

Formal Hypothesis Tests for Additive Structure in Random Forests

Lucas Mentch*

and

Giles Hooker

Department of Statistical Science
Cornell University

September 28, 2016

Abstract

While statistical learning methods have proved powerful tools for predictive modeling, the black-box nature of the models they produce can severely limit their interpretability and the ability to conduct formal inference. However, the natural structure of ensemble learners like bagged trees and random forests has been shown to admit desirable asymptotic properties when base learners are built with proper subsamples. In this work, we demonstrate that by defining an appropriate grid structure on the covariate space, we may carry out formal hypothesis tests for both variable importance and underlying additive model structure. To our knowledge, these tests represent the first statistical tools for investigating the underlying regression structure in a context such as random forests. We develop notions of total and partial additivity and further demonstrate that testing can be carried out at no additional computational cost by estimating the variance within the process of constructing the ensemble. Furthermore, we propose a novel extension of these testing procedures utilizing random projections in order to allow for computationally efficient testing procedures that retain high power even when the grid size is much larger than that of the training set.

*We would like to thank Cornell University's Lab of Ornithology for providing interesting data. Giles Hooker was partially supported by NSF grants DMS-1053252 and DEB-1353039 and NIH grant R03DA036683.

1 Introduction

As scientific data grows larger and becomes easier to collect, traditional statistical models often prove insufficient for fully capturing the underlying process. Learning algorithms, on the other hand, adapt well to a variety of data types and produce accurate predictions, but their inherent complexity and black-box nature makes addressing even the simplest scientific questions significantly more difficult. This work provides a formal statistical test for determining variable interactions whenever ensemble learning methods like random forests are used as the primary modeling tool.

Additive models were suggested by Friedman and Stuetzle (1981) and further developed and made popular by Stone (1985) and Hastie and Tibshirani (1990). An underlying regression function $F: \mathcal{X} \mapsto \mathbb{R}$ is said to be additive if

$$F(x_1, \dots, x_d) = \sum_{i=1}^d F_i(x_i)$$

for some functions F_1, \dots, F_d . If the regression function cannot be written as, or at least well-approximated by, a sum of univariate functions, then an interaction exists between some subset of the covariates. Many methods have been developed to estimate the additive functions F_1, \dots, F_d including a method based on marginal integration by Linton (1995), a wavelet method suggested by Amato and Antoniadis (2001), a tree-based method by Lou et al. (2013), and the most popular class based on backfitting algorithms as found in Buja et al. (1989), Opsomer and Ruppert (1998, 1999), and Mammen et al. (1999).

The popularity of additive models and their ease of interpretation has inspired hypothesis tests to assess whether observed data should be modeled in an additive fashion. Versions of these lack-of-fit tests have been proposed by Barry (1993), Eubank et al. (1995), Dette and Derbort (2001), Derbort et al. (2002), and De Canditiis and Sapatinas (2004). Fan and Jiang (2005) further extend these procedures to also evaluate whether the additive components belong to a particular parametric class. Even when additive models are not used as the primary analytical tool, scientists often utilize these and related interaction detection methods to determine which variables contribute additively to the response; when no interactions are detected, the levels of one feature may be changed without affecting the

contribution to the response made by the others.

Their utility notwithstanding, additive models can often fail to fully capture the signal hidden within modern complex data, even when relatively little signal results from variable interactions. On the other hand, learning algorithms like bagged trees and random forests introduced by Breiman (1996, 2001), are robust to a variety of regression functions and are considered something of a gold standard in terms of predictive accuracy. Though this accuracy continues to drive their popularity, little is understood about the underlying mathematical and statistical properties of these ensemble methods. Thus, while practitioners routinely rely on such methods to make predictions, when standard results such as confidence intervals or p-values from hypothesis tests for variable importance or interactions need reported, those practitioners are forced to move to an entirely different modeling technique and rely on more well-established procedures. At best, the ensembles might be used to better inform which hypotheses to test and/or which variables should be included in a simpler model.

Recently however, important progress has been made in understanding the asymptotic properties of these ensemble methods by considering a subsampling approach in lieu of the traditional bootstrapping procedure. Mentch and Hooker (2016) show that when proper subsamples are used to construct individual trees, the ensemble predictions can be seen as extensions of classical U-statistics and as such, are asymptotically normal. Wager et al. (2014) apply recent results on the infinitesimal jackknife (Efron, 2014) to produce estimates of standard errors for subsampled random forest predictions and Wager and Athey (2015) later demonstrate the consistency of such an approach. Most recently, Scornet et al. (2015) provided the first consistency results for Breiman’s original random forest procedure when subsampling is employed and the underlying regression function has an additive form.

This paper continues in this recent trend by developing formal hypothesis tests for additivity in ensemble learners like bagged trees and random forests. These tests allow practitioners to formally investigate the manner in which features contribute to the response when simpler, more direct tools are insufficient and to our knowledge, represent the first formal procedures for investigating the structure of the underlying regression function within the context of ensemble learning. That is, statistically valid results such as p-values

may be gathered directly from the ensemble instead of relying on *ad hoc* measures or appealing to a simplified model. In Section 2 we propose a formal test for feature significance by imposing a grid structure on the covariate space and in Section 3 we demonstrate that this additional structure further allows for tests of additivity. In Section 4 we incorporate random projections to extend our procedure to the situation where a large test grid is needed, so as to accommodate potential high dimensional settings. Finally, in Sections 5 and 6, we provide simulations to investigate the power of our hypothesis tests and apply our testing procedures to an ecological dataset.

2 Hypothesis tests for feature significance

Recent theory has demonstrated that a subsampling approach to constructing supervised ensembles like random forests may allow these learners to be reigned in within the realm of traditional statistical inference. Specifically, Mentch and Hooker (2016) show that by controlling the subsample growth rate, individual predictions are asymptotically normal thereby paving the way for a formal method of evaluating variable (feature) significance. As a simple example, consider a setting with just two features X_1 and X_2 where the response observed according to $Y = F(X_1, X_2) + \epsilon$. To test the significance of X_2 , we can generate a test set \mathbf{x}_{TEST} consisting of N points and build two subsampled ensembles \hat{F} and \hat{F}_1 . Both ensembles employ the same subsamples, but \hat{F} is constructed using both X_1 and X_2 whereas \hat{F}_1 is built using only X_1 . Predictions at each point in \mathbf{x}_{TEST} are then made with each ensemble and Mentch and Hooker (2016) show that the vector of differences in predictions has a multivariate normal limiting distribution with mean μ and variance Σ . Given consistent estimators of these parameters, $\hat{\mu}^T \hat{\Sigma}^{-1} \hat{\mu} \sim \chi_N^2$ can be used as a test statistic to formally evaluate the hypotheses

$$H_0 : F(x_1, x_2) = F_1(x_1) \quad \forall (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \tag{1}$$

$$H_1 : F(x_1, x_2) \neq F_1(x_1) \quad \text{for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \quad \text{for any } F_1.$$

Though asymptotically valid, this procedure requires building separate ensembles for each feature of interest. We demonstrate here that imposing additional structure on the test

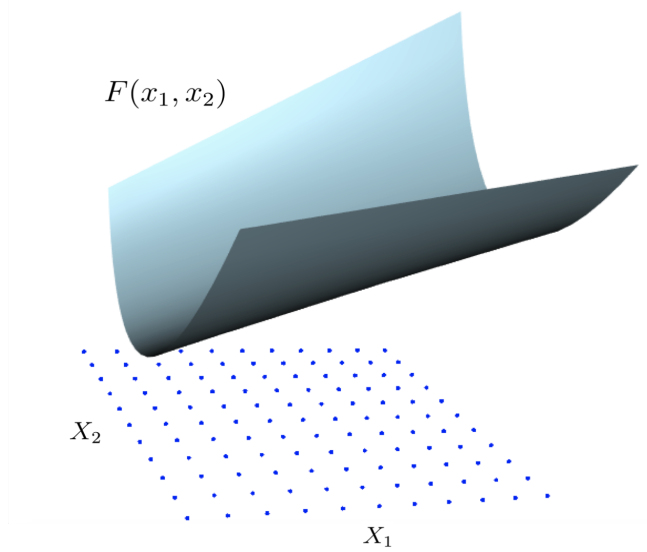


Figure 1: A grid of test points shown in the X_1X_2 plane below the response surface.

set allows us to both avoid training an additional set of trees and also perform tests for additivity.

Define a grid consisting of N total test points as in Figure 1 with N_1 levels x_{1_i} and N_2 levels x_{2_j} so that the $(i, j)^{th}$ point in the grid has true value $F_{ij} = F(x_{1_i}, x_{2_j})$ and predicted value \hat{F}_{ij} . In the case of categorical covariates, these grid levels are naturally occurring while in the case of continuous covariates, these levels can be specified as appropriate (e.g. based on quantiles of the observed data). Let V_F and $V_{\hat{F}}$ represent the vectorized versions of these true and predicted values so that $V_F = (F_{1,1}, \dots, F_{1,N_2}, \dots, F_{N_1,1}, \dots, F_{N_1,N_2})^T$ and define

$$\hat{f}_{i\cdot} = \frac{1}{N_2} \sum_{j=1}^{N_2} \hat{F}_{ij}$$

as the average response at the i^{th} level x_{1_i} across all grid levels x_{2_j} . For each point in the grid, the difference in predictions $\hat{F}_{ij} - \hat{f}_{i\cdot}$ can be written in vectorized form as $DV_{\hat{F}}$ for an $N \times N$ difference matrix D of rank $N - N_1$. In this case, $D = I_N - \left(I_{N_1} \otimes \frac{1}{N_2} \mathbf{1}_{N_2 \times N_2} \right)$ where I_C is the $C \times C$ identity matrix, $\mathbf{1}_{C \times C}$ is the $C \times C$ matrix of 1's, and \otimes denotes the standard tensor product. Let Σ denote the covariance of V_F and $\hat{\Sigma}$ a consistent covariance estimate of the predictions. Then we can define $\Sigma_D = cov(DV_F) = D\Sigma D^T$ so that $\hat{\Sigma}_D = D\hat{\Sigma}D^T$ forms a consistent estimate of the covariance of the projected predictions Σ_D . Then

$(DV_{\hat{F}})^T \hat{\Sigma}_D^{-1} DV_{\hat{F}} \sim \chi_{N-N_1}^2$ and since we can equivalently write the hypotheses in (1) as

$$\begin{aligned} H_0 : F_{ij} - f_i &= 0 \quad \forall (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \\ H_1 : F_{ij} - f_i &\neq 0 \quad \text{for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \end{aligned}$$

$(DV_{\hat{F}})^T \hat{\Sigma}_D^{-1} DV_{\hat{F}}$ can be used as a test statistic.

Asymptotically, this test statistic has a $\chi_{N-N_1}^2$ distribution and thus can be compared to the $1 - \alpha$ quantile to achieve a test with type 1 error rate α ; if the test statistic is larger than this critical value, we reject the null hypothesis and conclude that X_2 is significant.

This testing procedure readily extends to the more general case of d features X_1, \dots, X_d . Let \mathbf{X}_R and \mathbf{X}_A form a partition of $\{X_1, \dots, X_d\}$ so that \mathbf{X}_R and \mathbf{X}_A are disjoint and $\mathbf{X}_R \cup \mathbf{X}_A = \{X_1, \dots, X_d\}$; the set \mathbf{X}_R denotes the *reduced* set of features and \mathbf{X}_A represents the *additional* features that we want to test for significance. To test the hypotheses

$$\begin{aligned} H_0 : F(\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) &= F_R(\mathbf{x}_{R_i}) \quad \forall (\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) \in \mathbf{x}_{\text{TEST}} \\ H_1 : F(\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) &\neq F_R(\mathbf{x}_{R_i}) \quad \text{for some } (\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_R \end{aligned}$$

we simply repeat the testing procedure in the above example, replacing the levels x_{1_i} and x_{2_j} with appropriately redefined grid levels the feature sets \mathbf{X}_R and \mathbf{X}_A , respectively. Note that in this case, each grid point now corresponds to the value of a vector of features.

It is also worth noting that Mentch and Hooker (2016) suggest comparing predictions generated with the full training set to not only those produced with the reduced set \mathbf{X}_R , but also to those generated with \mathbf{X}_R and a permuted version of \mathbf{X}_A in order to rule out the possibility that the ensemble is simply making use of additional noise. The procedure we propose above avoids this potential confusion by utilizing the projection matrix D .

3 Tests for additivity

We now demonstrate that this grid structure also allows for formal tests of additivity.

Tests for total additivity

Again assume that our training set consists of only two features and that the response is observed according to $Y = F(X_1, X_2) + \epsilon$. Tests for *total* additivity assess whether the *entire* underlying regression function F is equal to, or at least well-approximated by, a sum of functions with disjoint domains. When each function is univariate, this simply means that there are no interactions between any covariates but a more general case is also discussed below. In the simple 2-dimensional case, the hypotheses of interest are

$$\begin{aligned} H_0 : & \exists F_1, F_2 \text{ such that } F(x_1, x_2) = F_1(x_1) + F_2(x_2) \quad \forall (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \\ H_1 : & F(x_1, x_2) \neq F_1(x_1) + F_2(x_2) \text{ for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, F_2. \end{aligned} \quad (2)$$

Again define a 2-dimensional grid of test points as in Figure 1 so that each point in the grid has true value F_{ij} , predicted value \hat{F}_{ij} , and vectorized versions V_F and $V_{\hat{F}}$. Define \bar{F} to be the mean of all predictions in the grid and define

$$\hat{f}_{i\cdot} = \frac{1}{N_2} \sum_{j=1}^{N_2} \hat{F}_{ij} \quad \text{and} \quad \hat{f}_{\cdot j} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{F}_{ij}$$

as the mean prediction at the i^{th} level x_{1_i} across all levels x_{2_j} , and the mean prediction at the j^{th} level x_{2_j} across all levels x_{1_i} , respectively. If the features are additive, (i.e. under the null hypothesis) all points (x_{1_i}, x_{2_j}) in the grid can be written as $F_{ij} = f_{i\cdot} + f_{\cdot j} - \mu$ where $\mu = \mathbb{E}\bar{F}$ is the true mean expected prediction across all points in the grid. Thus, we may equivalently write the hypotheses in (2) as

$$\begin{aligned} H_0 : & F_{ij} - f_{i\cdot} - f_{\cdot j} + \mu = 0 \text{ for all } (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \\ H_1 : & F_{ij} - f_{i\cdot} - f_{\cdot j} + \mu \neq 0 \text{ for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}}. \end{aligned}$$

The natural test statistic is then $\hat{F}_{ij} - \hat{f}_{i\cdot} - \hat{f}_{\cdot j} + \bar{F}$ which can be written as $D_2 V_{\hat{F}}$ where difference matrix is given by

$$D_2 = I_N - \left(I_{N_1} \otimes \frac{1}{N_2} \mathbf{1}_{N_2 \times N_2} \right) - \left(\mathbf{1}_{N_1 \times N_1} \otimes \frac{1}{N_2} I_{N_2} \right) - \left(\frac{1}{N} \mathbf{1}_{N \times N} \right).$$

Thinking of the N_1 and N_2 grid levels as factor levels of X_1 and X_2 , we have $P = 1 + (N_1 - 1) + (N_2 - 1)$ degrees of freedom and D_2 has rank $N - P$. As in Section 2, let Σ denote the covariance of V_F so that we can write $\Sigma_{D_2} = \text{cov}(D_2 V_F) = D_2 \Sigma D_2^T$ and use $(D_2 V_{\hat{F}})^T \hat{\Sigma}_{D_2}^{-1} D_2 V_{\hat{F}} \sim \chi_{N-P}^2$ as our test statistic. Note that this testing procedure for total additivity is identical to the procedure for testing significance but in the final two steps we calculate an alternative difference matrix and test statistic.

This procedure also naturally extends to the case of d features X_1, \dots, X_d . To test hypotheses of the form

$$H_0 : \exists F_1, \dots, F_d \text{ s.t. } F(x_1, \dots, x_d) = F_1(x_1) + \dots + F_d(x_d) \quad \forall (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}} \quad (3)$$

$$H_1 : F(x_1, \dots, x_d) \neq F_1(x_1) + \dots + F_d(x_d) \text{ for some } (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, \dots, F_d$$

we require a d -dimensional grid of test points so that given N_i levels of each feature X_i , our grid contains a total of $N = \prod_{i=1}^d N_i$ test points. Further, define

$$\hat{f}_{\dots j \dots} = \frac{1}{N_1 \dots N_{p-1} N_{p+1} \dots N_d} \sum_{i_1=1}^{N_1} \dots \sum_{i_{p-1}=1}^{N_{p-1}} \sum_{i_{p+1}=1}^{N_{p+1}} \dots \sum_{i_d=1}^{N_d} \hat{F}_{i_1 \dots j \dots i_d}$$

to be the average prediction over all points in the grid at the j^{th} level defined on the p^{th} feature, x_{pj} . As in the 2-dimensional case, we can rewrite the hypotheses in (3) as

$$H_0 : F_{i_1 \dots i_d} - f_{i_1 \dots} - f_{\dots i_2 \dots} - \dots - f_{\dots i_d} + (d-1)\mu = 0 \text{ for all } (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F_{i_1 \dots i_d} - f_{i_1 \dots} - f_{\dots i_2 \dots} - \dots - f_{\dots i_d} + (d-1)\mu \neq 0 \text{ for some } (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}}$$

and write $\hat{F}_{i_1 \dots i_d} - \hat{f}_{i_1 \dots} - \dots - \hat{f}_{\dots i_d} + (d-1)\bar{F}$ as $D_d V_{\hat{F}}$. Again, we define Σ to be the covariance of V_F so that $\Sigma_{D_d} = \text{cov}(D_d V_F) = D_d \Sigma D_d^T$ and we can use $(D_d V_{\hat{F}})^T \hat{\Sigma}_{D_d}^{-1} D_d V_{\hat{F}} \sim \chi_{N-P}^2$ as our test statistic, where $P = 1 + (N_1 - 1) + \dots + (N_d - 1)$.

Importantly, the additive functions need not be univariate. Define a (disjoint) partition of the feature space $\mathbf{S}_1, \dots, \mathbf{S}_q$ so that $\cup_{i=1}^q \mathbf{S}_i = \{X_1, \dots, X_d\}$. We can test hypotheses of the form

$$H_0 : \exists F_1, \dots, F_q \text{ such that } F(\mathbf{s}_1, \dots, \mathbf{s}_q) = F_1(\mathbf{s}_1) + \dots + F_q(\mathbf{s}_q) \quad \forall (\mathbf{s}_1, \dots, \mathbf{s}_q) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F(\mathbf{s}_1, \dots, \mathbf{s}_q) \neq F_1(\mathbf{s}_1) + \dots + F_q(\mathbf{s}_q) \text{ for some } (\mathbf{s}_1, \dots, \mathbf{s}_q) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, \dots, F_q$$

in exactly the same fashion by appropriately defining levels of an q -dimensional grid of test points.

Tests for partial additivity

We now handle the case where we are interested in testing only whether a proper subset of features contribute additively to the response. Suppose that our training set consists of three features X_1, X_2 , and X_3 and we are interested in testing

$$H_0 : \exists F_1, F_2 \text{ s.t. } F(x_1, x_2, x_3) = F_1(x_1, x_3) + F_2(x_2, x_3) \quad \forall (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}} \quad (4)$$

$$H_1 : F(x_1, x_2, x_3) \neq F_1(x_1, x_3) + F_2(x_2, x_3) \text{ for some } (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, F_2.$$

Rejecting this null hypothesis means that an interaction exists between X_1 and X_2 but implies nothing about potential interactions between X_1 and X_3 or between X_2 and X_3 . Hooker (2004) uses the size of the deviation of F from partial additivity as a means of identifying the bivariate and higher-order interactions required to reconstruct some percentage of the variation in the values of F . This is also referred to as the Sobol index for the X_1, X_2 interaction (Sobol, 2001). Define a 3-dimensional grid of test points with N_1, N_2 , and N_3 levels of X_1, X_2 and X_3 , respectively and continuing with the dot notation, define

$$\hat{f}_{i \cdot k} = \frac{1}{N_2} \sum_{j=1}^{N_2} \hat{F}_{ijk} \quad \text{and} \quad \hat{f}_{\cdot jk} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{F}_{ijk}$$

to be the average prediction over all levels of the feature X_2 in the grid at the i^{th} and k^{th} levels x_{1_i} and x_{3_k} , and the average prediction over all levels of the feature X_1 in the grid at the j^{th} and k^{th} levels x_{2_j} and x_{3_k} , respectively. If there is no interaction between X_1 and X_2 , then $F_{ijk} - f_{i \cdot k} - f_{\cdot jk} + f_{\cdot \cdot k} = 0$ at all levels $(x_{1_i}, x_{2_j}, x_{3_k})$ in the grid. Thus, we can

rewrite the hypotheses in (4) as

$$\begin{aligned} H_0 : F_{ijk} - f_{i.k} - f_{.jk} + f_{..k} &= 0 \quad \forall (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}} \\ H_1 : F_{ijk} - f_{i.k} - f_{.jk} + f_{..k} &\neq 0 \quad \text{for some } (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}} \end{aligned}$$

and use the empirical analogues of these parameters to conduct the testing procedure. Once again, we can write $\hat{F}_{ijk} - \hat{f}_{i.k} - \hat{f}_{.jk} + \hat{f}_{..k}$ as $D_3 V_{\hat{F}}$ for the appropriate difference matrix D_3 . Defining Σ as the covariance of V_F , we can write $\Sigma_{D_3} = \text{cov}(D_3 V_F) = D_3 \Sigma D_3^T$ and use $(D_3 V_{\hat{F}})^T \hat{\Sigma}_{D_3}^{-1} D_3 V_{\hat{F}} \sim \chi_{N-P}^2$ as our test statistic, where $N = N_1 N_2 N_3$. Note that since we must now account for two-way interactions, we have $P = 1 + (N_1 - 1) + (N_2 - 1) + (N_3 - 1) + (N_1 - 1)(N_3 - 1) + (N_2 - 1)(N_3 - 1)$ degrees of freedom and D_3 is of rank $N - P$. As was the case in testing for total additivity, the testing procedure remains identical with the appropriate difference matrix and test statistic calculated in the final steps.

This same testing procedure can also be performed when our training set consists of d features and we are interested in determining whether an interaction exists between X_i and X_j . Denote the set of all features except X_i and X_j as $\mathbf{X}_{-i,j}$ so that our hypotheses become

$$\begin{aligned} H_0 : \exists F_i, F_j \text{ such that } F(x_1, \dots, x_d) &= F_i(x_i, \mathbf{x}_{-i,j}) + F_j(x_j, \mathbf{x}_{-i,j}) \quad \forall (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}} \\ H_1 : F(x_1, \dots, x_d) &\neq F_i(x_i, \mathbf{x}_{-i,j}) + F_j(x_j, \mathbf{x}_{-i,j}) \text{ for some } (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_i, F_j. \end{aligned}$$

Now, instead of the third dimension of the grid containing levels of the single feature X_3 , these are now vector levels $\mathbf{x}_{-i,j}$ and the testing procedure remains identical. Likewise, X_i and X_j may be treated as vectors of features by redefining the grid levels as levels of the appropriate vector.

Remark: The testing procedures above as well as those defined in Section 2 were derived assuming equal weight is placed on each point in the test grid. In some cases, it may be advantageous to instead differentially weight grid points, for example based on the local density of observations. This alternative approach based on minimizing a weighted sum of squared errors is outlined in Appendix A. For a more thorough review of when such an

alternative may be preferred, we refer the reader to Hooker (2007).

4 Random Projections

The above procedures require estimating a covariance matrix of size proportional to the number of points in the test grid. However, estimating the variance parameters with too small an ensemble can result in a significant overestimate of the variance, thereby substantially reducing the power of our testing procedures; see Mentch and Hooker (2016) for a more complete discussion. Thus, in situations where large grids and/or complex additive forms are of interest, it may become computationally infeasible to directly obtain an accurate covariance estimate. In light of this, we further extend our above procedures to make use of random projections.

Random projections have a long-established history as a dimension-reduction method. The Johnson-Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984) provides that orthogonal projections from high dimensional spaces into lower dimensional spaces approximately preserve the distances between the projected elements. Lopes et al. (2011) and Srivastava et al. (2015) leverage this result to produce a high-dimensional extension of Hotelling's classic T^2 test to the $p > n$ case. Specifically, given two multivariate samples $X_{n_1 \times p}$ and $Y_{n_2 \times p}$, the data is projected via a random projection matrix R into a reduced dimension $r < n, p$ where an analogous testing procedure can be well-defined. In the latter work, the authors denote this test RAPTT (**R**andom **P**rojection **T**-**T**est) and for each projection matrix R_i , the projected test statistic and p-value are given by

$$T_{R_i}^2 = \frac{1}{n_1^{-1} + n_2^{-1}} (\bar{X} - \bar{Y})' R_i (R_i' S R_i)^{-1} R_i' (\bar{X} - \bar{Y})$$

and

$$\theta_{R_i} = 1 - F_{r, n-r+1} \left(\frac{n-r+1}{r} \frac{T_{R_i}^2}{n} \right)$$

where $n = n_1 + n_2 - 2$, S is the (pooled) sample covariance matrix, and $F_{a,b}$ denotes the F -distribution with numerator and denominator degrees of freedom a and b , respectively. RAPTT proceeds by sampling M random projection matrices R_1, \dots, R_M thereby obtaining

a total of M of the test statistics and p-values defined above. The final test statistic in the procedure with level α is defined as the average across the M p-values, $\theta = \frac{1}{M} \sum_{i=1}^M \theta_{R_i}$ and the null hypothesis of equal means is rejected whenever $\theta < u_\alpha$ where u_α is chosen such that $P \left[\theta < u_\alpha \middle| H_0 \right] = \alpha$.

In our context, we consider a training set of size n , an ensemble consisting of m trees, each of which is built with a subsample of size k , and we are interested in predicting at N total test points. Recall from Section 2 that the simplest form of test statistic that can be used to evaluate variable importance is given by $\hat{\mu}_N^T \hat{\Sigma}^{-1} \hat{\mu}_N \sim \chi_N^2$ where $\hat{\mu}$ is the vector of ensemble predictions, and $\hat{\Sigma}$ is the corresponding covariance matrix estimate. Given m predictions at each of N locations, we can think of our data as an $m \times N$ matrix so that for a set of M random projection matrices R_1, \dots, R_M and reduced dimension $r < m, N$, we can write each projected test statistic as

$$T_{R_i} = \hat{\mu}_N^T R_i (R_i^T \hat{\Sigma} R_i)^{-1} R_i^T \hat{\mu}_N \sim \chi_r^2. \quad (5)$$

The grid structure can also be incorporated in a straightforward manner. Though we utilize a difference matrix D to project into the space of additive models, so long as the elements of the R_i are independently generated continuous random variables, the overall projection has rank r with probability 1. The original test statistic is given by $(DV_{\hat{F}})^T \hat{\Sigma}_D^{-1} DV_{\hat{F}}$ where $\Sigma_D = \text{cov}(DV_F) = D \Sigma D^T$ and so the test statistic and p-value incorporating a random projection R_i become

$$T_{R_i} = (DV_{\hat{F}})^T R_i (R_i^T \hat{\Sigma}_D R_i)^{-1} R_i^T (DV_{\hat{F}}) \sim \chi_r^2$$

and

$$\theta_i = 1 - \Phi_r^2(T_{R_i})$$

respectively, where $r < N - P$ and Φ_r^2 denotes the cdf of the χ_r^2 . For M replicates of this randomized testing procedure, we can define our final test statistic as $\theta = \frac{1}{M} \sum_{i=1}^M \theta_i$ in the same fashion as RAPTT, where we reject H_0 whenever $\theta < u_\alpha$ and u_α is chosen such that $P \left[\theta < u_\alpha \middle| H_0 \right] = \alpha$.

4.1 Defining the Testing Parameters

The procedures developed in the preceding sections require a number of user-specified parameters. First, as noted in Srivastava et al. (2015), the choice of reduced dimension r is an important consideration that can influence the power of projection-based testing procedures. In our case, the covariance parameters are difficult to estimate accurately on large grids and thus, though the procedure is well-defined for $1 \leq r < m$, this practical restriction necessitates a relatively small projected dimension r . In many cases, we see a significant drop in power when testing on grids consisting of more than approximately 30 points, so choosing $5 \leq r \leq 15$ should be reasonable and computationally feasible in most situations. Further, note that because r is small, little dependence remains between the resulting p-values. Under the null hypothesis, each p-value is uniformly distributed on $[0, 1]$ and the mean of independent standard uniform random variables follows a Bates distribution, so the final cutoff u_α can be well approximated by the α quantile of this distribution.

The ideal method of sampling the random projection matrices is of less concern; Srivastava et al. (2015) show that any semi-orthogonal matrix R with elements generated from a continuous distribution with finite second moment satisfies the necessary conditions to perform the projection-based tests. For our situation, we recommend generating such matrices by sampling individual elements from a standard normal distribution, orthogonalizing via a process such as Gram-Schmidt, and selecting the appropriate submatrix. Such a procedure is straightforward and can be implemented in most software packages.

Algorithm 1 makes the random-projection-based testing procedure explicit, using the internal variance estimation procedure proposed in Mentch and Hooker (2016). For a particular query point of interest \mathbf{x} , the asymptotic variance of the prediction is given by

$$\frac{k^2}{m\alpha}\zeta_{1,k} + \frac{1}{m}\zeta_{k,k}$$

where $\zeta_{1,k} = \text{var}(E(T_i(\mathbf{x})|\tilde{\mathbf{x}}))$ represents the variance between tree-based predictions $T_i(\mathbf{x})$ at \mathbf{x} given a single common training point $\tilde{\mathbf{x}}$, $\zeta_{k,k} = \text{var}(T_i(\mathbf{x}))$ denotes the between-tree variance, and $\alpha = \lim n/m$. The algorithm makes use of the parameters $n_{\tilde{\mathbf{x}}}$ and n_{MC} in

order to structure the ensemble in such a fashion so as to readily post-compute consistent estimates of $\zeta_{1,k}$ and $\zeta_{k,k}$. The parameter $n_{\tilde{\mathbf{x}}}$ corresponds to the number of conditional expectation estimates $E(T_i(\mathbf{x})|\tilde{\mathbf{x}})$ computed in the definition of $\zeta_{1,k}$ and n_{MC} is the number of Monte Carlo samples used to estimate each conditional expectation so that in this case, $m = n_{\tilde{\mathbf{x}}} \times n_{MC}$. Though the ensemble need not be constructed in such a fashion, the internal estimation procedure allows us to easily select a small projected dimension r and also allows for the covariance estimates to be computed at no additional cost to the original ensemble.

Algorithm 1: Random Projection Testing Procedure

- 1 Compute difference matrix D
 - 2 Select reduced dimension r
 - 3 Generate random projection matrices R_1, \dots, R_M
 - 4 **for** i in 1 to $n_{\tilde{\mathbf{x}}}$
 - 5 Select initial fixed point $\tilde{\mathbf{x}}^{(i)}$
 - 6 **for** j in 1 to n_{MC}
 - 7 Select subsample $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)},j}$ of size k_n from training set that includes $\tilde{\mathbf{x}}^{(i)}$
 - 8 Build tree using subsample $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)},j}$
 - 9 Use tree to predict at each of the N grid points to obtain \hat{V}_j
 - 10 Apply each projection to \hat{V} to obtain $\hat{W}_{j,c} = (D\hat{V}_j)^T R_c$
 - 11 **end for**
 - 12 Record average of $\hat{W}_{j,c}$ over j for each projection
 - 13 **end for**
 - 14 Compute variance of each of the $n_{\tilde{\mathbf{x}}}$ averages to estimate each ζ_{1,k_n}
 - 15 Compute variance of all predictions from each projection to estimate each ζ_{k_n,k_n}
 - 16 Compute mean of all predictions from each projection to estimate each θ_{k_n}
 - 17 Compute each p-value $\theta_1, \dots, \theta_M$ by comparing to χ_r^2
 - 18 Record average p-value θ and compare to Bates α quantile
-

Finally, the levels of the test grid are an important consideration. As with all supervised learning procedures, these test points should be concentrated near the observed data so as to minimize the effects of extrapolation. However, with tree-based procedures, choosing grid points that appear in the original sample can also be problematic. Because the trees in random forests are grown to near full-depth without pruning, predictions made arbitrarily close to points in the training sample can suffer from overfitting and as a result, create the artificial appearance of interactions. Lastly, because predictions are based on localized

Method	n	α -level	Power
Linear Model	250	0.056	1.000
Subbagged Ensemble		0.065	0.954
Linear Model	500	0.048	1.000
Subbagged Ensemble		0.047	0.998
Linear Model	1000	0.046	1.000
Subbagged Ensemble		0.020	0.999

Table 1: Empirical α -levels and power for the linear model example.

averaging, grid points should be selected away from the boundary of the feature space to avoid edge effects. In most situations, uniformly spaced grid points in the interior of the feature space should produce tests with high power that preserve the level of the test.

5 Simulations

We now provide simulations to investigate the power of our proposed testing procedures. Suppose first that we have two features X_1 and X_2 and that our responses are generated according to $Y = X_1 + X_2 + \beta X_1 X_2 + \epsilon$ where we set $\beta = 0$ to assess α -level and $\beta = 1$ to evaluate power with $\epsilon \sim \mathcal{N}(0, 0.05^2)$. We first test for total additivity on 1000 datasets when $\beta = 0$ and 1000 datasets where $\beta = 1$, taking our empirical α -level as the proportion of tests that incorrectly reject the null hypothesis (when $\beta = 0$) and our estimate of power as the proportion of tests that correctly reject the null hypothesis (when $\beta = 1$). For reference, we also built 1000 linear regression models using the traditional t-test to determine whether the interaction is significant and recorded the empirical α -level and power of this testing procedure. This was repeated for data sets of size of 250, 500, and 1000 using subsample sizes of 30, 50, and 75 respectively and the results are shown in Table 1. The test grid was selected as a 4×4 grid with levels 0.2, 0.4, 0.6, and 0.8. In each case, our test for total additivity using a subbagged ensemble performed nearly exactly as well as the traditional t-test.

We also selected a number of more complex regression functions that have been used in previous publications related to testing additivity, such as De Canditiis and Sapatinas (2004) and Barry (1993), to further investigate α -level and power. Each estimate is the

Model	Test	Noise s.d.			Model	Test	Noise s.d.		
		0.5	0.25	0.05			0.5	0.25	0.05
(a) x_1	T	0.009 0.025	0.007 0.031	0.000 0.000	(h) x_1x_2	T	0.085 0.305	0.702 0.927	1.000 1.000
(b) e^{x_1}	T	0.002 0.028	0.000 0.011	0.000 0.000	(i) $x_1x_2x_3$	P	0.001 0.002	0.007 0.028	0.948 0.998
(c) $e^{x_1} + \sin(\pi x_2)$	T	0.008 0.045	0.008 0.060	0.007 0.059	(j) $\frac{\exp(5(x_1+x_2))}{1+\exp(5(x_1+x_2))} - 1$	T	0.006 0.021	0.029 0.089	0.948 0.999
(d) $x_1 + x_2 + x_3$	T	0.002 0.000	0.003 0.001	0.001 0.001	(k) $\frac{1+\sin(2\pi(x_1+x_2))}{2}$	T	1.000 1.000	1.000 1.000	1.000 1.000
(e) $e^{x_1} + e^{x_2} + e^{x_3}$	T	0.003 0.005	0.007 0.019	0.007 0.012	(l) $\frac{1+\sin(2\pi(x_1+x_2+x_3))}{2}$	P	0.158 0.011	0.874 0.222	1.000 0.959
(f) $x_1x_3 + x_2x_3$	P	0.000 0.000	0.000 0.001	0.002 0.008	(m) $64(x_1x_2)^3(1-x_1x_2)^3$	T	0.907 0.987	1.000 1.000	1.000 1.000
(g) $e^{x_1x_3} + e^{x_2x_3}$	P	0.000 0.000	0.001 0.006	0.014 0.066	(n) $64(x_1x_2x_3)^3(1-x_1x_2x_3)^3$	P	0.051 0.136	0.722 0.898	0.999 1.000

Table 2: Empirical α -level and power for a variety of underlying regression functions with noise levels of different standard deviations. Tests are either for Total (T) or Partial (P) additivity; for each model, the top result represents the power for a test without random projections, the bottom for the test employing random projections. The lettered labels beside each model are for comparison purposes to Figure 2.

result of 1000 simulations with a sample size of 500, subsample size of 50, and a 4×4 test grid (with levels 0.2, 0.4, 0.6, and 0.8) in the 2-dimensional tests for total additivity and a $3 \times 3 \times 3$ grid (with levels 0.3, 0.5, and 0.7) in the 3-dimensional tests for total and partial additivity. In each case the features were selected uniformly at random from $[0, 1]$, the responses generated according to $Y = F(\mathbf{X}) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with σ chosen to take values 0.05, 0.25 or 0.5, and the covariance estimated via the internal estimation procedure. The results are shown in Table 2 where the first line for each model gives the rejection probability for the tests defined here. Note that even though the response in the first two models does not depend on X_2 , this additional feature was still included in the training sets and the same test for total additivity was performed. In each case, we see that our false rejection rate is very conservative and we also maintain high power. Note that in each of these simulations, the variance estimation parameters were selected as $n_{\hat{\sigma}} = 50$ and $n_{MC} = 250$. These parameters assignments are smaller than those chosen in Mentch and Hooker (2016) and the authors note that these smaller ensemble sizes often lead to an overestimate of the variance thus resulting in the conservative test results (low α -levels) seen in Table 2.

Next, we repeated these simulations on the same functions, this time employing our

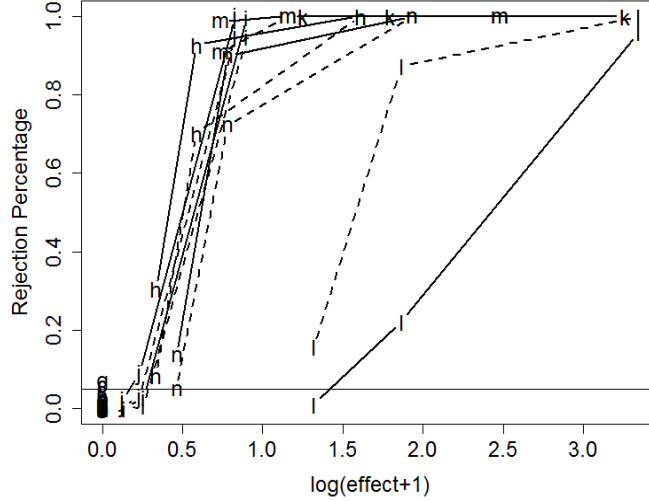


Figure 2: Graphical representation of the proportion of rejected tests out of 1000 trials corresponding to Table 2. The x -axis plots the non-centrality parameter for the non-random projection test for each case. Lines connect tests of the same model at different model variances; solid lines represent tests with random projections, dashed lines without.

tests that utilize random projections. In the 2-dimensional tests for total additivity, we use a 10×10 grid so that $N = 100$ and in the 3-dimensional tests for total additivity and the tests for partial additivity, we use a $5 \times 5 \times 5$ grid so that $N = 125$. The results are shown in Table 2 in the second row for each model. Note that in these tests, we maintain a reasonable type 1 error rate but achieve significantly more power due to the finer resolution of the test grid. These results are also presented graphically in Figure 2 where we can see that the tests utilizing random projections tend to have higher power. The only exception to this is model (l) with $y = 0.5(1 + \sin(2\pi(x_1 + x_2 + x_3))) + \epsilon$ where the complexity of the response surface and the choice of evaluation points likely affected the outcome.

The computational effort required to perform these tests is proportional to the dimension and overall size of the chosen grid. That is, tests of a particular form may be carried out at little additional computational cost for larger dimensions of the covariate space. To demonstrate this point, we first examine a test of total additivity. Here we again employ the model $Y = X_1 + X_2 + \beta X_1 X_2 + \epsilon$ where covariates are sampled uniformly from $[-1, 1]^{d+2}$, d takes values 5, 10 and 20, and ϵ is chosen to be either $N(0, 0.01)$ or $N(0, 1)$. Here d represents the dimension of nuisance covariates and β is taken to be one of 0, 0.1,

0.25, 0.5, 1, 2, giving the strength of the interaction. For each combination of β and d , we employ a 5×5 grid with points selected uniformly in $[-0.6, 0.6]$ and utilize 1000 random projections with $r = 5$. Selecting interior grid points in $[-0.6, 0.6]$ helps avoid the potential edge effects common in tree-based methods when predicting near the boundary of the feature space. For each of these settings, we generated 1000 datasets of 500 observations from which we obtained a random forest with subsamples of size 50 and conducted tests of total and partial additivity. The results of this simulation are given in left two panels of Figure 3 where we see that these tests achieve approximately the correct α level at $\beta = 0$, but quickly produce high power. We observe an expected drop in power with increasing error variance, but relative insensitivity to nuisance dimensions.

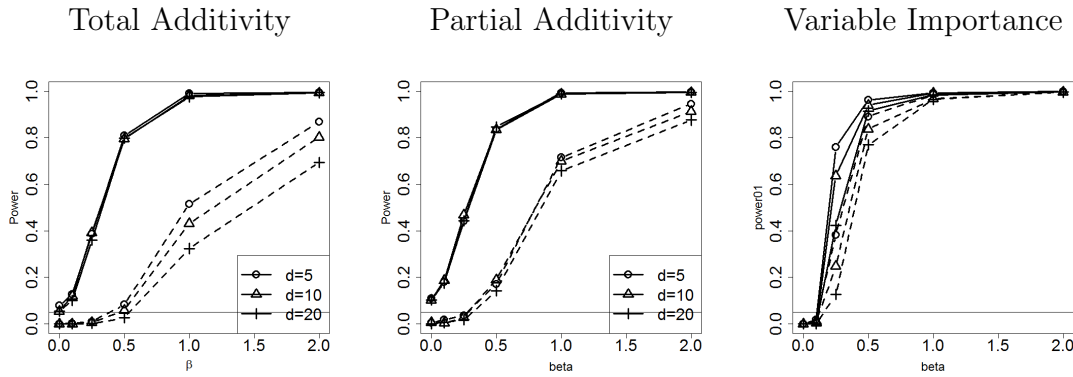


Figure 3: Results of a simulated power experiment. The left panel provides the power of a test of total additivity between two covariates with strength governed by β in the presence of d additional nuisance covariates. The middle panel repeats this procedure with tests of partial additivity. The right panel tests the importance of a three-dimensional set of covariates in the presence of an additional $d + 3$ covariates. In both cases, responses were generated with Gaussian errors with standard deviation 0.1 (solid lines) or 1.0 (dashed lines). Here we observe sensitivity to error variance, but a relatively small impact of nuisance covariate dimension.

We next extend this experiment to testing the importance of a *group* of variables. Here we employ the model

$$Y = \beta(X_1 + X_2 + X_3) + X_4 + X_5 + X_6 + \epsilon$$

under the same data generation scheme. Here we test the joint significance of (X_1, X_2, X_3) while also including further signal from (X_4, X_5, X_6) . As above, we used $d = 5, 10$ or 20 additional nuisance covariates and β is taken to be one of $0, 0.1, 0.25, 0.5, 1, 2$, giving the strength of the signal from the first three covariates. The ϵ_i are again normal with standard deviation 0.1 or 1 . The right panel in Figure 3 shows the empirical power of the test of importance for the vector (X_1, X_2, X_3) based on 1000 simulations of datasets of size 500 and subsamples of size 50. For each combination of β and d , we employ a $5 \times 5 \times 5$ grid with points selected uniformly in $[-0.6, 0.6]$ and utilize 1000 random projections with $r = 5$. We observe approximately the correct α -level at $\beta = 0$ with power increasing with β , resulting in power of approximately 0.8 at $\beta = 0.5$. These results are in agreement with Biau (2012) in which it is suggested that random forests are largely able ignore nuisance covariates with power decreasing only marginally with larger nuisance dimension d .

6 Application

We now demonstrate our testing procedures on a dataset provided by a team of ornithologists at the Cornell University Lab of Ornithology. This dataset was compiled in an effort to determine how pollution levels affect the change in Wood Thrush population. The data consists of 3 pollutant features, mercury deposition (*md*), acid deposition (*ad*), and soil PH level (*sph*) as well as 2 non-pollutant features, elevation (*elev*) and abundance (*ab*). We begin our analysis by testing whether the pollutant and non-pollutant features are additive:

$$H_0 : F(md, ad, sph, elev, ab) = F_P(md, ad, sph) + F_{NP}(elev, ab). \quad (6)$$

In this case we have two feature sets, $\mathbf{X}_1 = (md, ad, sph)$ and $\mathbf{X}_2 = (elev, ab)$ and we performed a test for total additivity using 4 levels of each set – the 0.20, 0.40, 0.60, and 0.80 quantiles of each feature – for a total of 16 test points. Our test statistic was 52.30, larger than the critical value, the 0.95 quantile of the χ^2_9 , of 16.92 so we reject the null hypothesis in (6) and conclude that an interaction exists between the pollutant and non-pollutant features. This result was confirmed by our random projection test, which consisted of 1000 random projections to a dimension of $r = 5$ using a 10×10 test grid. In

this case, the final averaged p-value was only 0.0043, far below the critical value of 0.485.

Next, we investigated how the pollutants contributed to the response. Based on preliminary investigations, ebird researchers suspected an interaction between mercury and acid deposition (md and ad) but were unsure of the relationship between soil PH (sph) and md and ad . In performing these tests for partial additivity, our test grid consisted of 3 points for each feature set, the 0.30, 0.50, and 0.70 quantiles of each feature for a total of 27 test points and a critical value, the 0.95 quantile of the χ^2_{12} , of 21.03. Our test for partial additivity between md and ad ,

$$H_0 : F(md, ad, sph, elev, ab) = F_1(md, sph, elev, ab) + F_2(ad, sph, elev, ab),$$

yielded a significant result with a test statistic of 41.00 so our test supports the belief that an interaction exists between md and ad . Again, this result was supported by our random projection test, which consisted of 1000 random projections to a dimension of $r = 5$ using a $5 \times 5 \times 5$ test grid, for a total of 125 test points. The final averaged p-value was only 0.0064, far below the critical value of 0.485.

Our test for partial additivity between sph and the vector (md, ad)

$$H_0 : F(md, ad, sph, elev, ab) = F_1(md, ad, elev, ab) + F_2(sph, elev, ab)$$

yielded a test statistic of 36.43, above the critical value of 21.03, so once again we reject the null hypothesis and conclude that an interaction exists between sph and (md, ad) . This result was again supported by the random projection test based on 1000 random projections to a dimension of $r = 5$ using a $5 \times 5 \times 5$ test grid. We find a final averaged p-value of 0.225, which, though larger than in the previous tests, is still far below the critical value of 0.485.

7 Discussion

This work harnesses desirable asymptotic properties of subsampled ensemble learners to develop formal hypothesis tests for additivity in random forests and suggests that tradi-

tional scientific and statistical questions need not be seen as a sacrifice of less interpretable learning procedures. Our tests require the definition of a reasonably sized test grid in order to achieve reasonably accurate covariance estimates while preserving power. When larger grids or more complex additive forms are required, we appeal to random projections and demonstrate that our tests still maintain very high power.

Many of the above demonstrations employed a version of random forests in which each covariate remains eligible at each split (subbagged ensembles), though we point out that the theory established in previous work such as Mentch and Hooker (2016), Wager and Athey (2015), and Scornet et al. (2015) allows for most general subsampled random forests implementations, or in fact any ensemble-type learner that conforms to the regularity conditions to be used. We caution however that the predictive improvement often seen with random forests is generally attributed to the increased independence between trees and thus should be expected to be less dramatic in these cases where subsamples are used in lieu of the traditional bootstrap samples.

Finally, it is important to note that the particular additive forms for which the testing procedures were developed were chosen only because of their scientific utility. Testing procedures for alternative additive forms can be developed in a similar manner by establishing appropriate model parameters from an ANOVA set-up and defining the difference matrix D accordingly. These methods can also be extended to provide formal statistical guarantees for the screening procedures described in Hooker (2004).

References

- Amato, U. and Antoniadis, A. (2001). Adaptive wavelet series estimation in separable nonparametric regression models. *Statistics and Computing*, 11(4):373–394.
- Barry, D. (1993). Testing for additivity of a regression function. *The Annals of Statistics*, pages 235–254.
- Biau, G. (2012). Analysis of a Random Forests Model. *The Journal of Machine Learning Research*, 98888:1063–1095.

- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.
- Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.
- De Canditiis, D. and Sapatinas, T. (2004). Testing for additivity and joint effects in multivariate nonparametric regression using fourier and wavelet methods. *Statistics and Computing*, 14(3):235–249.
- Derbort, S., Dette, H., and Munk, A. (2002). A test for additivity in nonparametric regression. *Annals of the Institute of Statistical Mathematics*, 54(1):60–82.
- Dette, H. and Derbort, S. (2001). Analysis of variance in nonparametric regression models. *Journal of multivariate analysis*, 76(1):110–137.
- Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007.
- Eubank, R., Hart, J. D., Simpson, D., Stefanski, L. A., et al. (1995). Testing for additivity in nonparametric regression. *The Annals of Statistics*, 23(6):1896–1920.
- Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association*, 100(471):890–907.
- Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43. CRC Press.
- Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580. ACM.
- Hooker, G. (2007). Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3).

- Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1.
- Linton, O. (1995). A kernel method of estimating structured nonparametric. *Biometrika*, 82(1):93–100.
- Lopes, M., Jacob, L., and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214.
- Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM.
- Mammen, E., Linton, O., Nielsen, J., et al. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27(5):1443–1490.
- Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17:1–41.
- Opsomer, J. D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, 93(442):605–619.
- Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, 8(4):715–732.
- Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.
- Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.
- Srivastava, R., Li, P., and Ruppert, D. (2015). RAPTT: An Exact Two-Sample Test in High Dimensions Using Random Projections. *Journal of Computational and Graphical Statistics*, (just-accepted).

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705.

Wager, S. and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*.

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

Appendix

A The generalized approach

The testing procedures developed in Sections 2 and 3 were derived by choosing the model parameters that minimized the sum of squared error (SSE) with equal weight placed on each point in the test grid. Instead, we may wish to differentially weight points on the grid. For example, in the above tests for partial additivity, we can select \hat{F}_1 and \hat{F}_2 to minimize the weighted SSE

$$WSSE = \sum_{i,j,k} w_{i,j,k} \left(F(x_{1_i}, x_{2_j}, x_{3_k}) - F_1(x_{1_i}, x_{3_k}) - F_2(x_{2_j}, x_{3_k}) \right)^2$$

where x_1, x_2, x_3 can be taken as individual features or interpreted more generally as vectors of features and the weights w_{ijk} are specified by the user. Hooker (2007) recommends basing such weights on an approximation to the density of observations near $(x_{1_i}, x_{2_j}, x_{3_k})$. This procedure takes the form of a weighted ANOVA. In particular, define \vec{F} to be the $N_1 N_3 + N_2 N_3$ vector concatenating the $\hat{F}_1(x_1, x_3)$ and $\hat{F}_2(x_2, x_3)$ and as in the previous sections let $V_{\hat{F}}$ be the vector containing the \hat{F}_{ijk} . Further, let Z be the $N \times (N_1 N_3 + N_2 N_3)$ matrix defined so that $Z\vec{F}$ produces the corresponding $\hat{F}_1(x_1, x_3) + \hat{F}_2(x_2, x_3)$ and let W be a diagonal matrix containing the weights. Then we can write

$$WSSE = (V_{\hat{F}} - Z\vec{F})^T W (V_{\hat{F}} - Z\vec{F})$$

and we know that the solution \vec{F} that minimizes this weighted SSE is given by

$$\vec{F} = (Z^T W Z)^{-1} Z^T W V_{\hat{F}}$$

so that under the null hypothesis

$$V_{\hat{F}} - Z\vec{F} = (I - Z(Z^T W Z)^{-1} Z^T W) V_{\hat{F}}$$

has mean 0. Further, letting Σ denote the covariance of V_F , the variance of $V_F - Z\vec{F}$ is given by

$$C = (I - Z(Z^T W Z)^{-1} Z^T W) \Sigma (I - Z(Z^T W Z)^{-1} Z^T W)^T$$

so that

$$[(I - Z(Z^T W Z)^{-1} Z^T W) V_{\hat{F}}]^T \hat{C}^{-1} [(I - Z(Z^T W Z)^{-1} Z^T W) V_{\hat{F}}]$$

has a χ^2_{N-P} distribution, where P remains as defined in the standard procedures developed in Sections 2 and 3. For equal weighting (W given by the identity matrix), these calculations reduce to the averages employed above, and for the sake of simplicity we have restricted ourselves to this choice. Note also that this generalized WLS approach can be applied to more general forms of additivity as well as those tests for total additivity developed in the previous section.