Journal of Machine Learning Research 20 (2019) 1-81 Submitted 11/18; Revised 8/19; Published 12/19

All Models are Wrong, but *Many* are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously

Aaron Fisher AFISHE27@ALUMNI.JH.EDU
Takeda Pharmaceuticals
Cambridge, MA 02139, USA

Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Departments of Computer Science and Electrical and Computer Engineering
Duke University
Durham, NC 27708, USA

Francesca Dominici FDOMINIC@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

(Authors are listed in order of contribution, with highest contribution listed first.)

Editor: Maya Gupta

Abstract

Variable importance (VI) tools describe how much covariates contribute to a prediction model's accuracy. However, important variables for one well-performing model (for example, a linear model $f(\mathbf{x}) = \mathbf{x}^T\beta$ with a fixed coefficient vector β) may be unimportant for another model. In this paper, we propose model class reliance (MCR) as the range of VI values across *all* well-performing model in a prespecified class. Thus, MCR gives a more comprehensive description of importance by accounting for the fact that many prediction models, possibly of different parametric forms, may fit the data well. In the process of deriving MCR, we show several informative results for permutation-based VI estimates, based on the VI measures used in Random Forests. Specifically, we derive connections between permutation importance estimates for a *single* prediction model, U-statistics, conditional variable importance, conditional causal effects, and linear model coefficients. We then give probabilistic bounds for MCR, using a novel, generalizable technique. We apply MCR to a public data set of Broward County criminal records to study the reliance of recidivism prediction models on sex and race. In this application, MCR can be used to help inform VI for unknown, proprietary models.

Keywords: Rashomon, permutation importance, conditional variable importance, U-statistics, transparency, interpretable models

1. Introduction

Variable importance (VI) tools describe how much a prediction model's accuracy depends on the information in each covariate. For example, in Random Forests, VI is measured by

prespecified 预先设定的

MCR:
in a prespecified class ?
parametric forms 参数的形式
informative 提供有用信息的

Variable importance VI

the decrease in prediction accuracy when a covariate is permuted (Breiman, 2001; Breiman et al., 2001; see also Strobl et al., 2008; Altmann et al., 2010; Zhu et al., 2015; Gregorutti et al., 2015; Datta et al., 2016; Gregorutti et al., 2017). A similar "Perturb" VI measure has been used for neural networks, where noise is added to covariates (Recknagel et al., 1997; Yao et al., 1998; Scardi and Harding, 1999; Gevrey et al., 2003). Such tools can be useful for identifying covariates that must be measured with high precision, for improving the transparency of a "black box" prediction model (see also Rudin, 2019), or for determining what scenarios may cause the model to fail.

However, existing VI measures do not generally account for the fact that many prediction models may fit the data almost equally well. In such cases, the model used by one analyst may rely on entirely different covariate information than the model used by another analyst. This common scenario has been called the "Rashomon" effect of statistics (Breiman et al., 2001; see also Lecué, 2011; Statnikov et al., 2013; Tulabandhula and Rudin, 2014; Nevo and Ritov, 2017; Letham et al., 2016). The term is inspired by the 1950 Kurosawa film of the same name, in which four witnesses offer different descriptions and explanations for the same encounter. Under the Rashomon effect, how should analysts give comprehensive descriptions of the importance of each covariate? How well can one analyst recover the conclusions of another? Will the model that gives the best predictions necessarily give the most accurate interpretation?

To address these concerns, we analyze the *set* of prediction models that provide near-optimal accuracy, which we refer to as a *Rashomon set*. This approach stands in contrast to training to select a *single* prediction model, among a prespecified class of candidate models. Our motivation is that Rashomon sets (defined formally below) summarize the range of effective prediction strategies that an analyst might choose. Additionally, even if the candidate models do not contain the true data generating process, we may hope that some of these models function in similar ways to the data generating process. In particular, we may hope there exist well performing candidate models that place the same importance on a variable of interest as the underlying data generating process does. If so, then studying sets of well-performing models will allow us to deduce information about the data generating process.

Applying this approach to study variable importance, we define *model class reliance* (MCR) as the highest and lowest degree to which any well-performing model within a given class may rely on a variable of interest for prediction accuracy. Roughly speaking, MCR captures the range of explanations, or mechanisms, associated with well-performing models. Because the resulting range summarizes many prediction models simultaneously, rather a single model, we expect this range to be less affected by the choices that an individual analyst makes during the model-fitting process. Instead of reflecting these choices, MCR aims to reflect the nature of the prediction problem itself.

We make several, specific technical contributions in deriving MCR. First, we review a core measure of how much an individual prediction model relies on covariates of interest for its accuracy, which we call *model reliance* (MR). This measure is based on permutation importance measures for Random Forests (Breiman et al., 2001; Breiman, 2001), and can be expanded to describe conditional importance (see Section 8, as well as Strobl et al. 2008). We draw a connection between permutation-based importance estimates (MR) and U-statistics, which facilitates later theoretical results. Additionally, we derive connections

Covariates 协变量

受生门效应：

不同的模型可能会依赖于不同的协变量信息

the set of prediction model

?

Rashomon set.

within a given class ?

Covariates of interest for its accuracy

其准确性相关的协变量.

{ model reliance (MR)
feature importance
variable importance

between MR, conditional causal effects, and coefficients for additive models. Expanding on MR, we propose MCR, which generalizes the definition of MR for a *class of models*. We derive finite-sample bounds for MCR, which motivate an intuitive estimator of MCR. Finally, we propose computational procedures for this estimator.

The tools we develop to study Rashomon sets are quite general, and can be used to make finite-sample inferences for arbitrary characteristics of well-performing models. For example, beyond describing variable importance, these tools can describe the range of risk predictions that well-fitting models assign to a particular covariate profile, or the variance of predictions made by well-fitting models. In some cases, these novel techniques may provide finite-sample confidence intervals (CIs) where none have previously existed (see Section 5).

MCR and the Rashomon effect become especially relevant in the context of criminal recidivism prediction. Proprietary recidivism risk models trained from criminal records data are increasingly being used in U.S. courtrooms. One concern is that these models may be relying on information that would otherwise be considered unacceptable (for example, race, sex, or proxies for these variables), in order to estimate recidivism risk. The relevant models are often proprietary, and cannot be studied directly. Still, in cases where the predictions made by these models are publicly available, it may be possible to identify alternative prediction models that are sufficiently similar to the proprietary model of interest.

In this paper, we specifically consider the proprietary model COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), developed by the company Northpointe Inc. (subsequently, in 2017, Northpointe Inc.,Courtview Justice Solutions Inc., and Constellation Justice Systems Inc. joined together under the name Equivant). Our goal is to estimate how much COMPAS relies on either race, sex, or proxies for these variables not measured in our data set. To this end, we apply a broad class of flexible, kernel-based prediction models to predict COMPAS score. In this setting, the MCR interval reflects the highest and lowest degree to which any prediction model in our class can rely on race and sex while still predicting COMPAS score relatively accurately. Equipped with MCR, we can relax the common assumption of being able to correctly specify the unknown model of interest (here, COMPAS) up to a parametric form. Instead, rather than assuming that the COMPAS model itself is contained in our class, we assume that our class contains at least one well-performing alternative model that relies on sensitive covariates to the same degree that COMPAS does. Under this assumption, the MCR interval will contain the VI value for COMPAS. Applying our approach, we find that race, sex, and their potential proxy variables, are likely not the dominant predictive factors in the COMPAS score (see analysis and discussion in Section 10).

The remainder of this paper is organized as follows. In Section 2 we introduce notation, and give a high level summary of our approach, illustrated with visualizations. In Sections 3 and 4 we formally present MR and MCR respectively, and derive theoretical properties of each. We also review related variable importance practices in the literature, such as retraining a model after removing one of the covariates. In Section 5, we discuss general applicability of our approach for determining finite-sample CIs for other problems. In Section 6, we present a general procedure for computing MCR. In Section 7, we give specific implementations of this procedure for (regularized) linear models, and linear models in a reproducing kernel Hilbert space. We also show that, for additive models, MR can be expressed in terms of the model's coefficients. In Section 8 we outline connections between

finite-sample inferences
有限样本推论

MR, causal inference, and conditional variable importance. In Section 9, we illustrate MR and MCR with a simulated toy example, to aid intuition. We also present simulation studies for the task of estimating MR for an unknown, underlying conditional expectation function, under misspecification. We analyze a well-known public data set on recidivism in Section 10, described above. All proofs are presented in the appendices.

2. Notation & Technical Summary

The label of "variable importance" measure has been broadly used to describe approaches for either inference (van der Laan, 2006; Díaz et al., 2015; Williamson et al., 2017) or prediction. While these two goals are highly related, we primarily focus on how much prediction models rely on covariates to achieve accuracy. We use terms such as "model reliance" rather than "importance" to clarify this context.

In order to evaluate how much prediction models rely on variables, we now introduce notation for random variables, data, classes of prediction models, and loss functions for evaluating predictions. Let $Z = (Y, X_1, X_2) \in \mathcal{Z}$ be a random variable with outcome $Y \in \mathcal{Y}$ and covariates $X = (X_1, X_2) \in \mathcal{X}$, where the covariate subsets $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$ may each be multivariate. We assume that observations of Z are *iid*, that $n \geq 2$, and that solutions to $\arg\min$ and $\arg\max$ operations exist whenever optimizing over sets mentioned in this paper (for example, in Theorem 4, below). Our goal is to study how much different prediction models rely on X_1 to predict Y.

We refer to our data set as $\mathbf{Z} = \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix}$, a matrix composed of a n-length outcome vector $\mathbf{y}$ in the first column, and a $n \times p$ covariate matrix $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$ in the remaining columns. In general, for a given vector $\mathbf{v}$, let $\mathbf{v}_{[j]}$ denote its j^{th} element(s). For a given matrix $\mathbf{A}$, let $\mathbf{A}'$, $\mathbf{A}_{[i,\cdot]}$, $\mathbf{A}_{[\cdot,j]}$, and $\mathbf{A}_{[i,j]}$ respectively denote the transpose of $\mathbf{A}$, the i^{th} row(s) of $\mathbf{A}$, the j^{th} column(s) of $\mathbf{A}$, and the element(s) in the i^{th} row(s) and j^{th} column(s) of $\mathbf{A}$.

We use the term *model class* to refer to a prespecified subset $\mathcal{F} \subset \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$ of the measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. We refer to member functions $f \in \mathcal{F}$ as *prediction models,* or simply as *models.* Given a model f, we evaluate its performance using a nonnegative *loss function* $L : (\mathcal{F} \times \mathcal{Z}) \to \mathbb{R}_{\geq 0}$. For example, L may be the squared error loss $L_{\text{se}}(f, (y, x_1, x_2)) = (y - f(x_1, x_2))^2$ for regression, or the hinge loss $L_{\text{h}}(f, (y, x_1, x_2)) = (1 - yf(x_1, x_2))_+$ for classification. We use the term *algorithm* to refer to any procedure $\mathcal{A} : \mathcal{Z}^n \to \mathcal{F}$ that takes a data set as input and returns a model $f \in \mathcal{F}$ as output.

2.1. Summary of Rashomon Sets & Model Class Reliance

Many traditional statistical estimates come from descriptions of a *single,* fitted prediction model. In contrast, in this section, we summarize our approach for studying a *set* of near-optimal models. To define this set, we require a prespecified "reference" model, denoted by f_{ref}, to serve as a benchmark for predictive performance. For example, f_{ref} may come from a flowchart used to predict injury severity in a hospital's emergency room, or from another quantitative decision rule that is currently implemented in practice. Given a reference model f_{ref}, we define a *population ϵ-Rashomon set* as the subset of models

inference / prediction

iid: independent and identically distributed

model class 是所有预定的模型集合，$f: \mathcal{X} \to \mathcal{Y}$

于是 model

with expected loss no more than ϵ above that of f_{ref}. We denote this set as $\mathcal{R}(\epsilon) := \{f \in \mathcal{F} : \mathbb{E}L(f, Z) \leq \mathbb{E}L(f_{\text{ref}}, Z) + \epsilon\}$, where $\mathbb{E}$ denotes expectations with respect to the population distribution. This set can be thought of as representing models that might be arrived at due to differences in data measurement, processing, filtering, model parameterization, covariate selection, or other analysis choices (see Section 4).

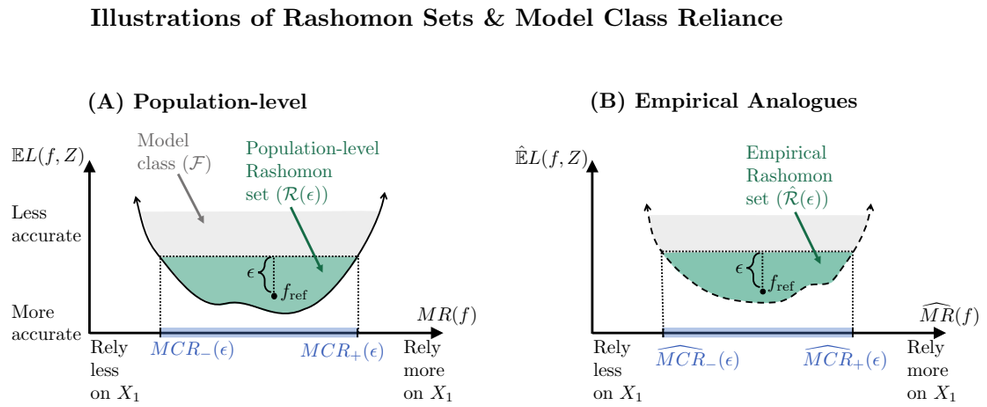

Figure 1: Rashomon sets and model class reliance – Panel (A) illustrates a hypothetical Rashomon set $\mathcal{R}(\epsilon)$, within a model class $\mathcal{F}$. The y-axis shows the expected loss of each model $f \in \mathcal{F}$, and the x-axis shows how much each model f relies on X_1 (defined formally in Section 3). Along the x-axis, the population-level MCR range is highlighted in blue, showing the values of MR corresponding to well-performing models (see Section 4). Panel (B) shows the in-sample analogue of Panel (A). Here, the y-axis denotes the in-sample loss, $\hat{\mathbb{E}}L(f, Z) := \frac{1}{n}\sum_{i=1}^{n} L(f, \mathbf{Z}_{[i,\cdot]})$; the x-axis shows the empirical model reliance of each model $f \in \mathcal{F}$ on X_1 (see Section 3); and the highlighted portion of the x-axis shows empirical MCR (see Section 4).

Figure 1-A illustrates a hypothetical example of a population ϵ-Rashomon set. Here, the y-axis shows the expected loss of each model $f \in \mathcal{F}$, and the x-axis shows how much each model relies on X_1 for its predictive accuracy. More specifically, given a prediction model f, the x-axis shows the percent increase in f's expected loss when noise is added to X_1. We refer to this measure as the *model reliance* (MR) of f on X_1, written informally as

$$MR(f) := \frac{\text{Expected loss of } f \text{ under noise}}{\text{Expected loss of } f \text{ without noise}}. \tag{2.1}$$

The added noise must satisfy certain properties, namely, it must render X_1 completely uninformative of the outcome Y, without altering the marginal distribution of X_1 (for details, see Section 3, as well as Breiman, 2001; Breiman et al., 2001).

Our central goal is to understand how much, or how little, models may rely on covariates of interest (X_1) while still predicting well. In Figure 1-A, this range of possible MR values

is shown by the highlighted interval along the x-axis. We refer to an interval of this type as a *population-level model class reliance* (MCR) range (see Section 4), formally defined as

$$[MCR_-(\epsilon),\ MCR_+(\epsilon)] := \left[\min_{f\in\mathcal{R}(\epsilon)} MR(f),\ \max_{f\in\mathcal{R}(\epsilon)} MR(f)\right]. \tag{2.2}$$

To estimate this range, we use empirical analogues of the population ϵ-Rashomon set, and of MR, based on observed data (Figure 1-B). We define an *empirical ϵ-Rashomon set* as the set of models with *in-sample* loss no more than ϵ above that of f_{ref}, and denote this set by $\hat{\mathcal{R}}(\epsilon)$. Informally, we define the *empirical* MR of a model f on X_1 as

$$\widehat{MR}(f) := \frac{\text{In-sample loss of } f \text{ under noise}}{\text{In-sample loss of } f \text{ without noise}}, \tag{2.3}$$

that is, the extent to which f appears to rely on X_1 in a given sample (see Section 3 for details). Finally, we define the *empirical model class reliance* as the range of empirical MR values corresponding to models with strong in-sample performance (see Section 4), formally written as

$$[\widehat{MCR}_-(\epsilon),\ \widehat{MCR}_+(\epsilon)] := \left[\min_{f\in\hat{\mathcal{R}}(\epsilon)} \widehat{MR}(f),\ \max_{f\in\hat{\mathcal{R}}(\epsilon)} \widehat{MR}(f)\right]. \tag{2.4}$$

In Figure 1-B, the above range is shown by the highlighted portion of the x-axis.

We make several technical contributions in the process of developing MCR.

- **Estimation of MR, and population-level MCR:** Given f, we show desirable properties of $\widehat{MR}(f)$ as an estimator of $MR(f)$, using results for U-statistics (Section 3.1 and Theorem 5). We also derive finite sample bounds for population-level MCR, some of which require a limit on the complexity of $\mathcal{F}$ in the form of a covering number. These bounds demonstrate that, under fairly weak conditions, empirical MCR provides a sensible estimate of population-level MCR (see Section 4 for details).

- **Computation of empirical MCR:** Although empirical MCR is fully determined given a sample, the minimization and maximization in Eq 2.4 require nontrivial computations. To address this, we outline a general optimization procedure for MCR (Section 6). We give detailed implementations of this procedure for cases when the model class $\mathcal{F}$ is a set of (regularized) linear regression models, or a set of regression models in a reproducing kernel Hilbert space (Section 7). The output of our proposed procedure is a closed-form, convex envelope containing $\mathcal{F}$, which can be used to approximate empirical MCR for any performance level ϵ (see Figure 2 for an illustration). Still, for complex model classes where standard empirical loss minimization is an open problem (for example, neural networks), computing empirical MCR remains an open problem as well.

- **Interpretation of MR in terms of model coefficients, and causal effects:** We show that MR for an additive model can be written as a function of the model's coefficients (Proposition 15), and that MR for a binary covariate X_1 can be written as a function of the conditional causal effects of X_1 on Y (Proposition 19).

- **Extensions to conditional importance:** We provide an extension of MR that is analogous to the notion of conditional importance (Strobl et al., 2008). This extension describes how much a model relies on the specific information in X_1 that cannot otherwise be gleaned from X_2 (Section 8.2).
- **Generalizations for Rashomon sets:** Beyond notions of variable importance, we also generalize our finite sample results for MCR to describe arbitrary characterizations of models in a population ϵ-Rashomon set. As we discuss in concurrent work (Coker et al., 2018), this generalization is analogous to the profile likelihood interval, and can, for example, be used to bound the range of risk predictions that well-performing prediction models may assign to a particular set of covariates (Section 5).

We begin in the next section by formally reviewing model reliance.

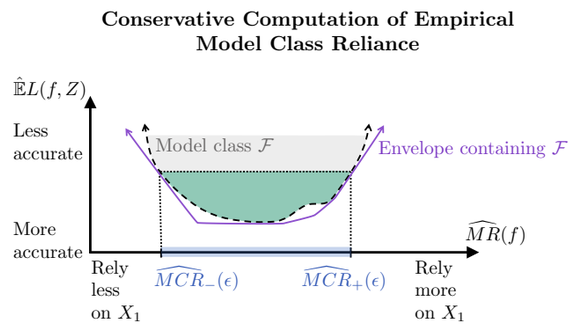

Figure 2: Illustration of output from our empirical MCR computational procedure – Our computation procedure produces a closed-form, convex envelope that contains $\mathcal{F}$ (shown above as the solid, purple line), which bounds empirical MCR for any value of ϵ (see Eq 2.4). The procedure works sequentially, tightening these bounds as much as possible near the ϵ value of interest (Section 6). The results from our data analysis (Figure 8) are presented in the same format as the above purple envelope.

3. Model Reliance

To formally describe how much the expected accuracy of a fixed prediction model f relies on the random variable X_1, we use the notion of a "switched" loss where X_1 is rendered uninformative. Throughout this section, we will treat f as a pre-specified prediction model of interest (as in Hooker, 2007). Let $Z^{(a)} = (Y^{(a)}, X_1^{(a)}, X_2^{(a)})$ and $Z^{(b)} = (Y^{(b)}, X_1^{(b)}, X_2^{(b)})$ be independent random variables, each following the same distribution as $Z = (Y, X_1, X_2)$. We define

$$e_{\text{switch}}(f) := \mathbb{E}L\{f, (Y^{(b)}, X_1^{(a)}, X_2^{(b)})\}$$

switched loss :
对x1被宣布为不提供信息的.
两个实例 $z^{(a)}$ $z^{(b)}$
对x1的值对调
switched expected loss

as representing the expected loss of model f across pairs of observations $(Z^{(a)}, Z^{(b)})$ in which the values of $X_1^{(a)}$ and $X_1^{(b)}$ have been switched. To see this interpretation of the above equation, note that we have used the variables $(Y^{(b)}, X_2^{(b)})$ from $Z^{(b)}$, but we have used the variable $X_1^{(b)}$ from an independent copy $Z^{(b)}$. This is why we say that $X_1^{(a)}$ and $X_1^{(b)}$ have been switched; the values of $(Y^{(b)}, X_1^{(a)}, X_2^{(b)})$ do not relate to each other as they would if they had been chosen together. An alternative interpretation of $e_{\text{switch}}(f)$ is as the expected loss of f when noise is added to X_1 in such a way that X_1 becomes completely uninformative of Y, but that the marginal distribution of X_1 is unchanged.

As a reference point, we compare $e_{\text{switch}}(f)$ against the standard expected loss when none of the variables are switched, $e_{\text{orig}}(f) := \mathbb{E}L(f, (Y, X_1, X_2))$. From these two quantities, we formally define *model reliance* (MR) as the ratio,

$$MR(f) := \frac{e_{\text{switch}}(f)}{e_{\text{orig}}(f)}, \tag{3.1}$$

as we alluded to in Eq 2.1. Higher values of $MR(f)$ signify greater reliance of f on X_1. For example, an $MR(f)$ value of 2 means that the model relies heavily on X_1, in the sense that its loss doubles when X_1 is scrambled. An $MR(f)$ value of 1 signifies no reliance on X_1, in the sense that the model's loss does not change when X_1 is scrambled. Models with reliance values strictly less than 1 are more difficult to interpret, as they rely less on the variable of interest than a random guess. Interestingly, it is possible to have models with reliance less than one. For instance, a model f' may satisfy $MR(f') < 1$ if it treats X_1 and Y as positively correlated when they are in fact negatively correlated. However, in many cases, the existence of a model $f' \in \mathcal{F}$ satisfying $MR(f') < 1$ implies the existence of another, better performing model $f'' \in \mathcal{F}$ satisfying $MR(f'') = 1$ and $e_{\text{orig}}(f'') \leq e_{\text{orig}}(f')$. That is, although models may exist with MR values less than 1, they will typically be suboptimal (see Appendix A.2).

Model reliance could alternatively be defined as a difference rather than a ratio, that is, as $MR_{\text{difference}}(f) := e_{\text{switch}}(f) - e_{\text{orig}}(f)$. In Appendix A.5, we discuss how many of our results remain similar under either definition.

3.1. Estimating Model Reliance with U-statistics, and Connections to Permutation-based Variable Importance

Given a model f and data set $\mathbf{Z} = \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix}$, we estimate $MR(f)$ by separately estimating the numerator and denominator of Eq 3.1. We estimate $e_{\text{orig}}(f)$ with the standard empirical loss,

$$\hat{e}_{\text{orig}}(f) := \frac{1}{n}\sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}. \tag{3.2}$$

We estimate $e_{\text{switch}}(f)$ by performing a "switch" operation across all observed pairs, as in

$$\hat{e}_{\text{switch}}(f) := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\}. \tag{3.3}$$

Above, we have aggregated over all possible combinations of the observed values for (Y, X_2) and for X_1, excluding pairings that are actually observed in the original sample. If

些处应是笔误。

original (standard) expected loss

How to compute MR as Ratio.

alluded 暗指；影射

当 MR 为2时，model 非常依赖 X_1，因为 X_1 受干扰时的 loss 是2倍。

当 MR 为1时，model 没有受到 X_1 的干扰而产生损失。

当 MR 小于1时，说明有更好的模型来替代模型。

estimating 估算。

standard empirical loss 标准经验损失。

the summation over all possible pairs (Eq 3.3) is computationally prohibitive due to sample size, another estimator of $e_{\text{switch}}(f)$ is

$$\hat{e}_{\text{divide}}(f) := \frac{1}{2\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left[L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i+\lfloor n/2 \rfloor, \cdot]}, \mathbf{X}_{2[i,\cdot]})\} \right. \tag{3.4}$$

$$\left. + L\{f, (\mathbf{y}_{[i+\lfloor n/2 \rfloor]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[i+\lfloor n/2 \rfloor, \cdot]})\} \right]. \tag{3.5}$$

Here, rather than summing over all pairs, we divide the sample in half. We then match the first half's values for (Y, X_2) with the second half's values for X_1 (Line 3.4), and vice versa (Line 3.5). All three of the above estimators (Eqs 3.2, 3.3 & 3.5) are unbiased for their respective estimands, as we discuss in more detail shortly.

Finally, we can estimate $MR(f)$ with the plug-in estimator

$$\widehat{MR}(f) := \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)}, \tag{3.6}$$

which we define as the *empirical model reliance* of f on X_1. In this way, we formalize the empirical MR definition in Eq 2.3.

Again, our definition of empirical MR is very similar to the permutation-based variable importance approach of Breiman (2001), where Breiman uses a single random permutation and we consider all possible pairs. To compare these two approaches more precisely, let $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$ be a set of n-length vectors, each containing a different permutation of the set $\{1, \ldots, n\}$. The approach of Breiman (2001) is analogous to computing the loss $\sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]}, \cdot]}, \mathbf{X}_{2[i,\cdot]})\}$ for a randomly chosen permutation vector $\boldsymbol{\pi}_l \in \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$. Similarly, our calculation in Eq 3.3 is proportional to the sum of losses over all possible $(n!)$ permutations, excluding the n unique combinations of the rows of $\mathbf{X}_1$ and the rows of $\begin{bmatrix} \mathbf{X}_2 & \mathbf{y} \end{bmatrix}$ that appear in the original sample (see Appendix A.3). Excluding these observations is necessary to preserve the (finite-sample) unbiasedness of $\hat{e}_{\text{switch}}(f)$.

The estimators $\hat{e}_{\text{orig}}(f)$, $\hat{e}_{\text{switch}}(f)$ and $\hat{e}_{\text{divide}}(f)$ all belong to the well-studied class of U-statistics. Thus, under fairly minor conditions, these estimators are unbiased, asymptotically normal, and have finite-sample probabilistic bounds (Hoeffding, 1948, 1963; Serfling, 1980; see also DeLong et al., 1988 for an early use of U-statistics in machine learning, as well as caveats in Demler et al., 2012). To our knowledge, connections between permutation-based importance and U-statistics have not been previously established.

While the above results from U-statistics depend on the model f being fixed a priori, we can also leverage these results to create *uniform* bounds on the MR estimation error for all models in a sufficiently regularized class $\mathcal{F}$. We formally present this bound in Section 4 (Theorem 5), after introducing required conditions on model class complexity. The existence of this uniform bound implies that it is feasible to train a model and to evaluate its importance using the *same data*. This differs from the classical VI approach of Random Forests (Breiman, 2001), which avoids in-sample importance estimation. There, each tree in the ensemble is fit on a random subset of data, and VI for the tree is estimated using the held-out data. The tree-specific VI estimates are then aggregated to obtain a VI estimate for the overall ensemble. Although sample-splitting approaches such as this are helpful in many cases, the uniform bound for MR suggests that they are not strictly necessary, depending on the sample size and the complexity of $\mathcal{F}$.