Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Nonparametric Assessment of Variable Selection and Ranking Algorithms

Zhou Tang & Ted Westling

To cite this article: Zhou Tang & Ted Westling (13 Oct 2025): Nonparametric Assessment of Variable Selection and Ranking Algorithms, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2025.2547064

To link to this article: https://doi.org/10.1080/10618600.2025.2547064

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2025, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/10618600.2025.2547064

Nonparametric Assessment of Variable Selection and Ranking Algorithms

Zhou Tang and Ted Westling

Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA

ABSTRACT

Selecting from or ranking a set of candidate variables in terms of their capacity for predicting an outcome of interest is an important task in many scientific fields. A variety of methods for variable selection and ranking have been proposed in the literature. In practice, it can be challenging to know which method is most appropriate for a given dataset. In this article, we propose methods of comparing variable selection and ranking algorithms. We first introduce measures of the quality of variable selection and ranking algorithms. We then define estimators of our proposed measures, and establish asymptotic results for our estimators. We use our results to conduct large-sample inference for our measures, and we propose a computationally efficient partial bootstrap procedure to potentially improve finite-sample inference. We assess the properties of our proposed methods using numerical studies, and we illustrate our methods with an analysis of data for predicting wine quality from its physicochemical properties. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2024
Accepted August 2025

KEYWORDS
Bootstrap; Cross-validation; Variable importance metric

1. Introduction

1.1. Background and Literature Review

In many scientific settings, researchers may wish to identify a subset of the available variables that are highly predictive of an outcome of interest or to rank the variables in terms of their predictive ability. For example, cell function is often regulated by only a small subset of genes, and identifying which genes control a specific cell function is an important research area in biology (Leclerc 2008). Similarly, the risk of many medical events only depends on a small subset of the collected covariates, and determining which covariates are risk factors is important for improving scientific understanding of an event of interest (Fan and Li 2002). Finally, in chemical and materials discovery, feature ranking techniques are an important tool for evaluating candidate materials (Janet and Kulik 2017).

Due to the importance of variable selection and ranking in an array of scientific fields, a variety of variable selection methods have been proposed. Some of the most well-known methods include best subset selection (Hocking and Leslie 1967), least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), elastic net (Zou and Hastie 2005), random forest (Breiman 2001), and Bayesian methods (Mitchell and Beauchamp 1988; George and McCulloch 1997; Park and Casella 2008). For a broad review of variable selection methods, we refer the reader to Heinze, Wallisch, and Dunkler (2018) and the references therein.

With a plethora of variable selection and ranking methods to choose from, it is not always clear which of these methods is best for a given dataset. Furthermore, applying different methods to the same data often results in different selected sets or variable ranks. It is therefore important to have a way of choosing the most appropriate selection or ranking method. Different methods have theoretical guarantees that rely on different assumptions about the true data-generating mechanism; for example, assumptions about the true regression function such as sparsity, linearity, or additivity, or assumptions about the marginal distribution of the predictors such as approximate independence or multivariate normality. If the relevant properties of the true distribution are known, it may be possible to narrow the set of viable methods to those that work well under the given conditions. However, since these properties are typically not known in practice, it is of interest to compare the performance of variable selection or ranking procedures using the data at hand. Several authors have studied the problem of comparing the performance of variable selection procedures. Heinze, Wallisch, and Dunkler (2018) suggested assessing the stability and sensitivity of candidate algorithms using quantities such as inclusion frequencies and root mean squared difference over bootstrap samples. Other authors have employed cross-validation to compare variable selection procedures (Stone 1974; Lachenbruch and Mickey 1968; Cover 1969; Refaeilzadeh, Tang, and Liu 2007; Murtaugh 2009; Sanchez-Pinto et al. 2018).

It is also important to quantify uncertainty when comparing variable selection and ranking procedures. For example, while selecting more variables typically reduces the cross-validated risk, a confidence interval may reveal that there is substantial uncertainty in the risk reduction. In this case, the method that selects a more parsimonious model may be preferred. Alternatively, if two different methods select different subsets of the

CONTACT Ted Westling *twestling@umass.edu* Department of Mathematics and Statistics, University of Massachusetts Amherst, Lederle Graduate Research Tower 1338, 710 N. Pleasant Street, Amherst, MA 01003.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

same size, testing the null hypothesis that the two sets have the same risk or obtaining a confidence interval for the difference or ratio of the risks allows the researcher to rigorously assess whether and to what extent one subset is more predictive than the other. Finally, when comparing variable ranking procedures across a range of subset sizes, uniform confidence bands for the risks or risk differences over the range allow the researcher to assess whether one method dominates another. To the best of our knowledge, no method yet exists for providing valid inference for comparisons of variable ranking procedures.

In this article, we propose a framework for empirically comparing variable selection and ranking procedures. Our framework is based on the nonparametric methods of assessing variable importance proposed in Williamson et al. (2022). However, Williamson et al. (2022) considered assessing the importance of fixed sets of variables, whereas here, the variable sets are random because the selection and ranking procedures use the data in potentially complicated ways. In Section 2, we review the framework of Williamson et al. (2022), define our proposed measures of the quality of a variable selection or ranking procedure, and discuss their interpretation. In Section 3, we propose estimators and provide conditions under which our estimators are asymptotically linear. In Section 4, we use our asymptotic results to derive large-sample confidence regions for our parameters of interest, and we also propose a computationally efficient modified bootstrap procedure to potentially improve finite-sample inference. In Section 5, we present a simulation study assessing the finite-sample properties of our methods, and in Section 6, we use the proposed methods to compare variable selection and ranking methods for predicting wine quality from its physicochemical properties. In Section 7, we provide a brief discussion. The proofs of all theorems and code implementing the methods in this article are provided in supplementary material.

2. Parameters of Interest and their Interpretation

2.1. Statistical Setting

We suppose that $X = (X_1, \ldots, X_p) \in \mathcal{X} \subseteq \mathbb{R}^p$ is a covariate vector, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is an outcome. We suppose that the observed data $\{(X_{1i}, \ldots, X_{pi}, Y_i) : i = 1, \ldots, n\}$ are drawn IID from an unknown distribution P_0. We assume that P_0 is known to lie in a model $\mathcal{M}$, which is typically a nonparametric model. With some abuse of notation, we also use P_0 as the true marginal distribution of $(X_{1i}, \ldots, X_{pi})$. The use of subscript 0 refers to evaluation under P_0; for example, we write E_0 to denote expectation under P_0. We define $\mathbb{P}_n$ as the empirical distribution of the observed data. For any measure P and P-integrable function f, we set $Pf := \int f \, \mathrm{d}P$. For any $S \subseteq \{1, \ldots, p\}$ and $X \in \mathcal{X}$, we denote X_{-S} as the elements of X whose indices *do not* fall in S, and we let $\mathcal{X}_{-S}$ be the sample space of X_{-S}.

2.2. Measures of Variable Importance

Williamson et al. (2022) proposed an approach to defining algorithm-agnostic, population-level measures of predictiveness of a subset of covariates. We will use their framework as part of our method of quantifying the quality of a variable selection or ranking algorithm, so we briefly describe the key elements of their approach. We first require a user-defined *predictiveness metric* $V : \mathcal{F} \times \mathcal{M} \to [0, 1]$, where $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to $\mathbb{R}$ endowed with a norm $\|\cdot\|_{\mathcal{F}}$. For example, $\mathcal{F}$ may consist of all $f : \mathcal{X} \to \mathbb{R}$ such that $\int f^2 \, \mathrm{d}P_0 < \infty$ and $\|f\|_{\mathcal{F}} := [\int f^2 \, \mathrm{d}P_0]^{1/2}$. Then, $V(f, P)$ is assumed to provide a measure of the predictiveness of a candidate prediction function $f \in \mathcal{F}$ when generating data from $P \in \mathcal{M}$, where higher values are assumed to correspond to better predictiveness. The *population maximizer* $f_0 \in \operatorname{argmax}_{f \in \mathcal{F}} V(f, P_0)$, is the best possible prediction function in $\mathcal{F}$ relative to V under sampling from P_0, and the *oracle predictiveness* $V(f_0, P_0)$ measures the best possible capacity of the entire covariate set X for predicting Y under sampling from P_0. Given $S \subset \{1, \ldots, p\}$, the *residual oracle predictiveness* is $V(f_{0,-S}, P_0)$, where $f_{0,-S} \in \operatorname{argmax}_{f \in \mathcal{F}_{-S}} V(f, P_0)$ and $\mathcal{F}_{-S}$ is the subset of $\mathcal{F}$ consisting of functions $f \in \mathcal{F}$ that *do not* depend on the covariates with indices in S. Thus, $V(f_{0,-S}, P_0)$ quantifies the remaining prediction potential after excluding the covariates with indices in S. The *population variable importance*, defined as $\psi_{0,S} := V(f_0, P_0) - V(f_{0,-S}, P_0)$, measures the amount of oracle predictiveness lost by excluding covariates with indices in S.

In this article, we make a simplifying assumption about the form of the predictiveness metric V. We assume there exist $\zeta : \mathcal{F} \times \mathcal{M} \to \mathbb{R}$ and $\eta : \mathcal{M} \to \mathbb{R}$ and a function $U : \operatorname{range}(\zeta) \times \operatorname{range}(\eta) \to \mathbb{R}$ such that $P \mapsto \zeta(f, P)$ is linear and $V(f, P) = U\big(\zeta(f, P), \eta(P)\big)$. Hence, we assume that V only depends on f and P together through a function that is linear in P. Such V are referred to as *standardized V-measures of degree one* in Williamson et al. (2022). This form greatly simplifies the technical conditions used for the theoretical results provided in Section 3.2, and so we sacrifice some generality for the sake of clarity and simplicity. Additional conditions on ζ, η, and U will be provided in Section 3.2. Among the four examples of predictiveness metrics V considered in Williamson et al. (2022), only the area under the ROC curve does not have this form. We now review three examples of predictiveness measures.

Example 1 (*R*-squared). Set $V(f, P) := 1 - E_P[Y - f(X)]^2 / \sigma_P^2$, where $\sigma_P^2 := \operatorname{var}_P(Y)$. This measure quantifies the proportion of variance in Y explained by $f(X)$. In this case, $\zeta(f, P) = E_P[Y - f(X)]^2$, $\eta(P) = \sigma_P^2$ and $U(v, w) = 1 - v/w$.

Example 2 (Deviance). For binary Y, let $V(f, P) := 1 - E_P[v(Y, f(X))]/v(\pi_P, \pi_P)$ for $v(u, v) := u \log v + (1-u)\log(1-v)$, and where $\pi_P := P(Y = 1)$. In this case, we have $\zeta(f, P) = E_P[v(Y, f(X))]$, $\eta(P) = v(\pi_P, \pi_P)$ and $U(v, w) = 1 - v/w$.

Example 3 (Classification accuracy). Suppose Y is binary, and define $V(f, P) := P(Y = f(X))$. This measure quantifies how often the prediction $f(X)$ coincides with Y. In this case, we have $\zeta(f, P) = P(Y = f(X))$, $\eta(P) = 1$ and $U(v, w) = v$.

In Examples 1–2, the maximizer of $f \mapsto V(f, P)$ over all f is the conditional mean function $\mu_P : x \mapsto E_P(Y \mid X = x)$, and in Example 3, it is the Bayes classifier $x \mapsto I\{\mu_P(x) > 0.5\}$. Similarly, the maximizer of $f \mapsto V(f, P)$ over $f \in \mathcal{F}_{-S}$ is $\mu_{P,-S} : x \mapsto E_P(Y \mid X_{-S} = x_{-S})$ in Examples 1–2 and $x \mapsto I\{\mu_{P,-S}(x) > 0.5\}$ in Example 3.

A central feature of this framework is that it is *algorithm-agnostic*, meaning that the population variable importance does not depend on the particular algorithm used for estimating f_0 or $f_{0,-S}$. Valid inference for the variable importance using the observed data requires estimating f_0 and $f_{0,-S}$, which does involve choosing an estimation algorithm, but the choice of an algorithm for purposes of estimation of the variable importance is separate from the definition of the parameter itself. We adopt this same approach when defining our measures of variable selection and ranking procedures.

2.3. Measures of Variable Selection Algorithms

Williamson et al. (2022) focused on assessing the importance of a fixed set of covariates S defined a priori by the researcher. Here, our focus is on assessing and comparing the performance of automatic variable selection algorithms—that is, algorithms that use the observed data to select a subset of the covariates. The variables selected by such an algorithm are a random subset of the covariates, which we denote $S_n \subseteq \{1, \ldots, p\}$. The subscript n indicates that this random subset depends on the n observed data points in some possibly complicated way. To produce an interpretable measure of the quality of a variable selection procedure, we first define a population parameter of interest for a given variable selection procedure, and we then tackle statistical inference for this population parameter. Our measures will be algorithm-agnostic in the sense that they will not be tied to the model or algorithm used by a given selection or ranking procedure.

In order to define a population parameter of interest, the simplest approach is to assume that the random subset S_n converges in probability to a fixed $S_0 \subseteq \{1, \ldots, p\}$. We will relax this condition later. We can then consider S_n as an estimator of S_0. We then define our population-level parameter of interest as the variable importance $\psi_{0,S_0} = V(f_0, P_0) - V(f_{0,-S_0}, P_0)$ of the limiting subset S_0.

We could then compare the random subsets S_n and S'_n produced by two different variable selection algorithms by comparing estimators of ψ_{0,S_0} and ψ_{0,S'_0}, where S_0 and S'_0 are the limiting subsets to which S_n and S'_n are assumed to converge. However, this natural approach suffers an important drawback. Variable importance metrics are nested: $S_0 \subset S'_0$ implies that $\psi_{0,S_0} \leq \psi_{0,S'_0}$. Hence, simply comparing ψ_{0,S_0} to ψ_{0,S'_0} is not sufficient to compare the quality of the selection algorithms (even if these population quantities were known exactly), because an algorithm that tends to select more variables will tend to produce higher variable importance. As an extreme example, the trivial algorithm that always selects all the variables will always have the maximal possible variable importance.

To resolve this drawback, we propose simply adding a second piece of information: the number of variables selected by the algorithm. Thus, our bivariate parameter of interest is $(|S_0|, \psi_{0,S_0})$. Furthermore, different algorithms can then be graphically compared by plotting the bivariate parameter in the coordinate plane $[1, p] \times [0, 1]$. This plot conveys how predictive the subset selected by each algorithm is against the number of covariates selected. If an algorithm achieves high variable importance with few selected variables, the point corresponding to the true parameter vector of the algorithm would be in the upper left region of the plot. For two limiting subsets S_0 and S'_0, if the point $(|S_0|, \psi_{0,S_0})$ is to the lower right of the point $(|S'_0|, \psi_{0,S'_0})$ on this plot, meaning that $|S_0| \geq |S'_0|$ and $\psi_{0,S_0} \leq \psi_{0,S'_0}$, and at least one of these inequalities is strict, then an algorithm with limiting subset S'_0 dominates an algorithm with limiting subset S_0 because its selected subset is smaller, yet has higher variable importance. On the other hand, if $|S_0| > |S'_0|$ and $\psi_{0,S_0} > \psi_{0,S'_0}$, then the subset S_0 is more important, but also larger, than the subset S'_0. In this case, whether the additional variables are worth the gain in importance is up to the user. One simple way to combine the two pieces of information to produce a single metric is by dividing the variable importance of the selected set by the size of the selected set; that is $\psi_{0,S_0}/|S_0|$. This measure can be interpreted as the average variable importance per selected variable, and graphically, can also be portrayed in the suggested diagram as the slope of the line connecting $(0, 0)$ and $(|S_0|, \psi_{0,S_0})$. An illustration of this diagram will be given in Section 2.5. We call this parameter the *predictiveness per selected variable* (PPSV) for short. If $\psi_{0,S_0}/|S_0| > \psi_{0,S'_0}/|S'_0|$, then an algorithm with limiting subset S_0 outperforms an algorithm with limiting subset S'_0 in terms of PPSV, meaning that the first algorithm produces higher predictiveness per selected variable than the second.

In the case where ψ_{0,S_0} increases linearly with $|S_0|$, each variable contributes equally to the total importance, resulting in a constant PPSV. This suggests that removing any predictors may lead to a significant loss in predictive capacity. Consequently, while the PPSV plot indicates the value of retaining a large set of variables, practical considerations such as interpretability or computational constraints may still necessitate variable reduction. In such cases, practitioners should use the PPSV plot in conjunction with these additional factors to guide their selection strategy.

We also note the conceptual similarity between our proposed PPSV metric and classical information criteria such as Akaike's Information Criterion (AIC) (Akaike 1998) and the Bayesian Information Criterion (BIC) (Schwarz 1978). Both AIC and BIC combine model fit and complexity into single-number summaries. Similarly, PPSV combines variable importance and the number of selected variables into one interpretable metric. An advantage of PPSV is its broader applicability, as it does not require a correctly-specified parametric or semiparametric model, making it especially useful for data-adaptive methods.

2.4. Measures of Variable Ranking Algorithms

We now extend the population parameters proposed in Section 2.3 for comparing variable selection algorithms to compare *variable ranking algorithms*; that is, algorithms that rank the p variables in terms of their potential for predicting the outcome. A variable ranking algorithm is a random ranking of the covariates. Specifically, we define a variable ranking algorithm R_n based on the n data points as a random permutation of $\{1, \ldots, p\}$. For each $j \in \{1, \ldots, p\}$, we denote $[j] := \{1, \ldots, j\}$ and $R_{n,[j]}$ as the first j elements in R_n, which are the indices of the j *most predictive* covariates according to the algorithm. For example, if $p = 3$, and $R_n = (3, 2, 1)$, then R_n ranks

X_3 as the most important for predicting Y, X_2 as the second most important, and X_1 as the least important, and $R_{n,[2]}$ would be $(3, 2)$. As with variable selection algorithms, we are most interested in situations where R_n depends on the data in a possibly complicated way. For instance, R_n may be the random variable ranking resulting from running a penalized regression algorithm on the data.

In order to define a population parameter of interest, the simplest approach is again to assume that the random ranking R_n converges in probability to a fixed rank R_0; that is, $P_0(R_n = R_0) \to 1$. This too will be relaxed later. We then consider R_n as an estimator of R_0. Intuitively, the quality of a variable ranking algorithm is higher if the variables that it tends to rank first have higher variable importance. Hence, for any variable rank R_0, we define the population *variable ranking operator characteristic* (VROC) as

$$\left(\psi_{0,R_{0,[1]}}, \ldots, \psi_{0,R_{0,[p]}}\right) = \left(V(f_0, P_0) - V(f_{0,-R_{0,[1]}}, P_0), \ldots, V(f_0, P_0) - V(f_{0,-R_{0,[p]}}, P_0)\right),$$

so that $\psi_{0,R_{0,[j]}}$ is the population variable importance of the variables indexed by $R_{0,[j]}$. We note that $\psi_{0,R_{0,[1]}} \leq \psi_{0,R_{0,[2]}} \leq \cdots \leq \psi_{0,R_{0,[p]}}$ because adding variables to the set used for prediction increases the size of the subset of $\mathcal{F}$ over which the optimization occurs.

We call the curve formed by plotting $\left(\psi_{0,R_{0,[1]}}, \ldots, \psi_{0,R_{0,[p]}}\right)$ on the vertical axis against $(1, \ldots, p)$ on the horizontal axis the *VROC curve*. Examples of population VROC curves will be provided in Section 2.5. As with the ROC curve in the context of prediction of a binary outcome (Woodward 1953), the closer the VROC curve is to the upper left corner of the rectangle $[1, p] \times [0, 1]$, the better the performance of the ranking algorithm. This is because the ideal ranking algorithm ranks the variables with the largest impact on the predictiveness metric first. Unlike the ROC curve, which necessarily ends at the point $(1, 1)$, the endpoint of the VROC curve is typically not $(p, 1)$ because the oracle predictiveness of all variables is usually not 1. However, the endpoint of the VROC curve is the same for all possible variable rankings because the endpoint represents the variable importance of all p covariates. In addition, there is not necessarily an optimal variable ranking in terms of the VROC curve in the sense that there may not exist any ranking such that the corresponding VROC curve is no smaller than that of any other ranking at all points. An exception is the case where the true conditional mean function μ is additive in the covariates and V is the R-squared predictiveness metric. In this case, the VROC curve is optimal if and only if it is concave. This will be demonstrated more in the examples below.

A one-number measure of the overall population performance of a ranking algorithm can be obtained as $\phi(\psi_{0,R_{0,[1]}}, \ldots, \psi_{0,R_{0,[p]}})$ for any order-preserving summary $\phi : \mathbb{R}^p \to \mathbb{R}$. Here, by order-preserving, we mean that $a_j \leq b_j$ for each $j \in \{1, \ldots, p\}$ implies that $\phi(a_1, \ldots, a_p) \leq \phi(b_1, \ldots, b_p)$. In particular, we will consider $\phi(\psi_{0,R_{0,[1]}}, \ldots, \psi_{0,R_{0,[p]}}) := \sum_{j=2}^{p} [\psi_{0,R_{0,[j]}} + \psi_{0,R_{0,[j-1]}}]/2$, which measures the area under the linear interpolation of the VROC curve. We call this *the area under the VROC curve* (AUVROC) for short.

2.5. Illustrative Example

Here, we illustrate the proposed measures of variable selection and ranking algorithms in a toy example. We suppose there are $p = 10$ covariates drawn from an independent uniform distributions on $[-1, 1]$. The outcome Y is generated according to $Y = 0.4X_1 + \sqrt{X_2 + 1} + 2X_3^2 + \varepsilon$, where ε follows a standard normal distribution independent of the covariates. We use the R-squared predictiveness metric V defined in Section 2.2. We consider two variable selection algorithms: the variable subset returned by marginal regression (MR), and the subset selected by multivariate adaptive regression splines (MARS) (Friedman 1991). For MR, we regress the outcome on each covariate separately using univariate linear regression and select the covariates whose resulting absolute standardized coefficients exceed a threshold. In this example, we use 0.1 and 0.2 as two thresholds for MR, and the corresponding selection algorithms are denoted MR(0.1) and MR(0.2). MARS uses regression splines to automatically model nonlinearities and interactions between variables.

By simulation, we find the limiting selected variables for MR(0.1), MR(0.2), and MARS are $S_{0,\text{MR}(0.1)} = \{1, 2\}$, $S_{0,\text{MR}(0.2)} = \{2\}$ and $S_{0,\text{MARS}} = \{1, 2, 3\}$. The corresponding population parameters are $(|S_{0,\text{MR}(0.1)}|, \psi_{0,S_{0,\text{MR}(0.1)}}) = (2, 0.11)$, $(|S_{0,\text{MR}(0.2)}|, \psi_{0,S_{0,\text{MR}(0.2)}}) = (1, 0.035)$ and $(|S_{0,\text{MARS}}|, \psi_{0,S_{0,\text{MARS}}}) = (3, 0.34)$, respectively. Figure 1 (left panel) shows the three resulting parameters in the rectangle $[1, p] \times [0, 1]$. Although the limiting subset selected by MARS has higher variable importance, MARS doesn't dominate MR(0.1) or MR(0.2) because these selected fewer variables than MARS. However, the PPSV corresponding to MARS, MR(0.1), and MR(0.2) are 0.11, 0.054, and 0.035, respectively, so the PPSV of MARS is larger than either of the MR methods. This can be determined using the left panel of Figure 1 by noting that the slope of the line corresponding to MARS is larger than that of MR(0.1) and MR(0.2).

To illustrate our proposed measures of variable ranking algorithms, we compare two ranking procedures: the rank obtained by sorting the absolute standardized coefficients from marginal regression (MR), and the rank obtained by sorting the p-values from smallest to largest from a generalized additive model (GAM) (Hastie and Tibshirani 1986) of Y on the covariates (using the default settings in the `mgcv` package in R). By simulation, we find that the limiting rankings of MR and GAM from most important to least important in predicting the outcome differ only on the first three variables. MR returns X_2, X_1, X_3 as the order of the first three variables, and GAM returns X_3, X_2, X_1 as the order of the first three important variables. Correspondingly, the population variable importances are $(0.073, 0.11, 0.34, \ldots, 0.34)$ and $(0.23, 0.31, 0.34, \ldots, 0.34)$, respectively. Figure 1 (right panel) shows the two resulting VROC curves for MR and GAM with values of AUVROCs equal to 2.7 and 3.0, respectively. In this case, the population ranking of GAM is better than that of MR because the first and first two variables selected by GAM have higher population R-squared values than the corresponding variables selected by MR. This is because the true data-generating process follows a GAM, but not a linear model, and MR fails to recognize the importance of the covariates with nonlinear effects. We also note that since the

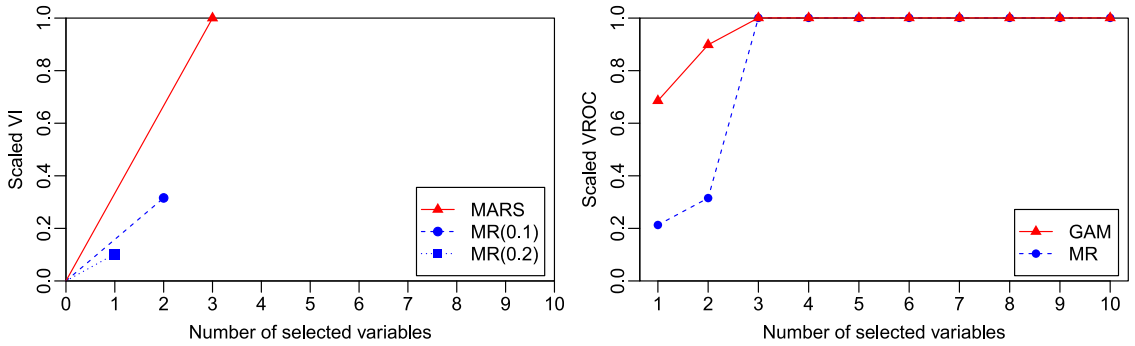

Figure 1. Illustration of the variable selection and ranking measures using the example described in Section 2.5. Left: The ratio of the variable importances of selected variables to the total variable importances for the MARS, MR(0.1), and MR(0.2) selection algorithms. Right: VROC curves scaled by the total variable importance for the MR and GAM ranking algorithms.

true model is a GAM, the optimal VROC curve is concave, as discussed above. An example where the optimal VROC curve is not concave will be provided in Section 5.

2.6. Asymptotically Stable Selection and Ranking Algorithms

When defining our measure of variable selection and ranking algorithms, we assumed that the algorithm converges in probability to a fixed limit. This is often too strong to hold in practice, so we now relax this assumption. One major reason that a ranking or selection algorithm may not converge to a fixed limit is the presence of null variables—that is, variables that do not change the value of the predictiveness function. For example, if there are $p = 3$ variables, but the outcome only depends on the first variable, then there is no true signal upon which to form a ranking of the last two variables, so these variables will be ranked on noise alone. As a result, the ranking R_n may not converge to a single fixed ranking as n increases, as small changes in the noise may change the ranking of the null variables. However, as long as the ranking algorithm asymptotically ranks the first variable first, it is reasonable to expect that R_n will asymptotically be contained in the *set* of rankings $\mathcal{R}_0$ with the first variable ranked first, that is, $\mathcal{R}_0 = \{(1,2,3),(1,3,2)\}$. Furthermore, the true oracle predictiveness function sequence $(f_{0,-R_{0,[1]}}, f_{0,-R_{0,[2]}}, f_{0,-R_{0,[3]}})$ is the same for each rank $R \in \mathcal{R}_0$ because both X_2 and X_3 are null variables. Similar situations can also happen for variable selection algorithms. For instance, if an algorithm is designed to select at least two variables, but there is only one true non-null variable, the algorithm may not converge to a fixed S_0 because it may add null variables to the selected set at random. However, the algorithm may be asymptotically contained in the set $\mathcal{S}_0$ with the first variable always selected, that is, $\mathcal{S}_0 = \{\{1,2\},\{1,3\}\}$. Furthermore, the true oracle predictiveness function is the same for each $S \in \mathcal{S}_0$. This leads us to the relaxed notion of *asymptotically stable* selection and ranking algorithms.

Definition 1. A variable selection algorithm S_n is asymptotically stable with limiting selection set $\mathcal{S}_0$ if $P_0(S_n \in \mathcal{S}_0) \to 1$ and $f_{0,-S} = f_{0,-S'}$ for all $S, S' \in \mathcal{S}_0$. A variable ranking algorithm R_n is asymptotically stable with limiting rank set $\mathcal{R}_0$ if $P_0(R_n \in \mathcal{R}_0) \to 1$ and $f_{0,-R_{[j]}} = f_{0,-R'_{[j]}}$ for all $R, R' \in \mathcal{R}_0$ and $j \in \{1,\ldots,p\}$.

A selection algorithm is asymptotically stable if it is asymptotically contained in a set of selections with common predictiveness function under P_0, and a ranking algorithm is asymptotically stable if it is asymptotically contained in a set of rankings with a common sequence of predictiveness functions under P_0. These definitions permit in particular that a selection algorithm includes random, non-converging sets of null variables in the selected set, and that a ranking algorithm ranks null variables in an arbitrary manner. Under this relaxed condition, our parameters of interest are still well-defined, and we will still be able to establish asymptotic results for our estimators.

Some selection or ranking algorithms may not be asymptotically stable. For example, suppose as above that there are $p = 3$ independent covariates, and that the true response model is $Y = X_1 + X_2^2 + \varepsilon$, for independent noise ε. If X_2 has a distribution symmetric around 0 and we rank the variables using MR defined in Section 2.5, X_2 and X_3 will both be regarded as null by the model, and hence once again the ranking may not converge to a fixed rank, but instead only be asymptotically contained in the set $\mathcal{R}_0$ defined above. However, unlike the previous example, the predictiveness function sequences of the two ranks in $\mathcal{R}_0$ are not the same because X_2 is not truly null. Situations like this may not be covered by our theoretical results.

3. Estimation and Asymptotic Results

3.1. Estimating the Proposed Measures

In this section, we introduce estimators of our parameters of interest and provide theoretical results guaranteeing convergence in distribution of our proposed estimators to mean-zero normal distributions. We use these results to construct asymptotically valid confidence sets for our parameters of interest based on our estimators.

We first propose estimators of our population parameters of interest. We recall that both of our parameters of interest involve $\psi_{0,S} := V(f_0,P_0) - V(f_{0,-S},P_0) = U(\zeta(f_0,P_0),\eta_0) - U(\zeta(f_{0,-S},P_0),\eta_0)$ for sets $S \subseteq \{1,\ldots,p\}$. To estimate $V(f_0,P_0)$ and $V(f_{0,-S},P_0)$, we will consider cross-fit plug-in estimators. We begin by obtaining the variable selection S_n and rank R_n on the whole dataset. Then, we randomly (independently of the data) partition the data into $K \geq 2$ folds with roughly equal sizes, meaning that $\lim_{n\to\infty} \max_{1\leq k\leq K} \left| \frac{n}{Kn_k} - 1 \right| = 0$, where n_k is the size of the kth fold. For simplicity, we assume that

K is fixed. For each $k \in \{1,\dots,K\}$, we construct estimators $f_{n,k}$, $f_{n,k,-S}$, $f_{n,k,-R_{[j]}}$, and $\eta_{n,k}$ of f_0, $f_{0,-S}$, $f_{0,-R_{[j]}}$, and η_0, respectively, based on the *training set* for fold k—that is, the data excluding fold k. For each k, we then define $V_{n,k} := U(\zeta(f_{n,k},\mathbb{P}_{n,k}),\eta_{n,k})$, $V_{n,k,-S_n} := U(\zeta(f_{n,k,-S_n},\mathbb{P}_{n,k}),\eta_{n,k})$, and $V_{n,k,-R_{n,[j]}} := U(\zeta(f_{n,k,-R_{n,[j]}},\mathbb{P}_{n,k}),\eta_{n,k})$, where $\mathbb{P}_{n,k}$ is the empirical distribution of the kth fold. Finally, we construct the K-fold cross-fitting estimators of $V(f_0,P_0)$, $V(f_{0,-S},P_0)$ and $V(f_{0,-R_{[j]}},P_0)$ as $\frac{1}{K}\sum_{k=1}^K V_{n,k}$, $\frac{1}{K}\sum_{k=1}^K V_{n,k,-S_n}$, and $\frac{1}{K}\sum_{k=1}^K V_{n,k,-R_{n,[j]}}$, respectively. We then define $\psi_{n,S_n} := \frac{1}{K}\sum_{k=1}^K (V_{n,k} - V_{n,k,-S_n})$.

For our proposed VROC measure $(\psi_{0,R_{0,[1]}},\dots,\psi_{0,R_{0,[p]}})$ of a variable ranking algorithm whose ranks R_n (which we recall is a random permutation of $\{1,\dots,p\}$) are converging to R_0, we propose the VROC estimator $(\psi_{n,R_{n,[1]}},\dots,\psi_{n,R_{n,[p]}})$ for $\psi_{n,R_{n,[j]}} := \frac{1}{K}\sum_{k=1}^K (V_{n,k} - V_{n,k,-R_{n,[j]}})$. For the AUVROC $\phi(\psi_{0,R_{0,[1]}},\dots,\psi_{0,R_{0,[p]}})$, we propose the analogous AUVROC estimator $\phi(\psi_{n,R_{n,[1]}},\dots,\psi_{n,R_{n,[p]}}) := \sum_{j=2}^p (\psi_{n,R_{n,[j]}} + \psi_{n,R_{n,[j-1]}})/2$.

We note that in theory, different algorithms could be used for the estimators $f_{n,k,-S}$ and $f_{n,k,-S'}$ of $f_{0,-S}$ and $f_{0,-S'}$, respectively, for two different sets of variables S and S'. Our theoretical results below would still hold as long as the two estimators both satisfied the required conditions. Nevertheless, in practice, we are not aware of a reason one would use two different algorithms, and for the sake of comparability and simplicity, we recommend using the same algorithm to construct $f_{n,k,-S}$ for each variable set S.

3.2. Asymptotic Linearity

We now provide conditions under which our estimators are asymptotically linear. Asymptotic linearity of an estimator implies $n^{-1/2}$-rate consistency and asymptotic normality, which facilitates large-sample statistical inference. Asymptotic linearity further enables joint asymptotic results, which facilitates joint large-sample inference. For a review of asymptotic linearity and semiparametric efficiency theory, we refer the reader to Le Cam, LeCam, and Yang (1990), van der Vaart (1998), and Kennedy (2016).

We recall that we assume $V(f,P) = U(\zeta(f,P),\eta(P))$. We introduce the following conditions, which are specific to a subset $S \subseteq \{1,\dots,p\}$.

(A1) There exists a function $\dot\zeta : \mathcal{F} \to L_2(P_0)$ such that $\zeta(f,P) = P[\dot\zeta(f)]$ and such that $f \mapsto \dot\zeta(f)$ is continuous at $f_{0,-S}$. Also, there exist $C,\delta \in (0,\infty)$ such that for all $f \in \mathcal{F}_{-S}$ with $\|f - f_{0,-S}\|_{\mathcal{F}} < \delta$, we have $\left|P_0\left[\dot\zeta(f) - \dot\zeta(f_{0,-S})\right]\right| \le C\|f - f_{0,-S}\|_{\mathcal{F}}^2$.

(A2) The map $(\zeta,\eta) \mapsto U(\zeta,\eta)$ is differentiable at $(\zeta(f_{0,-S},P_0),\eta_0)$ with $\dot U_\zeta(\zeta,\eta) := \frac{\partial U}{\partial\zeta}(\zeta,\eta)$ and $\dot U_\eta(\zeta,\eta) := \frac{\partial U}{\partial\eta}(\zeta,\eta)$.

(A3) It holds that $\|f_{n,k,-S} - f_{0,-S}\|_{\mathcal{F}} = o_{P_0}(n^{-1/4})$ for each $k \in \{1,\dots,K\}$.

(A4) It holds that $E_0\|\dot\zeta(f_{n,k,-S}) - \dot\zeta(f_{0,-S})\|_{L_2(P_0)} = o(1)$ for each $k \in \{1,\dots,K\}$.

(A5) It holds that $\eta_{n,k} = \eta_0 + \mathbb{P}_{n,k}\phi_0 + o_{P_0}(n_k^{-1/2})$.

Under these conditions, we have the following result.

Theorem 1. If the selection algorithm S_n is asymptotically stable with limiting selection set $\mathcal{S}_0$ and conditions (A1)–(A5) hold for $S \in \mathcal{S}_0$, then $\frac{1}{K}\sum_{k=1}^K V_{n,k,-S_n} = V(f_{0,-S},P_0) + \mathbb{P}_n\dot V_0(f_{0,-S}) + o_{P_0}(n^{-1/2})$ for

$$\dot V_0(f_{0,-S}) = \dot U_\zeta(\zeta(f_{0,-S},P_0),\eta_0)\left[\dot\zeta(f_{0,-S}) - P_0\dot\zeta(f_{0,-S})\right] + \dot U_\eta(\zeta(f_{0,-S},P_0),\eta_0)\phi_0.$$

If the ranking algorithm R_n is asymptotically stable with limiting ranking set $\mathcal{R}_0$ and conditions (A1)–(A5) hold for all $S = R_{[j]}$ where $R \in \mathcal{R}_0$ and $j \in \{1,\dots,p\}$, then $\frac{1}{K}\sum_{k=1}^K V_{n,k,-R_{n,[j]}} = V(f_{0,-R_{[j]}},P_0) + \mathbb{P}_n\dot V_0(f_{0,-R_{[j]}}) + o_{P_0}(n^{-1/2})$.

Theorem 1 provides conditions under which the cross-fitting estimators are asymptotically linear with influence functions equal to the nonparametric efficient influence functions established in Supplementary Material. Theorem 1 will be used to facilitate statistical inference for our parameters of interest in Section 4. Notably, asymptotic linearity of the cross-fitting estimators does not require any Donsker conditions. Theorem 1 differs from the results of Williamson et al. (2022) in that the set of covariates are random subsets of the p covariates. However, other than the limiting selection set $\mathcal{S}_0$ and limiting ranking set $\mathcal{R}_0$, the variable selection and ranking algorithms do not play a role in the influence functions of the estimators. This is because both the selected subset and variable ranks are discrete parameters, and the asymptotic stability assumption ensures that the true predictiveness of the estimated subsets equals a fixed limit asymptotically. However, in finite samples, the behavior of the selection or ranking algorithm can contribute to the sampling distribution of our estimators, which is not captured in the first-order asymptotic results of Theorem 1. This can result in under-coverage of confidence intervals. We propose an alternative bootstrap inference procedure in Section 4 to address this issue.

The estimators considered in Theorem 1 are based on the plug-in principle. Usually, plug-in estimators based on data-adaptive nuisance estimators inherit non-negligible asymptotic bias from the nuisance estimator, which hinders valid statistical inference for the parameter of interest. However, Theorem 1 demonstrates that plug-in estimators do not suffer from this problem in this case. As discussed in Williamson et al. (2022), this is because $f_{0,-S}$ is a maximizer of $f \mapsto V(f,P_0)$ over $\mathcal{F}_{-S}$, so that we may expect $\frac{\partial}{\partial\varepsilon}V(f_{\varepsilon,-S},P_0)\big|_{\varepsilon=0} = 0$ for a sufficiently smooth path P_ε through P_0 at $\varepsilon = 0$. As a result, we can expect that $V(f,P_0) - V(f_{0,-S},P_0)$ is bounded locally by $\|f - f_{0,-S}\|_{\mathcal{F}}^2$, as required by condition (A1). Hence, the plug-in bias $V(f_{n,k,-S},P_0) - V(f_{0,-S},P_0)$ is controlled by the behavior of $\|f_{n,k,-S} - f_{0,-S}\|_{\mathcal{F}}^2$, so that if $\|f_{n,k,-S} - f_{0,-S}\|_{\mathcal{F}} = o_{P_0}(n^{-1/4})$, as required by condition (A3), then the plug-in bias is $o_{P_0}(n^{-1/2})$.

Condition (A1) requires linearity of $P \mapsto \zeta(f,P)$. As mentioned earlier, this holds in many examples, and simplifies technical details. Condition (A1) also requires that the local behavior of $f \mapsto V(f,P_0)$ in a neighborhood of $f = f_{0,-S}$ is controlled by the quadratic norm of $f - f_{0,-S}$. This can be expected to hold because, as discussed in Williamson et al. (2022) and elsewhere, $f_{0,-S}$ is defined an optimizer of $f \mapsto V(f,P_0)$. Condition (A2) requires that the function U is differentiable so that the delta method can be used to obtain the influence

function of $V(f_{P,-S_P}, P)$. In supplementary material, we show that conditions (A1)–(A2) hold and provide $\dot{\zeta}$ and ϕ_0 for the three examples introduced in Section 2.2.

Condition (A3) requires that the maximal error of the nuisance estimators converges faster than $n^{-1/4}$. Since the rate $n^{-1/4}$ is slower than $n^{-1/2}$, condition (A3) can in principle be satisfied by data-adaptive estimators. However, the rate $n^{-1/4}$ is not achievable without some smoothness or structural assumptions about f_0, and the strength of these assumptions needs to increase with p due to the curse of dimensionality (Bühlmann and van De Geer 2011). For example, the minimax optimal rate of convergence of an estimator of f_0 in a model where f_0 is assumed to be m times differentiable is $n^{-m/(2m+p)}$ (Stone 1982). Hence, to achieve a rate faster than $n^{-1/4}$, one would need $m > p/2$. Similarly, if the covariates lie on a d-dimensional manifold in $\mathbb{R}^p$ for $d < p$, then the rate $n^{-1/4}$ can be achieved if f_0 belongs to a Sobolev class with smoothness α for $\alpha > d/2$ (Bickel and Li 2007). If f_0 is assumed to be additive and differentiable, the rate $n^{-1/4}$ can be achieved for any p (Stone 1985). As a final example, if f_0 is known to be a sparse function that depends only on $d \leq \min\{n, p\}$ variables, and f_0 belongs to a Hölder α-smooth class, then the rate $n^{-1/4}$ can be achieved if $\alpha > d/2$ and $n^{-1/2} d \log(p/d) \to 0$ (Yang and Tokdar 2015). However, whether the true function possesses these or other properties, and hence which regression estimator is best suited to the data, is typically unknown in practice. One approach to dealing with this uncertainty is to select between or combine several candidate estimators using cross-validation. For example, SuperLearner (van der Laan, Polley, and Hubbard 2007) is a generalization of the stacking algorithm that combines multiple candidate estimators, and achieves at least the best rate of convergence of the candidate estimators (van der Laan and Rose 2011).

Condition (A4) requires convergence in mean of the estimated influence function, which is used to control the empirical process term of the cross-fitting estimator. Condition (A4) follows from conditions (A1) and (A3) if $\sup_{f \in \mathcal{F}} |\dot{\zeta}(f)|$ is uniformly bounded. Condition (A5) requires that the estimator $\eta_{n,k}$ of η_0 is asymptotically linear with influence function ϕ_0 for each fold.

Finally, we establish asymptotic linearity of our estimator of the AUVROC parameter, which follows by the delta method.

Corollary 1. If the ranking algorithm R_n is asymptotically stable with limiting ranking set $\mathcal{R}_0$ and conditions (A1)–(A5) hold for all $S = R_{[j]}$ where $R \in \mathcal{R}_0$ and $j \in \{1, \ldots, p\}$, then $\phi(\psi_{n,R_{n,[1]}}, \ldots, \psi_{n,R_{n,[p]}}) = \phi(\psi_{0,R_{0,[1]}}, \ldots, \psi_{0,R_{0,[p]}}) + \mathbb{P}_n \sum_{j=2}^{p} [\dot{V}_0(f_0) - \dot{V}_0(f_{0,-R_{[j]}})/2 - \dot{V}_0(f_{0,-R_{[j-1]}})/2] + o_{P_0}(n^{-1/2})$.

4. Large-Sample Statistical Inference

Theorem 1 can be used to construct asymptotically valid confidence intervals for $\psi_{0,S}$, where $S \in \mathcal{S}_0$, as long as $\psi_{0,S} > 0$. If σ_n^2 is a consistent estimator of $\sigma_0^2 := P_0\left[\dot{V}_0(f_0) - \dot{V}_0(f_{0,-S})\right]^2$, then a Wald confidence interval for $\psi_{0,S}$ is given by $\psi_{n,S_n} \pm z_{1-\alpha/2} n^{-1/2} \sigma_n$, where z_p is the lower pth quantile of the standard normal distribution. Theorem 1 can also be used to construct asymptotically valid confidence intervals for each $\psi_{0,R_{[j]}}$, as long as $\psi_{0,R_{[j]}} > 0$, as well as uniformly valid confidence sets for $(\psi_{0,R_{[1]}}, \ldots, \psi_{0,R_{[p]}})$, where $R \in \mathcal{R}_0$, as long as $\psi_{0,R_{[1]}} > 0$. The asymptotic variance can be estimated using so-called *influence function-based* estimators. Since these methods are well-known, we omit the details here, but provide them in supplementary material. Asymptotically valid confidence intervals for the AUVROC parameter can also be constructed using an influence function-based variance estimator using Corollary 1.

The bootstrap is an alternative method of constructing confidence intervals (Efron 1982; Efron and Tibshirani 1994). In some settings, bootstrap confidence intervals have been shown to have higher-order accuracy and better finite-sample coverage than Wald intervals (Diciccio and Romano 1988; Hall 1988, 1992). In our setting, the bootstrap may be able to address at least two sources of potential finite-sample bias in the large-sample confidence intervals defined above. First, even if the selection or ranking algorithm is asymptotically stable, it may possess variability in finite samples that is not captured by the influence function-based variance estimators. Second, while the precise behavior of the prediction estimators does not play a role in the asymptotic distribution of our estimators as long as the prediction estimators satisfy the rate and complexity conditions, they may contribute to the finite-sample variability of our estimators. Accounting for these two sources of additional variability could improve the properties of our confidence intervals.

To implement a standard empirical bootstrap, we would generate n IID samples from the empirical distribution and use the bootstrap data to construct a bootstrap estimator in the exact same manner as the estimator was constructed using the original data. However, this standard approach has two shortcomings for our estimators. First, to avoid model misspecification, we advocate for using data-adaptive estimators for the prediction functions. Such estimators may be computationally intensive, and repeating this computationally intensive procedure for every bootstrap sample may be infeasible since the number of bootstrap samples B is typically in the hundreds or thousands. Second, the bootstrap can fail if the estimator is sensitive to replicated observations (Bickel, Götze, and van Zwet 1997), which may be the case for our estimators. Many data-adaptive estimators involve cross-validation as part of the procedure. When there are replicated observations in the data, the same observation can appear in the training and test sets, which breaks the independence of training and test sets and leads to overfitting (van der Laan and Rose 2018, chap. 28).

To address these two problems, we propose a modified *partial bootstrap* procedure. Specifically, when constructing our estimators using the bootstrap data, we use the prediction function estimator based on the *original data* rather than constructing new prediction function estimators using the bootstrap data. This reduces the computational burden of the bootstrap because the prediction function only needs to be estimated once per unique variable set generated over all bootstrap samples, rather than estimated for each variable set in each bootstrap sample. In addition, in practice many variable selection and ranking algorithms tend to concentrate on a small set of variables in different bootstrap samples, so the total number of variable sets for which the prediction function needs to be estimated is usually much smaller than all 2^p possible sets.

By fixing the prediction estimator, our proposed partial bootstrap does not account for variability in this estimator, which

may result in worse finite-sample performance than a procedure that does account for this variability. However, our partial bootstrap does account for variability in the selection or ranking algorithm. In addition, issues with replicated observations due to the prediction function estimator are also resolved. Since we are proposing to bootstrap the variable selection or ranking algorithm, we are implicitly assuming this algorithm is not sensitive to replicated observations.

Here are the details of our partial bootstrap procedure for the K-fold estimator based on cross-fitting. As in Section 3.2, we randomly partition the indices $\{1,\ldots,n\}$ into K folds $I_1,\ldots,I_K$ of roughly equal sizes, and for each k we construct the prediction function estimator $f_{n,k,-S}$ using the data excluding the indices in I_k. For each $b \in \{1,\ldots,B\}$, we construct the bth empirical bootstrap estimator as follows. We first draw n indices $(\alpha_{b,1},\ldots,\alpha_{b,n})$ IID from a uniform distribution on $\{1,\ldots,n\}$, and we define $\left\{\left(X_{\alpha_{b,i}}, Y_{\alpha_{b,i}}\right) : i = 1,\ldots,n\right\}$ as the bootstrap data (which typically contains replicates of the original observations). We then estimate the variable selection $S_n^{(b)}$ and variable rank $R_n^{(b)}$ using the bootstrap data. Next, for each $k \in \{1,\ldots,K\}$, we define $\mathbb{P}_{n,k}^{(b)}$ as the empirical distribution of the bootstrap data whose indices fall in I_k. We then define $\psi_{n,k,S}^{(b)} := V(f_{n,k}, \mathbb{P}_{n,k}^{(b)}) - V(f_{n,k,-S}, \mathbb{P}_{n,k}^{(b)})$ and $\psi_{n,S}^{(b)} := \sum_{k=1}^{K} w_{n,k}^{(b)} \psi_{n,k,S}^{(b)}$, where $w_{n,k}^{(b)} := \frac{1}{n}\sum_{i=1}^{n} I(\alpha_{b,i} \in I_k)$. Finally, we use the bootstrap estimates $\left(\psi_{n,S}^{(1)},\ldots,\psi_{n,S}^{(B)}\right)$ to construct bootstrap confidence intervals.

5. Numerical Studies

We conducted both a simple and a more complex simulation study to validate the large-sample results presented in Section 3 and to evaluate the finite-sample performance of the proposed methods. Due to space constraints, the simulation results and discussion for the simple case are provided in the supplementary material. For the more complex simulation setting, we considered the following data-generating process P_0. We first generated $(X_1, X_2, \ldots, X_5)$ from a multivariate normal distribution with $E_0[X_j] = 0$ and $\mathrm{Var}_0(X_j) = 1$ for all j and $\mathrm{Cor}_0(X_j, X_k) = 0.4$ for each $j \neq k$. Given $(X_1,\ldots,X_5)$, we then generated Y from a normal distribution with mean $\mu_0(x_1,\ldots,x_5) := (x_1+0.5)(x_2+1) + \sqrt{\max(x_2+5,0)} + 5\sqrt{(x_3-0.2)^2+1}$ and variance $1 + |x_4| + |x_5|$. Hence, our data-generating process involved an interaction term, nonlinear terms, null variables, correlated variables, and heteroscedasticity.

We considered three ranking algorithms. First, we considered the ranking obtained from the coefficient path from a LASSO regression (Tibshirani 1996) using the default settings in the `glmnet` package in R. Second, we considered the ranking obtained by ordering the p-values from smallest to largest from a generalized additive model (Hastie and Tibshirani 1986) using the default settings in the `mgcv` package in R. Finally, we considered the ranking obtained by ordering the estimated variable importances from smallest to largest from a multivariate adaptive regression splines regression (Friedman 1991) using the default settings in the `earth` package in R. Throughout this section, we abbreviate "generalized additive model" as "GAM" and "multivariate adaptive regression splines" as "MARS". By simulating $n = 10^5$ samples multiple times and inspecting the rankings for each procedure, we determined that MARS and GAM were converging to the limiting rank set $\mathcal{R}_0 = \{(3,1,2,4,5),(3,1,2,5,4)\}$ with true R-squared predictiveness sequence $(0.39, 0.55, 0.66, 0.70, 0.75)$ and AUVROC equal to 2.48, while LASSO was converging to the limiting rank set $\mathcal{R}_0 = \{(1,2,3,4,5),(1,2,3,5,4)\}$ with true R-squared predictiveness sequence $(0.14, 0.23, 0.66, 0.70, 0.75)$ and AUVROC 2.03.

We considered three selection algorithms. First, we considered the subset obtained from a LASSO regression using penalty parameter selected by ten-fold cross-validation and default settings in the `glmnet` package in R. Second, we considered the selection obtained by including the variables that have p-value smaller than 0.05 from a GAM using the default settings in the `mgcv` package in R. Finally, we considered the selection obtained from a MARS regression using the default settings in the `earth` package in R. By simulating a large number of samples, we determined that LASSO and MARS were converging to the limiting selection set $\mathcal{S}_0 = \{\{1,2,3\}\}$ with true R-squared predictiveness measure 0.66 and PPSV 0.22, and GAM was converging to the limiting selection set $\mathcal{S}_0 = \{\{1,2,3,4,5\}\}$ with true R-squared predictiveness measure 0.75 and PPSV 0.15.

For each sample size n equal to 500, 1K, 2K, 3K, 4K, 5K, 10K, and 20K, we simulated 500 samples of n IID observations from the above data-generating mechanism. We considered V equal to the R-squared predictiveness metric defined in Section 2.2. We estimated our proposed measures $\psi_{0,S_0}/|S_0|$ and AUVROC using both the cross-fitting estimator with $K = 5$ folds. To estimate the regression functions, we used SuperLearner (van der Laan, Polley, and Hubbard 2007) with 5-fold cross-validation and a library consisting of `xgboost` (Chen and Guestrin 2016; Chen et al. 2021), `gam` (Hastie and Tibshirani 1986; Hastie 2020), and `earth` (Friedman 1991; Milborrow 2021). We constructed 95% Wald intervals using the cross-fitting influence function-based variance estimator and the partial bootstrap procedure. We considered three types of bootstrap confidence intervals: the percentile method, percentile t-method, and Efron's percentile method (in the terminology of van der Vaart 1998). The true VROC curves for the three ranking algorithms along with cross-fitting estimators and pointwise Wald and uniform 95% confidence sets for a single simulation with $n = 1000$ are provided in supplementary material.

We now turn to the results of the simulation study. Figure 2 displays the properties of the AUVROC (top row) and PPSV (bottom row) estimators and corresponding Wald confidence intervals. The left column displays $n^{1/2}$ times the bias of the estimators. In all cases, the bias appears to tend to zero faster than $n^{-1/2}$ for large enough sample sizes, but there is considerable heterogeneity in the finite-sample bias of the estimators. For the PPSV, the absolute bias appears to decrease slower than $n^{-1/2}$ for n less than roughly 5000. As we discuss more below, this is because the sampling distribution of the PPSV estimators is multi-modal at these sample sizes. The middle column of Figure 2 displays $n^{1/2}$ times the standard deviation of the estimators. The standard deviations appear to stabilize at the $n^{-1/2}$ rate for all estimators except that of the PPSV with the GAM algorithm. This is because the subset selected by GAM is still not stable at sample size $20K$, and the standard deviation decreases as the selection algorithm stabilizes. The right column of Figure 2 displays the empirical coverage rate of 95%

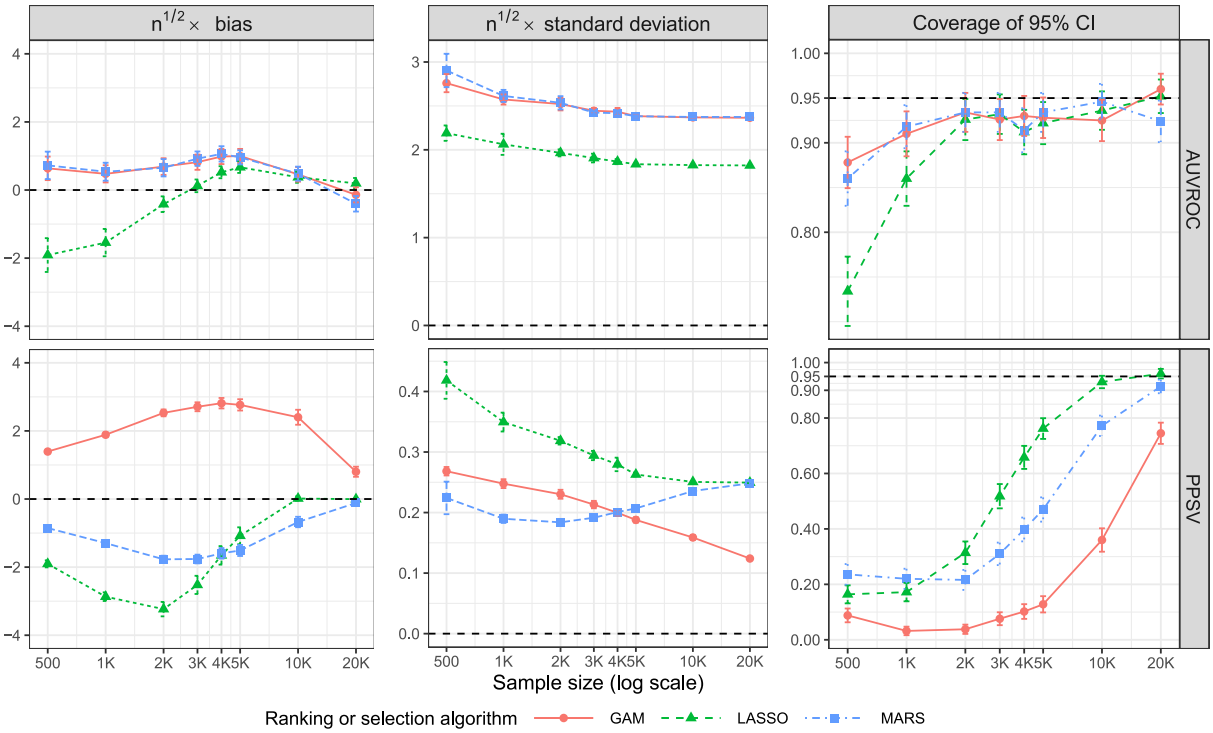

Figure 2. Properties of the AUVROC (top row) and PPSV (bottom row) estimators for the three algorithms. Vertical bars represent 95% CIs accounting for Monte Carlo error.

Wald confidence intervals for the estimators. The confidence intervals for the AUVROC are valid in large sample sizes and have generally good performance for $n \geq 2K$. However, for the PPSV, the confidence intervals have poor coverage. This is because the Wald confidence intervals do not take the variability of the selected set S_n into account. The denominator of the PPSV estimator is $|S_n|$, which is integer-valued and hence varies substantially unless the selection algorithm S_n is very stable. This results in a multi-modal sampling distribution of the PPSV estimator. The Wald interval uses a normal approximation to the sampling distribution, but the normal distribution is a bad approximation to the true multi-modal sampling distribution. The multi-modality of the sampling distribution is shown in additional figures in Supplementary Material. By sample size 20K, the selected set S_n is stable enough for the MARS and LASSO algorithms that the Wald confidence intervals have close to nominal coverage. However, the set selected by GAM is not stable even at this sample size, which results in poor coverage. As we will discuss below, our partial bootstrap procedure can in some cases improve the coverage of the PPSV.

Figure 3 displays the empirical coverage rate of 95% partial bootstrap confidence intervals. All bootstrap confidence intervals have close to 95% coverage for sample size 20K. The coverage of all types of bootstrap intervals for the AUVROC (top row) have comparable or better coverage than the Wald interval for small and moderate samples, and have coverage greater than 90% for all cases when $n \geq 1K$. The partial bootstrap yields better coverage than the Wald intervals for the AUVROC because the bootstrap procedure incorporates the variability of the ranking algorithm. For the PPSV (bottom row), the confidence intervals again have better coverage than the Wald intervals for the same reason. However, in this case, the percentile and percentile-t methods still have far from nominal coverage for $n < 10K$. Efron's percentile method, which uses the quantiles of the sampling distribution of the estimator directly to construct confidence intervals, has the best coverage by a wide margin. Efron's percentile method has close to nominal coverage for most sample sizes for the GAM algorithm, but requires larger sample sizes for the LASSO and MARS algorithms.

6. Application to Wine Quality Prediction

In this section, we use the methods developed in this article to compare variable selection and ranking algorithms for predicting the quality of wine. We use the data described in Cortez et al. (2009), which are publicly available at *https://archive.ics.uci.edu/ml/datasets/Wine+Quality*. This data contains 11 physicochemical properties of $n = 4898$ different *vinho verde* wines. We treat these as our covariates X. The data also contain a quality score between 0 and 10, which was computed as the median of at least three blind taste tests. We treat this as the outcome Y. Hence, our goal is to predict the subjective rating of a wine based on its physicochemical properties.

We used LASSO, GAM, and MARS to rank and select among the 11 physicochemical properties in terms of their importance for predicting wine quality. We refer the reader to Section 5 for precise explanations of these ranking and selection algorithms. We evaluated these algorithms using the R-squared predictiveness metric. As in Section 5, to estimate the regression function, we used SuperLearner with candidate library consisting of `xgboost`, `gam`, and `earth`. We used the estimators based on cross-fitting with $K = 5$ folds, and we constructed pointwise confidence intervals using the Wald method and equi-precision uniform confidence bands with influence function-based (co)variance estimator.

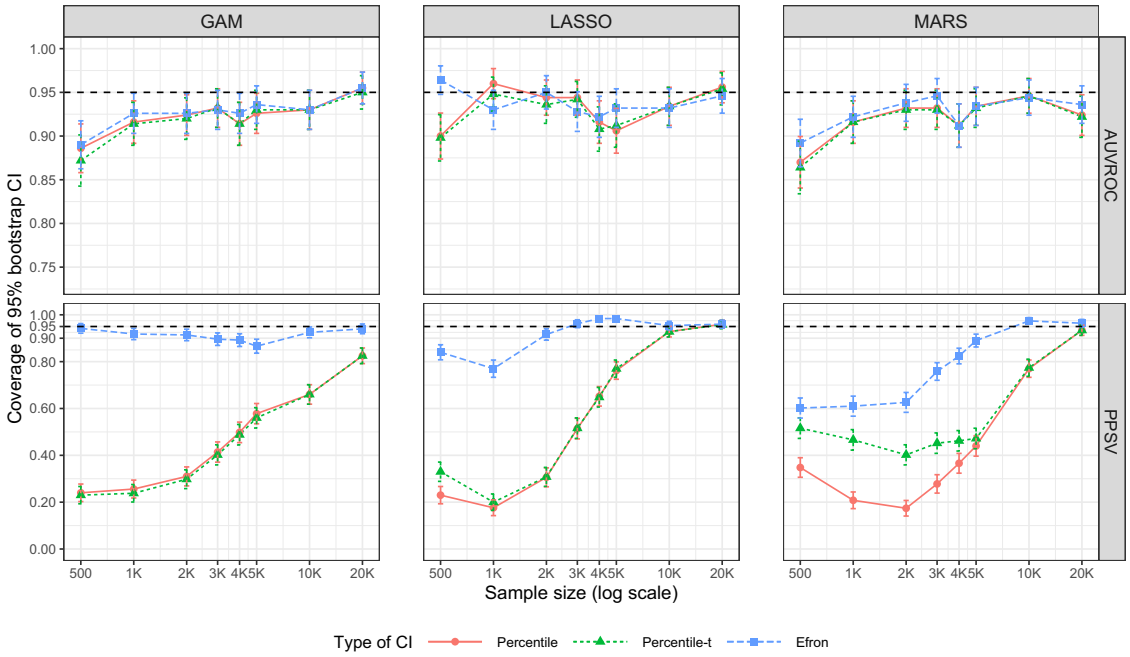

Figure 3. Empirical coverage of 95% confidence intervals using the partial bootstrap for the AUVROC (top row) and PPSV (bottom row) for the three algorithms. Vertical bars represent 95% CIs accounting for Monte Carlo error.

Table 1. Ranks of physicochemical properties for predicting wine rating.

	volatile acidity	free sulfur dioxide	residual sugar	alcohol	citric acid	fixed acidity	density	pH	sulphates	total sulfur dioxide	chlorides
MARS	1	2	3	4	5	6	7	8	9	10	11
GAM	1	2	3	8	6	9	4	5	7	10	11
LASSO	2	3	4	1	11	5	9	8	7	10	6

Table 1 displays the ranking of the physicochemical properties returned by the three ranking algorithms. All three algorithms ranked volatile acidity, free sulfur dioxide, and residual sugar among the top four most important variables, and MARS and GAM agreed on the ordering of these three. MARS and LASSO agree that alcohol is also important, but the rankings diverge somewhat after these first four rankings. All three algorithms ranked total sulfur dioxide second to last, and MARS and GAM both ranked chlorides last.

The left panel of Figure 4 displays the estimated VROC curves for the R-squared predictiveness metric scaled by the total R-squared predictiveness metric, along with the corresponding 95% pointwise and simultaneous confidence sets. We note that all three unscaled VROC curves ended at the point $(11, 0.44)$, implying that approximately 44% of the variance in wine ratings is accounted for by the 11 physicochemical predictors. None of the three scaled VROC curves were concave, suggesting that the 11 variables contribute roughly equally to the prediction, rather than a single or small set of variables outweighing the importance of the others. In fact, the largest increment for all three curves came when including variables 9, 10, and 11. This suggests that while the last few physicochemical properties alone may have had limited ability to explain the variation in wine quality, they were able to account for a substantial amount of variation when combined with the other physicochemical properties. This indicates the existence of interactions between the physicochemical properties. Furthermore, the MARS ranking algorithm generally had the largest estimated R-squared predictiveness of the three algorithms, which suggests that MARS was best able to assess the relative importance of the variables in the presence of interactions. Pairwise tests rejected the null hypothesis that there is no difference in the VROC curves with $p < 10^{-8}$ for all three pairs. We estimate that the AUVROC for MARS was 1.90, (1.75–2.06), the AUVROC for GAM was 1.62 (1.48–1.76), and the AUVROC for LASSO was 1.60 (1.46–1.74). Pairwise tests rejected the null hypothesis that there is no difference between the AUVROC of MARS and GAM ($p = 2.3 \times 10^{-7}$) and between MARS and LASSO ($p = 1.0 \times 10^{-6}$), but not that there was no difference between GAM and LASSO ($p = 0.71$).

The right panel of Figure 4 displays the R-squared predictiveness metric of the selected variables scaled by the total R-squared predictiveness metric, along with corresponding 95% confidence intervals. The GAM selection algorithm selected all variables and had an estimated PPSV of 0.039 (0.037–0.041). MARS selected ten variables and had an estimated PPSV of 0.035 (0.032–0.037). LASSO selected nine variables and had an estimated PPSV of 0.032 (0.029–0.035).

7. Conclusion

In this article, we proposed nonparametric, algorithm-agnostic measures of the quality of variable selection and ranking procedures. We proposed plug-in estimators of our measures, and provided conditions under which our estimators are asymptotically linear. Our theoretical results generalize those of Williamson et al. (2022) because our proposed measures

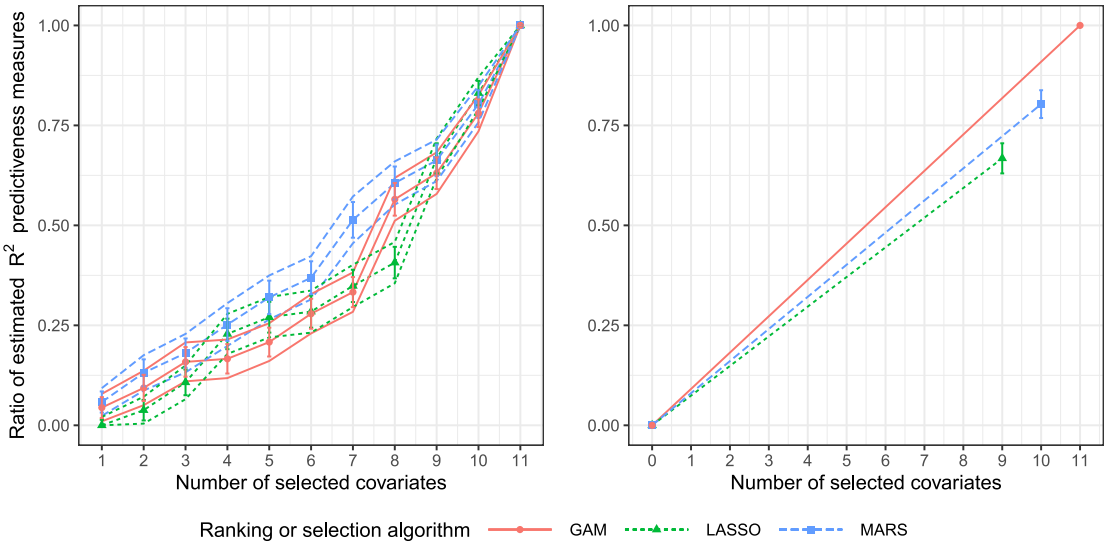

Figure 4. Left panel: Ratio of estimated VROC curve using *R*-squared predictiveness metric for prediction of wine rating by physicochemical properties with corresponding 95% pointwise and uniform confidence intervals. Right panel: Ratio of estimated *R*-squared predictiveness measure of selected physicochemical properties for prediction of wine rating with corresponding 95% pointwise confidence intervals.

are based on the variable importance framework introduced therein, but with random rather than fixed variable sets. We used our asymptotic results to construct large-sample confidence regions for our proposed measures. We also proposed a computationally efficient partial bootstrap procedure to account for finite-sample variability in the variable selection or ranking procedure not accounted for in the asymptotic results.

There are several natural extensions to our work. First, some variable selection and ranking procedures may not be asymptotically stable as defined in Section 2.6. In these cases, it is not clear how to even define a parameter of interest, or to achieve valid inference for the parameter. This is an important area of future research. Second, variable selection and ranking is of interest outside of classical regression analysis, such as causal inference and survival analysis (Fan, Feng, and Wu 2010; Shortreed and Ertefaie 2017), and our methods could in principle be extended to these areas as well.

Supplementary Materials

Appendices: Details on efficiency theory and construction of Wald confidence intervals; proofs of all theorems; verification of conditions for the examples; and additional results from numerical studies. (pdf)

R-code: R code implementing the methods described in the article. The code also contains replication code for the numerical studies and data analysis. (zip)

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The authors gratefully acknowledge support from NSF Award 2113171 (TW, ZT) and helpful comments from Brian Williamson and an anonymous referee.

ORCID

Ted Westling http://orcid.org/0000-0002-3362-1378

References

Akaike, H. (1998), *Information Theory and an Extension of the Maximum Likelihood Principle*, pp. 199–213, New York: Springer. [3]

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997), "Resampling Fewer than n Observations: Gains, Losses, and Remedies for Losses," *Statistica Sinica*, 7, 1–31. [7]

Bickel, P. J., and Li, B. (2007), "Local Polynomial Regression on Unknown Manifolds," *Lecture Notes-Monograph Series*, 54, 177–186. [7]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1]

Bühlmann, P., and van De Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*," Berlin, Heidelberg: Springer. [7]

Chen, T., and Guestrin, C. (2016), "Xgboost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. [8]

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2021), *xgboost: Extreme Gradient Boosting*, R package version 1.5.0.2. [8]

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling Wine Preferences by Data Mining from Physicochemical Properties," *Decision Support Systems*, 47, 547–553. Smart Business Networks: Concepts and Empirical Evidence. [9]

Cover, T. M. (1969), "Learning in Pattern Recognition," in *Methodologies of Pattern Recognition*, eds. S. Watanabe, pp. 111–132, London: Academic Press. [1]

Diciccio, T. J., and Romano, J. P. (1988), "A Review of Bootstrap Confidence Intervals," *Journal of the Royal Statistical Society*, Series B, 50, 338–354. [7]

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics. [7]

Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, New York: Chapman and Hall/CRC. [7]

Fan, J., Feng, Y., and Wu, Y. (2010), "High-Dimensional Variable Selection for Cox's Proportional Hazards Model," in J. O. Berger, T. T. Cai, and I. M. Johnstone (Eds.), *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown* (Vol. 6), pp. 70–87, Institute of Mathematical Statistics. [11]

Fan, J., and Li, R. (2002), "Variable Selection for Cox's proportional Hazards Model and Frailty Model," *The Annals of Statistics*, 30, 74–99. [1]

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67. [4,8]

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [1]

Hall, P. (1988), "Theoretical Comparison of Bootstrap Confidence Intervals," *The Annals of Statistics*, 16, 927–953. [7]

——— (1992), *The Bootstrap and Edgeworth Expansion*, New York, NY: Springer. [7]

Hastie, T. (2020), *gam: Generalized Additive Models*, R package version 1.20. [8]

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–310. [4,8]

Heinze, G., Wallisch, C., and Dunkler, D. (2018), "Variable Selection–A Review and Recommendations for the Practicing Statistician," *Biometrical Journal*, 60, 431–449. [1]

Hocking, R. R., and Leslie, R. N. (1967), "Selection of the Best Subset in Regression Analysis," *Technometrics*, 9, 531–540. [1]

Janet, J. P., and Kulik, H. J. (2017), "Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships," *The Journal of Physical Chemistry A*, 121, 8939–8954. [1]

Kennedy, E. H. (2016), "Semiparametric Theory and Empirical Processes in Causal Inference," in *Statistical Causal Inferences and Their Applications in Public Health Research*, eds. H. He, P. Wu, and D.-G. D. Chen, pp. 141–167, Cham: Springer. [6]

Lachenbruch, P. A., and Mickey, M. R. (1968), "Estimation of Error Rates in Discriminant Analysis," *Technometrics*, 10, 1–11. [1]

Le Cam, L., LeCam, L. M., and Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, New York, NY: Springer. [6]

Leclerc, R. D. (2008), "Survival of the Sparsest: Robust Gene Networks are Parsimonious," *Molecular systems biology*, 4, 213. [1]

Milborrow, S. (2021), *earth: Multivariate Adaptive Regression Splines*, R package version 5.3.1. [8]

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [1]

Murtaugh, P. A. (2009), "Performance of Several Variable-Selection Methods Applied to Real Ecological Data," *Ecology Letters*, 12, 1061–1068. [1]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [1]

Refaeilzadeh, P., Tang, L., and Liu, H. (2007), "On Comparison of Feature Selection Algorithms," in *Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II*, pp. 34–39. [1]

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., and Churpek, M. M. (2018), "Comparison of Variable Selection Methods for Clinical Predictive Modeling," *International Journal of Medical Informatics*, 116, 10–17. [1]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [3]

Shortreed, S. M., and Ertefaie, A. (2017), "Outcome-Adaptive Lasso: Variable Selection for Causal Inference," *Biometrics*, 73, 1111–1122. [11]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [7]

——— (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [7]

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–147. [1]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1,8]

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 6, 523–539. [7,8]

van der Laan, M. J., and Rose, S. (2011), *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York, NY: Springer. [7]

——— (2018), *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, Cham: Springer. [7]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press. [6,8]

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2022), "A General Framework for Inference on Algorithm-Agnostic Variable Importance," *Journal of the American Statistical Association*, 118, 1645–1658. [2,3,6,10]

Woodward, P. (1953), *Probability and Information Theory: With Applications to Radar* (Vol. 3), New York: McGraw-Hill. [4]

Yang, Y., and Tokdar, S. T. (2015), "Minimax-Optimal Nonparametric Regression in Hhigh Dimensions," *The Annals of Statistics*, 43, 652–674. [7]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1]