MODEL-FREE VARIABLE SELECTION VIA THE RULE-BASED VARIABLE PRIORITY

MIN LU
DIVISION OF BIOSTATISTICS
MILLER SCHOOL OF MEDICINE
UNIVERSITY OF MIAMI, USA
M.LU6@UMIAMI.EDU

HEMANT ISHWARAN
DIVISION OF BIOSTATISTICS
MILLER SCHOOL OF MEDICINE
UNIVERSITY OF MIAMI, USA
HISHWARAN@MED.MIAMI.EDU

Abstract

While achieving high prediction accuracy is a fundamental goal in machine learning, an equally important task is variable selection for finding a small number of features with high explanatory power. One of the more popular approaches used in machine learning is to measure variable importance by prediction accuracy; for example, random forests, gradient boosting and XGBoost are all procedures that adopt this idea. Permutation importance, specifically, calculates importance by permuting a variable and then considers how much the base learner prediction error changes as a result. In this paper, we show that this method has a hidden problem because it creates artificial data that may not match the true parent distribution. To avoid these issues, we instead propose a model-free framework of variable priority (VarPro) using rules acquired from externally constructed trees, that works with the existing data using elementary statistics without the need to create artificial data. The VarPro importance statistic for a set of variables S equals the difference between the estimator of conditional mean based on a rule and the estimator based on the corresponding released rule which is the rule obtained by removing any contraints on the features in S. The method is simple to use and applies to a broad range of data settings including regression, classification and survival. VarPro readily scales to large data requiring only calculating simple averages and is easily adapted to different targets of interest. We prove that VarPro is model-free consistent under relatively simple conditions that are generally agnostic to the type of rule-based procedure used for the rule generation step. Empirical studies using synthetic data and real world data demonstrates VarPro achieves excellent performance as good or better than state of the art methods.

Keywords Coordinate direction · Neighborhood · Released rule · Signal and noisy variables · Conditional expectation · Variable selection

1. Introduction

Although many machine learning procedures are capable of modeling a large number of variables and achieving high prediction accuracy, finding a small number of features with equivalent, or near equivalent prediction performance, is equally desirable. This allows the researcher to identify which variables play a prominent role in the problem setting, thus providing insight into the underlying mechanism for what otherwise might be considered a black box. In machine learning, variable selection is often performed using variable importance described by how much a prediction model accuracy depends on the information in each feature (Breiman, 2001; Friedman, 2001; Van der Laan, 2006; Ishwaran, 2007; Strobl et al., 2008;

Ishwaran et al., 2008; Doksum et al., 2008; Grömping, 2009; Genuer et al., 2010; Louppe et al., 2013). One of the most popular methods is permutation importance which was introduced by Leo Breiman in his famous random forests paper (Breiman, 2001). To calculate a variable's permutation importance, the given variable is randomly permuted in the out-of-sample data (i.e. out-of-bag OOB data) and the permuted OOB data is dropped down a tree. OOB prediction error is then calculated. The difference between this and the OOB error without permutation (i.e. from the original tree), averaged over all trees, is the importance of the variable. The larger the permutation importance of a variable, the more predictive the variable.

Besides using predictive performance, many other methods have also been developed in maching learning for selecting variables. However, these proposals tend to be specifically designed and tuned for the algorithm being developed and studied. Recently there has been attention given to developing variable importance that can apply more generally across different types of learning procedures (Wei et al., 2015; Lei et al., 2018; Fisher et al., 2019). This paper takes a broader approach in the spirit of these latter methods. We are interested in developing a simple and general procedure that can be used in a wide array of data settings. We will focus on rule-based procedures for generating our importance score; for example simple decision trees, or trees constructed from random forests. We are interested in describing how the rules obtained from these procedures can be used to assess variable importance.

Call Y the response variable and $X^{(1)}, \ldots, X^{(p)}$ the set of p potential explanatory features. We consider the setting where the researcher is interested in the conditional distribution of the response variable Y given the features $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})$. Our goal is to assess variable importance for functionals of the conditional distribution of Y given $\mathbf{X}$. We call the target of interest $\psi(\mathbf{X}) = \mathbb{E}(g(Y)|\mathbf{X})$, where g is some prechosen function that is context specific to the problem being studied. Some examples are given below:

1. *Classification.* For a categorical response with categories $c_1, \ldots, c_L$, interest could focus on the conditional probability $\psi(\mathbf{X}) = \mathbb{P}\{Y = c_l|\mathbf{X}\}$ for a specific category c_l, where $g(Y) = I\{Y = c_l\}$. For example, in studying the presence, absence, or recurrence of cancer, the researcher may focus on the recurrence of cancer to study the hypothesis that the probability of recurrence depends on certain features.
2. *Regression.* Here $\psi(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ for $g(Y) = Y$ where the goal is determining variables affecting the conditional mean.
3. *Time to event.* With survival analysis, the focus of interest can be the survival function $\psi(\mathbf{X}) = \mathbb{P}\{T^o > t|\mathbf{X}\}$, where $Y = T^o$ is the survival time, and notice this corresponds to $g(T^o) = I\{T^o > t\}$. Another popular way to summarize lifetime behavior is the restricted mean survival (RMST) (Royston and Parmar, 2011). As will be discussed in Section 5, the RMST evaluated at a time horizon $\tau > 0$ can be written as $\psi(\mathbf{X}) = \mathbb{E}(g(T^o)|\mathbf{X})$ where $g(T^o) = T^o \wedge \tau$.

From a high vantage point we can see that the goal is the same in each of our examples: even though the target depends on the problem being studied, the main idea is to identify variables of importance for a functional of interest, ψ. We expect in many settings that ψ will depend on a smaller subset of the p variables $\mathcal{S} \subset \{1, \ldots, p\}$ and we would like to find the minimal set $\mathcal{S}$ for which this holds, which we call the "signal variables", and simultaneously we would like to exclude the non-relevant variables, which we call "noisy variables". In order to make this idea more precise a formal definition is provided as follows. Write $\mathbf{X}^{(S)} = \{X^{(j)}\}_{j \in S}$ for the feature vector $\mathbf{X}$ restricted to coordinates $j \in S$ and $\mathbf{X}^{\setminus(S)} = \{X^{(j)}\}_{j \notin S}$ for coordinates not in S.

Definition 1. $\mathcal{S} \subset \{1, \ldots, p\}$ is the set of signal variables if $\mathcal{S}$ is the minimal set of coordinates that ψ depends on. Thus, $\mathcal{S}$ is the smallest subset of coordinates satisfying $\psi(\mathbf{x}) = \psi(\mathbf{x}^{(\mathcal{S})})$ for all $\mathbf{x} \in \mathcal{X}$. The complementary set $\mathcal{N} = \{1, \ldots, p\} \setminus \mathcal{S}$ is unrelated to ψ and therefore contains the noisy variables. If

$\mathcal{S} = \emptyset$, then $\psi = \mathbb{E}(g(Y))$ is constant and $\mathcal{N} = \{1, \ldots, p\}$. However, we rule this trivial case out and always assume $\mathcal{S} \neq \emptyset$.

Definition 1 implies a type of conditional independence. For example, if $p = 2$ and $\mathcal{S} = \{1\}$, then since $\mathbb{E}(g(Y)|\mathbf{X})$ depends only on the signal variable $X^{(1)}$, by the Tower Property:

$$\mathbb{E}(g(Y)|X^{(1)}, X^{(2)}) = \mathbb{E}(g(Y)|X^{(1)}). \tag{1}$$

Notice this does not depend on the relationship between the features. For example, even if $X^{(1)}$ and $X^{(2)}$ were highly correlated, as long as (1) holds almost surely, we still consider $X^{(2)}$ to be a noisy feature. We note that (1) is weaker than the conditional independence assumption typically employed with variable selection (discussed below; see (2)) since it depends upon the choice of g. For example in classification, if the analysts chooses $g(Y) = I\{Y = c_l\}$ for a specific class label c_l of interest, (1) implies a conditional independence of $X^{(2)}$ for class label c_l, but not necessarily for other class labels. Section 4.2 presents an example like this.

Our approach will be to work with the functional ψ, obtained by integration with respect to the conditional distribution of Y given $\mathbf{X}$, but an often used method for variable selection is to directly work with the conditional distribution. This general strategy posits a conditional independence between Y and the noise variables given the signal variables:

$$Y \perp \mathbf{X}^{\backslash(\mathcal{S})} | \mathbf{X}^{(\mathcal{S})}. \tag{2}$$

Let $S \subset \{1, \ldots, p\}$ be a set of variables of interest and the goal is to determine if S contains signal variables. The strategy is then to construct a test statistic $\hat{\theta}_n(S)$ using augmented features $(\mathbf{X}^{(S)}, \mathbf{X}^{\backslash(S)}, \tilde{\mathbf{X}}^{(S)}, \tilde{\mathbf{X}}^{\backslash(S)})$ where $\tilde{\mathbf{X}}^{(S)}$ and $\tilde{\mathbf{X}}^{\backslash(S)}$ are new artificial features. The test is constructed in such a way so that due to conditional independence, $\hat{\theta}_n(S)$ is statistically non-significant if $S \cap \mathcal{S} = \emptyset$.

As mentioned, permutation importance is a popular technique used for variable selection with random forests (Breiman, 2001). This can be seen to be a method that uses artificial data in combination with conditional independence. We will refer to this method as VIMP (variable importance) for short. In VIMP, the feature vector $\mathbf{X}^{\backslash(S)}$ is permuted to obtain $\tilde{\mathbf{X}}^{\backslash(S)}$ and then the predicted value for $(\mathbf{X}^{(S)}, \mathbf{X}^{\backslash(S)})$ is compared to the predicted value for $(\mathbf{X}^{(S)}, \tilde{\mathbf{X}}^{\backslash(S)})$ where this difference should be nearly zero if (2) holds. However, a well known problem with VIMP, to be discussed in detail shortly, is that the permuted sample does not have the same distribution as $\mathbf{X}$ which can lead to flawed variable selection in some settings (Strobl et al., 2008). A new different technique is the idea of knockoffs (Candes et al., 2018). In knockoffs, a simulation according to the distribution of $\mathbf{X}$ is used to obtain $(\tilde{\mathbf{X}}^{(S)}, \tilde{\mathbf{X}}^{\backslash(S)})$ where the artificial data is simulated so as to satisfy

$$(\mathbf{X}^{(S)}, \mathbf{X}^{\backslash(S)}, \tilde{\mathbf{X}}^{(S)}, \tilde{\mathbf{X}}^{\backslash(S)}) \stackrel{d}{=} (\mathbf{X}^{(S)}, \tilde{\mathbf{X}}^{\backslash(S)}, \tilde{\mathbf{X}}^{(S)}, \mathbf{X}^{\backslash(S)}). \tag{3}$$

This is used to compute a knockoff statistic for filtering variables. By making use of (2), the knockoff test statistic can achieve a desired false discovery level. This novel idea avoids the problems of permutation importance, however, this may rely on strong assumptions about the parent distribution and achieving (3) may not be easy to do in all situations.

1.1. A new rule-based approach

To avoid these issues, this paper proposes a new approach, called variable priority (VarPro), that comes from a completely different angle. Let S again represent the set of target variables of interest. Given a rule ζ, VarPro calculates a sample averaged estimator $\hat{\theta}_n(\zeta)$ for the target functional ψ of interest by using the data in ζ's neighborhood. Then a *released rule* ζ^S is constructed, by removing by removing any contraints on the indices in S, so that its sample averaged estimator $\hat{\theta}_n(\zeta^S)$ satisfies $\lim_{n\to\infty} |\hat{\theta}_n(\zeta) - \hat{\theta}_n(\zeta^S)| \stackrel{\text{p}}{\to} 0$ only if S

contains noisy variables. Many existing methodologies for variable importance rely in some form on either resampling, or refitting models, which can introduce finite sample bias, or they make use of artificial data as described above. The merit of this new approach is that it removes the need for resampling/refitting models and for artificial data creation. Rather, the *released estimator* $\hat{\theta}_n(\zeta^S)$ uses the existing data to estimate ψ over its neighborhood of data points. This is the same as $\hat{\theta}_n(\zeta)$, although the two regions are different due to the way ζ is manipulated to obtain ζ^S. Nevertheless, under certain conditions, due to the large sample properties of averages, and because of conditional independence, the difference between the estimators converges to zero for noisy variables; that is, if $S \cap \mathcal{S} = \emptyset$.

As can be seen, VarPro uses the idea of a *neighborhood* of a rule ζ, and we can think of such a neighborhood as being the region of the feature space obtained by a function R that maps $\zeta \mapsto \mathcal{X}$. In many cases, ζ can be associated with a series of univariate rules. The neighborhood mapped by R in this case is denoted by $R(\zeta) = \{\mathbf{x} \in \mathcal{X} : x^{(1)} \in R(\zeta^{X^{(1)}}), \ldots, x^{(p)} \in R(\zeta^{X^{(p)}})\}$, where $\zeta^{X^{(j)}}$ denotes the univariate rule for feature j. For example, if all the features are continuous, we can imagine a rule with a neighborhood $R(\zeta) = \{\mathbf{x} \in \mathbb{R}^p : a_1 \leq x^{(1)} \leq b_1, \ldots, a_p \leq x^{(p)} \leq b_p\}$. Boundaries like this naturally arise in machine learning methods constructed from decision rules. For example, in the case of decision trees, each branch of the tree represents a rule made up of a sequence of binary decisions, and the neighborhood occupied by that rule is its terminal node. With continuous features, this neighborhood can be described as a p-dimensional rectangle.

Now we describe VarPro in more detail. Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \in \mathcal{X} \bigotimes \mathcal{Y}$ be the i.i.d. learning data from a common distribution $\mathbb{P}$. Observations Y_i for $\mathbf{X}_i \in R(\zeta)$ are used to estimate the conditional mean of $g(Y)$ in $R(\zeta)$ via

$$\hat{\theta}_n(\zeta) = \frac{1}{\#\{\mathbf{X}_i \in R(\zeta)\}} \sum_{\mathbf{X}_i \in R(\zeta)} g(Y_i). \tag{4}$$

Given a set $S \subset \{1, \ldots, p\}$, define the released rule ζ^S for ζ by releasing the restrictions on the coordinates of $\mathbf{X}^{(S)}$ in $R(\zeta)$. The new released region $R(\zeta^S)$ gives the released estimator

$$\hat{\theta}_n(\zeta^S) = \frac{1}{\#\{\mathbf{X}_i \in R(\zeta^S)\}} \sum_{\mathbf{X}_i \in R(\zeta^S)} g(Y_i). \tag{5}$$

The absolute difference $|\hat{\theta}_n(\zeta^S) - \hat{\theta}_n(\zeta)|$ is used for determining the importance of S.

Definition 2. Let $R(\zeta) \subset \mathcal{X}$ be a region of the feature space. To check the importance of the variables $\mathbf{X}^{(S)}$ to the rule ζ, we introduce the concept of a released region $R(\zeta^S)$ for the released rule ζ^S obtained by removing the dependence of ζ on the coordinates $\mathbf{X}^{(S)}$:

$$R(\zeta^S) = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}^{\setminus(S)} \in R(\zeta)\}.$$

In other words, the released region $R(\zeta^S)$ is the set of all $\mathbf{x}$ whose S-coordinate values are unconstrained but with non-S-coordinate values that lie in $R(\zeta)$. In particular, $R(\zeta) \subseteq R(\zeta^S)$.

Fig. 1 provides an illustration for a rule ζ for $\mathcal{X} \subseteq \mathbb{R}^2$ obtained from the branch of a tree formed by splitting $x^{(1)} \leq -0.7$, $x^{(2)} \geq -0.8$, $x^{(2)} \leq 0.7$ and $x^{(1)} \geq -1.95$. We can think of the rule as being a product of indicator functions

$$\zeta = I\{x^{(1)} \leq -0.7\} I\{x^{(2)} \geq -0.8\} I\{x^{(2)} \leq 0.7\} I\{x^{(1)} \geq -1.95\}.$$

The region $R(\zeta)$ for ζ is the rectangle

$$R(\zeta) = \{(x^{(1)}, x^{(2)}) : -1.95 \leq x^{(1)} \leq -0.7, -0.8 \leq x^{(2)} \leq 0.7\}.$$

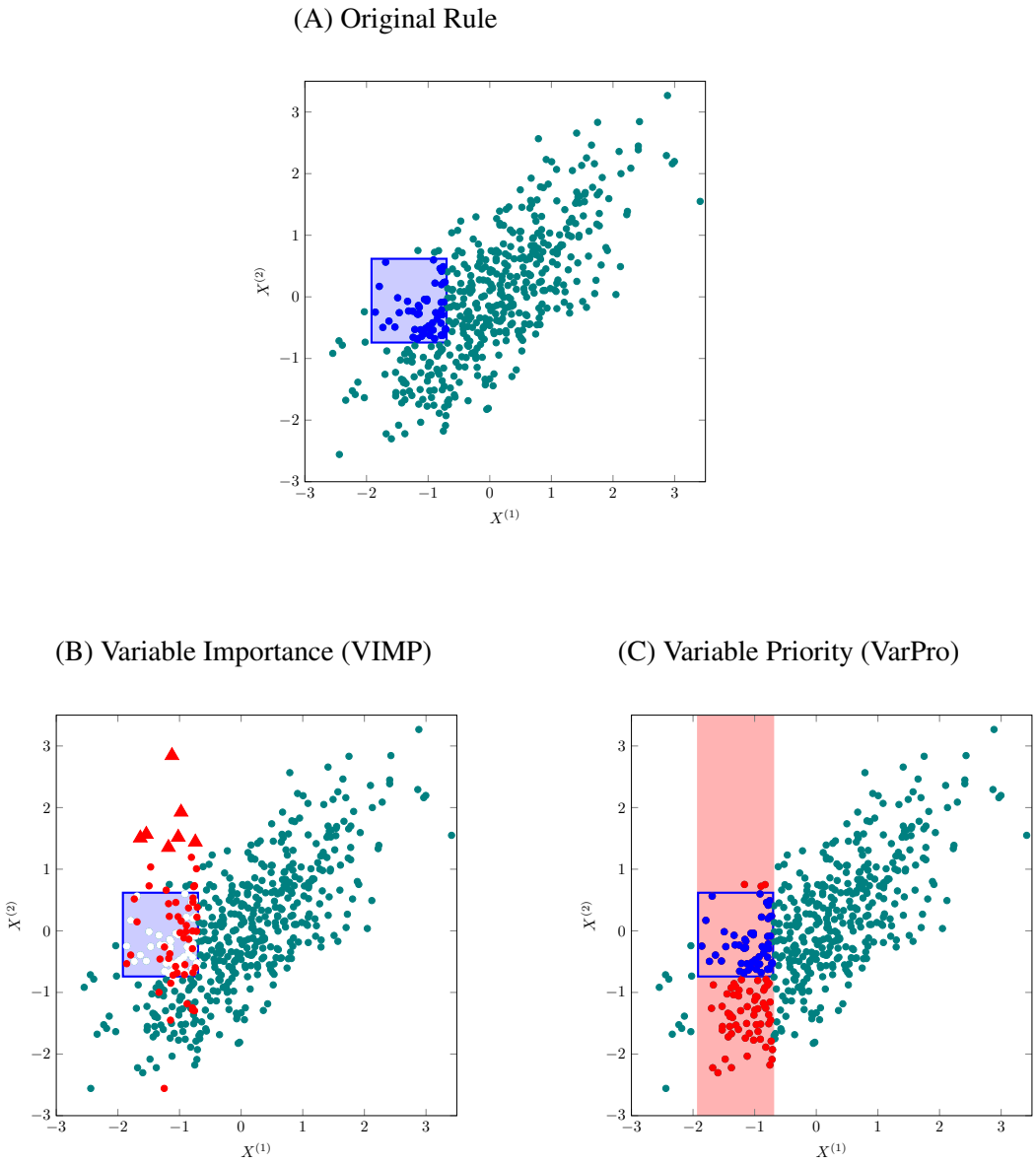

Figure 1: Two-dimensional illustration of how VarPro differs from artificial data methods. (A) The two-dimensional region for ζ takes the form of a rectangle. The data of interest are marked in blue. (B) Permutation variable importance (VIMP) for $X^{(2)}$. The data was permuted along $X^{(2)}$ and data marked in red with triangles identify values that do not match the joint distribution of $\mathbf{X}$. The model-based predicted values $\tilde{y} := \tilde{y}(x^{(1)}, \tilde{x}^{(2)})$ for these artificial points are extrapolated from a region of the feature space that could be from potentially different responses. (C) VarPro for $X^{(2)}$. The original rule is modified to the S-released rule ζ^S (where $S = \{2\}$) shown using a pink background color. No artificial data is created, and importance is defined using the estimator calculated using observed data values in blue compared to the estimator where the new released values in red are additionally used.

To test if $X^{(2)}$ is a noisy variable, we set $S = \{2\}$ and obtain the released rule ζ^S and its released region $R(\zeta^S)$ by releasing the dependence on coordinate $X^{(2)}$:

$$\zeta^S = I\{x^{(1)} \leq -0.7\}I\{x^{(1)} \geq -1.95\}, \quad R(\zeta^S) = \{(x^{(1)}, x^{(2)}) : -1.95 \leq x^{(1)} \leq -0.7\}.$$

For trees, the released rule is equivalent to altering the original rule, equal to a product of binary decision rules, such that whenever a binary decision is to be made based on a variable in S, the decision is always 1.

Fig. 1 shows the region $R(\zeta)$ by the blue rectangle in (A) and the released region $R(\zeta^S)$ by the pink area in (C). The data was permuted on coordinate $X^{(2)}$ in (B) and permuted data $(x^{(1)}, \tilde{x}^{(2)}) \in R(\zeta^S)$ are marked in red. In some of these cases (indicated by triangles) these data deviate strongly from the true distribution of $(X^{(1)}, X^{(2)})$. This creates problems for VIMP. This is because VIMP calculates the importance of $X^{(2)}$ by comparing the test statistic calculated using the observed data in $R(\zeta)$ (blue points in (A)) to that calculated using the permuted data in $R(\zeta^S)$ (red points in (B)). In a regression setting, where the goal is to estimate the conditional mean, this corresponds to averaging observations y_i for $i \in R(\zeta)$ and comparing them to averaged estimated values $\tilde{y}_i := \tilde{y}_i(x_i^{(1)}, \tilde{x}_i^{(2)})$. Unfortunately, since $\tilde{y}_i$ has to be model estimated (since it estimates $\psi(\mathbf{x})$ for $\mathbf{x}$ values not in the training data), this will produce atypical $\tilde{y}_i$ if $\tilde{x}_i^{(2)}$ is from a different region of the data space than the original data. This can result in large VIMP even when $X^{(2)}$ is a noisy variable.

We emphasize that the problem occurring above is not because the learning predictor is deficient but because the learner is being applied to artificially created data that deviates strongly from the true feature distribution. To avoid this, VarPro takes a more direct approach by avoiding the use of artificial data. Instead, it uses the original training data to form estimates of the conditional mean using the observed values $g(y_i)$ via (4) and (5), where for the example above, $g(y) = y$. In (C), $\hat{\theta}_n(\zeta)$ is calculated from observations in the blue box, which is compared to $\hat{\theta}_n(\zeta^S)$ calculated using the data in the pink rectangle, yielding the absolute difference

$$\left|\hat{\theta}_n(\zeta^S) - \hat{\theta}_n(\zeta)\right| = \left| \frac{\sum_{x_i^{(1)} \in I_1} y_i}{\#\{x^{(1)} \in I_1\}} - \frac{\sum_{(x_i^{(1)}, x_i^{(2)}) \in I_1 \bigotimes I_2} y_i}{\#\{(x^{(1)}, x^{(2)}) \in I_1 \bigotimes I_2\}} \right|,$$

where $I_1 = [-1.95, -0.7]$ and $I_2 = [-.8, 0.7]$. If $X^{(1)}$ is a signal feature, but $X^{(2)}$ is a noisy variable, then the two averages should converge to nearly the same value because $\psi(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X}) = \psi(X^{(1)})$ depends only on $X^{(1)}$. On the other hand, if $X^{(2)}$ is a signal feature, then the difference will not necessarily be zero. In fact, under certain conditions, it will be shown that $|\hat{\theta}_n(\zeta^S) - \hat{\theta}_n(\zeta)|$ converges to zero only for noisy variables.

VARPRO KEY IDEA. *Releasing a rule ζ along noisy coordinates S does not change ψ, therefore the difference between the average of g in the rule's neighborhood and the released neighborhood will be asymptotically the same, and subsequently we can expect a zero importance. For signal variables, the opposite happens, and since ψ changes along the released direction, we can expect a difference in average g values in the released and non-released neighborhood, and therefore we can expect a non-zero importance value asymptotically.*

1.2. Contributions of this work and outline of the paper

In Section 2 we build on the previous points to explain how VarPro is able to avoid the problems seen with permutation based importance. Surprisingly, we show that VarPro can be written as type of permuation test. We use this construction to give a theoretical explanation of what goes wrong with permutation importance

due to use of artificial data. While it has been demostrated empirically that permutation importance can be deficient, this new result is important because it explains theoretically why this happens. Section 2 also introduces the idea of an external estimator which is used later in Section 5 for survival analysis. Section 3 provides theoretical justification for VarPro. We prove that VarPro can consistently select variables in very general settings including regression and classification. We consider when the variable is noisy (null case) and when the variable has a signal (alternative case) and show that VarPro is able to consistently identify both types. These results hold without strong assumptions to the underlying model and rely only on relatively mild conditions such as a smoothness property for ψ and certain conditions needing to be met for the underlying rules used for the VarPro statistic. These latter conditions are expected to hold for any reasonably constructed tree structure, thereby making VarPro generally agnostic to the type of rule-based procedure used for the rule generation step. Empirical results demonstrating VarPro's effectiveness are provided in Section 4 using synthetic data for regression and classification and by considering real world data. In Section 5 we extend VarPro to the area of survival analysis, an important emerging area of interest for machine learning. This extension handles the problem of right censoring which causes the response to be potentially unobserved. Real data and high-dimensional simulations are given to illustrate the effectiveness of the proposed extension to survival. Section 6 concludes with a discussion. All proofs and supplementary information are located in the Appendix at the end of the manuscript.

2. Connection of VarPro to permutation tests

To expand on the previous points made in the Introduction, we use the following construction to gain insight into the dangers of using artificial data. It turns out that we can actually write the VarPro estimator as a specialized type of permutation test. We will use this construction to contrast this to model-based importance to explain what goes wrong when using artificial data and to explain how VarPro avoids this. This will also motivate the idea of using an external estimator, which is something we will take advantage of in Section 5 when we consider survival analysis.

Without loss of generality, assume that $\mathbf{x}$ is reordered so that $\mathbf{x} = (\mathbf{x}^{(S)}, \mathbf{x}^{\backslash(S)})$. Consider all possible permutations of the training data $\{(y_i, \mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)})\}_{1\le i,j\le n}$. Then it will be shown (see below) that

$$\frac{\sum_{i=1}^n g(y_i) I\{\mathbf{x}_i \in R(\zeta^S)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\}} = \frac{\sum_{i=1}^n \sum_{j=1}^n g(y_i) I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta)\}}{\sum_{i=1}^n \sum_{j=1}^n I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta)\}}. \tag{6}$$

The left-hand side is the VarPro released estimator $\hat{\theta}_n(\zeta^S)$ described in (5), whereas the right-hand side is an estimator using the permuted data.

Obviously in practice the permuted estimator on the right side of (6) is impractical since it involves $O(n^2)$ calculations and would be an inefficient means for calculating $\hat{\theta}_n(\zeta^S)$. However, its use will be in describing why permutation VIMP generally does not satisfy an identity like (6) which will then highlight the main problem.

First, however, let us explain why (6) holds. Let $\zeta^{\backslash(S)}$ be the rule that relaxes the constraints of ζ on the indices not in S. We have the following identity for rules expressible as products $\zeta = \prod_{s=1}^p I\{x^{(s)} \in I_s\}$ where $I_s \subseteq \mathbb{R}$ are real-valued intervals (this is like the example described in Fig. 1):

$$I\{\mathbf{x} \in R(\zeta)\} = I\{\mathbf{x}^{\backslash(S)} \in R(\zeta)\} I\{\mathbf{x}^{(S)} \in R(\zeta)\} = I\{\mathbf{x} \in R(\zeta^S)\} I\{\mathbf{x} \in R(\zeta^{\backslash(S)})\}. \tag{7}$$

The right side follows by Definition 2 since $\mathbf{x} \in R(\zeta^S)$ implies $\mathbf{x}^{\backslash(S)} \in R(\zeta)$ and $\mathbf{x} \in R(\zeta^{\backslash(S)})$ implies $\mathbf{x}^{(S)} \in R(\zeta)$.

Therefore for rules satisfying (7), notice that

$$\begin{aligned} I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta)\} &= I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta^S)\} I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta^{\backslash(S)})\} \\ &= I\{\mathbf{x}_i \in R(\zeta^S)\} I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})\} \end{aligned}$$

where the last line is because $(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta^S)$ depends only on $\mathbf{x}_i^{\backslash(S)}$ and $(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta^{\backslash(S)})$ depends only on $\mathbf{x}_j^{(S)}$. Hence we have

$$\begin{aligned} &\frac{\sum_{i=1}^n \sum_{j=1}^n g(y_i) I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta)\}}{\sum_{i=1}^n \sum_{j=1}^n I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta)\}} \\ &\qquad = \frac{\sum_{i=1}^n \sum_{j=1}^n g(y_i) I\{\mathbf{x}_i \in R(\zeta^S)\} I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})}{\sum_{i=1}^n \sum_{j=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\} I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})} \\ &\qquad = \frac{\sum_{i=1}^n g(y_i) I\{\mathbf{x}_i \in R(\zeta^S)\} \sum_{j=1}^n I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\} \sum_{j=1}^n I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})} \\ &\qquad = \frac{\sum_{i=1}^n g(y_i) I\{\mathbf{x}_i \in R(\zeta^S)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\}} := \hat{\theta}_n(\zeta^S). \end{aligned}$$

The first identity follows because of (7). The last identity is due to the cancellation of the common term in the numerator and denominator which is directly related to working with $g(y_i)$.

To make this last point more clear let us now consider how a model-based procedure such as VIMP would be applied in this setting. Recall that unlike VarPro, VIMP does not use the actual observed response value to estimate ψ but instead applies a model-based estimator. Let ψ_n be this estimator. For example, this could be the ensemble estimator from a random forest analysis or a boosted tree estimator using gradient boosting. Then using the exact same permuted data as above, the VIMP estimator is

$$\begin{aligned} \tilde{\theta}_{\text{VIMP}}(\zeta^S) &= \frac{\sum_{i=1}^n \sum_{j=1}^n \psi_n(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta)\}}{\sum_{i=1}^n \sum_{j=1}^n I\{(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) \in R(\zeta)\}} \\ &= \frac{\sum_{i=1}^n \sum_{j=1}^n \psi_n(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)}) I\{\mathbf{x}_i \in R(\zeta^S)\} I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\} \sum_{j=1}^n I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})} \end{aligned}$$

where the last identity is due to (7). Notice, however, that the cancellation that occured previously in the numerator and denominator is no longer guaranteed to hold and it is not true that $\tilde{\theta}_{\text{VIMP}}(\zeta^S)$ is the same as the non-permuted estimator

$$\tilde{\theta}_n(\zeta^S) = \frac{\sum_{i=1}^n \psi_n(\mathbf{x}_i) I\{\mathbf{x}_i \in R(\zeta^s)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^s)\}} \tag{8}$$

which is the analog to $\hat{\theta}_n(\zeta^S)$ since it replaces the observed value $g(y_i)$ with the model estimated value $\psi_n(\mathbf{x}_i)$. Just like $\hat{\theta}_n(\zeta^S)$, the estimator $\tilde{\theta}_n(\zeta^S)$ leads to consistent variable selection (to be shown in Theorem 5) so the equality would show that the model-based permutation importance has good properties. But this cancellation occurs and the two estimators become the same only if

$$\psi_n(\mathbf{x}) = \psi_n(\mathbf{x}^{\backslash(S)}) \tag{9}$$

because then

$$\begin{aligned} \tilde{\theta}_{\text{VIMP}}(\zeta^S) &= \frac{\sum_{i=1}^n \psi_n(\mathbf{x}_i^{\backslash(S)}) I\{\mathbf{x}_i \in R(\zeta^S)\} \sum_{j=1}^n I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\} \sum_{j=1}^n I\{\mathbf{x}_j \in R(\zeta^{\backslash(S)})} \\ &= \frac{\sum_{i=1}^n \psi_n(\mathbf{x}_i^{\backslash(S)}) I\{\mathbf{x}_i \in R(\zeta^S)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\}} := \tilde{\theta}_n(\zeta^S). \end{aligned}$$

If S consists entirely of noisy variables then $\psi(\mathbf{x}) = \psi(\mathbf{x}^{\backslash(S)})$, so in this case (9) is asserting that the model-based estimator has correctly eliminated all noisy variables. This is a very strong requirement and is asking a lot from the underlying procedure since after all this would mean that the procedure has achieved perfect variable selection on its own. This type of finite sample oracle behavior is obviously difficult to achieve, and especially so when the data is correlated. In fact, it is well known that machine learning methods have a lot of difficulty in selecting variables in such settings. We will present empirical evidence of this happening later on.

So since it is not reasonable to expect (9) to hold, this means $\tilde{\theta}_{\text{VIMP}}(\zeta^S)$ is calculated using model estimated values $\psi_n(\mathbf{x}_j^{(S)}, \mathbf{x}_i^{\backslash(S)})$ which could become very poor when the permuted data deviates strongly from the true $\mathbf{X}$ distribution. This is problematic and is at the root of the issues with the VIMP permutation method.

While the above arguments have identified the issue of using ψ_n with $\tilde{\theta}_{\text{VIMP}}(\zeta^S)$, we want to emphasize that this does not detract from the potential usefulness of an external estimator. A point we wish to explore is the idea of using the external estimator in a more coherent way than VIMP. We argue that the non-permuted estimator $\tilde{\theta}_n(\zeta^S)$ defined in (8) which uses ψ_n presents one such opportunity. After all, we know that machine learning methods like gradient boosting and random forests can produce highly accurate estimators in many settings. VIMP misuses ψ_n by applying it to data not representative of the true $\mathbf{X}$ distribution which can produce biased estimation since it will be difficult to estimate values over features not well supported by the training data. In contrast, the non-permuted estimator $\tilde{\theta}_n(\zeta^S)$, which was introduced as an analog to the VarPro estimator, only applies ψ_n to the training data. This is why it does not equal permutation VIMP and why it represents a legitimate alternate way to proceed.

In fact, in Section 5 we will present an estimator like this for survival analysis problems where the response may not be observed due to censoring. In such settings, an external estimator makes sense. However, at the same time, what we are also saying is that if the response is observed, then we prefer to use the VarPro estimator for importance

$$\left|\hat{\theta}_n(\zeta^S) - \hat{\theta}_n(\zeta)\right| = \left|\frac{\sum_{i=1}^n g(y_i) I\{\mathbf{x}_i \in R(\zeta^S)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\}} - \frac{\sum_{i=1}^n g(y_i) I\{\mathbf{x}_i \in R(\zeta)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta)\}}\right|$$

which makes use of the actual observed responses rather than the external estimator for importance

$$\left|\tilde{\theta}_n(\zeta^S) - \tilde{\theta}_n(\zeta)\right| = \left|\frac{\sum_{i=1}^n \psi_n(\mathbf{x}_i) I\{\mathbf{x}_i \in R(\zeta^S)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta^S)\}} - \frac{\sum_{i=1}^n \psi_n(\mathbf{x}_i) I\{\mathbf{x}_i \in R(\zeta)\}}{\sum_{i=1}^n I\{\mathbf{x}_i \in R(\zeta)\}}\right|.$$

The former leads to a simpler procedure that is based on averages of independent observations which is consistent under relatively mild conditions and also it is easily applied to different g functions without the need to fit a learning method each time g is changed by the researcher.

3. Methodology

How does one construct a rule ζ to identify variables informative for ψ? In practice, we have at our disposal all types of procedures to choose from that produce rules that can work with VarPro, including for example simple decision rules (Tan et al., 2005), rule learning (Fürnkranz, 1997), trees (Breiman et al., 1984), Bayesian trees (Nuti et al., 2021), Bayesian additive regression trees (Chipman et al., 2010), Bayesian forests (Liu et al., 2021) and random forests (Breiman, 2001). In the examples presented in this paper, the rules are chosen by randomly selecting branches from a tree. In general, the VarPro method can practically be used with any type of rule-based learner, and with some creativeness, the method could be used with other procedures as well.

We will assume hereafter that we have constructed a rule, or more generally, a collection of rules for a specific problem. These rules are assumed to be constructed independently of the data used by VarPro, and therefore without loss of generality, it will be assumed that all rules are deterministic. For each n, let $\zeta_{n,1}, \ldots, \zeta_{n,K_n}$ denote these rules. Notice that the number of rules K_n can vary with n and also that the rules themselves are allowed to change with n. For a given rule ζ, the rule-based estimator is

$$\hat{\theta}_n(\zeta) = m_n(\zeta)^{-1} \sum_{i=1}^{n} g(Y_i) I\{\mathbf{X}_i \in R(\zeta)\}, \quad m_n(\zeta) = \sum_{i=1}^{n} I\{\mathbf{X}_i \in R(\zeta)\},$$

which is a slightly more compact way of writing $\hat{\theta}_n$ than (4) (notice that $m_n(\zeta)$ equals the sample size of a rule ζ). In a likewise fashion, we can define $\hat{\theta}_n(\zeta^S) = m_n(\zeta^S)^{-1} \sum_{i=1}^{n} g(Y_i) I\{\mathbf{X}_i \in R(\zeta^S)\}$.

For notational ease, hereafter we let $m_{n,k} = m_n(\zeta_{n,k})$, $m^S_{n,k} = m_n(\zeta^S_{n,k})$ and $R_{n,k} = R(\zeta_{n,k})$, $R^S_{n,k} = R(\zeta^S_{n,k})$. The importance for S is defined as the weighted averaged difference

$$\begin{aligned} \Delta_n(S) &= \sum_{k=1}^{K_n} W_{n,k} |\hat{\theta}_n(\zeta^S_{n,k}) - \hat{\theta}_n(\zeta_{n,k})| \qquad (10) \\ &= \sum_{k=1}^{K_n} W_{n,k} \left| \frac{1}{m^S_{n,k}} \sum_{i=1}^{n} g(Y_i) I\{\mathbf{X}_i \in R^S_{n,k}\} - \frac{1}{m_{n,k}} \sum_{i=1}^{n} g(Y_i) I\{\mathbf{X}_i \in R_{n,k}\} \right| \end{aligned}$$

where $0 \leq W_{n,k} \leq 1$ are weights (deterministic or random) such that $\sum_{k=1}^{K_n} W_{n,k} = 1$. In the following sections, we study the asymptotic properties of (10), breaking this up into the case of noisy and signal variables.

3.1. Consistency for noisy variables

The following theorem shows, that under mild assumptions, VarPro is consistent for noisy variables.

Theorem 3. *Assume (A1), (A2), (A3) and (A4). If $K_n \leq O(\log n)$ and $m_{n,k} \geq m_n = n^{1/2}\gamma_n$ where $\gamma_n \uparrow \infty$ at a rate faster than $\log n$, then $\Delta_n(S) \xrightarrow{p} 0$ if $S \subseteq \mathcal{N}$.*

The assumptions used by Theorem 3 are given in Appendix C and are discussed in detail there. For convenience, however, we summarize some of the key points here and also add some further information. Assumption (A1) is an integrability condition needed for estimators to converge. Assumptions (A2) and (A4) are smoothness and boundedness conditions for ψ and its derivative. Roughly speaking, since $m_n \gg n^{1/2} \to \infty$, then due to the large sample properties of averages, $\hat{\theta}_n(\zeta^S_{n,k}) - \hat{\theta}_n(\zeta_{n,k})$ asympotically equals $\mathbb{E}(\psi(\mathbf{X})|\mathbf{X} \in R^S_{n,k}) - \mathbb{E}(\psi(\mathbf{X})|\mathbf{X} \in R_{n,k})$, and in order for this to provide useful information about a variable's importance, it is necessary for ψ to have certain smoothness properties. Assumptions (A2) and (A4) ensure that this happens. Capping the number of rules by $K_n \leq O(\log n)$ ensures that the approximation holds uniformly. Assumption (A3) is the only condition that specifically refers to rule neighborhoods. It requires that the diameter of a neighborhood shrinks to zero uniformly. However, the shrinkage rate is left unspecified, and the shrinkage only has to hold along the signal feature coordinates.

The shrinkage assumption and the rate condition m_n on the sample size of a neighborhood are expected to hold for rules obtained from any reasonably constructed procedure. Similar assumptions are fairly standard in the asymptotic analysis of trees. For tree consistency, a typical requirement is that the diameter of a cell converges to 0 and the number of points in the cell converges to ∞ in probability. See Theorem 6.1 from Devroye et al. (2013) and Theorem 4.2 from Györfi et al. (2002).

In fact, conditions (A3), and the rate condition m_n, hold even for a random tree construction as we now show. Assume that $\mathcal{X} = [0,1]^p$. Consider a tree construction where at the start of split $k \geq 1$, each of the

k leaves of the tree is a p-dimensional rectangle (when $k = 1$, the one leaf equals the root node of the tree, $[0, 1]^p$). Of these k leaves (rectangles), one is selected at random and its longest side is split at a random point; thus yielding at the end of step k, two new leaves, which are rectangles of reduced volume from the original rectangle. The tree construction is repeated for a total of K_n splits, yielding $K_n + 1$ branches which are the rules.

If $K_n = \log n$, then it can be shown that the following holds in probability (see the proof of Theorem 2 of Biau et al. (2008)):

(i) The number of data points in a rectangle is greater than $m_n = n^\delta \log n$ for any $1/2 < \delta < 1$.

(ii) The mean length of the largest side of a rectangle is less than $\mathbb{E}\left[(3/4)^{T_n}\right]$ where $T_n \xrightarrow{\text{p}} \infty$ does not depend on the specific rectangle; hence the diameter of each rectangle converges uniformly to zero.

Therefore, (ii) shows (A3) holds. Also, as somewhat of an added bonus, (i) shows that $m_n \gg n^{1/2}$ achieves the lower bound required by Theorem 3. This is an interesting consequence of the random construction and allows the bound to be achieved without supervision. Notice that the number of tree cuts K_n is order $\log n$. Since this naturally forces data to pile up in the terminal nodes we should expect $m_n \to \infty$. If tree node sizes are evenly distributed, then $m_n \asymp n/\log n$ which helps explain why the rate condition for m_n naturally holds for the random tree.

3.2. Limiting behavior for signal variables

For the analysis of signal variables, we assume that neighborhoods are rectangles. This is made for technical reasons to simplify arguments but does not limit applications of VarPro in practice. Thus, the neighborhood for a rule ζ can be written as $R(\zeta) = \bigotimes_{j=1}^{p} I_j$ where $I_j \subseteq \mathbb{R}$ are real-valued intervals. For notational simplicity, suppose that the first $|\mathcal{S}|$ coordinates are signal variables and the remaining $|\mathcal{N}|$ coordinates are noisy variables. Therefore, we can write $R(\zeta) = A \bigotimes B$ where $A = \bigotimes_{j=1}^{d} I_j$, $B = \bigotimes_{j=d+1}^{p} I_j$ and $d = |\mathcal{S}| \leq p$ is the number of signal variables.

Theorem 4. *Assume that the region for each rule is a rectangle contained within the connected space $\mathcal{X} \subseteq \mathbb{R}^p$, then under the same conditions as Theorem (3), if $S = \{s\}$ is a signal variable,*

$$\Delta_n(s) = \Big(1 + o_p(1)\Big) \sum_{k=1}^{K_n} W_{n,k} \Big| \mathbb{E}\left[\big(\psi^s_{n,k}(X^{(s)}) - \psi^s_{n,k}(x^{(s)}_{n,k})\big) \,\big|\, \mathbf{X} \in R^s_{n,k} \right] \Big| + o_p(1),$$

where $\psi^s_{n,k}(z) = \psi(x^{(1)}_{n,k}, \ldots, x^{(s-1)}_{n,k}, z, x^{(s+1)}_{n,k}, \ldots, x^{d}_{n,k})$ and $\mathbf{x}_{n,k} = (x^{(1)}_{n,k}, \ldots, x^{(p)}_{n,k})' \in R_{n,k}$ is a fixed point in each rectangle as defined in (A4).

To help understand Theorem 4 consider the case when ψ is an additive function, $\psi(\mathbf{x}) = \sum_{j=1}^{d} \phi_j(x^{(j)})$. Then $\psi^s_{n,k}(x) = \sum_{j \in \mathcal{S} \setminus s} \phi_j(x^{(j)}_{n,k}) + \phi_s(x)$ and

$$\mathbb{E}\left[\Big(\psi^s_{n,k}(X^{(s)}) - \psi^s_{n,k}(x^{(s)}_{n,k})\Big) \,\big|\, \mathbf{X} \in R^s_{n,k} \right] = \mathbb{E}\left[\Big(\phi_s(X^{(s)}) - \phi_s(x^{(s)}_{n,k})\Big) \,\big|\, \mathbf{X} \in R^s_{n,k} \right]. \tag{11}$$

This is the average difference between $\phi_s(X^{(s)})$ for a sampled point $\mathbf{X} \in R^s_{n,k}$ compared with $\phi_s(x^{(s)}_{n,k})$ for a fixed point $\mathbf{x}_{n,k} \in R_{n,k}$. Because $X^{(s)}$ is unrestricted, and can take any value in the s coordinate direction of $\mathcal{X}$, $\phi_s(X^{(s)}) - \phi_s(x^{(s)}_{n,k})$ should be non-zero on average. Also, even if this equals zero by chance, keep in mind this applies to each rectangle $R_{n,1}, \ldots, R_{n,K_n}$, thus we can expect an average nonzero effect when summing over all rules. This also explains why it is better to use many rules than just one rule. In fact, Theorem 4 allows for up to $O(\log n)$ rules.

It is helpful to note Theorem 4 continues to hold when S contains more than one signal feature: then $X^{(s)}$ is simply replaced with $\mathbf{X}^{(S)}$. For the additive model, (11) becomes $\sum_{s\in S}\mathbb{E}[(\phi_s(X^{(s)})-\phi_s(x^{(s)}_{n,k}))\,|\,\mathbf{X}\in R^S_{n,k}]$. In particular, if $S=\mathcal{S}$ equals the entire set of signal features, then $R^S_{n,k}$ releases along all signal coordinate directions, leaving only the noisy features in the conditional expectation. For simplification, if a rectangle is only constrained for signal features, conditioning is removed, so that $\mathbb{E}[\psi(\mathbf{X})-\psi(\mathbf{x}_{n,k})]$. Let $\mathbb{Q}_n$ be the distribution with support $\{\mathbf{x}_{n,1},\ldots,\mathbf{x}_{n,K_n}\}$ and probabilities $\{W_{n,1},\ldots,W_{n,K_n}\}$. Summing over all rectangles and taking absolute values gives

$$\sum_{k=1}^{K_n} W_{n,k}\big|\mathbb{E}[\psi(\mathbf{X})-\psi(\mathbf{x}_{n,k})]\big| = \int |C-\psi(\mathbf{x})|\,d\mathbb{Q}_n(\mathbf{x}), \quad \text{where } C=\mathbb{E}(\psi(\mathbf{x})).$$

This equals the $\mathbb{Q}_n$-mean absolute deviation of ψ compared to the mean for ψ taken with respect to $\mathbb{P}$.

To further expand on the argument for using many rules, data was drawn from the non-linear regression model

$$Y|\mathbf{x}=\psi(\mathbf{x})+\varepsilon, \qquad \text{where } \psi(\mathbf{x})=10\sin(x^{(1)}x^{(2)})$$

where ε is standard normal. In total, $p=100$ features were drawn uniformly from $[0,1]^p$ and then transformed so as to have a correlation of $\rho=.8$. Thus, there are two signal variables and all other variables are noisy and all variables have the same correlation. A sample size of 1000 was drawn.

Panel (A) of Fig. 2 displays a contour plot of $\psi(\mathbf{x})$ as a function of the two signal variables. Superimposed on (A) are the observed data points, which as expected exhibit high correlation due to the way the data was simulated. The bottom two panels display estimated values for

$$\left|\mathbb{E}\left[\left(\psi^s_{n,k}(X^{(s)})-\psi^s_{n,k}(x^{(s)}_{n,k})\right)\,\middle|\,\mathbf{X}\in R^s_{n,k}\right]\right| \tag{12}$$

for $s=\{1\}$ and $s=\{2\}$ in (B) and (C) respectively. The circles shown are located at $(x^{(1)}_{n,k},x^{(2)}_{n,k})$ for each region $R_{n,k}$ with the radius and color scaled proportional to the size of the asymptotic value (12) indicated by our theory. These values were obtained using the VarPro algorithm to be described in the next section using 500 trees where each tree produced up to 75 rules. In (B), values (12) that are especially large are achieved for large circles near the edges since releasing $x^{(1)}$ creates a released region that traverses through high and low values of ψ (as an example, see the black arrow pointing from left to right). The same is true for $x^{(2)}$ in panel (C); for example, large circles are observed near the bottom and releasing $x^{(2)}$ yields a released region traversing through areas where ψ changes rapidly (see black arrow pointing up). Keep in mind that besides these high value contributions, all values are used by VarPro when forming the importance estimate which is an average of all contributions from all points. This is the advantage of using many rules, as this allows VarPro to blanket the feature space so as to achieve a higher chance of acquiring a significant value for a signal variable. Thus, the general message of Theorem 4 is clear: if enough rules are constructed, each shrinking along the signal coordinate directions, then the variable importance for a signal variable will be an average of the absolute difference between ψ evaluated along the released direction compared with ψ evaluated inside a shrinking region, thus resulting in a large importance value.

4. Empirical results

In this section, we study the performance of VarPro in regression and classification problems. The following computational procedure was used.

1. First, the data of size N is split into two parts of proportion αN and $n=(1-\alpha)N$ for some $0<\alpha<1$. Typically, αN is the larger portion. For example, with rules constructed by trees as used here, we use $\alpha=.632$ which is the bootstrap sample size employed by tree-based procedures

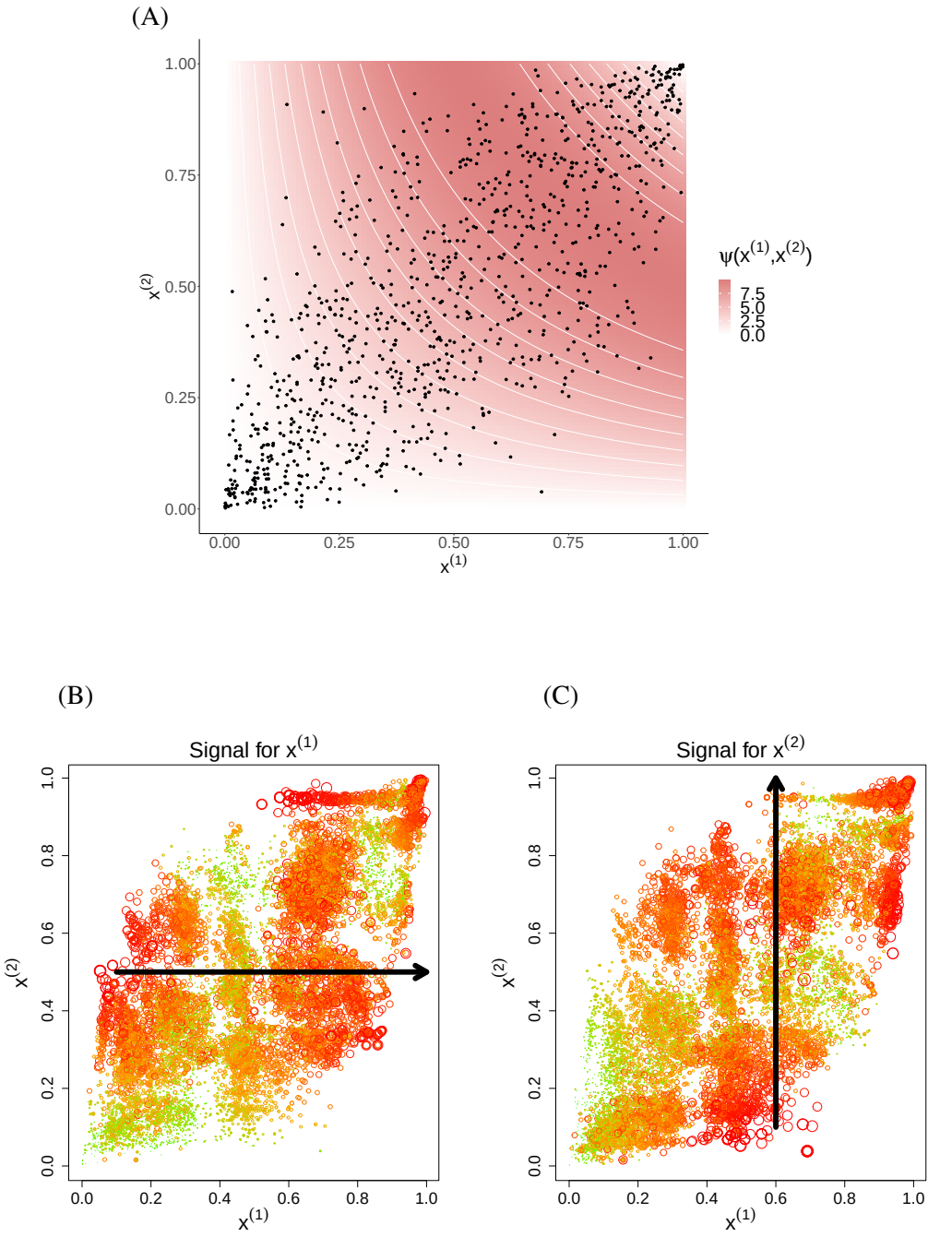

Figure 2: Nonlinear regression simulation $p = 100$ with two signal variables, $x^{(1)}, x^{(2)}$, and all other variables are noisy and all variables have a $\rho = .8$ correlation. (A) Contours for the true regression function $\psi(\mathbf{x}) = 10\sin(x^{(1)}x^{(2)})$ with observed data superimposed as points. (B), (C) Center of rules indicated by circles with radius of circle and color scaled to size of the asymptotic value (12) for a signal variable indicated by our theory. Black arrows show the direction of released regions that result in especially large values (12) due to the large change in ψ in that direction.

such as random forests. The larger portion αN is used for the rule generation step. The smaller portion n is used for the VarPro calculation.

2. A lasso model (Tibshirani, 1996) is fit using the αN data and the absolute coefficient value for each standardized variable is recorded (for classification, the mean absolute value across class labels is used). This is used for the split-weight for a variable in the tree-rule generation step to be discussed next (the point of this is so that the rule generation procedure is guided to use signal features as much as possible to accommodate assumption (A3)). We noticed in certain cases (such as nonparametric models) that the lasso could become overly sparse due to model misspecification. In some cases, it even happens that all coefficients are zero. Therefore, additionally, a small collection of shallow trees are constructed. The relative split frequency of a variable is recorded, and this value is added to the lasso coefficient estimate for the finalized split-weight.
3. Using the split-weights obtained from (b), a tree is constructed using random feature selection (again over the αN data). At each node of the tree, the tree is split by selecting a random subset of features to split on, with features randomly chosen with probability proportional to their split-weight. From the finalized tree, a random K_n of tree branches are selected, yielding rules $R_{n,1}, \ldots, R_{n,K_n}$.
4. Once the rules are obtained, the remaining data of size n is used for calculating $\Delta_n(s)$ using (10). Weights are defined by $W_{n,k} = m_{n,k} / \sum_{k=1}^{K_n} m_{n,k}$.

For the purpose of reducing variance and producing a "standardized" importance value, the rule generation and VarPro estimation is repeated $B > 1$ times. This yields an averaged $\Delta_n^*(s)$ and sample variance $\text{Var}_n^*(\Delta_n(s))$ that is converted to a standardized estimate:

$$\text{Standardized importance for variable } X^{(s)} = \frac{\Delta_n^*(s)}{\sqrt{\text{Var}_n^*(\Delta_n(s))}}, \quad s = 1, \ldots, p. \tag{13}$$

Variable $X^{(s)}$ is selected if (13) exceeds a cutoff value Z_0. This cutoff can be pre-chosen (for example $Z_0 = 2$) or selected by out-of-sample validation. Examples using both strategies are presented in the paper.

4.1. Regression

The performance of VarPro was tested using synthetic data from both linear and nonlinear regression models (see Appendix E). Sample size and dimension was set to $N = 2000$ and $p = 40$. Both uncorrelated and correlated features were considered. For the correlated scenarios, features retained the same marginal distribution as the uncorrelated scenarios, but were transformed using a copula so as to all have a correlation of $\rho = 0.9$. This was done for all simulations except *lm* and *lmi2* where the 15 signal features $X^{(1)}, \ldots, X^{(15)}$ were correlated within blocks of size 5 (1–5, 6–10 and 11–15).

Each run of VarPro used $B = 500$ trees with $K_n = 75$ rules extracted for each tree. Trees were split by random feature selection ($p/3$ candidate features selected for each split). Methods used for comparison included the lasso (Tibshirani, 1996), knockoffs (Candes et al., 2018), generalized boosted regression modeling (GBM) with trees; i.e. gradient boosted trees (Friedman, 2001) and permutation importance (Breiman, 2002) (referred to as Breiman-Cutler variable importance, abbreviated as BC-VIMP). These methods were implemented using the R packages `glmnet` (Friedman et al., 2010), `knockoff` (Patterson and Sesia, 2022), `gbm` (Greenwell et al., 2020) and `randomForestSRC` (Ishwaran and Kogalur, 2023). The lasso regularization parameter and GBM number of boosting iterations were obtained using 10-fold cross-validation. Knockoff test statistics were calculated by: (1) the difference between a lasso coefficient estimate and its knockoff coefficient estimate using `glmnet` with lasso parameter obtained by 10-fold validation; and (2) the difference between random forest impurity importance and knockoff impurity importance.

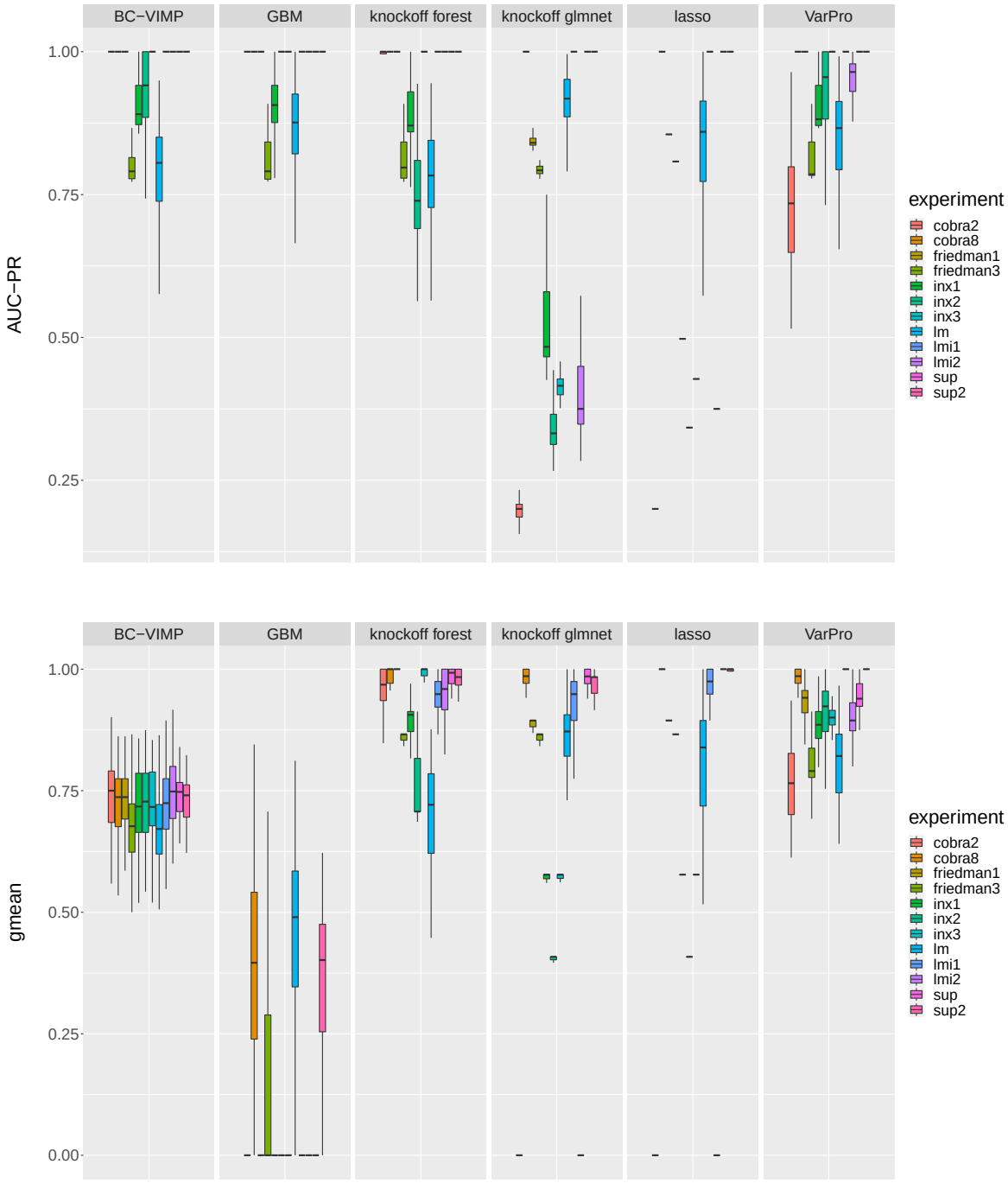

Figure 3: Area under the precision recall curve (AUC-PR) and gmean (geometric mean of TPR and TNR) for regression simulations where variables are uncorrelated.

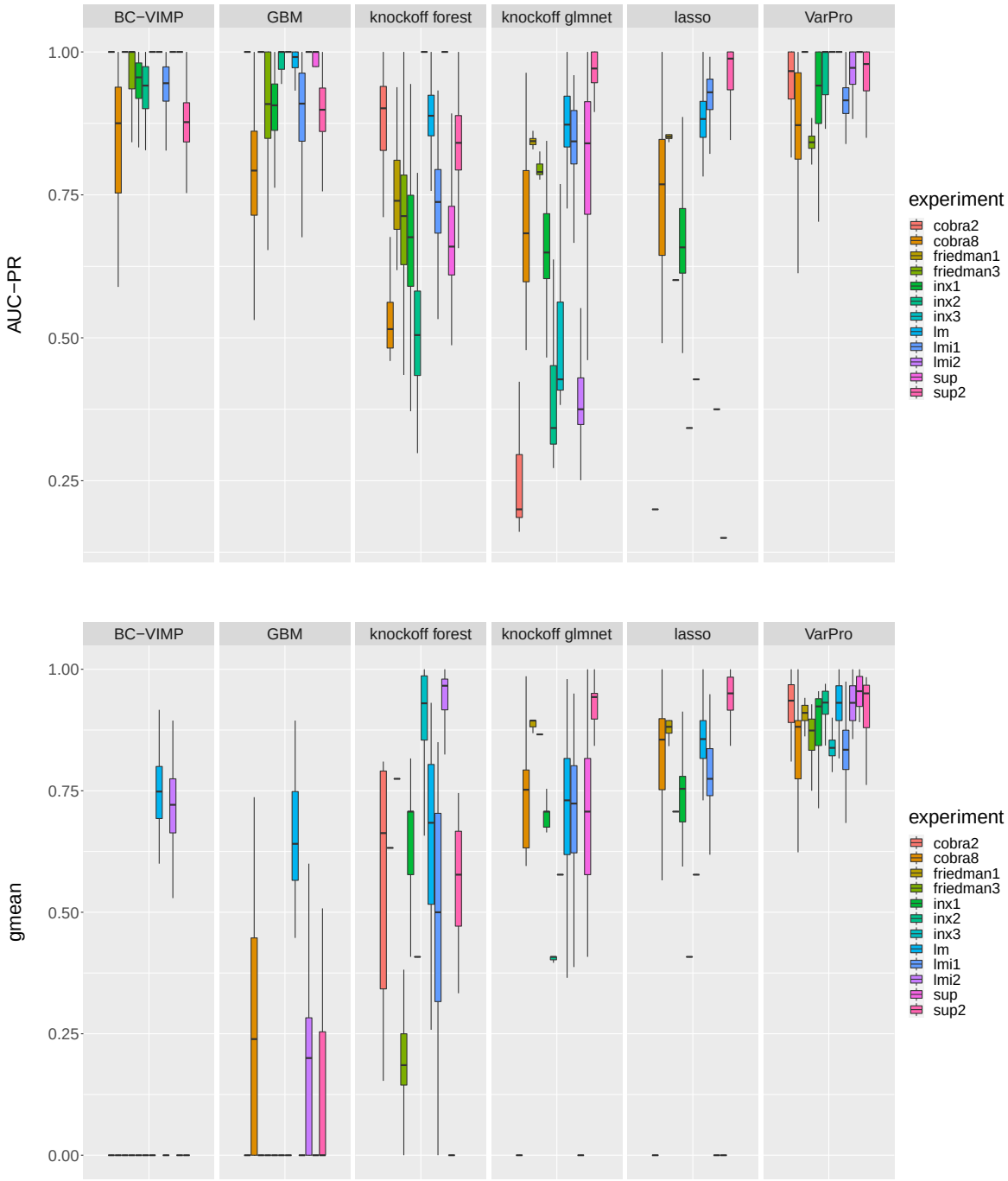

Figure 4: Similar to Fig. 3 but now using correlated variables.

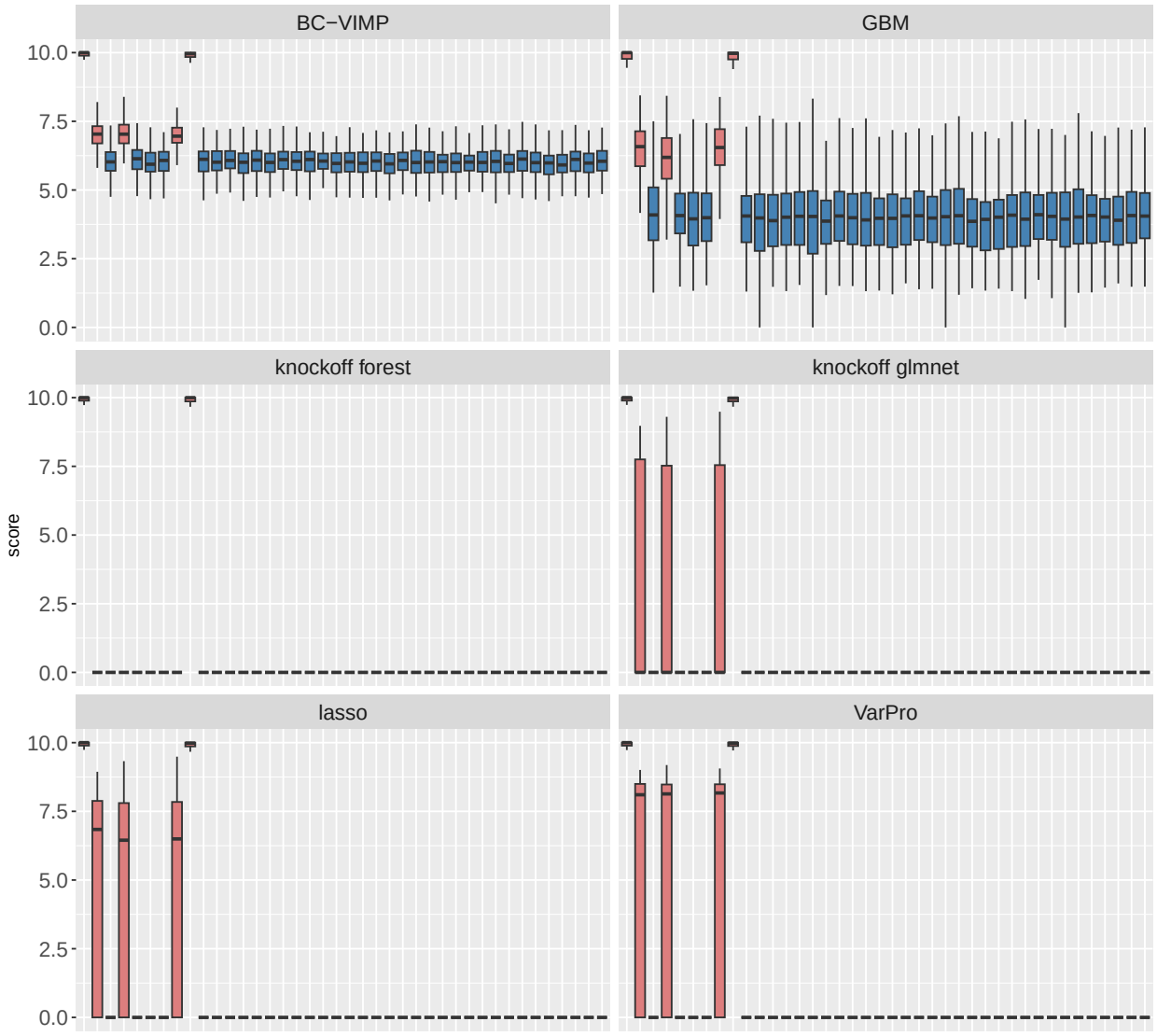

Figure 5: Threshold scores used for gmean calculation for *cobra8* simulation with correlated features. Scores have been scaled and monotonically transformed to have the same range where a zero score means a method classifies the variable as noise. True signal variables $X^{(1)}, X^{(2)}, X^{(4)}, X^{(8)}, X^{(9)}$ are shown in red while noise variables are shown in blue.

Performance of a method for selecting variables was assessed by area under the precision recall curve (AUC-PR) and the geometric mean (gmean), defined as the geometric mean of the TPR (true positive rate for selecting signal variables; i.e. the sensitivity) and TNR (true negative rate for selecting noise variables; i.e. the specificity). These metrics were selected since nearly all simulations have far more noise variables than signal variables, which creates an imbalanced classification problem. The AUC-PR summarizes the trade-off between the TPR (referred to as the recall) and the positive predictive value (refered to as the precision) for a method as the selection threshold is varied over all possible values. Its advantage is that it does not depend on the imbalanced ratio (frequency of signal to noise variables) and is thus suitable for imbalanced problems. The gmean measures the balance between classification performances on both the majority and minority classes and is therefore also an appropriate metric for imbalanced problems (Kubat et al., 1997).

To calculate AUC-PR, which does not require a threshold value, each method's output was converted to a continuous score with larger positive values indicating more likelihood of being a signal variable. For lasso this was the absolute value of the coefficient estimates; for GBM this was the relative influence (a non-negative value); for BC-VIMP this was the permutation importance (a value that can be both positive and negative); for knockoffs this was the absolute value of the knockoff test statistic; and for VarPro this was the standardized importance value (13) (a non-negative value).

To calculate gmean, a threshold value for selecting variables was required. A threshold of zero was used for all methods and a suitable score value was defined for this purpose. Because lasso achieves estimation and model selection simultaneously, this was the same score value used for the AUC-PR calculations (i.e. the absolute value of a coefficient). This also applies to GBM (which used cross-validation to tune the number of boosted trees); thus the same score was used as before: the relative influence. For BC-VIMP, negative importance values were converted to zero. For knockoffs, the knockoff test statistic was set to zero for variables screened under a target FDR value of 0.1. For VarPro, standardized importance was set to zero for values less than a cutoff value Z_0 where the latter was selected using an out-of-sample approach. In this strategy, a grid of cutoff values from $[0, 2]$ was formed. Then after ranking the features in the descending order of their VarPro importance, a random forest was fit to those features with VarPro importance meeting a given cutoff value. For each cutoff the out-of-bag error rate for the random forest fit was obtained. The optimized Z_0 value was defined as the cutoff with smallest out-of-bag error.

Each experiment was run 100 times independently. The results are given in Fig. 3 for the uncorrelated feature experiments and Fig. 4 for the correlated case. Looking at the two figures, we observe that BC-VIMP and GBM have similar behavior. Both achieve very high AUC-PR values but perform poorly in terms of gmean. This is because both methods tend to overfit by selecting too many noisy variables. They rank variables reasonably well with signal variables generally obtaining higher scores than noise variables, which yields high AUC-PR, however they are not able to threshold variables effectively and this leads to poor gmean performance, especially in the correlated variable setting. Consider Fig. 5 which displays the threshold scores for the *cobra8* correlated simulation. Scores have been standardized and monotonically transformed to a common scale to facilitate comparison where a value of zero means the method selects the variable to be noise. Observe how both BC-VIMP and GBM have trouble delineating noise compared to the other methods.

Returning to Fig. 3 and Fig. 4, we observe that while the lasso and knockoffs have a more balanced performance over AUC-PR and gmean, they struggle with certain models and do not achieve good performance across all experiments. For instance, lasso and knockoff glmnet do poorly in some of the non-linear problems. Also, performance for knockoffs generally degrades in the correlated problems. Overall, in comparison, we can see that VarPro is more consistent and generally performs well across all experiments. It

achieves good AUC-PR, and thus it is able to provide accurate ranking of variables, while at the same time it also achieves high gmean performance, and thus it is able to separate variables into signal versus noise. This property holds across both the uncorrelated and correlated feature simulations.

4.2. Classification

In classification there is an interesting difference that happens when using VarPro that is slightly more complicated than the regression case. This is related to the definition of a noisy variable due to the difference between weak conditional independence (1) used by Definition 1, which depends on the choice of g and ψ, compared with strong conditional independence (2), which is stated in terms of the response.

Consider Fig. 6 which displays data from an $L = 3$ multiclass problem with two variables. In this example there are three functions, ψ_1, ψ_2, ψ_3, where $\psi_l(\mathbf{x}) = \mathbb{P}\{Y = l | \mathbf{X} = \mathbf{x}\}$ and $\mathbf{X} = (X^{(1)}, X^{(2)})$. In this setting the target g function is $g(Y) = I\{Y = l\}$ for $l = 1, 2, 3$. Shown in Fig. 6 are the neighborhoods of 4 different rules (given by black rectangles). In (A), the rules are released along $X^{(2)}$ shown in pink which are the original neighborhoods but now unconstrained along $X^{(2)}$. The first box on the left lies within the classification boundary for class 1. Because the boundary is vertical, ψ_1 does not depend on $X^{(2)}$, and therefore $X^{(2)}$ is a noisy variable for this class. But this is not the case for the three boxes on the right where ψ_2 and ψ_3 depend on $X^{(2)}$. Therefore $X^{(2)}$ is a signal variable for class labels 2 and 3. In (B), the rules are released along $X^{(1)}$ shown by green regions. All regions vary with $X^{(1)}$ which shows that $X^{(1)}$ is a signal variable for all three classes. Thus what this illustrates, is that the definition of a noisy variable is dependent on the target function ψ. Here $X^{(2)}$ is noisy for class label 1, but not labels 2 and 3, and $X^{(1)}$ is a signal variable for all three labels.

(A) VarPro for variable $X^{(2)}$ (B) VarPro for variable $X^{(1)}$

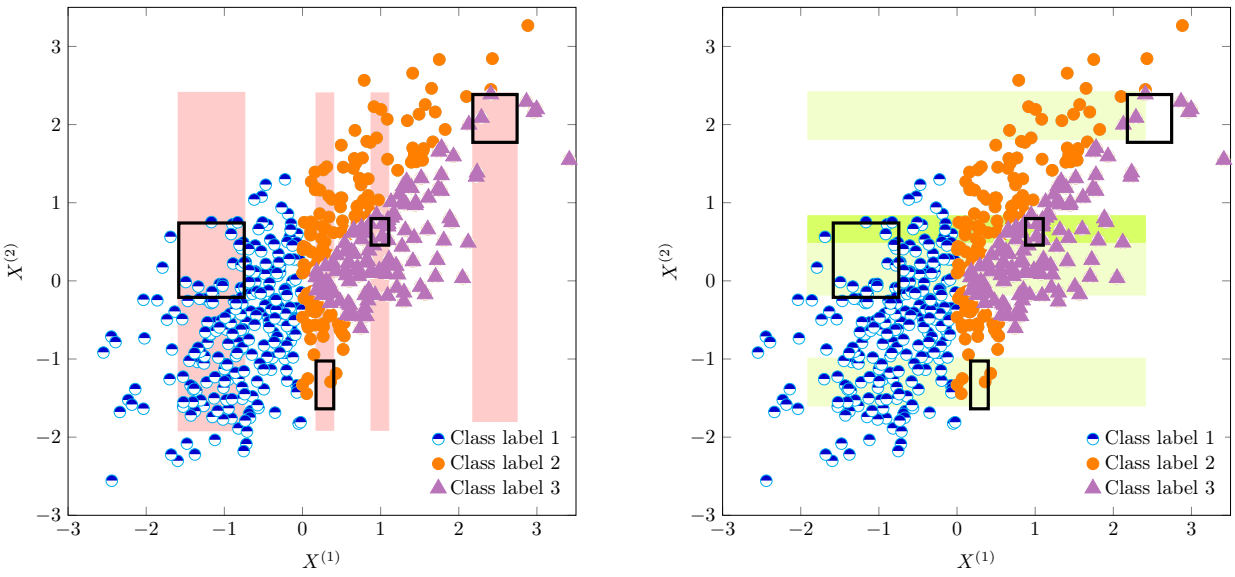

Figure 6: Illustration of VarPro in $L = 3$ classification problem (blue, orange and purple are class labels 1, 2 and 3). The classification boundary for class label 1 is vertical making $X^{(2)}$ a noisy variable for it but not for other class labels.

4.2.1. VarPro is acccurate for class specific selection and robust to correlation

To study how VarPro performs in multiclass problems, we simulated data according to the $L = 3$ multiclass model

$$\begin{aligned} y &= \operatorname*{argmax}_l \{\psi_l(\mathbf{x})\}, \quad \text{where } \psi_l(\mathbf{x}) = \frac{\phi_l(\mathbf{x})}{\sum_{l=1}^{L} \phi_l(\mathbf{x})}, \ \phi_l(\mathbf{x}) = \exp\left(\sum_{j=1}^{p} \beta_{j,l} x^{(j)}\right) \\ &\quad \beta_{j,1} = 1, \ j = 1,2,3, \text{ otherwise } \beta_{j,1} = 0 \\ &\quad \beta_{j,2} = 1, \ j = 4,5,6, \text{ otherwise } \beta_{j,2} = 0 \\ &\quad \beta_{j,3} = 1, \ j = 7,8,9, \text{ otherwise } \beta_{j,3} = 0. \end{aligned}$$

Due to the constraint $\sum_l \psi_l = 1$, all 9 features $X^{(1)}, \ldots, X^{(9)}$ are signal variables for all three classes. However, in this simulation, coordinates 1,2,3 are especially informative for class 1, coordinates 4,5,6 for class 2, and coordinates 7,8,9 for class 3. The features were drawn from a multivariate normal such that each coordinate $X^{(j)} \sim N(0,1)$ was independent of other coordinates except for pairs $(X^{(3)}, X^{(10)})$, $(X^{(6)}, X^{(15)})$, $(X^{(9)}, X^{(20)})$ which were correlated with correlation $\rho = .9$. Simulations used $N = 2000$ and $p = 20$.

VarPro estimators were calculated for each ψ_l yielding importance values $\Delta_{n,l}$ for $l = 1,2,3$. The procedure was repeated on split-samples, and the average importance was standardized using $\text{Var}_n^*(\Delta_{n,l})$ and then subtracted by a constant Z_0, yielding for each class l, a standardized importance value for each variable $X^{(s)}$, $s \in \{1, \ldots, p\}$.

The results from 250 independent runs are displayed in Fig. 7 using a cutoff value of $Z_0 = 2$. VarPro's results shown in the top panel are very good especially when contrasted with the bottom panel which displays permutation importance (BC-VIMP) obtained using random forests. The poor performance of BC-VIMP is due to the correlation between the signal variables $X^{(3)}, X^{(6)}, X^{(9)}$ and noise variables $X^{(10)}, X^{(15)}, X^{(20)}$ which causes permutation importance for signal to be degraded while artificially increasing values for noise. This differs from VarPro where signal variable importance is not degraded and noisy variables are substantially smaller and all are non-significant. Also, the group structure is clear (for example features $X^{(1)}, X^{(2)}, X^{(3)}$ are highly informative for class 1) and it is evident that each conditional probability depends on all 9 variables $X^{(1)}, \ldots, X^{(9)}$.

4.2.2. DNA methylation subtypes for adult glioma

To illustrate VarPro's ability to handle high-dimensional data, we reanalyze a subset of data used in Ceccarelli et al. (2016) for molecular profiling of adult diffuse gliomas. As part of the analysis, the authors developed a supervised analysis using DNA methylation data. Their original dataset was collected from a core set of 25,978 CpG probes which was reduced to eliminate sites that were methylated. We use their reduced set of 1206 probes for our analysis and consider $N = 880$ of the tissues from that study. One of the highlights of the study was that DNA methylation profiling was found to be effective in separating glioma subtypes. As one means to confirm the efficacy of methylation data for subtyping, and thus provide an informal validation of the studies' results, we used the supervised clusters from that analysis as outcomes in a multiclassification analysis using VarPro. In total, there were $L = 7$ labels: Classic-like, Codel, G-CIMP-high, G-CIMP-low, LGm6-GBM, Mesenchymal-like and PA-like. The 1206 CpG probes were used as variables. As an added challenge, we also included clinical data and other molecular data that were available for the samples. When added to the 1206 probes, this gave a total of $p = 1241$ variables. The data is available from the R package `TCGAbiolinksGUI.data`.

The standardized importance values from VarPro are given in Fig. 8 (an adaptive cutoff Z_0 was used applying the out-of-sample method of Section 4.1; here sequential models were fit using a random forest

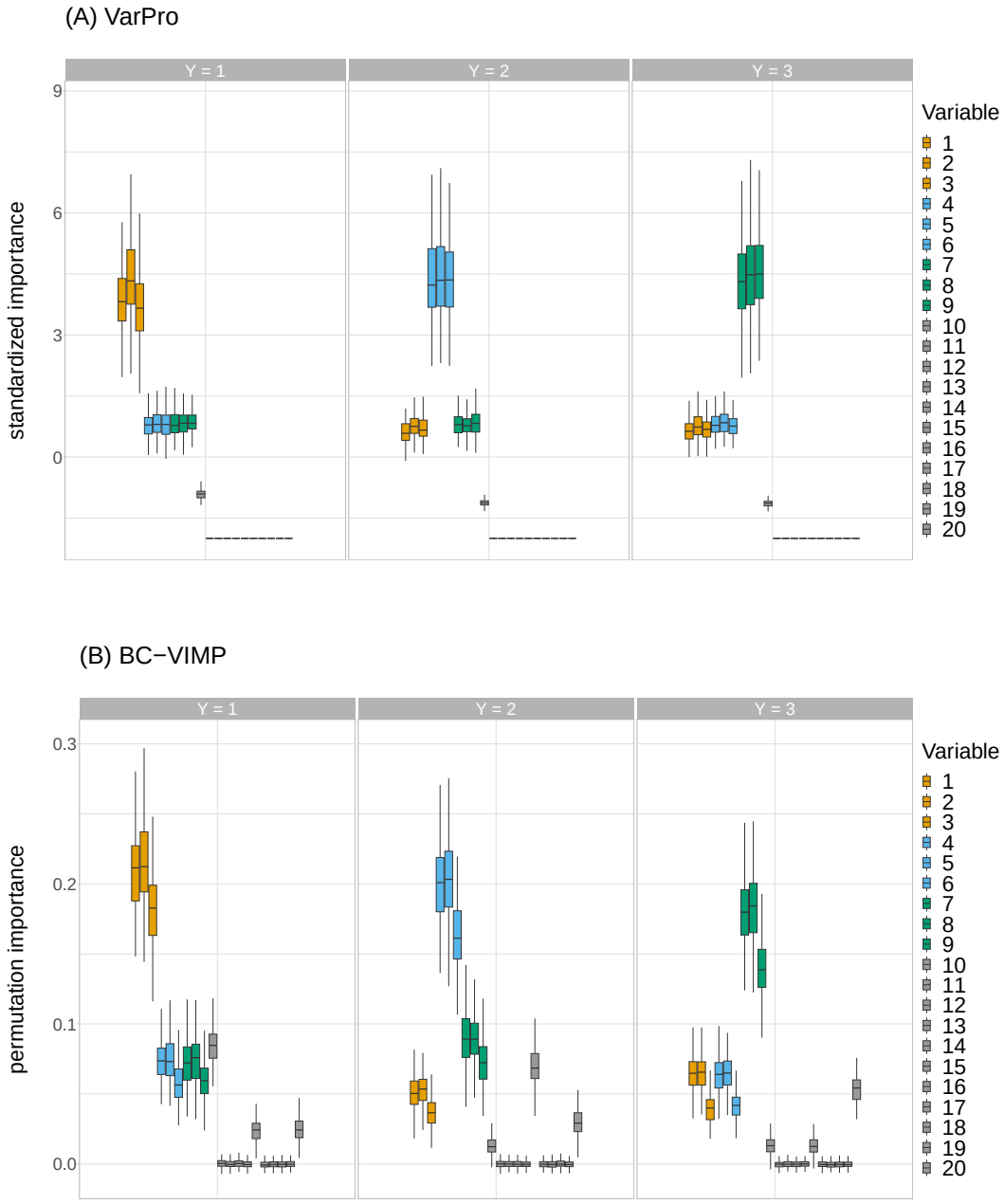

Figure 7: Results from a multiclass simulation where variables 1–3 are most informative for class 1, variables 4–6 for class 2 and variables 7–9 for class 3; variables 10–20 are noisy variables. Variables 3 and 10, 6 and 15, 9 and 20 are strongly correlated: thus there is correlation between signal and noisy features. (A) VarPro standardized importance performs very well and identifies the group structure and is not influenced by correlation. (B) BC-VIMP from random forests is influenced by correlation which degrades its performance.

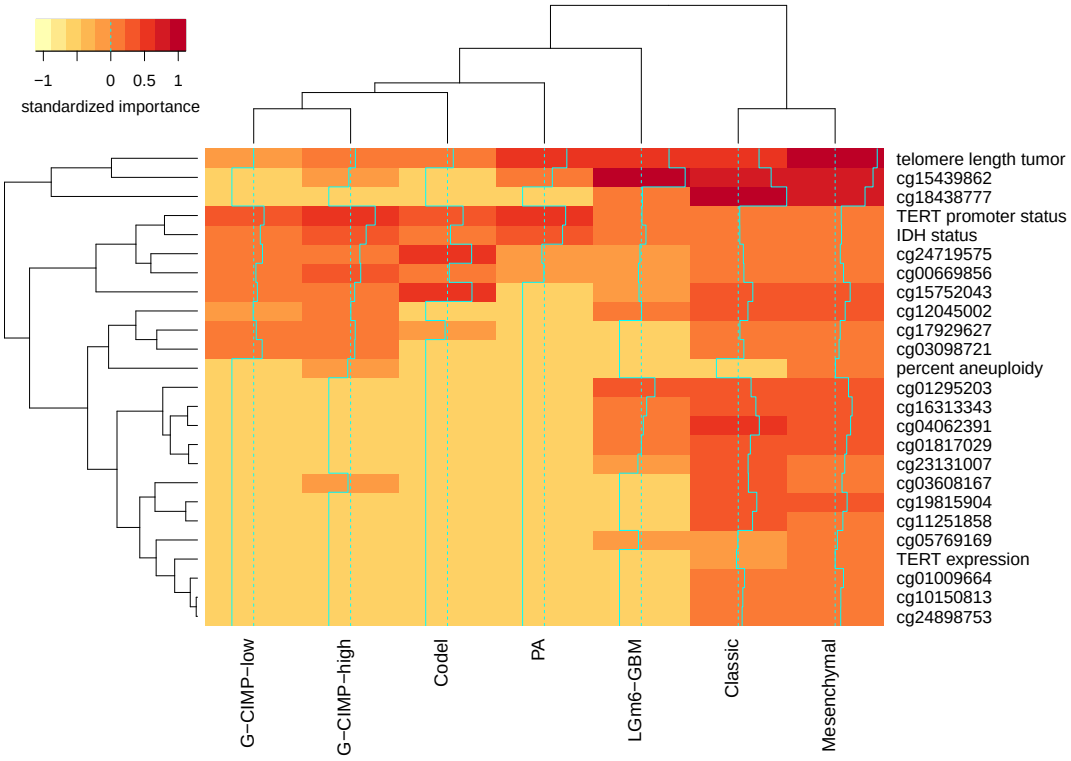

Figure 8: Heatmap of standardized importance from VarPro multiclass analysis of DNA methylation data ($N = 880, p = 1241$) for identifying subtypes for adult glioma.

classifier under misclassification error). IDH status, telomere length and TERT promotor status, which were three of the added molecular variables, are seen to be highly informative. However, of interest, there are several DNA methylation probes that are significant, and as can be seen, provide additional power to separate subtypes. For example, added benefit to separate along Codel, LGm6-GBM and Classic-like subtypes stand out in particular.

5. Time to event data using external estimators

Now we discuss the use of external estimators which is a way to extend the VarPro method. Recall that VarPro estimates the target functional $\psi(\mathbf{X}) = \mathbb{E}(g(Y)|\mathbf{X})$ by using a sample average of $g(Y)$ calculated locally within a region defined by a rule. However, depending on the setting and complexity of the problem, this may not always be suitable as it could be inefficient or difficult to estimate ψ this way.

This issue arises in time to event analysis like survival analysis. Such data presents an obstacle to VarPro, as it does not include a mechanism to account for right-censoring which can often occur in these analyses. Rather than studying a time-to-event outcome that is not always observed, we will make use of the idea of an external estimator that was touched upon in Section 2.

First, we remind readers of the survival analysis setting where right-censoring can occur. Due to censoring, the observed data is $(T, \delta, \mathbf{X})$, where $T = \min(T^o, C^o)$ is the time of event and $\delta = I\{T^o \leq C^o\}$ is the censoring indicator. Values T^o, C^o are the true survival and censoring times and the response is $Y = T^o$. Notice that Y is only observed if the observation is not censored. If censoring is not present than VarPro can be used as before, however when censoring occurs then this creates an obstacle for VarPro.

To explain what the issue is, consider variable selection for the survival function $S(t|\mathbf{X}) = \mathbb{P}\{T^0 > t|\mathbf{X}\}$ where $t > 0$ and notice this corresponds to $\psi(\mathbf{X}) = \mathbb{E}(g(T^o)|\mathbf{X})$ where $g(T^o) = I\{T^o > t\}$. Because T^o is potentially unobserved, we cannot estimate $S(t|\mathbf{X})$ locally using an estimator like

$$\hat{\theta}_n(R(\zeta)) := \frac{\sum_{i=1}^n g(Y_i) I\{\mathbf{X}_i \in R(\zeta)\}}{\sum_{i=1}^n I\{\mathbf{X}_i \in R(\zeta)\}} = \frac{\sum_{i=1}^n I\{T_i^o > t, \mathbf{X}_i \in R(\zeta)\}}{\sum_{i=1}^n I\{\mathbf{X}_i \in R(\zeta)\}}.$$

If we want to use a sample average of the observed data, as required by the VarPro methodology, then we have to account for censoring. For example, under conditional independence, assuming C^o is independent of $(T^o, \mathbf{X})$, one approach could be to use an IPCW (Inverse of Probability of Censoring Weighting) estimator such as

$$\hat{\theta}_n(\zeta) = \left(\sum_{i=1}^n I\{\mathbf{X}_i \in R(\zeta)\} \right)^{-1} \sum_{i=1}^n \frac{\delta_i}{G(T_i)} \Big[I\{T_i > t, \mathbf{X}_i \in R(\zeta)\} \Big]$$

where $G(u) = \mathbb{P}\{C^o > u\}$ is the unknown censoring distribution. However, there are known issues with IPCW estimators, such as estimation for G, and problems with inverse weights becoming large (Graf et al., 1999; Gerds and Schumacher, 2006).

As another example, consider variable selection for the restricted mean survival time (RMST) (Irwin, 1949; Andersen et al., 2004; Royston and Parmar, 2011; Kim et al., 2017). The RMST is a useful quantity summarizing lifetime behavior and is defined as the survival function integrated from $[0, \tau]$ for some fixed time point $0 < \tau < \infty$:

$$\int_0^\tau S(t|\mathbf{X})\, dt = \mathbb{E}\left[\int_0^\tau I\{T^o > t\}\, dt \Big| \mathbf{X} \right] = \mathbb{E}\left[\int_0^{T^o \wedge \tau} dt \Big| \mathbf{X} \right] = \mathbb{E}\left[T^o \wedge \tau | \mathbf{X} \right]. \tag{14}$$

Notice this can be written as $\psi(\mathbf{X}) = \mathbb{E}(g(T^o)|\mathbf{X})$ where $g(T^o) = T^o \wedge \tau$. The time horizon τ is usually selected to represent a clinically meaningful endpoint, such as 1 year or 5 year survival. However, just like the survival function, the RMST will be difficult to estimate locally in the presence of censoring.

The strategy we use to deal with this is to use an external estimator building on the idea discussed in Section 2. Thus, for each n, let $\psi_n : \mathcal{X} \to \mathbb{R}$ be an external estimator for ψ. In the examples above, ψ_n would be the estimator of the survival function at a given time point, or the RMST of an estimated survival function evaluated at a time horizon value $\tau > 0$. The estimator could be constructed from previously held out data, or it could be a deterministic value. In the former case, the theory developed below can be simply replaced with the corresponding assertion holding in probability. The modified VarPro estimator is defined as follows. For each rule $\zeta_{n,k}$, the modified procedure replaces $\hat{\theta}_n(\zeta_{n,k})$ with

$$\tilde{\theta}_n(\zeta_{n,k}) = \frac{1}{m_{n,k}} \sum_{i=1}^n \psi_n(\mathbf{X}_i) 1_{\{\mathbf{X}_i \in R_{n,k}\}}.$$

Likewise, the released rule $\hat{\theta}_n(\zeta_{n,k}^S)$, which releases the rule $\zeta_{n,k}$ along the coordinates of a set $S \subset \{1, \ldots, p\}$, is replaced with

$$\tilde{\theta}_n(\zeta_{n,k}^S) = \frac{1}{m_{n,k}^S} \sum_{i=1}^n \psi_n(\mathbf{X}_i) 1_{\{\mathbf{X}_i \in R_{n,k}^S\}}.$$

Therefore, $g(Y_i)$ used in the original formulation is replaced with $\psi_n(\mathbf{X}_i)$ which serves to estimate the conditional mean of $g(Y_i)$. Given rules $\zeta_{n,1}, \ldots, \zeta_{n,K_n}$, the modified VarPro measure of importance for S is

$$\tilde{\Delta}_n(S) = \sum_{k=1}^{K_n} W_{n,k} |\tilde{\theta}_n(\zeta_{n,k}^S) - \tilde{\theta}_n(\zeta_{n,k})|.$$

5.1. Consistency and large sample behavior of the modified procedure

In this paper we focus on applications to survival analysis but the modified VarPro procedure can be applied in general, and as shown below, is consistent under certain conditions. A sufficient condition that we adopt is the assumption that ψ_n converges uniformly to ψ. While this might appear to be a fairly strong assumption, we note that the rate of convergence is left unspecified and convergence only has to hold over a suitably defined subspace. This leads to Assumption (A5) used for the following theorem showing the modified VarPro procedure maintains the same properties as before. Assumption (A5) and conditions needed to satisfy the assumption are discussed in Appendix F.

Theorem 5. *The conclusions of Theorem 3 and Theorem 4 hold for $\tilde{\Delta}_n(S)$ under their stated conditions if additionally (A5) holds.*

5.2. High-dimensional, low sample size, variable selection for survival

For our first llustration, we consider a high-dimensional survival simulation. The survival time T^o was simulated by

$$T^o = \log\left[1 + V\exp\left(\sum_{j=1}^{p}\beta_j X^{(j)}\right)\right], \quad V \sim \frac{4}{10}\exp(.5) + \frac{1}{10}\exp(1) + \frac{2}{10}\exp(1.5) + \frac{3}{10}\exp(3)$$

where V was sampled independently of $\mathbf{X}$ and is a four-component mixture of exponential random variables with rate parameters (inverse means) $.5, 1, 1.5, 3$. The first $p_0 = 25$ features of $\mathbf{X}$ were signal variables: these were assigned coefficient values $\beta_j = \frac{1}{2}\log(1 + \sqrt{p/p_0})$. All features had marginal uniform $U(0, 1)$ distributions. The noisy features were uncorrelated; the signal features had pairwise correlation $\rho = 3/7$. Simulations were run using 50% and 75% random censoring. A sample size $N = 200$ was used while varying p.

For ψ, we use the integrated integrated cumulative hazard function (CHF) with the external ψ_n estimated using random survival forests (RSF) (Ishwaran et al., 2008). Calculations were performed using the `randomForestSRC` package (Ishwaran and Kogalur, 2023). Because the ensemble CHF is piecewise constant, no approximation was needed to integrate it.

Simulations were repeated 100 times independently. For each run we calculated the TPR, TNR and their geometric mean (gmean) and also total number of misclassified variables (miss). For comparison to VarPro we used Cox regression with lasso penalization (cox-glmnet) using the `glmnet` package (Simon et al., 2011). The cox-glmnet method was selected because it is designed for high-dimensional settings and like VarPro is computationally fast. Because of the challenging nature of the simulations, some careful tuning had to be used. For cox-glmnet, when selecting the lasso regularization parameter λ using 10-fold validation, the default one-standard error rule used by `glmnet` yielded overly sparse solutions, thus we chose instead to use the λ with the smallest out-of-sample error (the minimum rule). For VarPro, the Z_0 cutoff value was obtained using the out-of-sample methodology described in the regression simulations of Section 4.1. Sequential models used to select the optimized cutoff value were fit using RSF. Out-of-bag error was calculated using the continuous rank probability score (CRPS) (Graf et al., 1999; Gerds et al., 2008). The results are recorded in Table 1, which shows very good performance for VarPro. Compared to cox-glmnet, total number of misclassified variables is always smaller. Also, both TPR and gmean are larger in all examples.

Table 1: High-dimensional survival simulation ($N = 200$, $p = 500, 1000, 1500$ and $p_0 = 25$) where signal variables are pairwise correlated. Results are averaged from 100 runs.

		50% Censoring				75% Censoring			
	p	TPR	TNR	gmean	miss	TPR	TNR	gmean	miss
VarPro	500	0.81	1.00	0.90	5.67	0.71	0.99	0.84	12.33
	1000	0.83	1.00	0.91	6.12	0.73	0.99	0.84	16.14
	1500	0.86	1.00	0.93	6.79	0.69	0.99	0.82	21.62
cox-glmnet	500	0.71	0.97	0.83	19.89	0.52	0.97	0.71	26.77
	1000	0.73	0.98	0.85	22.70	0.55	0.98	0.73	30.55
	1500	0.75	0.99	0.86	25.88	0.57	0.98	0.74	35.13

5.3. Heart rate recovery

Exercise stress testing is commonly used to assess patients with known or suspected coronary artery disease. A useful predictor of mortality is fall in heart rate immediately after exercise stress testing, or heart rate recovery (Imai et al., 1994). Heart rate recovery is defined as the heart rate at peak exercise minus the heart rate measured 1 minute later. The hypothesis that heart rate recovery predicts mortality has been tested and validated in a number of cohorts (Cole et al., 2000). Here we study this issue by considering how predictive heart rate recovery is in the presence of other potentially useful features by making use of the modified varPro procedure.

For this analysis, we use data from the study considered in Ishwaran et al. (2004). The data involved N =23,701 patients referred for symptom-limited exercise testing. Each patient underwent an upright cool-down period for the first 2 minutes after recovery. Detailed data regarding reason for testing, symptoms, cardiac risk factors, other medical diagnoses, prior cardiac and noncardiac procedures, medications, resting electrocardiogram, resting heart rate, and blood pressure were recorded prospectively prior to testing. During each stage of exercise, and during the first 5 minutes of recovery, data were recorded regarding heart rate, blood pressure, ST-segment changes, symptoms, and arrhythmias. In total 85 variables were available for the analysis. All cause mortality was used for the survival outcome. Data was right-censored with mean follow-up among survivors of 5.7 years (range .75 to 10.1 years) during which 1,617 patients died.

For ψ, we use the RMST evaluated at $\tau = 3$ years. The RMST was calculated by (14) using the estimated survival function obtained using a RSF analysis. As before, calculations were performed using the `randomForestSRC` package (Ishwaran and Kogalur, 2023). The standardized importance values using the modified VarPro procedure are given in Fig. 9. These have been subtracted by a Z_0 cutoff value that was calculated using the same out-of-sample strategy described in the previous simulations. Fig. 9 shows heart rate recovery is significantly influential (see variable "hrrecov" on the left side of the plot). At the same time, it is also clear there are several other significant features, some with even larger standardized importance values, such as peak met, peak heart rate, copd and gender.

In order to compare these results and gain some insight into the effectiveness of VarPro, we re-analyzed the data using tree boosting with Cox regression. Computations were run using the R-package `gbm` (Greenwell et al., 2020). The top 5 variables from boosting were age, peak met, hrrecov, copd and heart rate which are in near agreement with the top 5 variables found by VarPro. However, there was some interesting differences between the the list of variables found by the two procedures. Most noticeable was how the two analyses strongly diverged for the variable gender. VarPro ranked sex as among the top 6 variables (and closely tied with number 5) whereas sex did not even make it into the top 12 variables for boosting. This is very

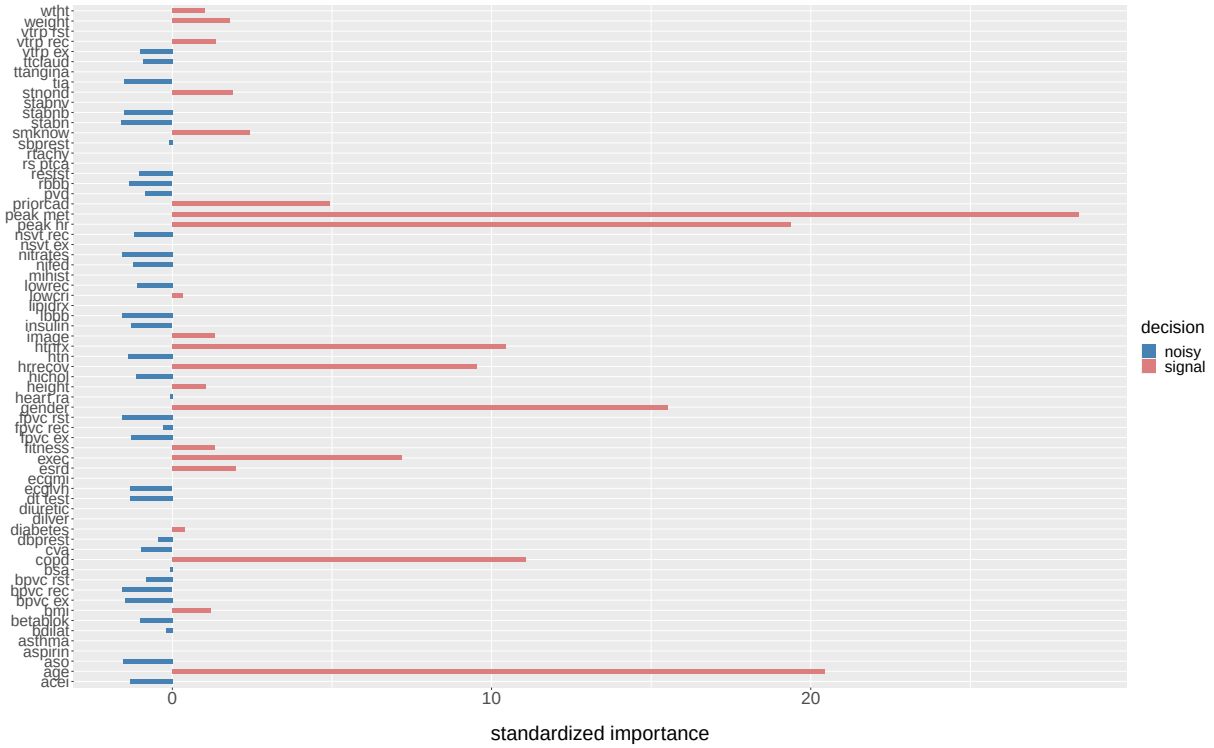

Figure 9: VarPro standardized importance for heart rate recovery survival analysis.

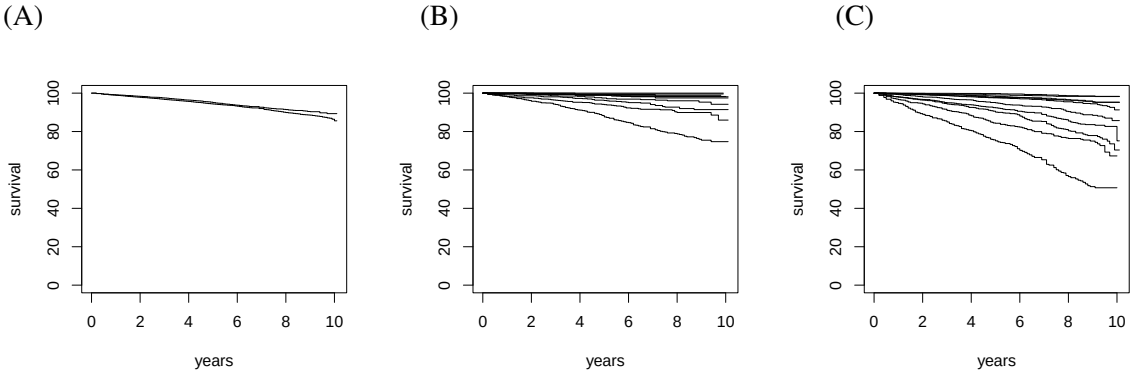

Figure 10: (A) Survival for men and women; (B) Survival for women, stratified by peak met; (C) Survival for men, stratified by peak met.

interesting as sex has been identified as a variable that is often overlooked in the heart failure literature. To study this more closely, we plot the Kaplan-Meier survival curves for men and women in Figure 10 (A). The two survival curves show hardly any difference, which is curious. However, we then looked at how gender might be interacting with other variables. We took peak met (one of the top variables) and broke this value up into 10 categories using its deciles. Then we plotted survival for each of these 10 categories, separately for women (B) and men (C). Now we observe a very large difference. For men (C) there is a noticeable decrease in survival compared with women (B) for lower values of peak met (survival curves going from top to bottom). This is an interesting finding since peak exercise capacity measured in metabolic equivalents (peak met) is often considered to be one the strongest predictors of the risk of death. Therefore this analysis provides strong evidence that men may be more susceptible to death if peak met is compromised. These

findings agree with a previous study which specifically looked at long term survival for men and women in terms of exercise/oxygen capacity (Hsich et al., 2007). In that study, women were found to be at lower risk for death than men for any given level of peak oxygen capacity.

6. Discussion

We have introduced VarPro as a new model-free variable selection method using rules and examined its empirical performance in regression, multiclassification and survival settings. Its merit is due to the fact that it does not use artificial data for inference; instead, it constructs released rules that are used with the original data, thus minimizing the risk of deviating from the feature distribution and introducing biased estimation for the target function of interest. As demonstrated, VarPro performs excellently. In a large cardiovascular survival study (Section 5), VarPro was able to identify a variable without a main effect but that had a strong interaction leading to meaningful and important differences in survival. Variables without main effects but having interactions are notoriously difficult to identify and are challenging to all but the most effective variable selection methods. This ability to identify interactions was also demonstrated in the synthetic experiments for regression (Section 4). In nearly all those models interactions terms were present. Not only did VarPro perform well in those experiments, but when features were correlated its performance was the best among all the procedures studied. This is promising because the presence of complicated interactions and correlations among features is exactly what we expect to encounter in real world data.

Another merit is that rule-based selection replaces the problem of building a high performance model (Li et al., 2005; Lee et al., 2016) into a series of lower-dimensional localized variable selection problems that can be solved computationally fast. All the examples presented in this paper can be computed efficiently and in our experience we have generally found VarPro to be fast in all kinds of data settings. Consider Fig. 11 which displays computational times for VarPro (red lines) using the Friedman 1 regression simulation of Section 4. Here we have simulated the data under different sample sizes varying from 250 to 10000 and different dimensions varying from 10 to 5000. Superimposed for comparison are computational times for BC-VIMP (see black lines) calculated using the R-package `ranger`. Observe that VarPro's compute time becomes better than BC-VIMP as N and p increase due to the heavy burden of having to calculate VIMP for so many variables and where each calculation gets more complex as sample size increases. This shows that VarPro is computationally efficient and can compete with current methods not only in usual data settings but also in big data settings.

Finally, the theory we have established lays the foundation for practical use. We provided an algorithm to calculate standardized importance, in which prior knowledge can be integrated via constructing split-weights in the rule generating phase. We put few assumptions on the rule construction and we also made external estimators possible for complex outcomes, ensuring its easy use and further extensions for machine learning, such as for survival analysis as discussed in Section 5. In this paper, only elementary tree structures were used for VarPro. It is likely that there are other ways to formulate a procedure to further tackle the problem. For example, how VarPro can be implemented in boosted trees is still pending further development and will be considered in future work.

Declarations

Consent to publish

The authors give explicit consent to submit this work and to have this work published and have obtained consent from the responsible authorities at the institute where the work has been carried out.

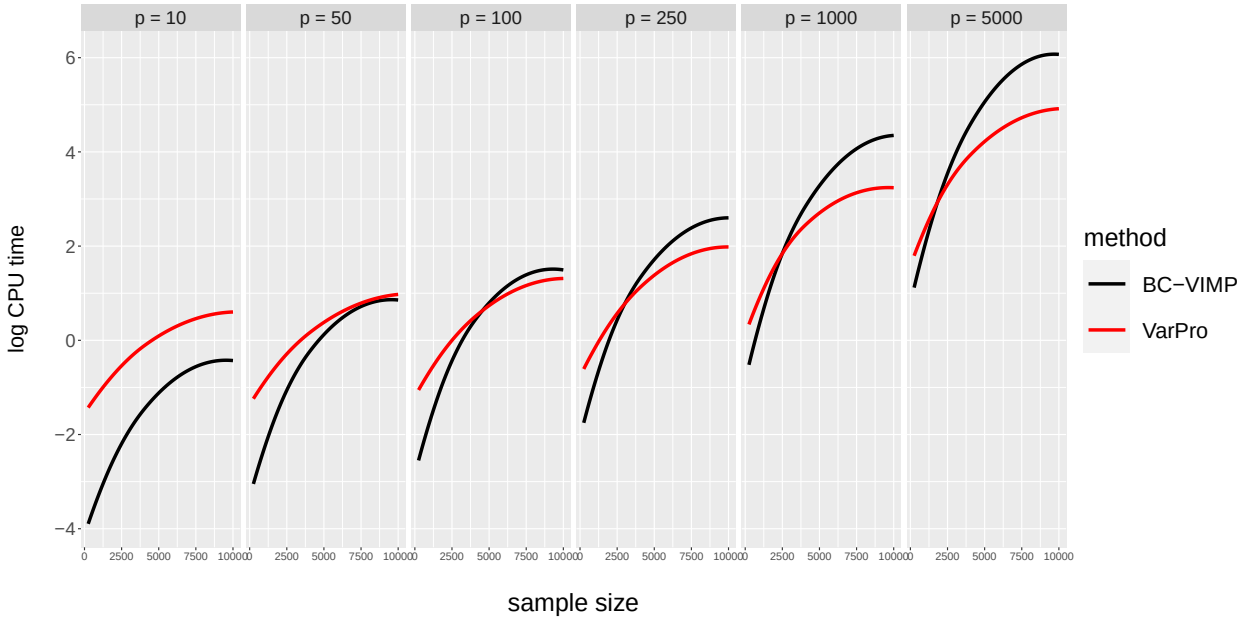

Figure 11: Log CPU times in seconds for VarPro (red lines) using Friedman 1 simulation with varying sample size N and dimension p. Also superimposed for comparison are log CPU times for BC-VIMP (black lines) calculated using the `ranger` R-package (500 trees used for each forest). Observe that VarPro's compute times becomes better than BC-VIMP as N and p increase.

Code availability

Our code is publicly available as an R package `varPro` and is available at the repository https://github.com/kogalur/varPro.

Funding

Research for the authors was supported by the National Institute Of General Medical Sciences of the National Institutes of Health, Award Number R35 GM139659.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Author contributions

All authors contributed equally to the study conception and design, material preparation, data collection and analysis and manuscript preparation. All authors read and approved the final manuscript.

Ethics approval and Consent to participate

The authors declare that this research did not require Ethics approval or Consent to participate since it does not concern human participants or human or animal datasets.

References

Andersen, P. K., Hansen, M. G. and Klein, J. P. (2004), Regression analysis of restricted mean survival time based on pseudo-observations, *Lifetime Data Analysis* 10(4), 335–350.

Biau, G., Devroye, L. and Lugosi, G. (2008), Consistency of random forests and other averaging classifiers, *Journal of Machine Learning Research* 9(Sep), 2015–2033.

Biau, G., Fischer, A., Guedj, B. and Malley, J. D. (2016), COBRA: A combined regression strategy, *Journal of Multivariate Analysis* 146, 18–28.

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

Breiman, L. (2002), Manual on setting up, using, and understanding random forests v3. 1, *Statistics Department University of California Berkeley, CA, USA* 1.

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

Candes, E., Fan, Y., Janson, L. and Lv, J. (2018), Panning for gold:model-x knockoffs for high dimensional controlled variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(3), 551–577.

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S. M. et al. (2016), Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma, *Cell* 164(3), 550–563.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010), BART: Bayesian additive regression trees, *The Annals of Applied Statistics* 4(1), 266–298.

Cole, C. R., Foody, J. M., Blackstone, E. H. and Lauer, M. S. (2000), Heart rate recovery after submaximal exercise testing as a predictor of mortality in a cardiovascularly healthy cohort, *Annals of Internal Medicine* 132(7), 552–555.

Devroye, L., Györfi, L. and Lugosi, G. (2013), *A Probabilistic Theory of Pattern Recognition*, Vol. 31, Springer.

Doksum, K., Tang, S. and Tsui, K.-W. (2008), Nonparametric variable selection: the EARTH algorithm, *Journal of the American Statistical Association* 103(484), 1609–1620.

Fisher, A., Rudin, C. and Dominici, F. (2019), All models are wrong, but many are useful: Learning a variables importance by studying an entire class of prediction models simultaneously., *Journal of Machine Learning Research* 20(177), 1–81.

Friedman, J. H. (1991), Multivariate adaptive regression splines, *The Annals of Statistics* 19(1), 1–67.

Friedman, J. H. (2001), Greedy function approximation: a gradient boosting machine, *Annals of Statistics* 29(5), 1189–1232.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* 33(1), 1–22.

Fürnkranz, J. (1997), Pruning algorithms for rule learning, *Machine Learning* 27(2), 139–172.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. (2010), Variable selection using random forests, *Pattern Recognition Letters* 31(14), 2225–2236.

Gerds, T. A., Cai, T. and Schumacher, M. (2008), The performance of risk prediction models, *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50(4), 457–479.

Gerds, T. A. and Schumacher, M. (2006), Consistent estimation of the expected brier score in general survival models with right-censored event times, *Biometrical Journal* 48(6), 1029–1040.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999), Assessment and comparison of prognostic classification schemes for survival data, *Statistics in Medicine* 18(17-18), 2529–2545.

Greenwell, B., Boehmke, B., Cunningham, J. and Developers, G. (2020), *gbm: Generalized Boosted Regression Models*. R package version 2.1.8. https://CRAN.R-project.org/package=gbm

Grömping, U. (2009), Variable importance assessment in regression: linear regression versus random forest, *The American Statistician* 63(4), 308–319.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H. et al. (2002), *A Distribution-Free Theory of Nonparametric Regression*, Vol. 1, Springer.

Hoeffding, W. (1963), Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* 58(301), 13–30.

Hsich, E., Chadalavada, S., Krishnaswamy, G., Starling, R. C., Pothier, C. E., Blackstone, E. H. and Lauer, M. S. (2007), Long-term prognostic value of peak oxygen consumption in women versus men with heart failure and severely impaired left ventricular systolic function, *The American journal of cardiology* 100(2), 291–295.

Imai, K., Sato, H., Hori, M., Kusuoka, H., Ozaki, H., Yokoyama, H., Takeda, H., Inoue, M. and Kamada, T. (1994), Vagally mediated heart rate recovery after exercise is accelerated in athletes but blunted in patients with chronic heart failure, *Journal of the American College of Cardiology* 24(6), 1529–1535.

Irwin, J. (1949), The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice, *Epidemiology & Infection* 47(2), 188–189.

Ishwaran, H. (2007), Variable importance in binary regression trees and forests, *Electronic Journal of Statistics* 1, 519–537.

Ishwaran, H., Blackstone, E. H., Pothier, C. E. and Lauer, M. S. (2004), Relative risk forests for exercise heart rate recovery as a predictor of mortality, *Journal of the American Statistical Association* 99(467), 591–600.

Ishwaran, H. and Kogalur, U. B. (2023), *Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 3.2.0. https://CRAN.R-project.org/package=randomForestSRC

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008), Random survival forests, *The Annals of Applied Statistics* 2(3), 841–860.

Kim, D. H., Uno, H. and Wei, L.-J. (2017), Restricted mean survival time as a measure to interpret clinical trial results, *JAMA Cardiology* 2(11), 1179–1180.

Kubat, M., Holte, R. and Matwin, S. (1997), Learning when negative examples abound, *in* European Conference on Machine Learning, Springer, pp. 146–153.

Lee, K.-Y., Li, B. and Zhao, H. (2016), Variable selection via additive conditional independence, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(5), 1037–1055.

Lei, J., GSell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018), Distribution-free predictive inference for regression, *Journal of the American Statistical Association* 113(523), 1094–1111.

Li, L., Dennis Cook, R. and Nachtsheim, C. J. (2005), Model-free variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), 285–299.

Liu, Y., Roková, V. and Wang, Y. (2021), Variable selection with ABC Bayesian forests, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83(3), 453–481.

Louppe, G., Wehenkel, L., Sutera, A. and Geurts, P. (2013), Understanding variable importances in forests of randomized trees, *Advances in Neural Information Processing Systems* 26, 431–439.

Nuti, G., Jiménez Rugama, L. A. and Cross, A.-I. (2021), An explainable Bayesian decision tree algorithm, *Frontiers in Applied Mathematics and Statistics* 7, 1–9.

Patterson, E. and Sesia, M. (2022), *knockoff: The Knockoff Filter for Controlled Variable Selection*. R package version 0.3.5. https://CRAN.R-project.org/package=knockoff

Royston, P. and Parmar, M. K. (2011), The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt, *Statistics in Medicine* 30(19), 2409–2421.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011), Regularization paths for Coxs proportional hazards model via coordinate descent, *Journal of Statistical Software* 39(5), 1–13.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. (2008), Conditional variable importance for random forests, *BMC Bioinformatics* 9(1), 1–11.

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L. and Geman, D. (2005), Simple decision rules for classifying human cancers from gene expression profiles, *Bioinformatics* 21(20), 3896–3904.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1), 267–288.

Van der Laan, M. J. (2006), Statistical inference for variable importance, *The International Journal of Biostatistics* 2(1).

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

Appendix A: Uniform approximation for the average size of a neighborhood

The following lemma will be used in several places and provides a uniform approximation for the average sample size of a neighborhood. The notation $b_n \gg a_n$ is used to signify $b_n/a_n \to \infty$.

Lemma 6. *Let $R_{n,k} \subseteq \mathcal{X}$ be a collection of sets such that $\mathbb{P}\{R_{n,k}\} > 0$, $k = 1, \ldots, K_n$, and define $m_{n,k} = \sum_{i=1}^n I\{\mathbf{X}_i \in R_{n,k}\}$ where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. random vectors over $\mathcal{X}$. If $K_n \leq O(\log n)$ and $m_{n,k} \geq m_n \gg \sqrt{n \log\log n}$, then the following identity holds over a set with probability tending to one uniformly over $k = 1, \ldots, K_n$:*

$$\frac{n}{m_{n,k}} = \frac{1}{\mathbb{P}(R_{n,k})}\Big(1 + \xi^*_{n,k}\Big), \quad \textit{where } |\xi^*_{n,k}| \leq \frac{\sqrt{\log\log n}}{n^{-1/2} m_n} \to 0. \tag{15}$$

Proof Observe that $m_{n,k} = \sum_{i=1}^n I\{\mathbf{X}_i \in R_{n,k}\}$ is a sum of i.i.d. Bernoulli random variables. Therefore by Hoeffding's inequality (Hoeffding, 1963),

$$\mathbb{P}\left\{\max_{1 \leq k \leq K_n} \left|n^{-1} m_{n,k} - \mathbb{P}(R_{n,k})\right| \geq \varepsilon\right\} \leq 2K_n \exp\left(-2\varepsilon^2 n\right).$$

Let $\mathcal{A}_n^c$ be the event inside the probability. If ε is allowed to depend on n so that $2\varepsilon^2 n \geq 2(\log\log n)$, then the bound on the right is order $K_n(\log n)^{-2} \to 0$. Therefore over $\mathcal{A}_n$, an event which occurs with probability tending to one,

$$|\xi_{n,k}| \leq \xi_n, \quad \text{where } \xi_{n,k} = \mathbb{P}(R_{n,k}) - n^{-1} m_{n,k},\ \xi_n = n^{-1/2}\sqrt{\log\log n}.$$

Using $m_{n,k} \geq m_n$, then with probability tending to one over $\mathcal{A}_n$:

$$\frac{n}{m_{n,k}} = \frac{1}{\mathbb{P}(R_{n,k})} + \frac{\xi^*_{n,k}}{\mathbb{P}(R_{n,k})}, \quad \text{where } \left|\xi^*_{n,k} := \frac{\xi_{n,k}}{n^{-1} m_{n,k}}\right| \leq \frac{\xi_n}{n^{-1} m_n}.$$

Upon rearrangement this gives (15). ■

Appendix B: Uniform approximation for sample averages with varying size

In order for $\Delta_n(S)$ to have good theoretical performance, $\hat{\theta}_n(\zeta_{n,1}), \ldots, \hat{\theta}_n(\zeta_{n,K_n})$ and the released values $\hat{\theta}_n(\zeta^S_{n,1}), \ldots, \hat{\theta}_n(\zeta^S_{n,K_n})$ must be controlled uniformly. This imposes a limit on the sequences $K_n, m_n(\zeta_{n,k})$ and $m_n(\zeta^S_{n,k})$. The lemma given below describes what this is.

To be able to use the lemma we will first need to center the estimators. For a given rule ζ, let $b(\zeta) = \mathbb{E}[\psi(\mathbf{X})I\{\mathbf{X} \in R(\zeta)\}]$ where recall that $\psi(\mathbf{X}) = \mathbb{E}(g(Y)|\mathbf{X})$. Notice that $\hat{\theta}_n$ can be rewritten as

$$\hat{\theta}_n(\zeta) = \frac{1}{m_n(\zeta)} \sum_{i=1}^{n} Z_i(\zeta) + \frac{nb(\zeta)}{m_n(\zeta)}, \tag{16}$$

where $Z_i(\zeta) = g(Y_i)I\{\mathbf{X}_i \in R(\zeta)\} - b(\zeta)$ are i.i.d. random variables with a mean of zero:

$$\begin{aligned} \mathbb{E}(Z_i(\zeta)) &= \mathbb{E}\Big\{ \mathbb{E}\Big[(g(Y) - \psi(\mathbf{X})I\{\mathbf{X} \in R(\zeta)\} \Big| \mathbf{X} \Big] \Big\} \\ &= \mathbb{E}\Big\{ I\{\mathbf{X} \in R(\zeta)\} \mathbb{E}\Big[(g(Y) - \psi(\mathbf{X})) \Big| \mathbf{X} \Big] \Big\} = 0. \end{aligned}$$

Recall that $m_{n,k} = m_n(\zeta_{n,k})$, $m^S_{n,k} = m_n(\zeta^S_{n,k})$ and $R_{n,k} = R(\zeta_{n,k})$, $R^S_{n,k} = R(\zeta^S_{n,k})$. Applying a similar centering as in (16) to the released rule, we have

$$\begin{aligned} \hat{\theta}_n(\zeta^S_{n,k}) - \hat{\theta}_n(\zeta_{n,k}) &= \left[\frac{1}{m^S_{n,k}} \sum_{i=1}^{n} Z_i(\zeta^S_{n,k}) - \frac{1}{m_{n,k}} \sum_{i=1}^{n} Z_i(\zeta_{n,k}) \right] \\ &\quad + \left[\frac{nb(\zeta^S_{n,k})}{m^S_{n,k}} - \frac{nb(\zeta_{n,k})}{m_{n,k}} \right], \end{aligned} \tag{17}$$

where (similar definitions apply to $\zeta^S_{n,k}$):

$$Z_i(\zeta_{n,k}) = g(Y_i)I\{\mathbf{X}_i \in R_{n,k}\} - b(\zeta_{n,k}), \quad b(\zeta_{n,k}) = \mathbb{E}\left[\psi(\mathbf{X})I\{\mathbf{X} \in R_{n,k}\}\right].$$

The sums appearing in (17) will be shown to converge to zero uniformly using the lemma given next. The quantity in the second square bracket will be dealt later: it represents a "bias" term that is asymptotically zero for noisy variables but non-zero for signal variables.

Key lemma

For each n, let $Z^{(n)}_{1,k}, \ldots, Z^{(n)}_{n,k}$ be independent random variables such that $\mathbb{E}(Z^{(n)}_{i,k}) = 0$ and $\mathbb{E}\left[(Z^{(n)}_{i,k})^2\right] \leq \sigma^2 < \infty$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K_n$. Let

$$S_{n,k} = \sum_{i=1}^{n} Z^{(n)}_{i,k}, \text{ and } T_{n,k} = \frac{1}{M_{n,k}} \sum_{i=1}^{n} Z^{(n)}_{i,k} = \frac{1}{M_{n,k}} S_{n,k}$$

where $M_{n,k}$ are random values (not necessarily independent of $Z^{(n)}_{1,k}, \ldots, Z^{(n)}_{n,k}$) satisfying $n \geq M_{n,k} \geq M_n > 0$ for $k = 1, \ldots, K_n$. We wish to identify conditions for the deterministic sequences M_n, K_n such that the asymptotic behavior of $T_{n,k}$ converges to zero uniformly over $k = 1, \ldots, K_n$ as $n \to \infty$.

Let $L_n = (nK_n)^{-1/2}$. Then by Chebyshev's inequality, for any constant $C > 0$,

$$\mathbb{P}\left\{ L_n \big| S_{n,k} \big| \geq C \right\} = \mathbb{P}\left\{ \left| \sum_{i=1}^{n} Z^{(n)}_{i,k} \right| \geq \frac{C}{L_n} \right\} \leq \frac{L_n^2}{C^2} \sum_{i=1}^{n} \mathbb{E}\left[(Z^{(n)}_{i,k})^2 \right] \leq \frac{\sigma^2}{C^2 K_n}. \tag{18}$$

Let $\mathcal{B}_{n,k} = \{\omega : |L_n S_{n,k}| \geq C\}$ and $\mathcal{B}_n = \bigcup_{k=1}^{K_n} \mathcal{B}_{n,k}$. Because $M_{n,k} \geq M_n$, setting $\delta_n = L_n^{-1} M_n^{-1} = M_n^{-1}(nK_n)^{1/2}$ we obtain

$$|T_{n,k}| = \frac{L_n^{-1}}{M_{n,k}} |L_n S_{n,k}| \, I_{\mathcal{B}_n^c} + |T_{n,k}| \, I_{\mathcal{B}_n} \leq C\delta_n + \max_{1 \leq k \leq K_n} |T_{n,k}| \, I_{\mathcal{B}_n}. \tag{19}$$

For the first term on the right of (19) to converge, M_n must converge at a rate faster than $(nK_n)^{1/2}$. For the second term, using (18), observe that

$$\begin{aligned} \mathbb{P}(\mathcal{B}_n) &= \mathbb{P}\left(\bigcup_{k=1}^{K_n} \mathcal{B}_{n,k} \right) \\ &= \mathbb{P}\left\{ \max_{1 \leq k \leq K_n} |L_n S_{n,k}| \geq C \right\} \\ &\leq \sum_{k=1}^{K_n} \mathbb{P}\left\{ |L_n S_{n,k}| \geq C \right\} \leq \frac{K_n \sigma^2}{C^2 K_n} = \frac{\sigma^2}{C^2}. \end{aligned}$$

Therefore using Markov's inequality, we have for each $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\left\{ \max_{1 \leq k \leq K_n} |T_{n,k}| \, I_{\mathcal{B}_n} \geq \varepsilon \right\} &\leq \frac{\sum_{k=1}^{K_n} \mathbb{E}\left(|T_{n,k}| I_{\mathcal{B}_n} \right)}{\varepsilon} \\ &\leq \frac{\sum_{k=1}^{K_n} \sqrt{\mathbb{E}\left(T_{n,k}^2 \right) \mathbb{P}(\mathcal{B}_n)}}{\varepsilon} \qquad \text{(Cauchy-Schwarz)} \\ &\leq \frac{\sigma}{C\varepsilon} \sum_{k=1}^{K_n} \sqrt{\mathbb{E}\left(T_{n,k}^2 \right)} \qquad \text{(bound from } \mathbb{P}(\mathcal{B}_n) \text{ above).} \end{aligned}$$

Because $\mathbb{E}(T_{n,k}^2) \leq M_n^{-2} \sum_{i=1}^n \mathbb{E}\left[(Z_{i,k}^{(n)})^2 \right] \leq M_n^{-2} n\sigma^2$,

$$\mathbb{P}\left\{ \max_{1 \leq k \leq K_n} |T_{n,k}| \, I_{\mathcal{B}_n} \geq \varepsilon \right\} \leq \frac{\sigma^2 n^{1/2} K_n}{C\varepsilon M_n}.$$

Therefore if $M_n^{-1} n^{1/2} K_n \to 0$, then by (19) we have shown

$$|T_{n,k}| \leq C\delta_n + o_p(C^{-1}\sigma^2 M_n^{-1} n^{1/2} K_n)$$

uniformly over k. But notice that $M_n^{-1} n^{1/2} K_n = K_n^{1/2} \delta_n \geq C\delta_n$ eventually, thus we have proven the following lemma.

Lemma 7. *If $M_n \uparrow \infty$ such that $M_n^{-1} n^{1/2} K_n \to 0$ then $|T_{n,k}| \leq o_p(1)$ uniformly over $k = 1, \ldots, K_n$.*

Appendix C: Consistency for noisy features (proof of Theorem 3)

The two sums in the first square bracket of (17) will be dealt with by Lemma 7 (for example, set $M_{n,k} = m_{n,k}$, $M_n = m_n$ and $Z_{i,k}^{(n)} = Z_i(\zeta_{n,k})$, then $T_{n,k}$ in Lemma 7 equals $m_{n,k}^{-1} \sum_{i=1}^n Z_i(\zeta_{n,k})$). The term inside the second square bracket of (17) is a bias term that will be shown to be asymptotically equal to

$$\beta_{n,k}(S) = \mathbb{E}(\psi(\mathbf{X}^{(S)}) | \, \mathbf{X} \in R_{n,k}^S) - \mathbb{E}(\psi(\mathbf{X}^{(S)}) | \, \mathbf{X} \in R_{n,k}). \tag{20}$$

We consider separately the case when $S \subseteq \mathcal{N}$ consists of only noisy variables and $S = \{s\}$ is a signal variable, $s \in \mathcal{S}$. Here we will consider the noisy scenario; Appendix D looks at the signal case.

The following assumptions will be used for both proofs. For Lemma 7 the following integrability condition will be needed:

(A1) $\mathbb{E}[g(Y)^2] < \infty$ and $\mathbb{E}[\psi(\mathbf{X})^2] < \infty$.

To deal with the bias term, ψ must satisfy a smoothness property. This is accommodated by the following assumption:

(A2) ψ is continuous and differentiable over the connected space $\mathcal{X} \subseteq \mathbb{R}^p$ and possesses a gradient $\mathbf{f} = \nabla\psi : \mathcal{X} \to \mathbb{R}^p$ satisfying the Lipschitz condition

$$|\mathbf{f}^{(\mathcal{S})}(\mathbf{x}_1) - \mathbf{f}^{(\mathcal{S})}(\mathbf{x}_2)| \leq C_0|\mathbf{x}_1^{(\mathcal{S})} - \mathbf{x}_2^{(\mathcal{S})}|, \qquad \text{for all } \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \tag{21}$$

for some constant $C_0 < \infty$ where $\mathbf{f}^{(S)}$ denotes the subvector of $\mathbf{f}$ with coordinates in S. Observe that the Lipschitz condition only applies to $\mathbf{f}^{(\mathcal{S})}$ as $\mathbf{f}$ is zero over the coordinates from $\mathcal{N}$.

We also require that each region shrinks to zero in a uniform sense:

(A3) For $k = 1, \ldots, K_n$, $\mathbb{P}\{\mathbf{X} \in R_{n,k}\} > 0$ and $\text{diam}_{\mathcal{S}}(R_{n,k}) \leq r_n$ for some sequence $r_n \to 0$ where $\text{diam}_{\mathcal{S}}(R_{n,k}) = \sup_{\{\mathbf{x}_1, \mathbf{x}_2 \in R_{n,k}\}} ||\mathbf{x}_1^{(\mathcal{S})} - \mathbf{x}_2^{(\mathcal{S})}||_2$.

Observe that shrinkage is only over the signal coordinates and that the shrinkage rate r_n is unspecified.

One last condition pertains to the weights. An integrability condition for ψ and $\mathbf{f}$ is required that ties it to the size of the weight:

(A4) For each n, there exists $\mathbf{x}_{n,k} \in R_{n,k}$ for $k = 1, \ldots, K_n$ such that

$$\sum_{k=1} W_{n,k}|\psi(\mathbf{x}_{n,k})| \leq C_1, \qquad \sum_{k=1} W_{n,k}||\mathbf{f}(\mathbf{x}_{n,k})||_2 \leq C_2, \tag{22}$$

for some constants $C_1, C_2 < \infty$.

Assumption (A4) stipulates a trade off between the weights and the behavior of ψ and $\mathbf{f}$ over the region of the feature space identified by a rule. Condition (22) is immediately satisfied if ψ and $\mathbf{f}$ are bounded. The condition also holds if all rules are constructed so their regions are contained within some closed bounded subspace in $\mathcal{X}$ defined by the signal coordinates. Then ψ and $\mathbf{f}$ will be bounded due to continuity. We note that assumption (A4) can also be formulated as an integrability condition, but where the integrability is taken with respect to a distribution that depends on the weights. Let $\mathbb{Q}_n$ be the distribution $\mathbb{Q}_n(A) = \sum_{k=1}^{K_n} W_{n,k} I\{\mathbf{x}_{n,k} \in A\}$. Then (22) can be written as

$$\int |\psi(\mathbf{x})| \, d\mathbb{Q}_n(\mathbf{x}) \leq C_1, \qquad \int ||\mathbf{f}(\mathbf{x})||_2 \, d\mathbb{Q}_n(\mathbf{x}) \leq C_2.$$

Also, it is also worth noting that by modifiying assumption (A4), we can accomodate a more general Lipschitz condition than (21). Assumption (A2) could be replaced with

$$|\mathbf{f}^{(\mathcal{S})}(\mathbf{x}) - \mathbf{f}^{(\mathcal{S})}(\mathbf{x}_{n,k})| \leq |\ell(\mathbf{x}_{n,k})| \cdot |\mathbf{x}^{(\mathcal{S})} - \mathbf{x}_{n,k}^{(\mathcal{S})}|, \qquad \text{for all } \mathbf{x} \in R_{n,k},$$

where $\ell : \mathcal{X} \to \mathbb{R}$ is a real-valued function satisfying $\int |\ell(\mathbf{x})| \, d\mathbb{Q}_n(\mathbf{x}) \leq C_3$ for some constant $C_3 < \infty$.

Finally, notice that we have implicitly assumed the existence of at least one signal feature, i.e. $\mathcal{S} \neq \emptyset$, in the above assumptions. We exclude $\mathcal{S} = \emptyset$ as this is a fairly trivial case. In this scenario, since ψ is constant, only condition (A1) is needed which amounts to the second moment condition $\mathbb{E}(g(Y)^2) < \infty$. Therefore the pure noisy case is easily dealt with.

We now prove Theorem 3 showing consistency of VarPro for noisy features.

Proof Apply Lemma 7 to each of the sums in the first square bracket of (17). The lemma applies since $\{Z_i(\zeta_{n,k}), Z_i(\zeta^S_{n,k})\}$ are centered i.i.d. variables with bounded second moment. The latter holds by Assumption (A1). For $\{Z_i(\zeta_{n,k})\}$, let $M_{n,k} = m_{n,k}, M_n = m_n$ where notice that $m_n^{-1}n^{1/2}K_n \to 0$ when $K_n \le O(\log n)$ and $m_n = n^{1/2}\gamma_n$ where $\gamma_n \gg \log n$; thus verifying the rate condition of the lemma. Moreover, because $m^S_{k,n} \ge m_{k,n}$ since $R_{n,k} \subseteq R^S_{n,k}$, the conditions of the lemma also hold for $\{Z_i(\zeta^S_{n,k})\}$ with $M_{n,k} = m^S_{n,k}, M_n = m_n$.

Therefore by Lemma 7, which holds uniformly,

$$\begin{aligned}\Delta_n(S) &\le \sum_{k=1}^{K_n} W_{n,k}|o_p(1) + b_{n,k}| \\ &\le o_p(1) + \sum_{k=1}^{K_n} W_{n,k}|b_{n,k}|, \quad \text{uniformly,} \end{aligned} \tag{23}$$

where (see the second term of (17) and use $\psi(\mathbf{X}) = \psi(\mathbf{X}^{(\mathcal{S})})$)

$$b_{n,k} = \frac{n}{m^S_{n,k}}\mathbb{E}[\psi(\mathbf{X}^{(\mathcal{S})})I\{\mathbf{X} \in R^S_{n,k}\}] - \frac{n}{m_{n,k}}\mathbb{E}[\psi(\mathbf{X}^{(\mathcal{S})})I\{\mathbf{X} \in R_{n,k}\}]. \tag{24}$$

Using Lemma 6 we will show (24) approximates

$$\begin{aligned}&\frac{\mathbb{E}[\psi(\mathbf{X}^{(\mathcal{S})})I\{\mathbf{X} \in R^S_{n,k}\}]}{\mathbb{P}(R^S_{n,k})} - \frac{\mathbb{E}[\psi(\mathbf{X}^{(\mathcal{S})})I\{\mathbf{X} \in R_{n,k}\}]}{\mathbb{P}(R_{n,k})} \\ &\quad = \mathbb{E}(\psi(\mathbf{X}^{(\mathcal{S})})|\,\mathbf{X} \in R^S_{n,k}) - \mathbb{E}(\psi(\mathbf{X}^{(\mathcal{S})})|\,\mathbf{X} \in R_{n,k}) \\ &\quad := \mathbb{E}^S_{n,k}(\psi) - \mathbb{E}_{n,k}(\psi), \end{aligned} \tag{25}$$

where for notational simplicity we write $\mathbb{E}^S_{n,k}$ and $\mathbb{E}_{n,k}$ for the conditional expectation of $\mathbf{X}$ given $R^S_{n,k}$ and $R_{n,k}$. Observe that (25) is the asymptotic bias $\beta_{n,k}(S)$ discussed earlier in (20).

Apply Lemma 6 noting $m_{n,k} \ge m_n = n^{1/2}\gamma_n \gg \sqrt{n \log\log n}$. By (15), there exists a set $\mathcal{A}_n$ with probability tending to one, such that

$$\frac{n}{m_{n,k}} = \frac{1}{\mathbb{P}(R_{n,k})} + \frac{\xi^*_{n,k}}{\mathbb{P}(R_{n,k})}, \quad \text{where } |\xi^*_{n,k}| \le \xi^*_n = \gamma_n^{-1}\sqrt{\log\log n} \to 0.$$

In a similar fashion, using $m^S_{n,k} \ge m_{n,k} \ge m_n = n^{1/2}\gamma_n$, there exists a set $\mathcal{A}^S_n$ with probability tending to one, such that

$$\frac{n}{m^S_{n,k}} = \frac{1}{\mathbb{P}(R^S_{n,k})} + \frac{\xi^{*S}_{n,k}}{\mathbb{P}(R^S_{n,k})}, \quad \text{where } |\xi^{*S}_{n,k}| \le \xi^*_n.$$

Thus from (24), over the set $\mathcal{A}_n \bigcap \mathcal{A}^S_n$ (an event with probability tending to 1), we have

$$|b_{n,k}| = \Big|\beta_{n,k} + \xi^{*S}_{n,k}\mathbb{E}^S_{n,k}(\psi) - \xi^*_{n,k}\mathbb{E}_{n,k}(\psi)\Big| \le |\beta_{n,k}| + \xi^*_n\Big(|\mathbb{E}^S_{n,k}(\psi)| + |\mathbb{E}_{n,k}(\psi)|\Big). \tag{26}$$

The smoothness assumption (A2) for ψ and the shrinking condition (A3) for $R_{n,k}$ are now used to expand $\mathbb{E}^S_{n,k}(\psi)$ and $\mathbb{E}_{n,k}(\psi)$ to first order which will show (25) is asymptotically zero and will enable us to further bound (26). Let $\mathbf{x}_{n,k}$ be an arbitrary point in $R_{n,k}$. By the mean-value theorem, for each $\mathbf{x} \in R_{n,k}$ there exists a $\lambda_{n,k} \in [0,1]$, such that

$$\psi(\mathbf{x}) - \psi(\mathbf{x}_{n,k}) = (\mathbf{x} - \mathbf{x}_{n,k})'\mathbf{f}(\mathbf{x}^*_{n,k})$$

where $\mathbf{x}^*_{n,k} = \mathbf{x}_{n,k} + \lambda_{n,k}(\mathbf{x} - \mathbf{x}_{n,k})$ (note that the dependence of $\lambda_{n,k}$ on $\mathbf{x}$ is suppressed). Using $\mathbf{f}(\mathbf{x}^*_{n,k}) = \mathbf{f}(\mathbf{x}_{n,k}) + [\mathbf{f}(\mathbf{x}^*_{n,k}) - \mathbf{f}(\mathbf{x}_{n,k})]$, the Lipshitz condition (21), and keeping in mind $\mathbf{f}$ is zero over the coordinates for $\mathcal{N}$,

$$|\psi(\mathbf{x}) - \psi(\mathbf{x}_{n,k})| \leq |(\mathbf{x}^{(\mathcal{S})} - \mathbf{x}^{(\mathcal{S})}_{n,k})' \mathbf{f}^{(\mathcal{S})}(\mathbf{x}_{n,k})| + C_0 |(\mathbf{x}^{(\mathcal{S})} - \mathbf{x}^{(\mathcal{S})}_{n,k})|' |(\mathbf{x}^{*(\mathcal{S})}_{n,k} - \mathbf{x}^{(\mathcal{S})}_{n,k})|.$$

Applying the Cauchy-Schwarz inequality to the first term on the right, and using assumption (A3), we have for $\mathbf{x} \in R_{n,k}$

$$\begin{aligned} |\psi(\mathbf{x}) - \psi(\mathbf{x}_{n,k})| &\leq ||(\mathbf{x}^{(\mathcal{S})} - \mathbf{x}^{(\mathcal{S})}_{n,k})||_2 ||\mathbf{f}^{(\mathcal{S})}(\mathbf{x}_{n,k})||_2 + C_0 |(\mathbf{x}^{(\mathcal{S})} - \mathbf{x}^{(\mathcal{S})}_{n,k})|' |(\mathbf{x}^{*(\mathcal{S})}_{n,k} - \mathbf{x}^{(\mathcal{S})}_{n,k})| \\ &\leq \text{diam}_{\mathcal{S}}(R_{n,k}) \Big[||\mathbf{f}^{(\mathcal{S})}(\mathbf{x}_{n,k})||_2 + C_0 \Big] \\ &\leq r_n \Big[||\mathbf{f}^{(\mathcal{S})}(\mathbf{x}_{n,k})||_2 + C_0 \Big] = r_n \Big[||\mathbf{f}(\mathbf{x}_{n,k})||_2 + C_0 \Big]. \end{aligned} \tag{27}$$

Therefore $\mathbb{E}_{n,k}(\psi) = \psi(\mathbf{x}_{n,k}) + r_{n,k}$, where

$$\Big| r_{n,k} := \mathbb{E}_{n,k}(\psi(\mathbf{X}) - \psi(\mathbf{x}_{n,k})) \Big| \leq r_n \Big[||\mathbf{f}(\mathbf{x}_{n,k})||_2 + C_0 \Big] := r^*_{n,k}.$$

By a similar argument, $\mathbb{E}^S_{n,k}(\psi) = \psi(\mathbf{x}_{n,k}) + r^S_{n,k}$ where $|r^S_{n,k}| \leq r^*_{n,k}$ satisfies the same bound as $r_{n,k}$. This is because (27) holds for $\mathbf{x} \in R^S_{n,k}$ because $R^S_{n,k}$ only differs from $R_{n,k}$ along the noisy coordinates (since $S \subseteq \mathcal{N}$).

Therefore $\mathbb{E}^S_{n,k}(\psi) = \psi(\mathbf{x}_{n,k}) + r^S_{n,k}$ and $\mathbb{E}_{n,k}(\psi) = \psi(\mathbf{x}_{n,k}) + r_{n,k}$, and hence

$$\beta_{n,k} = \big[\psi(\mathbf{x}_{n,k}) + r^S_{n,k}\big] - \big[\psi(\mathbf{x}_{n,k}) + r_{n,k}\big] = r^S_{n,k} - r_{n,k},$$

and (26) can be further bounded as follows:

$$\begin{aligned} |b_{n,k}| &\leq |r^S_{n,k}| + |r_{n,k}| + \xi^*_n \Big(|\psi(\mathbf{x}_{n,k}) + r^S_{n,k}| + |\psi(\mathbf{x}_{n,k}) + r_{n,k}| \Big) \\ &\leq 2(1 + \xi^*_n) r^*_{n,k} + 2\xi^*_n |\psi(\mathbf{x}_{n,k})|. \end{aligned}$$

Hence by (23), and assumption (A4), with probability tending to 1,

$$\begin{aligned} \Delta_n(S) &\leq o_p(1) + 2(1 + \xi^*_n) r_n \sum_{k=1}^{K_n} W_{n,k} \Big[||\mathbf{f}(\mathbf{x}_{n,k})||_2 + C_0 \Big] \\ &\qquad + 2\xi^*_n \sum_{k=1}^{K_n} W_{n,k} |\psi(\mathbf{x}_{n,k})| \\ &\leq o_p(1) + O(r_n) + O(\xi^*_n) = o_p(1), \end{aligned}$$

where the convergence is uniform. ■

Appendix D: Limiting behavior for signal features (proof of Theorem 4)

The proof for $S = \{s\}$ a signal variable is similar to the noisy variable case. The key difference is dealing with the bias term $\beta_{n,k}(s)$ (20) which is no longer asymptotically zero.

Proof Adopting the same notation as in the proof of Theorem 3, let $\mathbb{E}_{n,k}$ and $\mathbb{E}^s_{n,k}$ be the conditional expectation for $\mathbf{X}$ in $R_{n,k}$ and $R^s_{n,k}$. The same bound (27) used to derive $\mathbb{E}_{n,k}(\psi)$ applies here. Thus $\psi(\mathbf{x}) = \psi(\mathbf{x}_{n,k}) + r_{n,k}(\mathbf{x})$ where $|r_{n,k}(\mathbf{x})| \leq r^*_{n,k}$ for $\mathbf{x} \in R_{n,k}$ and $\mathbb{E}_{n,k}(\psi) = \psi(\mathbf{x}_{n,k}) + r_{n,k}$ where $r_{n,k} = \mathbb{E}_{n,k}(r_{n,k}(\mathbf{X})) \leq r^*_{n,k}$.

The previous argument used for $\mathbb{E}^S_{n,k}(\psi)$ however no longer applies because the released region now contains a signal variable. To deal with this, let $R_{n,k} = A_{n,k} \bigotimes B_{n,k}$ where $A_{n,k} = \bigotimes_{l=1}^d I_{n,k,l}$ is the subspace of $R_{n,k}$ defined by the signal features. By assumption (A3), $R_{n,k}$ is shrinking to zero in the signal features, thus $I_{n,k,l}$ are shrinking intervals for $l = 1, \ldots, d$. On the other hand, $R^s_{k,n}$ releases the coordinates in the direction of s and therefore it is shrinking in the signal coordinates in all directions except the s direction. This is the subspace $A^s_{n,k} = \bigotimes_{l \in \mathcal{S} \setminus s} I_{n,k,l}$. Notice that $A^s_{n,k}$ can be written as the union of two disjoint regions

$$A^s_{n,k} = A_{n,k} \dot{\bigcup} A^{*s}_{n,k}, \qquad \text{where } A^{*s}_{n,k} = \bigotimes_{l=1}^{s-1} I_{n,k,l} \bigotimes I^c_{n,k,s} \bigotimes_{l=s+1}^{d} I_{n,k,l}$$

and in particular this implies

$$I\{\mathbf{x}^{(\mathcal{S})} \in A^s_{n,k}\} = I\{\mathbf{x}^{(\mathcal{S})} \in A_{n,k}\} + I\{\mathbf{x}^{(\mathcal{S})} \in A^{*s}_{n,k}\}. \tag{28}$$

For $\mathbf{x} \in R^s_{n,k}$,

$$\psi(\mathbf{x}) = \big[\psi(\mathbf{x}_{n,k}) + r_{n,k}(\mathbf{x})\big] I\{\mathbf{x}^{(\mathcal{S})} \in A_{n,k}\} + \psi(\mathbf{x}) I\{\mathbf{x}^{(\mathcal{S})} \in A^{*s}_{n,k}\}.$$

Using (28), with some rearrangment, this implies for each $\mathbf{x} \in R^s_{n,k}$

$$\begin{aligned} \psi(\mathbf{x}) \;=\; & \psi(\mathbf{x}_{n,k}) I\{\mathbf{x}^{(\mathcal{S})} \in A^s_{n,k}\} \\ & + r_{n,k}(\mathbf{x}) I\{\mathbf{x}^{(\mathcal{S})} \in A_{n,k}\} \\ & - \left[\psi(\mathbf{x}) - \psi(\mathbf{x}_{n,k})\right] I\{\mathbf{x}^{(\mathcal{S})} \in A_{n,k}\} \\ & + \left[\psi(\mathbf{x}) - \psi(\mathbf{x}_{n,k})\right] I\{\mathbf{x}^{(\mathcal{S})} \in A^s_{n,k}\}. \end{aligned}$$

Therefore integrating with respect to $\mathbb{E}^s_{n,k}$,

$$\begin{aligned} \mathbb{E}^s_{n,k}(\psi(\mathbf{X})) \;=\; & \psi(\mathbf{x}_{n,k}) \\ & + r^s_{n,k} \\ & - \mathbb{E}^s_{n,k}\Big(\left[\psi(\mathbf{X}) - \psi(\mathbf{x}_{n,k})\right] I\{\mathbf{X}^{(\mathcal{S})} \in A_{n,k}\} \Big) \\ & + \mathbb{E}^s_{n,k}\Big(\psi(\mathbf{X}) - \psi(\mathbf{x}_{n,k}) \Big) \end{aligned} \tag{29}$$

where $r^s_{n,k} = \mathbb{E}^s_{n,k}(r_{n,k}(\mathbf{X}) I\{\mathbf{X}^{(\mathcal{S})} \in A_{n,k}\})$ and notice that

$$|r^s_{n,k}| \le \mathbb{E}^s_{n,k}(r^*_{n,k} I\{\mathbf{X}^{(\mathcal{S})} \in A_{n,k}\}) \le r^*_{n,k} \mathbb{E}^s_{n,k}(I\{\mathbf{X}^{(\mathcal{S})} \in A^s_{n,k}\}) = r^*_{n,k}.$$

In the proof of Theorem 3 it was shown $|\psi(\mathbf{x}) - \psi(\mathbf{x}_{n,k})| \le r^*_{n,k}$ for $\mathbf{x} \in R_{n,k}$. Therefore the third term of (29) is a remainder term of order $r^*_{n,k}$.

This leaves the fourth term in (29). To handle this, consider the local behavior of $\psi(\mathbf{x})$ for $\mathbf{x} \in A^s_{n,k}$ around the point $\tilde{\mathbf{x}}_{n,k} = (x^{(1)}_{n,k}, \ldots, x^{(s-1)}_{n,k}, x^{(s)}, x^{(s+1)}_{n,k}, \ldots, x^{(p)}_{n,k})' \in A^s_{n,k}$. By the mean-value theorem there exists a point $\tilde{\mathbf{x}}^*_{n,k} = \tilde{\mathbf{x}}_{n,k} + \lambda_{n,k}(\mathbf{x} - \tilde{\mathbf{x}}_{n,k})$ for some $0 \le \lambda_{n,k} \le 1$, such that

$$\psi(\mathbf{x}) - \psi(\tilde{\mathbf{x}}_{n,k}) = (\mathbf{x} - \tilde{\mathbf{x}}_{n,k})' \mathbf{f}(\tilde{\mathbf{x}}^*_{n,k}).$$

Because coordinate s of $\mathbf{x} - \tilde{\mathbf{x}}_{n,k}$ is zero,

$$\begin{aligned}
|\psi(\mathbf{x}) - \psi(\tilde{\mathbf{x}}_{n,k})| &= \left| \sum_{l \in \mathcal{S} \setminus s} \left((x^{(l)} - x^{(l)}_{n,k})\, \mathbf{f}^{(l)}(\tilde{\mathbf{x}}^*_{n,k}) \right) \right| \\
&\leq \sum_{l \in \mathcal{S} \setminus s} \left(|x^{(l)} - x^{(l)}_{n,k}| \cdot |\mathbf{f}^{(l)}(\tilde{\mathbf{x}}^*_{n,k})| \right) \\
&\leq \sum_{l \in \mathcal{S} \setminus s} \left(|x^{(l)} - x^{(l)}_{n,k}| \cdot |\mathbf{f}^{(l)}(\tilde{\mathbf{x}}^*_{n,k})| \right) + |\tilde{x} - x^{(s)}_{n,k}| \cdot |\mathbf{f}^{(s)}(\tilde{\mathbf{x}}^*_{n,k})| \\
&= |(\tilde{\mathbf{x}} - \mathbf{x}^{(\mathcal{S})}_{n,k})|' |\mathbf{f}^{(\mathcal{S})}(\tilde{\mathbf{x}}^*_{n,k})|
\end{aligned}$$

where $\tilde{\mathbf{x}} = (x^{(1)}, \ldots, x^{(s-1)}, \tilde{x}, x^{(s+1)}, \ldots, x^{(p)})'$ and $\tilde{x}$ can be chosen to be an arbitrary value in $I_{n,k,s}$. Notice that $\tilde{\mathbf{x}} \in A_{n,k}$. Therefore the right-hand side can be bounded using the argument of Theorem 3 from which it follows that

$$\left| \psi(\mathbf{x}) - \psi(\tilde{\mathbf{x}}_{n,k}) \right| \leq r^*_{n,k}, \qquad \text{for } \mathbf{x} \in A^s_{n,k}.$$

Recall that $\psi^s_{n,k}(z) = \psi(x^{(1)}_{n,k}, \ldots, x^{(s-1)}_{n,k}, z, x^{(s+1)}_{n,k}, \ldots, x^d_{n,k})$. Therefore, $\psi(\tilde{\mathbf{x}}_{n,k}) = \psi^s_{n,k}(x^{(s)})$ and $\psi(\mathbf{x}_{n,k}) = \psi^s_{n,k}(x^{(s)}_{n,k})$, from which it follows

$$\mathbb{E}^s_{n,k}\Big(\psi(\mathbf{X}) - \psi(\mathbf{x}_{n,k})\Big) = \mathbb{E}^s_{n,k}\left(\psi^s_{n,k}(X^{(s)}) - \psi^s_{n,k}(x^{(s)}_{n,k})\right) + O(r^*_{n,k}), \tag{30}$$

and hence using (29),

$$\begin{aligned}
\beta_{n,k} &= \mathbb{E}^S_{n,k}(\psi) - \mathbb{E}_{n,k}(\psi) \\
&= \left[\psi(\mathbf{x}_{k,n}) + \mathbb{E}^s_{n,k}\left(\psi^s_{n,k}(X^{(s)}) - \psi^s_{n,k}(x^{(s)}_{n,k})\right) + O(r^*_{n,k})\right] - \left[\psi(\mathbf{x}_{k,n}) + r_{n,k}\right] \\
&= \mathbb{E}^s_{n,k}\left(\psi^s_{n,k}(X^{(s)}) - \psi^s_{n,k}(x^{(s)}_{n,k})\right) + O(r^*_{n,k}).
\end{aligned}$$

Thus the bias does not vanish asymptotically as in the noisy variable case.

To finish the proof we follow the rest of the proof of Theorem 3. To simplify notation let $h_{n,k}(z) = \psi^s_{n,k}(z) - \psi^s_{n,k}(x^{(s)}_{n,k})$. Then

$$\begin{aligned}
\Delta_n(s) &= o_p(1) + \sum_{k=1}^{K_n} W_{n,k} \left| \beta_{n,k} + \xi^{*s}_{n,k} \mathbb{E}^s_{n,k}(\psi) - \xi^*_{n,k} \mathbb{E}_{n,k}(\psi) \right| \\
&= o_p(1) + \sum_{k=1}^{K_n} W_{n,k} \left| \beta_{n,k} + \xi^{*s}_{n,k} \mathbb{E}^s_{n,k}(\psi) \right| \\
&= o_p(1) + \sum_{k=1}^{K_n} W_{n,k} \left| (1 + \xi^{*s}_{n,k}) \mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)})) \right| \\
&= o_p(1) + (1 + o_p(1)) \sum_{k=1}^{K_n} W_{n,k} \left| \mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)})) \right|.
\end{aligned}$$

Going from line two to line three, we have used $\xi^{*s}_{n,k}\mathbb{E}^s_{n,k}(\psi) = \xi^{*s}_{n,k}\mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)})) + o_p(1)$, where the $o_p(1)$ term is uniform and is due to (29) combined with (30). Finally, the last line holds because

$$\begin{aligned}
&(1+\xi^*_n)\sum_{k=1}^{K_n} W_{n,k}\left|\mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)}))\right| \\
&\quad\geq \sum_{k=1}^{K_n} W_{n,k}\left|(1+\xi^{*s}_{n,k})\mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)}))\right| \\
&\quad\geq \sum_{k=1}^{K_n} W_{n,k}\left|\mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)}))\right| - \xi^*_n\sum_{k=1}^{K_n} W_{n,k}\left|\mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)}))\right| \\
&\quad= (1-\xi^*_n)\sum_{k=1}^{K_n} W_{n,k}\left|\mathbb{E}^s_{n,k}(h_{n,k}(X^{(s)}))\right|.
\end{aligned}$$

The right inequality is because $|a+b| \geq |a| - |b|$ for any real-valued a, b. ■

Appendix E: Regression simulation models used in empirical studies

Regression simulation models were used to test VarPro. Models used were of the form $Y|\mathbf{x} = \psi(\mathbf{x}) + \varepsilon$ or $Y|\mathbf{x} = \psi(\mathbf{x}, \varepsilon)$ and are listed below:

1. cobra2: $\psi(\mathbf{x}) = x^{(1)}x^{(2)} + (x^{(3)})^2 - x^{(4)}x^{(7)} + x^{(8)}x^{(10)} - (x^{(6)})^2$, $X^{(j)} \sim U(-1,1)$, $\varepsilon \sim N(0, 0.1^2)$.
2. cobra8: $\psi(\mathbf{x}, \varepsilon) = I\{x^{(1)} + (x^{(4)})^3 + x^{(9)} + \sin(x^{(2)}x^{(8)}) + \varepsilon > 0.38\}$, $X^{(j)} \sim U(-.25, 1)$, $\varepsilon \sim N(0, 0.1^2)$.
3. friedman1: $\psi(\mathbf{x}) = 10\sin(\pi x^{(1)}x^{(2)}) + 20(x^{(3)} - 0.5)^2 + 10x^{(4)} + 5x^{(5)}$, $X^{(j)} \sim U(0,1)$, $\varepsilon \sim N(0,1)$.
4. friedman3: $\psi(\mathbf{x}) = \arctan\left[\dfrac{x^{(2)}x^{(3)} - 1/(x^{(2)}x^{(4)})}{x^{(1)}}\right]$, $X^{(1)} \sim U(0, 100)$, $X^{(2)} \sim U(40\pi, 560\pi)$, $X^{(3)}, \ldots, X^{(p)} \sim U(0,1)$, $\varepsilon \sim N(0,1)$.
5. inx1: $\psi(\mathbf{x}) = x^{(1)}(x^{(2)})^2\sqrt{|x^{(3)}|} + \lfloor x^{(4)} - x^{(5)}x^{(6)}\rfloor$, $X^{(j)} \sim U(-1,1)$, $\varepsilon \sim N(0, 0.1^2)$.
6. inx2: $\psi(\mathbf{x}) = x^{(3)}(x^{(1)}+1)^{|x^{(2)}|} - \sqrt{\dfrac{(x^{(5)})^2}{|x^{(4)}| + |x^{(5)}| + |x^{(6)}|}}$, $X^{(j)} \sim U(-1,1)$, $\varepsilon \sim N(0, 0.1^2)$.
7. inx3: $\psi(\mathbf{x}) = \cos(x^{(1)} - x^{(2)}) + \arcsin(x^{(1)}x^{(3)}) - \arctan(x^{(2)} - (x^{(3)})^2)$, $X^{(j)} \sim U(-1,1)$, $\varepsilon \sim N(0, 0.1^2)$.
8. lm: $\psi(\mathbf{x}) = \sum_{j=1}^{15} x^{(j)}$, $X^{(j)} \sim N(0,1)$, $\varepsilon \sim N(0, 15^2)$.
9. lmi1: $\psi(\mathbf{x}) = .05f_1(\mathbf{x}) + \exp(.02f_1(\mathbf{x})f_2(\mathbf{x}))$, where $f_1(\mathbf{x}) = \sum_{j=1}^{10} x^{(j)}$, $f_2(\mathbf{x}) = \sum_{j=11}^{20} x^{(j)}$, $X^{(j)} \sim U(0,1)$, $\varepsilon \sim N(0, .1^2)$.
10. lmi2: $\psi(\mathbf{x}) = 3(\sum_{j=1}^{15} x^{(j)})^2$, $X^{(j)} \sim N(0,1)$, $\varepsilon \sim N(0, 15^2)$.
11. sup: $\psi(\mathbf{x}) = 10x^{(1)}x^{(2)} + .25\dfrac{1}{x^{(3)}x^{(4)} + 10x^{(5)}x^{(6)}}$, $X^{(j)} \sim U(0.05, 1)$, $\varepsilon \sim N(0, 0.5^2)$.
12. sup2: $\psi(\mathbf{x}) = \pi^{x^{(1)}x^{(2)}}\sqrt{2x^{(3)}} - \arcsin(x^{(4)}) + \log(x^{(3)} + x^{(5)}) - \dfrac{x^{(9)}}{x^{(10)}}\sqrt{\dfrac{x^{(7)}}{x^{(8)}}} - x^{(2)}x^{(7)}$, $X^{(j)} \sim U(0.5, 1)$, $\varepsilon \sim N(0, 0.5^2)$.

Simulations *cobra* are from Biau et al. (2016) and simulations *friedman* are from Friedman (1991). In a first set of runs, the $\mathbf{X}$ features were independently sampled as described above. In a second run, all features retained the same marginal distribution as before, but were transformed using a copula so as to make all features correlated with correlation $\rho = 0.9$. This was done for all simulations except *lm* and *lmi2* where the 15 signal features $X^{(1)}, \ldots, X^{(15)}$ were correlated within blocks of size 5 (1–5, 6–10 and 11–15).

Appendix F: Asymptotics for the modified procedure (proof of Theorem 5)

To derive the asymptotics of $\tilde{\Delta}_n(S)$ the following assumption will be used:

(A5) There exists a sequence $\tilde{r}_n \to 0$ and subspace $\mathcal{X}_n \subseteq \mathcal{X}$ containing all regions $\bigcup_{k=1}^{K_n} R_{k,n}$ and released regions $\bigcup_{k=1}^{K_n} R^S_{n,k}$ such that $|\psi_n(\mathbf{x}) - \psi(\mathbf{x})| \leq \tilde{r}_n$ for $\mathbf{x} \in \mathcal{X}_n$.

The assumption requires ψ_n to converge uniformly to ψ but some flexibility is allowed in that convergence only has to hold over a suitably defined subspace. For example, if $\psi_n(\mathbf{x}) = h(\sum_{l=1}^p \alpha_{n,l} x^{(l)})$ and $\psi(\mathbf{x}) = h(\sum_{l=1}^p \alpha_{0,l} x^{(l)})$ for h a real-valued function with derivative h', then by the mean value theorem

$$\begin{aligned} |\psi_n(\mathbf{x}) - \psi(\mathbf{x})| &\leq \left| h'\left(\sum_{l=1}^p \alpha^*_{n,l} x^{(l)}\right) \right| \sum_{l=1}^p \left| x^{(l)} \left(\alpha_{n,l} - \alpha_{0,l}\right) \right| \\ &\leq \left| h'\left(\sum_{l=1}^p \alpha^*_{n,l} x^{(l)}\right) \right| \cdot ||\mathbf{x}||_2 \cdot \sqrt{\sum_{l=1}^p |\alpha_{n,l} - \alpha_{0,l}|^2} \end{aligned}$$

where $\alpha^*_{n,l}$ is some value between $\alpha_{n,l}$ and $\alpha_{0,l}$. The simplest way to satisfy (A5) is to require boundedness where $\mathcal{X}_n \subseteq \mathcal{X}_0$ for $\mathcal{X}_0$ a closed bounded subspace of $\mathcal{X}$. Then (A5) holds under the relatively mild assumptions that h' is continuous and $\sum_{l=1}^p |\alpha_{n,l} - \alpha_{0,l}| \to 0$ where convergence can be at any rate. The boundedness condition is easily met as the size of a region is entirely controlled by the data analyst.

Proof For the proof we use a centering argument for $\tilde{\Delta}_n(S)$ similar to that used for $\Delta_n(S)$. Let $Z^\star_i(\zeta) = \psi(\mathbf{X}_i) I\{\mathbf{X}_i \in R(\zeta)\} - b(\zeta)$ where $b(\zeta) = \mathbb{E}[\psi(\mathbf{X}) I\{\mathbf{X} \in R(\zeta)\}]$. Using $\psi_n = \psi + (\psi_n - \psi)$, it follows that

$$\begin{aligned} \tilde{\theta}_n(\zeta^S_{n,k}) - \tilde{\theta}_n(\zeta_{n,k})| &= \left[\frac{1}{m^S_{n,k}} \sum_{i=1}^n Z^\star_i(\zeta^S_{n,k}) - \frac{1}{m_{n,k}} \sum_{i=1}^n Z^\star_i(\zeta_{n,k})\right] \\ &\quad + \left[\frac{n b(\zeta^S_{n,k})}{m^S_{n,k}} - \frac{n b(\zeta_{n,k})}{m_{n,k}}\right] \\ &\quad + \left[\frac{1}{m^S_{n,k}} \sum_{i=1}^n \tilde{Z}_{n,i}(\zeta^S_{n,k}) - \frac{1}{m_{n,k}} \sum_{i=1}^n \tilde{Z}_{n,i}(\zeta_{n,k})\right], \end{aligned} \tag{31}$$

where $\tilde{Z}_{n,i}(\zeta) = [\psi_n(\mathbf{X}_i) - \psi(\mathbf{X}_i)] I\{\mathbf{X}_i \in R(\zeta)\}$. Observe that the second term in (31) is the bias term asymptotically equal to $\beta_{n,k}(S)$ (20) worked out in the previous theorems. The terms in the first square bracket in (31) are sums of i.i.d. centered variables and therefore are similar to the sums in (17) and can be dealt with by Lemma 7 to show they converge to zero uniformly in probability. Therefore we only need consider the terms inside the third backet of (31).

Therefore, consider the bound

$$\sum_{k=1}^{K_n} W_{n,k} \left| \frac{1}{m_{n,k}^S} \sum_{i=1}^{n} \tilde{Z}_{n,i}(\zeta_{n,k}^S) - \frac{1}{m_{n,k}} \sum_{i=1}^{n} \tilde{Z}_{n,i}(\zeta_{n,k}) \right| \\ \leq \sum_{k=1}^{K_n} \frac{W_{n,k}}{m_{n,k}^S} \sum_{i=1}^{n} |\tilde{Z}_{n,i}(\zeta_{n,k}^S)| + \sum_{k=1}^{K_n} \frac{W_{n,k}}{m_{n,k}} \sum_{i=1}^{n} |\tilde{Z}_{n,i}(\zeta_{n,k})|. \tag{32}$$

Begin with the first sum on the right of (32). By (A5),

$$|\tilde{Z}_{n,i}(\zeta_{n,k}^S)| \leq |\psi_n(\mathbf{X}_i) - \psi(\mathbf{X}_i)| I\{\mathbf{X}_i \in R_{n,k}^S\} \leq \tilde{r}_n I\{\mathbf{X}_i \in R_{n,k}^S\}.$$

Therefore

$$\sum_{k=1}^{K_n} \frac{W_{n,k}}{m_{n,k}^S} \sum_{i=1}^{n} |\tilde{Z}_{n,i}(\zeta_{n,k}^S)| \leq r_n \sum_{k=1}^{K_n} \frac{W_{n,k}}{m_{n,k}^S} \sum_{i=1}^{n} I\{\mathbf{X}_i \in R_{n,k}^S\} = r_n \to 0.$$

The second sum on the right of (32) involving $\tilde{Z}_{n,i}(\zeta_{n,k})$ is dealt with similarly. ■