All Models are Wrong but *Many* are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using *Model Class Reliance*

Aaron Fisher
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA
aafisher@hsph.harvard.edu

Cynthia Rudin
Departments of Computer Science and Electrical and Computer Engineering
Duke University
Durham, NC 27708, USA

Francesca Dominici[1]
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

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

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, similar to the VI measures used in Random Forests. Specifically, we derive connections between permutation importance estimates for a *single* prediction model, U-statistics, conditional causal effects, and linear model coefficients. We then give probabilistic bounds for MCR, using a novel, generalizable technique. We apply MCR in a public dataset 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.

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

[1]Authors listed in order of contribution, with highest contribution listed first.

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 improving the transparency of a "black box" a prediction model, for determining what scenarios may cause the model to fail, or for identifying covariates that must be measured with high precision.

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 Statnikov et al., 2013; Tulabandhula and Rudin, 2014; Nevo and Ritov, 2015; 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?

To address this, 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.

We make several technical contributions in deriving MCR. First, we define a core importance measure, *model reliance* (MR), as the degree to which a *specific model* relies on covariates of interest to predict well. This measure is based on permutation importance measures for Random Forests (Breiman, 2001; Breiman et al., 2001). We draw a connection between permutation-based importance estimates and U-statistics, which facilitates later theoretical results. Additionally, we derive connections between permutation importance, conditional causal effects, and coefficients for additive models. Next, we expand on MR to propose MCR, which captures the reliance values for a *class of models*. We derive finite-sample bounds for MCR, which motivate an intuitive estimator of MCR.

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 Equivant (formerly known as Northpointe). Our goal is to estimate how much COMPAS relies on either race, sex, or proxies for these variables not measured in our dataset. 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 9).

The tools we use to derive bounds for MCR and quite powerful and general, and can be used to make finite-sample (non-asymptotic) inferences for many other summary descriptions of well-performing models. For example, rather than describing how much well-performing models may rely on covariates of interest, the same techniques allow inference for the range of risk predictions that well-performing models may assign to a given covariate profile. In some cases, these novel techniques may provide finite-sample confidence intervals where none have previously existed. In others, they shed light on why certain types of model classes (for example, Neural Networks) have resisted rigorous inferential theory (see Section 5).

The remainder of this paper is organized as follows. In Section 2 we introduce notation and terminology. In Sections 3 and 4 we present MR and MCR respectively, and derive theoretical properties of each. In Section 5, we discuss general applicability of our approach for determining finite-sample confidence intervals for other problems. In Section 6 we review additional related literature in more detail. In particular, we discuss the common practice of retraining a model after removing one of the covariates. In Section 7, we present a general procedure for computing MCR, with specific implementations for (regularized) linear models, and linear models in a reproducing kernel Hilbert space. In Section 8, 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 the unknown, underlying conditional expectation function, under misspecification. We analyze a well-known public dataset on recidivism in Section 9. All proofs are presented in the appendices.

2 Terminology & notation: "importance," "models," and "model classes."

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.

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. Our goal is to study how much different prediction models rely on X_1 to predict Y. We refer to our dataset as $\mathbf{Z} = \left[\begin{array}{cc} \mathbf{y} & \mathbf{X} \end{array} \right]$, a matrix composed of a n-length outcome vector $\mathbf{y}$ in the first column, and a $n \times p$ covariate matrix $\mathbf{X} = \left[\begin{array}{cc} \mathbf{X}_1 & \mathbf{X}_2 \end{array} \right]$ in the remaining columns. 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 6, below).

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 $L_{\text{se}}(f, (y, x)) = (y - f(x))^2$ for regression, or the hinge loss $L_{\text{h}}(f, (y, x)) = (1 - yf(x))_+$ for classification.

We use the term *algorithm* to refer to any procedure $\mathcal{A} : \mathcal{Z}^n \to \mathcal{F}$ that takes a dataset as input and returns a model $f \in \mathcal{F}$ as output. We illustrate these terms with two brief toy examples in Appendix A.1.

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 $\mathbf{A}$.

3 Model reliance

To 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. 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 $Z = (Y, X_1, X_2)$. Denote realizations of $Z^{(a)}$ and $Z^{(b)}$ by $z^{(a)} = (y^{(a)}, x_1^{(a)}, x_2^{(a)})$ and $z^{(b)} = (y^{(b)}, x_1^{(b)}, x_2^{(b)})$ respectively. Given the realizations $z^{(a)}$ and $z^{(b)}$, let $h_f(z^{(a)}, z^{(b)})$ be the loss of model f on $z^{(b)}$, if $x_1^{(b)}$ was first replaced with $x_1^{(a)}$:

$$\begin{aligned} h_f(z^{(a)}, z^{(b)}) &= h_f((y^{(a)}, x_1^{(a)}, x_2^{(a)}), (y^{(b)}, x_1^{(b)}, x_2^{(b)})) \\ &:= L(f, (y^{(b)}, x_1^{(a)}, x_2^{(b)})). \end{aligned}$$

For a given prediction function f, we wish to know the expectation of this quantity across pairs in the population, $e_{\text{switch}}(f) := \mathbb{E}h_f(Z^{(a)}, Z^{(b)})$.

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

$$MR(f) := \frac{e_{\text{switch}}(f)}{e_{\text{orig}}(f)}.$$

Higher values of $MR(f)$ signify greater reliance of f on X_1, and $MR(f) = 1$ signifies no reliance on X_1. Models with reliance values strictly less than 1 are more difficult to interpret, but are often accompanied by better performing models satisfying $MR(f) = 1$ (see Appendix A.2).

Model reliance could alternatively be defined as a difference rather than a ratio, that is, with $MR_{\text{difference}}(f) := e_{\text{switch}}(f) - e_{\text{orig}}(f)$. 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 dataset $\mathbf{Z} = \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix}$, we estimate $e_{\text{orig}}(f)$ with the standard empirical loss

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

We estimate $e_{\text{switch}}(f)$ with

$$\hat{e}_{\text{switch}}(f) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L\{f, (\mathbf{y}_{[j]}, \mathbf{X}_{1[i,\cdot]}, \mathbf{X}_{2[j,\cdot]})\}. \tag{3.1}$$

To illustrate the connection between Eq 3.1 and the permutation-based variable importance approach of Breiman (2001), 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 3.1 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).

As an alternative to Eq 3.1, if the summation over all possible pairs is computationally prohibitive due to sample size, another estimator of $e_{\text{switch}}(f)$ is

$$\hat{e}_{\text{divide}}(f) := \frac{1}{2\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \big\{ h_f(\mathbf{Z}_{[i,\cdot]}, \mathbf{Z}_{[i+\lfloor n/2 \rfloor,\cdot]}) + h_f(\mathbf{Z}_{[i+\lfloor n/2 \rfloor,\cdot]}, \mathbf{Z}_{[i,\cdot]}) \big\}. \tag{3.2}$$

The above estimator divides the sample into a single, specific grouping of $\lfloor n/2 \rfloor$ pairs, and averages across these pairs.

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

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

which we refer to as the "empirical" model reliance of f on X_1.

Both $\hat{e}_{\text{switch}}(f)$ and $\hat{e}_{\text{divide}}(f)$ belong to the well-studied class of U-statistics. Thus, under fairly minor conditions, $\hat{e}_{\text{switch}}(f)$ and $\hat{e}_{\text{divide}}(f)$ are unbiased, asymptotically normal, and have finite-sample probabilistic bounds (Hoeffding, 1948; Serfling, 1980). 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 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 8), 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.

3.2 Model reliance and causal effects

In this section, we show a connection between population-level model reliance and the conditional average causal effect. For consistency with the causal inference literature, we temporarily rename the random variables (Y, X_1, X_2) as (Y, T, C), with realizations (y, t, c). Here, $T := X_1$ represents a binary treatment indicator, $C := X_2$ represents a set of baseline covariates, and Y represents an outcome of interest. Under this notation, $e_{\text{orig}}(f)$ represents the expected loss of a prediction function f, and $e_{\text{switch}}(f)$ denotes the expected loss in a pair

of observations in which the treatment has been switched. Let $f_0(t,c) := \mathbb{E}(Y|C=c,T=t)$ be the (unknown) conditional expectation function for Y, where we place no restrictions on the functional form of f_0.

Let Y_1 and Y_0 be potential outcomes under treatment and control respectively, such that $Y = Y_0(1-T) + Y_1 T$. The treatment effect for an individual is defined as $Y_1 - Y_0$, and the average treatment effect is defined as $\mathbb{E}(Y_1 - Y_0)$. Let $\text{CATE}(c) := \mathbb{E}(Y_1 - Y_0|C=c)$ be the (unknown) conditional average treatment effect of T for all patients with $C=c$. Causal inference methods typically assume $(Y_1, Y_0) \perp T|C$ (conditional ignorability), and $0 < P(T=1|C=c) < 1$ for all values of c (positivity), in order for f_0 and CATE to be well defined and identifiable.

The next theorem quantifies the relation between the conditional average treatment effect function (CATE) and the model reliance of f_0 on X_1, as measured by $e_{\text{switch}}(f_0) - e_{\text{orig}}(f_0)$. We choose this form of the result for simplicity, although dividing both sides of Eq 3.3 (below) by $e_{\text{orig}}(f_0)$ immediately translates the result in terms of a model reliance ratio.

Theorem 1. *For any prediction model f, let $e_{orig}(f)$ and $e_{switch}(f)$ be defined based on the squared error loss $L(f,(y,t,c)) := (y - f(t,c))^2$. Under the assumptions that $(Y_1, Y_0) \perp T|C$ (conditional ignorability) and $0 < P(T=1|C=c) < 1$ for all values of c (positivity), we have*

$$e_{switch}(f_0) - e_{orig}(f_0) \quad = \quad Var(T) \sum_{t \in \{0,1\}} \mathbb{E}_{C|T=t}\left(CATE(C)^2\right), \tag{3.3}$$

where $Var(T)$ is the marginal variance of the treatment assignment.

Intuitively, we see that the difference between $e_{\text{switch}}(f_0)$ and $e_{\text{orig}}(f_0)$ depends on the treatment prevalence and the conditional average treatment effect of T on Y for each covariate profile c. For example, if all patients are treated, then scrambling the treatment in a random pair of observations has no effect on the loss. In this case we see that $\text{Var}(T) = 0$ and $e_{\text{switch}}(f_0) = e_{\text{orig}}(f_0)$. Likewise, if the conditional average treatment effect is zero for every covariate profile c, then $\mathbb{E}(\text{CATE}(C)^2) = 0$, $e_{\text{switch}}(f_0) = e_{\text{orig}}(f_0)$, and f_0 does not rely on T. If the conditional average treatment effect is positive and constant for all values of C, then a larger treatment effect will result in a larger value for $e_{\text{switch}}(f_0) - e_{\text{orig}}(f_0)$, and a higher reliance of f_0 on T.

Importantly, $\mathbb{E}(\text{CATE}(C)^2)$ will yield different conclusions than the average treatment effect $\mathbb{E}(\text{CATE}(C))$ when there is treatment effect heterogeneity (that is, when $\text{Var}(\text{CATE}(C)) = \mathbb{E}(\text{CATE}(C)^2) - \mathbb{E}(\text{CATE}(C))^2 > 0$). For example, if a treatment is harmful, on average, in one subpopulation, but helpful in another subpopulation, then the average treatment effect may be zero while the average of the squared conditional average treatment effect, $\mathbb{E}(\text{CATE}(C)^2)$, will be positive.

3.3 Model reliance for linear models and additive models.

For linear models and the squared error loss, we can show both an interpretable definition of model reliance, as measured by $e_{\text{switch}}(f) - e_{\text{orig}}(f)$, as well as a computationally efficient formula for $\hat{e}_{\text{switch}}(f)$. The result is similar to Theorem 1, and augments results from Gregorutti et al. (2017).

Theorem 2. *For any prediction model f, let $e_{orig}(f)$, $e_{switch}(f)$, $\hat{e}_{\text{orig}}(f)$, and $\hat{e}_{switch}(f)$ be defined based on the squared error loss $L(f,(y,x_1,x_2)) := (y - f(x_1,x_2))^2$ for $y \in \mathbb{R}$,*

$x_1 \in \mathbb{R}^{p_1}$, and $x_2 \in \mathbb{R}^{p_2}$, where p_1 and p_2 are positive integers. Let $\beta = (\beta_1, \beta_2)$ and f_β satisfy $\beta_1 \in \mathbb{R}^{p_1}$, $\beta_2 \in \mathbb{R}^{p_2}$, and $f_\beta(x) = x'\beta = x_1'\beta_1 + x_2'\beta_2$. Then

$$e_{switch}(f_\beta) - e_{orig}(f_\beta) = 2Cov(Y, X_1)\beta_1 - 2\beta_2' Cov(X_2, X_1)\beta_1, \tag{3.4}$$

and, for finite samples,

$$\hat{e}_{switch}(f_\beta) = \frac{1}{n}\left\{\mathbf{y}'\mathbf{y} - 2\begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}' \beta + \beta' \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \beta \right\}, \tag{3.5}$$

where $\mathbf{W} := \frac{1}{n-1}(\mathbf{1}_n\mathbf{1}_n' - \mathbf{I}_n)$, $\mathbf{1}_n$ is the n-length vector of ones, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Eq 3.4 shows that model reliance for linear models, as measured by $e_{\text{switch}}(f_\beta) - e_{\text{orig}}(f_\beta)$, can be interpreted in terms of the population covariances and the model coefficients. Gregorutti et al. (2017) show an equivalent formulation of Eq 3.4 under the stronger assumptions that f_β is equal to the conditional expectation function of Y (that is, $f_\beta(x) = \mathbb{E}(Y|X = x)$), and the covariates X_1 and X_2 are centered.

Although the number of terms in the definition of $\hat{e}_{\text{switch}}(f)$ grows quadratically in n (see Eq 3.1), in the case of linear models we see from Eq 3.5 that the computational complexity of $\hat{e}_{\text{switch}}(f)$ grows *only linearly* in n. Specifically, the terms $\mathbf{X}_1'\mathbf{W}\mathbf{y}$ and $\mathbf{X}_1'\mathbf{W}\mathbf{X}_2$ in Eq 3.5 can be computed as $\frac{1}{n-1}\{(\mathbf{X}_1'\mathbf{1}_n)(\mathbf{1}_n'\mathbf{y}) - (\mathbf{X}_1'\mathbf{y})\}$ and $\frac{1}{n-1}\{(\mathbf{X}_1'\mathbf{1}_n)(\mathbf{1}_n'\mathbf{X}_2) - (\mathbf{X}_1'\mathbf{X}_2)\}$ respectively, where the computational complexity each term in parentheses grows linearly in n.

As in Gregorutti et al. (2017), both results in Theorem 2 readily generalize to additive models of the form $f_{g_1,g_2}(X_1, X_2) := g_1(X_1) + g_2(X_2)$, since permuting X_1 is equivalent to permuting $g_1(X_1)$.

4 Model class reliance

Like many statistical procedures, our MR measure (Section 3) produces a description of a *single* predictive model. Given a model with high predictive accuracy, MR describes how much the model's performance hinges on covariates of interest (X_1). However, there will often be many other models that perform similarly well, and that rely on X_1 to different degrees. With this notion in mind, we now study how much *any* well-performing model from a prespecified class $\mathcal{F}$ may rely on covariates of interest.

To formally define the set of well-performing models, we require a "benchmark" or "reference" model, denoted by f_{ref}. For $\epsilon \geq 0$, and a reference model f_{ref}, let $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F}) := \{f \in \mathcal{F} : e_{\text{orig}}(f) \leq e_{\text{orig}}(f_{\text{ref}}) + \epsilon\}$ be the subset of models with expected loss no more than ϵ above that of f_{ref}. We refer to $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F})$ as a population-level "Rashomon set" around f_{ref}. 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.

While f_{ref} could be selected by minimizing the in-sample loss, the theoretical study of $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F})$ is simplified under the assumption that f_{ref} is prespecified. 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. The model f_{ref} can also be selected using sample splitting. In some cases it may be desirable

to fix f_{ref} equal to the best-in-class model $f^\star := \arg\min_{f \in \mathcal{F}} e_{\text{orig}}(f)$, but this is generally infeasible because $f^\star$ is unknown. Still, for any $f_{\text{ref}} \in \mathcal{F}$, the Rashomon set $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F})$ around f_{ref} will always contain the Rashomon set $\mathcal{R}(\epsilon, f^\star, \mathcal{F})$ around $f^\star$. Hereafter, we assume that f_{ref} and $\mathcal{F}$ are fixed, and typically omit their specification when writing $\mathcal{R}(\epsilon)$.

We define population-level model class reliance (MCR) by maximizing and minimizing MR over $\mathcal{R}(\epsilon)$:

$$MCR_+(\epsilon) := \max_{f \in \mathcal{R}(\epsilon)} MR(f) \quad \text{and} \quad MCR_-(\epsilon) := \min_{f \in \mathcal{R}(\epsilon)} MR(f). \tag{4.1}$$

Studying the above quantities is the main focus of this paper, as these measures provide a more comprehensive view of importance than traditional measures (for example, Section 3). If $MCR_+(\epsilon)$ is low, then no well-performing model exists that places high importance on X_1, and X_1 can be discarded at low cost regardless of future modeling decisions. If $MCR_-(\epsilon)$ is large, then every well-performing model must rely substantially on X_1, and X_1 should be given careful attention during the modeling process. Here, $\mathcal{F}$ may itself consist of several parametric model forms (for example, all linear models and all decision tree models with less than 6 single-split nodes). We stress that the range $(MCR_-(\epsilon), MCR_+(\epsilon))$ does not depend on the *fitting algorithm* used to select a model $f \in \mathcal{F}$. The range is valid for any algorithm producing models in $\mathcal{F}$, and applies for any $f \in \mathcal{F}$.

To study MCR empirically, we introduce the sample analogues

$$\widehat{MCR}_+(\epsilon) := \max_{f \in \hat{\mathcal{R}}(\epsilon)} \widehat{MR}(f) \quad \text{and} \quad \widehat{MCR}_-(\epsilon) := \min_{f \in \hat{\mathcal{R}}(\epsilon)} \widehat{MR}(f), \tag{4.2}$$

where $\hat{\mathcal{R}}(\epsilon) := \{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \hat{e}_{\text{orig}}(f_{\text{ref}}) + \epsilon\}$. We refer to $\widehat{MCR}_+(\epsilon)$ and $\widehat{MCR}_-(\epsilon)$ as "empirical" MCR measures, and to $\hat{\mathcal{R}}(\epsilon)$ as an "empirical Rashomon set." We discuss the utility of these empirical measures in the next section.

In Appendix B.10 we consider an alternate formulation of Rashomon sets and MCR where we replace the relative loss threshold in the definition of $\mathcal{R}(\epsilon)$ with an absolute loss threshold. This alternate formulation can be similar in practice, but still requires the specification of a reference function to ensure that $\mathcal{R}(\epsilon)$ and $\hat{\mathcal{R}}(\epsilon)$ are nonempty.

4.1 Finite-sample bounds for model class reliance.

In this section we derive finite-sample, probabilistic bounds for $MCR_+(\epsilon)$ and $MCR_-(\epsilon)$ based on empirical MCR. While the bounds we present are conservative, they imply that, under minimal assumptions, $\widehat{MCR}_+(\epsilon)$ and $\widehat{MCR}_-(\epsilon)$ are sensible point estimates for $MCR_+(\epsilon)$ and $MCR_-(\epsilon)$. Thus, in Sections 8.1 & 9, we use $\widehat{MCR}_+(\epsilon)$ and $\widehat{MCR}_-(\epsilon)$ as point estimates, and apply a bootstrap procedure to obtain confidence intervals.

To derive these results we introduce three bounded loss assumptions, each of which can be assessed empirically. Let $b_{\text{orig}}, B_{\text{ind}}, B_{\text{ref}}, B_{\text{switch}} \in \mathbb{R}$ be known constants.

Assumption 3. *(Bounded individual loss) For a given model $f \in \mathcal{F}$, assume that $0 \leq L(f, (y, x_1, x_2)) \leq B_{ind}$ for any $(y, x_1, x_2) \in (\mathcal{Y} \times \mathcal{X}_1 \times \mathcal{X}_2)$.*

Assumption 4. *(Bounded relative loss) For a given model $f \in \mathcal{F}$, assume that $|L(f, (y, x_1, x_2)) - L(f_{ref}, (y, x_1, x_2))| \leq B_{ref}$ for any $(y, x_1, x_2) \in \mathcal{Z}$.*

Assumption 5. *(Bounded aggregate loss) For a given model $f \in \mathcal{F}$, assume that $\mathbb{P}\{0 < b_{orig} \leq \hat{e}_{orig}(f)\} = \mathbb{P}\{\hat{e}_{switch}(f) \leq B_{switch}\} = 1$.*

Each assumption is a property of a specific model $f \in \mathcal{F}$. The notation B_{ind} and B_{ref} refer to bounds for any individual observation, and the notation b_{orig} and B_{switch} refer to bounds on the aggregated loss L in a sample.

We give example methods of determining B_{ind} in Sections 7.4.2 & 7.5.2. For Assumption 5, we can approximate b_{orig} by training a highly flexible model to the data, and setting b_{orig} equal to half (or any positive fraction) of the resulting cross-validated loss. To determine B_{switch} we can simply set $B_{\text{switch}} = B_{\text{ind}}$, although this may be conservative. For example, in the case of binary classification models for non-separated groups (see Section 8.1), no linear classifier can misclassify all observations, particularly after a covariate is permuted. Thus, it must hold that $B_{\text{ind}} > B_{\text{switch}}$. Similarly, if f_{ref} satisfies Assumption 3, then B_{ref} may be conservatively set equal to B_{ind}. If model reliance is redefined as a difference rather than a ratio, then a similar form of the results in this section will apply without Assumption 5 (see Appendix A.5).

Based on these assumptions, we can create a finite-sample upper bound for $MCR_+(\epsilon)$ and lower bound for $MCR_-(\epsilon)$.

Theorem 6. *Given a constant $\epsilon \geq 0$, let $f_{+,\epsilon} \in \arg\max_{\mathcal{R}(\epsilon)} MR(f)$ and $f_{-,\epsilon} \in \arg\min_{\mathcal{R}(\epsilon)} MR(f)$ be prediction models that attain the highest and lowest model reliance among models in $\mathcal{R}(\epsilon)$. If $f_{+,\epsilon}$ and $f_{-,\epsilon}$ satisfy Assumptions 3, 4 & 5, then*

$$\mathbb{P}\left(MCR_+(\epsilon) > \widehat{MCR}_+(\epsilon_1) + \mathcal{Q}_1\right) \leq \delta, \text{ and} \tag{4.3}$$

$$\mathbb{P}\left(MCR_-(\epsilon) < \widehat{MCR}_-(\epsilon_1) - \mathcal{Q}_1\right) \leq \delta, \tag{4.4}$$

where $\epsilon_1 := \epsilon + 2B_{ref}\sqrt{\frac{\log(3\delta^{-1})}{2n}}$, *and* $\mathcal{Q}_1 := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch} - B_{ind}\sqrt{\frac{\log(6\delta^{-1})}{n}}}{b_{orig} + B_{ind}\sqrt{\frac{\log(6\delta^{-1})}{2n}}}$.

Eq 4.3 states that, with high probability, $MCR_+(\epsilon)$ is no higher than $\widehat{MCR}_+(\epsilon_1)$ added to an error term $\mathcal{Q}_1$. As n increases, ϵ_1 approaches ϵ and $\mathcal{Q}_1$ approaches zero. A similar interpretation holds for Eq 4.4.

We provide a visualization of the result in Appendix A.4, and also give a brief sketch of the proof here. First, we enlarge the empirical Rashomon set by increasing ϵ to ϵ_1, such that, by Hoeffding's inequality, $f_{+,\epsilon} \in \hat{\mathcal{R}}(\epsilon_1)$ with high probability. When $f_{+,\epsilon} \in \hat{\mathcal{R}}(\epsilon_1)$, we know that $\widehat{MR}(f_{+,\epsilon}) \leq \widehat{MCR}_+(\epsilon_1)$ by the definition of $\widehat{MCR}_+(\epsilon_1)$. Next, the term $\mathcal{Q}_1$ accounts for estimator error of $\widehat{MR}(f_{+,\epsilon})$ relative to $MR(f_{+,\epsilon}) = MCR_+(\epsilon)$. Thus, we can relate $\widehat{MR}(f_{+,\epsilon})$ to both $\widehat{MCR}_+(\epsilon_1)$ and $MCR_+(\epsilon)$ in order to obtain Eq 4.3. Similar steps can be applied to obtain Eq 4.4.

The bounds in Theorem 6 naturally account for potential overfitting without an explicit limit on model class complexity, such as a covering number (Bousquet et al., 2004; Rudin and Schapire, 2009). Instead, these bounds depend on being able to fully optimize MR across sets in the form of $\hat{\mathcal{R}}(\epsilon)$. If we allow our model class $\mathcal{F}$ to become more flexible, then the size of $\hat{\mathcal{R}}(\epsilon)$ will also increase. Because the bounds in Theorem 6 result from optimizing over $\hat{\mathcal{R}}(\epsilon)$, increasing the size of $\hat{\mathcal{R}}(\epsilon)$ results in wider, more conservative bounds. In this way, Eqs 4.3 and 4.4 implicitly capture model class complexity.

A corollary of Theorem 6 is that we can create a probabilistic bound for the reliance of the (unknown) best-in-class model $f^\star$ on X_1.

Corollary 7. *Let $f^\star \in \arg\min_{f\in\mathcal{F}} e_{orig}(f)$ be a prediction model that attains the lowest possible expected loss, and let $f_{+,\epsilon}$ and $f_{-,\epsilon}$ be defined as in Theorem 6. If $f_{+,\epsilon}$ and $f_{-,\epsilon}$ satisfy Assumptions 3, 4 and 5, then*

$$\mathbb{P}\left(MR(f^\star) \in \left[\widehat{MCR}_-\left(\epsilon_2\right) - \mathcal{Q}_2, \quad \widehat{MCR}_+\left(\epsilon_2\right) + \mathcal{Q}_2\right]\right) \geq 1-\delta,$$

where $\epsilon_2 := 2B_{ref}\sqrt{\frac{\log(6\delta^{-1})}{2n}}$ *, and* $\mathcal{Q}_2 := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch}-B_{ind}\sqrt{\frac{\log(12\delta^{-1})}{n}}}{b_{orig}+B_{ind}\sqrt{\frac{\log(12\delta^{-1})}{2n}}}$.

The above result does not require that $f^\star$ be unique. If several models achieve the minimum possible expected loss, the above boundaries apply simultaneously for each of them. In the special case when the true conditional expectation function $\mathbb{E}(Y|X_1, X_2)$ is equal to $f^\star$, then we have a boundary for the reliance of the function $\mathbb{E}(Y|X_1, X_2)$ on X_1. This reliance bound can also be translated into a causal statement using Theorem 1.

So far, Theorem 6 allows us to cap the range of MR values corresponding to models that predict well, but it does not necessarily provide a complete understanding when this range is large. For example, even if the lower bound on $MCR_-(\epsilon)$ from Eq 4.4 is below 1, we are not able conclude that there exists a well-performing model $f_0 \in \mathcal{R}(\epsilon)$ with no reliance on X_1 (that is, with $MR(f_0) = 1$). To conclude that such a model exists, we need an upper bound on $MCR_-(\epsilon)$. Likewise, in order to make conclusions regarding the existence of well-performing models with high reliance on X_1, we need a lower bound on $MCR_+(\epsilon)$.

To create such bounds, we take a contrary approach to that of Theorem 6, where we had expanded the empirical Rashomon set by increasing ϵ to ϵ_1. We now contract the empirical Rashomon set by subtracting a buffer term from ϵ. This requires that we generalize the definition of an empirical Rashomon set to $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}) := \{f_{\text{ref}}\} \cup \{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \hat{e}_{\text{orig}}(f_{\text{ref}}) + \epsilon\}$ for $\epsilon \in \mathbb{R}$. The definition is unchanged for $\epsilon \geq 0$, but for $\epsilon < 0$ the explicit inclusion of f_{ref} now ensures that $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F})$ is nonempty. As before, we typically omit the notation f_{ref} and $\mathcal{F}$, writing $\hat{\mathcal{R}}(\epsilon)$ instead.

Creating these addition boundaries will also require a limit on the complexity of $\mathcal{F}$. We propose a complexity measure in the form of a covering number. Specifically, we define the set of functions $\mathcal{G}_r$ as an *r-margin-expectation-cover* if for any $f \in \mathcal{F}$ and any distribution D, there exists $g \in \mathcal{G}_r$ such that

$$\mathbb{E}_{Z\sim D}\left|L\left(f, Z\right) - L\left(g, Z\right)\right| \leq r. \tag{4.5}$$

We define the *covering number* $\mathcal{N}(\mathcal{F}, r)$ to be the size of the smallest r-margin-expectation-cover for $\mathcal{F}$. In general, we use $\mathbb{P}_{V\sim D}$ and $\mathbb{E}_{V\sim D}$ to denote probabilities and expectations with respect to a random variable V following the distribution D. We abbreviate these quantities accordingly when V or D are clear from context, for example, as $\mathbb{P}_D$, $\mathbb{P}_V$, or simply $\mathbb{P}$. Unless otherwise stated, all expectations and probabilities are taken with respect to the (unknown) population distribution.

Equipped with this complexity measure, we can now uniformly bound the estimation error of $\widehat{MR}(f)$ for all $f \in \mathcal{F}$. This, in turn, will let us determine the desired bounds for MCR.

Theorem 8. *(Uniform bound for $\widehat{MR}$) Given $r > 0$, if Assumptions 3 and 5 hold for all $f \in \mathcal{F}$, then*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\widehat{MR}(f) - MR(f)\right| > q(\delta, r, n)\right] \leq \delta,$$

where

$$q(\delta, r, n) := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch} - \left\{ B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r\sqrt{2}))}{n}} + 2r\sqrt{2} \right\}}{b_{orig} + \left\{ B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r \right\}}. \tag{4.6}$$

Theorem 8 states that, with high probability, the largest possible estimation error for $MR(f)$ across all models in $\mathcal{F}$ is bounded by $q(\delta, r, n)$, which can be made arbitrarily small by increasing n and decreasing r. This, in turn, means that it is possible to train a model and estimate it's reliance on variables without using sample-splitting. Theorem 8 also lets us derive a lower bound for $MCR_+(\epsilon)$, and upper bound for $MCR_-(\epsilon)$.

Theorem 9. *Given constants $\epsilon \geq 0$ and $r > 0$, if Assumptions 3, 4 and 5 hold for all $f \in \mathcal{F}$, and then*

$$\mathbb{P}\left(MCR_+(\epsilon) < \widehat{MCR}_+(\epsilon_3) - \mathcal{Q}_3 \right) \leq \delta, \text{ and} \tag{4.7}$$

$$\mathbb{P}\left(MCR_-(\epsilon) > \widehat{MCR}_-(\epsilon_3) + \mathcal{Q}_3 \right) \leq \delta, \tag{4.8}$$

where $\epsilon_3 := \epsilon - 2B_{ref}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} - 2r$, and $\mathcal{Q}_3 = q\left(\frac{\delta}{2}, r, n\right)$, as defined in Eq 4.6.

Eq 4.8 states that, with high probability, $MCR_-(\epsilon)$ is no higher than $\widehat{MCR}_-(\epsilon_3)$ added to an error term $\mathcal{Q}_3$, where both $\mathcal{Q}_3$ and $(\epsilon_3 - \epsilon)$ can be made arbitrarily small by shrinking r and increasing n. A similar interpretation holds for the upper bound on $MCR_+(\epsilon)$ in Eq 4.7. We provide a visualization of this result in Appendix A.4.

The proof for Theorem 9 follows a similar structure to that of Theorem 6, but incorporates Theorem 8's uniform bound on MR estimation error ($\mathcal{Q}_3$ term), as well as an additional uniform bound on the probability that any model has in-sample loss too far its expected loss (ϵ_3 term).

In some cases, it may be possible to improve the bounds in Theorem 9 by splitting the sample into two parts, using the first split to intelligently select a subset $\mathcal{F}_s \subset \mathcal{F}$, and using the second split to calculate boundaries. For any subset $\mathcal{F}_s \subset \mathcal{F}$, the value of $MCR_+(\epsilon, f_{\text{ref}}, \mathcal{F}_s)$ will naturally serve as a lower bound on $MCR_+(\epsilon, f_{\text{ref}}, \mathcal{F})$. Thus, rather than using Eq 4.7 to lower bound $MCR_+(\epsilon, f_{\text{ref}}, \mathcal{F})$ directly, we can search for a simpler model class $\mathcal{F}_s$ such that $MCR_+(\epsilon, f_{\text{ref}}, \mathcal{F}_s) \approx MCR_+(\epsilon, f_{\text{ref}}, \mathcal{F})$, and $\mathcal{N}(\mathcal{F}_s, r) \ll \mathcal{N}(\mathcal{F}, r)$. Balancing these two criteria may lead to a tighter bound from Theorem 9. For example, consider the special case where $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^d\}$ is indexed by a real-valued parameter $\boldsymbol{\theta}$. Given a training dataset, let $\boldsymbol{\theta}_{\text{orig}} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{e}_{\text{orig}}(f_{\boldsymbol{\theta}})$ and let $\boldsymbol{\theta}_+ \in \arg\max_{\{\boldsymbol{\theta} \,:\, f_{\boldsymbol{\theta}} \in \hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F})\}} \widehat{MR}(f_{\boldsymbol{\theta}})$, where $\hat{e}_{\text{orig}}(f_{\boldsymbol{\theta}})$ and $\widehat{MR}(f_{\boldsymbol{\theta}})$ are computed using training data. We can then set $\mathcal{F}_s$ as the subset of models resulting from convex combinations of $\boldsymbol{\theta}_{\text{orig}}$ and $\boldsymbol{\theta}_+$:

$$\mathcal{F}_s = \{f_{\boldsymbol{\theta}} \in \mathcal{F} : \boldsymbol{\theta} = w\boldsymbol{\theta}_{\text{orig}} + (1-w)\boldsymbol{\theta}_+;\ w \in [0,1]\}. \tag{4.9}$$

The bound from Eq 4.7 can be evaluated for $\mathcal{F}_s$ in a held-out dataset, potentially resulting in a tighter bound for $MCR_+(\epsilon, f_{\text{ref}}, \mathcal{F})$. In the case of linear classifiers, we can analytically derive a covering number for $\mathcal{F}_s$.

Proposition 10. *Consider the classification setting where $\mathcal{X} = \mathbb{R}^p$; $\mathcal{Y} = \{-1, 1\}$; L is the hinge loss $L(f, (y, x)) = (\delta - yf(x))_+$ for $\delta \in \mathbb{R}$; $\mathcal{F}$ is the set of linear classifiers*

$\{f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(x) = x'\boldsymbol{\theta}\,;\, \boldsymbol{\theta} \in \mathbb{R}^p\}$*; and* $\mathcal{F}_s$ *is defined as in Eq 4.9 for the prespecified vectors* $\boldsymbol{\theta}_{orig}, \boldsymbol{\theta}_+ \in \mathbb{R}^p$*. In this setting, if* $|f_{\boldsymbol{\theta}_+}(x_1, x_2) - f_{\boldsymbol{\theta}_{orig}}(x_1, x_2)| \leq c$ *for all* $(x_1, x_2) \in (\mathcal{X}_1 \times \mathcal{X}_2)$*, then* $\mathcal{N}(\mathcal{F}_s, r) \leq \left\lceil \frac{c}{2r} \right\rceil$ *holds for any* $r \in \mathbb{R}$*.*

Intuitively, Proposition 10 states that if the parameter vectors $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_{\text{orig}}$ produce similar predictions, then there exists a bound on the size of the r-margin-expectation-cover for the model class formed by convex combinations of $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_{\text{orig}}$ (as in Eq 4.9).

4.2 Model class reliance on imputation residuals of correlated predictors

One common scenario where multiple models achieve low loss is when the sets of predictors X_1 and X_2 are highly correlated, or contain redundant information. Models may predict well either through reliance on X_1, or through reliance on X_2, and so MCR will correctly identify a wide range of potential reliances on X_1. However, we may specifically be interested how much models that fully exhaust the information in X_2 can additionally rely on the information in X_1.

For example, age and accumulated wealth may be correlated, and both may be predictive of future promotion. We may wish to know the potential importance of wealth when predicting promotions, but only for models that fully incorporate age.

To achieve this, we propose a two-step procedure:

1. Impute the values of X_1 based on X_2 using a prespecified model $g_{\text{impute}} : \mathcal{X}_2 \to \mathcal{X}_1$. Replace X_1 with its imputation residual $\tilde{X}_1 := X_1 - g_{\text{impute}}(X_2)$.

2. Let $\mathcal{F}$ be a prespecified class of models that use $\tilde{X}_1$ and X_2 to predict Y. After computing $\tilde{X}_1$, estimate the model class reliance of $\mathcal{F}$ on $\tilde{X}_1$.

Such an approach will generally result in a smaller value of $MCR_+(\epsilon)$ relative to the result that would occur from proceeding without an imputation step.

5 General purpose, finite-sample CIs from Rashomon sets

Theorem 6 implies that Rashomon sets can be used to derive non-asymptotic confidence intervals (CIs) for the MR of best-in-class model(s). In this section we generalize the Rashomon set approach to create finite-sample CIs for other summary descriptions of best-in-class model(s). The generalization also helps to illustrate a core aspect of the argument underlying Theorem 6: best-in-class models tend to have relatively good performance in random samples.

Let $\phi : \mathcal{F} \to \mathbb{R}$ be a descriptor of interest for models in $\mathcal{F}$. For example, if $f_{\boldsymbol{\beta}}$ is the linear model $f_{\boldsymbol{\beta}}(x) = x'\boldsymbol{\beta}$, then ϕ may be defined as the norm of the associated coefficient vector (that is, $\phi(f_{\boldsymbol{\beta}}) = ||\boldsymbol{\beta}||_2^2$) or the prediction $f_{\boldsymbol{\beta}}$ would assign given a specific covariate profile x_{new} (that is, $\phi(f_{\boldsymbol{\beta}}) = f_{\boldsymbol{\beta}}(x_{\text{new}})$). For simplicity, we assume that $\phi(f)$ can be determined exactly for any model $f \in \mathcal{F}$. Note that this condition is not satisfied if we choose model reliance as our descriptor ϕ, and so, because of this simplification, the results of this section do not fully replace those of Section 4.1. We also temporarily assume that the best-in-class model $f^\star \in \mathcal{F}$ *uniquely* minimizes the expected loss. In this setting, the following proposition allows us to create finite-sample CIs for $\phi(f^\star)$ based on empirical Rashomon sets.

Proposition 11. *Let* $\hat{a}_-(\epsilon_4) := \min_{f \in \hat{\mathcal{R}}(\epsilon_4)} \phi(f)$ *and* $\hat{a}_+(\epsilon_4) := \max_{f \in \hat{\mathcal{R}}(\epsilon_4)} \phi(f)$*, where* $\epsilon_4 := 2B_{ref}\sqrt{\frac{\log(\delta^{-1})}{2n}}$*. Let* $f^\star \in \arg\min_{f \in \mathcal{F}} e_{orig}(f)$ *be the prediction model that uniquely attains the lowest possible expected loss. If* $f^\star$ *satisfies Assumption 4, then*

$$\mathbb{P}\{\phi(f^\star) \in [\hat{a}_-(\epsilon_4), \hat{a}_+(\epsilon_4)]\} \geq 1 - \delta.$$

Proposition 11 generates the finite-sample CI $[\hat{a}_-(\epsilon_4), \hat{a}_+(\epsilon_4)]$, which can be interpreted as the range of values $\phi(f)$ corresponding to models f with empirical loss not substantially above that of f_{ref}. Thus, the interval has both a rigorous coverage rate and a coherent in-sample interpretation. The proof of Proposition 11 uses Hoeffding's inequality to show that $f^\star$ is contained in $\hat{\mathcal{R}}(\epsilon_4)$ with high probability.

Our assumption that $f^\star$ uniquely minimizes $e_{\text{orig}}(f)$ over $f \in \mathcal{F}$ can be removed if we consider the following version of Proposition 11:

Proposition 12. *Let* $\hat{a}_-(\epsilon_5) := \min_{f \in \hat{\mathcal{R}}(\epsilon_5)} \phi(f)$ *and* $\hat{a}_+(\epsilon_5) := \max_{f \in \hat{\mathcal{R}}(\epsilon_5)} \phi(f)$*, where* $\epsilon_5 := 2B_{ref}\sqrt{\frac{\log(2\delta^{-1})}{2n}}$*. If Assumption 4 holds for all* $f \in \mathcal{R}(0)$*, then*

$$\mathbb{P}\left[\{\phi(f) : f \in \mathcal{R}(0)\} \subset [\hat{a}_-(\epsilon_5), \hat{a}_+(\epsilon_5)]\right] \geq 1 - \delta.$$

In contrast to Proposition 11, Proposition 12 creates a finite-sample CI *for the range* of values $\phi(f)$ corresponding to the models with expected loss no greater than f_{ref}. By definition, this range includes $\phi(f^\star)$ for any $f^\star$ minimizing the expected loss.

As demonstrated in this section, Rashomon sets allow us to reframe a wide set of statistical inference problems as in-sample optimization problems. The implied confidence intervals are not necessarily in closed form, but the approach still opens an exciting pathway for deriving non-asymptotic results. Propositions 11 & 12 also shed light on why certain types of modeling approaches, such as Neural Networks and Random Forests, have been associated with slow progress in inferential theory. For such model classes, it is not even guaranteed that the empirical risk can be globally minimized, much less objective functions in the form of $\phi(f)$, as in $\hat{a}_-$ and $\hat{a}_+$.

6 Limitations of existing variable importance methods

Several common approaches for variable selection, or for describing relationships between variables, do not necessarily capture a variable's importance. Null hypothesis testing methods may identify a relationship, but do not describe the relationship's strength. Similarly, checking whether a variable is included by a sparse model-fitting algorithm, such as the Lasso (Hastie et al., 2009), does not describe the extent to which the variable is relied on. Partial dependence plots (Breiman et al., 2001; Hastie et al., 2009) can be difficult to interpret if multiple variables are of interest, or if the prediction model contains interaction effects.

Another common VI procedure is to run a model-fitting algorithm twice, first on all of the data, and then again after removing X_1 from the dataset. The losses for the two resulting models are then compared to determine the importance, or "necessity," of X_1 (Gevrey et al., 2003). Because this measure is a function of two prediction models rather than one, it does not measure how much either individual model relies on X_1. We refer to this approach as measuring empirical *Algorithm Reliance* (AR) on X_1, as the model-fitting algorithm is the

common attribute between the two models. Related procedures were proposed by (Breiman et al., 2001; Breiman, 2001), which measure the sufficiency of X_1.

The permutation-based VI measure from RFs forms the inspiration for our definition of MR (see Section 3). This RF VI measure has been the topic of empirical studies (Archer and Kimes, 2008; Calle and Urrea, 2010; Wang et al., 2016), and several variations of the measure have been proposed (Strobl et al., 2007, 2008; Altmann et al., 2010; Hapfelmeier et al., 2014). Mentch and Hooker (2016) use U-statistics to study predictions of ensemble models fit to subsamples, similar to the bootstrap aggregation used in RFs. Procedures related to "Mean Difference Impurity," another VI measure derived for RFs, have been studied theoretically by (Louppe et al., 2013; Kazemitabar et al., 2017). All of this literature focuses on VI measures for RFs, for ensembles, or for individual trees. Our estimator for model reliance differs from the traditional RF VI measure (Breiman, 2001) in that we permute inputs to the overall model, rather than permuting the inputs to each individual ensemble member. Thus, our approach can be used generally, and is not limited to trees or ensemble models.

Outside of the context of RF VI, Zhu et al. (2015) propose an estimand similar to our definition of model reliance, and (Gregorutti et al., 2015, 2017) propose an estimand analogous to $e_{\text{switch}}(f) - e_{\text{orig}}(f)$. These recent works focus on model reliance specifically when f is equal to the conditional expectation function of Y (that is, $f(x) = \mathbb{E}[Y|X = x]$). In contrast, we consider model reliance for arbitrary prediction models f. Datta et al. (2016) study the extent to which a model's predictions are expected to change when a subset of variables is permuted, regardless of whether the permutation affects a loss function L. To our knowledge, connections between permutation-based importance and U-statistics have not been previously established.

In discussing the Rashomon effect and VI, Breiman et al. (2001) suggest that ensembling many well-performing models together may alleviate the Rashomon problem. However, this approach may only push the problem from the model level to the ensemble level, as there may be many different ensemble models that fit the data well.

The Rashomon effect has also been considered in several subject areas outside of VI, including those in non-statistical academic disciplines (Heider, 1988; Roth and Mehta, 2002). Tulabandhula and Rudin (2014) optimize a decision rule to perform well under the predicted range of outcomes from any well-performing model. Statnikov et al. (2013) propose an algorithm to discover multiple Markov boundaries, that is, minimal sets of covariates such that conditioning on any one set induces independence between the outcome and the remaining covariates. Nevo and Ritov (2015) report interpretations corresponding to a set of well-fitting, sparse linear models. Letham et al. (2016) search for a pair of well-fitting dynamical systems models that give maximally different predictions.

7 Calculating empirical model class reliance

In this Section, we propose a binary search procedure to bound the values of $\widehat{MCR}_-(\epsilon)$ and $\widehat{MCR}_+(\epsilon)$ (see Eq 4.2), where each step of the search consists of minimizing a linear combination of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$ across $f \in \mathcal{F}$. Our approach is related to the fractional programming approach of Dinkelbach (1967), but accounts for the fact that the problem is constrained by the value of the denominator, $\hat{e}_{\text{orig}}(f)$. We additionally show that, for many model classes, computing $\widehat{MCR}_-(\epsilon)$ only requires that we minimize *convex* combinations of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$, which is no more difficult than minimizing the average loss over an expanded and reweighted sample.

Computing $\widehat{MCR}_+(\epsilon)$ however will require that we are able to minimize arbitrary linear combinations of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$. Thus, in Sections 7.3, 7.4, and 7.5, we show specifically how to minimize arbitrary linear combinations of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$ when $\mathcal{F}$ is the class of linear models, regularized linear models, or linear models in a reproducing kernel Hilbert space (RKHS). Even when the associated objective functions are non-convex, we can tractably obtain global minima for these model classes. We also discuss procedures to determine an upper bound B_{ind} on the loss for any observation when using these model classes (see Assumption 3).

To simplify notation associated with the reference model f_{ref}, we present these computational results in terms of bounds on empirical MR subject to performance thresholds on the *absolute* scale. More specifically, we present bound functions b_- and b_+ satisfying $b_-(\epsilon_{\text{abs}}) \leq \widehat{MR}(f) \leq b_+(\epsilon_{\text{abs}})$ simultaneously for all $\{f, \epsilon_{\text{abs}} : \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}, f \in \mathcal{F}, \epsilon_{\text{abs}} > 0\}$ (Figure 7 shows an example). The binary search procedures we propose can be used to tighten these boundaries at a particular value ϵ_{abs} of interest.

We assume that $0 < \min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f)$, to ensure that MR is finite. In Sections 7.3, 7.4, and 7.5, we additionally assume that $\mathcal{X} \subset \mathbb{R}^p$ for $p \in \mathbb{Z}^+$, we assume that $\mathcal{Y} \subset \mathbb{R}^1$, and we set L equal to the squared error loss function $L(f, (y, x)) = (y - f(x))^2$.

7.1 Binary search for empirical MR lower bound

Before describing our binary search procedure, we introduce additional notation used in this section. Given a constant $\gamma \in \mathbb{R}$ and prediction model $f \in \mathcal{F}$, we define the linear combination $\hat{h}_{-,\gamma}$, and its minimizers (for example, $\hat{g}_{-,\gamma,\mathcal{F}}$), as

$$\hat{h}_{-,\gamma}(f) := \gamma \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f), \quad \text{and} \quad \hat{g}_{-,\gamma,\mathcal{F}} \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \hat{h}_{-,\gamma}(f).$$

We do not require that $\hat{h}_{-,\gamma}$ is uniquely minimized, and we frequently use the abbreviated notation $\hat{g}_{-,\gamma}$ when $\mathcal{F}$ is clear from context.

Our goal in this section is to derive a lower bound on $\widehat{MR}$ for subsets of $\mathcal{F}$ in the form of $\{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}\}$. We achieve this by minimizing a series of linear objective functions in the form of $\hat{h}_{-,\gamma}$, using a similar method to that of Dinkelbach (1967). Often, minimizing the linear combination $\hat{h}_{-,\gamma}(f)$ is more tractable than minimizing the MR ratio directly.

Almost all of the results shown in this section, and those in Section 7.2, also hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ throughout (see Eq 3.2), including in the definition of $\widehat{MR}$ and $\hat{h}_{-,\gamma}(f)$. The exception is Proposition 17, below, which we may still expect to approximately hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$.

Given an observed sample, we define the following condition for a pair of values $\{\gamma, \epsilon_{\text{abs}}\} \in \mathbb{R} \times \mathbb{R}_{>0}$, and argmin function $\hat{g}_{-,\gamma}$:

Condition 13. $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \geq 0$ and $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) \leq \epsilon_{\text{abs}}$.

Using this notation, the following lemma tells conditions on under which we can tractably create a lower bound for empirical MR.

Lemma 14. *(Lower bound for $\widehat{MR}$) If $\gamma \in \mathbb{R}$ satisfies $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \geq 0$, then*

$$\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{abs}} - \gamma \leq \widehat{MR}(f) \tag{7.1}$$

for all $f \in \mathcal{F}$ satisfying $\hat{e}_{orig}(f) \leq \epsilon_{abs}$. It also follows that $-\gamma \leq \widehat{MR}(f)$ for all $f \in \mathcal{F}$.

Additionally, if $f = \hat{g}_{-,\gamma}$ and at least one of the inequalities in Condition 13 holds with equality, then Eq 7.1 holds with equality.

Lemma 14 reduces the challenge of lower-bounding $\widehat{MR}(f)$ to the task of minimizing the linear combination $\hat{h}_{-,\gamma}(f)$. The result of Lemma 14 is not only a single boundary for a particular value of ϵ_{abs}, but a boundary *function* that holds all values of $\epsilon_{\text{abs}} > 0$, with lower values of ϵ_{abs} leading to more restrictive lower bounds on $\widehat{MR}(f)$.

In addition to the formal proof for Lemma 14, we provide a heuristic illustration of the result in Figures 1-3, to aid intuition.

It remains to determine which value of γ should be used in Eq 7.1. The following lemma implies that this value can be determined by a binary search, given a particular value of interest for ϵ_{abs}.

Lemma 15. *(Monotonicity for $\widehat{MR}$ lower bound binary search) The following monotonicity results hold:*

1. *$\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})$ is monotonically increasing in γ.*
2. *$\hat{e}_{orig}(\hat{g}_{-,\gamma})$ is monotonically decreasing in γ.*
3. *Given ϵ_{abs}, the lower bound from Eq 7.1, $\left\{\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{abs}} - \gamma\right\}$, is monotonically decreasing in γ in the range where $\hat{e}_{orig}(\hat{g}_{-,\gamma}) \leq \epsilon_{abs}$, and increasing otherwise.*

Given a particular performance level of interest, ϵ_{abs}, Point 3 of Lemma 15 tells us that the value of γ resulting in the tightest lower bound from Eq 7.1 occurs when γ is as low as possible while still satisfying Condition 13. Points 1 and 2 show that if γ_0 satisfies Condition 13, and one of the equalities in Condition 13 holds with equality, then Condition 13 holds for all $\gamma \geq \gamma_0$. Together, these results imply that we can use a binary search to determine the value of γ to be used in Lemma 14, reducing this value until Condition 13 is no longer met. In addition to the formal proof for Lemma 15, we provide an illustration of the result in Figure 4 to aid intuition.

Next we present a simple condition under which the binary search for values of γ can be restricted to the nonnegative real line. This result substantially extends the computational tractability of our approach, as minimizing $\hat{h}_{-,\gamma}$ for $\gamma \geq 0$ is equivalent to minimizing a reweighted empirical loss:

$$\hat{h}_{-,\gamma}(f) = \gamma \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}, \tag{7.2}$$

where $w_{ij} = \frac{\gamma 1(i=j)}{n} + \frac{1(i\neq j)}{n(n-1)} \geq 0$. Given $\mathcal{F}$ and L, we consider the condition that

Condition 16. For any distribution D satisfying $X_1 \perp_D (X_2, Y)$, there exists a function f_D satisfying

$$\mathbb{E}_D L\{f_D, (Y, X_1, X_2)\} = \min_{f \in \mathcal{F}} \mathbb{E}_D L\{f, (Y, X_1, X_2)\},$$

and

$$f_D(x_1^{(a)}, x_2) = f_D(x_1^{(b)}, x_2) \text{ for any } x_1^{(a)}, x_1^{(b)} \in \mathcal{X}_1 \text{ and } x_2 \in \mathcal{X}_2. \tag{7.3}$$

Intuitively, Condition 16 states that whenever X_1 contains no relevant information, there will exist a best-in-class model $f_D \in \mathcal{F}$ that is not influenced by X_1 in any way.

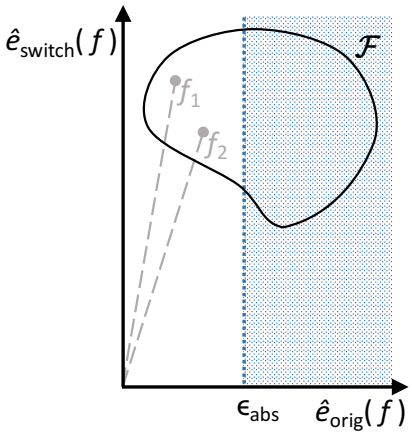

Figure 1: Geometric Representation of $\widehat{MR}$. Above, and in Figures 2 & 3, we illustrate the geometric intuition for Lemma 14. In this figure, we show an example of a hypothetical model class $\mathcal{F}$, marked by the enclosed region. For each model $f \in \mathcal{F}$, the x-axis shows $\hat{e}_{\text{orig}}(f)$ and the y-axis shows $\hat{e}_{\text{switch}}(f)$. Here, we can see that the condition $\min_{f\in\mathcal{F}} \hat{e}_{\text{orig}}(f) > 0$ holds. The blue dotted region marks models with higher empirical loss. We mark two example models within $\mathcal{F}$ as f_1 and f_2. The slopes of the lines connecting the origin to f_1 and f_2 are equal to $\widehat{MR}(f_1)$ and $\widehat{MR}(f_2)$ respectively. Our goal is to lower-bound the slope corresponding to $\widehat{MR}$ for any model f satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$.

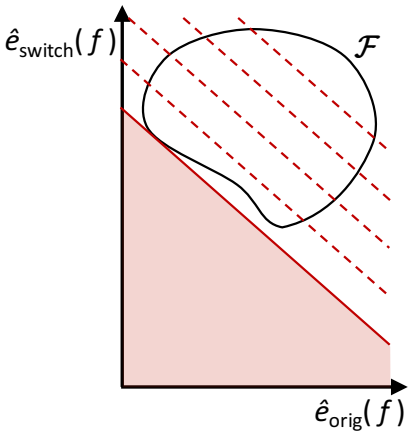

Figure 2: Minimizing Linear Combinations. Consider the linear combination $\hat{h}_{-,\gamma}(f) = \gamma\hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f)$ for $\gamma = 1$. Above, contour lines of $\hat{h}_{-,\gamma}$ are shown in red. The solid red line indicates the smallest possible value of $\hat{h}_{-,\gamma}$ across $f \in \mathcal{F}$. Specifically, its y-intercept equals $\min_{f\in\mathcal{F}} \hat{h}_{-,\gamma}(f)$. If we can determine this minimum, we can determine a linear border constraint on $\mathcal{F}$, that is, we will know that no points corresponding to models $f \in \mathcal{F}$ may lie in the shaded region above. Additionally, if $\min_{f\in\mathcal{F}} \hat{h}_{-,\gamma}(f) \geq 0$ (see Lemma 14), then we know that the origin is either excluded by this linear constraint, or is on the boundary.

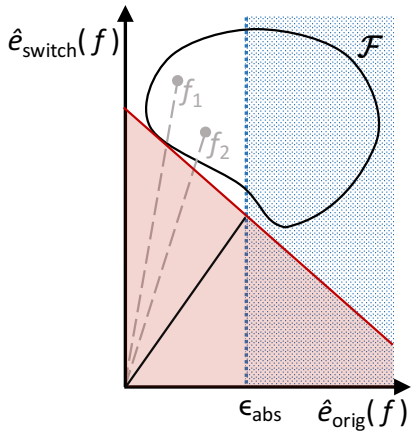

Figure 3: Combining Constraints. Combining the two constraints from Figures 1 & 2, we see that models $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ must correspond to points in the white, unshaded region above. Thus, as long as the unshaded region does not contain the origin, any line connecting the origin to the a model f satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ (for example, here, f_1,f_2) must have a slope at least as high as that of the solid black line above. It can be shown algebraically that the black line has slope equal to the left-hand side of Eq 7.1. Thus the left-hand side of Eq 7.1 is a lower bound for $\widehat{MR}(f)$ for all $\{f \in \mathcal{F} : \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}\}$.

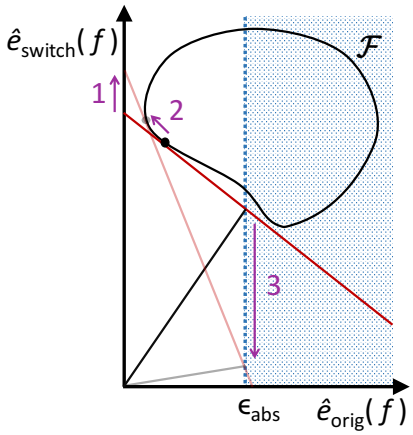

Figure 4: Monotonicity for binary search. Above we show a version of Figure 3 for two alternative values of γ. This figure is meant to add intuition for the monotonicity results in Lemma 15, in addition to the formal proof. Increasing γ is equivalent to *decreasing* the slope of the red line in Figure 3. We define two values $\gamma_1 < \gamma_2$, where γ_1 corresponds to the solid red line, above, and γ_2 corresponds to the semi-transparent red line. The y-intercept values of these lines are equal to $\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1})$ and $\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})$ respectively (see Figure 2 caption). The solid and semi-transparent black dots mark $\hat{g}_{-,\gamma_1}$ and $\hat{g}_{-,\gamma_2}$ respectively. Plugging γ_1 and γ_2 into Eq 7.1 yields two lower bounds for $\widehat{MR}$, marked by the slopes of the solid and semi-transparent black lines respectively (see Figure 3 caption). We see that (1) $\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})$, that (2) $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) \geq \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2})$, and that (3) the left-hand side of Eq 7.1 is decreasing in γ when $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) \leq \epsilon_{\text{abs}}$. These three conclusions are marked by arrows in the above figure, with numbering matching the enumerated list in Lemma 15.

Proposition 17. *(Convexity for $\widehat{MR}$ lower bound binary search) Assume that the loss $L\{f,(Y,X_1,X_2)\}$ depends on the covariates (X_1,X_2) only via the prediction function f, that is, $L\{f,(y,x_1^{(a)},x_2^{(a)})\} = L\{f,(y,x_1^{(b)},x_2^{(b)})\}$ whenever $f(x_1^{(a)},x_2^{(a)}) = f(x_1^{(b)},x_2^{(b)})$. If $\mathcal{F}$ and L satisfy Condition 16, and if $\gamma = 0$, then any function $\hat{g}_{-,\gamma}$ minimizing $\hat{h}_{-,\gamma}$ must also satisfy $\widehat{MR}(\hat{g}_{-,\gamma}) \leq 1$.*

The implication of Proposition 17 is that, when the conditions of Proposition 17 are met, the search region for γ can be limited to the nonnegative real line. To see this, recall that for a fixed value of ϵ_{abs} we can tighten the boundary in Lemma 14 by conducting a binary search for the smallest value of γ that satisfies Condition 13. If $\hat{e}_{\text{orig}}(g_{-,0}) > \epsilon_{\text{abs}}$, then setting γ equal to 0 does not satisfy Condition 13, and the search for γ may ignore the range of values below 0. If $\hat{e}_{\text{orig}}(g_{-,0}) \leq \epsilon_{\text{abs}}$, then we have identified a well-performing model $g_{-,0}$ with empirical MR no greater than 1, by Proposition 17. If $\epsilon_{\text{abs}} = \hat{e}_{\text{orig}}(f_{\text{ref}}) + \epsilon$, this implies that $\widehat{MCR}_-(\epsilon) \leq 1$, which is a sufficiently precise conclusion for most interpretational purposes (see Appendix A.2).

Because of the fixed pairing structure used in $\hat{e}_{\text{divide}}$, Proposition 17 will not necessarily hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ throughout (see Appendix C.3). However, since $\hat{e}_{\text{divide}}$ approximates $\hat{e}_{\text{switch}}$, we can expect Proposition 17 to hold approximately. The bound from Eq 7.1 still remains valid if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ and limit γ to the nonnegative

reals, although in some cases it may not be as tight.

7.2 Binary search for empirical MR upper bound

We now briefly present a binary search procedure to upper bound $\widehat{MR}$, which mirrors the procedure from Section 7.1. Given a constant $\gamma \in \mathbb{R}$ and prediction model $f \in \mathcal{F}$, we define the linear combination $\hat{h}_{+,\gamma}$, and its minimizers (for example, $\hat{g}_{+,\gamma,\mathcal{F}}$), as

$$\hat{h}_{+,\gamma}(f) := \hat{e}_{\text{orig}}(f) + \gamma \hat{e}_{\text{switch}}(f), \quad \text{and} \quad \hat{g}_{+,\gamma,\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \hat{h}_{+,\gamma}(f).$$

As in Section 7.1, $\hat{h}_{+,\gamma}$ need not be uniquely minimized, and we generally abbreviate $\hat{g}_{+,\gamma,\mathcal{F}}$ as $\hat{g}_{+,\gamma}$ when $\mathcal{F}$ is clear from context.

Given an observed sample, we define the following condition for a pair of values $\{\gamma, \epsilon_{\text{abs}}\} \in \mathbb{R}_{\leq 0} \times \mathbb{R}_{>0}$, and argmin function $\hat{g}_{+,\gamma}$:

Condition 18. $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) \geq 0$ and $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) \leq \epsilon_{\text{abs}}$.

We can now develop a procedure to upper bound $\widehat{MR}$, as shown in the next lemma.

Lemma 19. *(Upper bound for $\widehat{MR}$) If $\gamma \in \mathbb{R}$ satisfies $\gamma \leq 0$ and $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) \geq 0$, then*

$$\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{abs}} - 1 \right\} \gamma^{-1} \geq \widehat{MR}(f) \tag{7.4}$$

for all $f \in \mathcal{F}$ satisfying $\hat{e}_{orig}(f) \leq \epsilon_{abs}$. It also follows that $\widehat{MR}(f) \leq |\gamma^{-1}|$ for all $f \in \mathcal{F}$.

Additionally, if $f = \hat{g}_{+,\gamma}$ and at least one of the inequalities in Condition 18 holds with equality, then Eq 7.4 holds with equality.

As in Section 7.1, it remains to determine the value of γ to use in Lemma 19, given a value of interest for $\epsilon_{\text{abs}} \geq \min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f)$. The next lemma tells us that the boundary from Lemma 19 is tightest when γ is as low as possible while still satisfying Condition 18.

Lemma 20. *(Monotonicity for $\widehat{MR}$ upper bound binary search) The following monotonicity results hold:*

1. *$\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})$ is monotonically increasing in γ.*
2. *$\hat{e}_{orig}(\hat{g}_{+,\gamma})$ is monotonically decreasing in γ for $\gamma \leq 0$, and Condition 18 holds for $\gamma = 0$ and $\epsilon_{abs} \geq \min_{f \in \mathcal{F}} \hat{e}_{orig}(f)$.*
3. *Given ϵ_{abs}, the upper boundary $\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{abs}} - 1 \right\} \gamma^{-1}$ is monotonically increasing in γ in the range where $\hat{e}_{orig}(\hat{g}_{+,\gamma}) \leq \epsilon_{abs}$ and $\gamma < 0$, and decreasing in the range where $\hat{e}_{orig}(\hat{g}_{+,\gamma}) > \epsilon_{abs}$ and $\gamma < 0$.*

Together, the results from Lemma 20 imply that we can use a binary search across $\gamma \in \mathbb{R}$ to tighten the boundary on $\widehat{MR}$ from Lemma 19.

7.3 Linear models

In this section we consider MCR computation for the class of linear models

$$\mathcal{F}_{\text{lm}} := \left\{ f_\beta : f_\beta(x) = x'\beta, \quad \beta \in \mathbb{R}^p \right\}.$$

Calculating the boundaries proposed in Sections 7.1 and Section 7.2 requires us to be able to minimize arbitrary linear combinations of $\hat{e}_{\text{orig}}(f_\beta)$ and $\hat{e}_{\text{switch}}(f_\beta)$. For any $f_\beta \in \mathcal{F}_{\text{lm}}$ and $\xi_{\text{orig}}, \xi_{\text{switch}} \in \mathbb{R}$, we can show that the linear combination $\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\beta) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta)$ is proportional in β to the quadratic function $-2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta$, where

$$\mathbf{Q} := \xi_{\text{orig}}\mathbf{X}'\mathbf{X} + \xi_{\text{switch}} \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}, \qquad \mathbf{q} := \left(\xi_{\text{orig}}\mathbf{y}'\mathbf{X} + \xi_{\text{switch}} \begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}' \right)',$$

and $\mathbf{W} := \frac{1}{n-1}(\mathbf{1}_n\mathbf{1}_n' - \mathbf{I}_n)$. Thus, minimizing $\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\beta) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta)$ is equivalent to an unconstrained (possibly non-convex) quadratic program.

To show this, we apply Theorem 2 to see that

$$\begin{aligned}
&\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\beta) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta) \\
&\quad = \frac{\xi_{\text{orig}}}{n} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta) \\
&\quad = \frac{\xi_{\text{orig}}}{n} \left(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \right) \\
&\qquad\qquad + \frac{\xi_{\text{switch}}}{n} \left\{ \mathbf{y}'\mathbf{y} - 2 \begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}' \beta + \beta' \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \beta \right\} \\
&\quad \propto_\beta -2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta. \qquad (7.5)
\end{aligned}$$

This result allows for tractable computation of boundaries for MCR (see Sections 7.1 and 7.2), for the class of unconstrained linear models.

7.4 Regularized linear models

Next, we build on the results of Section 7.3 to calculate boundaries on $\widehat{MR}$ for regularized linear models. We consider the quadratically constrained subset of $\mathcal{F}_{\text{lm}}$, defined as

$$\mathcal{F}_{\text{lm},r_{\text{lm}}} := \left\{ f_\beta : f_\beta(x) = x'\beta, \quad \beta \in \mathbb{R}^p, \quad \beta'\mathbf{M}_{\text{lm}}\beta \le r_{\text{lm}} \right\}, \qquad (7.6)$$

where $\mathbf{M}_{\text{lm}}$ and r_{lm} are pre-specified. Again, this class describes linear models with a quadratic constraint on the coefficient vector.

7.4.1 Calculating MCR

As in Section 7.3, calculating bounds on $\widehat{MR}$ via Lemmas 14 & 19 requires minimizing linear combinations $\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\beta) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\beta)$ across $f_\beta \in \mathcal{F}_{\text{lm},r_{\text{lm}}}$ for arbitrary $\xi_{\text{orig}}, \xi_{\text{switch}} \in \mathbb{R}$. Applying Eq 7.5, we can again equivalently minimize $-2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta$ subject to the constraint in Eq 7.6:

$$\begin{aligned}
\text{minimize} \quad & -2\mathbf{q}'\beta + \beta'\mathbf{Q}\beta \\
\text{subject to} \quad & \beta'\mathbf{M}_{\text{lm}}\beta \le r_{\text{lm}}. \qquad (7.7)
\end{aligned}$$

The resulting optimization problem is a (possibly non-convex) quadratic program with one quadratic constraint (QP1QC). This problem is well studied, and is related to the trust region problem (Boyd and Vandenberghe, 2004; Pólik and Terlaky, 2007; Park and Boyd, 2017). Thus, the bounds on MCR presented in Sections 7.1 and 7.2 again become computationally tractable for the class of quadratically constrained linear models.

7.4.2 Upper bounding the loss

One benefit of constraining the coefficient vector ($\beta' \mathbf{M}_{\text{lm}} \beta \leq r_{\text{lm}}$) is that it facilitates determining an upper bound B_{ind} on the loss function $L(f_\beta, (y, x)) = (y - x'\beta)^2$, which automatically satisfies Assumption 3 for all $f \in \mathcal{F}_{\text{lm},r_{\text{lm}}}$. The following lemma gives sufficient conditions to determine B_{ind}.

Lemma 21. *If $\mathbf{M}_{lm}$ is positive definite, Y is bounded within a known range, and there exists a known constant $r_{\mathcal{X}}$ such that $x'\mathbf{M}_{lm}^{-1}x \leq r_{\mathcal{X}}$ for all $x \in (\mathcal{X}_1 \times \mathcal{X}_2)$, then Assumption 3 holds for the model class $\mathcal{F}_{lm,r_{lm}}$, the squared error loss function, and the constant*

$$B_{ind} = \max\left[\left\{\min_{y\in\mathcal{Y}}(y) - \sqrt{r_{\mathcal{X}} r_{lm}}\right\}^2, \left\{\max_{y\in\mathcal{Y}}(y) + \sqrt{r_{\mathcal{X}} r_{lm}}\right\}^2\right].$$

In practice, the constant $r_{\mathcal{X}}$ can be approximated by the empirical distribution of X and Y. The motivation behind the restriction $x'\mathbf{M}_{\text{lm}}^{-1}x \leq r_{\mathcal{X}}$ in Lemma 21 is to create complementary the constraints on X and β. For example, if $\mathbf{M}_{\text{lm}}$ is diagonal, then the smallest elements of $\mathbf{M}_{\text{lm}}$ correspond to directions along which β is least restricted by $\beta'\mathbf{M}_{\text{lm}}\beta \leq r_{\text{lm}}$ (Eq 7.7), as well as the directions along which x is most restricted by $x'\mathbf{M}_{\text{lm}}^{-1}x \leq r_{\mathcal{X}}$ (Lemma 21).

7.5 Regression in a reproducing kernel Hilbert space (RKHS)

In Sections 7.3 and 7.4 we considered additive linear models. In this section we expand our scope of model classes by considering regression in a reproducing kernel Hilbert space (RKHS). We show how, as in Section 7.4, minimizing a linear combination of $\hat{e}_{\text{orig}}(f)$ and $\hat{e}_{\text{switch}}(f)$ across $f \in \mathcal{F}$ can be expressed as a QP1QC, which allows us to implement the binary search procedure of Sections 7.1 & 7.2.

First we introduce notation required to describe regression in a RKHS. Let $\mathbf{D}$ be a $(R \times p)$ matrix representing a pre-specified dictionary of R reference points, such that each row of $\mathbf{D}$ is contained in $\mathcal{X} = \mathbb{R}^p$. Let k be a pre-specified positive definite kernel function, and let μ be a prespecified estimate of $\mathbb{E}Y$. Let $\mathbf{K}_{\mathbf{D}}$ be the $R \times R$ matrix with $\mathbf{K}_{\mathbf{D}[i,j]} = k(\mathbf{D}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]})$. We consider prediction models of the following form, where the distance to each reference point is used as a regression feature:

$$\mathcal{F}_{\mathbf{D},r_k} = \left\{ f_\alpha : f_\alpha(x) = \mu + \sum_{i=1}^{R} k\left(x, \mathbf{D}_{[i,\cdot]}\right)\alpha_{[i]}, \quad ||f_\alpha||_k \leq r_k, \quad \alpha \in \mathbb{R}^R \right\}. \tag{7.8}$$

Above, the norm $||f_\alpha||_k$ is defined as

$$||f_\alpha||_k := \sum_{i=1}^{R}\sum_{j=1}^{R} \alpha_{[i]}\alpha_{[j]} k(\mathbf{D}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]}) = \alpha' \mathbf{K}_{\mathbf{D}} \alpha. \tag{7.9}$$

We next show that bounds on MCR can again be tractably computed for this class, and that the loss for models in this class can be feasibly upper bounded.

7.5.1 Calculating MCR

Again, calculating bounds on $\widehat{MR}$ from Lemmas 14 & 19 requires us to be able to minimize arbitrary linear combinations of $\hat{e}_{\text{orig}}(f_\alpha)$ and $\hat{e}_{\text{switch}}(f_\alpha)$.

Given a size-n sample of test observations $\mathbf{Z} = \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix}$, let $\mathbf{K}_{\text{orig}}$ be the $n \times R$ matrix with elements $\mathbf{K}_{\text{orig}[i,j]} = k\left(\mathbf{X}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]}\right)$. Let $\mathbf{Z}_{\text{switch}} = \begin{bmatrix} \mathbf{y}_{\text{switch}} & \mathbf{X}_{\text{switch}} \end{bmatrix}$ be the $(n(n-1)) \times (1+p)$ matrix with rows that contain the set $\{(\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]}) : i, j \in \{1, \ldots, n\} \text{ and } i \neq j\}$. Finally, let $\mathbf{K}_{\text{switch}}$ be the $n(n-1) \times R$ matrix with $\mathbf{K}_{\text{switch}[i,j]} = k\left(\mathbf{X}_{\text{switch}[i,\cdot]}, \mathbf{D}_{[j,\cdot]}\right)$.

For any two constants $\xi_{\text{orig}}, \xi_{\text{switch}} \in \mathbb{R}$, we can show that minimizing the linear combination $\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\alpha) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\alpha)$ over $\mathcal{F}_{\mathbf{D},r_k}$ is equivalent to the minimization problem

$$\text{minimize} \quad \frac{\xi_{\text{orig}}}{n}||\mathbf{y} - \mu - \mathbf{K}_{\text{orig}}\alpha||_2^2 + \frac{\xi_{\text{switch}}}{n(n-1)}||\mathbf{y}_{\text{switch}} - \mu - \mathbf{K}_{\text{switch}}\alpha||_2^2 \tag{7.10}$$

$$\text{subject to} \quad \alpha'\mathbf{K}_{\mathbf{D}}\alpha < r_k. \tag{7.11}$$

Like Problem 7.7, Problem 7.10-7.11 is a QP1QC. To show Eqs 7.10-7.11, we first write $\hat{e}_{\text{orig}}(f_\alpha)$ as

$$\begin{aligned}
\hat{e}_{\text{orig}}(f_\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{y}_{[i]} - f_\alpha(\mathbf{X}_{[i,\cdot]})\right\}^2 &(7.12)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{y}_{[i]} - \mu - \sum_{j=1}^{R} k(\mathbf{X}_{[i,\cdot]}, \mathbf{D}_{[j,\cdot]})\alpha_{[j]}\right\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{y}_{[i]} - \mu - \mathbf{K}'_{\text{orig}[i,\cdot]}\alpha\right\}^2 \\
&= \frac{1}{n}||\mathbf{y} - \mu - \mathbf{K}_{\text{orig}}\alpha||_2^2. &(7.13)
\end{aligned}$$

Following similar steps, we can obtain

$$\hat{e}_{\text{switch}}(f_\alpha) = \frac{1}{n(n-1)}||\mathbf{y}_{\text{switch}} - \mu - \mathbf{K}_{\text{switch}}\alpha||_2^2.$$

Thus, for any two constants $\xi_{\text{orig}}, \xi_{\text{switch}} \in \mathbb{R}$, we can see that $\xi_{\text{orig}}\hat{e}_{\text{orig}}(f_\alpha) + \xi_{\text{switch}}\hat{e}_{\text{switch}}(f_\alpha)$ is quadratic in α. This means that we can tractably compute bounds on MCR for this class as well.

7.5.2 Upper bounding the loss

Using similar steps as in Section 7.4.2, the following lemma gives sufficient conditions to determine B_{ind} for the case of regression in a RKHS.

Lemma 22. *Assume that Y is bounded within a known range, and there exists a known constant $r_{\mathbf{D}}$ such that $v(x)'\mathbf{K}_{\mathbf{D}}^{-1}v(x) \leq r_{\mathbf{D}}$ for all $x \in (\mathcal{X}_1 \times \mathcal{X}_2)$, where $v : \mathbb{R}^p \to \mathbb{R}^R$ is the*

function satisfying $v(x)_{[i]} = k(x, \mathbf{D}_{[i,\cdot]})$. *Under these conditions, Assumption 3 holds for the model class* $\mathcal{F}_{\mathbf{D},r_k}$, *the squared error loss function, and the constant*

$$B_{ind} = \max\left[\left\{\min_{y\in\mathcal{Y}}(y) - (\mu + \sqrt{r_{\mathbf{D}} r_k})\right\}^2, \left\{\max_{y\in\mathcal{Y}}(y) + (\mu + \sqrt{r_{\mathbf{D}} r_k})\right\}^2\right].$$

We see that for regression models in a RKHS, we can satisfy Assumption 3 for all models in the class.

8 Simulations

In this section, we first present a toy example to illustrate the concepts of MR, MCR, and AR. We then present a Monte Carlo simulation studying the effectiveness of bootstrap confidence intervals for MCR.

8.1 Illustrative toy example with simulated data

To illustrate the concepts of MR, MCR, and AR (see Section 6), we consider a toy example where $X = (X_1, X_2) \in \mathbb{R}^2$, and $Y \in \{-1, 1\}$ is a binary group label. Our primary goal in this section is to build intuition for the differences between these three importance measures, and so we demonstrate them here only in a single sample. We focus on the empirical versions of our importance metrics ($\widehat{MR}$, $\widehat{MCR}_-$ and $\widehat{MCR}_+$), and compare them against AR, which is typically interpreted as an in-sample measure (Breiman, 2001), or as an intermediate step to estimate an alternate importance measure in terms of variable rankings (Gevrey et al., 2003; Olden et al., 2004).

We simulate $X|Y = -1$ from an independent, bivariate normal distribution with means $\mathbb{E}(X_1|Y = -1) = \mathbb{E}(X_2|Y = -1) = 0$ and variances $\text{Var}(X_1|Y = -1) = \text{Var}(X_2|Y = -1) = \frac{1}{9}$. We simulate $X|Y = 1$ by drawing from the same bivariate normal distribution, and then adding the value of a random vector $(C_1, C_2) := (\cos(U), \sin(U))$, where U is a random variable uniformly distributed on the interval $[-\pi, \pi]$. Thus, (C_1, C_2) is uniformly distributed across the unit circle.

Given a prediction model $f : \mathcal{X} \to \mathbb{R}$, we use the sign of $f(X_1, X_2)$ as our prediction of Y. For our loss function, we use the hinge loss $L(f, (x, y)) = (1 - yf(x))_+$, where $(a)_+ = a$ if $a \geq 0$ and $(a)_+ = 0$ otherwise. The hinge loss function is commonly used as a convex approximation to the zero-one loss $L(f, (x, y)) = 1[y = \text{sign}\{f(x)\}]$.

We simulate two samples of size 300 from the data generating process described above, one to be used for training, and one to be used for testing. Then, for the class of models used to predict Y, we consider the set of degree-3 polynomial classifiers

$$\begin{aligned}\mathcal{F}_{d3} = \big\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(x_1, x_2) &= \boldsymbol{\theta}_{[1]} + \boldsymbol{\theta}_{[2]} x_1 + \boldsymbol{\theta}_{[3]} x_2 \\ &+ \boldsymbol{\theta}_{[4]} x_1^2 + \boldsymbol{\theta}_{[5]} x_2^2 + \boldsymbol{\theta}_{[6]} x_1 x_2 \\ &+ \boldsymbol{\theta}_{[7]} x_1^3 + + \boldsymbol{\theta}_{[8]} x_2^3 + \boldsymbol{\theta}_{[9]} x_1^2 x_2 + \boldsymbol{\theta}_{[10]} x_1 x_2^2;\ \ ||\boldsymbol{\theta}_{[-1]}||_2^2 \leq r_{d3} \big\},\end{aligned}$$

where we set r_{d3} to the value that minimizes the 10-fold cross-validated loss in the training data. Let $\mathcal{A}_{d3}$ be the algorithm that minimizes the hinge loss over the (convex) feasible region $\{f_\theta : ||\boldsymbol{\theta}_{[-1]}||_2^2 \leq r_{d3}\}$. We apply $\mathcal{A}_{d3}$ to the training data to determine a reference model f_{ref}. Also using the training data, we set ϵ equal to 0.10 multiplied by the cross-validated loss of $\mathcal{A}_{d3}$, such that $\mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ contains all models in $\mathcal{F}_{d3}$ that exceed the

loss of f_{ref} by no more than approximately 10%. We then calculate empirical AR, MR, and MCR using the test observations.

We begin by considering the AR of $\mathcal{A}_{d3}$ on X_1. Calculating AR requires us to fit two separate models, first using all of the variables to fit a model on the training data, and then again using only X_2. In this case, the first model is equivalent to f_{ref}. We denote the second model as $\hat{f}_2$. To compute AR, we evaluate f_{ref} and $\hat{f}_2$ in the test observations. We illustrate this AR computation in Panel 1 of Figure 5, marking the classification boundaries for f_{ref} and $\hat{f}_2$ by the black dotted line and the blue dashed lines respectively, and marking the test observations by labelled points ("x" for $Y = 1$, and "o" for $Y = -1$). Comparing the loss associated with these two models gives one form of AR–an estimate of the necessity of X_1 for the algorithm $\mathcal{A}_{d3}$. Alternatively, to estimate the *sufficiency* of X_1, we can compare the reference model f_{ref} against the model resulting from retraining algorithm $\mathcal{A}_{d3}$ only using X_1. We refer to this third model as $\hat{f}_1$, and mark its classification boundary by the solid blue lines in Figure 5.

Each of the classifiers in Panel 1 of Figure 5 can also be evaluated for its reliance on X_1, as shown in Panel 3 of Figure 5. Here, we use $\hat{e}_{\text{divide}}$ in our calculation of $\widehat{MR}$ (see Eq 3.2). Unsurprisingly, the classifier fit without using X_1 (blue dashed line) has a model reliance of $\widehat{MR}(\hat{f}_2) = 1$. The reference model f_{ref} (dotted black line) has a model reliance of $\widehat{MR}(f_{\text{ref}}) = 3.47$. These values of $\widehat{MR}$ each have an interpretation contained to a single model–they compare the model's behavior under different data distributions, rather than the AR approach of comparing different models' behavior on the same data distribution.

We illustrate MCR in Panel 2 of Figure 5. In contrast to AR, MCR is only ever a function of well-performing prediction models. Here, we consider the empirical Rashomon set $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$, the subset of models in $\mathcal{F}_{d3}$ with test loss no more than ϵ above that of f_{ref}. We show the classification boundary associated with 15 well-performing models contained in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ by the gray solid lines. We also show two of the models in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ that approximately maximize and minimize empirical reliance on X_1 among models in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$. We denote these models as $\hat{f}_{+,\epsilon}$ and $\hat{f}_{-,\epsilon}$, and mark them by the solid green and dashed green lines respectively. For every model shown in Panel 2, we also mark its model reliance in Panel 3. We can then see from Panel 3 that $\widehat{MR}$ for each model in $\hat{\mathcal{R}}(\epsilon, f_{\text{ref}}, \mathcal{F}_{d3})$ is contained between $\widehat{MR}(\hat{f}_{-,\epsilon})$ and $\widehat{MR}(\hat{f}_{+,\epsilon})$, up to a small approximation error.

In summary, unlike AR, MCR is only a function of models that fit the data well.

8.2 Simulations of bootstrap confidence intervals

In our data analysis of recidivism models, our goal will be to estimate how much the proprietary model COMPAS relies on proxies for race and sex. Unfortunately, because COMPAS is unavailable, estimation of its attributes cannot be evaluated in simulations. In this section we instead study the related problem model class misspecification, where the goal is to estimate how much the conditional expectation function $f_0(x) = \mathbb{E}(Y|X = x)$ relies on subsets of covariates. Given a reference model f_{ref} and model class $\mathcal{F}$, our ability to describe $MR(f_0)$ will hinge on two conditions:

Condition 23. The class $\mathcal{F}$ contains a well-performing model $\tilde{f} \in \mathcal{R}(\epsilon, f_{\text{ref}}, \mathcal{F})$ satisfying $MR(\tilde{f}) = MR(f_0)$.

Condition 24. Bootstrap CIs for empirical MCR give appropriate coverage of population-level MCR.

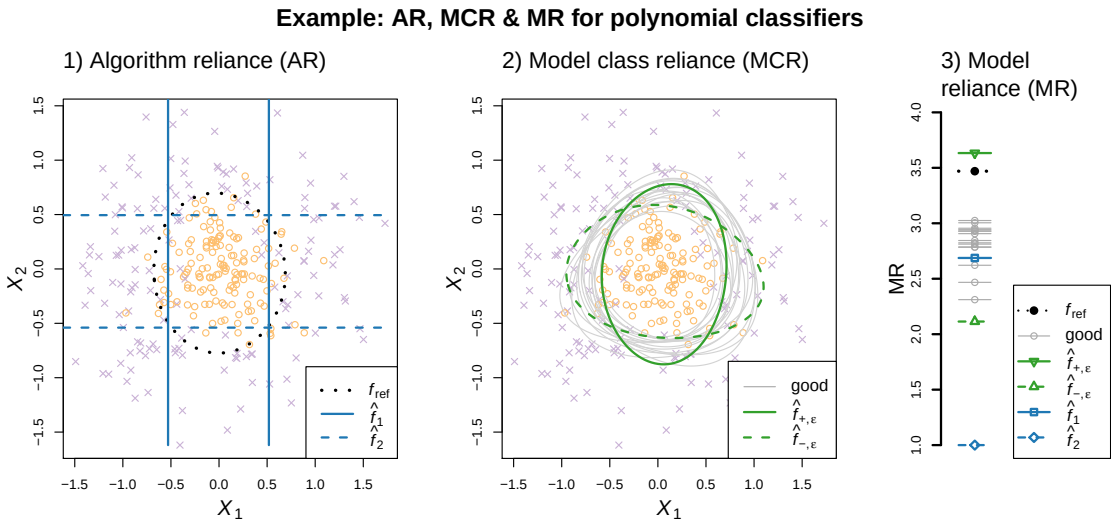

Figure 5: Example of AR, MCR & MR for polynomial classifiers. Panels 1 & 2 show the same 300 draws from a simulated dataset, with the classification of each data point marked by "x" for $Y = 1$, and "o" for $Y = -1$. Panel 3 shows the empirical model reliance on X_1 for each of the models in Panels 1 & 2.

Condition 23 ensures that the interval $[MCR_-(\epsilon), MCR_+(\epsilon)]$ contains $MR(f_0)$, and Condition 24 ensures that this interval can be estimated in finite samples. Condition 23 can also be interpreted as saying that the model reliance value of $MR(f_c)$ is "well supported" by the class $\mathcal{F}$, even if $\mathcal{F}$ does not contain f_0. Our primary goal is to assess whether confidence intervals derived from MCR can give appropriate coverage of $MR(f_0)$, which depends on both conditions. As a secondary goal, we also would like to be able to assess Conditions 23 & 24 individually.

Verifying the above conditions requires that we are able to calculate population-level MCR. To this end, we draw samples with replacement from a finite population of 20,000 observations, in which MCR can also be calculated directly. To derive a CI based on MCR, we divide each simulated sample $\mathcal{Z}_s$ into a training subset and analysis subset. We use the training subset to fit a reference model $f_{\text{ref},s}$, which is required for our definition of population-level MCR. We calculate a bootstrap CI by drawing 500 bootstrap samples from the analysis subset, and computing $\widehat{MCR}_-(\epsilon, f_{\text{ref},s})$ and $\widehat{MCR}_+(\epsilon, f_{\text{ref},s})$ in each bootstrap sample. We then take the 2.5% percentile of $\widehat{MCR}_-(\epsilon, f_{\text{ref},s})$ across bootstrap samples, and the 97.5% percentile of $\widehat{MCR}_+(\epsilon, f_{\text{ref},s})$ across bootstrap samples, as the lower and upper endpoints of our CI, respectively. We repeat this procedure for both X_1 and X_2.

We generate data according to a model with increasing amounts of nonlinearity. For $\gamma \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, we simulate continuous outcomes as $Y = f_0(X) + E$, where f_0 is the function $f_0(\mathbf{x}) = \sum_{j=1}^{p} j\mathbf{x}_{[j]} - \gamma\mathbf{x}_{[j]}^2$; the covariate dimension p is equal to 2, with X_1 and X_2 defined as the first and second elements of X; the covariates X are drawn from a multivariate normal distribution with $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$, $\text{Var}(X_1) = \text{Var}(X_2) = 0$, and $\text{Cov}(X_1, X_2) = 1/4$; and E is a normally distributed noise variable with mean zero and variance equal to $\sigma_E^2 := \text{Var}(f_0(X))$. We consider sample sizes of $n = 400$ and 800, of which $n_{tr} = 200$ or 300 observations are assigned to the training subset respectively.

To implement our approach, we use the model class $\mathcal{F}_{\text{lm}} = \{f_{\boldsymbol{\beta}} : f_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}_{[1]} +$

$\sum_{j=1}^{2} \mathbf{x}_{[j]}\boldsymbol{\beta}_{[j+1]}, \boldsymbol{\beta} \in \mathbb{R}^3\}$. We set the performance threshold ϵ equal to $0.1 \times \sigma_E^2$. We refer to this MCR implementation with $\mathcal{F}_{\text{lm}}$ as "MCR-Linear."

As a comparator method, we consider a simpler bootstrap approach, which we refer to as "Standard-Linear." Here, we take 500 bootstrap samples from the simulated data $\mathcal{Z}_s$. In each bootstrap sample, indexed by b, we set aside n_{tr} training points to train a model $f_b \in \mathcal{F}_{\text{lm}}$, and calculate $\widehat{MR}(f_b)$ from the remaining data points. We then create a 95% bootstrap percentile CI for $MR(f_0)$ by taking the 2.5% and 97.5% percentiles of $\widehat{MR}(f_b)$ across $b = 1, \ldots, 500$.

8.2.1 Results

Overall, we find that MCR provides more robust and conservative intervals for the reliance of f_0 on X_1 and X_2, relative to standard bootstrap approaches. We also find that higher sample size generally exacerbates coverage errors due to misspecification, as methods become more certain of biased results.

MCR-Linear gave proper coverage for up to moderate levels of misspecification ($\gamma = 0.3$), where Standard-Linear began to break down (Figure 6). For larger levels of misspecification ($\gamma \geq 0.4$), both MCR-Linear and Standard-Linear failed to give appropriate coverage.

The increased robustness of MCR comes at the cost of larger confidence intervals. Intervals for MCR-Linear were typically larger than intervals for Standard-Linear by a factor of approximately 2-4. This is partly due to the fact that CIs for MCR are meant to cover the range of values $[MCR_-(\epsilon, f_{\text{ref},s}), MCR_+(\epsilon, f_{\text{ref},s})]$, rather than to cover a single point.

When investigating Conditions 23 & 24 individually, we find that the coverage errors for MCR-Linear were largely attributable to violations of Condition 23. Condition 24 appears to hold conservatively for all scenarios studied–within each scenario, at least 95.9% of bootstrap CIs contained population-level MCR.

9 Data analysis: reliance of criminal recidivism prediction models on race and sex

Evidence suggests that bias exists among judges and prosecutors in the criminal justice system (Spohn, 2000; Blair et al., 2004; Paternoster and Brame, 2008). In an aim to counter this bias, machine learning models trained to predict recidivism are increasingly being used to inform judges' decisions on pretrial release, sentencing, and parole (Monahan and Skeem, 2016; Picard-Fritsche et al., 2017). Ideally, prediction models can avoid human bias and provide judges with empirically tested tools. But prediction models can also mirror the biases of the society that generates their training data, and perpetuate the same bias at scale. In the case of recidivism, if arrest rates across demographic groups are not representative of underlying crime rate (Beckett et al., 2006; Ramchand et al., 2006; U.S. Department of Justice - Civil Rights Devision, 2016), then bias can be created in both (1) the outcome variable, future crime, which is measured imperfectly via arrests or convictions, and (2) the covariates, which include the number of prior convictions on a defendant's record (Corbett-Davies et al., 2016; Lum and Isaac, 2016). Further, when a prediction model's behavior and mechanisms are an opaque black box, the model can evade scrutiny, and fail to offer recourse or explanations to individuals rated as "high risk."

We focus here on the issue of transparency, which takes an important role in the recent debate about the proprietary recidivism prediction tool COMPAS (Larson et al., 2016;

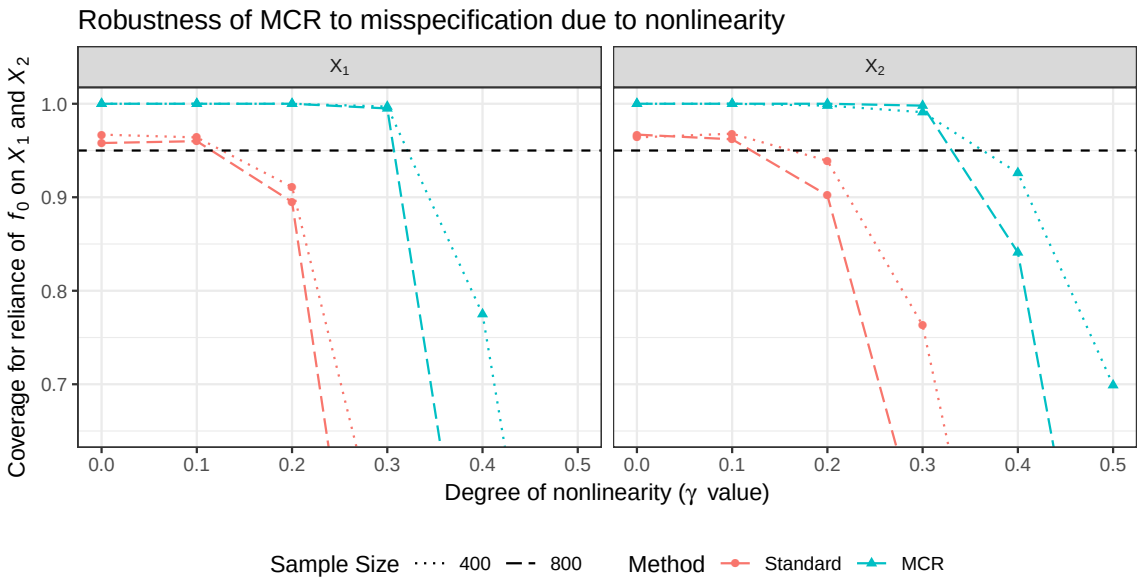

Figure 6: MR Coverage - The y-axis shows coverage rate for the reliance of f_0 on either X_1 (left column) or X_2 (right column), where X_2 is simulated to be more influential that X_1. The x-axis shows increasing levels of misspecification (γ). All methods aim to have at least 95% coverage for each scenario (dashed horizontal line).

Corbett-Davies et al., 2016). While COMPAS is known to not rely explicitly on race, there is concern that it may rely implicitly on race via proxies–variables statistically dependent with race (see further discussion in Section 9.2).

Our goal is to identify bounds for how much COMPAS relies on different covariate subsets, either implicitly or explicitly. We analyze a public dataset of defendants from Broward County, Florida, in which COMPAS scores have been recorded (Larson et al., 2016). Within this dataset, we only included defendants measured as African-American or Caucasian (3,373 in total) due to sparseness in the remaining categories. The outcome of interest (Y) is the COMPAS violent recidivism score. Of the available covariates, we consider three variables which we refer to as "admissible": an individual's age, their number of priors, and an indicator of whether the current charge is a felony. We also consider two variables which we refer to as "inadmissible": an individual's race and sex. Our labels of "admissible" and "inadmissible" are not intended to be legally precise–indeed, the boundary between these types of labels is not always clear (see Section 9.2). We compute empirical MCR and AR for each variable group, as well as bootstrap CIs for MCR (see Section 8.2).

To compute empirical MCR and AR, we consider a flexible class of linear models in a RKHS to predict the COMPAS score (described in more detail below). Given this class, the MCR range (See Eq 4.1) captures the highest and lowest degree to which any model in the class may rely on each covariate subset. We assume that our class contains at least one model that relies on "inadmissible variables" to the same extent that COMPAS relies either on "inadmissible variables" or on proxies that are unmeasured in our sample (analogous to Condition 23). We make the same assumption for "admissible variables." These assumptions can be interpreted as saying that the reliance values of COMPAS are relatively "well supported" by our chosen model class, and allows us to identify bounds on the MR

values for COMPAS. We also consider the more conventional, but less robust approach of AR (Section 6), that is, how much would the accuracy suffer for a model-fitting algorithm trained on COMPAS score if a variable subset was removed?

These computations require that we predefine our loss function, model class, and performance threshold. We define MR, MCR, and AR in terms of the squared error loss $L(f,(y,x)) = \{y - f(x)\}^2$. We define our model class $\mathcal{F}_{\mathbf{D},r_k}$ in the form of Eq 7.8, where we determine $\mathbf{D}$, μ, k, and r_k based on a subset $\mathcal{S}$ of 500 training observations. We set $\mathbf{D}$ equal to the matrix of covariates from $\mathcal{S}$; we set μ equal to the mean of Y in $\mathcal{S}$; we set k equal to the radial basis function $k_{\sigma_s}(\mathbf{x},\tilde{\mathbf{x}}) = \exp\left(-\frac{||\mathbf{x}-\tilde{\mathbf{x}}||^2}{2\sigma_s}\right)$, where we choose σ_s to minimize the cross-validated loss of a Nadaraya-Watson kernel regression (Hastie et al., 2009) fit to $\mathcal{S}$; and we select the parameters r_k by cross-validation on $\mathcal{S}$. We set ϵ equal to 0.1 times the cross-validated loss on $\mathcal{S}$. Also using $\mathcal{S}$, we train a reference model $f_{\text{ref}} \in \mathcal{F}_{\mathbf{D},r_k}$. Using the held-out 2,873 observations, we then estimate $MR(f_{\text{ref}})$ and MCR for $\mathcal{F}_{\mathbf{D},r_k}$. To calculate AR, we train models from $\mathcal{F}_{\mathbf{D},r_k}$ using $\mathcal{S}$, and evaluate their performance in the held-out observations.

9.1 Results

Our results imply that race and sex play somewhere between a null role and a modest role in determining COMPAS score, but that they are less important than "admissible" factors (Figure 7). As a benchmark for comparison, the empirical MR of f_{ref} is equal to 1.09 for "inadmissible variables," and 2.78 for "admissible variables." The AR is equal to 0.94 and 1.87 for "inadmissible" and "admissible" variables respectively, roughly in agreement with MR. The MCR range for "inadmissible variables" is equal to [1.00,1.56], indicating that for any model in $\mathcal{F}_{\mathbf{D},r_k}$ with empirical loss no more than ϵ above that of f_{ref}, the model's loss can increase by no more than 56% if race and sex are permuted. Such a statement cannot be made solely based on AR or MR methods, as these methods do not upper bound the reliance values of well-performing models. This MCR interval is truncated at 1, as it is often sufficiently precise to conclude that there exists a well-performing model with no reliance on the variables of interest (that is, MR equal to 1; see Appendix A.2). The bootstrap 95% CI for MCR on "inadmissible variables" is [1.00, 1.73]. Thus, under our assumptions, if COMPAS relied on sex, race, or their unmeasured proxies by a factor greater than 1.73, then intervals as low as what we observe would occur with probability < 0.05.

For "admissible variables" the MCR range is equal to [1.77,3.61], with a 95% bootstrap CI of [1.62, 3.96]. However, it is worth noting that the upper limit of 3.61 maximizes empirical MR on "admissible variables" not only among well-performing models, but globally across all models in the class (see Figure 7, and Lemma 19). Because the regularization constraints of $\mathcal{F}_{\mathbf{D},r_k}$ preclude higher levels of reliance, the MR of COMPAS on "admissible variables" may be underestimated by empirical MCR.

9.2 Discussion & limitations

Asking whether a proprietary model relies on sex and race, after adjusting for other covariates, is related to the fairness metric known as conditional statistical parity (CSP). A decision rule satisfies CSP if its decisions are independent of a sensitive variable, conditional on a set of "legitimate" covariates C (Corbett-Davies et al., 2017; see also Kamiran et al., 2013). Roughly speaking, CSP reflects the idea that "similar individuals are treated similarly" (Dwork et al., 2012), regardless of the sensitive variable (for example, race or sex).

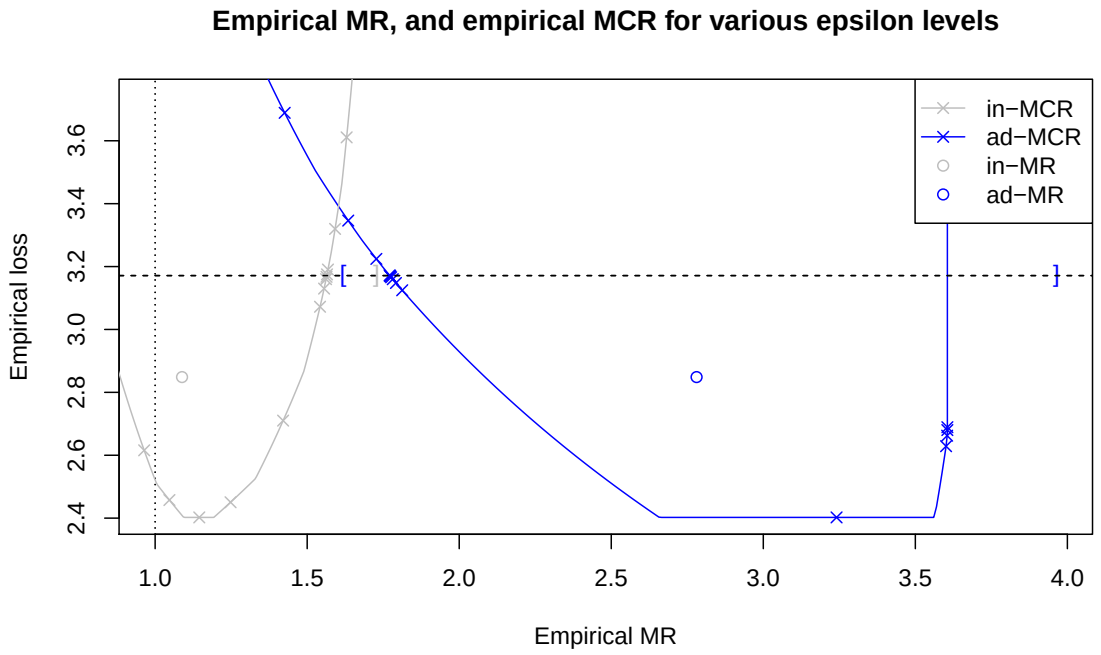

Figure 7: Empirical MR and MCR for Broward County criminal records dataset - For any prediction model f, the y-axis shows empirical loss ($\hat{e}_{\text{stnd}}(f)$) and the x-axis shows empirical reliance ($\widehat{MR}(f)$) on each covariate subset. Null reliance (MR equal to 1.0) is marked by the vertical dotted line. Reliances on different covariate subsets are marked by color ("admissible" = blue; "inadmissible" = gray). For example, model reliance values for f_{ref} are shown by the two circular points, one for "admissible" variables and one for "inadmissible" variables. MCR for different values of ϵ can be represented as boundaries on this coordinate space. To this end, for each covariate subset, we compute conservative boundary functions (shown as solid lines, or "bowls") guaranteed to contain *all models in the class* (see Section 7). Specifically, all models in $f \in \mathcal{F}_{\mathbf{D},r_k}$ are guaranteed to have an empirical loss ($\hat{e}_{\text{stnd}}(f)$) and empirical MR value ($\widehat{MR}(f)$) for "inadmissible variables" corresponding to a point within the gray bowl. Likewise, all models in $\mathcal{F}_{\mathbf{D},r_k}$ are guaranteed to have an empirical loss and empirical MR value for "admissible variables" corresponding to a point within the blue bowl. Points shown as "$\times$" represent additional models in $\mathcal{F}_{\mathbf{D},r_k}$ discovered during our computational procedure, and thus show where the "bowl" boundary is tight. The goal of our computation procedure (see Section 7) is to tighten the boundary as much as possible near the ϵ value of interest, shown by the dashed horizontal line above. This dashed line has a y-intercept equal to the loss of the reference model plus the ϵ value of interest. Bootstrap CIs for $MCR_{-}(\epsilon)$ and $MCR_{+}(\epsilon)$ are marked by brackets.

However, the criteria becomes superficial if too many variables are included in C, and care should be taken to avoid including proxies for the sensitive variables. Several other fairness metrics have also been proposed, which often form competing objectives (Kleinberg et al., 2016; Chouldechova, 2017; Nabi and Shpitser, 2018; Corbett-Davies et al., 2017). Here, if COMPAS was not influenced by race, sex, or variables related to race or sex (conditional on a set of "legitimate" variables), it would satisfy CSP.

Unfortunately, it is often difficult to distinguish between "legitimate" (or "admissible") variables and "illegitimate" variables. Some variables function both as part of a reasonable predictor for risk, and, separately, as a proxy for race. Because of disproportional arrest rates, particularly for misdemeanors and drug-related offenses (U.S. Department of Justice - Civil Rights Devision, 2016; Lum and Isaac, 2016), prior misdemeanor convictions may act as such a proxy (Corbett-Davies et al., 2016; Lum and Isaac, 2016).

Proxy variables for race (defined as being statistically dependent with race) that are unmeasured in our sample are also not the only reason that race could be predictive of COMPAS score. Other inputs to the COMPAS algorithm might be associated with race *only conditionally* on variables we categorize as "admissible." However, our result from Section 9.1 that race has limited predictive utility for COMPAS score suggests that such conditional relationships are also limited.

Conclusion

In this article, we propose MCR as the upper and lower limit on how important a set of variables can be to any well-performing model in a class. In this way, MCR provides a more comprehensive and robust measure of importance than traditional importance measures for a single model. We derive bounds on MCR, which motivate our choice of point estimates. We also derive connections between permutation importance, U-statistics, and conditional causal effects. We apply MCR in a dataset of criminal recidivism, in order to help inform the characteristics of the proprietary model COMPAS.

Several exciting areas remain open for future research. One research direction closely related to our current work is the development of MCR computation procedures for other model classes and loss functions. We have shown that, in model classes for which minimizing the standard loss is a convex optimization problem, computing $\widehat{MCR}_-$ can often be reduced to a series of convex optimization problems. General computation procedures for $\widehat{MCR}_+$ are still an open research area, and may depend on methods for difference in convex (DC) programming problems. Additional computational challenges exist in calculating MCR for model classes where globally minimizing even the standard loss is not tractable, as in neural networks.

Another direction is to consider MCR for variable selection. If MCR_+ is small for a variable, then no well-performing predictive model can heavily depend on that variable, indicating that it can be eliminated.

While we develop Rashomon sets with the goal of studying MR, Rashomon sets can also be useful for finite sample inferences about a wide variety of other attributes of best-in-class models (for example, Section 5). Characterizations of the Rashomon set itself may also be of interest. For example, in ongoing work, we are studying the size of the Rashomon set, and its connection to generalization of models and model classes.

Acknowledgements

Support for this work was provided by the National Institutes of Health (grants P01CA134294, R01GM111339, R01ES024332, R35CA197449, R01ES026217, P50MD010428, DP2MD012722, R01MD012769, & R01ES028033), by the Environmental Protection Agency (grants 83615601 & 83587201-0), and by the Health Effects Institute (grant 4953-RFA14-3/16-4).

Appendices

All labels for items in the following appendices begin with a letter (for example, Section A.1), while references to items in the main text contain only numbers (for example, Theorem 1).

A Miscellaneous supplemental sections

A.1 Notation examples: mean models & step function models

For simplicity, in this section we assume that $\mathcal{X} = \mathcal{Y} = \mathbb{R}^1$. As a first example, let $\mathcal{F}_{\text{intercept}} = \{f_\theta : f_\theta(x) = \theta; \quad \theta \in \mathbb{R}^1\}$ be the class of intercept-only regression models, or "mean-models." A standard algorithm to select a model from data is the simple procedure of setting θ equal to the sample mean:

$$\mathcal{A}_{\text{sample-mean}}(\mathbf{Z}) = f_\theta, \text{ where } \theta = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_{[i]}.$$

A second example of a model class is the set of all piecewise constant functions with a single discontinuity:

$$\mathcal{F}_{\text{piecewise}} = \{f_\theta : f_\theta(x) = \theta_0 + 1(x > m)\theta_1; \quad \boldsymbol{\theta} = (\theta_0, \theta_1, m) \in \mathbb{R}^3\}.$$

Given a dataset $\mathbf{Z} = [\mathbf{y}\ \mathbf{X}]$, one naive algorithm $\mathcal{A}_{\text{median}}$ to select a prediction model from $\mathcal{F}_{\text{piecewise}}$ is the following procedure:

1. Set $\hat{m}$ equal to the median of the $n \times 1$ matrix $\mathbf{X}$.
2. Set $\hat{\theta}_0 := \frac{\sum_{i=1}^{n} \mathbf{y}_{[i]} 1(\mathbf{X}_{[i,\cdot]} \leq m)}{\sum_{i=1}^{n} 1(\mathbf{X}_{[i,\cdot]} \leq m)}$, and $\hat{\theta}_1 := \frac{\sum_{i=1}^{n} \mathbf{y}_{[i]} 1(\mathbf{X}_{[i,\cdot]} > m)}{\sum_{i=1}^{n} 1(\mathbf{X}_{[i,\cdot]} > m)}$, where $1(E)$ is an indicator of the event E.
3. Return $f_{\hat{\theta}}$, where $\hat{\boldsymbol{\theta}} := (\hat{\theta}_0, \hat{\theta}_1, \hat{m})$.

A.2 Model reliance less than 1

While it is counterintuitive, it is possible for the expected loss of a prediction model to *decrease* when the information in X_1 is removed. Roughly speaking, a "pathological" model f_{silly} may use the information in X_1 to "intentionally" misclassify Y, such that $e_{\text{switch}}(f_{\text{silly}}) < e_{\text{orig}}(f_{\text{silly}})$ and $MR(f_{\text{silly}}) < 1$. The model f_{silly} may even be included in a Rashomon set (see Section 4) if it is still possible to predict Y sufficiently well from the information in X_2.

However, in these cases there will often exist another model that outperforms f_{silly}, and that has MR equal to 1 (i.e., no reliance on X_1). To see this, consider the case where

$\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^d\}$ is indexed by a parameter $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_{\text{silly}}$ and $\boldsymbol{\theta}^\star$ be parameter values such that $f_{\boldsymbol{\theta}_{\text{silly}}}$ is equivalent to f_{silly}, and $f_{\boldsymbol{\theta}^\star}$ is the best-in-class model. If $f_{\boldsymbol{\theta}^\star}$ satisfies $MR(f_{\boldsymbol{\theta}^\star}) > 1$ and if the model reliance function MR is continuous in $\boldsymbol{\theta}$, then there exists a parameter value $\boldsymbol{\theta}_1$ between $\boldsymbol{\theta}_{\text{silly}}$ and $\boldsymbol{\theta}^\star$ such that $MR(f_{\boldsymbol{\theta}_1}) = 1$. Further, if the loss function L is convex in $\boldsymbol{\theta}$, then $e_{\text{orig}}(f_{\boldsymbol{\theta}^\star}) \leq e_{\text{orig}}(f_{\boldsymbol{\theta}_1}) \leq e_{\text{orig}}(f_{\text{silly}})$, and any Rashomon set containing f_{silly} will also contain $f_{\boldsymbol{\theta}_1}$.

A.3 Relating $\hat{e}_{\text{switch}}(f)$ to all possible permutations of the sample

Following the notation in Section 3, let $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$ be a set of n-length vectors, each containing a different permutation of the set $\{1, \ldots, n\}$. We show in this section that $\hat{e}_{\text{switch}}(f)$ is equal to the product of

$$\sum_{l=1}^{n!} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]}, \mathbf{X}_{2[i,\cdot]})\} 1(\boldsymbol{\pi}_{l[i]} \neq i), \tag{A.1}$$

and a proportionality constant that is only a function of n.

First, consider the sum

$$\sum_{l=1}^{n!} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]}, \mathbf{X}_{2[i,\cdot]})\}, \tag{A.2}$$

which omits the indicator function found in Eq A.1.

The summation in Eq A.2 contains $n(n!)$ terms, each of which is a two-way combination of the form $L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}$ for $i, j \in \{1, \ldots n\}$. There are only n^2 unique combinations of this form, and each must occur in at least $(n-1)!$ of the $n(n!)$ terms in Eq A.2. To see this, consider selecting two integer values $\tilde{i}, \tilde{j} \in \{1, \ldots n\}$, and enumerating all occurrences of the term $L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\tilde{j},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\}$ within the sum in Eq A.2. Of the permutation vectors $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{n!}\}$, we know that $(n-1)!$ of them place $\tilde{i}$ in the $\tilde{j}^{th}$ position, i.e., that satisfy $\boldsymbol{\pi}_{l[\tilde{i}]} = \tilde{j}$. For each such permutation $\boldsymbol{\pi}_l$, the inner summation in Eq A.2 over all possible values of i must include the term $L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[\tilde{i}]},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\} = L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\tilde{j},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\}$. Thus, Eq A.2 contains at least $(n-1)!$ occurrences of the term $L\{f, (\mathbf{y}_{[\tilde{i}]}, \mathbf{X}_{1[\tilde{j},\cdot]}, \mathbf{X}_{2[\tilde{i},\cdot]})\}$.

So far, we have shown that each unique combination occurs at least $(n-1)!$ times, but it also follows that each unique combination must occur precisely $(n-1)!$ times. This is because each of the n^2 unique combinations must occur at least $(n-1)!$ times, which accounts for $n^2((n-1)!) = n(n!)$ terms in total. As noted above, Eq has A.2 has only $n(n)!$ terms, so there can be no additional terms. We can then simplify Eq A.2 as

$$\sum_{l=1}^{n!} \sum_{i=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]}, \mathbf{X}_{2[i,\cdot]})\} = (n-1)! \sum_{i=1}^{n} \sum_{j=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j,\cdot]}, \mathbf{X}_{2[i,\cdot]})\}.$$

By the same logic, we can simplify Eq A.1 as

$$
\begin{aligned}
&\sum_{l=1}^{n!}\sum_{i=1}^{n} L\{f,(\mathbf{y}_{[i]},\mathbf{X}_{1[\boldsymbol{\pi}_{l[i]},\cdot]},\mathbf{X}_{2[i,\cdot]})\}1(\boldsymbol{\pi}_{l[i]}\neq i)\\
=&(n-1)!\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} L\{f,(\mathbf{y}_{[i]},\mathbf{X}_{1[j,\cdot]},\mathbf{X}_{2[i,\cdot]})\}1(j\neq i)\right\}\\
=&(n-1)!\sum_{i=1}^{n}\sum_{j\neq i} L\{f,(\mathbf{y}_{[i]},\mathbf{X}_{1[j,\cdot]},\mathbf{X}_{2[i,\cdot]})\}, \qquad \text{(A.3)}
\end{aligned}
$$

and Line A.3 is proportional to $\hat{e}_{\text{switch}}(f)$ up to a function of n.

A.4 Illustration of Theorems 6 and 9

Figure 8 illustrates the different terms that compose the bounds in Theorems 6 and 9. Roughly speaking, to create the bounds for $MCR_+(\epsilon)$ (and $MCR_-(\epsilon)$) in Theorem 6, we expand the empirical Rashomon set by increasing ϵ to ϵ_1, such that $f_{+,\epsilon}$ (or $f_{-,\epsilon}$) is contained in $\hat{\mathcal{R}}(\epsilon_1)$ with high probability. We then add (or subtract) $\mathcal{Q}_1$ to account for estimation error of $\widehat{MR}(f_{+,\epsilon})$ (or $\widehat{MR}(f_{-,\epsilon})$). To create the bounds for $MCR_+(\epsilon)$ (and $MCR_-(\epsilon)$) in Theorem 9, we constrict the empirical Rashomon set by decreasing ϵ to ϵ_3, such that all models with high expected loss are simultaneously excluded from $\hat{\mathcal{R}}(\epsilon_3)$ with high probability. We then subtract (or add) $\mathcal{Q}_3$ to simultaneously account for MR estimation error for models in $\hat{\mathcal{R}}(\epsilon_3)$.

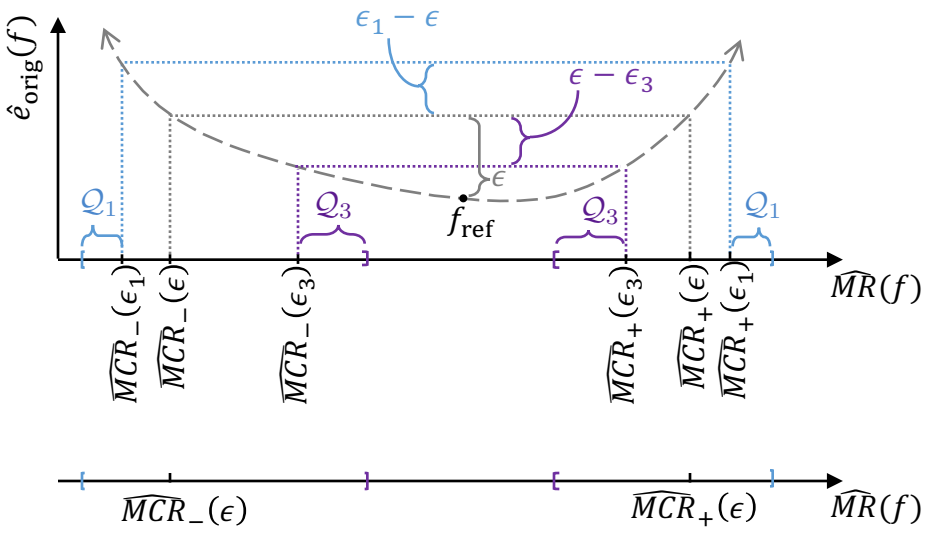

Figure 8: Illustration of terms in Theorems 6 and 9 – Above we show the relation between empirical MR (x-axis) and empirical loss (y-axis) for models f in a hypothetical model class $\mathcal{F}$. We mark f_{ref} by the black point. For each possible model reliance value $r \geq 0$, the curved, dashed line shows the lowest possible empirical loss for a function in $f \in \mathcal{F}$ satisfying $\widehat{MR}(f) = r$. The set $\hat{\mathcal{R}}(\epsilon)$ contains all models in $\mathcal{F}$ within the dotted gray lines. To create the bounds from Theorem 6, we increase ϵ to ϵ_1, compute $\widehat{MCR}_+(\epsilon_1)$ (and $\widehat{MCR}_-(\epsilon_1)$), and add (or subtract) $\mathcal{Q}_1$. All of these changes are illustrated above in blue, with the final bounds shown by the blue bracket symbols along the x-axis. To create the bounds from Theorem 9, we decrease ϵ to ϵ_3, compute $\widehat{MCR}_+(\epsilon_3)$ (and $\widehat{MCR}_-(\epsilon_3)$), and subtract (or add) $\mathcal{Q}_3$. These changes are illustrated above in purple, with the final bounds shown by the purple bracket symbols along the x-axis. For emphasis, below this figure we show a copy of the x-axis with selected annotations, from which it is clear that $\widehat{MCR}_-(\epsilon)$ and $\widehat{MCR}_+(\epsilon)$ are always within the bounds produced by Theorems 6 and 9.

A.5 Ratios versus differences for MR definition

We choose our ratio-based definition of model reliance, $MR(f) = \frac{e_{\text{switch}}(f)}{e_{\text{orig}}(f)}$, so that the measure can be comparable across problems, regardless of the scale of Y. However, several existing works define VI measures in terms of differences (Strobl et al., 2008; Datta et al., 2016; Gregorutti et al., 2017), analogous to

$$MR_{\text{difference}}(f) := e_{\text{switch}}(f) - e_{\text{orig}}(f). \tag{A.4}$$

While this difference measure is less readily interpretable, it has several computational advantages. The mean, variance, and asymptotic distribution of the estimator $\widehat{MR}_{\text{difference}}(f) :=$

$\hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$ can be easily determined using results for U-statistics, without the use of the delta method. Estimates in the form of $\widehat{MR}_{\text{difference}}(f)$ will also be more stable when $\min_{f \in \mathcal{F}} e_{\text{orig}}(f)$ is small, relative to estimates for the ratio-based definition of MR. To improve interpretability, we may also normalize $MR_{\text{difference}}(f)$ by dividing by the variance of Y, which can be easily estimated without the use of models, as in Williamson et al. (2017).

Under the difference-based definition for MR (Eq A.4), Theorem 6, Corollary 7, Theorem 9 will still hold under the following modified definitions of $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_3$:

$$\begin{aligned}
\mathcal{Q}_{1,\text{difference}} &:= \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{n}}, \\
\mathcal{Q}_{2,\text{difference}} &:= \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}} \sqrt{\frac{\log(12\delta^{-1})}{n}}, \text{ and} \\
\mathcal{Q}_{3,\text{difference}} &:= B_{\text{ind}} \left\{ \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\right))}{2n}} \right\} + 2r(\sqrt{2}+1).
\end{aligned}$$

Respectively replacing $\mathcal{Q}_1$, $\mathcal{Q}_2$, $\mathcal{Q}_3$, MR, and $\widehat{MR}$ with $\mathcal{Q}_{1,\text{difference}}$, $\mathcal{Q}_{2,\text{difference}}$, $\mathcal{Q}_{3,\text{difference}}$, $MR_{\text{difference}}$ and $\widehat{MR}_{\text{difference}}$ entails only minor changes to the corresponding proofs (see Appendices B.5, B.6, and B.7). The results will also hold without Assumption 5, as is suggested by the fact that b_{orig} and B_{switch} do not appear in $\mathcal{Q}_{1,\text{difference}}$, $\mathcal{Q}_{2,\text{difference}}$ or $\mathcal{Q}_{3,\text{difference}}$.

We also prove an analogous version of Theorem 8 in Theorem 27.

B Proofs for statistical results

B.1 Proof of Theorem 1

Proof. First we consider $e_{\text{orig}}(f_0)$. We briefly recall that the notation $f_0(t,c)$ refers to the *true* conditional expectation function for *both* potential outcomes Y_1, Y_0, rather than the expectation for Y_0 alone.

Under the assumption that $(Y_1, Y_0) \perp T|C$, we have $f_0(t,c) = \mathbb{E}(Y|C=c, T=t) = \mathbb{E}(Y_t|C=c)$. Applying this, we see that

$$\begin{aligned}
e_{\text{orig}}(f_0) &= \mathbb{E}L(f_0, (Y,T,C)) \\
&= \mathbb{E}L(f_0, (Y_T,T,C)) \\
&= \mathbb{E}_T \mathbb{E}_{C|T} \mathbb{E}_{Y_T|C}[\{Y_T - \mathbb{E}(Y_T|C)\}^2] \\
&= \mathbb{E}_T \mathbb{E}_{C|T} \text{Var}(Y_T|C) \\
&= q\mathbb{E}_{C|T=0} \text{Var}(Y_0|C) + p\mathbb{E}_{C|T=1} \text{Var}(Y_1|C), \qquad \text{(B.1)}
\end{aligned}$$

where $p := \mathbb{P}(T=1)$ and $q := \mathbb{P}(T=0)$.

Now we consider $e_{\text{switch}}(f_0)$. Let $(Y_0^{(a)}, Y_1^{(a)}, T^{(a)}, C^{(a)})$ and $(Y_0^{(b)}, Y_1^{(b)}, T^{(b)}, C^{(b)})$ be a pair of independent random variable vectors, each with the same distribution as (Y_0, Y_1, T, C). Then

$$
\begin{aligned}
e_{\text{switch}}(f_0) &= \mathbb{E}_{T^{(b)},T^{(a)},C^{(b)},Y^{(b)}_{T^{(b)}}}[\{Y^{(b)}_{T^{(b)}} - f_0(T^{(a)}, C^{(b)})\}^2] \\
&= \mathbb{E}_{T^{(b)},T^{(a)},C^{(b)},Y^{(b)}_{T^{(b)}}}[\{Y^{(b)}_{T^{(b)}} - \mathbb{E}(Y_{T^{(a)}}|C = C^{(b)})\}^2] \\
&= \mathbb{E}_{T^{(b)},T^{(a)}}\mathbb{E}_{C^{(b)}|T^{(b)}}\mathbb{E}_{Y^{(b)}_{T^{(b)}}|C^{(b)}}[\{Y^{(b)}_{T^{(b)}} - \mathbb{E}(Y_{T^{(a)}}|C = C^{(b)})\}^2].
\end{aligned}
$$

First we expand the outermost expectation, over $T^{(b)}, T^{(a)}$:

$$
\begin{aligned}
&e_{\text{switch}}(f_0) \\
&= \sum_{i,j\in\{0,1\}} \mathbb{P}(T^{(b)} = i, T^{(a)} = j)\mathbb{E}_{C^{(b)}|T^{(b)}=i}\mathbb{E}_{Y^{(b)}_i|C^{(b)}}[\{Y^{(b)}_i - \mathbb{E}(Y_j|C = C^{(b)})\}^2]. \quad \text{(B.2)}
\end{aligned}
$$

Since $T^{(b)} \perp T^{(a)}$, we can write

$$
\begin{aligned}
\mathbb{P}(T^{(b)} = i, T^{(a)} = j) &= \mathbb{P}(T^{(b)} = i)\mathbb{P}(T^{(a)} = j) \\
&= p^{i+j}q^{2-i-j}.
\end{aligned}
$$

Plugging this into Eq B.2 we get

$$
e_{\text{switch}}(f_0) = \sum_{i,j\in\{0,1\}} p^{i+j}q^{2-i-j}\mathbb{E}_{C^{(b)}|T^{(b)}=i}\mathbb{E}_{Y^{(b)}_i|C^{(b)}}[\{Y^{(b)}_i - \mathbb{E}(Y_j|C = C^{(b)})\}^2].
$$

Since $(Y^{(b)}_0, Y^{(b)}_1, C^{(b)}, T^{(b)})$ are the only random variables remaining, we can omit the superscript notation (e.g., replace $C^{(b)}$ with C) to get

$$
\begin{aligned}
e_{\text{switch}}(f_0) &= \sum_{i,j\in\{0,1\}} p^{i+j}q^{2-i-j}\mathbb{E}_{C|T=i}\mathbb{E}_{Y_i|C}[\{Y_i - \mathbb{E}(Y_j|C)\}^2] \\
&=: \sum_{i,j\in\{0,1\}} A_{ij},
\end{aligned}
$$

where $A_{ij} = p^{i+j}q^{2-i-j}\mathbb{E}_{C|T=i}\mathbb{E}_{Y_i|C}[\{Y_i - \mathbb{E}(Y_j|C)\}^2]$. First, we consider A_{00} and A_{11}:

$$
\begin{aligned}
A_{00} &= q^2\mathbb{E}_{C|T=0}\mathbb{E}_{Y_0|C}[\{Y_0 - \mathbb{E}(Y_0|C)\}^2] \\
&= q^2\mathbb{E}_{C|T=0}\text{Var}(Y_0|C),
\end{aligned}
$$

and, similarly, $A_{11} = p^2\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)$.

Next we consider A_{01} and A_{10}:

$$
\begin{aligned}
A_{01}: &= pq\mathbb{E}_{C|T=0}\mathbb{E}_{Y_0|C}[\{Y_0 - \mathbb{E}(Y_1|C)\}^2] \\
&= pq\mathbb{E}_{C|T=0}\left(\mathbb{E}(Y_0^2|C) - 2\mathbb{E}(Y_0|C)\mathbb{E}(Y_1|C) + \mathbb{E}(Y_1|C)^2\right) \\
&= pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + \mathbb{E}(Y_0|C)^2 - 2\mathbb{E}(Y_0|C)\mathbb{E}(Y_1|C) + \mathbb{E}(Y_1|C)^2\right) \\
&= pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + [\mathbb{E}(Y_1|C) - \mathbb{E}(Y_0|C)]^2\right) \\
&= pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + \text{CATE}(C)^2\right),
\end{aligned}
$$

and, following the same steps,

$$A_{10} = pq\mathbb{E}_{C|T=1}\left(\text{Var}(Y_1|C) + \text{CATE}(C)^2\right).$$

Plugging in A_{00}, A_{01}, A_{10}, and A_{11} we get

$$\begin{aligned}
e_{\text{switch}}(f_0) &= \{A_{00} + A_{11}\} \\
&\quad + [A_{01} + A_{10}] \\
&= \left\{q^2\mathbb{E}_{C|T=0}\text{Var}(Y_0|C) + p^2\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)\right\} \\
&\quad + \left[pq\mathbb{E}_{C|T=0}\left(\text{Var}(Y_0|C) + \text{CATE}(C)^2\right) + pq\mathbb{E}_{C|T=1}\left(\text{Var}(Y_1|C) + \text{CATE}(C)^2\right)\right] \\
&= \left\{q(q+p)\mathbb{E}_{C|T=0}\text{Var}(Y_0|C) + p(p+q)\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)\right\} \quad \text{(B.3)} \\
&\quad + pq\left[\mathbb{E}_{C|T=0}\left(\text{CATE}(C)^2\right) + \mathbb{E}_{C|T=1}\left(\text{CATE}(C)^2\right)\right] \quad \text{(B.4)} \\
&= \{e_{\text{orig}}(f_0)\} \quad \text{(B.5)} \\
&\quad + \text{Var}(T)\left[\mathbb{E}_{C|T=0}\left(\text{CATE}(C)^2\right) + \mathbb{E}_{C|T=1}\left(\text{CATE}(C)^2\right)\right]. \quad \text{(B.6)}
\end{aligned}$$

In Lines B.3 and B.4, we consolidate terms involving $\mathbb{E}_{C|T=0}\text{Var}(Y_0|C)$ and $\mathbb{E}_{C|T=1}\text{Var}(Y_1|C)$. In Line B.5, we use $p + q = 1$ to reduce Line B.3 to the right-hand side of Eq B.1. Finally, in Line B.6, we use $qp = \text{Var}(T)$. □

B.2 Proof of Theorem 2

Proof. To show Eq 3.4 we start with $e_{\text{orig}}(f_\beta)$,

$$\begin{aligned}
e_{\text{orig}}(f_\beta) &= \mathbb{E}[\{Y - X_1'\beta_1 - X_2'\beta_2\}^2] \\
&= \mathbb{E}[\{(Y - X_2'\beta_2) - X_1'\beta_1\}^2] \\
&= \mathbb{E}[(Y - X_2'\beta_2)^2] - 2\mathbb{E}[(Y - X_2'\beta_2)X_1']\beta_1 + \beta_1'\mathbb{E}[X_1X_1']\beta_1.
\end{aligned}$$

For $e_{\text{switch}}(f_\beta)$, we can follow the same steps as above:

$$\begin{aligned}
e_{\text{switch}}(f_\beta) &= \mathbb{E}_{Y^{(b)},X_1^{(a)},X_2^{(b)}}[\{Y^{(b)} - X_1^{(a)'}\beta_1 - X_2^{(b)'}\beta_2\}^2] \\
&= \mathbb{E}[(Y^{(b)} - X_2^{(b)'}\beta_2)^2] - 2\mathbb{E}[Y^{(b)} - X_2^{(b)'}\beta_2]\mathbb{E}[X_1^{(a)'}]\beta_1 + \beta_1'\mathbb{E}[X_1^{(a)}X_1^{(a)'}]\beta_1.
\end{aligned}$$

Since $(Y^{(b)}, X_1^{(b)}, X_2^{(b)})$ and $(Y^{(a)}, X_1^{(a)}, X_2^{(a)})$ each have the same distribution as (Y, X_1, X_2), we can omit the superscript notation to show Eq 3.4:

$$\begin{aligned}
e_{\text{switch}}(f_\beta) &= \mathbb{E}[(Y - X_2'\beta_2)^2] - 2\mathbb{E}[Y - X_2'\beta_2]\mathbb{E}[X_1']\beta_1 + \beta_1'\mathbb{E}[X_1X_1']\beta_1 \\
e_{\text{switch}}(f_\beta) &= e_{\text{orig}}(f_\beta) - 2\mathbb{E}[Y - X_2'\beta_2]\mathbb{E}[X_1']\beta_1 + 2\mathbb{E}[(Y - X_2'\beta_2)X_1']\beta_1 \\
e_{\text{switch}}(f_\beta) &= e_{\text{orig}}(f_\beta) + 2\text{Cov}(Y - X_2'\beta_2, X_1)\beta_1 \\
e_{\text{switch}}(f_\beta) &= e_{\text{orig}}(f_\beta) + 2\text{Cov}(Y, X_1)\beta_1 - 2\beta_2\text{Cov}(X_2, X_1)\beta_1 \\
e_{\text{switch}}(f_\beta) - e_{\text{orig}}(f_\beta) &= 2\text{Cov}(Y, X_1)\beta_1 - 2\beta_2\text{Cov}(X_2, X_1)\beta_1.
\end{aligned}$$

Next, we can use a similar approach to show Eq 3.5:

$$\begin{aligned}
\hat{e}_{\text{switch}}(f_\beta) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} h_f(\mathbf{Z}_{[i,\cdot]}, \mathbf{Z}_{[j,\cdot]}) \\
n(n-1)\hat{e}_{\text{switch}}(f_\beta) &= \sum_{i=1}^{n} \sum_{j \neq i} (\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2 - \mathbf{X}_{1[i,\cdot]}\beta_1)^2 \\
&= \sum_{i=1}^{n} \sum_{j \neq i} \left\{ (\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2)^2 - 2(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2)(\mathbf{X}_{1[i,\cdot]}\beta_1) + (\mathbf{X}_{1[i,\cdot]}\beta_1)^2 \right\} \\
&= (n-1) \sum_{i=1}^{n} (\mathbf{y}_{[i]} - \mathbf{X}_{2[i,\cdot]}\beta_2)^2 \\
&\qquad - 2 \left\{ \sum_{i=1}^{n} \sum_{j \neq i} (\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2) \right\} + (n-1) \sum_{i=1}^{n} (\mathbf{X}_{1[i,\cdot]}\beta_1)^2.
\end{aligned} \tag{B.7}$$

Focusing on the term in braces,

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j \neq i} (\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2) \\
&\qquad = \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2) - \sum_{i=1}^{n} (\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[i]} - \mathbf{X}_{2[i,\cdot]}\beta_2) \\
&\qquad = \sum_{i=1}^{n} (\mathbf{X}_{1[i,\cdot]}\beta_1) \sum_{j=1}^{n} (\mathbf{y}_{[j]} - \mathbf{X}_{2[j,\cdot]}\beta_2) - \sum_{i=1}^{n} (\mathbf{X}_{1[i,\cdot]}\beta_1)(\mathbf{y}_{[i]} - \mathbf{X}_{2[i,\cdot]}\beta_2) \\
&\qquad = \left\{ (\mathbf{X}_1\beta_1)'\mathbf{1}_n \right\} \left\{ \mathbf{1}_n'(\mathbf{y} - \mathbf{X}_2\beta_2) \right\} - (\mathbf{X}_1\beta_1)'(\mathbf{y} - \mathbf{X}_2\beta_2) \\
&\qquad = (\mathbf{X}_1\beta_1)'(\mathbf{1}_n\mathbf{1}_n' - \mathbf{I}_n)(\mathbf{y} - \mathbf{X}_2\beta_2).
\end{aligned} \tag{B.8}$$

Plugging this into Eq B.7, and applying the sample linear algebra representations as in Eq B.8, we get

$$\begin{aligned}
n(n-1)\hat{e}_{\text{switch}}(f_\beta) &= (n-1)||\mathbf{y} - \mathbf{X}_2\beta_2||_2^2 \\
&\qquad - 2(\mathbf{X}_1\beta_1)'(\mathbf{1}_n\mathbf{1}_n' - \mathbf{I}_n)(\mathbf{y} - \mathbf{X}_2\beta_2) \\
&\qquad + (n-1)||\mathbf{X}_1\beta_1||_2^2 \\
n\hat{e}_{\text{switch}}(f_\beta) &= ||\mathbf{y} - \mathbf{X}_2\beta_2||_2^2 \\
&\qquad - 2(\mathbf{X}_1\beta_1)'\mathbf{W}(\mathbf{y} - \mathbf{X}_2\beta_2) \\
&\qquad + ||\mathbf{X}_1\beta_1||_2^2 \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}_2\beta_2 + \beta_2'\mathbf{X}_2'\mathbf{X}_2\beta_2 \\
&\qquad - 2\beta_1'\mathbf{X}_1'\mathbf{W}\mathbf{y} + 2\beta_1'\mathbf{X}_1'\mathbf{W}\mathbf{X}_2\beta_2 \\
&\qquad + \beta_1'\mathbf{X}_1'\mathbf{X}_1\beta_1 \\
&= \mathbf{y}'\mathbf{y} - 2 \begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}' \beta + \beta' \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \beta.
\end{aligned}$$

□

B.3 Lemma relating empirical and population Rashomon sets

Throughout the remaining proofs, it will be useful to express the definition of Rashomon sets in terms of the expectation of a single loss function, rather than a comparison of two loss functions. To do this, we simply introduce the "standardized" loss function $\tilde{L}$, defined as

$$\tilde{L}(f) := L(f) - L(f_{\text{ref}}). \tag{B.9}$$

Because we assume f_{ref} is prespecified and fixed, we omit notation for f_{ref} in the definition of $\tilde{L}$. We can now write

$$\begin{aligned}\mathcal{R}(\epsilon) &= \{f_{\text{ref}}\} \cup \{f \in \mathcal{F} : \mathbb{E}L(f, Z) \leq \mathbb{E}L(f_{\text{ref}}, Z) + \epsilon\} \\ &= \{f_{\text{ref}}\} \cup \left\{f \in \mathcal{F} : \mathbb{E}\tilde{L}(f, Z) \leq \epsilon\right\},\end{aligned}$$

and, similarly,

$$\hat{\mathcal{R}}(\epsilon) = \{f_{\text{ref}}\} \cup \left\{f \in \mathcal{F} : \hat{\mathbb{E}}\tilde{L}(f, Z) \leq \epsilon\right\}.$$

With this definition, and for a given model $f_1 \in \mathcal{R}(\epsilon, f_{\text{ref}})$, the following lemma allows us to limit the probability that f_1 is excluded from empirical Rashomon sets.

Lemma 25. *For $\epsilon \in \mathbb{R}$ and $\delta \in (0, 1)$, let $\epsilon_4 := \epsilon + 2B_{ref}\sqrt{\frac{\log(\delta^{-1})}{2n}}$, and let $f_1 \in \mathcal{R}(\epsilon)$ denote a specific, possibly unknown prediction model. If f_1 satisfies Assumption 4, then*

$$\mathbb{P}\{f_1 \in \hat{\mathcal{R}}(\epsilon_4)\} \geq 1 - \delta.$$

Proof. If f_{ref} and f_1 are the same function, then the result holds trivially. Otherwise, the proof follows from Hoeffding's inequality. First, note that if f_1 satisfies Assumption 4, then $\tilde{L}(f_1)$ is bounded within an interval of length $2B_{\text{ref}}$. Applying this in line B.11, below, we see that

$$\begin{aligned}\mathbb{P}\{f_1 \notin \hat{\mathcal{R}}(\epsilon_4)\} &= \mathbb{P}\left[\hat{\mathbb{E}}\tilde{L}(f_1, Z) > \epsilon_4\right] && \text{from } f_1 \notin \{f_{\text{ref}}\} \\ &= \mathbb{P}\left[\hat{\mathbb{E}}\tilde{L}(f_1, Z) - \epsilon > 2B_{\text{ref}}\sqrt{\frac{\log(\delta^{-1})}{2n}}\right] && \text{from definition of } \epsilon_4 \qquad \text{(B.10)} \\ &\leq \mathbb{P}\left[\hat{\mathbb{E}}\tilde{L}(f_1, Z) - \mathbb{E}\tilde{L}(f_1, Z) > 2B_{\text{ref}}\sqrt{\frac{\log(\delta^{-1})}{2n}}\right] && \text{from } \mathbb{E}\tilde{L}(f_1, Z) \leq \epsilon \\ &\leq \exp\left\{-\frac{2n}{(2B_{\text{ref}})^2}\left[2B_{\text{ref}}\sqrt{\frac{\log(\delta^{-1})}{2n}}\right]^2\right\} && \text{from Hoeffding's inequality} \qquad \text{(B.11)} \\ &= \delta. && \qquad \text{(B.12)}\end{aligned}$$

For Hoeffding's inequality, see Pg 201 of Serfling, 1980, Theorem A. □

B.4 Lemma to transform between bounds

The following lemma will help us translate from bounds for variables to bounds for differences and ratios of those variables. We will apply this lemma to transform from bounds on empirical losses to bounds on empirical model reliance, defined either in terms of a ratio or in terms of a difference.

Lemma 26. *Let $X, Z, \mu_X, \mu_Z, k_X, k_Z \in \mathbb{R}$ be constants satisfying $|Z - \mu_Z| \leq k_Z$ and $|X - \mu_X| \leq k_X$, then*

$$|(Z - X) - (\mu_Z - \mu_X)| \leq q_{difference}(k_Z, k_X), \tag{B.13}$$

where $q_{difference}$ is the function

$$q_{difference}(k_Z, k_X) := k_Z + k_X. \tag{B.14}$$

Further, if there exists constants b_{orig} and B_{switch} such that $0 < b_{orig} \leq X, \mu_X$ and $Z, \mu_Z \leq B_{switch} < \infty$, then

$$\left| \frac{Z}{X} - \frac{\mu_Z}{\mu_X} \right| \leq q_{ratio}(k_Z, k_X), \tag{B.15}$$

where q_{ratio} is the function

$$q_{ratio}(k_Z, k_X) := \frac{B_{switch}}{b_{orig}} - \frac{B_{switch} - k_Z}{b_{orig} + k_X}. \tag{B.16}$$

Proof. Showing Eq B.13,

$$\begin{aligned} |(Z - X) - (\mu_Z - \mu_X)| &\leq |Z - \mu_Z| + |\mu_X - X| \\ &\leq k_Z + k_X. \end{aligned}$$

Showing Eq B.15, let $A_Z = \max(Z, \mu_Z)$, $a_X = \min(X, \mu_X)$, $d_Z = |Z - \mu_Z|$, and $d_X = |X - \mu_X|$. This implies that $\max(X, \mu_X) = a_X + d_X$ and $\min(Z, \mu_Z) = A_Z - d_Z$. Thus, $\frac{Z}{X}$ and $\frac{\mu_Z}{\mu_X}$ are both bounded within the interval

$$\left[\frac{\min(Z, \mu_Z)}{\max(X, \mu_X)}, \frac{\max(Z, \mu_Z)}{\min(X, \mu_X)} \right] = \left[\frac{A_Z - d_Z}{a_X + d_X}, \frac{A_Z}{a_X} \right],$$

which implies

$$\left| \frac{Z}{X} - \frac{\mu_Z}{\mu_X} \right| \leq \frac{A_Z}{a_X} - \frac{A_Z - d_Z}{a_X + d_X}. \tag{B.17}$$

Taking partial derivatives of the right-hand side, we get

$$\frac{\partial}{\partial a_X}\left(\frac{A_Z}{a_X}-\frac{A_Z-d_Z}{a_X+d_X}\right)=\frac{-A_Z}{a_X^2}+\frac{A_Z-d_Z}{(a_X+d_X)^2}\leq 0,$$
$$\frac{\partial}{\partial A_Z}\left(\frac{A_Z}{a_X}-\frac{A_Z-d_Z}{a_X+d_X}\right)=\frac{1}{a_X}-\frac{1}{a_X+d_X}\geq 0,$$
$$\frac{\partial}{\partial d_X}\left(\frac{A_Z}{a_X}-\frac{A_Z-d_Z}{a_X+d_X}\right)=\frac{A_Z-d_Z}{(a_X+d_X)^2}>0,$$
$$\text{and } \frac{\partial}{\partial d_Z}\left(\frac{A_Z}{a_X}-\frac{A_Z-d_Z}{a_X+d_X}\right)=\frac{1}{a_X+d_X}>0.$$

So the right-hand side of B.17 is maximized when d_Z, d_X, and A_Z are maximized, and when a_X is minimized. Thus, in the case where $|Z-\mu_Z|\leq k_Z$; $|X-\mu_X|\leq k_X$; $0<b_{\text{orig}}\leq X,\mu_X$; and $Z,\mu_Z\leq B_{\text{switch}}<\infty$, we have

$$\begin{aligned}\left|\frac{Z}{X}-\frac{\mu_Z}{\mu_X}\right| &\leq \frac{A_Z}{a_X}-\frac{A_Z-d_Z}{a_X+d_X}\\ &\leq \frac{B_{\text{switch}}}{b_{\text{orig}}}-\frac{B_{\text{switch}}-k_Z}{b_{\text{orig}}+k_X}.\end{aligned}$$

□

B.5 Proof of Theorem 6

Proof.

Step 1: Show that $\mathbb{P}\left[\widehat{MR}(f_{+,\epsilon})\leq \widehat{MCR}_+(\epsilon_1)\right]\geq 1-\frac{\delta}{3}$.

Consider the event that

$$\widehat{MR}(f_{+,\epsilon})\leq \widehat{MCR}_+(\epsilon_1). \tag{B.18}$$

Eq B.18 will always hold if $f_{+,\epsilon}\in\hat{\mathcal{R}}(\epsilon_1)$, since $\widehat{MCR}_+(\epsilon_1)$ upper bounds the empirical model reliance for models in $\hat{\mathcal{R}}(\epsilon_1)$ by definition. Applying the above reasoning in Line B.19, below, we get

$$\mathbb{P}\left[\widehat{MR}(f_{+,\epsilon})>\widehat{MCR}_+(\epsilon_1)\right]\leq \mathbb{P}\left[f_{+,\epsilon}\notin\hat{\mathcal{R}}(\epsilon_1)\right] \tag{B.19}$$
$$\leq \frac{\delta}{3} \qquad \text{from } \epsilon_1 \text{ definition and Lemma 25.} \tag{B.20}$$

Step 2: Conditional on $\widehat{MR}(f_{+,\epsilon})\leq\widehat{MCR}_+(\epsilon_1)$, upper bound $MR(f_{+,\epsilon})$ by $\widehat{MCR}_+(\epsilon_1)$ added to an error term.

When Eq B.18 holds we have,

$$\begin{aligned}\widehat{MR}(f_{+,\epsilon}) &\leq \widehat{MCR}_+(\epsilon_1)\\ \widehat{MR}(f_{+,\epsilon}) &\leq \widehat{MCR}_+(\epsilon_1)+\{MR(f_{+,\epsilon})-MR(f_{+,\epsilon})\}\\ MR(f_{+,\epsilon}) &\leq \widehat{MCR}_+(\epsilon_1)+[MR(f_{+,\epsilon})-\widehat{MR}(f_{+,\epsilon})].\end{aligned} \tag{B.21}$$

Step 3: Probabilistically bound the error term from Step 2.

Next we show that the bracketed term in Line B.21 is less than or equal to $\mathcal{Q}_1$ with high probability. For $k \in \mathbb{R}$, let $q_{\text{difference}}$ and q_{ratio} be defined as in Eqs B.14 and B.16. Let $q : \mathbb{R} \to \mathbb{R}$ be the function such that $q(k) = q_{\text{ratio}}\left(k, \frac{k}{\sqrt{2}}\right)$. Then

$$\begin{aligned}
\mathcal{Q}_1 &= \frac{B_{\text{switch}}}{b_{\text{orig}}} - \frac{B_{\text{switch}} - B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}}{b_{\text{orig}} + B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{2n}}} \\
&= q_{\text{ratio}}\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}, B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{2n}}\right) \\
&= q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right).
\end{aligned}$$

Applying this relation below, we have

$$\begin{aligned}
&\mathbb{P}\left[MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon}) > \mathcal{Q}_1\right] && \text{(B.22)} \\
&\leq \mathbb{P}\left[\left|MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon})\right| > q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right] \\
&\leq \mathbb{P}\left[\left\{|\hat{e}_{\text{orig}}(f_{+,\epsilon}) - e_{\text{orig}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{2n}}\right\}\right. \\
&\qquad \left.\bigcup\left\{|\hat{e}_{\text{switch}}(f_{+,\epsilon}) - e_{\text{switch}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right\}\right] \quad \text{from Lemma 26} \\
&\leq \mathbb{P}\left[|\hat{e}_{\text{orig}}(f_{+,\epsilon}) - e_{\text{orig}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{2n}}\right] \\
&\qquad + \mathbb{P}\left[|\hat{e}_{\text{switch}}(f_{+,\epsilon}) - e_{\text{switch}}(f_{+,\epsilon})| > B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{n}}\right] \quad \text{from the Union bound} \\
&\leq 2\exp\left\{-\frac{2n}{(B_{\text{ind}}-0)^2}\left[B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{2n}}\right]^2\right\} \\
&\qquad + 2\exp\left\{-\frac{n}{(B_{\text{ind}}-0)^2}\left[B_{\text{ind}}\sqrt{\frac{\log\left(6\delta^{-1}\right)}{n}}\right]^2\right\} \quad \text{from Hoeffding's inequality} \\
& && \text{(B.23)} \\
&= \frac{2\delta}{6} + \frac{2\delta}{6} = \frac{2\delta}{3}. && \text{(B.24)}
\end{aligned}$$

In Line B.23, above, recall that $\mathbb{E}\left[\hat{e}_{\text{switch}}(f_{+,\epsilon})\right] = e_{\text{switch}}(f_{+,\epsilon})$ because $\hat{e}_{\text{switch}}(f_{+,\epsilon})$ is an average of terms, and each term has expectation equal to $e_{\text{switch}}(f_{+,\epsilon})$. For the same reason, $\mathbb{E}\left[\hat{e}_{\text{orig}}(f_{+,\epsilon})\right] = e_{\text{orig}}(f_{+,\epsilon})$.

Alternatively, if we have defined model reliance as $MR(f) = e_{\text{switch}}(f) - e_{\text{orig}}(f)$ (see Section A.5), with $\widehat{MR}(f) := \hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$, and

$$\mathcal{Q}_{1,\text{difference}} := \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{n}} = q_{\text{difference}} \left(B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{n}}, B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{2n}}\right),$$

then the same proof holds without Assumption 5 if we replace $\mathcal{Q}_1$ with $\mathcal{Q}_{1,\text{difference}}$, and redefine $q : \mathbb{R} \to \mathbb{R}$ as the function $q(k) = q_{\text{difference}}\left(k, \frac{k}{\sqrt{2}}\right)$.

Eqs B.22-B.24 also hold if we replace $\hat{e}_{\text{switch}}$ throughout with $\hat{e}_{\text{divide}}$, including in Assumption 5, since the same bound from Hoeffding's inequality can be used for both $\hat{e}_{\text{switch}}$ and $\hat{e}_{\text{divide}}$.

Step 4: Combine results to show Eq 4.3

Finally, we connect the above results to show Eq 4.3. We know from Eq B.20 that Eq B.18 holds with high probability. Eq B.18 implies Eq B.21, which bounds $MCR_+(\epsilon) = MR(f_{+,\epsilon})$ up to a bracketed residual term. We also know from Eq B.24 that, with high probability, the residual term in Eq B.21 is less than $\mathcal{Q}_1 = q\left(B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{n}}\right)$. Putting this together, we can show Eq 4.3:

$$\begin{aligned}
&\mathbb{P}\left(MCR_+(\epsilon) > \widehat{MCR}_+(\epsilon_1) + \mathcal{Q}_1\right) \\
&\quad = \mathbb{P}\left(MR(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_1) + \mathcal{Q}_1\right) \\
&\quad \leq \mathbb{P}\left[\left(\widehat{MR}(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_1)\right) \bigcup \left(MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon}) > \mathcal{Q}_1\right)\right] \quad \text{from Step 2} \\
&\quad \leq \mathbb{P}\left[\widehat{MR}(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_1)\right] + \mathbb{P}\left[MR(f_{+,\epsilon}) - \widehat{MR}(f_{+,\epsilon}) > \mathcal{Q}_1\right] \\
&\quad \leq \frac{\delta}{3} + \frac{2\delta}{3} = \delta. \qquad \text{from Steps 1 \& 3}
\end{aligned} \tag{B.25}$$

This completes the proof for Eq 4.3. For Eq 4.4 we can use the same approach, shown below for completeness. Analogous to Eq B.20, we have

$$\mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) < \widehat{MCR}_-(\epsilon_1)\right] \leq \frac{\delta}{3}.$$

Analogous to Eq B.21, when $\widehat{MR}(f_{-,\epsilon}) \geq \widehat{MR}(\hat{f}_{-,\epsilon_1})$ we have

$$\begin{aligned}
\widehat{MR}(f_{-,\epsilon}) &\geq \widehat{MCR}_-(\epsilon_1) \\
\widehat{MR}(f_{-,\epsilon}) &\geq \widehat{MCR}_-(\epsilon_1) + \{MR(f_{-,\epsilon}) - MR(f_{-,\epsilon})\} \\
MR(f_{-,\epsilon}) &\geq \widehat{MCR}_-(\epsilon_1) - \left[\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon})\right].
\end{aligned}$$

Analogous to Eq B.24, we have

$$\mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon}) > q\left(B_{\text{ind}} \sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right] \leq \frac{2\delta}{3}. \tag{B.26}$$

Finally, analogous to Eq B.25, we have

$$\begin{aligned}
&\mathbb{P}\left(MCR_-(\epsilon) < \widehat{MCR}_-(\epsilon_1) - \mathcal{Q}_1\right) \\
&\quad \leq \mathbb{P}\left[\left(\widehat{MR}(f_{-,\epsilon}) < \widehat{MCR}_-(\epsilon_1)\right) \bigcup \left(\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon}) > q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right)\right] \qquad \text{(B.27)} \\
&\quad \leq \mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) < \widehat{MCR}_-(\epsilon_1)\right] + \mathbb{P}\left[\widehat{MR}(f_{-,\epsilon}) - MR(f_{-,\epsilon}) > q\left(B_{\text{ind}}\sqrt{\frac{\log(6\delta^{-1})}{n}}\right)\right] \\
&\quad \leq \frac{\delta}{3} + \frac{2\delta}{3} = \delta.
\end{aligned}$$

Again, the same proof holds without Assumption 5 if we replace $\mathcal{Q}_1$ with $\mathcal{Q}_{1,\text{difference}}$, and redefine q as the function with $q(k) = q_{\text{difference}}\left(k, \frac{k}{\sqrt{2}}\right)$ in Eqs B.26 & B.27. □

B.6 Proof of Corollary 7

Proof. By definition, $MR(f_{-,\epsilon_2}) \leq MR(f^\star) \leq MR(f_{+,\epsilon_2})$. Applying this relation in Line B.28, below, we see

$$\begin{aligned}
&\mathbb{P}\left(MR(f^\star) \in \left[\widehat{MCR}_-(\epsilon_2) - \mathcal{Q}_2, \quad \widehat{MCR}_+(\epsilon_2) + \mathcal{Q}_2\right]\right) \\
&\quad = 1 - \mathbb{P}\left\{MR(f^\star) < \widehat{MCR}_-(\epsilon_2) - \mathcal{Q}_2 \bigcup MR(f^\star) > \widehat{MCR}_+(\epsilon_2) + \mathcal{Q}_2\right\} \\
&\quad \geq 1 - \mathbb{P}\left\{MR(f^\star) < \widehat{MCR}_-(\epsilon_2) - \mathcal{Q}_2\right\} - \mathbb{P}\left\{MR(f^\star) > \widehat{MCR}_+(\epsilon_2) + \mathcal{Q}_2\right\} \\
&\quad \geq 1 - \mathbb{P}\left\{MR(f_{-,\epsilon}) < \widehat{MCR}_-(\epsilon_2) - \mathcal{Q}_2\right\} - \mathbb{P}\left\{MR(f_{+,\epsilon}) > \widehat{MCR}_+(\epsilon_2) + \mathcal{Q}_2\right\} \qquad \text{(B.28)} \\
&\quad \geq 1 - \frac{\delta}{2} - \frac{\delta}{2} \qquad \text{from Theorem 6.} \qquad \text{(B.29)}
\end{aligned}$$

To apply Theorem 6 in Line B.29, above, we note that $\mathcal{Q}_2$ and ϵ_2 are equivalent to the definitions of $\mathcal{Q}_1$ and ϵ_1 in Theorem 6, but with δ replaced by $\frac{\delta}{2}$.

Alternatively, if we have defined model reliance as $MR(f) = e_{\text{switch}}(f) - e_{\text{orig}}(f)$ (see Section A.5), with $\widehat{MR}(f) = \hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$, and

$$\mathcal{Q}_{2,\text{difference}} := \left(1 + \frac{1}{\sqrt{2}}\right) B_{\text{ind}}\sqrt{\frac{\log(12\delta^{-1})}{n}},$$

then $\mathcal{Q}_{2,\text{difference}}$ is similarly equivalent to $\mathcal{Q}_{1,\text{difference}}$ but with δ replaced with $\frac{\delta}{2}$. Under this difference-based definition of model reliance, Theorem 6 holds without Assumption 5 if we replace $\mathcal{Q}_1$ with $\mathcal{Q}_{1,\text{difference}}$ (see Section B.5), and so we can apply this altered version of Theorem 6 in Line B.29. Thus, Theorem 7 also holds without Assumption 5, under the difference-based definition of model reliance, with $\mathcal{Q}_2$ replaced by $\mathcal{Q}_{2,\text{difference}}$. □

B.7 Proof of Theorem 9

B.7.1 Proof of Theorem 8, and other limits on estimation error based on covering number

The following theorem uses the covering number based on r-margin-expectation-covers to jointly bound empirical losses for any function $f \in \mathcal{F}$. Theorem 8 in the main text follows directly from Eq B.33, below.

Theorem 27. *If Assumptions 3, 4 and 5 hold for all $f \in \mathcal{F}$, then for any $r > 0$*

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}}|\hat{e}_{orig}(f) - e_{orig}(f)| > B_{ind}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right] \leq \delta, \tag{B.30}$$

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}}\left|\hat{\mathbb{E}}\tilde{L}(f,Z) - \mathbb{E}\tilde{L}(f,Z)\right| > 2B_{ref}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right] \leq \delta, \tag{B.31}$$

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}}|\hat{e}_{switch}(f) - e_{switch}(f)| > B_{ind}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{n}} + 2r\right] \leq \delta, \tag{B.32}$$

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{\hat{e}_{orig}(f)}{\hat{e}_{switch}(f)} - \frac{e_{orig}(f)}{e_{switch}(f)}\right| > \mathcal{Q}_4\right] \leq \delta, \tag{B.33}$$

$$\mathbb{P}_D\left[\sup_{f\in\mathcal{F}}|\{\hat{e}_{switch}(f) - \hat{e}_{orig}(f)\} - \{e_{switch}(f) - e_{orig}(f)\}| > \mathcal{Q}_{4,difference}\right] \leq \delta, \tag{B.34}$$

where

$$\mathcal{Q}_4 := q_{ratio}\left(B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r\sqrt{2}))}{n}} + 2r\sqrt{2}, B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right), \tag{B.35}$$

$$\mathcal{Q}_{4,difference} := q_{difference}\left(B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r\sqrt{2}))}{n}} + 2r\sqrt{2}, B_{ind}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right), \tag{B.36}$$

and q_{ratio} and $q_{difference}$ are defined as in Lemma 26. For Eq B.34, the result is unaffected if we remove Assumption 5.

Proof of Eq B.30

Proof. Let $\mathcal{G}_r$ be a r-margin-expectation-cover for $\mathcal{F}$ of size $\mathcal{N}(\mathcal{F},r)$. Let D_p denote the population distribution, let D_s be the sample distribution, and let $D^\star$ be the uniform mixture of D_p and D_s, i.e., for any $z \in \mathcal{Z}$,

$$\mathbb{P}_{D^\star}(Z \leq z) = \frac{1}{2}\mathbb{P}_{D_p}(Z \leq z) + \frac{1}{2}\mathbb{P}_{D_s}(Z \leq z). \tag{B.37}$$

Unless otherwise stated, we take expectations and probabilities with respect to D_p. Since $\mathcal{G}_r$ is a r-margin-expectation-cover, we know that for any $f \in \mathcal{F}$ we can find a function $g \in \mathcal{G}_r$ such that $\mathbb{E}_{D^\star}|L(g,Z) - L(f,Z)| = \mathbb{E}_{D^\star}\left|\tilde{L}(g,Z) - \tilde{L}(f,Z)\right| \leq r$, and

$$\begin{aligned}
\left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z)\right| &= \left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z) + \left\{\hat{\mathbb{E}}L(g,Z) - \hat{\mathbb{E}}L(g,Z)\right\} + \{\mathbb{E}L(g,Z) - \mathbb{E}L(g,Z)\}\right| \qquad \text{(B.38)}\\
&\leq \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + |\mathbb{E}L(g,Z) - \mathbb{E}L(f,Z)| + \left|\hat{\mathbb{E}}L(f,Z) - \hat{\mathbb{E}}L(g,Z)\right|\\
&\leq \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + \mathbb{E}_{D_p}|L(g,Z) - L(f,Z)| + \mathbb{E}_{D_s}|L(f,Z) - L(g,Z)|\\
&= \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + 2\mathbb{E}_{D^\star}|L(g,Z) - L(f,Z)|\\
&\leq \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + 2r.
\end{aligned}$$

Applying the above relation in Line B.39 below, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right)\\
&\quad= \mathbb{P}\left(\exists f\in\mathcal{F}:\ \left|\hat{\mathbb{E}}L(f,Z) - \mathbb{E}L(f,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right)\\
&\quad\leq \mathbb{P}\left(\exists g\in\mathcal{G}_r:\ \left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| + 2r > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right) \qquad \text{(B.39)}\\
&\quad= \mathbb{P}\left(\bigcup_{g\in\mathcal{G}_r}\left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}\right)\\
&\quad\leq \sum_{g\in\mathcal{G}_r}\mathbb{P}\left(\left|\hat{\mathbb{E}}L(g,Z) - \mathbb{E}L(g,Z)\right| > B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}\right) \qquad \text{from the Union bound}\\
&\quad\leq \mathcal{N}(\mathcal{F},r)\,2\exp\left[-\frac{2n}{(B_{\text{ind}})^2}\left\{B_{\text{ind}}\sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}\right\}^2\right] \qquad \text{from Hoeffding's inequality}\\
&\qquad\qquad \text{(B.40)}\\
&\quad= \delta. \qquad \text{(B.41)}
\end{aligned}$$

To apply Hoeffding's inequality in Line B.40, above, we use the fact that $L(g,Z)$ is bounded within an interval of length B_{ind}. □

Proof of Eq B.31

Proof. The proof for Eq B.31 is nearly identical to the proof for Eq B.30. Simply replacing L and B_{ind} respectively with $\tilde{L}$ and $(2B_{\text{ref}})$ in Eqs B.38-B.41 yields a valid proof for Eq B.31. □

Proof of Eq B.32

Proof. Let F_D denote the cumulative distribution function for a distribution D. Let $\tilde{D}_p$ be the distribution such that

$$F_{\tilde{D}_p}(Y = y, X_1 = x_1, X_2 = x_2) = F_{D_p}(Y = y, X_2 = x_2)F_{D_p}(X_1 = x_1).$$

Let $\tilde{D}_s$ be the distribution satisfying

$$\mathbb{P}_{\tilde{D}_s}(Y = y, X_1 = x_1, X_2 = x_2) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}(\mathbf{y}_{[j]} = y, \mathbf{X}_{1[i,\cdot]} = x_1, \mathbf{X}_{2[j,\cdot]} = x_2).$$

Let $\tilde{D}^\star$ be the uniform mixture of $\tilde{D}_p$ and $\tilde{D}_s$, as in Eq B.37. Replacing e_{orig}, $\hat{e}_{\text{orig}}$, D_p, D_s, and $D^\star$ respectively with e_{switch}, $\hat{e}_{\text{switch}}$, $\tilde{D}_p$, $\tilde{D}_s$, and $\tilde{D}^\star$, we can follow the same steps as in the proof for Eq B.30. For any $f \in \mathcal{F}$, we know that there exists a function $g \in \mathcal{G}_r$ satisfying $\mathbb{E}_{\tilde{D}^\star} |L(g, Z) - L(f, Z)| \leq r$, which implies

$$\begin{aligned}
|\hat{e}_{\text{switch}}(f) - e_{\text{switch}}(f)| &= |\hat{e}_{\text{switch}}(f) - e_{\text{switch}}(f) + \{\hat{e}_{\text{switch}}(g) - \hat{e}_{\text{switch}}(g)\} + \{e_{\text{switch}}(g) - e_{\text{switch}}(g)\}| \\
&\leq |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + \mathbb{E}_{\tilde{D}_p} |L(g, Z) - L(f, Z)| + \mathbb{E}_{\tilde{D}_s} |L(f, Z) - L(g, Z)| \\
&= |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + 2\mathbb{E}_{\tilde{D}^\star} |L(g, Z) - L(f, Z)| \\
&\leq |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + 2r.
\end{aligned}$$

As a result,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{e}_{\text{switch}}(f) - e_{\text{switch}}(f)| > B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} + 2r \right) \\
&\quad \leq \mathbb{P}\left(\exists g \in \mathcal{G}_r : \ |\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| + 2r > B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} + 2r \right) \\
&\quad \leq \sum_{g \in \mathcal{G}_r} \mathbb{P}\left(|\hat{e}_{\text{switch}}(g) - e_{\text{switch}}(g)| > B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} \right) \\
&\quad \leq \mathcal{N}(\mathcal{F}, r)\, 2 \exp\left[-\frac{n}{(B_{\text{ind}} - 0)^2} \left\{ B_{\text{ind}} \sqrt{\frac{\log(2\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{n}} \right\}^2 \right] \\
&\quad = \delta.
\end{aligned}$$

□

Proof for Eq B.33

Proof. We apply Lemma 26 and Eq B.35 in Line B.43, below, to obtain

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{\hat{e}_{\text{orig}}(f)}{\hat{e}_{\text{switch}}(f)}-\frac{e_{\text{orig}}(f)}{e_{\text{switch}}(f)}\right|>\mathcal{Q}_4\right] \tag{B.42}$$

$$=\mathbb{P}\left(\exists f\in\mathcal{F}:\ \left|\frac{\hat{e}_{\text{orig}}(f)}{\hat{e}_{\text{switch}}(f)}-\frac{e_{\text{orig}}(f)}{e_{\text{switch}}(f)}\right|>\mathcal{Q}_4\right)$$

$$\leq\mathbb{P}\left(\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{orig}}(f)-e_{\text{orig}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}+2r\right\}\right. \tag{B.43}$$

$$\left.\bigcup\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{switch}}(f)-e_{\text{switch}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}}+2r\sqrt{2}\right\}\right)$$

$$=\mathbb{P}\left(\sup_{f\in\mathcal{F}}|\hat{e}_{\text{orig}}(f)-e_{\text{orig}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}+2r\right)$$

$$+\mathbb{P}\left(\sup_{f\in\mathcal{F}}|\hat{e}_{\text{switch}}(f)-e_{\text{switch}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}}+2r\sqrt{2}\right)$$

$$\leq\frac{\delta}{2}+\frac{\delta}{2}. \qquad \text{from Eqs B.30 and B.32} \tag{B.44}$$

□

Proof for Eq B.34

Proof. Finally, to show B.34, we apply the same steps as in Eqs B.42 through B.44. We apply Eq B.36 & Lemma 26 to obtain

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}|\{\hat{e}_{\text{switch}}(f)-\hat{e}_{\text{orig}}(f)\}-\{e_{\text{switch}}(f)-e_{\text{orig}}(f)\}|>\mathcal{Q}_{4,\text{difference}}\right]$$

$$\leq\mathbb{P}\left(\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{orig}}(f)-e_{\text{orig}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}}+2r\right\}\right.$$

$$\left.\bigcup\left\{\exists f\in\mathcal{F}:\ |\hat{e}_{\text{switch}}(f)-e_{\text{switch}}(f)|>B_{\text{ind}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}\left(\mathcal{F},r\sqrt{2}\right))}{n}}+2r\sqrt{2}\right\}\right)$$

$$\leq\frac{\delta}{2}+\frac{\delta}{2}.$$

□

B.7.2 Implementing Theorem 27 to show Theorem 9

Proof. Consider the event that

$$\exists\hat{f}_{+,\epsilon_3}\in\underset{f\in\hat{\mathcal{R}}(\epsilon_3)}{\arg\max}\,\widehat{MR}(f)\text{ such that }MCR_+(\epsilon)<MR\left(\hat{f}_{+,\epsilon_3}\right). \tag{B.45}$$

A brief outline of our proof for Eq 4.7 is as follows. We expect Eq B.45 to be unlikely due to the fact that $\epsilon_3 < \epsilon$. If Eq B.45 does not hold, then the only way that $MCR_+(\epsilon) < \widehat{MCR}_+(\epsilon_3) - \mathcal{Q}_3$ holds is if there exists $\hat{f}_{+,\epsilon_3} \in \arg\max_{f \in \hat{\mathcal{R}}(\epsilon_3)} \widehat{MR}(f)$ which has an empirical MR that differs from its population-level MR by at least $\mathcal{Q}_3$.

To show that Eq B.45 is unlikely, we apply Theorem 27:

$$
\begin{aligned}
&\mathbb{P}\left(\exists \hat{f}_{+,\epsilon_3} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_3)}{\arg\max}\, \widehat{MR}(f) :\, MCR_+(\epsilon) < MR\left(\hat{f}_{+,\epsilon_3}\right)\right) \\
&\leq \mathbb{P}\left(\exists f \in \hat{\mathcal{R}}(\epsilon_3) :\, MCR_+(\epsilon) < MR\left(f\right)\right) \\
&= \mathbb{P}\left(\exists f \in \hat{\mathcal{R}}(\epsilon_3) \backslash f_{\text{ref}} :\, MCR_+(\epsilon) < MR\left(f\right)\right) && \text{by } MCR_+(\epsilon) \geq MR(f_{\text{ref}}) \\
&\leq \mathbb{P}\left(\exists f \in \hat{\mathcal{R}}(\epsilon_3) \backslash f_{\text{ref}} :\, \mathbb{E}\tilde{L}(f,Z) > \epsilon\right) && \text{by } MCR_+(\epsilon) \text{ Def} \\
&= \mathbb{P}\left(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f,Z) > \epsilon :\, \hat{\mathbb{E}}\tilde{L}(f,Z) \leq \epsilon_3\right) && \text{by } \hat{\mathcal{R}}(\epsilon) \text{ Def} \\
&= \mathbb{P}\Bigg(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f,Z) > \epsilon : \\
&\qquad \hat{\mathbb{E}}\tilde{L}(f,Z) - \epsilon \leq -2B_{\text{ref}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} - 2r\Bigg) && \text{by } \epsilon_3 \text{ Def} \quad \text{(B.46)} \\
&\leq \mathbb{P}\Bigg(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f,Z) > \epsilon : \\
&\qquad \hat{\mathbb{E}}\tilde{L}(f,Z) - \mathbb{E}\tilde{L}(f,Z) \leq -2B_{\text{ref}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} - 2r\Bigg) \\
&\leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}\tilde{L}(f,Z) - \mathbb{E}\tilde{L}(f,Z)| \geq 2B_{\text{ref}}\sqrt{\frac{\log(4\delta^{-1}\mathcal{N}(\mathcal{F},r))}{2n}} + 2r\right) \\
&= \frac{\delta}{2} && \text{by Thm 27.} \\
& && \text{(B.47)}
\end{aligned}
$$

If Eq B.45 does not hold, we have

$$
\begin{aligned}
MCR_+(\epsilon) &\geq MR\left(\hat{f}_{+,\epsilon_3}\right) && \text{for all } \hat{f}_{+,\epsilon_3} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_3)}{\arg\max}\, \widehat{MR}(f) \\
&= \widehat{MR}\left(\hat{f}_{+,\epsilon_3}\right) - \left\{\widehat{MR}\left(\hat{f}_{+,\epsilon_3}\right) - MR\left(\hat{f}_{+,\epsilon_3}\right)\right\} && \text{for all } \hat{f}_{+,\epsilon_3} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_3)}{\arg\max}\, \widehat{MR}(f) \\
&= \widehat{MCR}_+(\epsilon_3) - \left\{\widehat{MR}\left(\hat{f}_{+,\epsilon_3}\right) - MR\left(\hat{f}_{+,\epsilon_3}\right)\right\} && \text{for all } \hat{f}_{+,\epsilon_3} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_3)}{\arg\max}\, \widehat{MR}(f) \\
&\geq \widehat{MCR}_+(\epsilon_3) - \sup_{f \in \mathcal{F}} |\widehat{MR}(f) - MR(f)|. && \text{(B.48)}
\end{aligned}
$$

Let q_{ratio} and $q_{\text{difference}}$ be defined as in Lemma 26. Then

$$\mathcal{Q}_3 = \frac{B_{\text{switch}}}{b_{\text{orig}}} - \frac{B_{\text{switch}} - \left\{ B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}(\mathcal{F}, r\sqrt{2}))}{n}} + 2r\sqrt{2} \right\}}{b_{\text{orig}} + \left\{ B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}(\mathcal{F}, r))}{2n}} + 2r \right\}}$$
$$= q_{\text{ratio}} \left(B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + 2r\sqrt{2}, B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\right))}{2n}} + 2r \right) \tag{B.49}$$

Theorem 27 implies that the sup term in Eq B.48 is less than $\mathcal{Q}_3$ with probability at least $1 - \frac{\delta}{2}$. Now, examining the left-hand side of Eq 4.7, we see

$$\begin{aligned}
&\mathbb{P}\left(MCR_+(\epsilon) < \widehat{MCR}_+(\epsilon_3) - \mathcal{Q}_3 \right) \\
&\quad \leq \mathbb{P}\left[\left\{ \exists \hat{f}_{+,\epsilon_3} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_3)}{\arg\max}\, \widehat{MR}(f) : MCR_+(\epsilon) < MR\left(\hat{f}_{+,\epsilon_3}\right) \right\} \right. \\
&\qquad\qquad \left. \bigcup \left\{ \sup_{f \in \mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_3 \right\} \right] \qquad \text{from Eq B.48} \\
&\quad \leq \mathbb{P}\left[\exists \hat{f}_{+,\epsilon_3} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_3)}{\arg\max}\, \widehat{MR}(f) : MCR_+(\epsilon) < MR\left(\hat{f}_{+,\epsilon_3}\right) \right] \\
&\qquad\qquad + \mathbb{P}\left[\sup_{f \in \mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_3 \right] \qquad \text{from the Union bound} \\
&\quad = \frac{\delta}{2} + \frac{\delta}{2} \qquad \text{from Eq B.47, Eq B.49, \& Theorem 27.}
\end{aligned} \tag{B.50}$$

This completes the proof for Eq 4.7.

Alternatively, if we have defined model reliance as $MR(f) = e_{\text{switch}}(f) - e_{\text{orig}}(f)$ (see Section A.5), with $\widehat{MR}(f) = \hat{e}_{\text{switch}}(f) - \hat{e}_{\text{orig}}(f)$, and

$$\mathcal{Q}_{3,\text{difference}} = B_{\text{ind}} \left\{ \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\right))}{2n}} \right\} + 2r(\sqrt{2}+1)$$
$$= q_{\text{difference}} \left(B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\sqrt{2}\right))}{n}} + 2r\sqrt{2}, B_{\text{ind}} \sqrt{\frac{\log(8\delta^{-1}\mathcal{N}\left(\mathcal{F}, r\right))}{2n}} + 2r \right),$$

then same proof of Eq 4.7 holds without Assumption 5 if we replace $\mathcal{Q}_3$ with $\mathcal{Q}_{3,\text{difference}}$, and apply Eq B.34 in Eq B.50.

For Eq 4.8 we can use the same approach. Consider the event that

$$\exists \hat{f}_{-,\epsilon_3} \in \underset{f \in \hat{\mathcal{R}}(\epsilon_3)}{\arg\min}\, \widehat{MR}(f) : MCR_-(\epsilon) > MR\left(\hat{f}_{-,\epsilon_3}\right). \tag{B.51}$$

Applying steps analogous to those used to derive Eq B.47, we have

$$\mathbb{P}\left(\exists \hat{f}_{-,\epsilon_3} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_3)}{\arg\min}\, \widehat{MR}(f) :\, MCR_-(\epsilon) > MR\left(\hat{f}_{-,\epsilon_3}\right)\right)$$
$$\leq \quad \mathbb{P}\left(\exists f \in \mathcal{F}, \mathbb{E}\tilde{L}(f,Z) > \epsilon : \hat{\mathbb{E}}\tilde{L}(f,Z) \leq \epsilon_3\right) \quad \leq \quad \frac{\delta}{2}.$$

Analogous to B.48, when Eq B.51 does not hold, we have have

$$\begin{aligned}
MCR_-(\epsilon) &\leq MR\left(\hat{f}_{-,\epsilon_3}\right) && \text{for all } \hat{f}_{-,\epsilon_3} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_3)}{\arg\min}\, \widehat{MR}(f)\\
&= \widehat{MR}\left(\hat{f}_{-,\epsilon_3}\right) + \left\{MR\left(\hat{f}_{-,\epsilon_3}\right) - \widehat{MR}\left(\hat{f}_{-,\epsilon_3}\right)\right\} && \text{for all } \hat{f}_{-,\epsilon_3} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_3)}{\arg\min}\, \widehat{MR}(f)\\
&= \widehat{MCR}_-\left(\epsilon_3\right) + \left\{MR\left(\hat{f}_{-,\epsilon_3}\right) - \widehat{MR}\left(\hat{f}_{-,\epsilon_3}\right)\right\} && \text{for all } \hat{f}_{-,\epsilon_3} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_3)}{\arg\min}\, \widehat{MR}(f)\\
&\leq \widehat{MCR}_-\left(\epsilon_3\right) + \sup_{f\in\mathcal{F}} |MR(f) - \widehat{MR}(f)|
\end{aligned}$$

Finally, analogous to Eq B.50,

$$\begin{aligned}
&\mathbb{P}\left(MCR_-(\epsilon) > \widehat{MCR}\left(\epsilon_3\right) + \mathcal{Q}_3\right)\\
&\quad\leq \mathbb{P}\left[\left\{\exists \hat{f}_{-,\epsilon_3} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_3)}{\arg\min}\, \widehat{MR}(f) :\, MCR_-(\epsilon) > MR\left(\hat{f}_{-,\epsilon_3}\right)\right\}\right.\\
&\qquad\qquad \left.\bigcup \left\{\sup_{f\in\mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_3\right\}\right]\\
&\quad\leq \mathbb{P}\left[\exists \hat{f}_{-,\epsilon_3} \in \underset{f\in\hat{\mathcal{R}}(\epsilon_3)}{\arg\min}\, \widehat{MR}(f) :\, MCR_-(\epsilon) > MR\left(\hat{f}_{-,\epsilon_3}\right)\right]\\
&\qquad\qquad +\mathbb{P}\left[\sup_{f\in\mathcal{F}} |\widehat{MR}(f) - MR(f)| > \mathcal{Q}_3\right]\\
&\quad= \frac{\delta}{2} + \frac{\delta}{2}.
\end{aligned} \tag{B.52}$$

Under the difference-based definition of model reliance (see Section A.5), the same proof for Eq 4.8 holds without Assumption 5 if we replace $\mathcal{Q}_3$ with $\mathcal{Q}_{3,\text{difference}}$, and apply Eq B.34 in Eq B.52. □

B.8 Proof of Proposition 10

Proof. For a given value of r, let $Q = \left\lceil \frac{c}{2r} \right\rceil$. Let $I_1, \ldots I_Q$ be the intervals $I_j = [\frac{j-1}{Q}, \frac{j}{Q}]$, with midpoints denoted by m_j. These intervals form a partition of the $[0,1]$ interval. Let $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q\}$ be the set of vectors with $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{\text{orig}} + m_j(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_{\text{orig}})$, and let $\mathcal{G} := \{f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(x) = x'\boldsymbol{\theta};\, \boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q\}\}$. It suffices to show that $\mathcal{G}$ is a *r-margin-expectation-cover* for $\mathcal{F}_s$.

For any function $f_{\tilde{\boldsymbol{\theta}}} \in \mathcal{F}_s$, there exists $\tilde{w} \in [0,1]$ such that

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \tilde{w}\boldsymbol{\theta}_{\text{orig}} + (1-\tilde{w})\boldsymbol{\theta}_+ \\ &= \boldsymbol{\theta}_{\text{orig}} + (1-\tilde{w})(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_{\text{orig}}).\end{aligned}$$

Let $\tilde{j}$ be the index such that $(1-\tilde{w}) \in I_{\tilde{j}}$. Since $I_{\tilde{j}}$ has length $\frac{1}{Q}$, each point in $I_{\tilde{j}}$ is within $\frac{1}{2Q}$ of the midpoint $m_{\tilde{j}}$, and so $|(1-\tilde{w}) - m_{\tilde{j}}| \leq \frac{1}{2Q}$. Now, for any $x \in (\mathcal{X}_1 \times \mathcal{X}_2)$ and $y \in \{-1, 1\}$,

$$\begin{aligned}\left|L\left(f_{\tilde{\boldsymbol{\theta}}}, (y,x)\right) - L\left(f_{\boldsymbol{\theta}_{\tilde{j}}}, (y,x)\right)\right| &= \left|\left(\delta - yx'\tilde{\boldsymbol{\theta}}\right)_+ - \left(\delta - yx'\boldsymbol{\theta}_{\tilde{j}}\right)_+\right| \\ &\leq \left|\delta - yx'\tilde{\boldsymbol{\theta}} - \left(\delta - yx'\boldsymbol{\theta}_{\tilde{j}}\right)\right|.\end{aligned}$$

The last line above holds since $|(a)_+ - (b)_+| \leq |a-b|$ for any $a, b \in \mathbb{R}$. Thus

$$\begin{aligned}\left|L\left(f_{\tilde{\boldsymbol{\theta}}}, (y,x)\right) - L\left(f_{\boldsymbol{\theta}_{\tilde{j}}}, (y,x)\right)\right| &\leq \left|\delta - yx'\tilde{\boldsymbol{\theta}} - \left(\delta - yx'\boldsymbol{\theta}_{\tilde{j}}\right)\right| \\ &= \left|-yx'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\tilde{j}})\right| \\ &= \left|x'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\tilde{j}})\right| \\ &= \left|x'\left[\{\boldsymbol{\theta}_{\text{orig}} + (1-\tilde{w})(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_{\text{orig}})\} - \left\{\boldsymbol{\theta}_{\text{orig}} + m_{\tilde{j}}(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_{\text{orig}})\right\}\right]\right| \\ &= \left|\left[(1-\tilde{w}) - m_{\tilde{j}}\right] x'(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_{\text{orig}})\right| \\ &= \left|(1-\tilde{w}) - m_{\tilde{j}}\right| \left|x'(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_{\text{orig}})\right| \\ &= \left|(1-\tilde{w}) - m_{\tilde{j}}\right| \left|f_{\boldsymbol{\theta}_+}(x) - f_{\boldsymbol{\theta}_{\text{orig}}}(x)\right| \\ &\leq \frac{1}{2Q}c \\ &\leq r.\end{aligned}$$

It follows from the monotonicity property of expectations that $\mathbb{E}_{Z\sim D}\left|L\left(f_{\tilde{\boldsymbol{\theta}}}, Z\right) - L\left(f_{\boldsymbol{\theta}_{\tilde{j}}}, Z\right)\right| \leq r$ for any distribution D, and so $\mathcal{G}$ is a r-margin-expectation-cover for $\mathcal{F}_s$. □

B.9 Proof of Propositions 11 and 12

First we show that the result of Proposition 11 holds not only for $f^\star$, but also for any individual function $f_1 \in \mathcal{R}(0)$ satisfying Assumption 4.

Lemma 28. *Let ϵ_4, $\hat{a}_-(\epsilon_4)$, and $\hat{a}_+(\epsilon_4)$ be defined as in Proposition 11. Given a function $f_1 \in \mathcal{R}(0)$, if Assumption 4 holds for f_1, then*

$$\mathbb{P}\{\phi(f_1) \in [\hat{a}_-(\epsilon_4), \hat{a}_+(\epsilon_4)]\} \geq 1 - \delta.$$

Proof. Consider the event that

$$\phi(f_1) \in [\hat{a}_-(\epsilon_4), \hat{a}_+(\epsilon_4)]. \tag{B.53}$$

Eq B.53 will always hold if $f_1 \in \hat{\mathcal{R}}(\epsilon_4)$, since the interval $[\hat{a}_-(\epsilon_4), \hat{a}_+(\epsilon_4)]$ contains $\phi(f)$ for any $f \in \hat{\mathcal{R}}(\epsilon_4)$ by definition. Thus,

$$\begin{aligned}\mathbb{P}\left\{\phi(f_1) \notin [\hat{a}_-(\epsilon_4), \hat{a}_+(\epsilon_4)]\right\} &\leq \mathbb{P}\left\{f_1 \notin \hat{\mathcal{R}}(\epsilon_4)\right\} \\ &\leq \delta \qquad \text{from Lemma 25.}\end{aligned}$$

□

Proof of Proposition 11

Proof. Since $f^\star \in \mathcal{R}(0)$, Proposition 11 follows immediately from Lemma 28. □

Proof of Proposition 12

Proof. Let $f_{-,\phi} \in \arg\min_{f \in \mathcal{R}(0)} \phi(f)$ and $f_{+,\phi} \in \arg\max_{f \in \mathcal{R}(0)} \phi(f)$ respectively denote functions that attain the lowest and highest values of $\phi(f)$ among models $f \in \mathcal{R}(0)$. Applying the definitions of $f_{-,\phi}$ and $f_{+,\phi}$ in Line B.54, below, we have

$$\begin{aligned}&\mathbb{P}\left(\left\{\phi(f) : f \in \mathcal{R}(0)\right\} \not\subset [\hat{a}_-(\epsilon_5), \hat{a}_+(\epsilon_5)]\right) \\ &\quad= \mathbb{P}\left([\phi(f_{-,\phi}), \phi(f_{+,\phi})] \not\subset [\hat{a}_-(\epsilon_5), \hat{a}_+(\epsilon_5)]\right) \qquad \text{(B.54)} \\ &\quad= \mathbb{P}\left(\phi(f_{-,\phi}) \notin [\hat{a}_-(\epsilon_5), \hat{a}_+(\epsilon_5)] \bigcup \phi(f_{+,\phi}) \notin [\hat{a}_-(\epsilon_5), \hat{a}_+(\epsilon_5)]\right) \\ &\quad\leq \mathbb{P}\left(\phi(f_{-,\phi}) \notin [\hat{a}_-(\epsilon_5), \hat{a}_+(\epsilon_5)]\right) + \mathbb{P}\left(\phi(f_{+,\phi}) \notin [\hat{a}_-(\epsilon_5), \hat{a}_+(\epsilon_5)]\right) \\ &\quad\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta \qquad \text{from Lemma 28, and the definition of } \epsilon_5.\end{aligned}$$

□

B.10 Absolute losses versus relative losses in definition of the Rashomon set

In this paper we primarily define Rashomon sets as the models that perform well *relative* to a reference model f_{ref}. We can also study an alternate formulation of Rashomon sets by replacing the relative loss $\tilde{L}$ with the non-standardized loss L throughout. This results in a new interpretation of the Rashomon set $\mathcal{R}(\epsilon_{\text{abs}}, f_{\text{ref}}, \mathcal{F}) = \{f_{\text{ref}}\} \cup \{f \in \mathcal{F} : \mathbb{E}L(f, Z) \leq \epsilon_{\text{abs}}\}$ as the union of f_{ref} and the subset of models with *absolute* loss L no higher than ϵ_{abs}, for $\epsilon_{\text{abs}} > 0$. The process of computing empirical MCR is largely unaffected by whether L or $\tilde{L}$ is used, as it is simple to transform from one optimization problem to the other.

We still require the explicit inclusion of f_{ref} in empirical and population-level Rashomon sets to ensure that they are nonempty. However, in many cases, this inclusion becomes redundant when interpreting the Rashomon set (e.g., when $\epsilon \geq 0$, and $\mathbb{E}L(f_{\text{ref}}, Z) \leq \epsilon_{\text{abs}}$).

Under the replacement of $\tilde{L}$ with L, we also replace Assumption 4 with Assumption 3 (whenever this is not redundant), and replace $2B_{\text{ref}}$ with B_{ind} in the definitions of ϵ_1, ϵ_2, ϵ_3, ϵ_4, and ϵ_5 in Theorem 6, Corollary 7, Theorem 9, and Propositions 11-12. This is because the motivation for the $2B_{\text{ref}}$ term is that $\tilde{L}(f_1)$ is bounded within an interval of length $2B_{\text{ref}}$ when f_1 satisfies Assumption 4. However, under Assumption 3, $L(f_1)$ is bounded within an interval of length B_{ind}.

C Proofs for computation results

All of the proofs in this section are unchanged if we replace $\hat{e}_{\text{orig}}(f)$ with $\hat{e}_{\text{divide}}(f)$ in our definitions of $\hat{h}_{-,\gamma}$, $\hat{h}_{+,\gamma}$, $\hat{g}_{-,\gamma}$, $\hat{g}_{+,\gamma}$, and $\widehat{MR}$. Throughout the following proofs, we will make use of the fact that, for constants $a, b, c, d \in \mathbb{R}$ satisfying $a \geq c$, the relation $a + b \leq c + d$ implies

$$\begin{aligned} a + b &\leq c + d \\ a - c &\leq d - b \\ 0 &\leq d - b && \text{since } 0 \leq a - c \\ b &\leq d. \end{aligned} \tag{C.1}$$

We also make use of the fact that for any $\gamma_1, \gamma_2 \in \mathbb{R}$, the definitions of $\hat{g}_{+,\gamma_1}$ and $\hat{g}_{-,\gamma_1}$ imply

$$\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) \leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}), \quad \text{and} \quad \hat{h}_{-,\gamma_1}(g_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_1}(g_{-,\gamma_2}). \tag{C.2}$$

Finally, for any two values $\gamma_1, \gamma_2 \in \mathbb{R}$, we make use of the fact that

$$\begin{aligned} \hat{h}_{+,\gamma_1}(f) &= \hat{e}_{\text{orig}}(f) + \gamma_1 \hat{e}_{\text{switch}}(f) \\ &= \hat{e}_{\text{orig}}(f) + \gamma_2 \hat{e}_{\text{switch}}(f) + \{\gamma_1 \hat{e}_{\text{switch}}(f) - \gamma_2 \hat{e}_{\text{switch}}(f)\} \\ &= \hat{h}_{+,\gamma_2}(f) + (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(f), \end{aligned} \tag{C.3}$$

and, by the same steps,

$$\hat{h}_{-,\gamma_1}(f) = \hat{h}_{-,\gamma_2}(f) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(f). \tag{C.4}$$

C.1 Proof of Lemma 14 (lower bound for MR)

Proof. **Part 1: Showing Eq 7.1 holds for all $f \in \mathcal{F}$ satisfying $\hat{e}_{\textbf{orig}}(f) \leq \epsilon_{\textbf{abs}}$.**

If $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \geq 0$, then for any function $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ we know that

$$\begin{aligned} \frac{1}{\epsilon_{\text{abs}}} &\leq \frac{1}{\hat{e}_{\text{orig}}(f)} \\ \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} &\leq \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(f)}. \end{aligned} \tag{C.5}$$

Now, for any $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$, the definition of $\hat{g}_{-,\gamma}$ implies that

$$\begin{aligned} \hat{h}_{-,\gamma}(f) &\geq \hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \\ \gamma \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) &\geq \hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) \\ \gamma + \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(f)} \\ \gamma + \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} && \text{from Eq C.5} \\ \widehat{MR}(f) &\geq \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} - \gamma. \end{aligned}$$

Part 2: Showing that if $f = \hat{g}_{-,\gamma}$, and at least one of the inequalities in Condition 13 holds with equality, then Eq 7.1 holds with equality.

We consider each of the two inequalities in Condition 13 separately. If $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma}) = 0$, then

$$\begin{aligned} 0 &= \gamma \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma}) \\ \frac{-\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} &= \gamma. \end{aligned}$$

As a result

$$\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} - \gamma = \frac{0}{\epsilon_{\text{abs}}} - \left\{ \frac{-\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} \right\} = \widehat{MR}(\hat{g}_{-,\gamma}).$$

Alternatively, if $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) = \epsilon_{\text{abs}}$, then

$$\frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\text{abs}}} - \gamma = \frac{\gamma \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} - \gamma = \gamma + \frac{\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma})} - \gamma = \widehat{MR}(\hat{g}_{-,\gamma}).$$

□

C.2 Proof of Lemma 15 (monotonicity for MR lower bound binary search)

Proof. **Part 1: $\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})$ is monotonically increasing in γ.**

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2$. We have assumed that $0 < \hat{e}_{\text{orig}}(f)$ for any $f \in \mathcal{F}$. Thus, for any $f \in \mathcal{F}$ we have

$$\begin{aligned} \gamma_1 \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) &< \gamma_2 \hat{e}_{\text{orig}}(f) + \hat{e}_{\text{switch}}(f) \\ \hat{h}_{-,\gamma_1}(f) &< \hat{h}_{-,\gamma_2}(f). \end{aligned} \tag{C.6}$$

Applying this, we have

$$\begin{aligned} \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) &\leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.2} \\ &\leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.6.} \end{aligned}$$

This result is analogous to Lemma 3 from Dinkelbach (1967).

Part 2: $\hat{e}_{\textbf{orig}}(\hat{g}_{-,\gamma})$ is monotonically decreasing in γ.

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2$. Then

$$\begin{aligned} \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) &\leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.2} \\ \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_1}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) &\leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.4} \\ (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) &\leq (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) && \text{from Eqs C.1 \& C.2} \\ \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) &\geq \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}). \end{aligned}$$

Part 3: $\left\{ \frac{\hat{h}_{-,\gamma}(\hat{g}_{-,\gamma})}{\epsilon_{\textbf{abs}}} - \gamma \right\}$ is monotonically decreasing in γ in the range where $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma}) \leq \epsilon_{\text{abs}}$, and increasing otherwise.

Suppose $\gamma_1 < \gamma_2$ and $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) \leq \epsilon_{\text{abs}}$. Then, from Eq C.2,

$$\begin{aligned}
&\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_1}) \\
&\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) \leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) + (\gamma_2 - \gamma_1)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) && \text{from Eq C.4} \\
&\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) \leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) + (\gamma_2 - \gamma_1)\epsilon_{\text{abs}} && \text{from } \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) \leq \epsilon_{\text{abs}} \\
&\frac{\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})}{\epsilon_{\text{abs}}} - \gamma_2 \leq \frac{\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1})}{\epsilon_{\text{abs}}} - \gamma_1.
\end{aligned}$$

Similarly, if $\gamma_1 < \gamma_2$ and $\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) \geq \epsilon_{\text{abs}}$. Then, from Eq C.2

$$\begin{aligned}
&\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_2}) \\
&\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) && \text{from Eq C.4} \\
&\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1}) \leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) + (\gamma_1 - \gamma_2)\epsilon_{\text{abs}} && \text{from } \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}) \geq \epsilon_{\text{abs}} \\
&\frac{\hat{h}_{-,\gamma_1}(\hat{g}_{-,\gamma_1})}{\epsilon_{\text{abs}}} - \gamma_1 \leq \frac{\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2})}{\epsilon_{\text{abs}}} - \gamma_2.
\end{aligned}$$

□

C.3 Proof of Proposition 17 (convexity for MR lower bound binary search)

Proof. Let $\gamma_1 := \frac{1}{n-1}$. First we show that there exists a function $\hat{g}_{-,\gamma_1}$ minimizing $\hat{h}_{-,\gamma_1}$ such that $\widehat{MR}(g_1) = 1$. Let D_s denote the sample distribution of the data, and let D_m be the distribution satisfying

$$\begin{aligned}
\mathbb{P}_{D_m}\{(Y, X_1, X_2) = (y, x_1, x_2)\} &= \mathbb{P}_{D_s}\{(Y, X_2) = (y, x_2)\} \times \mathbb{P}_{D_s}(X_1 = x_1) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} 1(\mathbf{y}_{[i]} = y \text{ and } \mathbf{X}_{2[i]} = x_2)\sum_{j=1}^{n} 1(\mathbf{X}_{1[j]} = x_1).
\end{aligned}$$

From $\gamma_1 = \frac{1}{n-1}$ and Eq 7.2, we see that

$$\begin{aligned}
\hat{h}_{-,\gamma_1}(f) &= \sum_{i=1}^{n}\sum_{j=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j]}, \mathbf{X}_{2[i]})\} \times \left\{\frac{\gamma_1 1(i=j)}{n} + \frac{1(i \neq j)}{n(n-1)}\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} L\{f, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j]}, \mathbf{X}_{2[i]})\} \times \left\{\frac{1}{n(n-1)}\right\}. \\
&\propto \mathbb{E}_{D_m} L\{f, (Y, X_1, X_2)\}.
\end{aligned}$$

Thus, from Condition 16, we know there exists a function $\hat{g}_{-,\gamma_1}$ that minimizes $\hat{h}_{-,\gamma_1}$ with $\hat{g}_{-,\gamma_1}(x_1^{(a)}, x_2) = \hat{g}_{-,\gamma_1}(x_1^{(b)}, x_2)$ for any $x_1^{(a)}, x_1^{(b)} \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$. The assumption of Proposition 17 then implies that $L\{\hat{g}_{-,\gamma_1}, (y, x_1^{(a)}, x_2)\} = L\{\hat{g}_{-,\gamma_1}, (y, x_1^{(b)}, x_2)\}$ for any $x_1^{(a)}, x_1^{(b)} \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, and $y \in \mathcal{Y}$. We apply this result in Line C.7, below, to show that

loss of model $\hat{g}_{-,\gamma_1}$ is unaffected by permuting X_1 within our sample:

$$\begin{aligned}
\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_1}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j\neq i} L\{\hat{g}_{-,\gamma_1}, (\mathbf{y}_{[i]}, \mathbf{X}_{1[j]}, \mathbf{X}_{2[i]})\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j\neq i} L\{\hat{g}_{-,\gamma_1}, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i]}, \mathbf{X}_{2[i]})\} \\
&= \frac{1}{n}\sum_{i=1}^{n} L\{\hat{g}_{-,\gamma_1}, (\mathbf{y}_{[i]}, \mathbf{X}_{1[i]}, \mathbf{X}_{2[i]})\} \\
&= \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}).
\end{aligned} \tag{C.7}$$

It follows that $\widehat{MR}(\hat{g}_{-,\gamma_1}) = 1$. To show the result of Proposition 17, let $\gamma_2 = 0$. For any function $\hat{g}_{-,\gamma_2}$ minimizing $\hat{h}_{-,\gamma_2}$, we know that

$$\begin{aligned}
\hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_2}) &\leq \hat{h}_{-,\gamma_2}(\hat{g}_{-,\gamma_1}) && \text{from the definition of } \hat{g}_{-,\gamma_2} \\
0 + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_2}) &\leq 0 + \hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_1}) && \text{from } \gamma_2 = 0 \text{ and the definition of } \hat{h}_{-,\gamma_2}.
\end{aligned} \tag{C.8}$$

From $\gamma_2 \leq \gamma_1$, and part 2 of Lemma 15, we know that

$$\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2}) \geq \hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1}). \tag{C.9}$$

Combining Eqs C.8 and C.9, we have

$$\widehat{MR}(\hat{g}_{-,\gamma_2}) = \frac{\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_2})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_2})} \leq \frac{\hat{e}_{\text{switch}}(\hat{g}_{-,\gamma_1})}{\hat{e}_{\text{orig}}(\hat{g}_{-,\gamma_1})} = \widehat{MR}(\hat{g}_{-,\gamma_1}) = 1.$$

The same result does not necessarily hold if we replace $\hat{e}_{\text{switch}}$ with $\hat{e}_{\text{divide}}$ in our definitions of $\hat{h}_{-,\gamma}$, $\widehat{MR}$, and $\widehat{MCR}_{-}$. This is because $\hat{e}_{\text{divide}}$ does not correspond to the expectation over a distribution in which X_1 is independent of X_2 and Y, due to the fixed pairing structure used in $\hat{e}_{\text{divide}}$. Thus, Condition 16 will not apply. □

C.4 Proof of Lemma 19 (upper bound for MR)

Proof. **Part 1: Showing Eq 7.4 holds for all $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$.**

If $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) \geq 0$, then for any function $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$ we know that

$$\begin{aligned}
\frac{1}{\epsilon_{\text{abs}}} &\leq \frac{1}{\hat{e}_{\text{orig}}(f)} \\
\frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} &\leq \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(f)}.
\end{aligned} \tag{C.10}$$

Now, if $\gamma \leq 0$, then for any $f \in \mathcal{F}$ satisfying $\hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$, the definition of $\hat{g}_{+,\gamma}$ implies

$$\begin{aligned}
\hat{h}_{+,\gamma}(f) &\geq \hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right) \\
\hat{e}_{\text{orig}}(f) + \gamma \hat{e}_{\text{switch}}(f) &\geq \hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right) \\
1 + \gamma \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right)}{\hat{e}_{\text{orig}}(f)} \\
1 + \gamma \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} &\geq \frac{\hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right)}{\epsilon_{\text{abs}}} && \text{from Eq C.10} \\
1 + \gamma \widehat{MR}(f) &\geq \frac{\hat{h}_{+,\gamma}\left(\hat{g}_{+,\gamma}\right)}{\epsilon_{\text{abs}}} \\
\widehat{MR}(f) &\leq \left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1 \right\} \gamma^{-1}.
\end{aligned}$$

Part 2: Showing that if $f = \hat{g}_{+,\gamma}$, and at least one of the inequalities in Condition 18 holds with equality, then Eq 7.4 holds with equality.

We consider each of the two inequalities in Condition 18 separately. If $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma}) = 0$, then

$$\begin{aligned}
0 &= \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) + \gamma \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma}) \\
-\gamma \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma}) &= \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) \\
-\frac{\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} &= \frac{1}{\gamma}.
\end{aligned}$$

As a result,

$$\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1 \right\} \gamma^{-1} = \left\{ \frac{0}{\epsilon_{\text{abs}}} - 1 \right\} \left\{ -\frac{\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} \right\} = \widehat{MR}(\hat{g}_{+,\gamma}).$$

Alternatively, if $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) = \epsilon_{\text{abs}}$, then

$$\begin{aligned}
\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1 \right\} \gamma^{-1} = \left\{ \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) + \gamma \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} - 1 \right\} \gamma^{-1} &= \left\{ 1 + \gamma \frac{\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma})}{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})} - 1 \right\} \gamma^{-1} \\
&= \widehat{MR}(\hat{g}_{+,\gamma}).
\end{aligned}$$

□

C.5 Proof of Lemma 20 (monotonicity for MR upper bound binary search)

Proof. **Part 1: $\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})$ is monotonically increasing in γ.**

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2$. We have assumed that $0 \leq \hat{e}_{\text{switch}}(f)$ for any $f \in \mathcal{F}$. Thus, for any $f \in \mathcal{F}$ we have

$$\begin{aligned}
\hat{e}_{\text{orig}}(f) + \gamma_1 \hat{e}_{\text{switch}}(f) &< \hat{e}_{\text{orig}}(f) + \gamma_2 \hat{e}_{\text{switch}}(f) \\
\hat{h}_{+,\gamma_1}(f) &< \hat{h}_{+,\gamma_2}(f). \qquad \text{(C.11)}
\end{aligned}$$

Applying this, we have

$$\begin{aligned} \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) && \text{from Eq C.2} \\ &< \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) && \text{from Eq C.11.} \end{aligned}$$

Part 2: $\hat{e}_{\textbf{orig}}(\hat{g}_{+,\gamma})$ is monotonically decreasing in γ for $\gamma \leq 0$, and Condition 18 holds for $\gamma = 0$ and $\epsilon_{\textbf{abs}} \geq \min_{f \in \mathcal{F}} \hat{e}_{\textbf{orig}}(f)$.

Let $\gamma_1, \gamma_2 \in \mathbb{R}$ satisfy $\gamma_1 < \gamma_2 \leq 0$. Then

$$\begin{aligned} \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) && \text{from Eq C.2} \\ \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_1}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) + (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(g_{\gamma_2}) && \text{from Eq C.3} \\ (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\leq (\gamma_1 - \gamma_2)\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) && \text{from Eqs C.1 \& C.2} \\ \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) \\ \gamma_2\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) &\leq \gamma_2\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) && \text{from } \gamma_2 \leq 0. \end{aligned} \tag{C.12}$$

Now we are equipped to show the result that $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma})$ is monotonically decreasing in γ for $\gamma \leq 0$:

$$\begin{aligned} \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) &\leq \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_1}) && \text{from Eq C.2} \\ \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) + \gamma_2\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) &\leq \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) + \gamma_2\hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) \\ \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) &\leq \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) && \text{from Eqs C.1 \& C.12.} \end{aligned} \tag{C.13}$$

To show that Condition 18 holds for $\gamma = 0$ and $\min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}$, we first note that $h_{0,+}(g_{0,+}) = \hat{e}_{\text{orig}}(g_{0,+})$, which is positive by assumption. Second, we note that

$$\hat{e}_{\text{orig}}(g_{0,+}) = h_{0,+}(g_{0,+}) = \min_{f \in \mathcal{F}} h_{0,+}(f) = \min_{f \in \mathcal{F}} \hat{e}_{\text{orig}}(f) \leq \epsilon_{\text{abs}}.$$

Part 3: $\left\{ \frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\textbf{abs}}} - 1 \right\} \gamma^{-1}$ is monotonically increasing in γ in the range where $\hat{e}_{\textbf{orig}}(\hat{g}_{+,\gamma}) \leq \epsilon_{\textbf{abs}}$ and $\gamma < 0$, and decreasing in the range where $\hat{e}_{\textbf{orig}}(\hat{g}_{+,\gamma}) > \epsilon_{\textbf{abs}}$ and $\gamma < 0$.

To prove the first result, suppose that $\gamma_1 < \gamma_2 < 0$ and $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) \leq \epsilon_{\text{abs}}$. This implies

$$\begin{aligned} \frac{1}{\gamma_2} &< \frac{1}{\gamma_1} \\ \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_2} &> \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1}. \end{aligned} \tag{C.14}$$

Then, starting with Eq C.2,

$$\begin{aligned}
\hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) &\leq \hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_1}) \\
\hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) &\leq \gamma_2 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) + \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) \\
\frac{\hat{h}_{+,\gamma_2}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2} &\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_2} \\
&\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_1}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1} \qquad \text{from Eq C.14} \\
&= \frac{\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1}.
\end{aligned}$$

Dividing both sides of the above equation by ϵ_{abs} proves that $\left\{\frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\boldsymbol{\epsilon_{\text{abs}}}} - 1\right\}\gamma^{-1}$ is monotonically increasing in γ in the range where $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) \leq \epsilon_{\text{abs}}$ and $\gamma < 0$.

To prove the second result we proceed in the same way. Suppose that $\gamma_1 < \gamma_2 < 0$ and $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_1}), \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) \geq \epsilon_{\text{abs}}$. This implies

$$\begin{aligned}
\frac{1}{\gamma_2} &< \frac{1}{\gamma_1} \\
\frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2} &< \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_1}.
\end{aligned} \tag{C.15}$$

Then, starting with Eq C.2,

$$\begin{aligned}
\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) \\
\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) &\leq \gamma_1 \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) + \hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) \\
\frac{\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_1}) - \epsilon_{\text{abs}}}{\gamma_1} &\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_1} \\
&\geq \hat{e}_{\text{switch}}(\hat{g}_{+,\gamma_2}) + \frac{\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2} \qquad \text{from Eq C.15} \\
&= \frac{\hat{h}_{+,\gamma_1}(\hat{g}_{+,\gamma_2}) - \epsilon_{\text{abs}}}{\gamma_2}.
\end{aligned}$$

Diving both sides of the above equation by ϵ_{abs} proves that $\left[\left\{\frac{\hat{h}_{+,\gamma}(\hat{g}_{+,\gamma})}{\epsilon_{\text{abs}}} - 1\right\}\gamma^{-1}\right]$ is monotonically decreasing in γ in the range where $\hat{e}_{\text{orig}}(\hat{g}_{+,\gamma}) > \epsilon_{\text{abs}}$ and $\gamma < 0$. □

C.6 Proof of Lemma 21 (loss upper bound for linear models)

Proof. Under the conditions in Lemma 21 and Eq 7.7, we can construct an upper bound on $L(f_\beta, (y, x)) = (y - x'\beta)^2$ by either maximizing or minimizing $x'\beta$. First, we consider the maximization problem

$$\max_{\beta, x \in \mathbb{R}^p} x'\beta \text{ subject to } x'\mathbf{M}_{\text{lm}}^{-1}x \leq r_{\mathcal{X}} \text{ and } \beta'\mathbf{M}_{\text{lm}}\beta \leq r_{\text{lm}}. \tag{C.16}$$

We can see that both constraints hold with equality at the solution to this problem. Next, we apply the change of variables $\tilde{x} = \frac{1}{\sqrt{r_{\mathcal{X}}}}\mathbf{D}^{\frac{-1}{2}}\mathbf{U}'x$ and $\tilde{\beta} = \frac{1}{\sqrt{r_{\text{lm}}}}\mathbf{D}^{\frac{1}{2}}\mathbf{U}'\beta$, where $\mathbf{UDU}' = \mathbf{M}_{\text{lm}}$ is the eigendecomposition of $\mathbf{M}_{\text{lm}}$. We obtain

$$\max_{\tilde{\beta},\tilde{x}\in\mathbb{R}^p} \tilde{x}'\tilde{\beta}\sqrt{r_{\mathcal{X}}r_{\text{lm}}} \text{ subject to } \tilde{x}'\tilde{x}=1 \text{ and } \tilde{\beta}'\tilde{\beta}=1,$$

which has an optimal objective value equal to $\sqrt{r_{\mathcal{X}}r_{\text{lm}}}$. By negating the objective in Eq C.16, we see that the minimum possible value of $x'\beta$, subject to the constraints in Eq 7.7 and Lemma 21, is found at $-\sqrt{r_{\mathcal{X}}r_{\text{lm}}}$. Thus, we know that

$$L(f,(y,x_1,x_2)) \leq \max\left[\left\{\left(\min_{y\in\mathcal{Y}} y\right)-\sqrt{r_{\mathcal{X}}r_{\text{lm}}}\right\}^2, \left\{\left(\max_{y\in\mathcal{Y}} y\right)+\sqrt{r_{\mathcal{X}}r_{\text{lm}}}\right\}^2\right],$$

for any $(y,x_1,x_2)\in(\mathcal{Y}\times\mathcal{X}_1\times\mathcal{X}_2)$. □

C.7 Proof of Lemma 22 (loss upper bound for regression in a RKHS)

This proofs follows a similar structure as the proof in Section C.6. From the assumptions of Lemma 22, we know from Eq 7.9 that the largest possible output from a model $f_\alpha \in \mathcal{F}_{\mathbf{D},r_k}$ is

$$\begin{aligned}
&\mu + \max_{x\in\mathbb{R}^p,\alpha\in\mathbb{R}^R} \sum_{i=1}^{R} k(x,\mathbf{D}_{[i,\cdot]})\alpha_{[i]} && \text{subject to } v(x)'\mathbf{K}_{\mathbf{D}}^{-1}v(x)\leq r_{\mathbf{D}} \text{ and } \alpha'\mathbf{K}_{\mathbf{D}}\alpha\leq r_k\\
=&\mu + \max_{x\in\mathbb{R}^p,\alpha\in\mathbb{R}^R} v(x)'\alpha && \text{subject to } v(x)'\mathbf{K}_{\mathbf{D}}^{-1}v(x)\leq r_{\mathbf{D}} \text{ and } \alpha'\mathbf{K}_{\mathbf{D}}\alpha\leq r_k\\
\leq&\mu + \max_{\mathbf{z},\alpha\in\mathbb{R}^R} \mathbf{z}'\alpha && \text{subject to } \mathbf{z}'\mathbf{K}_{\mathbf{D}}^{-1}\mathbf{z}\leq r_{\mathbf{D}} \text{ and } \alpha'\mathbf{K}_{\mathbf{D}}\alpha\leq r_k.
\end{aligned}$$

The above problem can be solved in the same way as Eq C.16, and has a solution at $\left(\mu+\sqrt{r_{\mathbf{D}}r_k}\right)$. The smallest possible model output will similarly be lower bounded by $-\left(\mu+\sqrt{r_{\mathbf{D}}r_k}\right)$. Thus, B_{ind} is less than or equal to

$$\max\left[\left\{\min_{y\in\mathcal{Y}}(y)-(\mu+\sqrt{r_{\mathbf{D}}r_k})\right\}^2, \left\{\max_{y\in\mathcal{Y}}(y)+(\mu+\sqrt{r_{\mathbf{D}}r_k})\right\}^2\right].$$

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260.

Beckett, K., Nyrop, K., and Pfingst, L. (2006). Race, drugs, and policing: Understanding disparities in drug delivery arrests. *Criminology*, 44(1):105–137.

Blair, I. V., Judd, C. M., and Chapleau, K. M. (2004). The influence of afrocentric facial features in criminal sentencing. *Psychological science*, 15(10):674–679.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

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

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Calle, M. L. and Urrea, V. (2010). Letter to the editor: stability of random forest importance measures. *Briefings in bioinformatics*, 12(1):86–89.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 598–617. IEEE.

Díaz, I., Hubbard, A., Decker, A., and Cohen, M. (2015). Variable importance and prediction methods for longitudinal problems with missing variables. *PloS one*, 10(3):e0120031.

Dinkelbach, W. (1967). On nonlinear fractional programming. *Management science*, 13(7):492–498.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Gevrey, M., Dimopoulos, I., and Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678.

Hapfelmeier, A., Hothorn, T., Ulm, K., and Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, 24(1):21–34.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning 2nd edition*. New York: Springer.

Heider, K. G. (1988). The Rashomon effect: When ethnographers disagree. *American Anthropologist*, 90(1):73–81.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pages 293–325.

Kamiran, F., Žliobaitė, I., and Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644.

Kazemitabar, J., Amini, A., Bloniarz, A., and Talwalkar, A. S. (2017). Variable importance using decision trees. In *Advances in Neural Information Processing Systems*, pages 425–434.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica*.

Letham, B., Letham, P. A., Rudin, C., and Browne, E. P. (2016). Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063110.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, pages 431–439.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Monahan, J. and Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12:489–513.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access.

Nevo, D. and Ritov, Y. (2015). Identifying a minimal class of models for high-dimensional data. *arXiv preprint arXiv:1504.00494*.

Olden, J. D., Joy, M. K., and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3):389–397.

Park, J. and Boyd, S. (2017). General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint arXiv:1703.07870*.

Paternoster, R. and Brame, R. (2008). Reassessing race disparities in maryland capital cases. *Criminology*, 46(4):971–1008.

Picard-Fritsche, S., Rempel, M., Tallon, J. A., Adler, J., and Reyes, N. (2017). Demystifying risk assessment, key principles and controversies. Technical report. Available at `https://www.courtinnovation.org/publications/demystifying-risk-assessment-key-principles-and-controversies`.

Pólik, I. and Terlaky, T. (2007). A survey of the s-lemma. *SIAM review*, 49(3):371–418.

Ramchand, R., Pacula, R. L., and Iguchi, M. Y. (2006). Racial differences in marijuana-users' risk of arrest in the united states. *Drug and alcohol dependence*, 84(3):264–272.

Recknagel, F., French, M., Harkonen, P., and Yabunaka, K.-I. (1997). Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling*, 96(1-3):11–28.

Roth, W. D. and Mehta, J. D. (2002). The Rashomon effect: Combining positivist and interpretivist approaches in the analysis of contested events. *Sociological Methods & Research*, 31(2):131–173.

Rudin, C. and Schapire, R. E. (2009). Margin-based ranking and an equivalence between adaboost and rankboost. *Journal of Machine Learning Research*, 10(Oct):2193–2232.

Scardi, M. and Harding, L. W. (1999). Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological modelling*, 120(2):213–223.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

Spohn, C. (2000). Thirty years of sentencing reform: The quest for a racially neutral sentencing process. *Criminal justice*, 3:427–501.

Statnikov, A., Lytkin, N. I., Lemeire, J., and Aliferis, C. F. (2013). Algorithms for discovery of multiple markov boundaries. *Journal of Machine Learning Research*, 14(Feb):499–566.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25.

Tulabandhula, T. and Rudin, C. (2014). Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*.

U.S. Department of Justice - Civil Rights Devision (2016). Investigation of the Baltimore City Police Department. Available at `https://www.justice.gov/crt/file/883296/download`.

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

Wang, H., Yang, F., and Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC bioinformatics*, 17(1):60.

Williamson, B. D., Gilbert, P. B., Simon, N., and Carone, M. (2017). Nonparametric variable importance assessment using machine learning techniques.

Yao, J., Teng, N., Poh, H.-L., and Tan, C. L. (1998). Forecasting and analysis of marketing data using neural networks. *J. Inf. Sci. Eng.*, 14(4):843–862.

Zhu, R., Zeng, D., and Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784.