervals for model parameters.

In practice, larger sample sizes naturally lend themselves to larger subsample sizes. As such, we would like to establish theoretical results for random forests trained on growing subsamples. If we impose the standard condition that the subsample size k_n grow slower than $\sqrt{n}$, we obtain the following intuitive result.

Lemma 1. *Consider a collection of* B_n *trees built from a training dataset of size* n *on subsamples of size* k_n*, say* $\{T_{j,k_n}\}_{j=1}^{B_n}$*, satisfying Condition 1. Then, as long as* $k_n/\sqrt{n} \to 0$ *and*

$$\binom{B_n}{2} \log\left[\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right] \to 0$$

the infinite sample sequence of trees, $\{T_{1,\infty,k_\infty}, ..., T_{B,\infty,k_\infty}, ...\}$ *is an infinite sequence of pairwise independent random functions.*

The condition placed on the number of trees built, B_n, is likely not of much practical importance. For finite B_n, the probability sequence has the form of a_n^K, so because $a_n \to 1$, a_n^K also converges to 1. However, if we let B_n grow with n, the number of trees may overwhelm the independence induced by subsampling. Thus, we must let the log probability of an individual pair being independent go to 0 faster than $\binom{B_n}{2} \approx B_n^2/2$ goes to infinity.

Lemma 1 establishes asymptotic pairwise independence, but not that the limiting sequence is iid. Intuition here can also be somewhat misleading - it would seem that across a broad collection of trees trained on many relatively small subsamples, that the trees may not be independent on the whole. However, this turns out to not be the case. We now recall a result from Aldous (1985).

Lemma 2. *(Aldous, 1985) Let* $Z_1, Z_2, ...$ *be an infinitely exchangeable sequence. Then, if the* Z_j *are pairwise independent, the sequence is a sequence of iid random variables.*

An immediate consequence of the preceding lemmas is the following corollary.

Corollary 1. *Let* $\{T_{j,k_n}\}_{j=1}^{B_n}$ *be a collection of* B_n *trees trained on subsamples from* $\mathcal{D}_n$*, satisfying the conditions of Lemma 1. Then,* $\{T_{j,\infty}\}_{j=1}^{\infty} := \lim_{n\to\infty}\{T_{j,k_n}\}_{j=1}^{B_n}$ *is an iid sequence of functions.*

The infinite sequence of subsampled trees enjoys many properties that the finite sequence does not. In particular, we are able to obtain a pointwise central limit theorem under the further restriction that tree predictions have finite, non-zero variance.

Corollary 2. *Let* $\{T_{j,k_n}\}_{j=1}^{B_n}$ *be a sequence of trees on subsamples from* $\mathcal{D}_n$*, satisfying the conditions of Lemma 1 and Condition 1. Further, assume* $\boldsymbol{x} \in \mathcal{X}$ *is such that* $0 < Var(T_\infty(\boldsymbol{x})) = \sigma^2(\boldsymbol{x}) < \infty$*. Then*

$$\sqrt{B_n}\left[\frac{1}{B_n}\sum_{i=1}^{B_n} T_{i,k_n}(\boldsymbol{x}) - \mathbb{E}\left(\frac{1}{B_n}\sum_{i=1}^{B_n} T_{i,k_n}(\boldsymbol{x})\right)\right] \xrightarrow{d} \sigma(\boldsymbol{x})Z \tag{3.2}$$

as $n \to \infty$*, and* $Z \sim \mathcal{N}(0,1)$ *.*

Corollary 2 follows directly from applying the Central Limit Theorem to the sequence of univariate random variables $\{T_{j,\infty}(\boldsymbol{x})\}_{j=1}^{\infty}$, which are iid by Corollary 1.

Remark: Corollary 2 applies to any ensemble model whose components are asymptotically independent. There has been extensive research into establishing central limit theorems for exchangeable sequences, such as Chow and Teicher (2012) and Klass and Teicher (1987). Thus, the regularity conditions required for the trees to be iid are likely stronger than

needed and a central limit theorem may be possible to obtain in a wider variety of cases.

Remark: For a collection of test points, $\boldsymbol{x}_1, ...\boldsymbol{x}_{N_t}$, we can also consider the sequence of vectors $\boldsymbol{T}_{i,k_n} = [T_{i,k_n}(\boldsymbol{x}_1), ..., T_{i,k_n}(\boldsymbol{x}_{N_t})]^T$, which are iid by Corollary 1. If we assume that $\Sigma = \mathbb{E}\big[(\boldsymbol{T}_{i,k_n} - \mathbb{E}(\boldsymbol{T}_{i,k_n}))(\boldsymbol{T}_{i,k_n} - \mathbb{E}(\boldsymbol{T}_{i,k_n}))^T\big]$ has finite entries, the multivariate central limit theorem gives that

$$\sqrt{B_n}\left[\frac{1}{B_n}\sum_{i=1}^{B_n}\boldsymbol{T}_{i,k_n} - \mathbb{E}\left(\frac{1}{B_n}\sum_{i=1}^{B_n}\boldsymbol{T}_{i,k_n}\right)\right] \xrightarrow{d} Z_\Sigma \sim \mathcal{N}(0, \Sigma)$$

as $n \to \infty$. Along with Corollary 2, this result is similar in spirit to the results of Mentch and Hooker (2016) and Wager and Athey (2017) in establishing asymptotic normality of random forest predictions.

The notation of Equation 3.2 is to indicate that the CLT applies to the limiting tree predictions $T_\infty(\boldsymbol{x})$. Further, we intentionally leave $\sigma(\boldsymbol{x})$ as an abstraction; as noted above, direct estimation of $\sigma(\boldsymbol{x})$ is not necessarily straightforward and may incur substantial computational burden. Instead, this result will be used as the basis for asymptotic validity of our permutation test which, uncharacteristically, is far more computationally efficient.

Further, variance estimation issues aside, the distribution in Equation 3.2 cannot serve as the basis for valid inference about the underlying regression function, $m(\boldsymbol{x})$. While this result guarantees that the variance of random forests vanishes, it makes no comment about the bias of either an individual tree or forest prediction. Much work, e.g. Meinshausen (2006); Biau (2012); Scornet et al. (2015); Wager and Athey (2017) has been put into understanding the bias of random forests. Regardless, it should be no surprise that the variance of a subsampled random forest vanishes; Friedman et al. (2010) provide the following variance calculation for a random forest prediction

$$\text{Var}[RF_{B,k_n}(\boldsymbol{x})] = \rho_n(\boldsymbol{x})\text{Var}[T_{k_n}(\boldsymbol{x})] + \frac{1-\rho_n(\boldsymbol{x})}{B}\text{Var}[T_{k_n}(\boldsymbol{x})]$$

where $\rho_n(\boldsymbol{x}) = \text{Cor}[T_{k_n}(\boldsymbol{x}), T'_{k_n}(\boldsymbol{x})]$ is the theoretical correlation between two trees trained on subsamples of size k_n from the same dataset. Lemma 1 guarantees that $\rho_n(\boldsymbol{x}) \to 0$, as $n \to \infty$, so that the variance of the forest vanishes.

Any transformation of tree predictions is also an iid sequence, as $n \to \infty$. In particular, the residual of tree predictions, $r(T(\boldsymbol{x}); Y) = (Y - T(\boldsymbol{x}))$ are also asymptotically iid, conditional on the observed Y in the test set. Typically, we are interested in the generalization error of a random forest, which is the expected loss, with expectation taken over both test and training sets. In the following discussion, we consider the test set known and fixed, so that the expectations are taken over just the distribution of the training data and all

randomization parameters of the models. In practice, this is estimated by holding out a portion of the data from the training process, and then calculating an error rate on the held out observations.

Unfortunately, the conditional MSE is *not* a linear function of exchangeable random variables – we can expand the conditional MSE as

$$MSE_{RF}(\boldsymbol{x}; Y) = \frac{1}{B^2}\sum_{j=1}^{B}\sum_{i=1}^{B} T_j(\boldsymbol{x})T_i(\boldsymbol{x}) - \frac{2Y}{B}\sum_{j=1}^{B} T_j(\boldsymbol{x}) + Y^2. \tag{3.3}$$

The first term can be seen as a V-statistic (Mises, 1947), and the second term as a linear rescaling of the quantity described in Equation 2.1. We now proceed to derive the asymptotic (in B) distribution of Equation 3.3. To do this, we consider the asymptotic case, so that the sequence of tree predictions behaves like an iid sequence. Further, in the infinite sample case, the number of trees can be made arbitrarily large, and so we allow B to go to infinity with the understanding that it does so in such a way that respects the requirements of Lemma 1. This is mostly a matter of notation - we could include the dependence on n in each of the following statements and stress that the limiting distributions only hold as $n \to \infty$. With these conventions, we will show that the quantity in Equation 3.3 has an asymptotically Gaussian distribution, albeit one whose parameters are not easily estimated. To begin, write the expanded MSE as

$$MSE_{RF}(\boldsymbol{x}; Y) = V_{B,2}(\boldsymbol{x}) - 2Y\,RF_B(\boldsymbol{x}) + Y^2.$$

We analyze each term separately. Corollary 2 gives that the second term is asymptotically Gaussian, i.e. $-2Y\sqrt{B}(RF_B(\boldsymbol{x}) - \mathbb{E}(RF_B(\boldsymbol{x})) \xrightarrow{d} \mathcal{N}(0, 4Y^2\sigma^2(\boldsymbol{x}))$. Now, we want to calculate the asymptotic distribution of

$$\sqrt{B}[V_{B,2}(\boldsymbol{x}) - \mathbb{E}(V_{B,2}(\boldsymbol{x}))].$$

Much of the theory of V-statistics is rooted in the study of their corresponding U-statistics (Mises, 1947; Hoeffding, 1948; Van der Vaart, 2000). As such, it is easier to analyze the corresponding U-statistic

$$U_{B,2} = \frac{1}{\binom{B}{2}}\sum_{1\le i<j\le B} T_j(\boldsymbol{x})T_i(\boldsymbol{x}).$$

The quantity $U_{B,2}$ is an unbiased estimate of $\mathbb{E}(T(\boldsymbol{x}))^2$, or the square of the mean tree prediction at $\boldsymbol{x}$. Standard results due to Hoeffding (1948) regarding U-statistics give that, under the further assumption that $\mathbb{E}[T_j^2(\boldsymbol{x})T_i^2(\boldsymbol{x})] < \infty$

$$\sqrt{B}(U_{B,2} - \mathbb{E}(T(\boldsymbol{x}))^2) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}(T(\boldsymbol{x}))^2\sigma^2(\boldsymbol{x}))$$

as $B \to \infty$. Further, in Appendix A, we show that $\sqrt{B}(V_{B,2} - U_{B,2}) \xrightarrow{p} 0$, so that $\sqrt{B}V_{B,2}$ has the same limiting distribution as $\sqrt{B}U_{B,2}$. Care must be taken to ensure that the quantities in Equation 3.3 are jointly asymptotically Gaussian. To do this, we introduce a mild assumption on the quantities used in MSE_{RF}. We refer to this condition as *ellipticity* of the U-statistics, to suggest that the level curves of the joint distribution of the U-statistics roughly follows an ellipse.

Definition 2. *(Elliptic U-Statistics)* Let $\left\{[T_{1,i}, ..., T_{m,i}]^T\right\}_{i=1}^{B} \overset{iid}{\sim} P$ be a random sample of vectors from a distribution P, and let $\phi_1, ..., \phi_k$ (with $\mathbb{E}_{X_1,...,X_{r_l} \sim P_j}(\phi_l^2(X_1, ..., X_{r_l})) < \infty$ for each l and each marginal P_j of P) be permutation symmetric kernel functions, with fixed but perhaps distinct orders $r_1, ..., r_k$, so that each has a corresponding U-statistic

$$U_{l,B}(X_1, ..., X_B) = \frac{1}{\binom{B}{r_l}} \sum_{1 \le i_1 < ... < i_{r_l} \le B} \phi_l(X_1, ..., X_{r_l}).$$

We say that $[U_l]_{l=1}^{k}$ is *elliptic* for the random vector $\left\{[T_{1,i}, ..., T_{m,i}]^T\right\}_{i=1}^{B}$ if the corresponding $km \times 1$ vector of Hájek projections

$$\tilde{U}_{k,m,B} = \begin{bmatrix} \tilde{U}_1(T_{1,1}, ..., T_{1,B}) \\ \vdots \\ \tilde{U}_k(T_{1,1}, ..., T_{1,B}) \\ \vdots \\ \tilde{U}_k(T_{m,1}, ..., T_{m,B}) \end{bmatrix} \tag{3.4}$$

satisfies $\mathbb{E}\left[\left[\tilde{U}_{k,m,B} - \mathbb{E}(\tilde{U}_{k,m,B})\right]\left[\tilde{U}_{k,m,B} - \mathbb{E}(\tilde{U}_{k,m,B})\right]^T\right] = \Sigma$ where Σ is a positive definite matrix with finite entries.

The utility of this condition is summarized in the following lemma.

Lemma 3. *Let $[U_l]_{l=1}^{k}$ be elliptic for $\left\{[T_{1,i}, ..., T_{m,i}]^T\right\}_{i=1}^{B} \overset{iid}{\sim} P$. Then, the vector $U_{k,m,B}$, defined analogously to $\tilde{U}_{k,m,B}$ in Equation 3.4, satisfies*

$$\sqrt{B}[U_{k,m,B} - \mathbb{E}(U_{k,m,B})] \xrightarrow{d} Z_\Sigma \sim \mathcal{N}(0, \Sigma) \quad \text{as } B \to \infty.$$

The main takeaway is that if we assume ellipticity of $RF_B(\boldsymbol{x})$ (which is a U-statistic of order 1) and $U_{B,2}(\boldsymbol{x})$ with respect to the random vectors $\{[T_i(\boldsymbol{x})]\}_{i=1}^{B}$ we can conclude that, because $\sqrt{B}(V_{B,2}(\boldsymbol{x}) - U_{B,2}(\boldsymbol{x})) \xrightarrow{p} 0$, as $B \to \infty$

$$\sqrt{B}\left[\begin{pmatrix} RF_B(\boldsymbol{x}) \\ V_{B,2}(\boldsymbol{x}) \end{pmatrix} - \mathbb{E}\begin{pmatrix} RF_B(\boldsymbol{x}) \\ V_{B,2}(\boldsymbol{x}) \end{pmatrix}\right] \xrightarrow{d} \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2(\boldsymbol{x}) & \rho \\ \rho & \mathbb{E}(T(\boldsymbol{x}))^2\sigma^2(\boldsymbol{x}) \end{bmatrix}\right]$$

where ρ is the limiting covariance between the V-statistic and random forest estimate. Now, if ρ satisfies

$$|\rho| < \sqrt{\sigma^2(\boldsymbol{x})}\sqrt{\mathbb{E}(T(\boldsymbol{x}))^2\sigma^2(\boldsymbol{x})}$$

then

$$\sqrt{B}\big(MSE_{RF}(\boldsymbol{x};Y) - \mathbb{E}(MSE_{RF}(\boldsymbol{x};Y))\big) \xrightarrow{d} \left[\sqrt{4Y^2\sigma^2(\boldsymbol{x}) + \mathbb{E}(T(\boldsymbol{x}))^2\sigma^2(\boldsymbol{x}) - 2\rho}\right] Z \quad (3.5)$$

where $Z \sim \mathcal{N}(0,1)$.

Remark. The limiting distribution in Equation 3.5 is quite intractable. It requires knowledge of the variance of a tree prediction, the expected value of the tree prediction, and the correlation between the random forest prediction and a V-statistic of the trees. There may be approaches to estimating this but as with the various quantities discussed above, accurate estimation would require extreme computational overhead. We will later show that the permutation distribution approaches this unconditional distribution, so that we can avoid this potentially expensive computation.

Recall that one of the major advantages of the procedure described in Section 2 is the ability to compare random forest accuracies at large numbers of test points. Consider a situation where instead of predicting at a single point, we make predictions at many test points, say $\mathcal{T} = \{(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_{N_t}, Y_{N_t})\}$. Applying a random forest to each of these points and calculating the squared error for each point yields the vector

$$MSE_{RF}(\mathcal{T}) = (MSE_{RF}(\boldsymbol{x}_1; Y_1), ..., MSE_{RF}(\boldsymbol{x}_{N_t}; Y_{N_t}))^T$$

Assuming ellipticity for each U-statistic in the vectors above (with respect to the tree predictions at each point in $\mathcal{T}$), we see that $\sqrt{B}(MSE_{RF}(\mathcal{T}) - \mathbb{E}(MSE_{RF}(\mathcal{T})) \xrightarrow{d} \boldsymbol{Z}_{N_t}$ where $\boldsymbol{Z}_{N_t}$ is a multivariate normal random variable with mean vector 0 and covariance matrix Σ. Then, let $\mathbf{1}/\boldsymbol{N_t}$ be the $N_t \times 1$ vector of the quantity $1/N_t$, so that as $n \to \infty$

$$\sqrt{B}(\mathbf{1}/\boldsymbol{N_t}^T MSE_{RF}(\mathcal{T}) - \mathbb{E}(\mathbf{1}/\boldsymbol{N_t}^T MSE_{RF}(\mathcal{T}))) \xrightarrow{d} Z\sqrt{\mathbf{1}/\boldsymbol{N_t}^T \Sigma \mathbf{1}/\boldsymbol{N_t}} \quad (3.6)$$

where $Z \sim \mathcal{N}(0,1)$. This is the limiting unconditional distribution of the mean squared error of a subsampled random forest, conditional on the test points.

For application to the testing procedure of Section 2, we now identify the distribution of the differences in MSE. Recalling the notation of Section 2, let $MSE_{RF}(\mathcal{T})$ be the MSE of a full random forest at a set of test points $\mathcal{T}$, and let $MSE_{RF^\pi}(\mathcal{T})$ of a forest trained on a reduced feature space. By the results above, we see that, under the hypothesis that $\mathbb{E}(MSE_{RF}(\mathcal{T})) = \mathbb{E}(MSE_{RF^\pi}(\mathcal{T}))$ we have that

$$\sqrt{B}\mathbf{1}/\boldsymbol{N_t}^T(MSE_{RF^\pi}(\mathcal{T}) - MSE_{RF}(\mathcal{T})) \xrightarrow{d} \mathcal{N}(0, \tau^2)$$

for some positive constant τ^2 which is dependent on Σ and the correlation between the two MSEs.

3.2.1 Tree-specific results

Until now, our discussion of the proposed test has remained largely agnostic on the type of base-learners employed, subject to the regularity conditions needed for asymptotic normality. We now argue that the trees typically grown in a random forest indeed satisfy these conditions. The following result follows a similar strategy as Lemma 2 in Meinshausen (2006) with regularity conditions similar to those imposed in Wager and Athey (2017).

Proposition 1. *Assume that* $Y = m(\boldsymbol{X}) + \epsilon$, *where* $m(\cdot)$ *is continuous on the unit cube. Let* $\mathcal{X} = [0,1]^p$, *and assume that* $X_{i,j} \overset{iid}{\sim} Unif(0,1)$ *for* $i = 1, ..., n$ *and* $j = 1, ..., p$. *Then, let* $T_n(\boldsymbol{x})$ *be a tree trained on iid pairs* $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$ *such that each leaf of the tree contains a single observation. Further, assume the trees satisfy the following two conditions:*

(i) $\exists \gamma > 0$ *such that* $P(\text{variable } j \text{ is split on}) > \gamma$ *for* $j \in \{1, ..., p\}$

(ii) Each split leaves at least γn *observations in each node.*

Then, for each $\boldsymbol{x} \in \mathcal{X}$

$$T_n(\boldsymbol{x}) \xrightarrow{d} Y | \boldsymbol{X} = \boldsymbol{x} \text{ as } n \to \infty.$$

The tree predictions thus asymptotically behave like the conditional samples of Y, and thus should, in practice, have finite non zero variance. Breiman (2001) recommends building trees to full depth, so that for most random forests used in practice, Condition 1 is automatically satisfied.

3.3 Extension to Permutation Tests

The unconditional sampling distribution of the difference in test MSE was shown to be asymptotically Gaussian, with an intractable variance. Now, we focus on showing that the permutation distribution generated according to the procedure of Section 2 converges to the unconditional distribution of the MSE differences. Recall the hypotheses of interest are

$$H_0 : \mathbb{E}(MSE_{RF}(\mathcal{T})) = \mathbb{E}(MSE_{RF^\pi}(\mathcal{T}))$$
$$H_1 : \mathbb{E}(MSE_{RF}(\mathcal{T})) < \mathbb{E}(MSE_{RF^\pi}(\mathcal{T})).$$

If a computationally efficient, consistent estimator of $\text{Var}[MSE_{RF^\pi}(\mathcal{T}) - MSE_{RF}(\mathcal{T})]$ were available, we could use the results from the previous section directly in a Z-test. Instead, the complicated structure of our test statistic incentivizes a permutation test, which, for

the correct null hypotheses, circumvents this need. Here, as is common, we appeal to a classical theorem from Hoeffding.

Theorem 2. *(Hoeffding, 1952) For a sequence of data $\{X_i\}_{i=1}^n$ and a statistic $S : \mathbb{R}^n \to \mathbb{R}$, define the permutation distribution function as*

$$\hat{J}_n(t) = \frac{1}{|\mathcal{G}_n|} \sum_{\pi \in \mathcal{G}_n} I\big\{S(X_{\pi(1)}, ..., X_{\pi(n)}) \leq t\big\}$$

where $\mathcal{G}_n$ is the group of all permutations of $\{1, ..., n\}$. Let π, π' be two permutations drawn independently and uniformly over $\mathcal{G}_n$, and suppose that

$$\big(S(X_{\pi(1)}, ..., X_{\pi(n)}), S(X_{\pi'(1)}, ..., X_{\pi'(n)})\big) \xrightarrow{d} (S, S') \tag{3.7}$$

where S, S' are iid with cdf $R(\cdot)$. Then for all t at which $R(\cdot)$ is continuous

$$\hat{J}_n(t) \xrightarrow{p} R(t). \tag{3.8}$$

We now verify Equation 3.7 for a variety of scenarios. Importantly, we note that the converse of Theorem 2 was shown to hold in Chung and Romano (2013), so Equation 3.7 and Equation 3.8 can be used interchangeably.

For a single test point, Equation 3.3 gives that the MSE differences can be expanded as

$$\begin{aligned} MSE_{RF^\pi}(\boldsymbol{x}; Y) - MSE_{RF}(\boldsymbol{x}; Y) &= \frac{1}{B^2} \sum_{j=1}^{B} \sum_{i=1}^{B} T_j^\pi(\boldsymbol{x}) T_i^\pi(\boldsymbol{x}) - \frac{2Y}{B} \sum_{j=1}^{B} T_j^\pi(\boldsymbol{x}) + Y^2 \\ &\quad - \left(\frac{1}{B^2} \sum_{j=1}^{B} \sum_{i=1}^{B} T_j(\boldsymbol{x}) T_i(\boldsymbol{x}) - \frac{2Y}{B} \sum_{j=1}^{B} T_j(\boldsymbol{x}) + Y^2 \right) \\ &= \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \big[T_j^\pi(\boldsymbol{x}) T_i^\pi(\boldsymbol{x}) - T_j(\boldsymbol{x}) T_i(\boldsymbol{x}) \big] + \frac{2Y}{B} \sum_{j=1}^{B} \big[T_j(\boldsymbol{x}) - T_j^\pi(\boldsymbol{x}) \big]. \end{aligned} \tag{3.9}$$

The extension to a multipoint test set (of N_t test points) is given by

$$\begin{aligned} \mathbf{1}/\boldsymbol{N_t}^T (MSE_{RF^\pi}(\mathcal{T}) - MSE_{RF}(\mathcal{T})) = \frac{1}{N_t} \sum_{(\boldsymbol{x}, Y) \in \mathcal{T}} \Bigg(\frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \big[T_j^\pi(\boldsymbol{x}) T_i^\pi(\boldsymbol{x}) - T_j(\boldsymbol{x}) T_i(\boldsymbol{x}) \big] \\ + \frac{2Y}{B} \sum_{j=1}^{B} \big[T_j(\boldsymbol{x}) - T_j^\pi(\boldsymbol{x}) \big] \Bigg). \end{aligned}$$

To analyze the permutation distribution of this quantity, we examine the permutation distribution of each component of Equation 3.9. Let $D_B = \frac{2Y}{B} \sum_{j=1}^{B} \big[T_j(\boldsymbol{x}) - T_j^\pi(\boldsymbol{x}) \big]$. Then,

because under H_0, $\mathbb{E}(T(\boldsymbol{x})) = \mathbb{E}(T^\pi(\boldsymbol{x}))$ and $\lim_{B\to\infty} \frac{B}{2B} \equiv 1/2$ we can use Example 2.1 of Chung and Romano (2013) to say that the permutation distribution of $\sqrt{B}D_B$ converges to a normal distribution with the same asymptotic variance as the unconditional distribution of $\sqrt{B}D_B$.

We analyze the second term in two steps. First note that we have only discussed the unconditional distribution of the V-statistic in the MSE differences through the lens of its corresponding U-statistic. The results of Chung and Romano (2016) only apply to the permutation distribution of the U-statistic so first we demonstrate that results about the permutation distribution of the U-statistic apply to the V-statistic:

Lemma 4. *Let $X_1, ..., X_B$ be an exchangeable sequence and define $T_1 = T_1(X_1, ..., X_B)$ to be a statistic such that, for iid permutations π, π':*

$$(T_1(X_{\pi(1)}, ..., X_{\pi(B)}), T_1(X_{\pi'(1)}, ..., X_{\pi'(B)})) \xrightarrow{d} (T, T') \quad \text{as } B \to \infty$$

where $T, T' \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$, so that by Theorem 15.2.3 of Lehmann and Romano (2006):

$$\sup_t |\hat{J}_B^{T_1}(t) - \Phi(t/\tau)| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where $\hat{J}_B^{T_1}$ is the permutation distribution defined in Section 1. Then, let $T_2 = T_2(X_1, ..., X_B)$ be another statistic. If $T_2 \xrightarrow{p} T_1$, then

$$\sup_t |\hat{J}_B^{T_2}(t) - \Phi(t/\tau)| \xrightarrow{p} 0 \quad \text{as } B \to \infty.$$

Lemma 4 gives that the V-statistic that appears in the MSE expansion (3.3) has the same limiting permutation distribution as the corresponding U-statistic. Under mild assumptions, we show now that the permutation distribution of the U-statistic is the same as the unconditional distribution in the large sample limit.

Theorem 3. *Let $P(\boldsymbol{x}) = P$ denote the distribution of tree predictions at $\boldsymbol{x}$ for trees trained on on $\mathcal{D}_n$, and let $Q(\boldsymbol{x}) = Q$ denote the distribution for tree predictions at $\boldsymbol{x}$ for trees trained on $\mathcal{D}_n^\pi$. Assume that the subsample size and number of trees is such that the trees become iid in the large sample limit (i.e. the trees satisfy Lemma 1). If $\mathbb{E}_P(T(\boldsymbol{x})) = \mathbb{E}_Q(T(\boldsymbol{x})) = \mu$ and $\mathbb{E}_P(T^2(\boldsymbol{x})) = \mathbb{E}_Q(T^2(\boldsymbol{x})) = \sigma^2$, then the randomization distribution of $U = \sqrt{B}\frac{1}{\binom{B}{2}} \sum_{1\le i<j\le B}(T_i^\pi(\boldsymbol{x})T_j^\pi(\boldsymbol{x}) - T_i(\boldsymbol{x})T_j(\boldsymbol{x}))$ satisfies*

$$\sup_t |\hat{J}_B^{U}(t) - \Phi(t/\tau)| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where τ is such that $U \xrightarrow{d} \tau Z$ for Z a standard normal random variable.

Denoting $\mu_P = \mathbb{E}_P(T(\boldsymbol{x}))$ and $\sigma_P^2 = \mathbb{E}_P(T^2(\boldsymbol{x}))$, we see that the permutation test converges to the unconditional distribution under the null hypothesis

$$H_0 : \{\mu_P = \mu_Q\} \cap \{\sigma_P^2 = \sigma_Q^2\}. \tag{3.10}$$

Note that this is not quite equivalent to H_0 in Equation 2.2 at a single point. As such, we now connect these hypotheses about the tree predictions to those about the MSE of a random forest. The conditional MSE can be decomposed as

$$\begin{aligned}\mathbb{E}_P(MSE_{RF}(\boldsymbol{x}; Y)) &= \text{Var}_P(RF(\boldsymbol{x})) + (\mathbb{E}_P(RF(\boldsymbol{x})) - Y)^2 \\ &= \frac{1}{B}(\sigma_P^2 + \mu_P^2) + (\mu_P - Y)^2\end{aligned}$$

where the equality follows because (in the $n \to \infty$ case), $\text{Var}(RF(\boldsymbol{x})) = \frac{1}{B}\text{Var}(T(\boldsymbol{x}))$, and $\mathbb{E}(RF(\boldsymbol{x})) = \mathbb{E}(T(\boldsymbol{x}))$. So, if Equation 3.10 holds, $\mathbb{E}_P(MSE_{RF}(\boldsymbol{x}; Y)) = \mathbb{E}_Q(MSE_{RF}(\boldsymbol{x}; Y))$. The converse is true if we additionally assume

$$\text{Var}_P(RF(\boldsymbol{x})) - \text{Var}_Q(RF(\boldsymbol{x})) \neq -(\mathbb{E}_P(RF(\boldsymbol{x})) - Y)^2 + (\mathbb{E}_Q(RF(\boldsymbol{x})) - Y)^2.$$

The above has a straightforward interpretation - the test is valid for H_0 in Equation 2.2 if and only if the change in variance from the full feature space is not equal to the negative change in (squared) bias from using the full feature space. Recalling Proposition 1, the asymptotic variance of a fully grown tree prediction at $\boldsymbol{x}$ is equal to that of an observation drawn at random at from $Y|\boldsymbol{X} = \boldsymbol{x}$, so that if a feature is unimportant, meaning $Y|\boldsymbol{X} = \boldsymbol{x} \overset{d}{=} Y|\boldsymbol{X}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}$, for $\mathcal{S} \subset \{1, ..., p\}$, the asymptotic variances should be equivalent.

Note that the test MSE is a linear combination of N_t V-statistics of order 2 and N_t sample means, each of which has an asymptotically normal permutation distribution. We now present a result guaranteeing the convergence of this linear combination to a normal distribution.

Lemma 5. *Define*

$$\tilde{T}_{2B} = [T_1(X_{\pi(2B)}), T_1(X_{\pi'(2B)}), \cdots, T_M(X_{\pi(2B)}), T_M(X_{\pi'(2B)})]^T$$

where T_j is a U-statistic for each j, and further for a vector $t \in \mathbb{R}^{2M}$, define $T_{t,2B} = t^T\tilde{T}_{2B}$. In particular, consider $c = [c_i]_{i=1}^{2M}$ and $c' = [c'_i]_{i=1}^{2M}$, where $c_i = I\{i \text{ is odd}\}$ and $c'_i = I\{i \text{ is even}\}$. If each pair $\left(T_j(X_{\pi(2B)}), T_j(X_{\pi'(2B)})\right)$ satisfies

$$\left(T_j(X_{\pi(2B)}), T_j(X_{\pi'(2B)})\right) \overset{d}{\to} T_j, T'_j \overset{iid}{\sim} \mathcal{N}(0, \tau_j^2)$$

such that $\tilde{T}_{2B}$ are elliptic for $X_{\pi(2B)}$ for any permutation π, then we see that

$$(T_{c,2B}, T_{c',2B}) \overset{d}{\to} (T, T')$$

where $T, T' \overset{iid}{\sim} \mathcal{N}(0, \tau_M^2)$*, so that*

$$\sup_t |\hat{J}_B^T(t) - \Phi(t/\tau_T)| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where $\tau_M^2 = \sum_{j=1}^M \tau_j^2 + 2\sum_{1 \le j < k \le M} Cov(T_j, T_k)$.

The upshot of these technical results is that, under some mild assumptions, the permutation distribution of the MSE difference between forests (under H_0) is identical to its unconditional counterpart. As such, we may evaluate the hypotheses without the need to estimate the unwieldly variance of Equation 3.6. We state this result in its entirety below.

Theorem 4. *Let* $P(\{T(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{T}}) = P$ *denote the joint distribution of tree predictions at the points in* $\mathcal{T}$ *for trees trained on* $\mathcal{D}_n$*, and let* $Q(\{T(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{T}}) = Q$ *denote the joint distribution for tree predictions at points in* $\mathcal{T}$ *for trees trained on* $\mathcal{D}_n^\pi$*. Assume that the subsample size and number of trees is such that the trees become iid in the large sample limit (i.e. the trees satisfy the conditions of Lemma 1). Further, assume that for all* $(\boldsymbol{x}, y) \in \mathcal{T}$*,* $B \in \mathbb{N}$

$$Var_P(RF_B(\boldsymbol{x})) - Var_Q(RF_B(\boldsymbol{x})) \neq -(\mathbb{E}_P(RF_B(\boldsymbol{x})) - y)^2 + (\mathbb{E}_Q(RF_B(\boldsymbol{x})) - y)^2.$$

Define the test statistic

$$\Delta_B := \sqrt{B}\mathbf{1}/\boldsymbol{N}_{\boldsymbol{t}}^T(MSE_{RF^\pi}(\mathcal{T}) - MSE_{RF}(\mathcal{T})).$$

Assume that the collection of terms in the summation of Δ_B *are elliptic for the vector of tree predictions* $\left\{[T_{1,i}, ..., T_{N_t,i}]^T\right\}_{i=1}^B$*. Then, under* $H_0 : \mathbb{E}(\Delta_B) = 0$ *the permutation distribution satisfies*

$$\sup_t |\hat{J}_B^\Delta(t) - \Phi(t/\tau_\Delta)| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where τ_Δ *is such that* $\Delta_B \xrightarrow{d} Z_\Delta \sim \mathcal{N}(0, \tau_\Delta^2)$*, and* $\Phi(\cdot)$ *is the standard normal cdf.*

It should be noted that a similar procedure could be used in more standard situations such as the two-sample mean estimation problem. However, in these simpler situations, the variance estimation is straightforward, so a permutation test approach generally requires more computational effort than the unconditional test.

4 Simulations

We now demonstrate that the theoretical guarantees, though asymptotic in nature, are supported by empirical evidence. We simulate data from four models, summarized in Table 1, with the covariate structure summarized in Table 2. For each of our simulations,

we train random forests using the `randomForest` package in R (Liaw and Wiener, 2002) using the default `mtry` parameters.

Model #	Data Generating Model	Covariate Structure
1	$Y = \beta X_1 + \beta I(X_6 = 2) + \epsilon$	M1
2	$Y = \beta \sin(\pi I(X_7 = 2) X_1) + 2\beta(X_3 - .05)^2 + \beta X_4 + \beta X_2 + \epsilon$	M1
3	$Y = RF_{\texttt{eBird}}(\boldsymbol{X}) + \epsilon$	`eBird`
4	$P(Y = 1 \vert \boldsymbol{X}) = \text{expit}\left[\beta \sum_{j=2}^{5} X_j\right]$	M2

Table 1: Distributions of $Y|\boldsymbol{X}$ for each model used in the simulation. In Models 1-3, ϵ is mean 0, Gaussian noise with variance σ^2 (varies across simulations), and we let $\text{expit}(z) = \frac{1}{1+e^z}$.

Model #	Covariate Structure
M1	$X_1, ..., X_5 \overset{iid}{\sim} Unif(0,1)$, $X_6, ..., X_{10} \overset{iid}{\sim}$ Multionimial$(1, [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T)$
M2	$X_1, ..., X_{500} \sim \text{AR}_1(0.15)$
`eBird`	Data from Coleman et al. (2017) - 12 variables + 2 proxy variables

Table 2: Distribution of $\boldsymbol{X}$ for various simulation studies.

Model 1 is a standard ANCOVA model, which is intended to include both an important discrete and continuous predictor, to demonstrate the robustness of the proposed procedure to covariate type. In Model 1, we test the importance of (X_1, X_6, X_2, X_7), of which the first two are important, and X_1, X_2 are continuous while X_6, X_7 are categorical. Model 2 is a modification of the MARS data generating model (Friedman, 1991) to include an important discrete covariate. The MARS data generating model has been used as a common "complex" model in simulation studies of random forests, such as in Mentch and Hooker (2016). Model 4 is an adaptation of the model used in Candes et al. (2016) for high-dimensional correlated data.

In Model 4, we test for the significance of X_2, which is important, and also X_1 and X_{500}, which are unimportant, but X_1 is highly correlated with X_2 and X_{500} is much more weakly correlated. Candes et al. (2016) demonstrated that the standard logistic regression p-values in this situation are far from uniform under H_0, so that standard parametric inference may not be valid. Random forests, on the other hand, have been shown (Biau, 2012; Scornet et al., 2015) to be insensitive to the dimension of the ambient feature space, and instead are sensitive "strong" feature space. As such, we want to explore the utility of our method in the high dimensional, sparse signal case.

Model 3 is a situation where the true data generating model is a random forest. To construct this, we use a subset of data from the `ebird` project, which stores user inputted information about bird sightings (Sullivan et al., 2009, 2014). We focus on a particular dataset used in Coleman et al. (2017) about the occurrence of tree swallows. To construct $RF_{\texttt{eBird}}$, we draw 5000 points from the data, and train $RF_{\texttt{eBird}}$, a random forest with $\texttt{mtry} = 9$ and 1000 total trees.

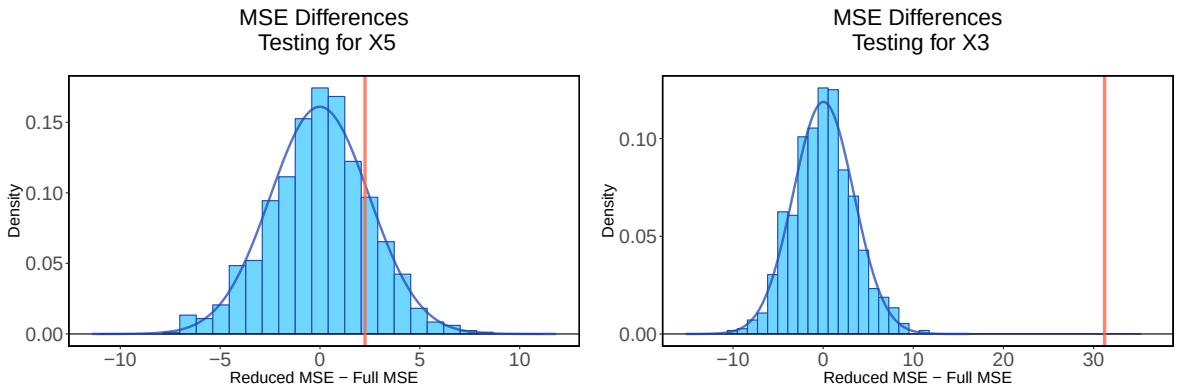
To simulate from this model, we draw (without replacement) samples of size n from the remaining 20727 points, predict at them using $RF_{\texttt{eBird}}$, and then add in a noise term. We test for the effect of two variables: `eff.hours`, which corresponds to the number of hours a user expended upon a hike, and `dfs`, which is a fractional measurement of day of year. We further include two proxy variables, each not used to train $RF_{\texttt{eBird}}$. These variables (`eff.hours.proxy`, `dfs.proxy`) are defined as

$$\texttt{eff.hours.proxy} = \frac{\texttt{eff.hours} + Z_{0.5}}{\sqrt{\text{Var}(\texttt{eff.hours})} + 0.5}$$
$$\texttt{dfs.proxy} = \frac{\texttt{dfs} + Z_{0.025}}{\sqrt{\text{Var}(\texttt{dfs})} + 0.025}$$

where Z_σ is a standard normal random variable with variance σ^2. The purpose of this construction is that the proxy variables' relationship with Y is solely dictated by their dependence on their original copy. The division step is to standardize the variance of the proxies to 1.

4.1 Normality of Permutation Distributions

We first provide a concise simulation demonstrating that the permutation approximation of the Gaussian proposed in Theorem 3 is valid in practice. We simulate $n = 2000$ training observations from Model 2 (with $\beta = 10, \sigma = 10$), along with $N_t = 100$ test observations and apply our procedure to test for the significance of X_3 (important) and X_5 (unimportant). The random forests each consist of $B = 200$ trees trained on subsamples of size $k_n = n^{0.6}$, with $\texttt{mtry} = 3$. The resulting permutation distributions are shown in Figure 1.



(a) Under H_0 (b) Under H_1

Figure 1: Permutation distributions of Δ_B. Red line indicates observed value, and histograms are overlayed by an estimated normal density.

These plots demonstrate that the permutation distributions do approximate a Gaussian distribution. Moreover, in the null case, the observed Δ_B lies squarely in the center of the distribution, while in the alternative case, Δ_B lies far away from the center. Next, we more formally investigate the power/validity of the testing procedure.

4.2 Signal to Noise Ratio Power Curves

For simulating from Models 1-3, we draw $n = 2000$ points from the joint distribution of $(\boldsymbol{X}, Y)$, subsample sizes of $k_n = n^{0.6} \approx 95$ and built $B = 125$ trees in each forest. Predictions were made at $N_t = 100$ test points, each drawn from the same joint distribution as the training data. Note that the null hypothesis, as defined in Equation 2.2, is conditional on the test points used. These simulations change the null hypothesis each time, because the conditioning points change. Thus, the simulations mimic the common practice of random splitting the data into a training and validation fold. In this section, we use the notation `seq(a, b, length.out = c)` as a sequence a_k such that $a_1 = a, a_c = b$ and the other a_k are evenly spaced.

For Models 1 and 2, we focus on a marginal signal to noise ratio, which is controlled by the parameters β and σ. In particular, we fix $\beta = 10$ across all simulations, and evaluate the simulation for σ at the following values

$$\sigma \in \left\{ \frac{10}{k} : k \in \ \{0.005, \texttt{ seq(0.01, 2.25, length.out = 9)}\} \right\}$$

such that for small k, the signal to noise ratio is small. The results for the tests for Model 1 are shown in Figure 2, and the results for Model 2 are shown in Figure 3.

The test maintains the nominal Type I error rate, and attains high power for marginal signal to noise ratios around 1, for all variables except X_7 in Model 2. An immediate takeaway is that the the Type I error rate is insensitive to the covariate structure. Further, in the MARS model, the test has more power against X_3 than X_7, because X_7 is only important insofar as it interacts with X_1.

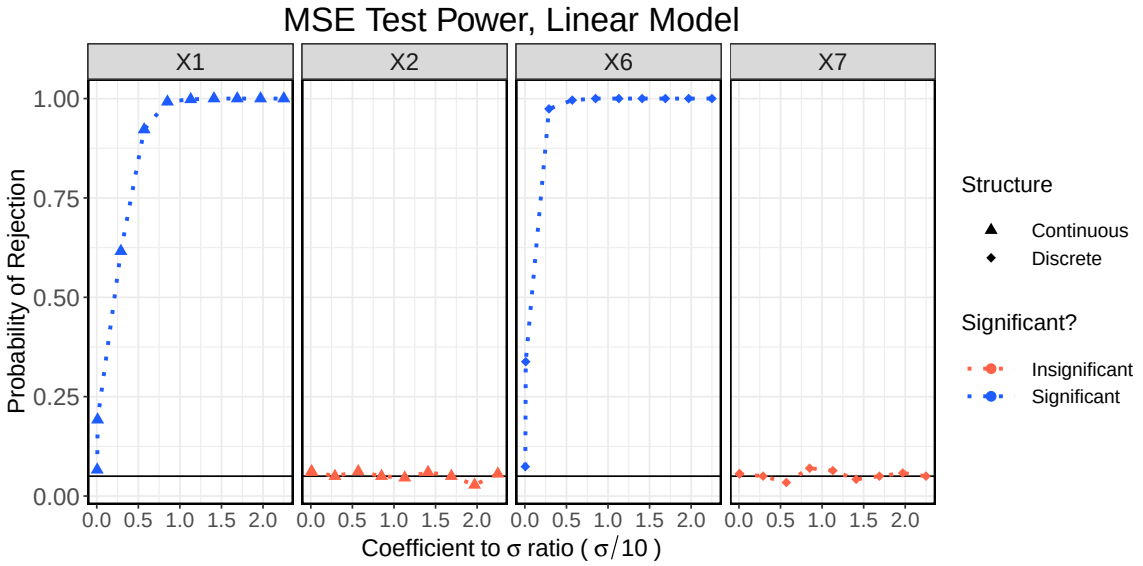


Figure 2: Model 1 power curves for 500 simulations. The Y-axis represents $P(p \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

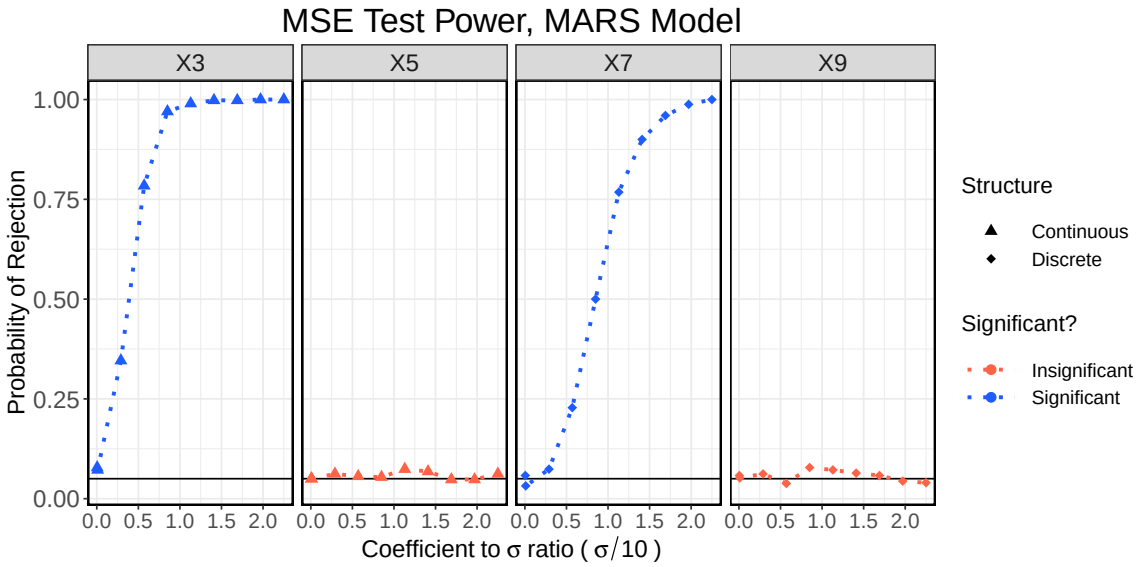


Figure 3: Model 2 power curves for 500 simulations. The Y-axis represents $P(p \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

Next, we consider the situation from Model 3. In this case, there are no linear parameters to vary. Instead, we vary the variance of the noise term, ϵ. We again let $n = 2000$, $k_n = n^{0.6}$, $B = 125$, and $N_t = 100$. We vary σ according to

$$\sigma \in \left\{ e^{-k} : \; k \in \{\texttt{seq(1, 5, length.out = 10)}\} \right\}.$$

The results of this simulation are show in Figure 4. We see that the test still maintains the nominal Type I error rate and modest power for the true variables. Moreover, the testing procedure correctly identifies the true variables as important over the proxy variables.

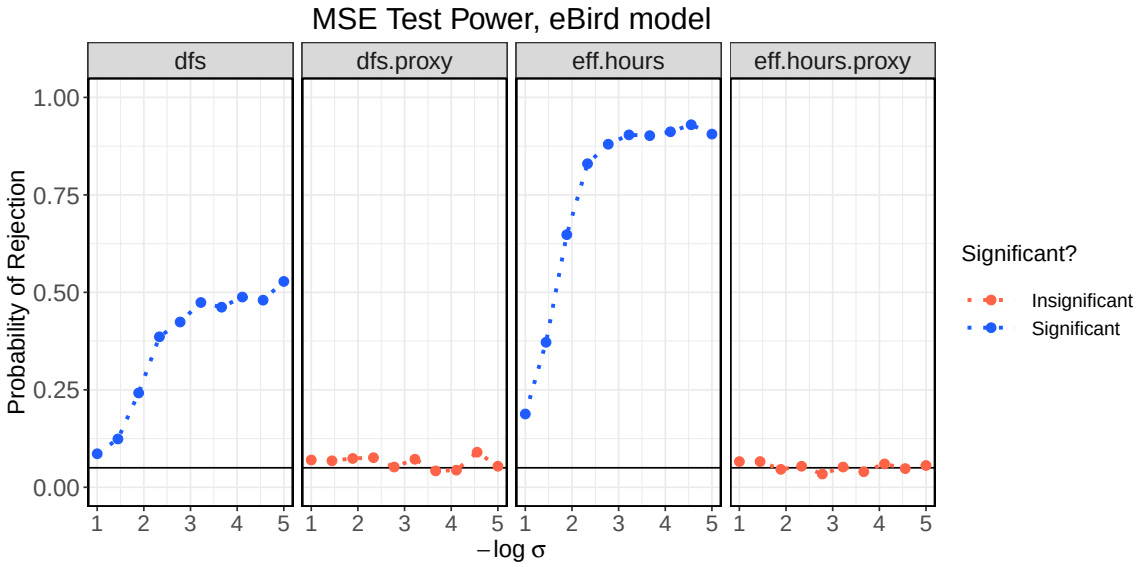


Figure 4: Model 3 power curves for 500 simulations. The Y-axis represents $P(p \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

Finally, we present results for the simulation from Model 4. In this case, we limit $n = 600$, so that p/n is not small, though the dimension of the strong features is still small relative to n. We let $k_n = n^{0.6} \approx 46$, $B = 125$, and $N_t = 100$. To develop power curves, we vary the β coefficient according to

$$\beta \in \left\{ \texttt{seq(0.01, 2.5, length.out = 8)} \cup \texttt{seq(5, 20, length.out = 7)} \right\}.$$

The results are shown in Figure 5. We see that the proposed test resolves the biased p-value issue associated with the standard glm procedure and is still able to attain reasonable power for the effect of X_2. The power is likely limited by the fact that for large β, the change in the marginal effect of each covariate likely only changes $P(Y = 1|\boldsymbol{X})$ slightly due to the rapidly decaying first derivative of the expit(z) function.

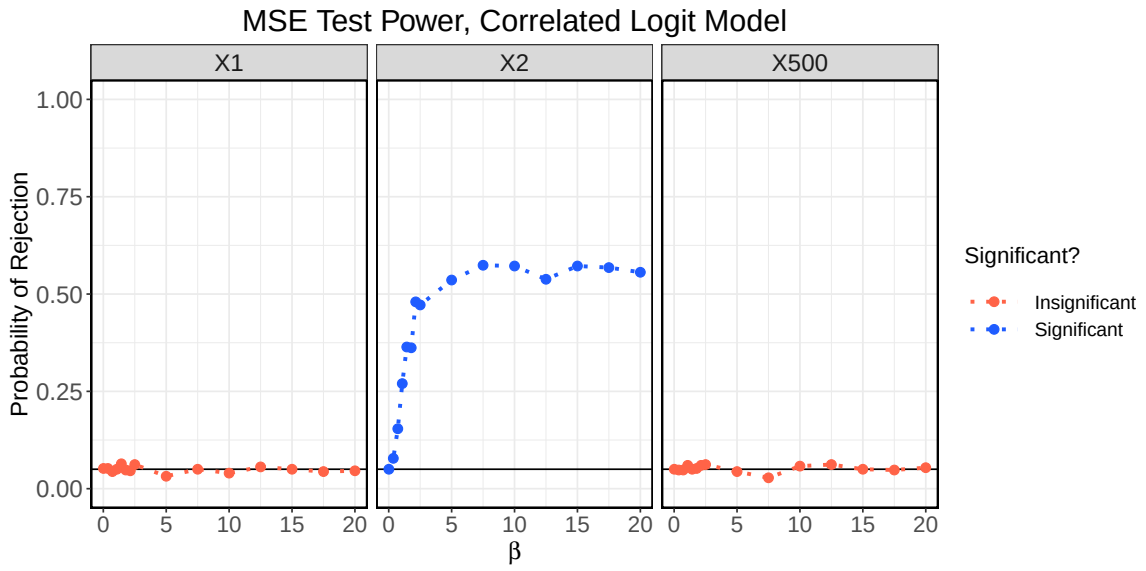


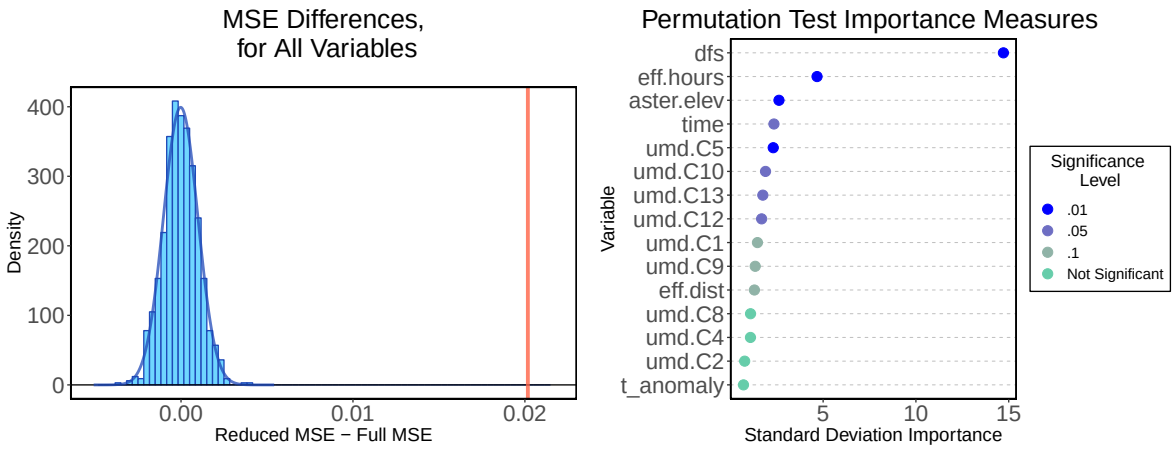
Figure 5: Model 4 power curves for 500 simulations. The Y-axis represents $P(\tilde{p} \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

5 Applications to Ecological Data

We now apply our proposed testing procedure on two ecological datasets where random forests have been shown to perform well in recent work.

eBird

We first consider the eBird data described in the previous section to constructed a simulated random forest model. Here we utilize the original data as considered in Coleman et al. (2017). The standard task to predict tree swallow `occurrence` during the fall migration season in a particular geographic area referred to as Bird Conservation Region (BCR) 30. This is a Citizen Science project where observers submit reports detailing when and where they recorded observations. The response in each row of the data is either 0 or 1 corresponding to whether a tree swallow was observed during that particular outing. Features include information about latitude, longitude, time of year, as well as observer, environmental, temperature, and land cover characteristics. The data consists of $n = 25727$ observations on 23 features, gathered between 2008 and 2013. Coleman et al. (2017) argue that the data contain several nonlinearities and complex higher-order interactions and demonstrate that random forests provide a more accurate fit than numerous alternatives.



(a) MSE Test for all covariates

(b) Marginal Importance Scores

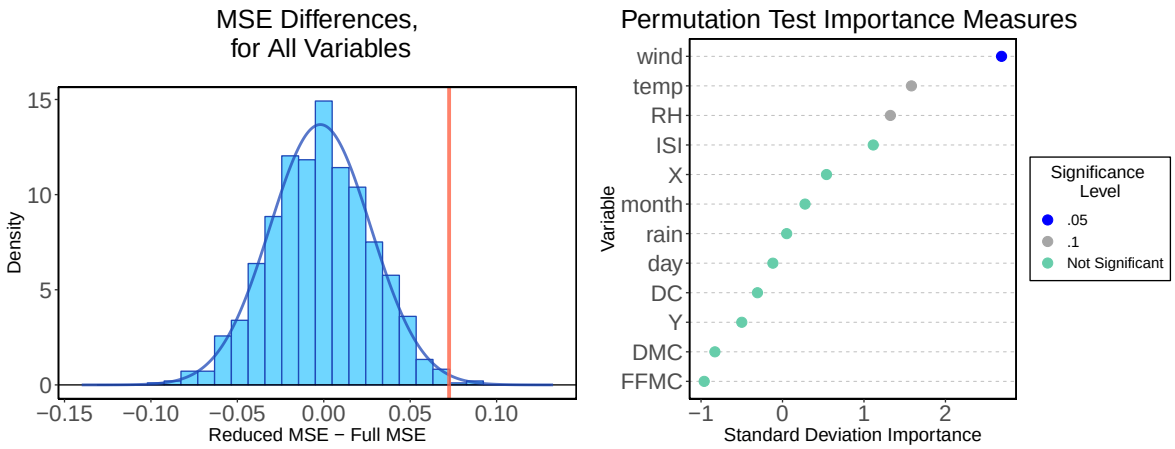
Figure 6: Results on the eBird data from (Sullivan et al., 2009, 2014). Red line indicates observed value, and histograms of differences in MSE after permutation are overlayed by an estimated normal density.

The authors carry out a testing procedure based on the parametric approach in Mentch and Hooker (2016) but due to the limitations described in previous sections, are limited to a test sample of only 25 points.

We first apply Algorithm 1 to test the importance of any variables in predicting `occurrence`, analogous to an overall F-test in multiple linear regression. Here we select 15% of the available observations ($\approx$ 3800 points) uniformly at random to serve as the test set where the hypotheses will be evaluated. The random forests were trained with the `ranger` package using the default `mtry` $= 4$, subsamples of size $k_n = n^{0.6}$, and consisting of $B = 250$ trees in each. The results are shown in Figure 6. There is clear evidence for signal in the data, with an overall p-value of $p < 0.0001$. We see that `dfs`, which corresponds to the day of the year, `eff.hours`, which corresponds to a users' effort (in time), and `aster.elev`, which corresponds to elevation, are the most important features. Time of year (`dfs`) and elevation (`aster.elev`) have an intuitive relationship with occurrence, and likely serve as proxies for climate conditions. Larger `eff.hours` suggest that a user spent more time out in the field on a particular day, meaning they were more likely to observe a tree swallow because of increased birding time.

Forest Fires

We turn now to forest fire data from Cortez and Morais (2007) where the task is to predict log(1 + `area`) burned by several fires in northern Portugal using covariate information



(a) MSE Test for all covariates

(b) Marginal Importance Scores

Figure 7: Results on the forest fire data from Cortez and Morais (2007). Red line indicates observed value, and histograms of differences in MSE after permutation are overlayed by an estimated normal density.

on location, time of year, and local weather characteristics. The data contains $n = 537$ observations on 13 features. Cortez and Morais (2007) found that a naive mean predictor attained the lowest RMSE - suggesting that there is weak signal in the data. We carry out our testing procedure in exactly the same fashion as the eBird data, using `mtry` $= 12$ and $k_n = n^{0.6} \approx 43$, $B = 250$ trees for the importance test and $B = 500$ trees for the overall test; results are shown in Figure 7. The overall test suggests that there is signal in the data ($p = 0.0040$), albeit a weaker effect than in the eBird case study. The importance procedure suggests that `wind` and `temp`, which respectively correspond to the wind speeds and temperatures at the location of the fire.

6 Discussion

The work here presents a formal hypothesis testing framework for evaluating the predictive significance of covariates in a random forests model which, unlike existing approaches, is both computationally efficient and statistically valid, placing hypothesis tests with random forests firmly within the grasp of applied researchers. Existing formal parametric approaches are computationally prohibitive and place severe restrictions on the number of points where the hypotheses can be evaluated, while the popular heuristic out-of-bag (oob) approaches are easily computed but also easily fooled by correlated and/or categorical covariates. We note further that while the ensemble nature of random forests presents a

natural context for such tests, much of the theoretical backing for this procedure is largely agnostic to the particular class of base-learner models being constructed.

Besides its feasibility, this permutation approach also offers some flexibility in the kinds of problems open to investigation by applied scientists and practitioners. Consider, for example, the mediator detection problem in psychology where a covariate X_1 is called a *mediator* for another covariate X_2 whenever the effect of X_2 on the response is nullified (or at least lessened) by including X_1 in the model. The same two-step process often employed with linear models can now be carried out in a random forest context: first determine whether X_2 is significant without X_1 in the model, then test whether the significance of X_2 disappears whenever X_1 is included.

One potential criticism of the approach presented here may be that it becomes more computationally burdensome whenever one wishes to evaluate the significance of all available covariates one at a time, transforming the test statistic into something akin to the oob importance measures. Note however that by construction, we need only build relatively few trees to conduct each test and thus in small or even moderate dimensions, simply repeating our permutation approach p times is still more computationally efficient than carrying out even a single parametric test. Furthermore, in practice, it may be reasonable to consider more heuristic methods for generating hypotheses about feature importance and formally evaluating them via the procedure laid out here.

Finally, we note that the validity of our approach was verified by arguing that the random forest trees behave like an iid sequence asymptotically. It may be possible to establish the necessary conditions under alternative scenarios given that central limit theorems can sometimes be established in more complex settings. The resulting variance may not line up with the variance estimate suggested by the permutation distribution but the limiting distributions should be similar so long as the number of trees built B is small relative to the correlation between them. In particular, if we assume that the tree predictions at $\boldsymbol{x}$ have variance σ^2, and pairwise correlation ρ, then Friedman et al. (2001) provides that the variance of the random forest prediction is given by

$$\text{Var}\left[\frac{1}{B}\sum_{i=1}^{B} T_i(\boldsymbol{x})\right] = \sigma^2\frac{B}{B^2} + 2\frac{\frac{B^2-B}{2}}{B^2}\rho\sigma^2 = \sigma^2\left(\frac{1+(B-1)\rho}{B}\right).$$

For $\rho = 0$, the random forest variance reduces to $\frac{\sigma^2}{B}$ – the same variance used in the permutation test theory above by treating the trees as iid. Thus, the difference between the true variance and the theoretical variance is $\sigma^2\big(\frac{(B-1)\rho}{B}\big)$. For large B, this term approaches $\rho\sigma^2$, which is the maximum difference between the variances. However, for small B, i.e. $1/\rho \gg B$, the difference term is very near 0, so the iid approximation is reasonable. Recall that both ρ and B depend on n, so that a weaker and more reasonable requirement for an iid-like CLT to hold on the number of trees may be $\rho_n = o(1/B_n)$.

SUPPLEMENTARY MATERIAL

R code for implementing the procedure for a variety of base learners is provided, as well as all simulation examples. A package is under development, and will be submitted to the CRAN.

ACKNOWLEDGEMENTS

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. We specifically acknowledge the assistance of Kim Wong. LM was supported in part by NSF DMS-1712041. We thank the thousands of eBird participants for their contributions and support for eBird from The Wolf Creek Foundation and the National Science Foundation (ABI sustaining: DBI-1356308; computing support from CNS-1059284).

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII1983*, pp. 1–198. Springer.

Altmann, A., L. Toloşi, O. Sander, and T. Lengauer (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics 26*(10), 1340–1347.

Athey, S., J. Tibshirani, and S. Wager (2016). Generalized random forests. *arXiv preprint arXiv:1610.01271*.

Barber, R. F., E. J. Candès, et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055–2085.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research 13*(Apr), 1063–1095.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Candes, E., Y. Fan, L. Janson, and J. Lv (2016). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*.

Chow, Y. S. and H. Teicher (2012). *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media.

Chung, E. and J. P. Romano (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 484–507.

Chung, E. and J. P. Romano (2016). Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference 168*, 97–105.

Coleman, T., L. Mentch, D. Fink, F. La Sorte, G. Hooker, W. Hochachka, and D. Winkler (2017). Statistical inference on tree swallow migrations. *arXiv preprint arXiv:1710.09793*.

Cortez, P. and A. d. J. R. Morais (2007). A data mining approach to predict forest fires using meteorological data.

Cui, Y., R. Zhu, M. Zhou, and M. Kosorok (2017). Some asymptotic results of survival tree and forest models. *arXiv preprint arXiv:1707.09631*.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, Volume 7, pp. 1–68.

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

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim (2014). Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research 15*(1), 3133–3181.

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1–67.

Good, P. I. (2002). Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods 1*(2), 34.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, 293–325.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 169–192.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics 15*(3), 651–674.

Ishwaran, H. and M. Lu (2008). Random survival forests. *Wiley StatsRef: Statistics Reference Online*, 1–13.

Ishwaran, H. and M. Lu (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine 38*(4), 558–582.

Janitza, S., E. Celik, and A.-L. Boulesteix (2016). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 1–31.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & probability letters 36*(1), 9–21.

Janssen, A. (2005). Resampling student'st-type statistics. *Annals of the Institute of Statistical Mathematics 57*(3), 507–529.

Klass, M. and H. Teicher (1987). The central limit theorem for exchangeable random variables without moments. *The Annals of Probability*, 138–153.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.

Lehmann, E. L. and J. P. Romano (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lehmann, E. L., C. Stein, et al. (1949). On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics 20*(1), 28–45.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*(3), 18–22.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research 7*(Jun), 983–999.

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

Mentch, L. and G. Hooker (2017). Formal hypothesis tests for additive structure in random forests. *Journal of Computational and Graphical Statistics*, 1–9.

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

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, 1760–1779.

Nicodemus, K. K., J. H. Callicott, R. G. Higier, A. Luna, D. C. Nixon, B. K. Lipska, R. Vakkalanka, I. Giegling, D. Rujescu, D. S. Clair, et al. (2010). Evidence of statistical epistasis between disc1, cit and ndel1 impacting risk for schizophrenia: biological validation with functional neuroimaging. *Human genetics 127*(4), 441–452.

Pesarin, F. and L. Salmaso (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association 85*(411), 686–692.

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis 146*, 72–83.

Scornet, E., G. Biau, J.-P. Vert, et al. (2015). Consistency of random forests. *The Annals of Statistics 43*(4), 1716–1741.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics 8*(1), 25.

Sullivan, B. L., J. L. Aycrigg, J. H. Barry, R. E. Bonney, N. Bruns, C. B. Cooper, T. Damoulas, A. A. Dhondt, T. Dietterich, A. Farnsworth, et al. (2014). The ebird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation 169*, 31–40.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling (2009). ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation 142*(10), 2282–2292.

Toloşi, L. and T. Lengauer (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics 27*(14), 1986–1994.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* (just-accepted).

Wager, S., T. Hastie, and B. Efron (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research 15*(1), 1625–1651.

Wright, M. N. and A. Ziegler (2015). Ranger: a fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

Zhu, R., D. Zeng, and M. R. Kosorok (2015). Reinforcement learning trees. *Journal of the American Statistical Association 110*(512), 1770–1784.

A Proofs and Calculations

We now provide the technical details and proofs for theoretical discussion in Section 3. For completeness, theorems and lemmas are restated.

Theorem 1. *Under the exchangeability conditions outlined in Section 3.1, denote a sequence of (potentially randomized) trees trained on subsamples from $\mathcal{D}_n$ as $\{T_k(\cdot)\}_1^\infty$. Moreover, consider an independently drawn test point, $\boldsymbol{Z}^* = (\boldsymbol{X}^*, Y^*) \sim F$. Then, the residuals, defined as:*

$$r_k = T_k(\boldsymbol{X}) - Y^*$$

form an infinitely exchangeable sequence of random variables.

Proof. Let Ξ be the distribution of randomization parameters, and let $\mathcal{S}_{k_n}(\mathcal{D}_n)$ be the distribution of subsamples of size k_n drawn uniformly from the original data. Then, to construct a tree, we have the following procedure:

1. Draw $\mathcal{D}^*_{k_n} \sim \mathcal{S}_{k_n}(\mathcal{D}_n)$
2. Draw $\xi \sim \Xi$
3. Draw $\boldsymbol{Z}^* \sim F$
4. Construct a tree according to some combining function, say ϕ , of $\xi, \mathcal{D}^*_{k_n}$, i.e. $T = \phi(\xi, \mathcal{D}^*_{k_n})$.

Each draw is done independent of the other draws. Repeating (1) and (2) independently gives iid sequences $\{\mathcal{D}^*_{l,k_n})\}_{l=1}^{\infty}$ and $\{\xi_j\}_{j=1}^{\infty}$. Then, the sequence

$$T_1 = \phi(\xi_1, \mathcal{D}^*_{1,k_n}),\ T_2 = \phi(\xi_2, \mathcal{D}^*_{2,k_n}), ...$$

is a mixture of iid sequences, where the mixture is directed (in the sense of Aldous (1985)) by $\mathcal{D}_n$. So, $\{T_l \mid \mathcal{D}_n\}$ is exactly an iid sequence of functions. Further, $\{r_l \mid \mathcal{D}_n, \boldsymbol{Z}^*\}$ is an iid sequence of random variables, and thus the conclusion follows from the converse of DeFinetti's Theorem. □

See Aldous (1985) page 29 for more details on this construction.

We turn now to Lemma 1 from Section 3.2, which establishes asymptotic pairwise independence.

Lemma 1. *Consider a collection of B_n trees built from a training dataset of size n on subsamples of size k_n, say $\{T_{j,k_n}\}_{j=1}^{B_n}$, satisfying Condition 1. Then, as long as $k_n/\sqrt{n} \to 0$ and*

$$\binom{B_n}{2} \log\left[\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right] \to 0$$

the infinite sample sequence of trees, $\{T_{1,\infty,k_\infty}, ..., T_{B,\infty,k_\infty}, ...\}$ is an infinite sequence of pairwise independent random functions.

Proof. Condition 1 guarantees the existence of a limiting random variable.

It is sufficient to show that asymptotically, the trees are trained using independent training samples, because we have assumed that our original data are iid. Define the indices of a subsample in the following way:

$$\text{ind}(\mathcal{D}^*_{k_n}) := \{j \in \{1, ..., n\} : Z_j \in \mathcal{D}^*_{k_n}\}.$$

Then, by the assumption that the Z_k are independent,

$$\mathcal{D}^*_{k_n,j} \perp\!\!\!\perp \mathcal{D}^*_{k_n,l} \iff |\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0$$

so, it is sufficient to show that

$$\lim_{n\to\infty} P(|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0) = 1,\ \forall\ j \neq l.$$

Note that if $k_n \geq n/2$, this event has probability 0, so choose n so that $n > 2k_n$. Then

$$\begin{aligned}
P(|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0) &= \frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}} \\
&= \frac{((n-k_n)!)^2}{n!(n-2k_n)!} \\
&= \frac{(n-k_n)!}{n!} \times \frac{(n-k_n)!}{(n-2k_n)!} \\
&= \frac{(n-k_n)(n-k_n-1)...(n-2k_n+1)}{n(n-1)...(n-k_n+1)}.
\end{aligned}$$

There are k_n terms in both the numerator and denominator here, so we can separate the product in the term above as

$$\begin{aligned}
P(|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0) &= \frac{n-k_n}{n} \times \frac{n-k_n-1}{n-1} \times ... \times \frac{n-2k_n+1}{n-k_n+1}. \\
&\geq \left(\frac{n-2k_n+1}{n}\right)^{k_n} \\
&= \left(1 - \frac{2k_n+1}{n}\right)^{k_n} \\
&= \exp\left[k_n \log\left(1 - \frac{2k_n+1}{n}\right)\right] \\
&\approx \exp\left[k_n\left(-\frac{2k_n+1}{n}\right) - \frac{k_n}{2}\left(\frac{2k_n+1}{n}\right)^2\right] \\
&\approx \exp\left[-\frac{2k_n^2+k_n}{n}\right] \\
&\approx 1
\end{aligned}$$

where $a_n \approx b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$, and we have used the Taylor expansion of $\log(1-x)$ in the above.

This means that two pre-specified subsamples will be independent in the limit. Next, we need to ensure that this holds for all subsamples, i.e.

$$P\left(\bigcap_{j\neq l}\{|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0\}\right) \to 1.$$

For B_n trees, there are $\binom{B_n}{2}$ subsample pairings, each drawn independently. Thus

$$P\left(\bigcap_{j\neq l}\{|\text{ind}(\mathcal{D}^*_{k_n,j})\cap\text{ind}(\mathcal{D}^*_{k_n,l})|=0\}\right)=\prod_{j\neq l}P(|\text{ind}(\mathcal{D}^*_{k_n,j})\cap\text{ind}(\mathcal{D}^*_{k_n,l})|=0)$$
$$=\left(\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right)^{\binom{B_n}{2}}.$$

Next, by assumption,

$$\log P\left(\bigcap_{j\neq l}\{|\text{ind}(\mathcal{D}^*_{k_n,j})\cap\text{ind}(\mathcal{D}^*_{k_n,l})|=0\}\right)=\binom{B_n}{2}\log\left[\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right]\to 0$$

so that the probability of this event goes to 1. □

Next, we formalize the discussion and calculations from subsection 3.2. Recall that the conditional RF MSE can be expanded as

$$MSE_{RF}(\boldsymbol{x};Y)=\underbrace{\frac{1}{B^2}\sum_{j=1}^{B}\sum_{i=1}^{B}T_j(\boldsymbol{x})T_i(\boldsymbol{x})}_{V_{2,B}}-\underbrace{\frac{2Y}{B}\sum_{j=1}^{B}T_j(\boldsymbol{x})}_{-2YRF_B(\boldsymbol{x})}+Y^2$$

where the $V_{2,B}$ is a V-statistic of order 2 and the second term is a rescaled RF prediction (that depends on $\boldsymbol{x}$, but we suppress that dependence in what follows.) Corollary 2 gives that $\sqrt{B}(-2YRF_B(\boldsymbol{x}))$ is asymptotically normal. Now, we formalize two claims made in subsection 3.2: *(i)* $\sqrt{B}V_{2,B}$ is asymptotically normal and *(ii)* $\sqrt{B}V_{2,B}$, $\sqrt{B}(-2YRF_B(\boldsymbol{x}))$ converge jointly to a bivariate normal distribution.

(i) We see that $V_{2,B}$ is a V-statistic for the kernel function $\phi(x_1,x_2)=x_1x_2$, which is permutation symmetric. As such, we can also consider the U-statistic:

$$U_{2,B}=\frac{1}{\binom{B}{2}}\sum_{1\leq i<j\leq k}T_iT_j=\frac{2}{B^2-B}\sum_{1\leq i<j\leq k}T_iT_j$$

and further define the Hájek projection of U_2 as

$$\tilde{U}_{2,B}=\sum_{i=1}^{B}\mathbb{E}(U_2|T_i)=\frac{2}{B}\sum_{i=1}^{B}\mathbb{E}(T_iT_j|T_i)=\frac{2}{B}\sum_{i=1}^{B}T_i\mathbb{E}(T_j)$$

where the second equality follows from the proof of Theorem 12.3 in Van der Vaart (2000). Letting $\mu=\mathbb{E}(T_j)$ and $\sigma^2=\text{Var}(T_j)$, the central limit theorem then gives that $\sqrt{B}\tilde{U}_{2,B}\xrightarrow{d}$

$\tilde{Z} \sim \mathcal{N}(2\mu^2, 4\mu^2\sigma^2)$ as $B \to \infty$. Thus, it suffices to show that $\sqrt{B}(V_{2,B} - \tilde{U}_{2,B}) \xrightarrow{p} 0$ to establish asymptotic normality of $V_{2,B}$.

By Theorem 12.3 in Van der Vaart (2000), $\sqrt{B}(U_{2,B} - \tilde{U}_{2,B}) \xrightarrow{p} 0$, so that it is sufficient to show that $\sqrt{B}(V_{2,B} - U_{2,B}) \xrightarrow{p} 0$. We first note that $\frac{2}{B^2 - B} \sim \frac{2}{B^2}$, as $B \to \infty$, so that

$$U_{2,B} \sim U^*_{2,B} = \frac{2}{B^2} \sum_{1 \le i < j \le B} T_i T_j.$$

Then

$$\begin{aligned}
\sqrt{B} V_{2,B} &= \frac{\sqrt{B}}{B^2} \sum_{1 \le i < j \le B} T_i T_j + \frac{\sqrt{B}}{B^2} \sum_{1 \le j < i \le B} T_i T_j + \frac{2\sqrt{B}}{B^2} \sum_{i=1}^{B} T_i^2 \\
&= \sqrt{B} U^*_{2,B} + \frac{2}{\sqrt{B}} \left(\frac{1}{B} \sum_{i=1}^{B} T_i^2 \right)
\end{aligned}$$

The quantity in the parentheses converges in probability to $\sigma^2 + \mu^2 < \infty$ (by assumption), so by Slutsky's lemma, the entire quantity converges in probability to 0. Thus, $\sqrt{B} V_{2,B} \xrightarrow{p} \sqrt{B} U^*_{2,B} \sim \sqrt{B} U_{2,B}$ and the result follows.

(ii) To show $\sqrt{B} V_{2,B}$, $\sqrt{B}(-2Y RF_B(\boldsymbol{x}))$ converge jointly to a bivariate normal distribution, we need only prove Lemma 3, stated again below.

Lemma 3. *Let $[U_l]_{l=1}^k$ be elliptic for $\left\{[T_{1,i}, ..., T_{m,i}]^T\right\}_{i=1}^B \overset{iid}{\sim} P$. Then, the vector $U_{k,m,B}$, defined analogously to $\tilde{U}_{k,m,B}$ in Equation 3.4, satisfies*

$$\sqrt{B}[U_{k,m,B} - \mathbb{E}(U_{k,m,B})] \xrightarrow{d} Z_\Sigma \sim \mathcal{N}(0, \Sigma) \quad \text{as } B \to \infty.$$

Proof. The Hájek vector is defined as

$$\begin{aligned}
\tilde{U}_{k,m,B} &= \sum_{i=1}^{B} \mathbb{E}\big(U_{k,m,B} \big|\ [T_{1,i}, ..., T_{m,i}]^T\big) \\
&= \frac{1}{B} \sum_{i=1}^{B} \begin{bmatrix} r_1 \mathbb{E}\big(\phi_1(T_{1,i}, ..., T_{r_1}) \big|\ [T_{1,i}, ..., T_{m,i}]^T\big) \\ \vdots \\ r_k \mathbb{E}\big(\phi_k(T_{1,i}, ..., T_{r_k}) \big|\ [T_{1,i}, ..., T_{m,i}]^T\big) \\ \vdots \\ r_k \mathbb{E}\big(\phi_k(T_{m,i}, ..., T_{r_k}) \big|\ [T_{1,i}, ..., T_{m,i}]^T\big) \end{bmatrix}
\end{aligned}$$

which is an iid sum with finite, positive definite covariance matrix Σ. Thus, the Multivariate Central Limit Theorem gives that

$$\sqrt{B}[\tilde{U}_{k,m,B} - \mathbb{E}(\tilde{U}_{k,m,B})] \xrightarrow{d} Z_\Sigma \sim \mathcal{N}(0, \Sigma) \quad \text{as } B \to \infty$$

Next, by assumption that each U-statistic kernel ϕ_l has bounded second moments, Theorem 12.3 in Van der Vaart (2000), gives that $\sqrt{B}||U_{k,m,B} - \tilde{U}_{k,m,B}|| \xrightarrow{p} 0$ as $B \to \infty$, where $||\cdot||$ is any norm on $\mathbb{R}^{km}$. Finally, the result follows from the fact that $\mathbb{E}(\tilde{U}_{k,m,B}) = \mathbb{E}(U_{k,m,B})$. □

Next, we move on to the proof of Proposition 1, restated below.

Proposition 1. *Assume that $Y = m(\boldsymbol{X}) + \epsilon$, where $m(\cdot)$ is continuous on the unit cube. Let $\mathcal{X} = [0,1]^p$, and assume that $X_{i,j} \overset{iid}{\sim} Unif(0,1)$ for $i = 1, ..., n$ and $j = 1, ..., p$. Then, let $T_n(\boldsymbol{x})$ be a tree trained on iid pairs $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$ such that each leaf of the tree contains a single observation. Further, assume the trees satisfy the following two conditions:*

(i) $\exists \gamma > 0$ *such that* $P(\text{variable } j \text{ is split on}) > \gamma$ *for* $j \in \{1, ..., p\}$

(ii) *Each split leaves at least γn observations in each node.*

Then, for each $\boldsymbol{x} \in \mathcal{X}$

$$T_n(\boldsymbol{x}) \xrightarrow{d} Y|\boldsymbol{X} = \boldsymbol{x} \text{ as } n \to \infty$$

Proof. Each tree divides $\mathcal{X}$ into a partition of rectangular subspaces, corresponding to leaves of the tree. Following Meinshausen (2006), for each point $\boldsymbol{x}$ (with coordinates $[x_1, ..., x_p]$), let $\ell(\boldsymbol{x})$ denote the unique leaf of the tree containing $\boldsymbol{x}$. Let $R_\ell(\boldsymbol{x})$ be the rectangular subspace of $[0,1]^p$ corresponding to a particular leaf $\ell(\boldsymbol{x})$. The rectangular nature of the subspaces means that for each input feature, R_ℓ can be expressed as

$$R_\ell(\boldsymbol{x}) = \bigotimes_{i=1}^{p} [a(\boldsymbol{x}, i), b(\boldsymbol{x}, i)]$$

where $0 \leq a(\boldsymbol{x}, i) \leq x_i \leq b(\boldsymbol{x}, i) \leq 1$ are scalars inducing an interval in dimension i. Then, the tree (by the existence of the requisite γ) satisfies the conditions of Lemma 2 in Meinshausen (2006), so that $\max_i |a(\boldsymbol{x}, i) - b(\boldsymbol{x}, i)| \xrightarrow{p} 0$. Let $\boldsymbol{a}(\boldsymbol{x}) = [a(\boldsymbol{x}, 1), ..., a(\boldsymbol{x}, p)]$ and similarly define $\boldsymbol{b}(\boldsymbol{x})$, so that the previous sentence implies: $\boldsymbol{a}(\boldsymbol{x}) \xrightarrow{p} \boldsymbol{b}(\boldsymbol{x})$. We therefore also see that $a(\boldsymbol{x}, i), b(\boldsymbol{x}, i) \xrightarrow{p} x_i$ for all i.

The trees are fully grown, so the tree prediction at the point $\boldsymbol{x}$ is given by

$$T_n(\boldsymbol{x}) = \sum_{k=1}^{n} I(\boldsymbol{X}_k \in R_\ell(\boldsymbol{x}))Y_k$$

i.e. the response for the observation whose leaf contains $\boldsymbol{x}$. As such, let k^* be the index corresponding to the observation who shares a leaf with $\boldsymbol{x}$, so that $T_n(\boldsymbol{x}) = Y_{k^*}$. We can deconstruct the event $\boldsymbol{X}_{k^*} \in R_\ell(\boldsymbol{x})$ as

$$\{\boldsymbol{X}_{k^*} \in R_\ell(\boldsymbol{x})\} = \bigcap_{i=1}^{p} \{a(\boldsymbol{x}, i) \leq X_{i,k^*} \leq b(\boldsymbol{x}, i)\}.$$

Thus, in the limit, $a(\boldsymbol{x}, i), b(\boldsymbol{x}, i) \xrightarrow{p} X_{i,k^*}$, and so $X_{i,k^*} \xrightarrow{p} x_i$ for all i. Further, continuity of m yields that $m(\boldsymbol{X}_{k^*}) \xrightarrow{p} m(\boldsymbol{x})$. Thus, we see that, in the limit

$$Y_{k^*} = m(\boldsymbol{x}) + \epsilon_{k^*} \xrightarrow{d} m(\boldsymbol{x}) + \epsilon \xrightarrow{d} Y | \boldsymbol{X} = \boldsymbol{x}$$

because ϵ_{k^*} is independent of the location of $\boldsymbol{X}$. □

A.1 Permutation Test Proofs

We now present the proofs from subsection 3.3. We begin with Lemma 4.

Lemma 4. *Let $X_1, ..., X_B$ be an exchangeable sequence and define $T_1 = T_1(X_1, ..., X_B)$ to be a statistic such that, for iid permutations π, π':*

$$(T_1(X_{\pi(1)}, ..., X_{\pi(B)}), T_1(X_{\pi'(1)}, ..., X_{\pi'(B)})) \xrightarrow{d} (T, T') \quad \text{as } B \to \infty$$

where $T, T' \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$, so that by Theorem 15.2.3 of Lehmann and Romano (2006):

$$\sup_t |\hat{J}_B^{T_1}(t) - \Phi(t/\tau)| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where $\hat{J}_B^{T_1}$ is the permutation distribution defined in Section 1. Then, let $T_2 = T_2(X_1, ..., X_B)$ be another statistic. If $T_2 \xrightarrow{p} T_1$, then

$$\sup_t |\hat{J}_B^{T_2}(t) - \Phi(t/\tau)| \xrightarrow{p} 0 \quad \text{as } B \to \infty.$$

Proof. By the exchangeability assumption, the distribution of T_2 is invariant under permutation of the observations, so that

$$\begin{pmatrix} T_2(X_{\pi(1)}, ..., X_{\pi(B)}) \\ T_2(X_{\pi'(1)}, ..., X_{\pi'(B)}) \end{pmatrix} \xrightarrow{p} \begin{pmatrix} T_1 \\ T_1' \end{pmatrix}$$

as $B \to \infty$, where convergence in probability here is with respect to the metric $|A - B| = |a_1 - b_1| + |a_2 - b_2|$ for 2×1 vectors A, B. Thus, (because convergence in probability implies convergence in distribution) we see that

$$\begin{pmatrix} T_2(X_{\pi(1)}, ..., X_{\pi(B)}) \\ T_2(X_{\pi'(1)}, ..., X_{\pi'(B)}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} T_1 \\ T_1' \end{pmatrix} \stackrel{d}{=} (T, T')$$

where T, T' are as in the statement of the lemma. Thus, T_2 satisfies the conditions of Theorem 15.2.3 in Lehmann and Romano (2006). □

Theorem 3. *Let $P(\boldsymbol{x}) = P$ denote the distribution of tree predictions at $\boldsymbol{x}$ for trees trained on on $\mathcal{D}_n$, and let $Q(\boldsymbol{x}) = Q$ denote the distribution for tree predictions at $\boldsymbol{x}$ for trees trained on $\mathcal{D}_n^\pi$. Assume that the subsample size and number of trees is such that the trees become iid in the large sample limit (i.e. the trees satisfy Lemma 1). If $\mathbb{E}_P(T(\boldsymbol{x})) = \mathbb{E}_Q(T(\boldsymbol{x})) = \mu$ and $\mathbb{E}_P(T^2(\boldsymbol{x})) = \mathbb{E}_Q(T^2(\boldsymbol{x})) = \sigma^2$, then the randomization distribution of $U = \sqrt{B}\frac{1}{\binom{B}{2}} \sum_{1 \le i < j \le B} (T_i^\pi(\boldsymbol{x}) T_j^\pi(\boldsymbol{x}) - T_i(\boldsymbol{x}) T_j(\boldsymbol{x}))$ satisfies*

$$\sup_t |\hat{J}_B^U(t) - \Phi(t/\tau)| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where τ is such that $U \xrightarrow{d} \tau Z$ for Z a standard normal random variable.

Proof. We rely on variance calculations associated with Theorem 3.1 of Chung and Romano (2016). Note that our kernel is of the form

$$\varphi(X_1, X_2, Y_1, Y_2) = X_1 X_2 - Y_1 Y_2$$

so that it satisfies the antisymmetry condition. Next, we note that the asymptotic unconditional variance of the U-statistic is given by the following:

$$\tau^2 = 2^2 \Bigg[\iiiint (X_1 X_2 - Y_1 Y_2)^2 dP(X_1) dP(X_2) dQ(Y_1) dQ(Y_2) + \frac{1/2}{1 - 1/2} \iiiint (X_1 X_2 - Y_1 Y_2)^2 dP(X_1) dP(X_2) dQ(Y_1) dQ(Y_2) \Bigg]$$

After tedious calculations, this yields $\tau^2 = 16(\sigma^4 - \mu^4)$. Next, let $\bar{P} = \frac{1}{2}(P + Q)$. Then, by Theorem 3.1 of Chung and Romano (2016), the randomization distribution satisfies

$$\sup_t |\hat{J}_B^U(t) - \Phi(t/\bar{\tau})| \xrightarrow{p} 0$$

where $\bar{\tau}^2 = \frac{2^2}{1/2}\iiiint (X_1X_2 - Y_1Y_2)^2 d\bar{P}(X_1)d\bar{P}(X_2)d\bar{P}(Y_1)d\bar{P}(Y_2)$. We can expand this quantity as follows

$$\bar{\tau}^2 = \frac{8}{16}\iiiint (X_1X_2 - Y_1Y_2)^2 (dP(X_1) + dQ(X_1))(dP(X_2) + dQ(X_2)) \\ (dP(Y_1) + dQ(Y_1))(dP(Y_2) + dQ(Y_2))$$

Now, we note that the above can be broken down into a sum of integrals with respect to differing joint measures, each of which is a term in the expansion of $(dP + dQ)^4$, so that we arrive at the following expression for $\bar{\tau}^2$

$$\begin{aligned}
\bar{\tau}^2 = &\frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dP(X_1)dP(X_2)dP(Y_1)dP(Y_2) + \\
&\frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dQ(X_1)dP(X_2)dP(Y_1)dP(Y_2) + \\
&\cdots + \frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dP(X_1)dP(X_2)dP(Y_1)dQ(Y_2) + \\
&\frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dP(X_1)dP(X_2)dQ(Y_1)dQ(Y_2) + \\
&\cdots + \frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dQ(X_1)dQ(X_2)dP(Y_1)dP(Y_2) + \\
&\frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dP(X_1)dQ(X_2)dQ(Y_1)dQ(Y_2) + \\
&\cdots + \frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dQ(X_1)dQ(X_2)dQ(Y_1)dP(Y_2) + \\
&\frac{1}{2}\iiiint (X_1X_2 - Y_1Y_2)^2 dQ(X_1)dQ(X_2)dQ(Y_1)dQ(Y_2).
\end{aligned}$$

The number of terms between each portion of the sum corresponds to the binomial coefficient so that the entire sum above is over 16 total terms. The integrand does not change for each term; in particular, it is expanded as

$$(X_1X_2 - Y_1Y_2)^2 = X_1^2X_2^2 - X_1X_2Y_1Y_2 + Y_1^2Y_2^2.$$

By assumption, integration with respect to either P or Q of any of these terms is the same (namely $2\sigma^4 - 2\mu^4$), so that $\bar{\tau}^2 = 16(\sigma^4 - \mu^4)$, which is equal to τ^2, thus completing the proof. □

Now for Lemma 5:

Lemma 5. *Define*

$$\tilde{T}_{2B} = [T_1(X_{\pi(2B)}), T_1(X_{\pi'(2B)}), \cdots, T_M(X_{\pi(2B)}), T_M(X_{\pi'(2B)})]^T$$

where T_j is a U-statistic for each j, and further for a vector $t \in \mathbb{R}^{2M}$, define $T_{t,2B} = t^T \tilde{T}_{2B}$. In particular, consider $c = [c_i]_{i=1}^{2M}$ and $c' = [c'_i]_{i=1}^{2M}$, where $c_i = I\{i \text{ is odd}\}$ and $c'_i = I\{i \text{ is even}\}$. If each pair $\left(T_j(X_{\pi(2B)}), T_j(X_{\pi'(2B)})\right)$ satisfies

$$\left(T_j(X_{\pi(2B)}), T_j(X_{\pi'(2B)})\right) \xrightarrow{d} T_j, T'_j \overset{iid}{\sim} \mathcal{N}(0, \tau_j^2)$$

such that $\tilde{T}_{2B}$ are elliptic for $X_{\pi(2B)}$ for any permutation π, then we see that

$$(T_{c,2B}, T_{c',2B}) \xrightarrow{d} (T, T')$$

where $T, T' \overset{iid}{\sim} \mathcal{N}(0, \tau_M^2)$, so that

$$\sup_t |\hat{J}_B^T(t) - \Phi(t/\tau_T)| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where $\tau_M^2 = \sum_{j=1}^M \tau_j^2 + 2\sum_{1 \le j < k \le M} Cov(T_j, T_k)$.

Proof. Let $\tilde{T} = [T_1, T'_1, \cdots, T_M, T'_M]$ be the collection of limiting random variables. We begin by noting that because the U-statistics in $\tilde{T}_{2B}$ are elliptic for the permuted X terms, Lemma 3 gives that $\tilde{T}$ has a multivariate normal distribution. In particular, we see that

$$\tilde{T}_{2B} \xrightarrow{d} Z_0 \sim \mathcal{N}(0, \Sigma) \tag{A.1}$$

where Σ is a $2M \times 2M$ matrix with the structure

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1M} \\ \vdots & \ddots & \vdots \\ \Sigma_{1M} & \cdots & \Sigma_{MM} \end{bmatrix}.$$

By the independence of the permutations, each Σ_{jk} has the structure

$$\Sigma_{jk} = \begin{bmatrix} \mathrm{Cov}(T_j, T_k) & 0 \\ 0 & \mathrm{Cov}(T'_j, T'_k) \end{bmatrix}.$$

By assumption, each pair $T_j, T'_j \overset{iid}{\sim} \mathcal{N}(0, \tau_j^2)$ so that

$$\rho_{jk} := \mathrm{Cov}(T_j, T_k) = \mathrm{Cov}(T'_j, T'_k) \quad \forall\ j, k.$$

Thus, $\Sigma_{jk} = \rho_{jk} I_2$ for all j, k. For any vector $t \in \mathbb{R}^{2M}$

$$T_{t,2B} \xrightarrow{d} Z_t \sim \mathcal{N}(0, t^T \Sigma t) \tag{A.2}$$

by the converse of the Cramer-Wold device. In particular for c, c' defined earlier, we see that

$$T_{c,2B} \xrightarrow{d} Z_c \sim \mathcal{N}\left(0, \sum_{j=1}^{M} \tau_j^2 + 2 \sum_{1 \le j < k \le M} \rho_{jk}\right)$$
$$T_{c',2B} \xrightarrow{d} Z_{c'} \sim \mathcal{N}\left(0, \sum_{j=1}^{M} \tau_j^2 + 2 \sum_{1 \le j < k \le M} \rho_{jk}\right).$$

Now it suffices to show that the above convergences hold jointly, e.g. that for any $t_1, t_2 \in \mathbb{R}$

$$t_1 T_{c,2B} + t_2 T_{c',2B} \xrightarrow{d} Z_{t_1,t_2} \sim \mathcal{N}(0, (t_1^2 + t_2^2)\tau_M^2)$$

but, this follows immediately from the conclusion of Equation A.2, applied to the vectors $\boldsymbol{t}_1 = t_1 c$ and $\boldsymbol{t}_2 = t_2 c'$ (multiplication here is taken to be scalar multiplication.) □

Finally, the proof of Theorem 4.

Theorem 4. *Let $P(\{T(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{T}}) = P$ denote the joint distribution of tree predictions at the points in $\mathcal{T}$ for trees trained on $\mathcal{D}_n$, and let $Q(\{T(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{T}}) = Q$ denote the joint distribution for tree predictions at points in $\mathcal{T}$ for trees trained on $\mathcal{D}_n^{\pi}$. Assume that the subsample size and number of trees is such that the trees become iid in the large sample limit (i.e. the trees satisfy the conditions of Lemma 1). Further, assume that for all $(\boldsymbol{x}, Y) \in \mathcal{T}$, $B \in \mathbb{N}$*

$$Var_P(RF_B(\boldsymbol{x})) - Var_Q(RF_B(\boldsymbol{x})) \neq -(\mathbb{E}_P(RF_B(\boldsymbol{x})) - Y)^2 + (\mathbb{E}_Q(RF_B(\boldsymbol{x})) - Y)^2.$$

Define the test statistic

$$\Delta_B := \sqrt{B}\mathbf{1}/\boldsymbol{N}_{\boldsymbol{t}}^T (MSE_{RF^{\pi}}(\mathcal{T}) - MSE_{RF}(\mathcal{T})).$$

Assume that the collection of terms in the summation of Δ_B are elliptic for the vector of tree predictions $\left\{[T_{1,i}, ..., T_{N_t,i}]^T\right\}_{i=1}^{B}$. Then, under $H_0 : \mathbb{E}(\Delta_B) = 0$ the permutation distribution satisfies

$$\sup_t |\hat{J}_B^{\Delta}(t) - \Phi(t/\tau_{\Delta})| \xrightarrow{p} 0 \quad \text{as } B \to \infty$$

where τ_{Δ} is such that $\Delta_B \xrightarrow{d} Z_{\Delta} \sim \mathcal{N}(0, \tau_{\Delta}^2)$, and $\Phi(\cdot)$ is the standard normal cdf.

Proof. Recall the expansion of Δ_B

$$\begin{aligned}\Delta_B &= \frac{\sqrt{B}}{N_t}\sum_{(\boldsymbol{x}_\ell,Y_\ell)\in\mathcal{T}}\left(\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\left[T_j^\pi(\boldsymbol{x}_\ell)T_i^\pi(\boldsymbol{x}_\ell)-T_j(\boldsymbol{x}_\ell)T_i(\boldsymbol{x}_\ell)\right]+\frac{2Y_\ell}{B}\sum_{j=1}^{B}\left[T_j(\boldsymbol{x}_\ell)-T_j^\pi(\boldsymbol{x}_\ell)\right]\right)\\ &= \frac{1}{N_t}\sum_{(\boldsymbol{x}_\ell,Y_\ell)\in\mathcal{T}}\left(V_2\big(T_1(\boldsymbol{x}_\ell),...,T_B(\boldsymbol{x}_\ell),T_1^\pi(\boldsymbol{x}_\ell),...,T_B^\pi(\boldsymbol{x}_\ell)\big)+V_1\big(T_1(\boldsymbol{x}_\ell),...,T_B(\boldsymbol{x}_\ell),T_1^\pi(\boldsymbol{x}_\ell),...,T_B^\pi(\boldsymbol{x}_\ell)\big)\right)\\ &= \frac{1}{N_t}\sum_{\ell=1}^{N_t}(V_{2,\ell}+V_{1,\ell}).\end{aligned}$$

Lemma 4 gives that for all ℓ, $V_{2,\ell}$ has the same permutation distribution as $U_{2,\ell}$ where $U_{2,\ell}$ is defined as

$$U_{2,\ell} = \frac{\sqrt{B}}{\binom{B}{2}}\sum_{1\le i<j\le B}(T_i^\pi(\boldsymbol{x}_\ell)T_j^\pi(\boldsymbol{x}_\ell)-T_i(\boldsymbol{x}_\ell)T_j(\boldsymbol{x}_\ell)).$$

Thus, applying Theorem 3 and the assumption that $\mathbb{E}(\Delta_B)=0$ (which implies Equation 3.10), for iid permutations π,π' of $\{1,...,2B\}$, $V_{2,\ell}$ satisfies

$$\left(V_2\big(T^*_{\pi(1)}(\boldsymbol{x}_\ell),...,T^*_{\pi(2B)}(\boldsymbol{x}_\ell)\big),V_2\big(T^*_{\pi'(1)}(\boldsymbol{x}_\ell),...,T^*_{\pi'(2B)}(\boldsymbol{x}_\ell)\big)\right)\xrightarrow{d}\mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix},\tau^2_{V_2,\ell}I_2\right]$$

where $\tau^2_{V_2,\ell}$ is the asymptotic, unconditional variance of $U_{2,\ell}$. For $V_{1,\ell}$, we see that this is a rescaling of the sample mean statistic. By Theorem 15.2.5 of Lehmann and Romano (2006), $V_{1,\ell}$ satisfies, for iid permutations π,π' of $\{1,...,2B\}$

$$\left(V_1\big(T^*_{\pi(1)}(\boldsymbol{x}_\ell),...,T^*_{\pi(2B)}(\boldsymbol{x}_\ell)\big),V_1\big(T^*_{\pi'(1)}(\boldsymbol{x}_\ell),...,T^*_{\pi'(2B)}(\boldsymbol{x}_\ell)\big)\right)\xrightarrow{d}\mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix},\tau^2_{V_1,\ell}I_2\right]$$

where $\tau^2_{V_1,\ell}=4Y^2[\mathrm{Var}_P(T(\boldsymbol{x}_\ell)+\mathrm{Var}_Q(T(\boldsymbol{x}_\ell))]$. Standard calculations (such as those in Example 2.1 of Chung and Romano (2013)) show that this is identical to the unconditional variance of $V_{1,\ell}$. Then applying Lemma 5 to the statistics $\{(V_{2,\ell},V_{1,\ell})\}_{\ell=1}^{N_t}$, we attain the desired result. □

B More Simulations

We include some more simulations now to demonstrate two phenomena:

1. The instability of the variance estimation procedures laid out in Wager et al. (2014); Mentch and Hooker (2016). Clearly variance estimation is useful for developing confidence intervals about random forest predictions, which in the case of pointwise

consistency (as in the honest trees proposed by Wager and Athey (2017)), are also valid for the underlying regression function. However, in the hypothesis testing framework, these estimates are useful only insofar as they allow for calculation of a test statistic. These variance estimates, such as the infinitesmal jackknife of Wager et al. (2014), recommend building $B = \mathcal{O}(n^{\beta})$ trees where $\beta \geq 1$. We demonstrate that this recommendation cannot be violated.

2. The robustness (and potential weaknesses) of the proposed procedure to different random forest implementations. In particular, we want to study the effect of larger subsamples/more trees. The theoretical results presented in Section 3 rely on treating the tree predictions as iid. Clearly, this is never true in practice, and some theoretical justification for the effects of this being small were presented in Section 6.

B.1 Variance Estimation Instability

Here, we use the infinitesmal jackknife (IJ), as implemented in the `ranger` package (Wright and Ziegler, 2015), to estimate the variance of a random forest prediction at a given point. In particular, we simulate data from Model 2 from Table 1, train a subsampled random forest, and record the IJ variance estimate of random forest prediction at $X_1 = ... = X_5 = 0.5$ and $X_6 = ... = X_{10} = 2$. We use $n = 2000$, $k_n = n^{0.5} \approx 44$, and vary the number of trees. Often times, the IJ variance estimate is negative, leading to a `NaN` output from the IJ software. These instances represent a case when the IJ estimate is useless to a practitioner, and as such, we report the percentage of times that a `NaN` output is returned for each number of trees. For each number of trees, we repeat the simulation 100 times, and results are shown in Figure 8.

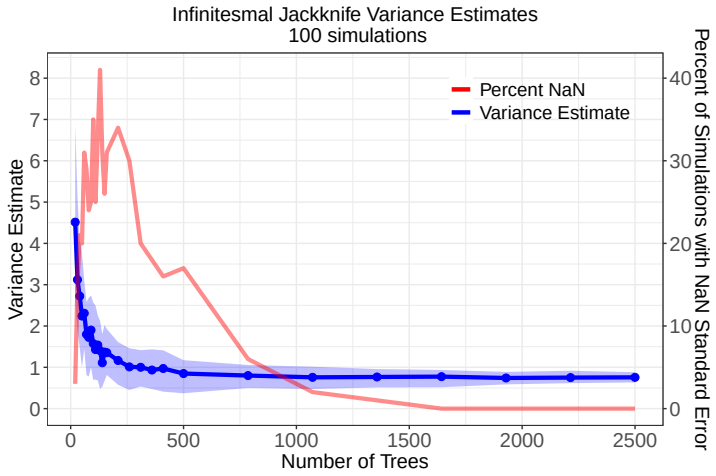


Figure 8: `ranger` IJ variance estimate. Blue ribbon plot indicates central 90% of variance estimates (corresponds to left axis), and red line (corresponds to right axis) represents percentage of runs that return `NaN`.

The IJ estimate provides overwhelmingly large variance estimates for small numbers of trees, leading to overly conservative confidence intervals and tests with exceptionally low power. Moreover, the ribbon remains quite wide until around $B = 2000$ trees, suggesting that at least $\mathcal{O}(n)$ trees are necessary to attain a stable variance estimate. A similar number of trees is necessary to ensure that a `NaN` is never returned. We should note that this is the simplest possible case of variance estimation, i.e. the estimation is only at a single point. The problem grows exponentially more complex as more test points are considered and covariance estimates are needed. Mentch and Hooker (2016) note that the procedure is infeasible for more than 20-30 test points. The authors demonstrate in follow-up work (Mentch and Hooker, 2017) that an approximate test can be produced by utilizing random projections which allows for slightly larger test sets but at the cost added computational strain. In contrast, besides the minimal overhead required to form the additional predictions, the testing procedure proposed here is almost entirely immune to the number of points in the test set. Once the initial predictions are formed, the only remaining work is to shuffle predictions (trees) and re-compute the difference in MSE between forests.

B.2 Test Robustness

We now present more figures similar to the power curves presented in Section 4. The goal here is to present the proposed procedure's robustness to the number of trees B and the subsample size k_n. To do so, we modify the simulation study plotted in Figure 3. Here,

we fix the error variance at $\sigma^2(\epsilon) = 16$, and again simulate $n = 2000$ training observations and $N_t = 100$ test observations. First, we vary the number of trees built, according to

$$B \in \{20, 50, 75, 125, 250, 375, 500, 750, 1000\}$$

and let $k_n = n^{0.6}$. The resulting simulations are plotted in Figure 9.

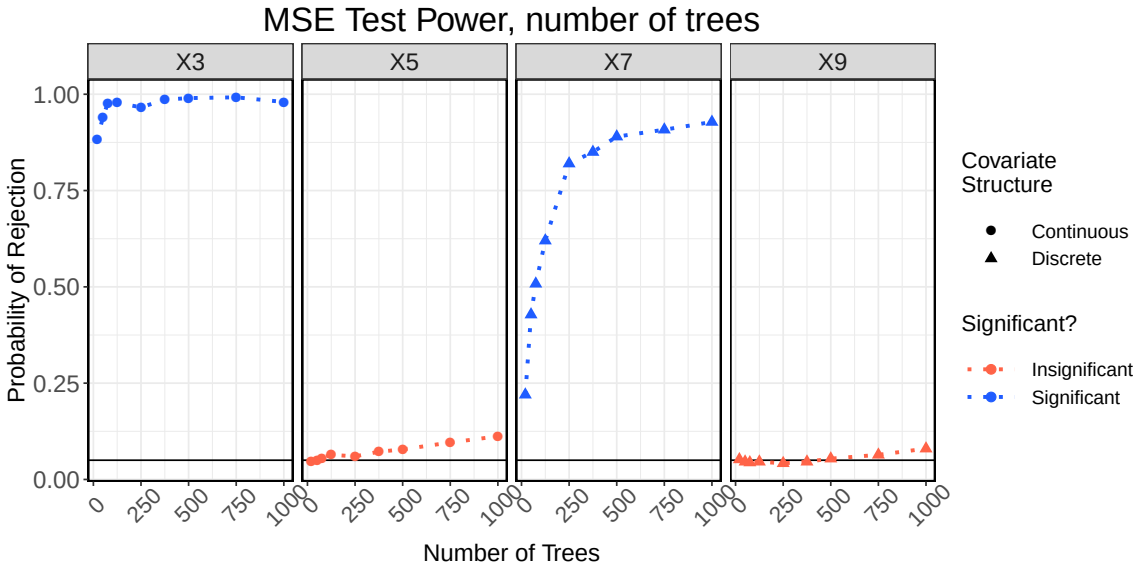


Figure 9: Model 2 power curves for 500 simulations, by number of trees. The Y-axis represents $P(\tilde{p} \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

Two clear patterns are clear in the figure - the power and type I error rate of the test both increase as the number of trees grows. However, the rate of growth for each of them is markedly different - the test attains high power around $B \approx 250$ trees, but deviations from the nominal level are only noticeable around $B \approx 500$ trees. Even when $B = 1000$, the observed level is still within nearly 5% of the baseline. Thus, while the level of the test may be slightly inflated for large numbers of trees, the procedure remains valid for limited, but realistic tree sizes.

Recall that the subsample size is a key limiting factor of Lemma 1 - it is required that $k_n = o(\sqrt{n})$ - to establish asymptotic normality. Other work (Wager and Athey, 2017) weakens these conditions, but places restrictions on the types of trees allowed in the ensemble. We now examine the behavior of our procedure under larger sample sizes. We use the same simulation parameters as in Figure 9, but now fix $B = 125$ and let $k_n = n^p$, and we vary p according to

$$p \in \{\texttt{seq(0.1, 0.99, length.out = 10)}\}$$

The resulting simulation is shown in Figure 10. We see that for $p \leq 0.75$, the Type I error rate is maintained, but for larger subsamples, we begin to see a severe deviation. Though severe, this is not necessarily surprising as such large subsampling rates correspond directly to a more severe violation of the iid approximation.

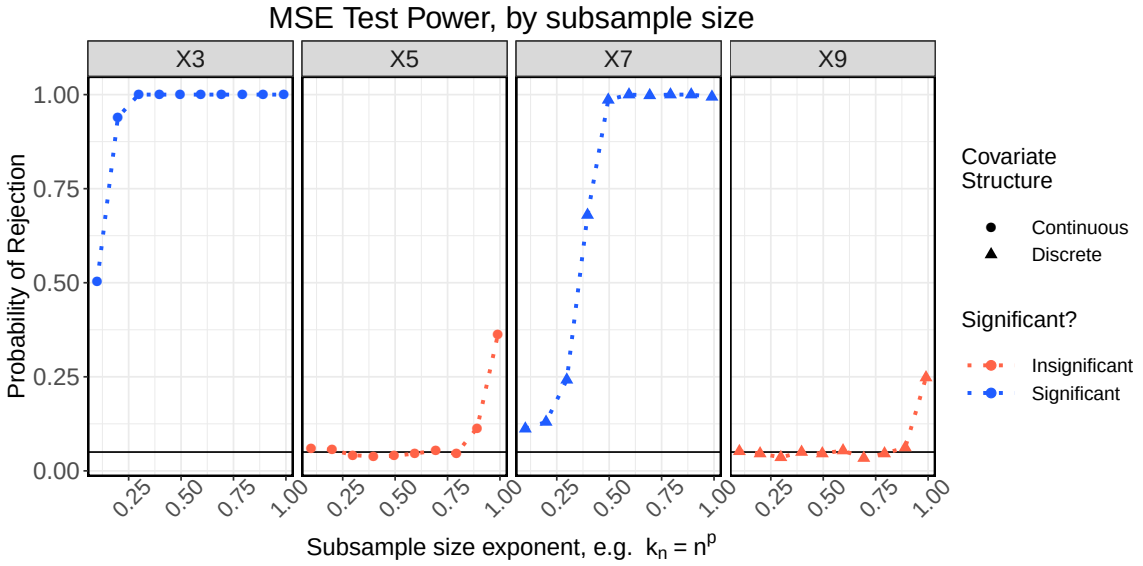


Figure 10: Model 2 power curves for 500 simulations, by subsample exponent. The Y-axis represents $P(\tilde{p} \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.