Journal of Machine Learning Research 23 (2022) 1-35 Submitted 9/20; Revised 11/21; Published 6/22

Scalable and Efficient Hypothesis Testing with Random Forests

Tim Coleman TSC35@PITT.EDU
Wei Peng WEP15@PITT.EDU
Lucas Mentch LKM31@PITT.EDU
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15215, USA

Editor: Genevra Allen

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 often precludes their practical scientific use. Here we propose a hypothesis test to formally assess feature significance, which uses permutation tests to circumvent computationally infeasible estimates of nuisance parameters. This test is intended to be analogous to the F-test for linear regression. We establish asymptotic validity of the test via exchangeability arguments and show that the test maintains high power with orders of magnitude fewer computations. Importantly, the procedure scales easily to big data settings where large training and testing sets may be employed, conducting statistically valid inference without the need to construct additional models. Simulations and applications to ecological data, where random forests have recently shown promise, are provided.

Keywords: Ensemble Methods, 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 are largely impenetrable to traditional statistical analysis. The random forests algorithm (Breiman, 2001) is among the relatively few supervised procedures for which formal statistical properties have recently been developed, paving the way for inference procedures. As detailed below, however, methods proposed to this point for assessing variable importance have either been *ad hoc* and susceptible to producing misleading and inconsistent results even in simple settings or have come with severe restrictions on the test-

ing 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, at least asymptotically. Given 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 $1, ..., N$, for $N = m + n$. Let $T = r(Z_1, ...Z_N)$ be a statistic of interest and let T_0 denote the statistic calculated on the original data. A p-value for the hypothesis the null hypothesis $H_0 : X \stackrel{d}{=} Y$ 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, often 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.

1.1 Permutation Tests

Classical work on permutation tests from Hoeffding (1952) and Lehmann et al. (1949) demonstrates the convergence of the permutation distribution to the sampling distribution for a wide variety of test statistics. Much of the 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 to which it is then shown that the permutation distribution also converges. This idea has underpinned results in Neuhaus (1993) and Janssen (2005), who provide various sufficient conditions for the convergence to the unconditional distribution.

Permutation tests are exact tests for 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 prove results regarding the asymptotic validity of our procedure for more general hypotheses. The individual models (base learners) in supervised ensembles, such as decision trees in a random forest, naturally lend themselves to the permutation framework by being exchangeable in many practical cases.

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 can overfit the data without careful pruning. Bagging stabilizes the variance by training many individual learners on bootstrap samples. Random forests (Breiman, 2001) augment the bagging procedure by introducing auxiliary randomness in the construction of each individual 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.

Recent years have seen something of a surge in the development of formal statistical analyses of random forests. 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 under restrictions on subsample size. Wager and Athey (2018) proved both consistency and asymptotic normality for subsampled random forests whenever trees are restricted to being built according to honesty and regularity conditions and large numbers of trees are constructed. The original random forest formulation has also been extended to various setups including quantile regression (Meinshausen, 2006), survival analysis (Ishwaran and Lu, 2008; Cui et al., 2017), reinforcement learning (Zhu et al., 2015), and a generalized framework allowing random forests weights to be used for general local parameter estimation (Athey et al., 2016).

In addition to their robust history of empirical success and these newly-developed statistical properties, the availability of *ad hoc* tools for evaluating variable importance has also been a major contributing factor to their continued widespread practical use. Among these, thanks in large part to their computational feasibility, the out-of-bag (oob) measures proposed by Breiman (2001) remain the most popular by a substantial margin, with versions of this measure available in nearly every major statistical software. Unfortunately, in the decades since their introduction, a substantial amount of literature has repeatedly demonstrated their inadequacy and inconsistency; see Strobl et al. (2007) and Toloşi and Lengauer (2011) as popular, representative examples. Among the issues with oob measures are is a tency to inflate the relative importance of categorical covariates with many levels as well as those with high correlation to others. The latter issue is particularly problematic as variables deemed most important may have relatively little impact on the response but be highly dependent only on each other. Recent work by Hooker and Mentch (2019) gives an explanation for this behavior based on extrapolation.

In light of these issues, recent work has sought to cast the issue of variable importance more formally in a classical hypothesis testing framework. Notably, Mentch and Hooker (2016) showed an equivalence between subsampled random forests and infinite-order U-statistics,

allowing for asymptotic normality to be established from which a formal hypothesis testing procedure for evaluating variable importance can be derived. 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. Accurate estimation of the covariance necessitates constructing a very large number of trees and becomes computationally infeasible for more than 20-30 test points, even when the original dataset is relatively small. Mentch and Hooker (2017) extend 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 (2018) requires the number of trees constructed be at least on the order of n to be valid. Thus, while these kinds of procedures can be shown to successfully alleviate the troublesome issues with the classical oob measures, their computational complexity precludes their use in the vast majority of practical settings where flexible procedures like random forests may hold the most promise.

In contrast with these previous approaches, this work develops a formal testing framework for variable importance that is both computationally efficient and statistically valid. In particular, our procedure is almost entirely computationally agnostic to the number of test points utilized. 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 the predictive importance of any given subset of covariates 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. We note also that while our focus here is on random forests, only a small portion of the theory we provide is tree-specific and thus ensembles consisting of other kinds of base learners could easily fit within this testing framework as well.

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. In Section 4, we present simulation studies of the testing procedure for a variety of underlying regression functions, as well as a comparison with two different knockoff statistics. In Section 5, we apply our procedure to multiple ecological datasets where random forests have been successfully employed in recent 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

Consider a sample $\mathcal{D}_n = \{Z_1, Z_2, ..., Z_n\}$, with $Z_i = (\boldsymbol{X}_i, Y_i)$ 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 the random forest procedure is to accurately estimate $m(\boldsymbol{x})$. Each tree in a 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. This process is repeated B times and the random forest prediction at some point $\boldsymbol{x} \in \mathcal{X}$ is given by

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

We can similarly evaluate the RF prediction accuracy at a single fixed test location $\boldsymbol{x}$ and with true response value y via its mean squared error

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

Similarly, we can write 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 Eq. (1), but with $\mathcal{D}_n$ replaced by $\mathcal{D}_n^\pi$, where $\mathcal{D}_n^\pi$ replaces some subset of features with an alternate copy drawn independent of Y given the rest of the covariates. To make this concrete, suppose that this subset consists of just a single feature X_j. We can then evaluate whether X_j is important by conducting a test of the following hypotheses

$$\begin{aligned} H_0^j &: \mathbb{E}(MSE_{RF}(\mathcal{T})) = \mathbb{E}(MSE_{RF^\pi}(\mathcal{T})) \\ H_1^j &: \mathbb{E}(MSE_{RF}(\mathcal{T})) < \mathbb{E}(MSE_{RF^\pi}(\mathcal{T})) \end{aligned} \tag{2}$$

where the expectation is taken over the training data and auxiliary randomness. Though conditional on $\mathcal{T}$, we stress that the computational complexity of the testing procedure we employ is almost entirely immune to the size of this test set, thus effectively allowing practitioners to evaluate the hypothesis at as many locations as are desired. We call X_j important if we are able to reject H_0^j, and correspondingly measure its importance as the difference in MSEs, $MSE_{RF^\pi}(\mathcal{T}) - MSE_{RF}(\mathcal{T})$. This definition of importance is model based and therefore different 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

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

The standard knockoff procedure controls the False Discovery Rate (FDR) for hypotheses about each of the covariates for arbitrary distributions over (X, Y). It should be noted that conditional independence of X_j and Y is neither necessary nor sufficient for H_0^j. However, in practice, the test statistic utilized in the knockoff procedure is generally taken as the difference in importance measures between original and knockoff variables and thus the outcome of the procedure itself remains highly model dependent. We also note that our procedure, while it could use knockoffs, does not require knowledge of the distribution

of the covariates. We also want to highlight another popular means of non-parametric hypothesis testing, *conformal inference* introduced in Vovk et al. (2005), and applied to similar problems to the one considered here in Lei et al. (2018). Conformal inference analyzes the quantity (X_{n+1}, Y_{n+1}), which is an observation drawn from the joint distribution of the covariates and response, rather than the conditional quantity presented in Eq. (2), which is the focus of our work.

2.1 Testing Procedure

Intuitively, 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 ensembles 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 of response 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. 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. Importantly, this procedure requires only $2B$ trees, regardless of the size of the test set.

Pseudo-code for the permutation test is provided in Algorithm 1. We use $\oplus$ to denote concatenation of data matrices by column, $\uplus$ to denote concatenation by row, and $\ominus$ to denote the removal of columns from a dataset. 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. This suffices to make the testing procedure slightly more conservative, but more amenable to potential p-value transforming procedure, like an FDR filter; see, for example, Phipson and Smyth (2010) for a more thorough discussion. 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.

3. Establishing Statistical Validity

We now develop the theoretical backing for the hypothesis testing procedure outlined above. In Section 3.1, we make explicit the connection between bagged-models and exchangable random variables. Then, in Section 3.2, we use these results establish asymptotic normality for subsampled random forest predictions, under mild conditions. Next, in Section 3.3, we extend these results to the fixed-test set MSE, establishing a CLT for this quantity. To use these distributions directly is unwieldy - there is no obvious consistent estimator of the variance parameters available. Thus, in Section 3.4, we prove that the proposed permutation test is asymptotically equivalent to the computationally infeasible parametric alternative, building on recent arguments from Chung and Romano (2013). For readability, most proofs are reserved for Appendix A.

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, $\boldsymbol{X}_S$, n_{perm} number of permutations to evaluate

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

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

end

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

3.1 Exchangeable Random Variables & Permutation Tests

Recall that a sequence of random variables $X_1, X_2, ...$ is exchangeable if $(X_{i_1}, X_{i_2}, ..., X_{i_k}) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, ..., X_{\pi(k)})$ for every finite sub-collection indexed by $i_1, ..., i_k$ and every permutation of the indices $\pi(\cdot)$, see Aldous (1985) for a thorough review.

Permutation tests naturally lend themselves to exchangeable data by providing a means of evaluating the 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

$$\mu(X_1, ..., X_n, Y_1, ..., Y_m) = \prod_{i=1}^{n} P(X_i) \prod_{j=1}^{m} Q(Y_j)$$

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

Modern work for permutation tests has focused largely on modifications needed to account for 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 a two sample problem, with $X_1, ..., X_n \overset{iid}{\sim} P_X = \mathcal{N}(0, 5)$ and independently let $Y_1, ..., Y_m \overset{iid}{\sim} P_Y = \mathcal{N}(0, 1)$. Clearly, median$(P_X)$ = median(P_Y), but the data are no longer exchangeable and so an unstudentized permutation test of H_0 : median(P_X) = median(P_Y) is no longer valid at a pre-specified level (see Chung and Romano (2013) for details). However, note that exchangeability is violated only because the data are no longer identically distributed; permutation tests can remain valid for data that are correlated but identically distributed so long as the pairwise dependence is

constant. The upshot of this is that random forest ensembles possess this property, and so can be shown to be exchangeable, as formalized in Theorem 1.

Theorem 1 *Denote a sequence of (potentially randomized) subsampled trees as* $\{T_k(\cdot)\}_1^\infty$. *Under the conditions outlined above, the residuals at* $\boldsymbol{Z}^* = (\boldsymbol{X}^*, Y^*) \sim F$ *given by*

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

form an infinitely exchangeable sequence of random variables.

In the case of a single random forest, exchangeability is readily apparent as the order in which trees are trained has no bearing on their structure. Indeed, Theorem 1 can be extended to any bagged learning method.

Given a dataset $\mathcal{D}_n$ with $n \times p$ design matrix $\boldsymbol{X}$, let $\mathcal{S} \subset \{1, ..., p\}$ and define $\boldsymbol{X}_\mathcal{S} = \{X_j : j \in \mathcal{S}\}$ and $\boldsymbol{X}_{-\mathcal{S}} = \{X_j : \ j \notin \mathcal{S}\}$; we take $\boldsymbol{X}_\mathcal{S}$ to be the covariates of interest. We then create a randomized version of $\boldsymbol{X}_\mathcal{S}$ independent of Y, denoted 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 $[\boldsymbol{X}_{-\mathcal{S}}, \boldsymbol{X}_\mathcal{S}] \stackrel{d}{=} [\boldsymbol{X}_{-\mathcal{S}}, \boldsymbol{X}_\mathcal{S}^\pi]$. By construction, $\boldsymbol{X}_\mathcal{S}^\pi$ ind $Y|\boldsymbol{X}_{-\mathcal{S}}$ 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 so that $T(\boldsymbol{x}; \mathcal{D}_n^\pi)$ ind $Y|\ \boldsymbol{X}_{-\mathcal{S}}$, allowing for the testing of a null hypothesis of conditional independence.

3.2 Asymptotic Behavior of Trees

Within-forest exchangeability is not sufficient to justify the proposed testing procedure at the nominal level. Instead, we need to establish sufficient conditions to justify exchanging trees between forests. An important step in this direction is to establish the existence of a limiting sequence of subsampled trees that behave like an iid sequence.

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

In Section 3.3.1 we provide sufficient conditions for this to hold. We note that this condition is similar in spirit to Assumption 15.7.1 in Lehmann and Romano (2006), which is fundamental to the validity of subsampling based intervals for model parameters.

In practice, we would like to establish results for random forests trained on growing subsamples. If we insist that the subsample size k_n grow slower than $\sqrt{n}$, we obtain the following intuitive result.

Lemma 2 *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 on the number of trees B_n is likely not of much practical importance. For finite B_n, the probability sequence has the form of a_n^B, where $a_n = \frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}$, so because $a_n \to 1$, a_n^B 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. For this, we turn to a result from Aldous (1985).

Lemma 3 *(Aldous, 1985) Let $Z_1, Z_2, ...$ be an infinitely exchangeable sequence. If $Z_i \perp\!\!\!\perp Z_j, i \neq j$, then $Z_1, Z_2, ...$ is a sequence of iid random variables.*

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

Corollary 4 *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 can obtain the following pointwise central limit theorem.

Corollary 5 *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 as $n \to \infty$*

$$\sqrt{B_n}\left[\frac{1}{B_n}\sum_{i=1}^{B_n} T_{i,k_n}(\boldsymbol{x}) - \mathbb{E}\left(\frac{1}{B_n}\sum_{i=1}^{B_n} T_{i,k_n}(\boldsymbol{x})\right)\right] \xrightarrow{d} \mathcal{N}(0, \sigma^2(\boldsymbol{x})). \tag{3}$$

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 6 *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 as $n \to \infty$*

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

Remark 7 *We can generalize the independence results to a collection of two sets of trees. In particular, suppose that we now train $B_n/2$ trees on $\mathcal{D}_n = \{Z_i\}_{i=1}^n$ and $\mathcal{D}_n^\pi = \{Z_i^\pi\}_{i=1}^n$, where $Z_i^\pi = ([\boldsymbol{X}_\mathcal{S}, \boldsymbol{X}_{-\mathcal{S}}^\pi]_i, Y_i)$. Note that $Z_i^\pi \perp\!\!\!\perp Z_j, \forall\, i \neq j$, so there is the same independence structure between the datasets as within. Thus, the probability that a pair of trees trained on subsamples of size k_n, one from $\mathcal{D}_n$ and one from $\mathcal{D}_n^\pi$, are independent is the same as*

the probability that a pair of trees within forest are independent. As such, $\{T_{i,k_n}(\boldsymbol{x})\}_{i=1}^{B_n}$ *and* $\{T^{\pi}_{i,k_n}(\boldsymbol{x})\}_{i=1}^{B_n}$*, where* B_n, k_n *satisfy the conditions of Lemma 1, behave like two independently iid samples.*

We intentionally leave $\sigma(\boldsymbol{x})$ as an abstraction since estimation of $\sigma(\boldsymbol{x})$ is not straightforward. Instead, this result will be used as the basis for asymptotic validity of our permutation test which, uncharacteristically, is far more computationally efficient. Going forward, 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 largely a matter of notational convenience; we could explicitly include the dependence on n in each of the following statements and stress that the limiting distributions only hold as $n \to \infty$.

3.3 Asymptotic Distribution of MSEs

Unfortunately, the MSE is not a linear function of exchangeable random variables and thus requires more careful attention before being used as a test statistic in a permutation test. In this subsection we establish the asymptotic normality of the MSE which we then utilize to show that the difference in MSEs between two forests is asymptotically normal. To begin, consider a single test point $(\boldsymbol{x}, y)$. We can write the MSE as

$$MSE_{RF}(\boldsymbol{x}; y) = g\left(\frac{1}{B}\sum_{i=1}^{B} T_i(\boldsymbol{x}), y\right) \tag{4}$$

where $g(a, b) = (a - b)^2$. In what follows, we suppress the dependence on y, writing just $MSE_{RF}(\boldsymbol{x}; y) = g\left(RF_B(\boldsymbol{x})\right)$. We derive the asymptotic distribution of the MSE via the delta method, which we belabor here for its intuitive value. We can then appeal to the mean value theorem to say

$$g(RF_B(\boldsymbol{x})) = g(\mathbb{E}RF_B(\boldsymbol{x})) + g'(\tilde{R}_B(\boldsymbol{x}))[RF_B(\boldsymbol{x}) - \mathbb{E}RF_B(\boldsymbol{x})]$$

where $\tilde{R}_B(\boldsymbol{x})$ is a random quantity bounded between $RF_B(\boldsymbol{x}), \mathbb{E}RF_B(\boldsymbol{x})$. The law of large numbers gives that $RF_B(\boldsymbol{x}) = \mathbb{E}RF_B(\boldsymbol{x}) + o_P(1)$ and further $\tilde{R}_B(\boldsymbol{x}) \xrightarrow{p} \mathbb{E}RF_B(\boldsymbol{x})$. Next, continuity of g' gives that $g'(\tilde{R}_B(\boldsymbol{x})) \xrightarrow{p} g(\mathbb{E}RF_B)$. Thus,

$$\begin{aligned}\sqrt{B}\left[g(RF_B(\boldsymbol{x})) - g(\mathbb{E}RF_B(\boldsymbol{x}))\right] &= g'(\tilde{R}_B(\boldsymbol{x}))\sqrt{B}\left[RF_B(\boldsymbol{x}) - \mathbb{E}RF_B(\boldsymbol{x})\right] \\ &\xrightarrow{d} \mathcal{N}\left(0, g'(\mathbb{E}RF_B(\boldsymbol{x}))^2\sigma^2\right) \\ &\stackrel{d}{=} \mathcal{N}\left(0, 4(\mathbb{E}RF_B(\boldsymbol{x}) - y)^2\sigma^2\right) \text{ for } g(z) = (z-y)^2.\end{aligned}$$

The calculation above is more informative - we see that the MSE is asymptotically a linear function of the random forest prediction. An issue is that the above quantity is centered around $g(\mathbb{E}RF_B(\boldsymbol{x}))$ rather than $\mathbb{E}g(RF_B(\boldsymbol{x}))$, which we now address. In particular, suppose we begin by centering around $\mathbb{E}g(RF_B(\boldsymbol{x}))$ rather than $g(\mathbb{E}RF_B(\boldsymbol{x}))$. Then,

$$\begin{aligned}&\sqrt{B}\left[g\left(RF_B(\boldsymbol{x})\right) - \mathbb{E}g\left(RF_B(\boldsymbol{x})\right)\right] = \\ &\qquad \sqrt{B}\left[g\left(RF_B(\boldsymbol{x})\right) - g(\mathbb{E}RF_B(\boldsymbol{x}))\right] + \sqrt{B}\left[g\left(\mathbb{E}RF_B(\boldsymbol{x})\right) - \mathbb{E}g\left(RF_B(\boldsymbol{x})\right)\right]\end{aligned} \tag{5}$$

so that if $\sqrt{B}\left[g\left(\mathbb{E}RF_B(\boldsymbol{x})\right) - \mathbb{E}g\left(RF_B(\boldsymbol{x})\right)\right] = o(1)$, then the same distributional result holds. This is shown in Lemma 3.

Lemma 8 *Assume the conditions needed from Corollary 2. Additionally, assume that g has at least k derivatives for some $k \geq 3$, and that $g^{(k)}(x) < \infty$ for all x. Further, assume that $\mathbb{E}|T_i(\boldsymbol{x})|^k < \infty$. Then,*

$$\sqrt{B}\left[\mathbb{E}g(RF_B(\boldsymbol{x})) - g(\mathbb{E}RF_B(\boldsymbol{x}))\right] = \frac{g''(\mathbb{E}RF_B(\boldsymbol{x}))\sigma^2}{2\sqrt{B}} + o(B^{-3/2}) = o(1).$$

Since the MSE function defined as $g(RF_B(\boldsymbol{x})) = \left(RF_B(\boldsymbol{x}) - y\right)^2$ satisfies the conditions posited by Lemma 3, we can conclude that

$$\sqrt{B}\left[g\left(RF_B(\boldsymbol{x})\right) - \mathbb{E}g\left(RF_B(\boldsymbol{x})\right)\right] \xrightarrow{d} \mathcal{N}\left(0, g'(\mathbb{E}RF_B(\boldsymbol{x}))^2\sigma^2\right).$$

Application of the mean value theorem requires that $g'(\mathbb{E}RF_B(\boldsymbol{x})) \neq 0$ if and only if $\mathbb{E}RF_B \neq y$. The expected prediction can be written as $\mathbb{E}RF_B(\boldsymbol{x}) = m(\boldsymbol{x}) + \delta(\boldsymbol{x})$, where $\delta(\boldsymbol{x})$ is the pointwise bias of the random forest. Recalling that the response is given by $Y = m(\boldsymbol{x}) + \epsilon$, if it holds for all $\boldsymbol{x}$ that $P(\epsilon \neq \delta(\boldsymbol{x})) = 1$, then the result holds for the squared error calculated with respect to almost all Y and thus is trivially satisfied for continuous errors. A similar result could be applied to any continuously differentiable loss function $g(\cdot, \cdot)$, again under the condition that g' is almost surely non zero.

Remark 9 *Corollary 2 is not a necessary condition for the asymptotic normality of MSE's to hold. In fact, a similar argument could be used to justify the asymptotic normality of the MSE for any random forest who satisfies a central limit theorem and a law of large numbers (with respect to its own expectation), such as the results in Mentch and Hooker (2016) and Wager and Athey (2018).*

We can extend this result to the two forest case, where we compare the MSE of $RF_B(\boldsymbol{x})$ against that of $RF_B^{\pi}(\boldsymbol{x})$. In particular, if $\mathbb{E}MSE_{RF}(\boldsymbol{x}; y) = \mathbb{E}MSE_{RF^{\pi}}(\boldsymbol{x}; y)$, we see that

$$\sqrt{B}\left[MSE_{RF}(\boldsymbol{x}; y) - MSE_{RF^{\pi}}(\boldsymbol{x}; y)\right] \xrightarrow{d} \mathcal{N}\left(0, g'(\mathbb{E}RF_B(\boldsymbol{x}))^2\sigma^2 + g'(\mathbb{E}RF_B^{\pi}(\boldsymbol{x}))^2\sigma_{\pi}^2\right) \quad (6)$$

where $\sigma_{\pi}^2 = \mathrm{Var}(T^{\pi}(\boldsymbol{x}))$. This extension uses a similar argument as before to justify centering around the expected MSE instead of the MSE of the expected forest.

Now consider a test set with many points, denoted $\mathcal{T} = [(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{N_t}, y_{N_t})]^T$. Given the vector of random forest predictions, $RF_B(\mathcal{T})$, we can calculate the pointwise squared errors as $MSE_{RF}(\mathcal{T}) = \left[(RF_B(\boldsymbol{x}_i) - y_i)^2\right]_{i=1}^{N_t}$.

Finally, to connect back to the testing procedure proposed earlier, we now derive the asymptotic distribution of the differences in MSE between two forests. Let $MSE_{RF}(\mathcal{T})$ be the MSE of a random forest at a set of test points $\mathcal{T}$ and let $MSE_{RF^{\pi}}(\mathcal{T})$ denote the MSE of a forest trained on the partially randomized data. By the results above, under the hypothesis that $\mathbb{E}MSE_{RF}(\mathcal{T}) = \mathbb{E}MSE_{RF^{\pi}}(\mathcal{T})$, we have that as $B \to \infty$,

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

for some $\tau^2 > 0$ that does not necessarily have a form that is amenable to analysis. To calculate τ^2, for the MSE at each point in $\mathcal{T}$, let $g_j(RF_B(\boldsymbol{X}_j)) = (RF_B(\boldsymbol{X}_j) - Y_j)^2$, by continuity, $g_j'(\tilde{R}_B(\boldsymbol{X}_j)) = g_j'(\mathbb{E}RF_B(\boldsymbol{X}_j)) + o_P(1)$. Thus, we see that

$$\begin{aligned}
MSE_{RF}(\mathcal{T}) - \mathbb{E}MSE_{RF}(\mathcal{T}) &= \frac{1}{N_t}\sum_{j=1}^{N_t} MSE_{RF}(\boldsymbol{X}_j, Y_j) \\
&= \frac{1}{N_t}\sum_{j=1}^{N_t} g_j'(\mathbb{E}RF_B(\boldsymbol{X}_j))\left[RF_B(\boldsymbol{X}_j) - \mathbb{E}RF_B(\boldsymbol{X}_j)\right] + o_P(1) \\
&= \frac{1}{N_t}\sum_{j=1}^{N_t} g_j'(\mathbb{E}RF_B(\boldsymbol{X}_j))\left[\frac{1}{B}\sum_{i=1}^{B}\left[T_i(\boldsymbol{X}_j) - \mathbb{E}RF_B(\boldsymbol{X}_j)\right]\right] + o_P(1) \\
&= \frac{1}{B}\sum_{i=1}^{B}\underbrace{\frac{1}{N_t}\sum_{j=1}^{N_t} g_j'(\mathbb{E}RF_B(\boldsymbol{X}_j))\left[T_i(\boldsymbol{X}_j) - \mathbb{E}RF_B(\boldsymbol{X}_{\boldsymbol{j}})\right]}_{\bar{T}_i} + o_P(1)
\end{aligned}$$

where $g_j(\cdot)$ is used to suggest that the squared difference is calculated with respect to Y_j. $\bar{T}_i$ is an iid sequence, so that $\sqrt{B}[MSE_{RF}(\mathcal{T}) - \mathbb{E}MSE_{RF}(\mathcal{T})]$ is asymptotically an iid sum with mean 0 and variance $\sigma^2_{\bar{T}}$ given by

$$\sigma^2_{\bar{T}} = \frac{1}{N_t}\sum_{j=1}^{N_t}\sigma_j^2\left(g_j'(\mathbb{E}RF_B(\boldsymbol{X}_j))\right)^2 + \frac{2}{N_t}\sum_{i<j} g_j'(\mathbb{E}RF_B(\boldsymbol{X}_j))g_i'(\mathbb{E}RF_B(\boldsymbol{X}_i))\rho_{ij} \tag{7}$$

where $\rho_{ij} = \mathrm{Cov}(T(\boldsymbol{X}_i), T(\boldsymbol{X}_j))$ and $\sigma_j^2 = \mathrm{Var}(T(\boldsymbol{X}_j))$. We can obtain a similar variance $(\sigma^2_{\bar{T}^\pi})$ for $MSE_{RF^\pi}(\mathcal{T})$, so that under the hypothesis that $\mathbb{E}MSE_{RF}(\mathcal{T}) = \mathbb{E}MSE_{RF^\pi}(\mathcal{T})$, τ^2 can be seen to be

$$\tau^2 = \sigma^2_{\bar{T}} + \sigma^2_{\bar{T}^\pi}.$$

That the $\bar{T}_i$ and $\bar{T}_i^\pi$ are two indenpendently iid sequences follows from Lemma 2. Independence of the two samples follows from a similar argument to the second remark after Corollary 2. Crucially, there are many complicated quantities in this Eq. (7), i.e. $\sigma_j^2, \sigma^2_{\pi,j}, \rho_{ij}, \rho^\pi_{ij}$, for which there are not obvious estimators available and thus this result alone is not clearly practical. In the following sections, we verify the validity of our proposed permutation procedure, which avoids the necessary explicit estimation of these quantities.

3.3.1 Tree-specific results

Until now, our discussion has remained largely agnostic to 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 satisfy such 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 (2018).

Proposition 10 *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 a constant proportion of observations in the original 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 as a result, should have finite non zero variance. Note that Breiman (2001) recommends building trees to full depth in which case Condition 1 is automatically satisfied.

3.4 Extension to Permutation Tests

In Section 3.3 we established that the sampling distribution of MSE differences between forests was asymptotically Gaussian, but with a computationally intractable variance. Here we show that the permutation distribution converges to that sampling distribution. We begin by restating a classical theorem from Hoeffding.

Theorem 11 *(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 as $N \to \infty$

$$\big(S(X_{\pi(1)}, ..., X_{\pi(N)}), S(X_{\pi'(1)}, ..., X_{\pi'(N)})\big) \xrightarrow{d} (S, S') \tag{8}$$

where S, S' are iid with cdf $R(\cdot)$. Then for all t at which $R(\cdot)$ is continuous, $\hat{J}_N(t) \xrightarrow{p} R(t)$.

Direct application of Theorem 2 is often challenging. Suppose $\{X_i\}_{i=1}^n \overset{iid}{\sim} P_X$ and independently $\{Y_i\}_{i=1}^m \overset{iid}{\sim} P_Y$, and we calculate the statistic $\sqrt{n+m}\,[S(X_1, ..., X_n) - S(Y_1, ..., Y_m)]$, and further define $p = \lim_{n\to\infty} \frac{n}{n+m}$. Theorem 2.1 of Chung and Romano (2013) states that if there exists a function ψ_{P_Z} (which may depend on the distribution of the data, P_Z) such that

$$\sqrt{N}\,[S(Z_1, ..., Z_N) - \mathbb{E}S(Z_1, ..., Z_N)] = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_{P_Z}(Z_i) + o_{P_Z}(1) \tag{9}$$

(i.e. the statistic is asymptotically linear), then the permutation distribution of the aforementioned statistic is asymptotically normal with mean 0 and variance given by

$$\tau^2 = \frac{1}{p(1-p)} \mathrm{Var}(\psi(Z)) = \frac{1}{p(1-p)}\,[p\mathrm{Var}(\psi(X)) + (1-p)\mathrm{Var}(\psi(Y))] \tag{10}$$

where $Z \sim pP_X + (1-p)P_Y$. A key challenge is that τ^2 is often not equal to the variance of the unconditional distribution without additional assumptions on P_X and P_Y.

A canonical example of this phenomenon is the permutation distribution of the difference in sample means. Given two independent iid samples $X_1, ..., X_n$ and $Y_1, ..., Y_m$, with $\text{Var}(X) = \sigma_X^2 < \infty$, $\text{Var}(Y) = \sigma_Y^2 < \infty$, and $\mathbb{E}X = \mathbb{E}Y$, the central limit theorem gives that $\sqrt{n+m}\left(\bar{X}_n - \bar{Y}_m\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{p}\sigma_X^2 + \frac{1}{1-p}\sigma_Y^2\right)$ where $p = \lim_{n\to\infty} \frac{n}{n+m}$. The conclusion of Eq. (10), however, is that the permutation distribution of the statistic $\sqrt{n+m}\left(\bar{X}_n - \bar{Y}_m\right)$ approaches a normal disribution with mean 0 and variance $\frac{1}{1-p}\sigma_X^2 + \frac{1}{p}\sigma_Y^2$ (Lehmann and Romano, 2006). Thus, unless $\sigma_X^2 = \sigma_Y^2$ or $p = \frac{1}{2}$, the permutation distribution fails to match the unconditional distribution.

The goal here is thus to provide a general result combining the delta method with the results of Chung and Romano (2013). First, we note that the finite forest centered MSE is equal to the original difference rescaled by $g'(\tilde{R}_B(\boldsymbol{x})) = g'(\mathbb{E}RF_B(\boldsymbol{x})) + o_P(1)$, so that

$$\sqrt{B}\left[MSE_{RF}(\boldsymbol{x}; y) - \mathbb{E}MSE_{RF}(\boldsymbol{x}; y)\right] = \sqrt{B}g'(\mathbb{E}RF_B(\boldsymbol{x}))\left[RF_B(\boldsymbol{x}) - \mathbb{E}RF_B(\boldsymbol{x})\right] + o_P(1)$$

and therefore the MSE at a single point satisfies Eq. (9) for

$$\begin{aligned}\psi(T(\boldsymbol{x})) &= g'(\mathbb{E}RF_B(\boldsymbol{x}))\left[T(\boldsymbol{x}) - \mathbb{E}RF_B(\boldsymbol{x})\right] \\ \psi^\pi(T^\pi(\boldsymbol{x})) &= g'(\mathbb{E}RF_B^\pi(\boldsymbol{x}))\left[T^\pi(\boldsymbol{x}) - \mathbb{E}RF_B^\pi(\boldsymbol{x})\right].\end{aligned}$$

Thus, the single point MSE satisfies the conditions needed to apply Theorem 2.1 of Chung and Romano (2013). The calculation of the permutation distribution variance follows immediately from Eq. (10); the permutation distribution of the statistic $\sqrt{2B}[MSE_{RF}(\boldsymbol{x}; y) - MSE_{RF^\pi}(\boldsymbol{x}; y)]$ converges to a normal distribution with mean 0 and variance

$$\tau^2 = \frac{1}{1/4}\left[\frac{1}{2}\text{Var}(g'(\mathbb{E}RF_B(\boldsymbol{x}))T(\boldsymbol{x})) + \frac{1}{2}\text{Var}(g'(\mathbb{E}RF_B^\pi(\boldsymbol{x}))T^\pi(\boldsymbol{x}))\right].$$

This is double the variance of Eq. (6), because the previous calculations were done for a $\sqrt{B}$ rescaling, and so the conditional and unconditional variances agree. Because the ensemble sizes used in Algorithm 1 are assumed to be the same, $p = \frac{1}{2}$, so that the permutation test for equivalence of forest predictions is automatically valid in the sense of matching the permutation and unconditional distributions. This argument is formalized in the following result.

Theorem 12 *Let $T_{1,k_n}, ..., T_{B,k_n}$ and $T^\pi_{1,k_n}, ..., T^\pi_{B,k_n}$ be two collections of trees satisfying the conditions of Lemma 1 and Lemma 3, and fix a test point with location $\boldsymbol{X}$ and response Y. Consider a test of the null hypothesis*

$$H_0 : \mathbb{E}\left[MSE_{RF}(\boldsymbol{X}; Y)\middle|\ \boldsymbol{X}, Y\right] = \mathbb{E}\left[MSE_{RF^\pi}(\boldsymbol{X}; Y)\middle|\ \boldsymbol{X}, Y\right]$$

using the statistic $\hat{\Delta} = MSE_{RF}(\boldsymbol{X}; Y) - MSE_{RF^\pi}(\boldsymbol{X}; Y)$. Then under H_0, the permutation distribution of $\sqrt{B}\hat{\Delta}$ converges to a normal distribution with mean 0 and variance

$$\tau^2 = g'(\mathbb{E}RF_B(\boldsymbol{x}))^2\sigma^2 + g'(\mathbb{E}RF_B^\pi(\boldsymbol{x}))^2\sigma_\pi^2$$

which is also the variance of the unconditional distribution of $\sqrt{B}\hat{\Delta}$, as $n \to \infty$. Thus, the permutation test attains the asymptotic Type I error rate.

Proof The only claim that remains to be verified is that the permutation test attains the Type I error rate. Let $\Phi(\cdot)$ be the standard normal cdf, and let $\hat{J}_B(t)$ be the (random) cdf of the permutation distribution, with corresponding quantile function $\hat{J}_B^{-1}(q)$. By the argument preceding the theorem statement, we have that $\sup_t |\hat{J}_B(t) - \Phi(t/\tau)| \xrightarrow{p} 0$. Then, by Lemma 11.2.1 of Lehmann and Romano (2006), for any number $q \in (0,1)$, $\hat{J}_B^{-1}(q) \xrightarrow{p} \tau\Phi^{-1}(q)$. In particular, for a given significance level α, the 1-sided permutation test of H_0 at the level α has a critical value $\hat{J}_B^{-1}(1-\alpha)$ which converges in probability to $\tau\Phi^{-1}(1-\alpha)$. Thus, as $B \to \infty$,

$$P(\sqrt{B}\hat{\Delta} \geq \hat{J}_B^{-1}(1-\alpha)|H_0) \to P(\sqrt{B}\hat{\Delta} \geq \tau\Phi^{-1}(1-\alpha)|H_0) \to \alpha.$$

■

We now must extend this result to multipoint test sets. However, Theorem 2.1 of Chung and Romano (2013) deals only with the scalar case. As such, recall that the multipoint MSE can be broken down into a sum of iid components. In particular, letting $\mathcal{T}$ be a test set consisting of N_t points, it was shown in Section 3.3 that

$$\sqrt{B}\left[MSE_{RF}(\mathcal{T}) - \mathbb{E}MSE_{RF}(\mathcal{T})\right] = \frac{1}{\sqrt{B}}\sum_{i=1}^{B}\bar{T}_i + o_P(1)$$

where $\bar{T}_i$ is an iid sequence of random variables, each with mean 0 and variance presented in Eq. (7). Thus, the scaled and centered MSE satisfies the linearity condition presented in Eq. (9). In particular, $\bar{T}_1, ..., \bar{T}_B \overset{iid}{\sim} P$ and $\bar{T}_1^\pi, ..., \bar{T}_B^\pi \overset{iid}{\sim} P^\pi$, and we are testing $H_0 : \mathbb{E}\bar{T}_i = \mathbb{E}\bar{T}_i^\pi$. Thus, because each is calculated with B trees, the same results hold and the test is asymptotically valid at multiple test points. This leads naturally to the following culminating theorem, the proof of which follows an identical argument to that of Theorem 3.

Theorem 13 *Let $T_{1,k_n}, ..., T_{B,k_n}$ and $T^\pi_{1,k_n}, ..., T^\pi_{B,k_n}$ be two collections of trees satisfying the conditions of Lemma 1 and Lemma 3, and fix a collection of test points $\mathcal{T}$. Consider a test of the null hypothesis*

$$H_0 : \mathbb{E}\left[MSE_{RF}(\mathcal{T}) \mid \mathcal{T}\right] = \mathbb{E}\left[MSE_{RF^\pi}(\mathcal{T}) \mid \mathcal{T}\right]$$

using the statistic $\hat{\Delta} = MSE_{RF}(\mathcal{T}) - MSE_{RF^\pi}(\mathcal{T})$. Then, assuming H_0, the permutation distribution of $\sqrt{B}\hat{\Delta}$ converges to a normal distribution with mean 0 and variance given by Eq. (7) which is also the variance of the unconditional distribution of $\sqrt{B}\hat{\Delta}$, as $n \to \infty$. Thus, the permutation test attains the asymptotic Type I error rate.

3.4.1 Beyond the iid Approximation

We note that the conditions of Lemma 1 are likely far stronger than needed to attain the result in Theorem 4. The proofs of validity for the permutation tests rely on projecting the random forest (which is a correlated sum $\frac{1}{B}\sum_{i=1}^{B} T_i(\boldsymbol{x})$) onto a sum of iid random variables, $\sum_{i=1}^{n}\psi_n(Z_i)$ for some function ψ_n, to which a central limit theorem can then

apply. Indeed, this is exactly the approach of the Hájek projection and H-decomposition used respectively by Mentch and Hooker (2016) and Wager and Athey (2018). In these works, roughly speaking, it is shown that under constraints on the forest construction, the random forest prediction at a point $\boldsymbol{x}$ satisfies

$$\frac{1}{\sqrt{B}}\sum_{i=1}^{B}[T_i(\boldsymbol{x}) - \mathbb{E}RF_B(\boldsymbol{x})] = \sum_{i=1}^{n}\psi_n(Z_i) + o_P(1).$$

For example, if the Hájek projection is used, $\psi_n(Z_i) = \sqrt{B}\mathbb{E}\left[RF_B(\boldsymbol{x}) \mid Z_i\right] - \mathbb{E}RF_B(\boldsymbol{x})$. Moreover, as mentioned in the remark following Lemma 3, the fact that the MSE is asymptotically linear is independent of the iid approximation, and thus the MSE for these forests is also asymptotically linear.

4. Simulations

We now apply our testing procedure in a number of settings with varying regression functions and covariate structures. We simulate data from four models summarized in Table 1, with covariate structures 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.

4.1 Power and Error Control

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	$P(Y = 1\|\boldsymbol{X}) = \text{expit}\left[\beta\sum_{j=2}^{5} X_j\right]$	M2
4	$Y = RF_{\texttt{eBird}}(\boldsymbol{X}) + \epsilon$	`eBird`

Table 1: Distributions of $Y|\boldsymbol{X}$ for each model. $\text{expit}(z) = \frac{1}{1+e^z}$.

Model #	Covariate Structure
M1	$X_1, ..., X_5 \overset{iid}{\sim} Unif(0,1)$, $X_6, ..., X_{10} \overset{iid}{\sim}$ Multinomial$(1, [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T)$
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.

Jordon et al. (2018) Romano et al. (2020) 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. Here we test the importance of (X_1, X_6, X_2, X_7) where X_1, X_6 are important, X_1, X_2 are continuous, and X_6, X_7 are categorical. Model 2 resembles the MARS data generating model (Friedman, 1991) commonly used in random forest studies, but with a modification to include an important discrete covariate. In both settings, we draw $n = 2000$ points from the joint distribution of $(\boldsymbol{X}, Y)$,

subsample sizes of $k_n = n^{0.6} \approx 95$, and build $B = 125$ 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 Eq. (2), is conditional on the test points used. These simulations change the null hypothesis each time, because the validation set changes. Thus, the simulations mimic the common practice of random splitting the data into a training and validation fold.

For Models 1 and 2, we focus on a marginal signal to noise ratio, which is controlled by the parameters β and σ. We fix $\beta = 10$ across all simulations let $\sigma = 10/j$ where j takes 9 equally spaced values between 0.005 and 2.25 so that for small k, the signal to noise ratio (SNR) is small. The results are shown in Fig. 1. We see that the test maintains the nominal type I error rate and attains high power for marginal SNRs near 1 for all variables except X_7 in Model 2. Note also that the type I error rate appears insensitive to the covariate structure. In the MARS model, we see that the test has more power against X_3 than X_7, because X_7 is only important insofar as it interacts with X_1.

Model 3 is an adaptation of the model used in Candes et al. (2016) for high-dimensional correlated data. Here 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 largely insensitive to the dimension of the ambient feature space, and instead sensitive only to the "strong" feature space. This setting helps to explore the utility of our method in the high dimensional sparse signal 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 $k_n = n^{0.6} \approx 46$, $B = 125$, $N_t = 100$, and vary the β coefficient according to 8 equally spaced values between 0.01 and 2.5 and also for 7 equally spaced values between 5 and 20. The results are shown in the bottom panel of Fig. 1. Note that the 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 only changes $P(Y = 1|\boldsymbol{X})$ slightly due to the rapidly decaying first derivative of the expit(z) function.

Finally, we turn to Model 4 where the true data generating model is a random forest. We utilize a dataset from Coleman et al. (2017) describing the occurrence of tree swallows and 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 add Gaussian noise. 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 (not used to train $RF_{\texttt{eBird}}$), defined as $\texttt{eff.hours.proxy} = \frac{\texttt{eff.hours}+Z_{0.5}}{\sqrt{\text{Var}(\texttt{eff.hours})+0.5}}$ and $\texttt{dfs.proxy} = \frac{\texttt{dfs}+Z_{0.025}}{\sqrt{\text{Var}(\texttt{dfs})+0.025}}$ 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.

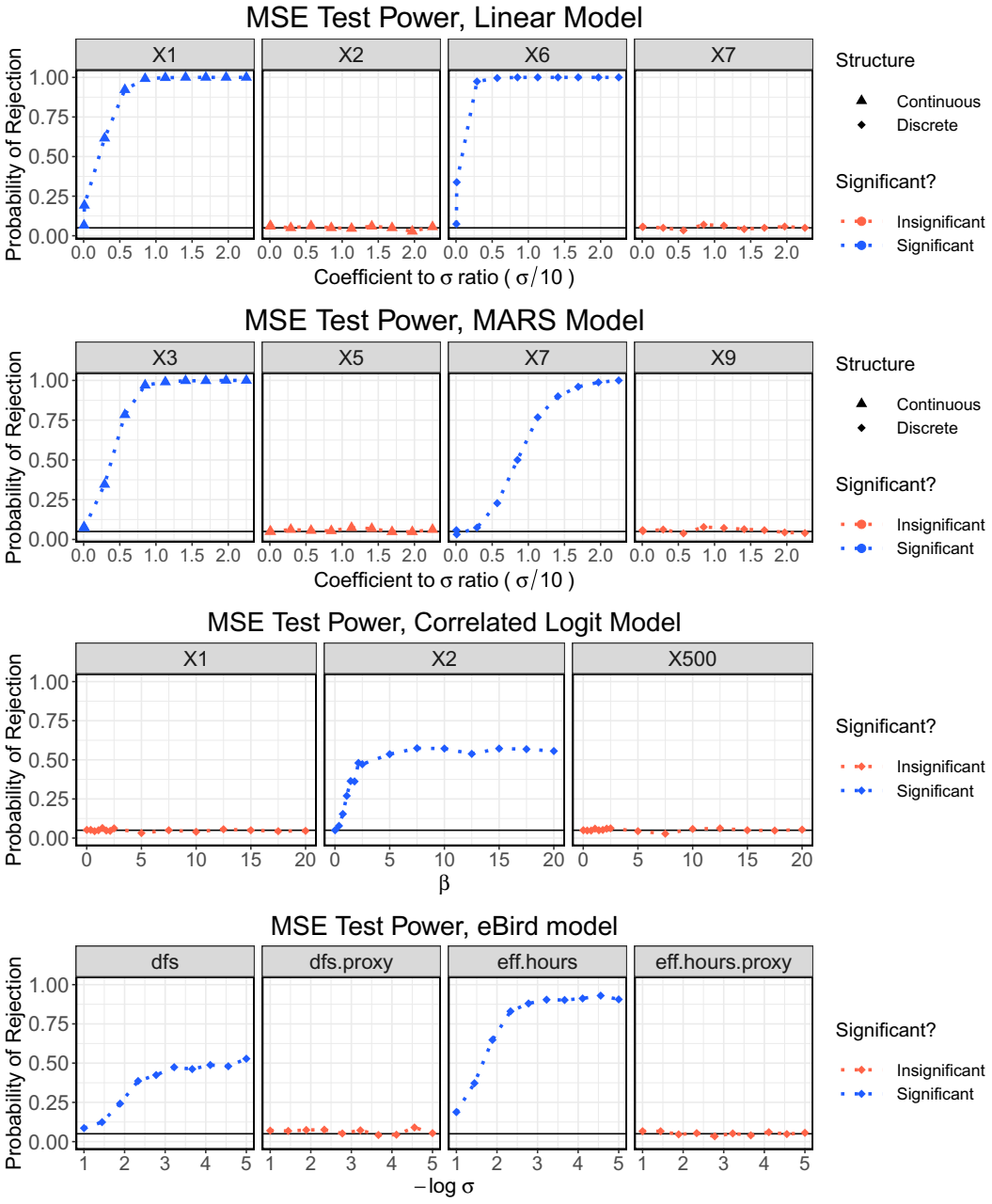


Figure 1: Simulation results for each of the models from Table 1. Black line corresponds to $\alpha = 0.05$, the nominal level

In Model 4, we let $n = 2000$, $k_n = n^{0.6}$, $B = 125$, $N_t = 100$, and let $\sigma = e^{-j}$ for 10 values of j equally spaced between 1 and 5. The results of this simulation are show in Fig. 1. We see that again the test maintains the nominal type I error rate with modest power for signal variables. Moreover, the procedure correctly identifies the true variables as important over their highly correlated proxies.

4.2 Formal Comparison with Knock-offs

In this section, we formally compare the proposed procedure with several implementations of the knockoff framework (Barber et al., 2015; Candes et al., 2016), an exciting new method for statistically valid variable selection in a model-free way. As noted in Section 2, the null hypothesis tested by knockoffs is slightly different from that in our procedure. To a practitioner, however, the procedures would likely be used in a similar way, and as such we leave this subtlety out of our subsequent discussion. In general, we conclude that our method is largely complementary to knockoffs.

A key assumption of the knockoff framework is that the distribution of the covariates X is known (also referred to as the model-X assumption) , which, crucially, our method does not require. Candes et al. (2016) proposes a second order method for generating knockoffs via a Gaussian analogue for X (i.e. a Gaussian random vector with the same covariance and mean as X). As of now, it is unclear how well knockoffs perform, both in terms of power and Type I Error control, when an approximation is employed. Finally, our method is designed to be powerful in situations where the response has a complex relationship with the data. To tackle these diverse scenarios, we use the following simulation set-up, with 4 different pairings:

- We fix $p = 25$, and generate covariates according to the following data distributions, one where the model-X assumption is satisfied, and one where it is not:

  Gaussian $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{ij} = \rho^{|i-j|}$ and we choose $\rho = 0.25$.

  Fish Toxicity We simulate X from the UCI fish toxicity data set provided by Cassotti et al. (2015), which comes with $n = 908$ observations on 6 covariates with information regarding chemicals that are believed to be toxic to a species of minnow. These covariates are quite non-Gaussian. To fill in the remaining 19 covariates, we randomly sample 19 columns (with replacement) from the original 6, and then sample rows of those 19 columns from the original data, so that there is no replication between the original 6 columns and the synthetic 19. X is also scaled and centered, to account for differing units.

- Our responses are generated according to the following two regression functions. In both cases, $\epsilon \sim \mathcal{N}(0, 1/\text{SNR})$.

  Linear $Y = \sum_{j=1}^{s} X_j + \epsilon$

  Flattened Sine $Y = \frac{1}{\sqrt{\sum_{j=1}^{s} X_j^2}} \sin\left(\pi \sqrt{\sum_{j=1}^{s} X_j^2}\right) + \epsilon$. Note that in this set up, each variable has little linear effect but quite a strong nonlinear joint effect.

- For the responses, we vary the parameters s and SNR, to respectively control the density of the model (in terms of the number of important features) and the strength of the signal present in the data.

Each set up is evaluated at 6 different values of s, spaced evenly between 2 and 25, for a very sparse to a dense model, and 10 different signal to noise ratios, spaced evenly from 0.5 to 5, for a total of 60 simulation pairs. We apply the knockoff filter with the both standard lasso coefficient difference statistic and random forest out-of-bag importance statistic, to use both a linear and nonlinear statistic. We apply our procedure and the two knockoff approaches to 100 repetitions of $n = 908$ observations from the 4 different model set ups listed above. For our procedure, we build 125 trees, holdout 90 observations at random for testing, and take subsamples of size $k = \sqrt{908} \approx 30$. All tests here are conducted with respect to evaluating the marginal importance of X_1, which is important (in the sense of conditional independence) in each scenario. We define the power of the knockoff procedure to be $\frac{1}{100}\sum_{l=1}^{100} I(X_1 \in \mathcal{S}_l)$, where $\mathcal{S}_l$ is the selection of variables produced by the knockoff filter.

The results are plotted in Fig. 2. Several patterns are shared across the plots. First, the proportion of important variables appears to be more important for attaining good power than the SNR in both our method and the knockoff procedure. However, the directionality is inverted - our procedure performs much better in sparse models, while knockoffs seem to require a dense model to select any variables. Next, the Gaussian flattened sine presents a challenge for both procedures, but our procedure is able to attain good power in all other scenarios, while knockoffs really only succeed (with either statistic) in the linear Gaussian case. While throughout these simulations the FDR is controlled at the nominal level, a steep price is paid in terms of power for losing the knowledge of the distribution of X. Both knockoff statistics exhibit almost identical performance, which suggests further that the oob importance measures are unlikely to be useful as a nonlinear test statistics.

We conclude that knockoffs are a powerful method when there are many covariates suspected to be important to the response. In these cases, the knockoff procedure can efficiently identify a dense model. However, the overwhelming dependence on s and the model-X assumption being satisfied suggests the need for more direct alternatives like that proposed here. Our procedure exhibits qualitatively similar behavior in 3/4 of the set ups, attaining good power even for signal to noise ratios below 1 in the sparse model. Knockoffs maintain a computational edge over our method, needing only a single model fit to provide FDR controlled variable selection, while a naive implementation of our method would require $2p$ model fits, followed by a FDR filter such as Benjamini-Hochberg that accepts p-values (Benjamini and Hochberg, 1995).

5. Applications to Ecological Data

We now apply our 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 construct a simulated random forest model. Here we utilize the original data as considered in Coleman

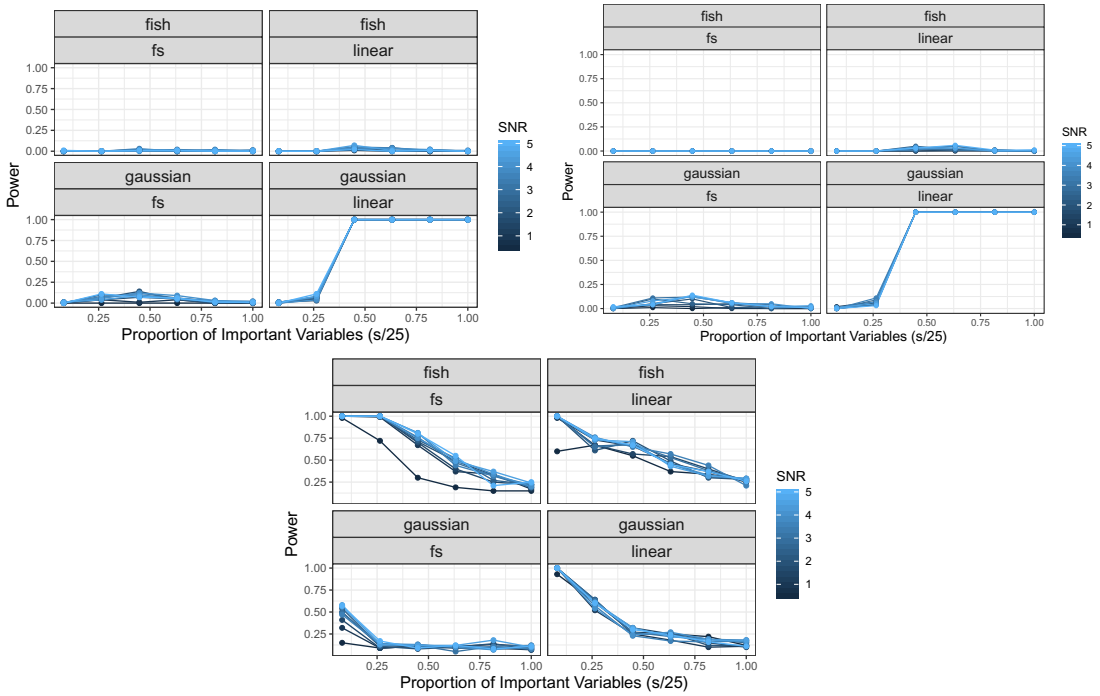
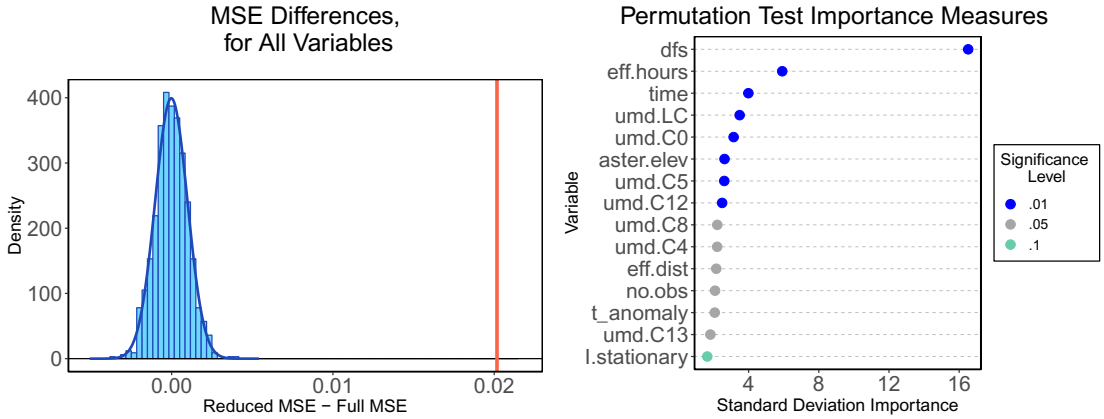


Figure 2: Simulation results for the knockoff comparison, showing the associated power curves, calculated with respect to a nominal Type I error rate of $\alpha = 0.10$. The knockoff procedure is run with the FDR threshold set to α. Light shades of blue indicate more powerful signal. `fs` refers to the flattened sine model. Bottom: our procedure. Top left: Knockoffs with the lasso statistic. Top right: Knockoffs with the random forest out-of-bag importance statistic.

et al. (2017). The standard task is to predict tree swallow `occurrence` during the fall migration season in a particular geographic area referred to as Bird Conservation Region (BCR) 30. eBird 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) 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 on the left hand side of Fig. 3. There is clear evidence



(a) MSE Test for all covariates

(b) Marginal Importance Scores

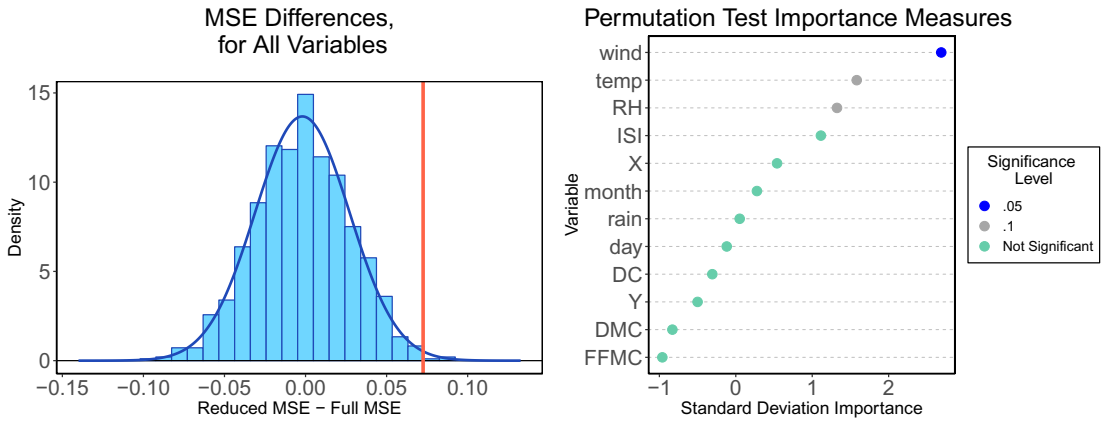
Figure 3: Results on the eBird data from (Sullivan et al., 2009, 2014). Red line indicates observed value, and histograms of differences in MSE after permutation are overlayed by an estimated normal density.

for signal in the data, with an overall p-value of $p < 0.0001$. Next, to produce an output similar to the out-of-bag importance scores traditionally computed, we repeat the testing procedure for each covariate individually, recording the marginal importance for each as the number of standard deviations away that the original MSE difference is from the center of the permutation distribution. The results are shown in the right hand side of Fig. 3. 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, serving 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: Cortez and Morais (2007) sought to predict $\log(1 + \texttt{area})$ burned by several fires in northern Portugal using covariate information 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 Fig. 4. The overall test suggests that there is signal in the data ($p = 0.0040$), albeit a weaker effect than in the preceding eBird case study. The importance procedure suggests that only `wind` – the wind speed at the location of the fire – is significant at the 0.05 level.

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,



(a) MSE Test for all covariates

(b) Marginal Importance Scores

Figure 4: 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.

is both computationally efficient and statistically valid, placing hypothesis tests with random forests firmly within the grasp of applied researchers. Previously suggested parametric approaches are computationally prohibitive and place severe restrictions on 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 practitioners. Consider, for example, the mediator detection problem arising frequently in medical studies wherein a covariate X_1 is a *mediator* for another covariate X_2 whenever the effect of X_2 on the response is nullified (or substantially lessened) by including X_1 in the model. The same two-step process often employed with linear models can be carried out with random forests using the tests developed here: 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. Moreover, our procedure attains good power in a wide variety of model set ups, and as such is likely usable off-the-shelf by practitioners interested in the nonlinear regression inference problem.

The primary goal of this work is to identify covariates that produce statistically significant improvements in model accuracy. To assess this, we considered 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 rendered independent of Y, conditional on the rest of the features. This muting of $\boldsymbol{X}_S$ can be achieved in various ways:

- Outright exclusion: $\boldsymbol{X}_S$ is 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.

The manner in which covariates are randomized or muted will subtly alter the underlying null hypotheses in Eq. (2), but crucially, the Type I error control of our procedure holds for each of these null hypotheses. Indeed, it is possible to reject Eq. (2) because of artifacts in the covariate distribution, rather than a notion of conditional independence. Assuming $\boldsymbol{X}_\mathcal{S}$ ind $Y|\boldsymbol{X}_{-\mathcal{S}}$, we would expect predictions from trees trained on $\mathcal{D}_n$ to have the same distribution as those generated from trees trained on $\mathcal{D}_n^\pi$. In this case, a rejection of the null hypothesis of equal MSE's suggests that $\boldsymbol{X}_\mathcal{S}$ and Y are not conditionally independent. This is the case if the distribution of $\boldsymbol{X}$ is known (or can be estimated easily), so that a knock-off version of $\boldsymbol{X}_\mathcal{S}$ can be employed. However, practically speaking, our method provides valid model based inference even without any knowledge of the covariate distribution. In such cases, formally investigating whether particular covariates significantly improve predictive accuracy beyond permuted analogues, for example, can still provide valuable insight into their relative value and utility.

As mentioned earlier, a 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. 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 often far more computationally efficient than carrying out even a single parametric test. Indeed, while knockoffs maintain a computational edge in these cases, they still require the unwieldy model-X assumption, and require a large number of important features to be present in the data.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637).

A. Proofs of Technical Results

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*

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

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.

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

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

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

Next, by assumption,

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

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

After Lemma 1, we next need to prove Lemma 3, whose purpose is to show that the observed MSE is asymptotically centered around its own expectation.

Lemma 3 *Assume the conditions needed from Corollary 2. Additionally, assume that g has at least k derivatives for some $k\geq 3$, and that $g^{(k)}(x)<\infty$ for all x. Further, assume that $\mathbb{E}|T_i(\boldsymbol{x})|^k<\infty$. Then,*

$$\sqrt{B}\left[\mathbb{E}g(RF_B(\boldsymbol{x})-g(\mathbb{E}RF_B(\boldsymbol{x}))\right]=\frac{g''(\mathbb{E}RF_B(\boldsymbol{x}))\sigma^2}{2\sqrt{B}}+o(B^{-3/2}).$$

Proof We rely on a result presented in Oehlert (1992), which states that under the conditions presented in the lemma statement,

$$\mathbb{E}g(RF_B(\boldsymbol{x}))=g(\mathbb{E}RF_B(\boldsymbol{x}))+\frac{g''(\mathbb{E}RF_B(\boldsymbol{x}))\sigma^2}{2B}+o(B^{-2}).\tag{11}$$

Thus, the result follows from multiplying both sides of Eq. (11) by $\sqrt{B}$ and rearranging terms. ■

Next, we move on to the proof of Proposition 1 from Section 3.3.1, which gives that the trees typically utilized in a random forest obey the necessary regularity conditions for Corollary 2.

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* $P(\textit{variable } j \textit{ is split on}) > \gamma$ *for* $j \in \{1, ..., p\}$

(ii) *Each split leaves at least a constant proportion of observations in the original 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

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

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

B. Additional Simulations

We include some additional simulations here to demonstrate the following points.

1. The accuracy of the permutation distribution approximation of the Gaussian. The theory outlined in Section 3 establishes that the difference in MSEs between forests is asymptotically Gaussian but the difficulty in estimating the resulting variance largely restricts its direct usage in practical settings. We go on to demonstrate that the permutation distribution approaches this distribution, thereby circumventing the need for a direct variance estimate. The simulations below present empirical evidence that this approximation is reasonable in practical settings.

2. The instability of the variance estimation procedures laid out in Wager et al. (2014) and 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 (2018)), 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.

3. 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.

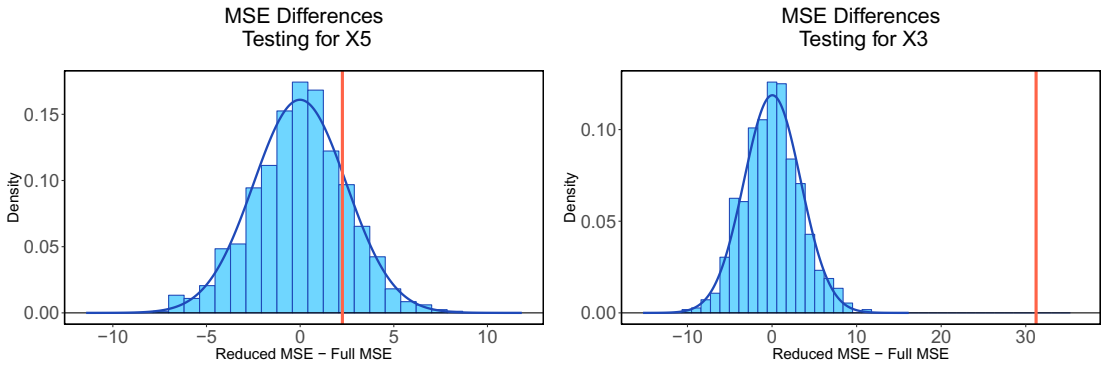
B.1 Normality of Permutation Distributions

Here we provide a concise simulation demonstrating the accuracy of the permutation distribution approximation of the Gaussian in a practical setting. We simulate $n = 2000$ training observations from Model 2 with covariate structure M1 as described in Section 4. Specifically, we consider the model $Y = \beta \sin(\pi I(X_7 = 2)X_1) + 2\beta(X_3 - .05)^2 + \beta X_4 + \beta X_2 + \epsilon$ where we sample covariates according to $X_1, ..., X_5 \overset{iid}{\sim} Unif(0, 1)$ and $X_6, ..., X_{10} \overset{iid}{\sim}$ Multionimial$(1, [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T)$. Here we use $\beta = 10, \sigma = 10$, 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 `mtry` $= 3$. The resulting permutation distributions are shown in Fig. 5.

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.

B.2 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



(a) Under H_0 (b) Under H_1

Figure 5: Permutation distributions of Δ_B. Red line indicates observed value, and histograms are overlayed by an estimated normal density.

$X_6 = ... = X_{10} = 2$. 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 Fig. 6.

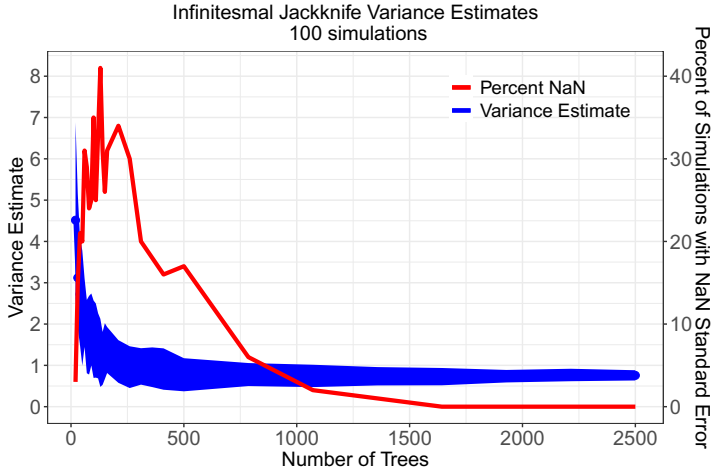


Figure 6: `ranger` IJ variance estimate. Blue ribbon plot indicates central 90% of variance estimates (corresponds to left axis), and red line (corresponds to right axis) represents percentage of runs that return `NaN`.

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.3 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 the second panel of Fig. 1. 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 Fig. 7.

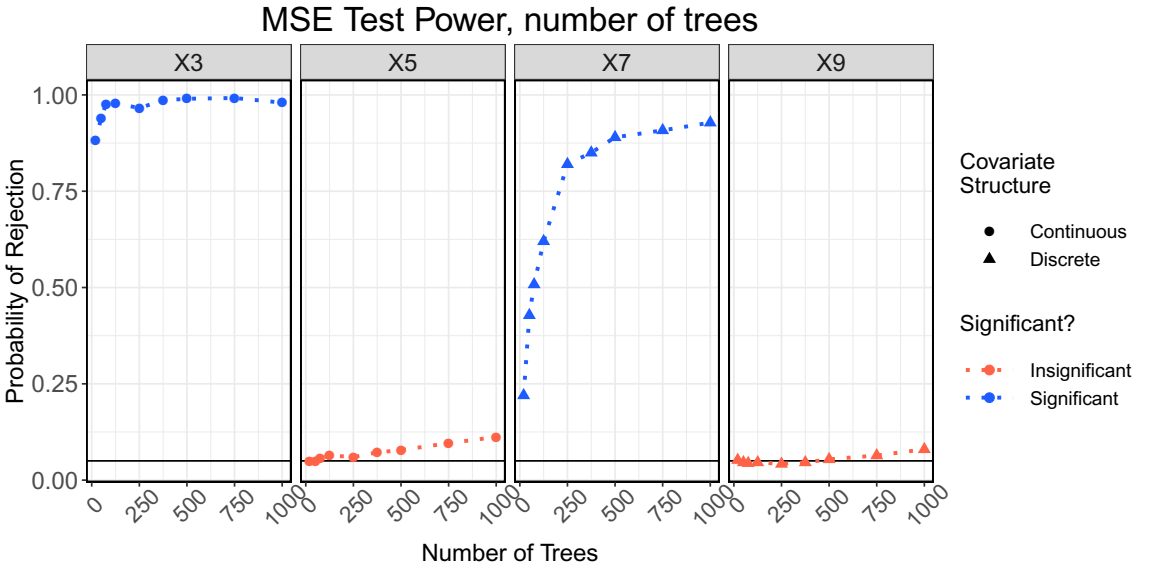


Figure 7: Model 2 power curves for 500 simulations, by number of trees. The Y-axis represents $P(\tilde{p} \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

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, 2018) weakens these conditions, but places explicit 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 Fig. 7, but now fix $B = 125$ and let $k_n = n^p$, and we vary p at 10 equally spaced values between 0.1 and 0.99.

The resulting simulation is shown in Fig. 8. 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.

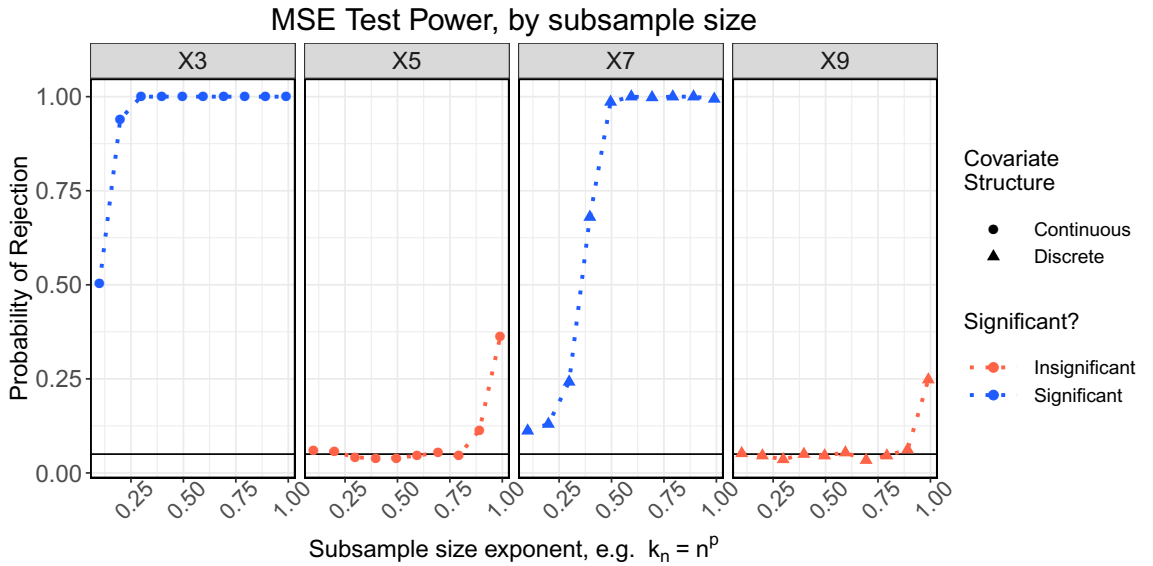


Figure 8: Model 2 power curves for 500 simulations, by subsample exponent. The Y-axis represents $P(\tilde{p} \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

References

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *arXiv preprint arXiv:1610.01271*, 2016.

Rina Foygel Barber, Emmanuel J Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

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

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

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

M Cassotti, D Ballabio, R Todeschini, and V Consonni. A similarity-based qsar model for predicting acute toxicity towards the fathead minnow (pimephales promelas). *SAR and QSAR in Environmental Research*, 26(3):217–243, 2015.

EunYi Chung and Joseph P Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, pages 484–507, 2013.

Tim Coleman, Lucas Mentch, Daniel Fink, Frank La Sorte, Giles Hooker, Wesley Hochachka, and David Winkler. Statistical inference on tree swallow migrations. *arXiv preprint arXiv:1710.09793*, 2017.

Paulo Cortez and Aníbal de Jesus Raimundo Morais. A data mining approach to predict forest fires using meteorological data. 2007.

Yifan Cui, Ruoqing Zhu, Mai Zhou, and Michael Kosorok. Some asymptotic results of survival tree and forest models. *arXiv preprint arXiv:1707.09631*, 2017.

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

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

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

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

Giles Hooker and Lucas Mentch. Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

Hemant Ishwaran and Min Lu. Random survival forests. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2008.

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

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*, 2018.

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

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

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

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

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

Lucas Mentch and Giles Hooker. Formal hypothesis tests for additive structure in random forests. *Journal of Computational and Graphical Statistics*, pages 1–9, 2017.

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

Gary W Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

Fortunato Pesarin and Luigi Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

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

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

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

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

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

Brian L Sullivan, Jocelyn L Aycrigg, Jessie H Barry, Rick E Bonney, Nicholas Bruns, Caren B Cooper, Theo Damoulas, André A Dhondt, Tom Dietterich, Andrew Farnsworth, et al. The ebird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014.

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

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

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

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

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

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