# Nonparametric Variable Screening with Optimal Decision Stumps

Jason M. Klusowski* and Peter M. Tian†

Department of Operations Research and Financial Engineering, 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 attempt to directly 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

*jason.klusowski@princeton.edu
Supported in part by NSF DMS-1915932 and NSF TRIPODS DATA-INSPIRE CCF-1934924.
†ptian@princeton.edu
Supported in part by the Gordon Wu Fellowship of Princeton University.

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 [11].

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 [9] (for Lasso [36]) and Sure Independence Screening (SIS) [11] 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) [10] and Sparse Additive Models (SPAM) [34] have also enjoyed a similar history of success.

## 1.1 Tree based variable selection procedures

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 construction and its 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 [16, 25], the last point being of relevance to the variable selection problem.

Conventional tree structured models such as CART [5], random forests [3], ExtraTrees [14], and gradient tree boosting [13] are also equipped with heuristic variable importance measures that can be used to rank and identify relevant predictor variables 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 [4, 32, 7, 8, 19], identify loyal customers and clients [6, 37, 26], detect network intrusion [43, 44], and understand income redistribution preferences [24], 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) [13, Section 8.1], [5, Section 5.3.4], [16, Sections 10.13.1 & 15.3.2], which calculates an importance score for a variable by summing the weighted impurity reductions over all non-terminal nodes split with that variable,

averaged over all trees in the ensemble. Another commonly used and sometimes more accurate measure is the Mean Decrease in Accuracy (MDA) or permutation importance measure, defined as the average difference in out-of-bag errors before and after randomly permuting the data values of a variable in out-of-bag samples over all trees in the ensemble [3]. However, from a computational perspective, MDI is preferable to MDA since it can be computed as each tree is grown with no additional cost. While we view the analysis of MDA as an important endeavor, it is outside the scope of the present paper and therefore not our current focus.

Note that in general, MDI and MDA may not necessarily be derived from accurate tree based predictors—which can occur with CART and random forests. On the other hand, as we show in this paper, 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.

## 1.3 New contributions

In contrast to the aforementioned variable selection procedures like Lasso, SIS, NIS, and SPAM, little is known about the finite sample performance of MDI. Theoretical results are mainly limited to showing what would be expected from a reasonable measure of variable importance. For example, it is shown in [30] that the MDI importance of a 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. Furthermore, the average bias of MDI for irrelevant variables is known to be small when each tree is not too deep [28].

This lack of theoretical development is likely because the training mechanism involves complex steps—which present new theoretical challenges—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 [23]. 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 line search step, which will prove to be a key ingredient of the methodology. With this context in mind, two natural questions we seek to answer are as follows.

- What notion of importance does MDI measure?
- Does MDI rank the variables according to their relevance?

Specifically, we derive rigorous finite sample guarantees for the *Single-level Decrease in*

*Impurity* (SDI) importance measure, which is a special case of MDI for a single-level CART decision tree or "decision stump" [20]. This is similar in spirit to the approach of DSTUMP [23] 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 too inherits the so-called sure screening property [11] under suitable assumptions.

Unlike SIS, 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 particularly 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 a polynomial spline family) 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, Lasso, NIS, and SPAM. We find that SDI is competitive and performs favorably in cases where the model exhibits more irregularity. Furthermore, we empirically verify the similarity between SDI and SIS for the linear model case, and confirm the model selection consistency properties of SDI for various types of nonparametric models.

Let us close this section by saying a few words about the primary goals of this paper. In practice, tree based variable importance measures such as MDI and MDA are most commonly used to *rank variables*, in order to determine which ones are worthy of further examination. This is a less ambitious endeavor than that sought by the aforementioned variable selection methods, which aim for consistent *model selection*, namely, determining exactly which variables are relevant and irrelevant. Though we do study the model selection problem, our priority is to demonstrate the power of tree based variable importance measures as interpretable, accurate, and efficient variable ranking tools.

## 1.4 Organization

The paper is organized according to the following schema. First, in Section 2, we formally describe our learning setting and problem, discuss prior art, and introduce the SDI algorithm. We outline some key lemmas and ideas used in the proofs in Section 3. In the next two sections, we establish our main results: in Section 4, we establish that for linear mod-

els, SDI is asymptotically equivalent to SIS (up to logarithmic factors in the sample size) and in Section 5, we establish the variable ranking and model selection consistency properties of SDI for more general nonparametric models. Lastly, we compare the performance of SDI to other well-known variable selection techniques from a theoretical perspective in Section 6 and from an empirical perspective in Section 7. We conclude with a few remarks in Section 8. Due to space constraints, we include the full proofs of our main results in Sections 4 and 5 in the appendix.

# 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 $\{(U_1, V_1), \ldots, (U_n, V_n))\}$ 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 $\mathbf{X} = (X_1, \ldots, X_p)^\top$, and $\varepsilon$ is statistical noise. While our results are valid for general nonparametric models, the canonical model class we have in mind is *additive models*, i.e.,

$$g(X_1, \ldots, X_p) = g_1(X_1) + \cdots + g_p(X_p) \tag{1}$$

for some univariate component functions $g_1(\cdot), \ldots, g_p(\cdot)$. As is standard with additive modeling [16, 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 $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ with the $i^{\text{th}}$ sample point $(\mathbf{X}_i, Y_i) = (X_{i1}, \ldots, X_{ip}, Y_i)$ 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* $\mathcal{S} \subset \{1, \ldots, p\}$ 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 5.3 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

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

with the ultimate goal of studying their variances (i.e., the individual contribution of $X_j$ to the variance in $Y$, ignoring the effects of the other variables) or their correlations with the response variable.[1] To accomplish this, SIS, NIS, and [15] rank the variables according to the correlations between the response values and least squares fits over a univariate model class $\mathcal{H}$, i.e.,

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

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, 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 $g_j(X_j)$ include SPAM, which uses a smooth back-fitting algorithm with soft-thresholding, and [18], 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

$$\beta_1 \, \mathbf{1}(X_j \le z) + \beta_2 \, \mathbf{1}(X_j > z), \qquad z, \beta_1, \beta_2 \in \mathbb{R}. \tag{3}$$

Unlike the previous expressive models, a one-level decision tree, realized by the model (3) 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

[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.

are based on linear estimators, decision stumps (3) 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.

Finally, we mention that one benefit of such a simple model as (3) is that it is completely free of tuning parameters. In contrast, other methods such as the ones listed here require careful calibration of, for example, variable bandwidth smoothers or the number of terms in the basis expansions (e.g., SPAM and NIS).

## 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 $j^{\text{th}}$ variable. Define the mean of the left daughter node to be $\overline{Y}_L = \frac{1}{N_L}\sum_{i:X_{ij}\le z} Y_i$ 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 $N_L = \#\{i : X_{ij} \le z\}$ 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 $j^{\text{th}}$ variable is defined to be

$$\widehat{\Delta}(z; X_j, Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \frac{1}{n}\sum_{i:X_{ij}\le z}(Y_i - \overline{Y}_L)^2 - \frac{1}{n}\sum_{i:X_{ij}>z}(Y_i - \overline{Y}_R)^2. \quad (4)$$

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

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

and for convenience, we denote the largest impurity reduction by

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

[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.

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 $\gamma_n$ 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 [5, Section 9.3], using a sum of squares decomposition, we can rewrite the impurity reduction (4) as

$$\widehat{\Delta}(z; X_j, Y) = \frac{N_L}{n} \frac{N_R}{n} (\overline{Y}_L - \overline{Y}_R)^2, \tag{5}$$

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.

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

**Input:** Dataset $\mathcal{D} = \{(X_{i1}, \ldots, X_{ip}, Y_i)\}_{i=1}^n$
**for** $j = 1, \ldots, p$ **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** $i = 1, \ldots, n-1$ **do**
    Update $\overline{Y}_L \leftarrow \frac{i-1}{i}\overline{Y}_L + \frac{Y_i}{i}$, $\quad \overline{Y}_R \leftarrow \frac{n-i+1}{n-i}\overline{Y}_R - \frac{Y_i}{n-i}$
    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 $(\hat{\jmath}_1, \ldots, \hat{\jmath}_p)$ such that $\widehat{\Delta}(X_{\hat{\jmath}_1}, Y) \geq \cdots \geq \widehat{\Delta}(X_{\hat{\jmath}_p}, Y)$

## 2.5 Data-driven choices of $\gamma_n$

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 $\gamma_n$ to modulate the number of selected variables. Here we propose two data-driven methods to determine the threshold $\gamma_n$.

*Permutation method* The first thresholding method is similar to the Iterative Nonparametric Independence Screening (INIS) method based on NIS [10, Section 4]. The first step

is to choose a random permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ 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 $\gamma_n$ 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 $\pi$ 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 $\gamma_n$ 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 $\gamma_n$. Interestingly, this method has parallels to MDA in that we permute the data values of a given variable and calculate the resulting change in the quality of the fit.

*Elbow method* One problem with the permutation method is that it performs poorly when there is correlation between relevant and irrelevant variables. This is because, in the decoupled dataset $\mathcal{D}^\pi$, there is essentially no correlation between $Y$ and any variable; whereas in the original dataset, $Y$ may be correlated with some of the irrelevant variables. In this case, the threshold $\gamma_n$ will be too small and the selected set of variables will include many irrelevant variables (false positives). An alternative is to employ visual inspection via the *elbow method*, which is typically used to determine the number of clusters in cluster analysis. Here we plot the largest impurity reductions $\widehat{\Delta}(X_j, Y)$ in decreasing order and search for an "elbow" in the curve. We then let $\widehat{\mathcal{S}}$ consist of the variable indices that come before the elbow. Instead of choosing the cutoff point visually, we can also automate the process by clustering the impurity reductions with, for example, a two-component Gaussian mixture model. Both of these methods are generally more robust to correlation between relevant and irrelevant variables than the permutation method.

We illustrate the empirical performance of both the permutation method and the elbow method in Section 7.

## 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 [9, 16]. While approximate methods like coordinate descent for Lasso can reduce the complexity to $\mathcal{O}(pn)$ at each iteration, their convergence properties are unclear. Thus, as SPAM is a generalization of Lasso for nonparametric additive models, its implementation

(via a functional version of coordinate descent for Lasso) 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 (3). Next, by Lemma A.1 in [25], we have:

$$\begin{aligned} \widehat{\Delta}(z, X_j, Y) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widetilde{Y}(X_{ij}))^2 \\ &= \widehat{\mathrm{Var}}(Y) \times \widehat{\rho}^{\,2}(\widetilde{Y}(X_j), Y), \end{aligned} \tag{6}$$

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 (6) 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 [15] via (2). In fact, as we shall see in Section 3, SDI is for all intents and purposes also similar to ranking the variables according to the correlation between the response data and the marginal projections $f_j(X_j)$.

Like $r^2$ for linear models, (6) 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. In fact, the equivalence between ranking with correlations and ranking with squared error goodness-of-fit is a ubiquitous trait among most models [15, Theorem 1].

[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 7, i.e., for an individual decision tree $T$, the MDI for a variable $X_j$ can be found by summing the largest impurity reductions (weighted by the fraction of samples in the node) over all non-terminal nodes split with $X_j$. More succinctly, the MDI of $T$ for $X_j$ equals

$$\sum_{\mathrm{t}} \frac{N(\mathrm{t})}{n} \widehat{\Delta}(X_j, Y | \mathbf{X} \in \mathrm{t}), \tag{7}$$

where the sum extends over all non-terminal nodes $\mathrm{t}$ in which $X_j$ was split, $N(\mathrm{t})$ is the number of sample points in $\mathrm{t}$, and $\widehat{\Delta}(X_j, Y | \mathbf{X} \in \mathrm{t})$ is the largest reduction in impurity for samples in $\mathrm{t}$. Note that if $T$ is a decision stump with split along $X_j$, then (7) equals $\widehat{\Delta}(X_j, Y)$, the largest reduction in impurity at the root node. In that sense, SDI is perhaps most akin to MDI for gradient tree boosting in which each base learner is a shallow tree.

# 3 Preliminaries

Our first lemma, developed by the first author in recent work, reveals the crucial role that optimization (of a nonlinear model) plays in assessing whether a particular variable is relevant or irrelevant—by relating the impurity reduction for a particular variable $X_j$ to the sample correlation between the response variable $Y$ and *any* function of $X_j$. This lemma also highlights a key departure from other approaches in past decision tree literature that do not consider splits that depend on *both* input and output data (see, for example, DSTUMP [23]).

In order to state the lemma, we will need to introduce the concept of stationary intervals. We define a *stationary interval* of a univariate function $h(\cdot)$ to be a maximal interval $I$ such that $h(I) = c$, where $c$ is a local extremum of $h(\cdot)$ ($I$ is maximal in the sense that there does not exist an interval $I'$ such that $I \subset I'$ and $h(I') = c$). In particular, note that a monotone function does not have any stationary intervals.

**Lemma 1** (Lemma A.4, Supplementary Material in [25])**.** *Almost surely, uniformly over all functions $h(\cdot)$ of $X_j$ that have at most $M$ stationary intervals, we have*

$$\widehat{\Delta}(X_j, Y) \geq \frac{1}{D^{-1} M n + \log(2n) + 1} \times \widehat{\mathrm{Cov}}^2 \left( \frac{h(X_j)}{\sqrt{\widehat{\mathrm{Var}}(h(X_j))}}, Y \right), \tag{8}$$

*where $D \geq 1$ is the smallest number of data points in a stationary interval of $h(\cdot)$ that contains at least one data point.*[4]

[4]More precisely, if $I_1, \ldots, I_M$ are the stationary intervals of $h(\cdot)$ and $D_k = \#\{X_{ij} \in I_k\}$, then $D = \min_k \{D_k : D_k \geq 1\}$.

**Remark 1.** *Note that $M$ can also be thought of as the number of times $h(\cdot)$ changes from strictly increasing to decreasing (or vice versa).*

*Proof sketch of Lemma 1.* For self-containment, we sketch the proof when $h(\cdot)$ is differentiable. The essential idea is to construct an empirical prior $\Pi$ on the split points $z$ and lower bound $\widehat{\Delta}(X_j, Y)$ by

$$\int \widehat{\Delta}(z; X_j, Y) d\Pi(z).$$

Recall from Section 2.4 that $N_L = N_L(z)$ and $N_R = N_R(z)$ are the number of samples in the left and right daughter nodes, respectively, if the $j^{\text{th}}$ variable is split at $z$. The special prior we choose has density

$$\frac{d\Pi(z)}{dz} = \frac{|h'(z)|\sqrt{N_L(z)N_R(z)}}{\int |h'(z')|\sqrt{N_L(z')N_R(z')}dz'},$$

with support between the minimum and maximum values of the data $\{X_{ij}\}$. This then yields $\widehat{\Delta}(X_j, Y) \geq C(h) \times \widehat{\text{Cov}}^2\left(\frac{h(X_j)}{\sqrt{\widehat{\text{Var}}(h(X_j))}}, Y\right)$. The factor $C(h)$ can be minimized (by solving a simple quadratic program) over all functions $h(\cdot)$ under the constraint that they have at most $M$ stationary intervals containing at least $D$ data points, yielding the desired result (8). We direct the reader to [25, Lemma A.4, Supplementary Material] for the full proof. □

Ignoring the factor $(D^{-1}Mn + \log(2n) + 1)^{-1}$ in (8) and focusing only on the squared sample covariance term, note that choosing $h(\cdot)$ to be the marginal projection $f_j(\cdot)$, we have

$$\widehat{\text{Cov}}^2\left(\frac{f_j(X_j)}{\sqrt{\widehat{\text{Var}}(f_j(X_j))}}, Y\right) \approx \text{Cov}^2\left(\frac{f_j(X_j)}{\sqrt{\text{Var}(f_j(X_j))}}, Y\right) = \text{Var}(f_j(X_j)),$$

where the last equality can be deduced from the fact that the marginal projection $f_j(X_j)$ is orthogonal to the residual $Y - f_j(X_j)$. Thus, in an ideal setting, Lemma 1 enables us to asymptotically lower bound $\widehat{\Delta}(X_j, Y)$ by a multiple of the variance of the marginal projections—which can then be used to screen for important variables.

To summarize, the previous lemma shows that $\widehat{\Delta}(X_j, Y)$ is large for variables $X_j$ such that $f_j(X_j)$ is strongly correlated with $Y$—or equivalently—variables that have large signals in terms of the variance of the marginal projection. Conversely, our next lemma shows that $\widehat{\Delta}(X_j, Y)$ is with high probability not greater than the variance of the marginal projection. A special instance of this lemma, namely, when $Y$ is independent of $X_j$, was stated in [28, Lemma 1] and serves as the inspiration for our proof. Due to space constraints, we include the proof in Appendix A.

**Lemma 2.** *Suppose that $Y$ is conditionally sub-Gaussian with variance parameter $\sigma_Y^2$, i.e., $\mathbb{E}[\exp(\lambda Y)|\mathbf{X}] \leq \exp(\lambda^2\sigma_Y^2/2)$ for all $\lambda \in \mathbb{R}$. With probability at least $1-4n\exp(-n\xi^2/(12\sigma_Y^2))$,*

$$\widehat{\Delta}(X_j, Y) \leq 3\widehat{\mathrm{Var}}(f_j(X_j)) + \xi^2.$$

In other words, Lemmas 1 and 2 together imply that SDI is a proxy for the variance of the marginal projection and therefore it roughly ranks the variables accordingly, up to constant factors. These lemmas are a key ingredient of the proofs for model selection and may be of independent interest.

# 4 SDI for 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 (6) 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 $X_j$, behaves roughly like the correlation between a linear model $Y$ and a coordinate $X_j$.

**Theorem 1** (SDI is asymptotically equivalent to 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 $1 - 4n\exp(-n\delta^2/12) - 2\exp(-(n-1)/16)$,*

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

*Proof sketch of Theorem 1.* We only sketch the proof due to space constraints, but a more complete version is provided in Appendix B.

The first step in proving the lower bound (9) is to apply Lemma 1 with $h(X_j) = X_j$ (a monotone function) to see that

$$\widehat{\Delta}(X_j, Y) \geq \frac{\widehat{\mathrm{Var}}(Y)}{\log(2n)+1} \times \widehat{\rho}^{\,2}(X_j, Y), \tag{11}$$

since $M = 0$. Next, we can apply asymptotic tail bounds for Pearson's sample correlation coefficient $\widehat{\rho}\,(X_j, Y)$ between two correlated Gaussian distributions [17] to show that with

high probability, $|\widehat{\rho}(X_j, Y)| \geq (1-\delta)|\rho(X_j, Y)|$. Finally, we divide (11) by $\widehat{\mathrm{Var}}(Y)$, use (6), and take square roots to complete the proof of the high probability lower bound (9).

To prove the upper bound (10), notice that since $X_j$ and $Y$ are jointly Gaussian with mean zero, we have $f_j(X_j) = \rho_j \frac{\sigma_Y}{\sigma_{X_j}} X_j$, where $\rho_j = \rho(X_j, Y)$. Thus, by Lemma 2 with $\xi^2 = \sigma_Y^2 \delta^2$, with probability at least $1 - 4n\exp(-n\delta^2/12)$,

$$\widehat{\Delta}(X_j, Y) \leq 3\widehat{\mathrm{Var}}(f_j(X_j)) + \sigma_Y^2 \delta^2 = 3\rho_j^2 \frac{\sigma_Y^2}{\sigma_{X_j}^2} \widehat{\mathrm{Var}}(X_j) + \delta^2 \sigma_Y^2. \tag{12}$$

We further upper bound (12) by obtaining high probability upper and lower bounds, respectively, for $\widehat{\mathrm{Var}}(X_j)$ and $\widehat{\mathrm{Var}}(Y)$ in terms of $\sigma_{X_j}^2$ and $\sigma_Y^2$, with a standard chi-squared concentration bound, per the Gaussian assumption. This yields that with high probability,

$$\widehat{\Delta}(X_j, Y) \lesssim \rho_j^2 \widehat{\mathrm{Var}}(Y) + \delta^2 \widehat{\mathrm{Var}}(Y). \tag{13}$$

Finally, dividing both sides of (13) by $\widehat{\mathrm{Var}}(Y)$, using (6), and taking square roots proves (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 7.

One corollary of Theorem 1 is that, like SIS, SDI also enjoys the sure screening property, under the same assumptions as [11, Conditions 1-4], which include mild conditions on the eigenvalues of the design covariance matrices and minimum signals of the parameters $\beta_j$. Similarly, like SIS, SDI can also be paired with lower dimensional variable selection methods such as Lasso or SCAD [2] 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 the rigidity of 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.

# 5 SDI for 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 $n \to \infty$. We describe the assumptions needed in Section 5.1 and outline the main consistency results and their proofs in Section 5.3. Finally, we describe how to modify the main results for binary classification in Section 5.5.

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.

## 5.1 Assumptions

In this section, 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 [10, 11], but are weaker than most past work on tree based variable selection [28, 23].

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

**Assumption 2** (Smoothness of marginal projections)**.** *Let $r$ be a positive integer, let $0 < \alpha \leq 1$ be such that $d := r + \alpha$ is at least $5/8$, and let $0 < L < \infty$. The $r^{\text{th}}$ order derivative of $f_j(\cdot)$ exists and is $L$-Lipschitz of order $\alpha$, 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** (Monotonicity of marginal projections)**.** *The marginal projection $f_j(\cdot)$ is monotone on $\mathbb{R}$.*

**Assumption 4** (Partial orthogonality of predictor variables)**.** *The collections $\{X_j\}_{j \in \mathcal{S}}$ and $\{X_j\}_{j \in \mathcal{S}^c}$ are independent of each other.*

**Assumption 5** (Uniform marginals of relevant variables)**.** *The marginal distribution of each $X_j$, for $j \in \mathcal{S}$, is uniform on the unit interval.*

**Assumption 6** (Sub-Gaussian error distribution)**.** *The error distribution is conditionally sub-Gaussian, 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 $\sigma^2 > 0$.*

## 5.2 Discussion of the assumptions

Assumption 2 is a standard smoothness assumption for variable selection in nonparametric additive models [10, Assumption A] and [18, Section 3] except that, for technical reasons, we have the condition $d \geq 5/8$ instead of $d > 1/2$. 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. Alternatively, instead of Assumption 2, as we shall see, stronger conclusions can be provided if we impose a monotonicity constraint, namely, Assumption 3. Note that this monotonicity assumption encompasses many important "shape constrained" statistical models such as linear or isotonic regression.

Assumption 4 is essentially the so-called "partial orthogonality" condition in marginal screening methods [12]. Importantly, it allows for correlation between the relevant variables $\{X_j\}_{j\in\mathcal{S}}$, unlike previous works on tree based variable selection [23, 28]. 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.

Assumption 5 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 [10, 18]. In fact, even these assumptions are not required. If the marginal projection is monotone (i.e., Assumption 3), *no marginal distributional assumptions are required*, that is, each $X_j$ for $j \in \mathcal{S}$ could be continuous, discrete, have unbounded support, or have a density that vanishes or is unbounded. More generally, similar 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 $[0, 1]$. See Remark 2 for details.

**Remark 2.** *Let $q_j(\cdot)$ and $F_j(\cdot)$ be the quantile function and the cumulative distribution function of $X_j$, respectively. Recall the Galois inequalities state that $q_j(w) \leq z$ if and only if $w \leq F_j(z)$ and furthermore that $q_j(F_j(X_j)) = X_j$ almost surely. Then, choosing $w = F_j(X_j)$ we see that, almost surely, $X_j \leq z$ if and only if $F_j(X_j) \leq F_j(z)$. Therefore, almost surely,*

$$\max_z \widehat{\Delta}(z; X_j, Y) = \max_z \widehat{\Delta}(F_j(z); F_j(X_j), Y) = \max_{w\in[0,1]} \widehat{\Delta}(w; F_j(X_j), Y).$$

*This means that for continuous data, the problem can be reduced to the uniform case by pre-applying the marginal cumulative distribution function $F_j(\cdot)$ to each variable, since $F_j(X_j) \sim \text{Uniform}([0, 1])$. Note that the marginal projections now equal the composition of the original $f_j(\cdot)$ with $q_j(\cdot)$, i.e., $(f_j \circ q_j)(\cdot)$. By the chain rule from calculus, if $q_j(\cdot)$ satisfies Assumption 2, then so does $(f_j \circ q_j)(\cdot)$.*

## 5.3 Theory for variable ranking and model selection

Here our goal will be to provide variable ranking and model selection guarantees of SDI using the assumptions in Section 5.1. Again, in this section, we sketch the proofs, but the full versions can be found in Appendices D and E.

### 5.3.1 Preliminary results

The high level idea will be to show that the impurity reductions for relevant variables dominate those for irrelevant variables with high probability, meaning that relevant variable are correctly ranked.

The following two propositions provide high probability lower bounds on the impurity reduction for relevant variables, the size of which depend on whether we assume Assumption 2 or Assumption 3.

Our first result deals with general, smooth marginal projections.

**Proposition 1.** *Under Assumptions 1, 2, 5, and 6, with probability at least* $1 - (4n + 2)\exp\big(-nC_1\mathrm{Var}(f_j(X_j))\big)$*, we have*

$$\widehat{\Delta}(X_j, Y) \geq \frac{C_2(\mathrm{Var}(f_j(X_j)))^{6/5+1/d}}{\log(n)},$$

*for some positive constants* $C_1$ *and* $C_2$ *which depend only on* $B$*,* $\sigma$*,* $r$*, and* $\alpha$*.*

Next, we state an analogous bound, but under a slightly different assumption on the marginal projection. It turns out that if the marginal projection is monotone, we can obtain a stronger result.

**Proposition 2.** *Under Assumptions 1, 3, and 6, with probability at least* $1 - 2\exp\Big(-\frac{(n-1)\mathrm{Var}(f_j(X_j))}{32(B^2+\sigma^2)}\Big)$*,*

$$\widehat{\Delta}(X_j, Y) \geq \frac{\mathrm{Var}(f_j(X_j))}{16(1 + \log(2n))}.$$

*Proof sketch of Propositions 1 and 2.* We sketch the proof of Proposition 1; the proof of Proposition 2 is based on similar arguments. The main idea is to apply Lemma 1 with $h(\cdot)$ equal to a modified polynomial approximation $\widetilde{f}_j(\cdot)$ to $f_j(\cdot)$. This is done to temper the effect of the factor $(D^{-1}Mn + \log(2n) + 1)^{-1}$ from Lemma 1, by controlling $M$ and $D$ individually.

To construct such a function, we first employ a Jackson-type estimate [22] in conjunction with Assumption 2 and Bernstein's theorem for polynomials [21] to show the existence of a good polynomial approximation $P_M(\cdot)$ (of degree $M + 1$) to $f_j(\cdot)$. We then construct $\widetilde{f}_j(\cdot)$ by redefining $P_M(\cdot)$ to be constant in a small neighborhood around each of its local

extrema, which ensures that each resulting stationary interval of $\widetilde{f}_j(\cdot)$ has a sufficiently large length. Since $P_M(\cdot)$ has at most $M$ local extrema, the number of stationary intervals of $\widetilde{f}_j(\cdot)$ will also be at most $M$.

Next, we use concentration of measure to ensure that each stationary interval of $\widetilde{f}_j(\cdot)$ is saturated with enough data, effectively providing a lower bound on $D$. Executing this argument reveals that valid choices of $D$ and $M$ (which come from optimizing a bound) are:

$$D \gtrsim n \times (\mathrm{Var}(f_j(X_j)))^{1/5+1/(2d)}, \quad M \lesssim (\mathrm{Var}(f_j(X_j)))^{-1/(2d)}.$$

Plugging these values into the lower bound (8) in Lemma 1, we find that with high probability,

$$\begin{aligned} \widehat{\Delta}(X_j, Y) &\geq \frac{1}{D^{-1}Mn + \log(2n) + 1} \times \widehat{\mathrm{Cov}}^2\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y\right) \\ &\gtrsim \frac{(\mathrm{Var}(f_j(X_j)))^{1/5+1/d}}{\log(n)} \times \widehat{\mathrm{Cov}}^2\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y\right). \end{aligned} \tag{14}$$

Thus, the lower bound in Proposition 1 will follow if we can show that the squared sample covariance factor in (14) exceeds $\mathrm{Var}(f_j(X_j))$ with high probability. To this end, note that

$$\begin{aligned} \widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y\right) = &\underbrace{\widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, f_j(X_j)\right)}_{\text{(I)}} \\ &+ \underbrace{\widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y - f_j(X_j)\right)}_{\text{(II)}}. \end{aligned} \tag{15}$$

With high probability, (I) can be lower bounded by a multiple of $\sqrt{\mathrm{Var}(f_j(X_j))}$, using the approximation properties of $\widetilde{f}_j(\cdot)$ for $f_j(\cdot)$ and a concentration inequality for the sample variance of $f_j(X_j)$. Furthermore, since $Y - f_j(X_j)$ has conditional mean zero, a Hoeffding type concentration inequality shows that, with high probability, (II) is larger than any (strictly) negative constant, including a multiple of $-\sqrt{\mathrm{Var}(f_j(X_j))}$. Combining this analysis from (14) and (15), we obtain the high probability lower bound on $\widehat{\Delta}(X_j, Y)$ given in Proposition 1. □

Next, we need to ensure that there is a sufficient separation in the impurity reductions between relevant and irrelevant variables. To do so, we use Lemma 2 along with the partial orthogonality assumption in Section 5.1 to show that the impurity reductions for irrelevant variables will be small with high probability.

**Lemma 3.** *Under Assumptions 1, 4 and 6, for each $j \in \mathcal{S}^c$, with probability at least $1 - 4n\exp(-n\xi^2/(12(B^2+\sigma^2)))$*

$$\widehat{\Delta}(X_j, Y) \leq \xi^2.$$

*In other words, if $j \in \mathcal{S}^c$, then $\widehat{\Delta}(X_j, Y) = \mathcal{O}(n^{-1}\log(n))$ with probability at least $1 - n^{-\Omega(1)}$.*

*Proof of Lemma 3.* Under Assumptions 1 and 6, observe that

$$\begin{aligned}
\mathbb{E}[\exp(\lambda Y)|\mathbf{X}] &= \mathbb{E}\big[\exp(\lambda(g(\mathbf{X}) + \varepsilon))\big|\mathbf{X}\big] \\
&= \exp(\lambda g(\mathbf{X}))\mathbb{E}\big[\exp(\lambda\varepsilon)\big|\mathbf{X}\big] \\
&\leq \mathbb{E}\big[\exp(\lambda g(\mathbf{X}))\big]\exp(\lambda^2\sigma^2/2) \\
&\leq \mathbb{E}\big[\exp(\lambda^2(B^2+\sigma^2)/2)\big], \qquad (16)
\end{aligned}$$

where we used Hoeffding's Lemma in the last inequality (16). Using Assumption 4 along with Lemma 2 with $\sigma_Y^2 = B^2 + \sigma^2$ proves Lemma 3. □

### 5.3.2 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 $s$ variables. Alternatively, if $s$ is unknown, we instead choose an asymptotic threshold $\gamma_n$ 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 the minimum signal strength of the relevant variables:

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

which is the same as the minimum variance parameter in independence screening papers (e.g., [10, Assumption C]). Note that $v$ measures the minimum contribution of each relevant variable alone to the variance in $Y$, ignoring the effects of the other variables.

**Theorem 2.** *Suppose Assumptions 1, 2, 4, 5, and 6 hold. Then the top $s$ variables ranked by Algorithm 1 equal the correct set $\mathcal{S}$ of relevant variables with probability at least*

$$1 - s(4n+2)\exp(-C_1 nv) - 4n(p-s)\exp\Big(-\frac{nC_2 v^{6/5+1/d}}{24\log(n)(B^2+\sigma^2)}\Big), \qquad (17)$$

*where $C_1$ and $C_2$ are the same constants in Proposition 1.*

As mentioned earlier, stronger results can be obtained if we assume that the marginal projections are monotone. In our next theorem, notice that we do not have to make any additional smoothness assumptions, nor do we require any distributional assumptions on the relevant variables, other than that they are independent of the irrelevant ones.

**Theorem 3.** *Suppose Assumptions 1, 3, 4, and 6 hold. Then the top $s$ variables ranked by Algorithm 1 equal the correct set $\mathcal{S}$ of relevant variables with probability at least*

$$1 - 2s \exp\Big( - \frac{(n-1)v}{32(B^2+\sigma^2)} \Big) - 4n(p-s) \exp\Big( - \frac{nv}{96(1+\log(2n))(B^2+\sigma^2)} \Big). \quad (18)$$

**Remark 3.** *When $s$ is unknown, Propositions 1 and 2 and Lemma 3 imply that the oracle threshold choices*

$$\gamma_n = \frac{C_2 v^{6/5+1/d}}{2\log(n)} \quad \text{and} \quad \frac{v}{8(1+\log(2n))} \quad (19)$$

*ensure that*

$$\max_{j \in \mathcal{S}^c} \widehat{\Delta}(X_j, Y) < \gamma_n \leq \min_{j \in \mathcal{S}} \widehat{\Delta}(X_j, Y)$$

*and hence will yield the same high probability bounds (17) and (18), respectively. Thus, while the permutation and elbow methods from Section 2.5 are somewhat ad-hoc, if they, at the very least, produce thresholds that are close to (19), then high probability performance guarantees are still possible to obtain.*

## 5.4 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, $v = \min_{j \in \mathcal{S}} \mathrm{Var}(g_j(X_j))$, which will always be strictly greater than zero. As Theorems 2 and 3 show, $v$ 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

$$v \gtrsim \Big( \frac{\log(n)\log(n(p-s))}{n} \Big)^{\frac{5d}{6d+5}}, \quad (20)$$

up to constants that depend on $B$, $\sigma$, $r$, and $\alpha$. That is, (20) is a sufficient condition on the signal of all relevant variables so that $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$ as $n \to \infty$. Similarly, we see from Theorem 3 that

$$v \gtrsim \frac{\log(n)\log(n(p-s))}{n} \quad (21)$$

is sufficient to guarantee model selection consistency for monotone marginal projections. A particularly striking aspect of (21) is that the rate is independent of the smoothness of the marginal projection. This means that the "difficulty" of detecting a signal from a general monotone marginal projection is essentially no more than if the marginal projection was linear.

### 5.5 SDI for binary classification

Our theory for SDI can also naturally be extended to the context of classification, as we now describe. For simplicity, we focus on the problem of binary classification where $Y \in \{0, 1\}$. To begin, we first observe that Gini impurity and variance impurity are equivalent up to a factor of two [29, Section 3]. Thus, we can use the same criterion $\Delta(X_j, Y)$ to rank the variables. Next, we identify the marginal projection as

$$f_j(X_j) = \mathbb{E}[Y|X_j] = \mathbb{P}(Y = 1|X_j).$$

Because of these connections to the regression setting, the results in Section 5.3 hold verbatim for classification under the same assumptions therein with $v = \min_{j \in \mathcal{S}} \mathrm{Var}(\mathbb{P}(Y = 1|X_j))$, $B = 1$, and $\sigma^2 = 0$. An interesting special case arises when $\mathbb{P}(Y = 1|\mathbf{X}) = h(\beta_1 X_1 + \cdots \beta_p X_p)$ for some monotone link function $h(\cdot)$ and $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ for some positive semi-definite matrix $\mathbf{\Sigma}$. Using the fact that Gaussian random vectors are conditionally Gaussian distributed, it can be shown that the marginal projection $\mathbb{P}(Y = 1|X_j)$ is monotone and therefore satisfies Assumption 3. Consequently, for example, logistic regression with Gaussian data inherits the same stronger conclusions as Theorem 3.

## 6 Comparing SDI with other model selection methods

In this section, we compare the finite sample guarantees of SDI given in Section 5.3 and Section 4 to those of NIS, SPAM, and Lasso. 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 [10] to uncover the corresponding threshold $v$ for NIS. In order to have model selection consistency, the probability bound in [10, Theorem 2] must approach one as $n \to \infty$, which necessitates

$$d_n \lesssim \left(\frac{n^{1-4\kappa}}{\log(np)}\right)^{1/3}, \tag{22}$$

where $d_n$ is the number of spline basis functions and $\kappa$ is a free parameter (in the notation of [10]). However notice that by [10, Assumption F], we must also have that $d_n \gtrsim n^{\frac{2\kappa}{2d+1}}$, and combining this with (22) shows that we must have

$$n^{-\kappa} \gtrsim \left(\frac{\log(np)}{n}\right)^{\frac{2d+1}{8d+10}}. \tag{23}$$

Now substituting (22) and (23) into $v \gtrsim d_n n^{-2\kappa}$ ([10, Assumption C]), it follows that we have

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

for NIS, which is a larger minimum signal than our (20).

***Minimum signal strength for SPAM*** In the case where $d = 2$, by [34, 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 $v \gtrsim \left(\frac{\log(n)\log(np)}{n}\right)^{10/17}$, which is obtained by setting $d = 2$ in (20).

***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 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 Appendix C. 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

$$n \gg s\log(p-s) \times (\log(s) + \log\log(p-s)). \tag{24}$$

Now, it is shown in [38, 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 [39]. Amazingly, (24) 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 and 3 show 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 [10, 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 $p = \exp(o(n^{1/6}))$ [34, Equation (45)], which is again lower than the dimensionality $p = \exp(o(n))$ for SDI.

# 7 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 [23, 28, 29, 30, 31, 35, 40, 41], we do not aim for comprehensiveness.

Our first set of experiments compare the performance of SDI with SPAM, SIS, and Lasso. In Section 7.1, we assess performance based on *partial recovery* (proportion of the true support recovered), while in Section 7.2, we instead assess performance based on *exact recovery* (probability that the support is recovered exactly). To ensure a fair comparison between SDI and the other variable selection algorithms, we assume a priori knowledge of the true sparsity level $s$, and we incorporate $s$ into Lasso and SPAM by specifying the model degrees of freedom in advance. These simulations are similar to those in [23] and [34], and were conducted in R using the packages `rpart` for SDI, `SAM` for SPAM, `SIS` for SIS, and `glmnet` for Lasso with default settings. Finally, our second set of experiments, in Section 7.3, deals with the case when the sparsity level $s$ is unknown, whereby we demonstrate the empirical performance of the permutation and elbow methods from Section 2.5.

In all our experiments, we generate $n$ samples from an $s$-sparse additive model $g(\mathbf{X}) = (1/\sqrt{s}) \sum_{j=1}^{s} g_j(X_j)$ for various types of components $g_j(X_j)$. Notice that for convenience, we slightly abuse notation and leave out the $1/\sqrt{s}$ factor in the $g_j(X_j)$ notation of (1). The error distribution is $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with standard deviation $\sigma = 0.1$ and the ambient dimension is fixed at $p = 200$. We consider the following two model types.

**Model 1** Assume $g_j(X_j) = X_j$ and $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, 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 consider $\rho = 0$ (*Model 1a*) or $\rho = 0.5$ (*Model 1b*).

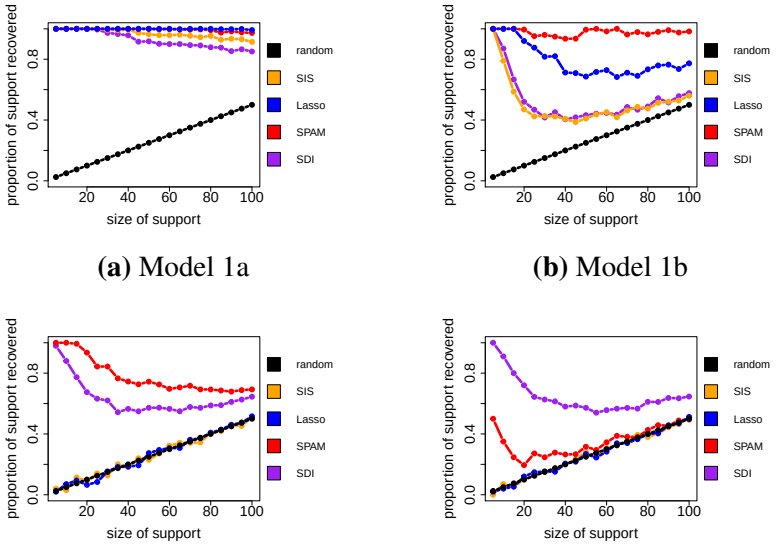
**Model 2** Assume $\mathbf{X} \sim \text{Uniform}([0, 1]^p)$, where all predictor variables are independent. We consider nonlinear additive components $g_j(X_j) = (X_j - 1/2)^2$ (*Model 2a*) or $g_j(X_j) = \cos(4\pi X_j)$ (*Model 2b*).

In our simulations, Models 1a and 1b test the Gaussian linear model setting of Theorem 1 with and without correlation, respectively. Models 2a and 2b are examples of 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.

## 7.1 Partial recovery

For our experiment on partial recovery, we fix $n = 1000$ and compute the proportion of relevant variables selected (averaged over 10 independent replications) at various sparsity levels, namely, $s \in \{5, 10, 15, \ldots, 100\}$. In Figure 1, we plot the proportion of the support

recovered against the true sparsity level for all model types. In agreement with Theorem 1, in Figures 1a and 1b, we observe that SDI and SIS exhibit similar behavior for linear models across varying levels of correlation among the predictor variables. Furthermore, as we can see from Figures 1c and 1d, SDI and SPAM both appear to significantly outperform SIS and Lasso when the underlying model is nonlinear. For component functions that are more irregular, such as a sinusoid in Figure 1d, SDI performs even better than SPAM.



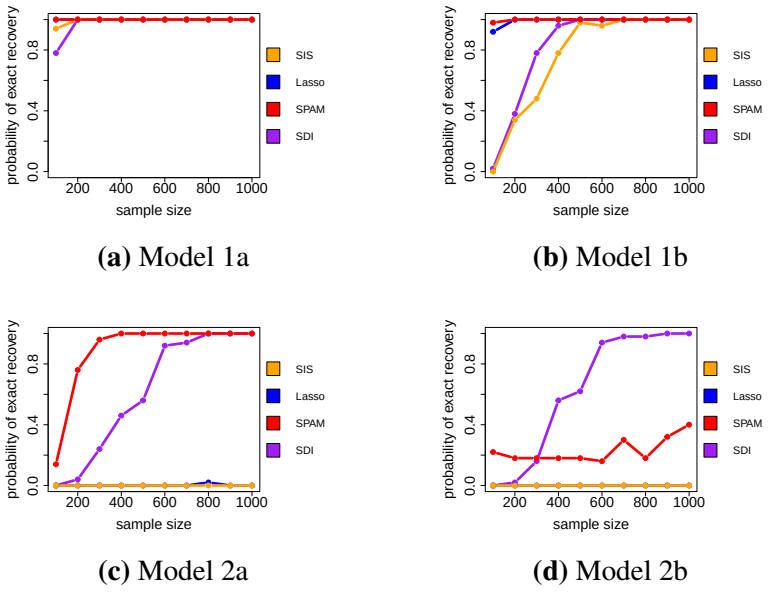

**(a)** Model 1a **(b)** Model 1b

**(c)** Model 2a **(d)** Model 2b

**Figure 1:** Plots of the proportion of the support selected as the sparsity level varies. All methods are compared to the baseline random selection proportion, which is $s/p$.

## 7.2 Exact recovery

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 of the model. In Figure 2, we plot this estimated probability against various sample sizes, namely, $n \in \{100, 200, 300, \ldots, 1000\}$. Again, as expected, Figure 2 shows that SDI and SPAM appear to outperform SIS and Lasso unless the model is linear, in which case all methods appear to achieve model selection consistency. For more irregular component functions such as sinusoids, SDI again appears to outperform even SPAM, as seen in Figure 2d. In summary, in agreement with Theorems 2 and 3, Figure 2 demonstrates the robustness of SDI across various additive models.




**(a)** Model 1a

**(b)** Model 1b

**(c)** Model 2a

**(d)** Model 2b

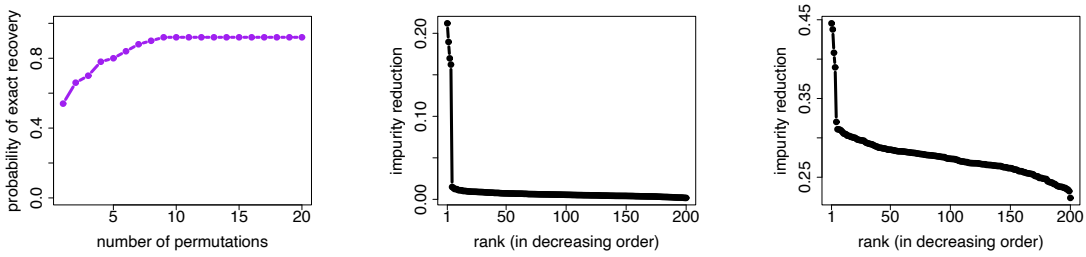
**Figure 2:** Plots of the probability of exact recovery as the sample size varies.

## 7.3 Data driven choices of $\gamma_n$

We now consider the case when $s$ is unknown. As we have already demonstrated the performance of SDI in Sections 7.1 and 7.2, here we examine how well the data-driven thresholding methods discussed in Section 2.5 can estimate the true sparsity level $s$.

We first consider Model 1a and fix the sparsity level $s = 4$ and the sample size $n = 1000$. In Figure 3a, we plot the probability of exact support recovery (averaged over 50 independent replications) using the permutation method against the number of permutations. After a very small number of permutations, the significance level and performance of the algorithm appear to stabilize. For comparison, in Figure 3b, we show the graph of the ranked impurity reductions (SDI importance measures) used for the elbow method. When there is no correlation between irrelevant and relevant variables, as is the case with Model 1a, the permutation method may be more precise than the elbow method.

However, as discussed in Section 2.5, the permutation method may be inaccurate if there is correlation between irrelevant and relevant variables. To illustrate this, we now consider Model 1b with sparsity level $s = 4$. Using 10 permutations, the permutation method chooses a threshold $\gamma_n = 0.021$, which will lead to all variables being selected, as can be seen from ranked impurity reductions shown in Figure 3c. In other words, the permutation method underestimates the threshold $\gamma_n$, which in turn, creates too many false positives. In contrast, as can be seen from Figure 3c, the ranked impurity reductions still exhibit a distinct "elbow", from which the relevant and irrelevant variables can be discerned.

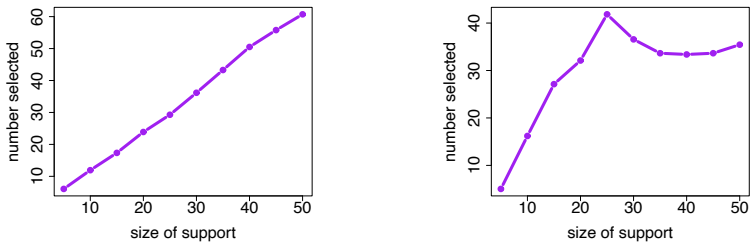



**(a)** Permutation method **(b)** Elbow method **(c)** Elbow method with correlation

**Figure 3:** The probability of exact recovery by the number of random permutations used in the permutation method are shown in Figure 3a. Figure 3b shows plots of the corresponding ranked impurity reductions used for the elbow method. When the relevant and irrelevant variables are correlated, the ranked impurity reductions used for the elbow method are shown in Figure 3c.

As seen from Figures 3b and 3c, the impurity reductions for irrelevant variables appear to be tightly clustered at lower values, while the relevant variables tend to be spread out at higher values.

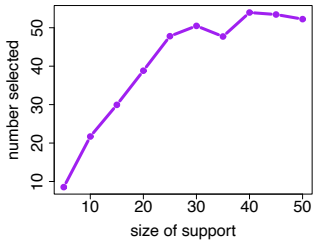
Since the irrelevant and relevant variable clusters have very different shapes and imbalanced sizes, we select the relevant variable cluster using a two-component Gaussian mixture model with the R package `EMCluster`. For our last experiment, we fix the sample size $n$ at 1000 and consider various sparsity levels, namely, $s \in \{5, 10, 15, \ldots, 50\}$ for all models. We run 50 independent trials and average the size of the selected cluster. From Figure 4, we find that the average number of variables selected is close to the true sparsity level $s$ for all models, though the performance appears to be worse when there is correlation between variables.



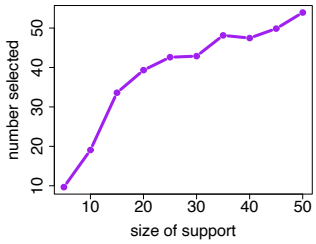

**(a)** Model 1a




**(b)** Model 1b




**(c)** Model 2a




**(d)** Model 2b

**Figure 4:** Plots of the number of variables selected using a two-component Gaussian mixture model against the true sparsity level.

# 8 Discussion and conclusion

In this paper, we developed a theoretically rigorous approach for variable selection based on decision trees. The underlying approach is simple, intuitive, and interpretable—we test whether a variable is relevant/irrelevant by fitting a decision stump to that variable and then determining how much it explains the variance of the response variable. Despite its simplicity, SDI performs favorably relative to its less interpretable competitors. Furthermore, due to the parsimony of the model, there is also no need to perform variable bandwidth selection or calibrate the number terms in basis expansions.

On the other hand, we have sacrificed generality for analytical tractability. That is, decision stumps are poor at capturing interaction effects and, therefore, an importance measure built from a multi-level decision tree (such as MDI) may be more appropriate for models with more than just main effects. However, the presence of multi-level splits adds an additional layer of complexity to the analysis that, at the moment, we do not know how to overcome. It is also unclear how to leverage these additional splits to strengthen the theory. While out of the scope of the present paper, we view this as an important problem for future investigation. It is our hope that the tools in this paper can be used by other scholars to address this important issue.

To conclude, we hope that our analysis of single-level decision trees for variable selection will shed further light on the unique benefits of tree structured learning.

# 9 Appendix

# A Appendix

In this appendix, we first prove Lemma 2 in detail in Appendix A. We then prove Theorem 1 on linear models with Gaussian variates in Appendix B and then use Theorem 1 to determine a sufficient sample size for model selection consistency (mentioned in Section 6) in Appendix C. Finally, we prove Propositions 1 and 2 in Appendix D and then prove Theorems 2 and 3 in Appendix E.

# A Proof of Lemma 2

Let $\pi$ be a permutation of the data such that $X_{\pi(1)j} \leq X_{\pi(2)j} \leq \cdots \leq X_{\pi(n)j}$. Recall from the representation (5) that we have

$$\widehat{\Delta}(X_j, Y) = \max_{1 \leq k \leq n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\Bigg(\underbrace{\frac{1}{k}\sum_{X_{ij} \leq X_{\pi(k)j}} Y_i - \frac{1}{n-k}\sum_{X_{ij} > X_{\pi(k)j}} Y_i}_{\text{(III)}}\Bigg)^2.$$

Now, since

$$\sum_{i=1}^{n}\left(\frac{\mathbf{1}(X_{ij} \leq X_{\pi(k)j})}{k} - \frac{\mathbf{1}(X_{ij} > X_{\pi(k)j})}{n-k}\right) = 0,$$

we can rewrite (III) as

$$\underbrace{\frac{1}{k}\sum_{X_{ij} \leq X_{\pi(k)j}} (Y_i - \mathbb{E}[Y|X_{ij}])}_{\text{(a)}} - \underbrace{\frac{1}{n-k}\sum_{X_{ij} > X_{\pi(k)j}} (Y_i - \mathbb{E}[Y|X_{ij}])}_{\text{(b)}}$$
$$+ \underbrace{\sum_{i=1}^{n}\left(\mathbb{E}[Y|X_{ij}] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y|X_{ij}]\right)\left(\frac{\mathbf{1}(X_{ij} \leq X_{\pi(k)j})}{k} - \frac{\mathbf{1}(X_{ij} > X_{\pi(k)j})}{n-k}\right)}_{\text{(c)}}.$$

Therefore, we have that

$$\begin{aligned}\widehat{\Delta}(X_j, Y) &= \max_{1 \leq k \leq n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)(\text{(a)} - \text{(b)} + \text{(c)})^2 \\ &\leq 3 \max_{1 \leq k \leq n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\text{(a)}^2 + 3 \max_{1 \leq k \leq n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\text{(b)}^2 + \\ &\quad 3 \max_{1 \leq k \leq n} \frac{k}{n}\Big(1 - \frac{k}{n}\Big)\text{(c)}^2,\end{aligned} \tag{A.1}$$

where we use, in succession, the inequality $(x-y+z)^2 \leq 3(x^2+y^2+z^2)$ for any real numbers $x$, $y$, and $z$, and the fact that the maximum of a sum is at most the sum of the maxima. To finish the proof, we will bound the terms involving $(\text{a})^2$, $(\text{b})^2$, and $(\text{c})^2$ separately.

For the last term in (A.1), notice that by the Cauchy-Schwartz inequality we have

$$\begin{aligned}
\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\text{c})^2 &= \frac{k}{n}\Big(1-\frac{k}{n}\Big)\Bigg[\sum_{i=1}^n\Big(\mathbb{E}[Y|X_{ij}]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}[Y|X_{ij}]\Big)\times \\
&\qquad\qquad \Big(\frac{\mathbf{1}(X_{ij}\leq X_{\pi(k)j})}{k}-\frac{\mathbf{1}(X_{ij}>X_{\pi(k)j})}{n-k}\Big)\Bigg]^2 \\
&\leq \frac{k}{n}\Big(1-\frac{k}{n}\Big)\sum_{i=1}^n\Big(\mathbb{E}[Y|X_{ij}]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}[Y|X_{ij}]\Big)^2\times \\
&\qquad\qquad \sum_{i=1}^n\Big(\frac{\mathbf{1}(X_{ij}\leq X_{\pi(k)j})}{k}-\frac{\mathbf{1}(X_{ij}>X_{\pi(k)j})}{n-k}\Big)^2,
\end{aligned}$$

which is exactly equal to

$$\frac{k}{n}\Big(1-\frac{k}{n}\Big)\Big[n\widehat{\text{Var}}(f_j(X_j))\Big(k\cdot\frac{1}{k^2}+(n-k)\cdot\frac{1}{(n-k)^2}\Big)\Big]=\widehat{\text{Var}}(f_j(X_j)).$$

Therefore we have shown that

$$\max_{1\leq k\leq n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\text{c})^2\leq\widehat{\text{Var}}(f_j(X_j)). \tag{A.2}$$

To bound the first term in (A.1), by a union bound we have that

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\leq k\leq n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\text{a})^2>\frac{\xi^2}{6}\Big) \\
&\leq\sum_{k=1}^n\mathbb{P}\Big(\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\text{a})^2>\frac{\xi^2}{6}\Big) \\
&=\sum_{k=1}^n\mathbb{P}\Big(\frac{k}{n}\Big(1-\frac{k}{n}\Big)\Big(\frac{1}{k}\sum_{X_{ij}\leq X_{\pi(k)j}}(Y_i-\mathbb{E}[Y|X_{ij}])\Big)^2>\frac{\xi^2}{6}\Big).
\end{aligned}\tag{A.3}$$

Next, notice that, conditional on $X_{1j},\ldots,X_{nj}$, $\sum_{X_{ij}\leq X_{\pi(k)j}}(Y_i-\mathbb{E}[Y|X_{ij}])$ is a sum of $k$ independent, sub-Gaussian, mean zero random variables. Thus, by the law of total probability, we have that (A.3) is equal to

$$\sum_{k=1}^n\mathbb{E}\Bigg[\mathbb{P}\Bigg(\Big|\frac{1}{k}\sum_{X_{ij}\leq X_{\pi(k)j}}(Y_i-\mathbb{E}[Y|X_{ij}])\Big|>\xi\sqrt{\frac{n^2}{6k(n-k)}}\quad\Big|\quad X_{1j},\ldots,X_{nj}\Bigg)\Bigg]$$

and, by Hoeffding's inequality for sub-Gaussian random variables, is bounded by

$$\sum_{k=1}^{n} 2\exp\Big(-k\frac{\xi^2 n^2}{12k(n-k)\sigma_Y^2}\Big) \le 2n\exp\Big(-\frac{\xi^2 n}{12\sigma_Y^2}\Big).$$

Note that here we have implicitly used the fact that $\mathbb{E}[\exp(\lambda Y)|\mathbf{X}] \le \exp(\lambda^2\sigma_Y^2/2)$ implies $\mathbb{E}[\exp(\lambda Y)|X_j] \le \exp(\lambda^2\sigma_Y^2/2)$, which is true by the law of iterated expectation. It thus follows that with probability at least $1-2n\exp\big(-\frac{\xi^2 n}{12\sigma_Y^2}\big)$ that

$$\max_{1\le k\le n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\mathrm{a})^2 \le \frac{\xi^2}{6}. \tag{A.4}$$

A similar argument shows that with probability at least $1-2n\exp\Big(-\frac{\xi^2 n}{12\sigma_Y^2}\Big)$, the second terms in (A.1) obeys

$$\max_{1\le k\le n}\frac{k}{n}\Big(1-\frac{k}{n}\Big)(\mathrm{b})^2 \le \frac{\xi^2}{6}. \tag{A.5}$$

Therefore, substituting (A.2), (A.4), and (A.5) into (A.1) and using a union bound, it follows that with probability at least $1-4n\exp\big(-\frac{\xi^2 n}{12\sigma_Y^2}\big)$,

$$\widehat{\Delta}(X_j, Y) \le 3\widehat{\mathrm{Var}}(f_j(X_j)) + \xi^2.$$

# B Proof of Theorem 1

The goal of this section is to prove Theorem 1. In the first section, we prove the lower bound (9) and in the second section, we prove the upper bound (10).

Throughout this section, for brevity, we let $\rho_j = \mathrm{Cor}(X_j, Y) \neq 0$.

## B.1 Proof of the lower bound (9)

Choosing $h(X_j) = X_j$ (which is monotone) in Lemma 1 to get that

$$\widehat{\Delta}(X_j, Y) \ge \frac{1}{\log(2n)+1} \times \widehat{\mathrm{Cov}}^2\left(\frac{X_j}{\sqrt{\widehat{\mathrm{Var}}(X_j)}}, Y\right) = \frac{\widehat{\mathrm{Var}}(Y)}{\log(2n)+1} \times \widehat{\rho}^{\,2}(X_j, Y).$$

Now observe that $\widehat{\rho}(X_j, Y)$ is the empirical Pearson sample correlation between two correlated normal distributions. If $\rho_j > 0$, by [17, Equation (44)], we have that

$$\begin{aligned}
&\mathbb{P}(\widehat{\rho}(X_j, Y) > (1-\delta)\rho_j) \\
&= 1 - \mathbb{P}(\widehat{\rho}(-X_j, Y) > -(1-\delta)\rho_j) \\
&\sim 1 - (2\pi)^{-1/2}\frac{\Gamma(n)}{\Gamma(n+1/2)}(1-\rho_j^2)^{n/2}(1-[(1-\delta)\rho_j]^2)^{(n-1)/2} \\
&\qquad \times (-(1-\delta)\rho_j - (-\rho_j))^{-1}(1-(-\rho_j)(-(1-\delta)\rho_j))^{-n+3/2}(1+\mathcal{O}(n^{-1})).
\end{aligned} \tag{B.6}$$

If $\rho_j < 0$ we can show the same bound on $\mathbb{P}(\widehat{\rho}\,(X_j, Y) < (1-\delta)\rho_j)$. Again by [17, Equation (44)], we have the similar bound

$$\begin{aligned}
&\mathbb{P}(\widehat{\rho}\,(X_j, Y) < (1-\delta)\rho_j) \\
&= 1 - \mathbb{P}(\widehat{\rho}\,(X_j, Y) > (1-\delta)\rho_j) \\
&\sim 1 - (2\pi)^{-1/2} \frac{\Gamma(n)}{\Gamma(n+1/2)} (1-\rho_j^2)^{n/2} (1 - [(1-\delta)\rho_j]^2)^{(n-1)/2} \\
&\qquad \times ((1-\delta)\rho_j - \rho_j)^{-1} (1 - \rho_j \times (1-\delta)\rho_j))^{-n+3/2} (1 + \mathcal{O}(n^{-1})).
\end{aligned} \tag{B.7}$$

Therefore because of (B.6) and (B.7), regardless of the sign of $\rho_j$, it follows that there exists a universal constant $C_0$ for which

$$\begin{aligned}
&\mathbb{P}(|\widehat{\rho}\,(X_j, Y)| > (1-\delta)|\rho_j|) \\
&\geq 1 - \frac{C_0}{\sqrt{2\pi}\delta|\rho_j|} \frac{\Gamma(n)}{\Gamma(n+1/2)} (1-\rho_j^2)^{\frac{n}{2}} (1 - (1-\delta)^2\rho_j^2)^{\frac{n-1}{2}} (1 - (1-\delta)\rho_j^2)^{-n+\frac{3}{2}} \\
&\geq 1 - \frac{C_0}{\sqrt{n}\delta|\rho_j|} \exp\big( -\rho_j^2 n/2 - (1-\delta)^2\rho_j^2(n-1)/2 + (1-\delta)\rho_j^2(n-3/2)\big) \\
&= 1 - \frac{C_0}{\sqrt{n\delta^2\rho_j^2}} \exp\big( -\rho_j^2 n\delta^2/2 + \rho_j^2(1-\delta)^2/2 - 3(1-\delta)\rho_j^2/2\big) \\
&\geq 1 - \frac{C_0}{\sqrt{n\delta^2\rho_j^2}} \exp(-\rho_j^2 n\delta^2/2),
\end{aligned} \tag{B.8}$$

where we used $\exp(x) \geq 1+x$ and Wendel's inequality [42] $\frac{\Gamma(n)}{\Gamma(n+1/2)} \leq \sqrt{\frac{n+1/2}{n}} \frac{1}{\sqrt{n}} \leq \sqrt{\frac{2\pi}{n}}$ in the second inequality (B.8).

Thus, we have that with probability at least $1 - \frac{C_0}{\sqrt{n\delta^2\rho_j^2}} \exp(-\rho_j^2 n\delta^2/2)$ that

$$\widehat{\Delta}(X_j, Y) \geq \frac{(1-\delta)^2 \widehat{\mathrm{Var}}(Y)\rho_j^2}{\log(2n)+1} \iff \widehat{\rho}^{\,2}(\widehat{Y}(X_j), Y) \geq \frac{(1-\delta)^2\rho_j^2}{\log(2n)+1}.$$

This completes the first half of the proof of Theorem 1.

## B.2 Proof of the upper bound (10)

We first state the following sample variance concentration inequality, which will be helpful.

**Lemma B.1.** *Let $Z_1, \ldots, Z_n$ be i.i.d. $\mathcal{N}(0, \sigma_Z^2)$. For any $0 < \delta < 1$, we have*

$$\mathbb{P}(\widehat{\mathrm{Var}}(Z) \geq (1-\delta)\frac{n-1}{n}\sigma_Z^2) \geq 1 - \exp(-\delta^2(n-1)/4) \tag{B.9}$$

*and*

$$\mathbb{P}(\widehat{\mathrm{Var}}(Z) \leq (1+\delta)\frac{n-1}{n}\sigma_Z^2) \geq 1 - \exp(-(n-1)(1+\delta-\sqrt{1+2\delta})/2). \tag{B.10}$$

*Proof of Lemma B.1.* Since $Z_i$ are independent and normally distributed, by Cochran's theorem we have $\widehat{\mathrm{Var}}(Z) \sim \frac{\sigma_Z^2}{n}\chi^2_{n-1}$. In the notation of [27], choosing $D = n-1$ and $x = \delta^2(n-1)/4$ for the chi-squared concentration inequality (4.4) in [27], we have that

$$\begin{aligned}\mathbb{P}\Big(\widehat{\mathrm{Var}}(Z) \geq (1-\delta)\frac{n-1}{n}\sigma_Z^2\Big) &= 1 - \mathbb{P}(\chi^2_{n-1} < (1-\delta)(n-1)) \\ &\geq 1 - \exp(-\delta^2(n-1)/4),\end{aligned}$$

proving (B.9). For (B.10), choosing $D = n-1$ and $x = (n-1)(1+\delta-\sqrt{1+2\delta})/2$ in [27, Equation (4.3)] we see that

$$\begin{aligned}\mathbb{P}\Big(\widehat{\mathrm{Var}}(Z) \leq (1+\delta)\frac{n-1}{n}\sigma_Z^2\Big) &= 1 - \mathbb{P}(\chi^2_{n-1} > (1+\delta)(n-1)) \\ &\geq 1 - \exp(-(n-1)(1+\delta-\sqrt{1+2\delta})/2). \qquad \square\end{aligned}$$

Now we are ready to prove the upper bound (10). We begin with the inequality (12), as shown in the proof sketch of Theorem 1. We aim to upper bound the right hand side of (12) using Lemma B.1. Since the samples $X_{1j}, \ldots, X_{nj}$ are i.i.d., using (B.10) and choosing $\delta = 1$, we find that with probability at least $1 - \exp\left(-(n-1)\frac{2-\sqrt{3}}{2}\right) \geq 1 - \exp(-(n-1)/16)$, we have that $\widehat{\mathrm{Var}}(X_j) \leq 2\sigma^2_{X_j}$. Similarly, choosing $\delta = 1/2$ in (B.9), we also have that with probability at least $1 - \exp(-(n-1)/16)$ that $\widehat{\mathrm{Var}}(Y) \geq \sigma_Y^2/4$. Substituting these concentration inequalities into the RHS of (12), it follows by a union bound that with probability at least $1 - 4n\exp(-n\delta^2/12) - 2\exp(-(n-1)/16)$,

$$\widehat{\Delta}(X_j, Y) \leq 24\widehat{\mathrm{Var}}(Y)\rho_j^2 + 4\delta^2\widehat{\mathrm{Var}}(Y) \iff \widehat{\rho}^{\,2}(\widehat{Y}(X_j), Y) \leq 24\rho_j^2 + 4\delta^2.$$

Finally, noticing that $\sqrt{24\rho_j^2 + 4\delta^2} < 5|\rho_j| + 2\delta$ completes the proof.

# C Proof of Model Selection Consistency for Linear Models

Recall the setting mentioned under the heading "***Minimum sample size for consistency***" in Section 6, which considers the same linear model with Gaussian variates from Theorem 1. To reiterate, we assume that $\mathbf{\Sigma} = \mathbf{I}_{p\times p}$ is the $p \times p$ identity matrix, $\sum_{k=1}^p \beta_k^2 = \mathcal{O}(1)$, and $\min_{j\in\mathcal{S}} |\beta_j|^2 \asymp 1/s$, all of which are special cases of the more general setting considered in [38, Corollary 1]. Under these assumptions, we then have $\rho^2(X_j, Y) = \beta_j^2/(\sigma^2 + \sum_{k=1}^p \beta_k^2) \gtrsim 1/s$ for any $j \in \mathcal{S}$ and $\rho(X_j, Y) = 0$ for $j \in \mathcal{S}^c$. Our goal is to show that $n \asymp s\log(n)\log(n(p-s))$ samples suffice for high probability model selection consistency.

Choosing $\delta = 1/2$ in (9) applied to $j \in \mathcal{S}$ and using (6), there exists a universal positive constant $C_0$ such that with probability at least $1 - \frac{2C_0}{\sqrt{n\rho^2(X_j,Y)}}\exp(-n\rho^2(X_j,Y)/8)$, we

have

$$\begin{aligned}\widehat{\Delta}(X_j, Y) &\geq \widehat{\mathrm{Var}}(Y) \times \frac{\rho^2(X_j, Y)}{4(\log(2n)+1)} \\ &= \frac{\widehat{\mathrm{Var}}(Y)}{4(\log(2n)+1)} \times \frac{\beta_j^2}{\sigma^2 + \sum_{k=1}^p \beta_k^2} \\ &\gtrsim \frac{\widehat{\mathrm{Var}}(Y)}{s\log(n)}.\end{aligned}$$

Therefore by a union bound over all $s$ relevant variables, we have that with probability at least $1 - s\max_{j\in\mathcal{S}}\{\frac{2C_0}{\sqrt{n\rho^2(X_j,Y)}}\exp(-n\rho^2(X_j,Y)/8)\}$,

$$\widehat{\Delta}(X_j, Y) \gtrsim \frac{\widehat{\mathrm{Var}}(Y)}{s\log(n)} \qquad \forall j \in \mathcal{S}. \tag{C.11}$$

Furthermore, by applying (10) for $j \in \mathcal{S}^c$ (and noting that $\rho(X_j, Y) = 0$) and using (6), with probability at least $1 - 4n\exp(-\delta^2 n/12) - 2\exp(-(n-1)/16)$, we have

$$\widehat{\Delta}(X_j, Y) \leq 4\widehat{\mathrm{Var}}(Y)\delta^2.$$

Therefore by a union bound over all $p - s$ irrelevant variables we have that with probability at least $1 - 4n(p-s)\exp(-\delta^2 n/12) - 2(p-s)\exp(-(n-1)/16)$,

$$\widehat{\Delta}(X_j, Y) \leq 4\widehat{\mathrm{Var}}(Y)\delta^2 \qquad \forall j \in \mathcal{S}^c.$$

Now, choosing $\delta^2 = \frac{C_3}{s\log(n)}$ for some appropriate constant $C_3 > 0$ which only depends on $\sigma^2$ to match (C.11), we see by a union bound that

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \geq 1 &- \max_{j\in\mathcal{S}}\Big\{\frac{2C_0 s}{\sqrt{n\rho^2(X_j,Y)}}\exp\Big(-\frac{n\rho^2(X_j,Y)}{8}\Big)\Big\} \\ &- 4n(p-s)\exp\Big(-\frac{C_3 n}{12s\log(n)}\Big) - 2(p-s)\exp\Big(-\frac{(n-1)}{16}\Big).\end{aligned} \tag{C.12}$$

Since $\rho^2(X_j, Y) \gtrsim 1/s$ for all $j \in \mathcal{S}$, (C.12) implies that if $n(p-s)\exp\big(-\frac{C_3 n}{12s\log(n)}\big) \to 0$, then $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$. Hence, a sufficient sample size for consistent support recovery is

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

as desired.

# D Proof of Propositions 1 and 2

This section will mainly be devoted to proving Proposition 1. First we will present the machinery which will be used to prove Proposition 1. At the end of the section we will complete the proof of Proposition 1 and prove Proposition 2 by recycling and simplifying the proof of Proposition 1.

First, we state and prove a lemma that will be used in later proofs. Though stated in terms of general probability measures, we will be specifically interested in the case where $\mathbb{P}$ is the empirical probability measure $\mathbb{P}_n$ and $\mathbb{E}$ is the empirical expectation $\mathbb{E}_n$, both with respect to a sample of size $n$.

**Lemma D.2.** *For any random variables $U$ and $V$ with finite second moments with respect to a probability measure $\mathbb{P}$,*

$$\mathrm{Cov}_{\mathbb{P}}(U,V) \geq \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\big(\sqrt{\mathrm{Var}_{\mathbb{P}}(V)} - 2\sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]}\big).$$

*Proof of Lemma D.2.* First note that by the triangle inequality,

$$\big|\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| \leq \sqrt{\mathrm{Var}_{\mathbb{P}}(U-V)} \leq \sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]}. \tag{D.13}$$

To complete the proof, we write

$$\mathrm{Cov}_{\mathbb{P}}(U,V) = \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathrm{Var}_{\mathbb{P}}(V)} + \big(\mathrm{Cov}_{\mathbb{P}}(U,V) - \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big) \tag{D.14}$$

and apply (D.13) to arrive at

$$\begin{aligned}
&\big|\mathrm{Cov}_{\mathbb{P}}(U,V) - \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| && \text{(D.15)}\\
&= \big|\mathrm{Cov}_{\mathbb{P}}(U,V) - \mathrm{Var}_{\mathbb{P}}(U) + \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\big(\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big)\big| \\
&\leq \big|\mathrm{Cov}_{\mathbb{P}}(U,V) - \mathrm{Var}_{\mathbb{P}}(U)\big| + \sqrt{\mathrm{Var}_{\mathbb{P}}(U)} \times \big|\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| \\
&\leq 2\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} \times \big|\sqrt{\mathrm{Var}_{\mathbb{P}}(U)} - \sqrt{\mathrm{Var}_{\mathbb{P}}(V)}\big| && \text{(D.16)}\\
&\leq 2\sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]},
\end{aligned}$$

where the penultimate line (D.16) follows from the Cauchy-Schwarz inequality. Substituting (D.15) into (D.14), we get

$$\mathrm{Cov}_{\mathbb{P}}(U,V) \geq \sqrt{\mathrm{Var}_{\mathbb{P}}(U)}\big(\sqrt{\mathrm{Var}_{\mathbb{P}}(V)} - 2\sqrt{\mathbb{E}_{\mathbb{P}}[(U-V)^2]}\big),$$

which proves the claim. □

The following sample variance concentration inequality will also come in handy.

**Lemma D.3** (Equation 5, [33])**.** *Let $U$ be a random variable bounded by $B$. Then for all $\gamma > 0$,*

$$\mathbb{P}\Big(\frac{n}{n-1}\widehat{\mathrm{Var}}(U) \geq \mathrm{Var}(U) - \gamma\Big) \geq 1 - \exp\Big(-\frac{(n-1)\gamma^2}{8B^2\mathrm{Var}(U)}\Big).$$

As explained in the main text, the key step in the proof of Proposition 1 is to apply Lemma 1 with a good approximation $\widetilde{f}_j(\cdot)$ to the marginal projection $f_j(\cdot)$ that also has a sufficiently large number of data points in every one of its stationary intervals. The following lemma provides the precise properties of such an approximation.

**Lemma D.4.** *Suppose $f_j(\cdot)$ satisfies Assumption 2. Let $a > 0$ be a positive constant and let $M$ be a positive integer such that $Ma \leq 1$. There exists a function $\widetilde{f}_j(\cdot)$ with at most $M$ stationary intervals such that, with probability at least $1 - 2n\exp(-na/12)$, both of the following statements are simultaneously true:*

1. *The number of data points in any stationary interval of $\widetilde{f}_j(\cdot)$ is between $na/2$ and $3na/2$.*
2. *$\sqrt{\mathbb{E}_n[(f_j(X_j) - \widetilde{f}_j(X_j))^2]} \leq K_0 M^{-d} + K_1(Ma)^{5/2}$, where $K_0$ and $K_1$ are some constants depending on $d$, $B$, $\alpha$, and $r$, and $\mathbb{E}_n$ denotes the empirical expectation.*

*Proof of Lemma D.4.* Recall that $f_j(\cdot)$ is defined over $\mathbb{R}$, so even though we restrict our attention $[0,1]$, we can assume it is bounded over the larger interval $[-1,2]$ for all $j$. By [22, Theorem VIII], there exists a polynomial $P_M(\cdot)$ of degree $M+1 > r$ such that

$$\sup_{x\in[-1,2]} |f_j(x) - P_M(x)| \leq 3^d LA(M+1)^{-r}(M+1-r)^{-\alpha},$$

where $A = \frac{(r+1)^{r-1}(K/2)^r(K/2+2)}{r!}$, $L$ is the Lipschitz constant from Assumption 2, and $K$ is a universal constant given in [22, Theorem I]. Since $\frac{M+1}{M+1-r} \leq r+1$ whenever $M+1 > r$, we also have that

$$\sup_{x\in[-1,2]} |f_j(x) - P_M(x)| \leq 3^d(r+1)^\alpha LA(M+1)^{-d} \leq K_0 M^{-d}, \tag{D.17}$$

where $K_0 = 3^d(r+1)^\alpha LA$ is a constant.

Let $\{\xi_k\}_{1\leq k\leq M}$ be the collection of (at most) $M$ stationary points of $P_M(\cdot)$ in $[0,1]$. We assume that $a < \min_k(\xi_k - \xi_{k-1})$; otherwise we remove points from $\{\xi_k\}_{1\leq k\leq M}$ until this holds. Let $I_k = [\xi_k, \xi_k + a]$.

We will now show that Properties 1 and 2 of Lemma D.4 are satisfied by the function

$$\widetilde{f}_j(x) = \begin{cases} P_M(x) & x \notin \bigcup_k I_k \\ P_M(\xi_k) & x \in I_k \text{ for some } k. \end{cases} \tag{D.18}$$

First, it is clear that $\widetilde{f}_j(\cdot)$ has at most $M$ stationary intervals. Next, notice that by the multiplicative version of Chernoff's inequality [1, Proposition 2.4], since each stationary interval $I_k$ has length $a$, we have

$$\mathbb{P}(\#\{X_{ij} \in I_k\} \geq na/2) > 1 - \exp(-na/8),$$

and

$$\mathbb{P}(\#\{X_{ij} \in I_k\} \leq 3na/2) > 1 - \exp(-na/12).$$

Thus, by a union bound, the probability that

$$na/2 \leq \#\{X_{ij} \in I_k\} \leq 3na/2 \qquad \forall\, k$$

is at least $1-2M\exp(-na/12)$. Since all stationary intervals are disjoint and are contained in $[0,1]$, we have $Ma \leq 1$ or $M \leq 1/a$, which implies that $1 - 2M\exp(-na/12) \geq 1 - (2/a)\exp(-na/12)$. Therefore, we know that

$$\mathbb{P}(3na/2 \geq \#\{X_{ij} \in I_k\} \geq na/2 \text{ for all } k) \geq \max\{0, 1 - (2/a)\exp(-na/12)\}. \quad \text{(D.19)}$$

Notice that if $a \leq 1/n$, then

$$\max\{0, 1 - (2/a)\exp(-na/12)\} = 0 \geq 1 - 2n\exp(-na/12)$$

and if $a > 1/n$, then

$$\begin{aligned}\max\{0, 1 - (2/a)\exp(-na/12)\} &\geq 1 - (2/a)\exp(-na/12) \\ &\geq 1 - 2n\exp(-na/12).\end{aligned}$$

Thus, for all $a \geq 0$, we have

$$\max\{0, 1 - (2/a)\exp(-na/12)\} \geq 1 - 2n\exp(-na/12),$$

which when combined with (D.19) proves Property 1 of Lemma D.4.

To prove Property 2 of Lemma D.4 we use the triangle inequality and part of Property 1:

$$\mathbb{P}(\#\{X_{ij} \in I_k\} \leq 3na/2 \text{ for all } k) \geq 1 - 2n\exp(-na/12). \quad \text{(D.20)}$$

In view of (D.17), to bound $\sqrt{\mathbb{E}_n[(f_j(X_j) - \widetilde{f}_j(X_j))^2]}$ we aim to bound $\frac{1}{n}\sum_{i=1}^n (P_M(X_{ij}) - \widetilde{f}_j(X_{ij}))^2$.

Notice that by Bernstein's theorem for polynomials [21, Theorem B2a], we also have

$$|P_M''(x)| \leq \frac{(M+1)^2 \sup_{x\in[-1,2]} |P_M(x)|}{(2-x)(x+1)}, \quad -1 < x < 2. \quad \text{(D.21)}$$

By (D.17), $\sup_{x\in[-1,2]} |P_M(x)| \leq K_0 M^{-d} + \sup_{x\in[-1,2]} |f_j(X_j)| \leq K_0 + B$. Thus, by (D.21),

$$\sup_{x\in[0,1]} |P_M''(x)| \leq \frac{(M+1)^2(K_0+B)}{2} \leq K_1 M^2,$$

where $K_1 = 2(K_0 + B)$ and we used the fact that $(2 - x)(x + 1) \geq 2$ for $x \in [0, 1]$. Additionally, because each $\xi_k$ is a stationary point, $P'_M(\xi_k) = 0$, and hence by a second order Taylor approximation, we have

$$|P_M(x) - P_M(\xi_k)| \leq K_1 M^2 a^2/2 \tag{D.22}$$

for $x \in I_k$.

Therefore, by (D.18), (D.20), and (D.22) we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(P_M(X_{ij}) - \widetilde{f}_j(X_{ij}))^2 &= \sum_k \frac{1}{n}\sum_{X_{ij}\in I_k}(P_M(X_{ij}) - P_M(\xi_k))^2 \\
&\leq \frac{K_1^2 M^4 a^4}{4n}\sum_{k=1}^{M}\#\{X_{ij} \in I_k\} \\
&\leq \frac{K_1^2 M^4 a^4}{4n}\sum_{k=1}^{M}\frac{3na}{2} \\
&= \frac{3K_1^2(Ma)^5}{8},
\end{aligned}$$

with probability at least $1 - 2n\exp(-na/12)$. Combining this with (D.17) and using the triangle inequality to bound $\sqrt{\mathbb{E}_n[(f_j(X_j) - \widetilde{f}_j(X_j))^2]}$ by $\sqrt{\mathbb{E}_n[(P_M(X_j) - \widetilde{f}_j(X_j))^2]} + \sqrt{\mathbb{E}_n[(P_M(X_j) - f_j(X_j))^2]}$ proves Property 2. □

Returning to the proof of Proposition 1, by Lemma 1 along with Lemma D.4, we have that with probability at least $1 - 2n\exp(-na/12)$ that

$$\widehat{\Delta}(X_j, Y) \geq \frac{1}{2M/a + \log(2n) + 1} \times \widehat{\mathrm{Cov}}^2\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y\right). \tag{D.23}$$

Now we choose

$$a = \left(\frac{\tau^2 \mathrm{Var}(f_j(X_j))}{4(2K_0 + K_1)^2}\right)^{(2d+5)/(10d)}, \quad M = \lfloor a^{-5/(2d+5)} \rfloor, \tag{D.24}$$

where $\tau$ is a constant to be specified in (D.26), so that Property 2 of Lemma D.4 becomes

$$\begin{aligned}
\sqrt{\mathbb{E}_n[(f_j(X) - \widetilde{f}_j(X))^2]} &\leq K_0 M^{-d} + K_1(Ma)^{5/2} \\
&\leq (2K_0 + K_1)a^{5d/(2d+5)} \\
&\leq \frac{\tau\sqrt{\mathrm{Var}(f_j(X_j))}}{2},
\end{aligned} \tag{D.25}$$

with probability at least $1 - 2n\exp(-na/12)$.

Recall the condition $M + 1 > r$ as part of [22, Theorem VIII], which states that the degree of the polynomial must be greater than the order of Lipschitz derivative in order to approximate the function well. Since, $\mathrm{Var}(f_j(X_j)) \leq B^2$, this condition will be satisfied if we make the following choice:

$$\tau := \min\Big(\frac{2(2K_0 + K_1)}{r^d B}, \frac{1}{4}\Big). \tag{D.26}$$

It follows by Lemma D.2 along with (D.25) that

$$\widehat{\mathrm{Cov}}(\widetilde{f}_j(X_j), f_j(X_j)) \geq \sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}\Big(\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))} - \tau\sqrt{\mathrm{Var}(f_j(X_j))}\Big) \tag{D.27}$$

with probability at least $1 - 2n\exp(-na/12)$.

In the next Lemma, we use (D.27) along with Lemma D.3 to obtain a lower bound on the right hand side of (D.23).

**Lemma D.5.** *With probability at least* $1 - \exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big) - \exp\Big(-\frac{n\tau^2\mathrm{Var}(f_j(X_j))}{8(B^2+\sigma^2)}\Big) - 2n\exp(-na/12)$, *we have that*

$$\widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y\right) \geq (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))}.$$

*Proof of Lemma D.5.* Recalling (15), we will prove Lemma D.5 by first getting a concentration bound on (I) and then getting a concentration bound on (II).

To get a concentration bound on (I), we need Lemma D.3 to lower bound the sample variance on the right hand side of inequality (D.27). Choosing $U = f_j(X_j) \in [-B, B]$ and $\gamma = \mathrm{Var}(f_j(X_j))(1 - 8\tau^2)$ (which is greater than zero by the choice of $\tau$ in (D.26)), notice that Lemma D.3 gives us

$$\begin{aligned}
&\mathbb{P}\Big(\widehat{\mathrm{Var}}(f_j(X_j)) \geq \frac{8\tau^2(n-1)}{n}\mathrm{Var}(f_j(X_j))\Big) \\
&\geq 1 - \exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big),
\end{aligned}$$

so that by (D.27),

$$\begin{aligned}
\widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, f_j(X_j)\right) &\geq \sqrt{\widehat{\mathrm{Var}}(f_j(X_j))} - \tau\sqrt{\mathrm{Var}(f_j(X_j))} \\
&\geq \tau(\sqrt{8}\sqrt{1 - 1/n} - 1)\sqrt{\mathrm{Var}(f_j(X_j))} \\
&\geq \tau\sqrt{\mathrm{Var}(f_j(X_j))},
\end{aligned}$$

with probability at least $1 - \exp\left(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\right) - 2n\exp(-na/12)$.

Now we need to get a concentration bound for (II). Let $s_i = \frac{\widetilde{f}_j(X_{ij}) - \frac{1}{n}\sum_{k=1}^n \widetilde{f}_j(X_{kj})}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}$. We need to bound

$$\widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y - f_j(X_j)\right) = \frac{1}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}) + \varepsilon_i).$$

For notational simplicity, we let $\underline{\mathbf{X}} = (X_{ij})$ be the $n \times p$ data matrix with $\mathbf{X}_i$ as rows. First notice that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}) + \varepsilon_i)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}) + \varepsilon_i)\right)\middle|\underline{\mathbf{X}}\right]\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i\varepsilon_i\right)\middle|\underline{\mathbf{X}}\right]\right].
\end{aligned}$$

Now, by the sample independence of the errors $\epsilon_i$, we can write the above as

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda}{n}s_i\varepsilon_i\right)\middle|\underline{\mathbf{X}}\right]\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\prod_{i=1}^n \exp\left(\frac{\lambda^2 s_i^2\sigma^2}{2n^2}\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\right]\exp\left(\frac{\lambda^2\sigma^2}{2n}\right),
\end{aligned}$$

where we used Assumption 6 and the fact that $\frac{1}{n}\sum_{i=1}^n s_i^2 = 1$. Recalling that $s_i$ depends on $(X_{1j}, X_{2j}, \ldots, X_{nj})^\top$, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\middle|X_{1j}, X_{2j}, \ldots, X_{nj}\right]\right] \qquad \text{(D.28)} \\
&= \mathbb{E}\left[\prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda}{n}s_i(g(\mathbf{X}_i) - f_j(X_{ij}))\right)\middle|X_{1j}, X_{2j}, \ldots, X_{nj}\right]\right],
\end{aligned}$$

where we used sample independence in the second equality. Finally, applying Hoeffding's Lemma along with the fact that $\|g\|_\infty \leq B$, we have that (D.28) is bounded above by

$$\mathbb{E}\left[\prod_{i=1}^{n} \exp\left(\frac{\lambda^2 s_i^2 B^2}{2n^2}\right)\right] \leq \exp\left(\frac{\lambda^2 B^2}{2n}\right).$$

Having bounded the moment generating function, we can now apply Markov's inequality to see that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} s_i(g(X) - f_j(X_j) + \varepsilon_i) \leq -\gamma\right) \\
&= \mathbb{P}\left(\exp\left(-\frac{\lambda}{n}\sum_{i=1}^{n} s_i(g(X) - f_j(X_j) + \varepsilon_i)\right) \geq \exp(\lambda\gamma)\right) \\
&\leq \exp\left(\frac{\lambda^2(B^2+\sigma^2)}{2n} - \gamma\lambda\right) \\
&\leq \exp\left(-\frac{n\gamma^2}{2(B^2+\sigma^2)}\right),
\end{aligned}$$

where the last inequality follows by maximizing over $\lambda$. Choosing $\gamma = (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))}$, we have by a union bound that,

$$\begin{aligned}
\widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y\right) &= \widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, f_j(X_j)\right) \\
&\quad + \widehat{\mathrm{Cov}}\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y - f_j(X_j)\right) \\
&\geq \tau\sqrt{\mathrm{Var}(f_j(X_j))} - (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))} \\
&= (\tau/2)\sqrt{\mathrm{Var}(f_j(X_j))},
\end{aligned}$$

with probability at least $1 - \exp\left(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\right) - \exp\left(-\frac{n\tau^2\mathrm{Var}(f_j(X_j))}{8(B^2+\sigma^2)}\right) - 2n\exp(-na/12)$. □

With this setup, we are now ready to finish the proofs of Proposition 1 and 2.

*Proof of Proposition 1.* Recalling our concentration bound (D.23) along with Lemma D.5, it follows that with probability at least $1 - \exp\left(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\right) - \exp\left(-\right.$

$\frac{n\tau^2\mathrm{Var}(f_j(X_j))}{8(B^2+\sigma^2)}\Big) - 4n\exp\Big(-\frac{n}{12}\Big(\frac{\tau^2\mathrm{Var}(f_j(X_j))}{4(2K_0+K_1)^2}\Big)^{(2d+5)/(10d)}\Big)$ that

$$\begin{aligned}\widehat{\Delta}(X_j,Y) &\geq \frac{1}{2M/a+\log(2n)+1}\times\widehat{\mathrm{Cov}}^2\left(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}},Y\right)\\ &\geq \frac{\tau^2\mathrm{Var}(f_j(X_j))}{8a^{-(2d+10)/(2d+5)}+4(\log(2n)+1)}\\ &\geq \frac{\tau^2\mathrm{Var}(f_j(X_j))a^{(2d+10)/(2d+5)}}{8+4(\log(2n)+1)a^{(2d+10)/(2d+5)}}\\ &\geq \frac{C_2(\mathrm{Var}(f_j(X_j)))^{(6d+5)/5d}}{\log(n)},\end{aligned} \tag{D.29}$$

where we used our choice of $M$ and $a$ in (D.24) and $\tau$ in (D.26) and $C_2$ is some constant which only depends on $L$, $B$, $r$, and $\alpha$. Notice in the last inequality (D.29) we bound $a$ in the denominator with $\mathrm{Var}(f_j(X_j)) \leq B^2$.

To conclude the proof, we simplify our probability bound. To this end, notice that

$$(\mathrm{Var}(f_j(X_j)))^{(2d+5)/(10d)} = \frac{\mathrm{Var}(f_j(X_j))}{(\mathrm{Var}(f_j(X_j)))^{(8d-5)/(10d)}} \geq \frac{\mathrm{Var}(f_j(X_j))}{B^{(8d-5)/(5d)}},$$

where the last inequality holds by assumption that $d \geq 5/8$ and the fact that $\mathrm{Var}(f_j(X_j)) \leq B^2$. Therefore we have $1 - \exp\Big(-\frac{(n-1)(1-8\tau^2)^2\mathrm{Var}(f_j(X_j))}{8B^2}\Big) - \exp\Big(-\frac{n\tau^2\mathrm{Var}(f_j(X_j))}{8(B^2+\sigma^2)}\Big) - 4n\exp\Big(-\frac{n}{12}\Big(\frac{\tau^2\mathrm{Var}(f_j(X_j))}{4(K_0+K_1)^2}\Big)^{(2d+5)/(10d)}\Big) \geq 1-(4n+2)\exp(-nC_1\mathrm{Var}(f_j(X_j)))$ for some constant $C_1$ which depends only on $L$, $B$, $\sigma$, $r$, and $\alpha$. □

*Proof of Proposition 2.* The main difference with Proposition 1 is that when $f_j(\cdot)$ is monotone we now have $M = 0$. We no longer need the approximations $P_M(\cdot)$ or $\widetilde{f}_j(\cdot)$ in Lemma D.4, $a$ in (D.24) and $\tau$ in (D.26), or Lemma D.2. We will only need Lemma 1 and a version of Lemma D.5. Instead of applying Lemma 1 with an approximation to the marginal projection, we can choose $h_j(\cdot)$ to equal $f_j(\cdot)$ directly to see that

$$\widehat{\Delta}(X_j,Y) \geq \frac{\widehat{\mathrm{Cov}}^2\left(\frac{f_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}},Y\right)}{\log(2n)+1}. \tag{D.30}$$

Next, we have

$$\begin{aligned}
&\widehat{\mathrm{Cov}}\left(\frac{f_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}}, Y\right) \\
&= \widehat{\mathrm{Cov}}\left(\frac{f_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}}, f_j(X_j)\right) + \widehat{\mathrm{Cov}}\left(\frac{f_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}}, Y - f_j(X_j)\right) \\
&= \underbrace{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}}_{\text{(IV)}} + \underbrace{\widehat{\mathrm{Cov}}\left(\frac{f_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}}, Y - f_j(X_j)\right)}_{\text{(V)}}.
\end{aligned} \tag{D.31}$$

Now, we follow the same steps as the proof of Lemma D.5 to lower bound (D.31). For (IV), again use Lemma D.3 with $U = f_j(X_j) \in [-B, B]$ but choose instead $\gamma = \mathrm{Var}(f_j(X_j))/2$ to get

$$\begin{aligned}
\mathbb{P}\Big(\text{(IV)} \geq \frac{\sqrt{\mathrm{Var}(f_j(X_j))}}{2}\Big) &\geq \mathbb{P}\Big(\widehat{\mathrm{Var}}(f_j(X_j)) \geq \frac{n-1}{2n}\mathrm{Var}(f_j(X_j))\Big) \\
&\geq 1 - \exp\Big(-\frac{(n-1)\mathrm{Var}(f_j(X_j))}{32B^2}\Big).
\end{aligned} \tag{D.32}$$

For (V), we can follow the same steps as the second half of the proof of Lemma D.5 to see that

$$\mathbb{P}(\text{(V)} \leq -\gamma) \leq \exp\left(-\frac{n\gamma^2}{2(B^2+\sigma^2)}\right).$$

However, for (V), we instead choose $\gamma = \sqrt{\mathrm{Var}(f_j(X_j))}/4$ to get

$$\mathbb{P}\left(\text{(V)} \geq -\frac{\sqrt{\mathrm{Var}(f_j(X_j))}}{4}\right) \geq 1 - \exp\left(-\frac{n\mathrm{Var}(f_j(X_j))}{32(B^2+\sigma^2)}\right). \tag{D.33}$$

Now using a union bound and substituting the events in (D.33) and (D.32) into (D.31), we see that with probability at least $1 - \exp\Big(-\frac{(n-1)\mathrm{Var}(f_j(X_j))}{32B^2}\Big) - \exp\Big(-\frac{n\mathrm{Var}(f_j(X_j))}{32(B^2+\sigma^2)}\Big) \geq 1 - 2\exp\Big(-\frac{(n-1)\mathrm{Var}(f_j(X_j))}{32(B^2+\sigma^2)}\Big)$, we have that

$$\widehat{\mathrm{Cov}}\left(\frac{f_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(f_j(X_j))}}, Y\right) \geq \frac{\sqrt{\mathrm{Var}(f_j(X_j))}}{4}. \tag{D.34}$$

Therefore, substituting (D.34) into (D.30), we have that with probability at least $1 - 2\exp\Big(-\frac{(n-1)\mathrm{Var}(f_j(X_j))}{32(B^2+\sigma^2)}\Big)$ that

$$\widehat{\Delta}(X_j, Y) \geq \frac{\widehat{\mathrm{Cov}}^2\Big(\frac{\widetilde{f}_j(X_j)}{\sqrt{\widehat{\mathrm{Var}}(\widetilde{f}_j(X_j))}}, Y\Big)}{\log(2n)+1} \geq \frac{\mathrm{Var}(f_j(X_j))}{16(\log(2n)+1)}. \quad \square$$

# E Proof of Theorems 2 and 3

In this section we use Proposition 1 and Proposition 2 along with Lemma 3 to complete the proofs of Theorem 2 and Theorem 3.

*Proof of Theorem 2.* The high-level idea is to show that the upper and lower bounds on the impurity reductions for irrelevant and relevant variables from Lemma 3 and Proposition 1, respectively, are well-separated.

By Proposition 1 for all variables $j \in \mathcal{S}$ and a union bound, we see that with probability at least $1 - s(4n+2)\exp(-C_1 nv)$, we have

$$\widehat{\Delta}(X_j, Y) \geq \frac{C_2 v^{6/5+1/d}}{\log(n)} \qquad \forall\, j \in \mathcal{S}. \tag{E.35}$$

By Lemma 3 and applying a union bound over all $p - s$ variables in $\mathcal{S}^c$, we have that with probability at least $1 - 4n(p-s)\exp(-n\xi^2/(12(B^2+\sigma^2)))$ that

$$\widehat{\Delta}(X_j, Y) \leq \xi^2 \qquad \forall\, j \in \mathcal{S}^c. \tag{E.36}$$

Recall that if we know the size $s$ of the support $\mathcal{S}$, then $\widehat{\mathcal{S}}$ consists of the top $s$ impurity reductions. Note that choosing $\xi^2 = \frac{C_2 v^{6/5+1/d}}{2\log(n)}$ in (E.36) will give us a high probability upper bound on $\widehat{\Delta}(X_j, Y)$ for irrelevant variables which is dominated by the lower bound on $\widehat{\Delta}(X_j, Y)$ for relevant variables in (E.35). Thus, by a union bound, it follows that with probability at least $1 - s(4n+2)\exp(-C_1 nv) - 4n(p-s)\exp\Big(-\frac{nC_2 v^{6/5+1/d}}{24\log(n)(B^2+\sigma^2)}\Big)$, we have $\widehat{\mathcal{S}} = \mathcal{S}$. □

*Proof of Theorem 3.* The proof is similar to that of Theorem 2 except that we use Proposition 2 in place of Proposition 1.

By Proposition 2 for all variables $j \in \mathcal{S}$ and a union bound, we see that with probability at least $1 - 2s\exp\Big(-\frac{(n-1)v}{32(B^2+\sigma^2)}\Big)$,

$$\widehat{\Delta}(X_j, Y) \geq \frac{v}{4(1+\log(2n))} \qquad \forall\, j \in \mathcal{S}.$$

Again, we have by Lemma 3 and a union bound over all $p - s$ variables in $\mathcal{S}^c$ that with probability at least $1 - 4n(p-s)\exp(-n\xi^2/(12(B^2+\sigma^2)))$ that

$$\widehat{\Delta}(X_j, Y) \leq \xi^2 \qquad \forall\, j \in \mathcal{S}^c. \tag{E.37}$$

Choosing $\xi^2 = \frac{v}{8(1+\log(2n))}$ in (E.37) and using a union bound, it follows that with probability at least $1 - 2s\exp\Big(-\frac{(n-1)v}{32(B^2+\sigma^2)}\Big) - 4n(p-s)\exp\Big(-\frac{nv}{96(1+\log(2n))(B^2+\sigma^2)}\Big)$, we have $\widehat{\mathcal{S}} = \mathcal{S}$. □

# References


[1] D. Angluin and L.G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18:155–193, 1979.

[2] Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–955, 2001.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

[5] Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 1984.

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

[7] Alexandre Bureau, Josée Dupuis, Kathleen Falls, Kathryn L Lunetta, Brooke Hayward, Tim P Keith, and Paul Van Eerdewegh. Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, 28:171–82, 2005.

[8] Ramón Díaz-Uriarte and Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics volume*, 7(3):171–82, 2006.

[9] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004.

[10] Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494), 2011.

[11] Jianqing Fan and Jinchi Lv. Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 70:849–911, 2008.

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

[13] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[14] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

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

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

[17] Harold Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B*, 15(2):193–232, 1953.

[18] Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313, 2010.

[19] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*, 2010.

[20] Wayne Iba and Pat Langley. Induction of one-level decision trees. *Machine Learning Proceedings*, pages 233–240, 1992.

[21] Dunham Jackson. Bernstein's theorem and trigonometric approximation. *Transactions of the American Mathematical Society*, 40(2):225–251, 1936.

[22] Dunham Jackson. *The Theory of Approximation*, volume 11. American Mathematical Society, 1991.

[23] Jalil Kazemitabar, Arash Amini, Adam Bloniarz, and Ameet S Talwalkar. Variable importance using decision trees. In *Advances in Neural Information Processing Systems*, pages 426–435, 2017.

[24] L.C. Keely and C.M. Tan. Understanding preferences for income redistribution. *Journal of Public Economics*, 92:944–961, 2008.

[25] Jason M. Klusowski. Sparse learning with CART. In *Advances in Neural Information Processing Systems*, 2020.

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

[27] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[28] Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A debiased MDI feature importance measure for random forests. In *Advances in Neural Information Processing Systems 32*, pages 8049–8059. Curran Associates, Inc., 2019.

[29] Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint: arXiv:1407.7502*, 2014.

[30] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, pages 431–439, 2013.

[31] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint: arXiv:1802.03888*, 2018.

[32] Kathryn L. Lunetta, L. Brooke Hayward, Jonathan Segal, and Paul Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, 5(32):199–231, 2004.

[33] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *COLT*, 2009.

[34] Pradeep Ravikumar, Han Liu, John Lafferty, and Larry Wasserman. Spam: Sparse additive models. *Journal of the Royal Statistical Society. Series B*, 71(5):1009–1030, 2009.

[35] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:563–564, 2007.

[36] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

[37] D. Van den Poel W. Buckinx. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research*, 164:252–268, 2005.

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

[39] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[40] Huazhen Wang, Fan Yang, and Zhiyuan Luo. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*, 17(60), 2016.

[41] Pengfei Wei, Zhenzhou Lu, and Jingwen Song. Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, 142:399–432, 2015.

[42] JG Wendel. Note on the gamma function. *The American Mathematical Monthly*, 55(9):563–564, 1948.

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

[44] J. Zhang, M. Zulkernine, and A. Haque. 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, 2008.