Nonparametric Variable Screening with Optimal Decision Stumps

Jason M. Klusowski
Princeton University

Peter M. Tian
Princeton University

Abstract

Decision trees and their ensembles are endowed with a rich set of diagnostic tools for ranking and screening variables in a predictive model. Despite the widespread use of tree based variable importance measures, pinning down their theoretical properties has been challenging and therefore largely unexplored. To address this gap between theory and practice, we derive finite sample performance guarantees for variable selection in nonparametric models using a single-level CART decision tree (a decision stump). Under standard operating assumptions in variable screening literature, we find that the marginal signal strength of each variable and ambient dimensionality can be considerably weaker and higher, respectively, than state-of-the-art nonparametric variable selection methods. Furthermore, unlike previous marginal screening methods that estimate each marginal projection via a truncated basis expansion, the fitted model used here is a simple, parsimonious decision stump, thereby eliminating the need for tuning the number of basis terms. Thus, surprisingly, even though decision stumps are highly inaccurate for estimation purposes, they can still be used to perform consistent model selection.

1 INTRODUCTION

A common task in many applied disciplines involves determining which variables, among many, are most important in a predictive model. In high-dimensional sparse models, many of these predictor variables may be irrelevant in how they affect the response variable. As a result, variable selection techniques are crucial for filtering out irrelevant variables in order to prevent overfitting, improve accuracy, and enhance the interpretability of the model. Indeed, algorithms that screen for relevant variables have been instrumental in the modern development of fields such as genomics, biomedical imaging, signal processing, image analysis, and finance, where high-dimensional, sparse data is frequently encountered (Fan and Lv, 2008).

Over the years, numerous parametric and nonparametric methods for variable selection in high-dimensional models have been proposed and studied. For linear models, the LARS algorithm (Efron et al., 2004) (for Lasso (Tibshirani, 1996)) and Sure Independence Screening (SIS) (Fan and Lv, 2008) serve as prototypical examples that have achieved immense success, both practically and theoretically. Other strategies for nonparametric additive models such as Nonparametric Independence Screening (NIS) (Fan et al., 2011) and Sparse Additive Models (SPAM) (Ravikumar et al., 2009) have also enjoyed a similar history of success.

1.1 Tree-based Variable Selection

Alternatively, because they are built from highly interpretable and simple objects, decision tree models are another important tool in the data analyst's repertoire. Indeed, after only a brief explanation, one is able to understand the tree output in terms of meaningful domain specific attributes of the variables. In addition to being interpretable, tree based model have good computational scalability as the number of data points grows, making them faster than many other methods when dealing with large datasets. In terms of flexibility, they can naturally handle a mixture of numeric variables, categorical variables, and missing values. Lastly, they require less preprocessing (because they are invariant to monotone transformations of the inputs), are quite robust to outliers, and are relatively unaffected by the inclusion of many irrelevant variables (Hastie et al., 2009; Klusowski, 2020), the last point being of relevance to the variable selection problem.

Conventional tree structured models such as CART (Breiman et al., 1984), random forests (Breiman, 2001a), ExtraTrees (Geurts et al., 2006), and gradient tree boosting (Friedman, 2001) are also equipped with heuristic variable importance measures that can be used to rank and identify relevant predictor variables

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

for further investigation (such as plotting their partial dependence functions). In fact, tree based variable importance measures have been used to discover genes in bioinformatics (Breiman, 2001b; Lunetta et al., 2004; Bureau et al., 2005; Díaz-Uriarte and de Andrés, 2006; Huynh-Thu et al., 2010), identify loyal customers and clients (Buckinx et al., 2007; W. Buckinx, 2005; Lariviere and den Poel, 2005), detect network intrusion (Zhang and Zulkernine, 2006; Zhang et al., 2008), and understand income redistribution preferences (Keely and Tan, 2008), to name a few applications.

1.2 Mean Decrease in Impurity

An attractive feature specific to CART methodology is that one can compute, essentially for free, measures of variable importance (or influence) using the optimal splitting variables and their corresponding impurities. The canonical CART-based variable importance measure is the Mean Decrease in Impurity (MDI) (Friedman, 2001, Section 8.1), (Breiman et al., 1984, Section 5.3.4), (Hastie et al., 2009, Sections 10.13.1 & 15.3.2), which calculates an importance score for a variable by summing the largest impurity reductions (weighted by the fraction of samples in the node) over all non-terminal nodes split with that variable, averaged over all trees in the ensemble. In contrast to the aforementioned variable selection procedures like SIS, NIS, Lasso, and SPAM, except for a few papers, little is known about the finite sample performance of MDI. Theoretical results are mainly limited to the asymptotic, fixed dimensional data setting and to showing what would be expected from a reasonable measure of variable importance. For example, it is shown in (Louppe et al., 2013) that the MDI importance of a categorical variable (in an asymptotic data and ensemble size setting) is zero precisely when the variable is irrelevant, and that the MDI importance for relevant variables remains unchanged if irrelevant variables are removed or added.

Recent complementary work by (Scornet, 2020) established the large sample properties of MDI for additive models by showing that it converges to sensible quantities like the variance of the component functions, provided the decision tree is sufficiently deep. However, these results are not fine-grained enough to handle the case where either the dimensionality grows or the marginal signals decay with the sample size. Furthermore, (Scornet, 2020) crucially relies on the assumption that the CART decision tree has a small approximation error, which is currently only known for additive models (Scornet et al., 2015). It is therefore unclear whether the techniques can be generalized to models beyond those considered therein. On the other hand, our results suggest that tree based variable importance measures can still have good variable selection properties even though the underlying tree model may be a poor predictor of the data generating process—which can occur with CART and random forests.

Lastly, we mention that important steps have also been taken to characterize the finite sample properties of MDI; (Li et al., 2019) show that MDI is less biased for irrelevant variables when each tree is shallow. This work therefore covers one facet of the variable selection problem, i.e., controlling the number of false positives, and will be employed in our proofs.

1.3 New Contributions

The lack of theoretical development for tree based variable selection is likely because the training mechanism involves complex steps such as bagging, boosting, pruning, random selections of the predictor variables for candidate splits, recursive splitting, and line search to find the best split points (Kazemitabar et al., 2017). The last consideration, importantly, means that the underlying tree construction (e.g., split points) depends on *both* the input and output data, which enables it to adapt to structural properties of the underlying statistical model (such as sparsity). This data adaptivity is a double-edged sword from a theoretical standpoint, though, since unravelling the data dependence is a formidable task.

Despite the aforementioned challenges, we advance the study of tree based variable selection by focusing on the following two fundamental questions:

- Do tree based methods enjoy finite sample guarantees for variable ranking?
- What are the benefits of tree based methods over other variable screening methods?

Specifically, we derive rigorous finite sample guarantees for what we call the SDI importance measure, a sobriquet for *Single-level Decrease in Impurity*, which is a special case of MDI for a single-level CART decision tree or "decision stump" (Iba and Langley, 1992). This is similar in spirit to the approach of DSTUMP (Kazemitabar et al., 2017) but, importantly, SDI incorporates the line search step by finding the *optimal* split point, instead of the empirical median, of every predictor variable. As we shall see, ranking variables according to their SDI is equivalent to ranking the variables according to the marginal sample correlations between the response data and the optimal decision stump with respect to those variables. This equivalence also yields connections with other variable selection methods: for linear models with Gaussian variates, we show that SDI is asymptotically equivalent to SIS (up to logarithmic factors in the sample size), and so SDI inherits the so-called sure screening property (Fan and Lv, 2008) under suitable assumptions.

Unlike SIS, however, SDI is accompanied by provable guarantees for nonparametric models. We show that under certain conditions, SDI achieves *model selection consistency*; that is, it correctly selects the relevant variables of the model with probability approaching one as the sample size increases. In fact, the minimum signal strength of each relevant variable and maximum dimensionality of the model are shown to be less restrictive for SDI than NIS or SPAM. In the linear model case with Gaussian variates, SDI is shown to nearly match the optimal sample size threshold (achieved by Lasso) for exact support recovery. These favorable properties are striking when one is reminded that the underlying model fit to the data is a simple, parsimonious decision stump—in particular, there is no need to specify a flexible function class (such as polynomial splines) and be concerned with calibrating the number of basis terms or bandwidth parameters. Finally, we empirically compare SDI to other contemporaneous variable selection algorithms, namely, SIS, NIS, Lasso, and SPAM, and find that it performs competitively.

2 SETUP AND ALGORITHM

In this section we introduce notation, formalize the learning setting, and give an explicit layout of our SDI algorithm. At the end of the section, we discuss its complexity and provide several interpretations.

2.1 Notation

For labeled data $\{(U_1, V_1), \dots, (U_n, V_n))\}$ drawn from a population distribution (U, V), we let $\widehat{\mathrm{Cov}}(U, V) = \frac{1}{n}\sum_{i=1}^n (U_i - \overline{U})(V_i - \overline{V})$, $\widehat{\mathrm{Var}}(U) = \frac{1}{n}\sum_{i=1}^n (U_i - \overline{U})^2$, $\overline{U} = \frac{1}{n}\sum_{i=1}^n U_i$, and $\widehat{\rho}(U, V) = \frac{\widehat{\mathrm{Cov}}(U,V)}{\sqrt{\widehat{\mathrm{Var}}(U)\widehat{\mathrm{Var}}(V)}}$ denote the sample covariance, variance, mean, and Pearson product-moment correlation coefficient respectively. The population level covariance, variance, and correlation are denoted by $\mathrm{Cov}(U, V)$, $\mathrm{Var}(U) = \sigma_U^2$, and $\rho(U, V)$, respectively.

2.2 Learning Setting

Throughout this paper, we operate under a standard regression framework where the statistical model is $Y = g(\mathbf{X}) + \varepsilon$, the vector of predictor variables is $\mathbf{X} = (X_1, \dots, X_p)^\top$, and ε is statistical noise. While our results are valid for general nonparametric models, for conceptual simplicity, the canonical model class we have in mind is *additive models*, i.e.,

$$g(X_1, \dots, X_p) = g_1(X_1) + \cdots + g_p(X_p) \qquad (1)$$

for some univariate component functions $g_1(\cdot), \dots, g_p(\cdot)$. As is standard with additive modeling (Hastie et al., 2009, Section 9.1.1), for identifiability of the components, we assume that the $g_j(X_j)$ have population mean zero for all j. This model class strikes a balance between flexibility and learnability—it is more flexible than linear models, but, by giving up on modeling interaction terms, it does not suffer from the curse of dimensionality.

We observe data $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ with the i^{th} sample point $(\mathbf{X}_i, Y_i) = (X_{i1}, \dots, X_{ip}, Y_i)$ drawn independently from the model above. Note that with this notation, X_{ij} are i.i.d. instances of the random variable X_j. We assume that the regression function $g(\cdot)$ depends only on a small subset of the variables $\{X_j\}_{j\in\mathcal{S}}$, which we call *relevant variables* with *support* $\mathcal{S} \subset \{1, \dots, p\}$ and *sparsity level* $s = |\mathcal{S}| \ll p$. Equivalently, $g_j(\cdot)$ is identically zero for the *irrelevant variables* $\{X_j\}_{j\in\mathcal{S}^c}$. In this paper, we consider the *variable ranking* problem, defined here as ranking the variables so that the top s coincide with $\mathcal{S}$ with high probability. As a corollary, this will enable us to solve the *variable selection* problem, namely, determining the subset $\mathcal{S}$. We pay special attention to the high-dimensional regime where $p \gg n$. In fact, in Section 4.3 we will provide conditions under which consistent variable selection occurs even when $p = \exp(o(n))$.

2.3 Prior Art

The conventional approach to marginal screening for nonparametric additive models is to directly estimate either the nonparametric components $g_j(X_j)$ or the marginal projections

$$f_j(X_j) := \mathbb{E}[Y|X_j],$$

with the ultimate goal of studying their variances or their correlations with the response variable.[1] To accomplish this, SIS, NIS, and (Hall and Miller, 2009) rank the variables according to the correlations between the response values and least squares fits over a univariate model class $\mathcal{H}$, i.e.,

$$\widehat{\rho}(\hat{h}(X_j), Y), \quad \hat{h}(\cdot) \in \operatorname*{arg\,min}_{h(\cdot)\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n (Y_i - h(X_{ij}))^2. \qquad (2)$$

The model class $\mathcal{H}$ is chosen to make the above optimization tractable, while at the same time, be sufficiently rich in order to approximate $f_j(X_j)$. For example, if $\mathcal{H}$ is the space of polynomial splines of a fixed degree, then $\hat{h}(\cdot)$ in (2) can be computed efficiently via a truncated B-spline basis expansion

$$\beta_1\Psi_1(X_j) + \beta_2\Psi_2(X_j) + \cdots + \beta_{d_n}\Psi_{d_n}(X_j), \quad \beta_j \in \mathbb{R}, \qquad (3)$$

as is done with NIS. Similarly, SIS takes $\mathcal{H}$ to be the family of linear functions in a single variable. Complementary methods that aim to directly estimate each

[1] Note that $f_j(X_j)$ need not be the same as $g_j(X_j)$ unless, for instance, the predictor variables are independent and the noise is independent and mean zero.

$g_j(X_j)$ include SPAM, which uses a smooth backfitting algorithm with soft-thresholding, and (Huang et al., 2010), which combines adaptive group Lasso with truncated B-spline basis expansions.

As we shall see, SDI is equivalent to ranking the variables according to (2) when $\mathcal{H}$ consists of the collection of all decision stumps in X_j of the form

$$\beta_1 \mathbf{1}(X_j \leq z) + \beta_2 \mathbf{1}(X_j > z), \qquad z, \beta_1, \beta_2 \in \mathbb{R}. \quad (4)$$

Unlike previous models such as polynomial splines (3), a one-level decision tree, realized by the model (4) above, severely underfits the data and would therefore be ill-advised for estimating $f_j(X_j)$, if that were the goal. Remarkably, we show that this rigidity does not hinder SDI for variable selection. What redeems SDI is that, unlike the aforementioned methods that are based on linear estimators, decision stumps (4) are *nonlinear* since the splits points can depend on the response data. These model nonlinearities equip SDI with the ability to discover nonlinear patterns in the data, despite its poor approximation capabilities.

2.4 The SDI Algorithm

In this section, we provide the details for the SDI algorithm. We first provide some high-level intuition.

In order to determine whether, say, X_j is relevant for predicting Y from $\mathbf{X}$, it is natural to first divide the data into two groups according to whether X_j is above or below some predetermined cutoff value and then assess how much the variance in Y changes before and after this division. A small change in the variability indicates a weak or nonexistent dependence of Y on X_j; whereas, a moderate to large change indicates heterogeneity in Y across different values of X_j. As we now explain, this is precisely what SDI does when the predetermined cutoff value is sought by a least squares fit over all possible ways of dividing the data.

Let z be a candidate split for a variable X_j that divides the response data Y into left and right daughter nodes based on the j^{th} variable. Define the mean of the left daughter node to be $\overline{Y}_L = \frac{1}{N_L}\sum_{i:X_{ij}\leq z} Y_i$ and the mean of the right daughter node to be $\overline{Y}_R = \frac{1}{N_R}\sum_{i:X_{ij}>z} Y_i$ and let the size of the left and right daughter nodes be $N_L = \#\{i : X_{ij} \leq z\}$ and $N_R = \#\{i : X_{ij} > z\}$, respectively. For CART regression trees, the *impurity reduction* (or variance reduction) in the response variable Y from choosing the split point z for the j^{th} variable is defined to be

$$\widehat{\Delta}(z; X_j, Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \frac{1}{n}\sum_{i:X_{ij}\leq z}(Y_i - \overline{Y}_L)^2 - \frac{1}{n}\sum_{i:X_{ij}>z}(Y_i - \overline{Y}_R)^2. \quad (5)$$

For each variable X_j, we choose a split point $\hat{z}_j$ that maximizes the impurity reduction

$$\hat{z}_j \in \arg\max_z \widehat{\Delta}(z; X_j, Y),$$

and for convenience, we denote the largest impurity reduction by

$$\widehat{\Delta}(X_j, Y) := \widehat{\Delta}(\hat{z}_j; X_j, Y).^2$$

We then rank the variables commensurate with the sizes of their impurity reductions, i.e., we obtain a ranking $(\hat{\jmath}_1, \ldots, \hat{\jmath}_p)$ where $\widehat{\Delta}(X_{\hat{\jmath}_1}, Y) \geq \cdots \geq \widehat{\Delta}(X_{\hat{\jmath}_p}, Y)$. If desired, these rankings can be repurposed to perform model selection (e.g., an estimate $\widehat{\mathcal{S}}$ of $\mathcal{S}$), as we now explain. If we are given the sparsity level s in advance, we can choose $\widehat{\mathcal{S}}$ to be the top s of these ranked variables; otherwise, we must find a data-driven choice of how many variables to include. Equivalently, the latter case is realized by choosing $\widehat{\mathcal{S}}$ to be the indices j for which $\widehat{\Delta}(X_j, Y) \geq \gamma_n$, where γ_n is a threshold to be described in Section 2.5. This is of course a delicate task as including too many variables may lead to more false positives.

By (Breiman et al., 1984, Section 9.3), using a sum of squares decomposition, we can rewrite the impurity reduction (5) as

$$\widehat{\Delta}(z; X_j, Y) = \frac{N_L}{n}\frac{N_R}{n}(\overline{Y}_L - \overline{Y}_R)^2, \quad (6)$$

which allows us to compute the largest impurity reductions for all possible split points with a single pass over the data by first ordering the data along X_j and then updating $\overline{Y}_L$ and $\overline{Y}_R$ in an online fashion. This alternative expression for the objective function facilitates its rapid evaluation and *exact* optimization. Pseudocode for SDI is given in Algorithm 1.

2.5 Data-driven Choices of γ_n

As briefly mentioned in Section 2.4, if we do not know the sparsity level s in advance, we can instead use a data-driven threshold γ_n to control the number of selected variables. Here we propose a data-driven methods to determine the threshold γ_n, which is similar to

[2] The impurity reduction can be highly non-concave and therefore the optimal split point need not be unique. In such cases, we break ties arbitrarily.

Algorithm 1: Single-level Decrease in Impurity (SDI)

```
Input: Dataset $\mathcal{D} = \{(X_{i1}, \dots, X_{ip}, Y_i)\}_{i=1}^n$
for $j = 1, \dots, p$ do
    Relabel $\mathcal{D}$ with $X_{ij}$ sorted in increasing order
    Initialize $\overline{Y}_L = 0$, $\overline{Y}_R = \overline{Y}$, $\widehat{\Delta}(X_j, Y) = 0$
    for $i = 1, \dots, n-1$ do
        Update $\overline{Y}_L \leftarrow \frac{i-1}{i}\overline{Y}_L + \frac{Y_i}{i}$,   $\overline{Y}_R \leftarrow \frac{n-i+1}{n-i}\overline{Y}_R - \frac{Y_i}{n-i}$
        Compute $\widehat{\Delta}(X_{ij}; X_j, Y) = \frac{i}{n}(1 - \frac{i}{n})(\overline{Y}_L - \overline{Y}_R)^2$
        if $\widehat{\Delta}(X_{ij}; X_j, Y) > \widehat{\Delta}(X_j, Y)$ then
            Update $\widehat{\Delta}(X_j, Y) \leftarrow \widehat{\Delta}(X_{ij}; X_j, Y)$
        end
    end
end
Output: Ranking $(\hat{\jmath}_1, \dots, \hat{\jmath}_p)$ such that
        $\widehat{\Delta}(X_{\hat{\jmath}_1}, Y) \geq \cdots \geq \widehat{\Delta}(X_{\hat{\jmath}_p}, Y)$
```

the Iterative Nonparametric Independence Screening (INIS) method based on NIS (Fan et al., 2011, Section 4). The first step is to choose a random permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ of the data to decouple $\mathbf{X}_i$ from Y_i so that the new dataset $\mathcal{D}^\pi = \{(\mathbf{X}_{\pi(i)}, Y_i)\}$ follows a null model. Then we choose the threshold γ_n to be the maximum of the impurity reductions $\widehat{\Delta}(X_j, Y; \mathcal{D}^\pi)$ over all j based on the dataset $\mathcal{D}^\pi$. We can also generate T different permutations π and take the maximum of $\widehat{\Delta}(X_j, Y; \mathcal{D}^\pi)$ over all such permuted datasets to get a more significant threshold, i.e., $\gamma_n = \max_{j,\,\pi} \widehat{\Delta}(X_j, Y; \mathcal{D}^\pi)$. With γ_n selected in this way, SDI will then output the variable indices $\widehat{\mathcal{S}}$ consisting of the indices j for which the original impurity reductions $\widehat{\Delta}(X_j, Y; \mathcal{D})$ are at least γ_n. Interestingly, this method is similar to MDA in that we permute the data values of a given variable and calculate the resulting change in the quality of the fit.

2.6 Computational Issues

We now briefly discuss the computational complexity of Algorithm 1—or equivalently—the computational complexity of growing a single-level CART decision tree. For each variable X_j, we first sort the input data along X_j with $\mathcal{O}(n \log n)$ operations. We then evaluate the decrease in impurity along n data points (as done in the nested for-loop of Algorithm 1), and finally find the maximum among these n values (as done in the nested if-statement of Algorithm 1), all with $\mathcal{O}(n)$ operations. Thus, the total number of calculations for all of the p variables is $\mathcal{O}(pn \log(n))$. This is only slightly worse than the complexity of SIS for linear models $\mathcal{O}(pn)$, comparable to NIS based on the complexity of fitting B-splines, and favorable to that of Lasso or stepwise regression $\mathcal{O}(p^3 + p^2 n)$, especially when p is large (Efron et al., 2004; Hastie et al., 2009). While approximate methods like coordinate descent for Lasso can reduce the complexity to $\mathcal{O}(pn)$ at each iteration, their convergence properties are unclear. As SPAM is a generalization of Lasso for nonparametric additive models, its implementation (via a functional version of coordinate descent) may be similarly expensive.

2.7 Interpretations of SDI

In this section, we outline two interpretations of SDI.

Interpretation 1. Our first interpretation of SDI is in terms of the sample correlation between the response and a decision stump. To see this, denote the decision stump that splits X_j at z by

$$\widetilde{Y}(X_j) \coloneqq \overline{Y}_L \, \mathbf{1}(X_j \leq z) + \overline{Y}_R \, \mathbf{1}(X_j > z)$$

and one at an optimal split value $\hat{z}_j$ by

$$\widehat{Y}(X_j) \coloneqq \overline{Y}_L \, \mathbf{1}(X_j \leq \hat{z}_j) + \overline{Y}_R \, \mathbf{1}(X_j > \hat{z}_j).$$

Note that $\widehat{Y}(X_j)$ equivalently minimizes the marginal sum of squares (2) over the collection of all decision stumps (4). Next, by Lemma A.1 in (Klusowski, 2020), we have:

$$\widehat{\Delta}(z, X_j, Y) = \widehat{\mathrm{Var}}(Y) \times \widehat{\rho}^{\,2}(\widetilde{Y}(X_j), Y), \quad \text{and}$$
$$\widehat{\rho}^{\,2}(\widetilde{Y}(X_j), Y) = 1 - \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \widetilde{Y}(X_{ij}))^2}{\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2}, \tag{7}$$

where

$$\widehat{\rho}(\widetilde{Y}(X_j), Y) \coloneqq \frac{\frac{1}{n}\sum_{i=1}^n (\widetilde{Y}(X_{ij}) - \overline{Y})(Y_i - \overline{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (\widetilde{Y}(X_{ij}) - \overline{Y})^2 \times \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2}} \geq 0$$

is the Pearson product-moment sample correlation coefficient between the data Y and decision stump $\widetilde{Y}(X_j)$. In other words, we see from (7) that an optimal split point $\hat{z}_j$ is chosen to maximize the Pearson sample correlation between the data Y and decision stump $\widetilde{Y}(X_j)$. This reveals that SDI is, at its heart, a correlation ranking method, in the same spirit as SIS, NIS, and (Hall and Miller, 2009) via (2).

Like r^2 for linear models, (7) reveals that the squared sample correlation $\widehat{\rho}^{\,2}(\widetilde{Y}(X_j), Y)$ equals the *coefficient of determination* R^2, i.e., the fraction of variance in Y explained by a decision stump $\widetilde{Y}(X_j)$ in X_j.[3] Thus, SDI is also equivalent to ranking the variables according to the goodness-of-fit for decision stumps of each variable.

[3] However, unlike linear models, for this relationship to be true, the decision stump $\widetilde{Y}(X_j)$ need not necessarily be a least squares fit, i.e., $\widehat{Y}(X_j)$.

Interpretation 2. The other interpretation is in terms of the aforementioned MDI importance measure. Recall the definition of MDI in Section 1.2, i.e., for an individual decision tree T, the MDI for X_j is the total reduction in impurity attributed to the splitting variable X_j. More succinctly, the MDI of T for X_j equals

$$\sum_{\mathrm{t}} \frac{N(\mathrm{t})}{n} \widehat{\Delta}(X_j, Y | \mathbf{X} \in \mathrm{t}), \tag{8}$$

where the sum extends over all non-terminal nodes t in which X_j was split, $N(\mathrm{t})$ is the number of sample points in t, and $\widehat{\Delta}(X_j, Y|\mathbf{X} \in \mathrm{t})$ is the largest reduction in impurity for samples in t. Note that if T is a decision stump with split along X_j, then (8) equals $\widehat{\Delta}(X_j, Y)$, the largest reduction in impurity at the root node. Because a split at the root node captures the main effects of the model, $\widehat{\Delta}(X_j, Y)$ can be seen as a first order approximation of (1.2) in which higher order interaction effects are ignored.

3 LINEAR MODELS

To connect SDI to other variable screening methods that are perhaps more familiar to the reader, we first consider a linear model with Gaussian distributed variables. We allow for any correlation structure between covariates. Recall from (7) that $\widehat{\Delta}(X_j, Y)$ is equal to $\widehat{\mathrm{Var}}(Y)$ times $\widehat{\rho}^2(\widehat{Y}(X_j), Y)$, so that SDI is equivalent to ranking by $\widehat{\rho}(\widehat{Y}(X_j), Y)$. Our first theorem shows that $\widehat{\rho}(\widehat{Y}(X_j), Y)$, the sample correlation between Y and an optimal decision stump in X_j, is, with high probability, sandwiched between constant multiples of $\rho(X_j, Y)$, the correlation between a linear model Y and a coordinate X_j. Because of space constraints, the proof is deferred to Supplement B.

Theorem 1 (SDI is asymptotically SIS) *Let $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$ and assume that $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ for some positive semi-definite matrix $\mathbf{\Sigma}$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$. Let $\delta \in (0, 1)$. There exists a universal positive constant C_0 such that, with probability at least $1 - \frac{C_0}{\sqrt{n\delta^2\rho^2(X_j,Y)}} \exp(-n\delta^2\rho^2(X_j, Y)/2)$,*

$$\widehat{\rho}(\widehat{Y}(X_j), Y) \geq \frac{(1-\delta)|\rho(X_j, Y)|}{\sqrt{\log(2n) + 1}}. \tag{9}$$

Furthermore, with probability at least $1 - 4n\exp(-n\delta^2/12) - 2\exp(-(n-1)/16)$,

$$\widehat{\rho}(\widehat{Y}(X_j), Y) \leq 5|\rho(X_j, Y)| + 2\delta. \tag{10}$$

Theorem 1 shows that with high-probability, SDI is asymptotically equivalent (up to logarithmic factors in the sample size) to SIS for linear models in that it ranks the magnitudes of the marginal sample correlations between a variable and the model, i.e., $\widehat{\rho}(X_j, Y) \approx \rho(X_j, Y)$. As a further parallel with decision stumps (see Section 2.7), the square of the sample correlation, $\widehat{\rho}^2(X_j, Y)$, is also equal to the coefficient of determination r^2 for the least squares linear fit of Y on X_j. We confirm the similarity between SDI and SIS empirically in Section 6.

One corollary of Theorem 1 is that, like SIS, SDI also enjoys the sure screening property, under the same assumptions as (Fan and Lv, 2008, Conditions 1-4), which include mild conditions on the eigenvalues of the design covariance matrices and minimum signals of the parameters β_j. Similarly, like SIS, SDI can also be paired with lower dimensional variable selection methods such as Lasso or SCAD (Antoniadis and Fan, 2001) for a complete variable selection algorithm in the correlated linear model case.

On the other hand, SDI, a nonlinear method, applies to broader contexts far beyond linear models. In the next section, we will investigate how SDI performs for general nonparametric models with additional assumptions on the distribution of the variables.

4 NONPARAMETRIC MODELS

In this section, we establish the variable ranking and selection consistency properties of SDI for general nonparametric models; that is, we show that for Algorithm 1, we have $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$ as $n \to \infty$.

Although our approach differs substantively, to facilitate easy comparisons with other marginal screening methods, our framework and assumptions will be similar. As mentioned earlier, SDI is based on a more parsimonious but significantly more biased model fit than those than underpin conventional methods. As we shall see, despite the decision stump severely underfitting the data, SDI nevertheless achieves model selection guarantees that are similar to, and in some cases stronger than, its competitors. This highlights a key difference between quantifying sensitivity and screening—in the latter case, we are not concerned with obtaining *consistent* estimates of the marginal projections $f_j(X_j)$ and their variances. Doing so demands more from the data and is therefore less efficient, when otherwise crude estimates would work equally well.

4.1 Assumptions

Here we describe the key assumptions and ideas which will be needed to achieve model selection consistency. The assumptions will be similar to those in the independence screening literature (Fan et al., 2011; Fan and Lv, 2008), but are weaker than most past work on tree based variable selection Li et al. (2019); Kazemitabar et al. (2017).

Assumption 1 (Bounded regression function) *The regression function $g(\cdot)$ is bounded with $B = \|g\|_\infty < \infty$.*

Assumption 2 (Smooth marginal projections) *Let r be a positive integer, let $0 < \alpha \leq 1$, let $d := r + \alpha$, and let $0 < L < \infty$. The r^{th} order derivative of $f_j(\cdot)$ exists and is L-Lipschitz of order α, i.e.,*

$$\left|f_j^{(r)}(x) - f_j^{(r)}(x')\right| \leq L|x - x'|^\alpha, \quad x, x' \in \mathbb{R}.$$

Assumption 3 (Partial orthogonal covariates) *The collections $\{X_j\}_{j \in \mathcal{S}}$ and $\{X_j\}_{j \in \mathcal{S}^c}$ are independent of each other.*

Assumption 4 (Uniform relevant variables) *The marginal distribution of each X_j, for $j \in \mathcal{S}$, is uniform on the unit interval.*

Assumption 5 (Sub-Gaussian error distribution) *The error distribution is conditionally sub-Gaussian given $\mathbf{X}$, i.e., $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$ and $\mathbb{E}[\exp(\lambda\varepsilon) \mid \mathbf{X}] \leq \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$ with $\sigma^2 > 0$.*

4.2 Discussion of the Assumptions

Assumption 2 is a standard smoothness assumption for variable selection in nonparametric additive models (Fan et al., 2011, Assumption A) and (Huang et al., 2010, Section 3). Because SDI does not involve tuning parameters that govern its approximation properties of the nonparametric constituents (such as with NIS and SPAM), Assumption 2 can be relaxed to allow for different levels of smoothness in different dimensions and, by straightforward modifications of our proofs, one can show that SDI adapts automatically.

Assumption 3 is essentially the so-called "partial orthogonality" condition in marginal screening methods (Fan and Song, 2010). Importantly, it allows for correlation between the relevant variables $\{X_j\}_{j \in \mathcal{S}}$, unlike previous works on tree based variable selection (Kazemitabar et al., 2017; Li et al., 2019). Notably, NIS and SPAM do allow for dependence between relevant and irrelevant variables, under suitable assumptions on the data matrix of basis functions. However, these assumptions are difficult to translate in terms of the joint distribution of the predictor variables and difficult to verify given the data. We do not believe this assumption is strictly necessary, however. The following example illustrates how SDI can be asymptotically impervious to confounders (see also the empirical evidence in Section 6).

Example 1 *Suppose $(X_1, X_2, \ldots, X_p) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Suppose that X_2 is an irrelevant variable that is uncorrelated with every relevant variable expect for one, say, X_1. Then, asymptotically, as long as X_1 and X_2 are not perfectly correlated, SDI will always rank X_1 above X_2; in other words $\Delta(X_2, Y) < \Delta(X_1, Y)$, where $\Delta(X_j, Y)$ is the population version of $\widehat{\Delta}(X_j, Y)$ (i.e., the almost sure limit of $\widehat{\Delta}(X_j, Y)$ as $n \to \infty$). This is quite remarkable given that Y can have arbitrary nonlinear dependence on both X_1 and X_2. Due to space constraints we include the full proof of this result in Supplement D.*

Assumption 4 is stated as is for clarity of exposition and is not strictly necessary for our main results to hold. For instance, we may assume instead that the marginal densities of the relevant variables are compactly supported and uniformly bounded above and below by a strictly positive constant, as in (Fan et al., 2011; Huang et al., 2010). More generally, other distributional relaxations are made possible by the fact that CART decision trees are invariant to monotone transformations, enabling us to reduce the general setting to the case where each predictor variable is uniformly distributed on $[0, 1]$.

4.3 Main Results

Assuming we know the size s of the support $\mathcal{S}$, we can use the SDI ranking from Algorithm 1 to choose the top s most important variables. Alternatively, if s is unknown, we instead choose an asymptotic threshold γ_n of the impurity reductions to select variables; that is, $\widehat{\mathcal{S}} = \{j : \widehat{\Delta}(X_j, Y) \geq \gamma_n\}$. We state our variable ranking guarantees in terms of the minimum signal strength of the relevant variables:

$$v := \min_{j \in \mathcal{S}} \mathrm{Var}(f_j(X_j)),$$

identical to the minimum variance parameter in independence screening papers (e.g., (Fan et al., 2011, Assumption C)). Note that v measures the minimum contribution of each relevant variable alone to the variance in Y, ignoring the effects of other variables.

Theorem 2 *Suppose Assumptions 1, 2, 3, 4, and 5 hold. Then the top s most important variables ranked by Algorithm 1 equal the correct set $\mathcal{S}$ of relevant variables with probability at least*

$$1 - 3s\exp(-C_1 nv) - 4n(p - s)\exp\Big(-\frac{nC_2 v^{1+1/d}}{24(B^2 + \sigma^2)}\Big), \tag{11}$$

where $C_1 = c_1(B^2 + \sigma^2)^{-1}$ and $C_2 = c_2 L^{-2/d}$ and c_1 and c_2 are universal positive constants.

Remark 1 *As a corollary of Theorem 2, we obtain a quantitative version of Proposition 1 in (Scornet et al., 2015) for the root node of a CART decision tree, which states more generally that a relevant variable is selected at each node with probability converging to one.*

4.4 Minimum Signal Strengths

Like all marginal screening methods, the theoretical basis for SDI is that each marginal projection for a relevant variable should be nonconstant, or equivalently, that $v > 0$. Note that when the relevant variables are independent and the underlying model is additive, per (1), the marginal projections equal the component functions of the additive model. Hence, $v = \min_{j \in \mathcal{S}} \operatorname{Var}(g_j(X_j))$, which will always be strictly greater than zero. As Theorem 2 shows, v controls the probability of a successful ranking of the variables. In practice, many of the relevant variables may have very small signals—therefore we are particularly interested in cases where v is allowed to become small when the sample size grows large, as we now discuss.

We see from Theorem 2 that in order to have model selection consistency with probability at least $1 - n^{-\Omega(1)}$, it suffices to have

$$v \gtrsim \Big(\frac{\log(n(p-s))}{n}\Big)^{\frac{d}{d+1}}, \tag{12}$$

up to constants that depend only on B, σ, L, r, and α. That is, (12) is a sufficient condition on the signal of all relevant variables so that $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$ as $n \to \infty$.

5 COMPARISONS

In this section, we compare the finite sample guarantees of SDI given in Section 4.3 and Section 3 to those of NIS, Lasso, and SPAM. To summarize, we find that SDI enjoys model selection consistency even when the marginal signal strengths of the relevant variables are smaller than those for NIS and SPAM. We also find that the minimum sample size of SDI for high probability support recovery is nearly what is required for Lasso, which is minimax optimal. Finally, we show that SDI can handle a larger number of predictor variables than NIS and SPAM.

Minimum signal strength for NIS. We analyze the details of Fan et al. (2011) to uncover the corresponding threshold v for NIS. In order to control the number of false positives, the probability bound in (Fan et al., 2011, Theorem 2) must approach one as $n \to \infty$, which necessitates

$$d_n \lesssim \Big(\frac{n^{1-4\kappa}}{\log(np)}\Big)^{1/3}, \tag{13}$$

where d_n is the number of spline basis functions and κ is a free parameter (in the notation of (Fan et al., 2011)). However notice that by (Fan et al., 2011, Assumption F), we must also have that $d_n \gtrsim n^{\frac{2\kappa}{2d+1}}$, and combining this with (13) shows that we must have

$$n^{-\kappa} \gtrsim \Big(\frac{\log(np)}{n}\Big)^{\frac{2d+1}{8d+10}}. \tag{14}$$

Now substituting (13) and (14) into $v \gtrsim d_n n^{-2\kappa}$ ((Fan et al., 2011, Assumption C)), it follows that we have

$$v \gtrsim \Big(\frac{\log(np)}{n}\Big)^{\frac{4d}{8d+10}}$$

for NIS, which is a significantly stronger minimum signal requirement than our (12).

Minimum signal strength for SPAM. When $d = 2$, by (Ravikumar et al., 2009, Section 6.1), we must have $v \gtrsim n^{-4/15} \log^{16/5}(np)$ for SPAM to achieve consistent model selection. For comparison, our algorithm allows for a smaller signal $v \gtrsim \big(\frac{\log(np)}{n}\big)^{2/3}$, which is obtained by setting $d = 2$ in (12).

Minimum sample size for consistency. Consider the linear model with Gaussian variates from Theorem 1, where for simplicity we additionally assume that $\mathbf{\Sigma} = \mathbf{I}_{p \times p}$ is the $p \times p$ identity matrix, yielding $\rho^2(X_j, Y) = \beta_j^2/(\sigma^2 + \sum_{k=1}^p \beta_k^2)$.

Following the same steps used to prove Theorem 2 but using Theorem 1 and Lemma A.2 instead, we can derive a result similar to Theorem 2 for the probability of exact support recovery, but for a linear model with Gaussian variates. The full details are in Supplement E. With the specifications $\sum_{k=1}^p \beta_k^2 = \mathcal{O}(1)$ and $\min_{j \in \mathcal{S}} |\beta_j|^2 \asymp 1/s$, we find that a sufficient sample size for high probability support recovery is

$$n \gg s \log(n) \log(n(p-s)),$$

which happens when

$$n \gg s \log(p-s) \times (\log(s) + \log\log(p-s)). \tag{15}$$

Now, it is shown in (Wainwright, 2009a, Corollary 1) that the minimax optimal threshold for support recovery under these parameter specifications is $n \asymp s \log(p-s)$, which is achieved by Lasso (Wainwright, 2009b). Amazingly, (15) coincides with this optimal threshold up to $\log(s)$ and $\log\log(p-s)$ factors, despite SDI not being tailored to linear models.

Maximum dimensionality. Suppose the signal strength v is bounded above and below by a positive constant when the sample size increases. Then Theorems 2 shows model selection consistency for SDI up to dimensionality $p = \exp(o(n))$. This is larger than the maximum dimensionality $p = \exp(o(n^{2(d-1)/(2d+1)}))$ for NIS (Fan et al., 2011, Section 3.2), thus applying to an even broader spectrum of ultra high-dimensional problems. Furthermore, when $d = 2$, SPAM is able to handle dimensionality up to $p = \exp(o(n^{1/6}))$ (Ravikumar et al., 2009, Equation (45)), which is again lower than the dimensionality $p = \exp(o(n))$ for SDI.

6 EXPERIMENTS

In this section, we conduct computer experiments of SDI with synthetic data. As there are many existing empirical studies of the related MDI measure (Kazemitabar et al., 2017; Li et al., 2019; Louppe, 2014; Louppe et al., 2013; Lundberg et al., 2018; Strobl et al., 2007; Wang et al., 2016; Wei et al., 2015), we do not aim for comprehensiveness.

Our experiments compare the performance of SDI with SIS, NIS, Lasso, SPAM, MDI for a single CART decision tree, and MDI for a random forest. Specifically, we assess performance based on the probability of *exact support recovery*. To ensure a fair comparison between SDI and the other algorithms, we assume a priori knowledge of the true sparsity level s, which is incorporated into Lasso and SPAM by specifying the model degrees of freedom in advance. These simulations were conducted in R using the packages `rpart` for SDI, `SAM` for SPAM, `SIS` for SIS, and `glmnet` for Lasso with default settings. We also compute two versions of MDI: MDI RF using the package `randomForest` with `ntrees = 100` and MDI CART (based on a pruned CART decision tree) using the package `rpart` with default settings. The source code from (Fan et al., 2011) was used to conduct experiments with NIS.

In all our experiments, we generate n samples from an s-sparse additive model $g(\mathbf{X}) = \sum_{j=1}^{s} g_j(X_j)$ for various types of components $g_j(X_j)$. The error distribution is $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, the sparsity level is fixed at $s = 4$, and the ambient dimension is fixed at $p = 2000$. We consider the following model types.

Model 1. Consider linear additive components $g_j(X_j) = X_j$ and variables $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where the covariance matrix $\mathbf{\Sigma}$ has diagonal entries equal to 1 and off-diagonal entries equal to some constant $\rho \in (-1, 1)$. We set the noise level $\sigma^2 = 1$ and consider correlation level $\rho = 0.5$.

Model 2. We consider additive components $g_j(X_j) = \cos(4\pi X_j)$, where $\mathbf{X} \sim \text{Uniform}([0,1]^p)$ (i.e., all predictor variables are independent) and $\sigma^2 = 1$.

Model 3. Consider nonlinear additive components

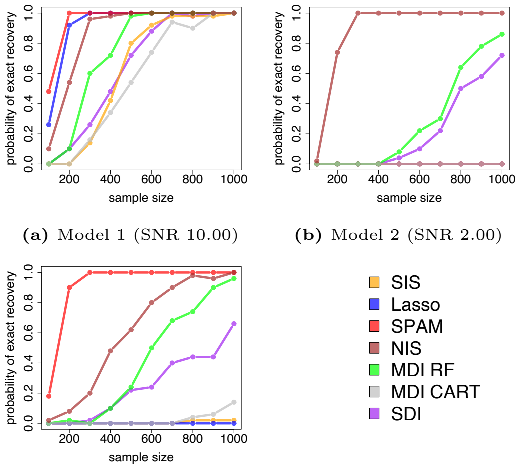
$$g_1(x) = 5x, \;\; g_2(x) = 3(2x-1)^2, \;\; g_3(x) = \frac{4\sin(2\pi x)}{2 - \sin(2\pi x)},$$

$$g_4(x) = 6(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)).$$

Let $\mathbf{X} \sim \text{Uniform}([0,1]^p)$ and set the noise level $\sigma^2 = 1.74$. This is the same model as Example 3 of (Fan et al., 2011) with $t = 0$ (and correlation 0).

Model 1 tests the correlated Gaussian linear model setting of Theorem 1 while Models 2 and 3 test the setting of Theorem 2 for general nonparametric models. Though our main results apply to general nonparametric models, we have chosen to focus our experiments on additive models to facilitate comparison with other methods designed for the same setting.

For our experiments on exact recovery, we fix the sparisty level $s = 4$ and estimate the probability of exact support recovery by running 50 independent replications and computing the fraction of replications which exactly recover the support. In Figure 1, we plot this estimated probability against various sample sizes $n < p$, namely, $n \in \{100, 200, 300, \ldots, 1000\}$. In agreement with Theorem 1, in Figure 1a, we observe that SDI and SIS exhibit similar behavior for correlated Gaussian linear models, a case in which all methods appear to achieve model selection consistency. As expected, Figure 1b and 1c show that SDI, NIS, SPAM, and MDI significantly outperform SIS and Lasso when the model has nonlinear and non-monotone additive components. For more irregular component functions such as sinusoids, SDI appears to outperform SPAM, as seen in Figure 1b. In general, for additive models, SDI appears to outperform its progenitor MDI CART though it seems to sacrifice a small amount of accuracy for simplicity compared to MDI RF.



(a) Model 1 (SNR 10.00)

(b) Model 2 (SNR 2.00)

(c) Model 3 (SNR 9.01)

Figure 1: Plots of estimated $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S})$ as n increases for various models (approximate signal to noise ratio in parentheses).

Acknowledgements

Financial support for JMK was provided in part by NSF DMS1915932 and NSF HDR TRIPODS DATA-INSPIRE DCCF-1934924. PMT was supported in part by the Gordon Wu Fellowship of Princeton University.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–955.

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

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and regression trees.* Chapman and Hall/CRC.

Buckinx, W., Verstraeten, G., and den Poel, D. V. (2007). Predicting customer loyalty using the internal transactional database. *Expert Systems with Applications*, 32:125–134.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Eerdewegh, P. V. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, 28:171–82.

Díaz-Uriarte, R. and de Andrés, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics volume*, 7(3):171–82.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494).

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 70:849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics. Springer New York.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*.

Iba, W. and Langley, P. (1992). Induction of one-level decision trees. *Machine Learning Proceedings*, pages 233–240.

Kazemitabar, J., Amini, A., Bloniarz, A., and Talwalkar, A. S. (2017). Variable importance using decision trees. In *Advances in Neural Information Processing Systems*, pages 426–435.

Keely, L. and Tan, C. (2008). Understanding preferences for income redistribution. *Journal of Public Economics*, 92:944–961.

Klusowski, J. M. (2020). Sparse learning with CART. In *Advances in Neural Information Processing Systems*.

Lariviere, B. and den Poel, D. V. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29:472–484.

Li, X., Wang, Y., Basu, S., Kumbier, K., and Yu, B. (2019). A debiased MDI feature importance measure for random forests. In *Advances in Neural Information Processing Systems 32*, pages 8049–8059. Curran Associates, Inc.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint: arXiv:1407.7502*.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, pages 431–439.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint: arXiv:1802.03888*.

Lunetta, K. L., Hayward, L. B., Segal, J., and Eerdewegh, P. V. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, 5(32):199–231.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2009). Spam: Sparse additive models. *Journal of the Royal Statistical Society. Series B*, 71(5):1009–1030.

Scornet, E. (2020). Trees, forests, and impurity-based variable importance. *arXiv preprint arXiv:2001.04295*.

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

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:563–564.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288.

W. Buckinx, D. V. d. P. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research*, 164:252–268.

Wainwright, M. J. (2009a). Information-theretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741.

Wainwright, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Wang, H., Yang, F., and Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*, 17(60).

Wei, P., Lu, Z., and Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, 142:399–432.

Zhang, J. and Zulkernine, M. (2006). A hybrid network intrusion detection technique using random forests. *Proceedings of the First International Conference on Availability, Reliability and Security*, pages 262–269.

Zhang, J., Zulkernine, M., and Haque, A. (2008). A hybrid network intrusion detection technique using random forests. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, 38(5):649–659.