The Rashomon Importance Distribution: Getting RID of Unstable, Single Model-based Variable Importance

Jon Donnelly*
Department of Computer Science
Duke University
Durham, NC 27708
jon.donnelly@duke.edu

Srikar Katta*
Department of Computer Science
Duke University
Durham, NC 27708
srikar.katta@duke.edu

Cynthia Rudin
Department of Computer Science
Duke University
Durham, NC 27708
cynthia.rudin@duke.edu

Edward P. Browne
Department of Medicine
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
epbrowne@email.unc.edu

Abstract

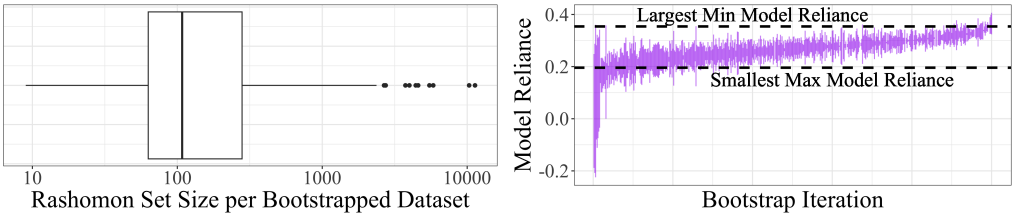
Quantifying variable importance is essential for answering high-stakes questions in fields like genetics, public policy, and medicine. Current methods generally calculate variable importance for a given model trained on a given dataset. However, for a given dataset, there may be many models that explain the target outcome equally well; without accounting for all possible explanations, different researchers may arrive at many conflicting yet equally valid conclusions given the same data. Additionally, even when accounting for all possible explanations for a given dataset, these insights may not generalize because not all good explanations are stable across reasonable data perturbations. We propose a new variable importance framework that quantifies the importance of a variable across the set of all good models and is stable across the data distribution. Our framework is extremely flexible and can be integrated with most existing model classes and global variable importance metrics. We demonstrate through experiments that our framework recovers variable importance rankings for complex simulation setups where other methods fail. Further, we show that our framework accurately estimates the *true importance* of a variable for the underlying data distribution. We provide theoretical guarantees on the consistency and finite sample error rates for our estimator. Finally, we demonstrate its utility with a real-world case study exploring which genes are important for predicting HIV load in persons with HIV, highlighting an important gene that has not previously been studied in connection with HIV.

1 Introduction

Variable importance analysis enables researchers to gain insight into a domain or a model. Scientists are often interested in understanding causal relationships between variables, but running randomized experiments is time-consuming and expensive. Given an observational dataset, we can use global variable importance measures to check if there is a predictive relationship between two variables. It is particularly important in high stakes real world domains such as genetics [44, 34], finance [37], and criminal justice [15, 24] where randomized controlled trials are impractical or unethical. Variable

*Jon Donnelly and Srikar Katta contributed equally to this work.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) The number of models in each Rashomon set (b) Model reliance range across each Rashomon set

Figure 1: Statistics of Rashomon sets computed across 500 bootstrap replicates of a given dataset sampled from the Monk 3 data generation process [42]. The original dataset consisted of 124 observations, and the Rashomon set was calculated using its definition in Equation 1, with parameters specified in Section D of the supplement. The Rashomon set size is the number of models with loss below a threshold. Model reliance is a measure of variable importance for a single variable — in this case, X_2 — and Model Class Reliance (MCR) is its range over the Rashomon set. Both the Rashomon set size and model class reliance are unstable across bootstrap iterations.

importance would ideally be measured as the importance of each variable to the data generating process. However, the data generating process is never known in practice, so prior work generally draws insight by analyzing variable importance for a surrogate model, treating that model and its variable importance as truth.

This approach can be misleading because there may be many good models for a given dataset – a phenomenon referred to as the Rashomon effect [7, 40] — and variables that are important for one good model on a given dataset are *not* necessarily important for others. As such, any insights drawn from a single model need not reflect the underlying data distribution or even the consensus among good models. Recently, researchers have sought to overcome the Rashomon effect by computing *Rashomon sets*, the set of all good (i.e., low loss) models for a given dataset [15, 12]. However, *the set of all good models is not stable across reasonable perturbations (e.g., bootstrap or jackknife) of a single dataset*, with stability defined as in [50]. This concept of stability is one of the three pillars of veridical data science [51, 13]. Note that there is wide agreement on the intuition behind stability, but not its quantification [22, 33]. As such, in line with other stability research, we do not subscribe to a formal definition and treat stability as a general notion [22, 33, 50, 51]. In order to ensure trustworthy analyses, variable importance measures must account for both the Rashomon effect and stability.

Figure 1 provides a demonstration of this problem: across 500 bootstrap replicates from *the same* data set, the Rashomon set varies wildly – ranging from ten models to over *ten thousand* — suggesting that we should account for its instability in any computed statistics. This instability is further highlighted when considering the Model Class Reliance (MCR) variable importance, which is the range of model reliance (i.e., variable importance) values across the Rashomon set for the given dataset [15] (we define MCR and the Rashomon set more rigorously in Sections 2 and 3 respectively). In particular, for variable X_2, one interval — ranging from -0.1 to 0.33 — suggests that there exist good models that do not depend on this variable at all (0 indicates the variable is not important); on the other hand, another MCR from a bootstrapped dataset ranges from 0.33 to 0.36, suggesting that this variable is essential to all good models. Because of this instability, different researchers may draw very different conclusions about the same data distribution even when using the same method.

In this work, we present a framework unifying concepts from classical nonparametric estimation with recent developments on Rashomon sets to overcome the limitations of traditional variable importance measurements. We propose a stable, model- and variable-importance-metric-agnostic estimand that quantifies variable importance across all good models for the empirical data distribution and a corresponding bootstrap-style estimation strategy. Our method creates a cumulative density function (CDF) for variable importance over all variables via the framework shown in Figure 2. Using the CDF, we can compute a variety of statistics (e.g., expected variable importance, interquartile range, and credible regions) that can summarize the variable importance distribution.

The rest of this work is structured as follows. After more formally introducing our variable importance framework, we theoretically guarantee the convergence of our estimation strategy and derive error bounds. We also demonstrate experimentally that our estimand captures the true variable importance for the data generating process more accurately than previous work. Additionally, we illustrate the

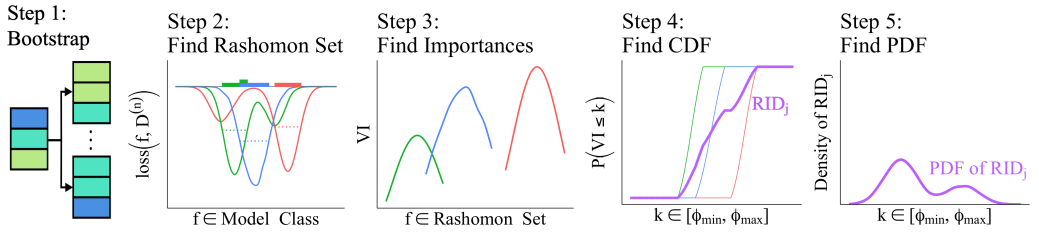


Figure 2: An overview of our framework. **Step 1:** We bootstrap multiple datasets from the original. **Step 2:** We show the loss values over the model class for each bootstrapped dataset, differentiated by color. The dotted line marks the Rashomon threshold; all models whose loss is under the threshold are in the Rashomon set for that bootstrapped dataset. On top, we highlight the number of bootstrapped datasets for which the corresponding model is in the Rashomon set. **Step 3:** We then compute the distribution of model reliance (variable importance – VI) values for variable j across the Rashomon set for each bootstrapped dataset. **Step 4:** We then average the corresponding CDF across bootstrap replicates into a single CDF (in purple). **Step 5:** Using the CDF, we compute the marginal distribution (PDF) of variable importance for variable j across the Rashomon sets of bootstrapped datasets.

generalizability of our variable importance metric by analyzing the reproducibility of our results given new datasets from the same data generation process. Lastly, we use our method to analyze which transcripts and chromatin patterns in human T cells are associated with high expression of HIV RNA. Our results suggest an unexplored link between the LINC00486 gene and HIV load.

Code is available at https://github.com/jdonnelly36/Rashomon_Importance_Distribution.

2 Related Work

The key difference between our work and most others is the way it incorporates model uncertainty, also called the Rashomon effect [7]. The Rashomon effect is the phenomenon in which many different models explain a dataset equally well. It has been documented in high stakes domains including healthcare, finance, and recidivism prediction [14, 30, 15]. The Rashomon effect has been leveraged to create uncertainty sets for robust optimization [43], to perform responsible statistical inference [11], and to gauge whether simple yet accurate models exist for a given dataset [40]. One barrier to studying the Rashomon effect is the fact that *Rashomon sets* are computationally hard to calculate for non-trivial model classes. Only within the past year has code been made available to solve for (and store) the full Rashomon set for any nonlinear function class – that of decision trees [49]. This work enables us to revisit the study of variable importance with a new lens.

A classical way to determine the importance of a variable is to leave it out and see if the loss changes. This is called algorithmic reliance [15] or leave one covariate out (LOCO) inference [25, 36]. The problem with these approaches is that the performance of the model produced by an algorithm will not change if there exist other variables correlated with the variable of interest.

Model reliance (MR) methods capture the global variable importance (VI) of a given feature *for a specific model* [15]. (Note that MR is limited to refer to permutation importance in [15], while we use the term MR to refer to any metric capturing global variable importance of a given feature and model. We use VI and MR interchangably when the relevant model is clear from context.) Several methods for measuring the MR of a model from a specific model class exist, including the variable importance measure from random forest which uses out-of-bag samples [7] and Lasso regression coefficients [20]. Lundberg et al. [28] introduce a way of measuring MR in tree ensembles using SHAP [27]. Williamson et al. [48] develop MR based on the change in performance between the optimal model and the optimal model using a subset of features.

In addition to the metrics tied to a specific model class, many MR methods can be applied to models *from any model class*. Traditional correlation measures [20] can measure the linear relationship (Pearson correlation) or general monotonic relationship (Spearman correlation) between a feature and predicted outcomes for a model from any model class. Permutation model reliance, as discussed by [1, 15, 21], describes how much worse a model performs when the values of a given feature are permuted such that the feature becomes uninformative. Shapley-based measures of MR, such as those

of [47, 27], calculate the average marginal contribution of each feature to a model's predictions. A complete overview of the variable importance literature is beyond the scope of this work; for a more thorough review, see, for example, [2, 31]. Rather than calculating the importance of a variable for a single model, our framework finds the importance of a variable for all models within a Rashomon set, although our framework is applicable to *all* of these model reliance metrics.

In contrast, model class reliance (MCR) methods describe how much a *class of models* (e.g., decision trees) relies on a variable. Fisher et al. [15] uses the Rashomon set to provide bounds on the possible *range* of model reliance for good models of a given class. Smith et al. [41] analytically find the range of model reliance for the model class of random forests. Zhang and Janson [52] introduce a way to compute confidence bounds for a specific variable importance metric over arbitrary models, which Aufiero and Janson [3] extend so that it is applicable to a broad class of surrogate models in pursuit of computational efficiency. These methods report MCR as a range, which gives no estimate of variable importance – only a range of what values are possible. In contrast, Dong and Rudin [12] compute and visualize the variable importance for every member of a given Rashomon set in projected spaces, calculating a set of points; however, these methods have no guarantees of stability to reasonable data perturbations. In contrast, our framework overcomes these finite sample biases, supporting stronger conclusions about the underlying data distribution.

Related to our work from the stability perspective, Duncan et al. [13] developed a software package to evaluate the stability of permutation variable importance in random forest methods; we perform a similar exercise to demonstrate that current variable importance metrics computed for the Rashomon set are not stable. Additionally, Basu et al. [5] introduced iterative random forests by iteratively reweighting trees and bootstrapping to find *stable* higher-order interactions from random forests. Further, theoretical results have demonstrated that bootstrapping stabilizes many machine learning algorithms and reduces the variance of statistics [18, 8]. We also take advantage of bootstrapping's flexibility and properties to ensure stability for our variable importance.

3 Methods

3.1 Definitions and Estimands

Let $\mathcal{D}^{(n)} = \{(X_i, Y_i)\}_{i=1}^n$ denote a dataset of n independent and identically distributed tuples, where $Y_i \in \mathbb{R}$ denotes some outcome of interest and $X_i \in \mathbb{R}^p$ denotes a vector of p covariates. Let g^* represent the data generating process (DGP) producing $\mathcal{D}^{(n)}$. Let $f \in \mathcal{F}$ be a model in a model class (e.g., a tree in the set of all possible sparse decision trees), and let $\phi_j\left(f, \mathcal{D}^{(n)}\right)$ denote a function that measures the importance of variable j for a model f over a dataset $\mathcal{D}^{(n)}$. This can be any of the functions described earlier (e.g., permutation importance, SHAP). Our framework is flexible with respect to the user's choice of ϕ_j and enables practitioners to use the variable importance metric best suited for their purpose; for instance, conditional model reliance [15] is best-suited to measure only the unique information carried by the variable (that cannot be constructed using other variables), whereas other metrics like subtractive model reliance consider the unconditional importance of the variable. Our framework is easily integrable with either of these. We only assume that the variable importance function ϕ has a bounded range, which holds for a wide class of metrics like SHAP [27], permutation model reliance, and conditional model reliance. Finally, let $\ell(f, \mathcal{D}^{(n)}; \lambda)$ represent a loss function given $f, \mathcal{D}^{(n)}$, and loss hyperparameters λ (e.g., regularization). We assume that our loss function is bounded above and below, which is true for common loss functions like 0-1 classification loss, as well as for differentiable loss functions with covariates from a bounded domain.

In an ideal setting, we would measure variable importance using g^* and the whole population, but this is impossible because g^* is unknown and data is finite. In practice, scientists instead use the empirical loss minimizer for a specific dataset $\hat{f}^* \in \arg\min_{f \in \mathcal{F}} \ell(f, \mathcal{D}^{(n)})$; however, several models could explain the same dataset equally well (i.e., the Rashomon effect). Rather than using a single model to compute variable importance, we propose using the entire Rashomon set. Given a single dataset $\mathcal{D}^{(n)}$, we define the **Rashomon set** for a model class $\mathcal{F}$ and parameter ε as the set of all models in $\mathcal{F}$ whose empirical losses are within some bound $\varepsilon > 0$ of the empirical loss minimizer:

$$\mathcal{R}(\epsilon, \mathcal{F}, \ell, \mathcal{D}^{(n)}, \lambda) = \left\{ f \in \mathcal{F} : \ell(f, \mathcal{D}^{(n)}; \lambda) \leq \min_{f' \in \mathcal{F}} \ell(f', \mathcal{D}^{(n)}; \lambda) + \varepsilon \right\}. \tag{1}$$

We denote this Rashomon set by $\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}$ or "Rset" (this assumes a fixed $\mathcal{F}$, ℓ and λ). As discussed earlier, the Rashomon set can be fully computed and stored for non-linear models (e.g., sparse decision trees [49]). For notational simplicity, we often omit λ from the loss function.

While the Rashomon set describes the set of good models for a *single* dataset, Rashomon sets vary across permutations (e.g., subsampling and resampling schemes) of the given data. We introduce a stable quantity for variable importance that accounts for all good models and all permutations from the data using the random variable ***RIV***. ***RIV*** is defined by its cumulative distribution function (CDF), the **Rashomon Importance Distribution (*RID*)**:

$$\begin{aligned} \textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda) &= \mathbb{P}_{\mathcal{D}_b^{(n)}\sim\mathcal{P}_n}(\textit{RIV}_j(\varepsilon,\mathcal{F},\ell,\mathcal{D}_b^{(n)};\lambda)\leq k) \qquad (2)\\ &:= \mathbb{E}_{\mathcal{D}_b^{(n)}\sim\mathcal{P}_n}\left[\frac{|\{f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}:\phi_j(f,\mathcal{D}_b^{(n)})\leq k\}|}{|\mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}|}\right]\\ &= \mathbb{E}_{\mathcal{D}_b^{(n)}\sim\mathcal{P}_n}\left[\frac{\text{vol of Rset s.t. variable } j\text{'s importance is at most } k}{\text{vol of Rset}}\right], \end{aligned}$$

where ϕ_j denotes the variable importance metric being computed on variable j, $k\in[\phi_{\min},\phi_{\max}]$. For a continuous model class (e.g., linear regression models), the cardinality in the above definition becomes the volume under a measure on the function class, usually ℓ_2 on parameter space. ***RID*** constructs the cumulative distribution function (CDF) for the distribution of variable importance across Rashomon sets; as k increases, the value of $\mathbb{P}(\textit{RIV}_j(\varepsilon,\mathcal{F},\ell,\mathcal{D}_b^{(n)};\lambda)\leq k)$ becomes closer to 1. The probability and expectation are taken with respect to datasets of size n sampled from the empirical distribution $\mathcal{P}_n$, which is the same as considering all possible resamples of size n from the originally observed dataset $\mathcal{D}^{(n)}$. Equation (2) weights the contribution of $\phi_j(f,\mathcal{D}_b^{(n)})$ for each model f by the proportion of datasets for which this model is a good explanation (i.e., in the Rashomon set). Intuitively, this provides greater weight to the importance of variables for stable models.

We now define an analogous metric for the loss function ℓ; we define the **Rashomon Loss Distribution (*RLD*)** evaluated at k as the expected fraction of functions in the Rashomon set with loss below k. Here, ***RLV*** is a random variable following this CDF.

$$\begin{aligned} \textit{RLD}(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda) &= \mathbb{P}_{\mathcal{D}_b^{(n)}\sim\mathcal{P}_n}\left(\textit{RLV}(\varepsilon,\mathcal{F},\ell,\mathcal{D}_b^{(n)};\lambda)\leq k\right)\\ &:= \mathbb{E}_{\mathcal{D}_b^{(n)}\sim\mathcal{P}_n}\left[\frac{|\{f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}:\ell(f,\mathcal{D}_b^{(n)})\leq k\}|}{|\mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}|}\right]\\ &= \mathbb{E}_{\mathcal{D}_b^{(n)}\sim\mathcal{P}_n}\left[\frac{|\mathcal{R}\left(k-\min_{f\in\mathcal{F}}\ell(f,\mathcal{D}_b^{(n)}),\mathcal{F},\ell;\lambda\right)|}{|\mathcal{R}(\varepsilon,\mathcal{F},\ell;\lambda)|}\right]. \end{aligned}$$

This quantity shows how quickly the Rashomon set "fills up" on average as loss changes. If there are many near-optimal functions, this will grow quickly with k.

In order to connect ***RID*** for model class $\mathcal{F}$ to the unknown DGP g^*, we make a Lipschitz-continuity-style assumption on the relationship between ***RLD*** and ***RID*** relative to a general model class $\mathcal{F}$ and $\{g^*\}$. To draw this connection, we define the loss CDF for g^*, called LD^*, over datasets of size n as:

$$LD^*(k;\ell,n,\mathcal{P}_n,\lambda) := \mathbb{E}_{\mathcal{D}_b^{(n)}\sim\mathcal{P}_n}\left[\mathbb{1}[\ell(g^*,\mathcal{D}_b^{(n)})\leq k]\right].$$

One could think of LD^* as measuring how quickly the DGP's Rashomon set fills up as loss changes. Here, LD^* is the analog of ***RLD*** for the data generation process.

Assumption 1. *If*

$$\begin{aligned} \rho\left(\textit{RLD}(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda),\textit{LD}^*(\cdot;\ell,n,\mathcal{P}_n,\lambda)\right)\leq\gamma \textit{ then}\\ \rho\left(\textit{RID}_j(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda),\textit{RID}_j(\cdot;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right)\leq d(\gamma) \end{aligned}$$

for a function $d:[0,\ell_{\max}-\ell_{\min}]\to[0,\phi_{\max}-\phi_{\min}]$ *such that* $\lim_{\gamma\to 0}d(\gamma)=0$. *Here,* ρ *represents any distributional distance metric (e.g., 1-Wasserstein).*

Assumption 1 says that a Rashomon set consisting of good approximations for g^* in terms of loss will also consist of good approximations for g^* in terms of variable importance. More formally, from this assumption, we know that as $\rho(LD^*, \textit{RLD}) \to 0$, the variable importance distributions will converge: $\rho\left(\textit{RID}_j(\cdot, \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda), \textit{RID}_j(\cdot, \varepsilon, \{g^*\}, \ell, \mathcal{P}_n, \lambda)\right) \to 0$. We demonstrate that this assumption is realistic for a variety of model classes like linear models and generalized additive models in Section C of the supplement.

3.2 Estimation

We estimate $\textit{RID}_j$ for each variable j by leveraging bootstrap sampling to draw new datasets from the empirical data distribution: we sample observations from an observed dataset, construct its Rashomon set, and compute the j-th variable's importance for each model in the Rashomon set. After repeating this process for B bootstrap iterations, we estimate *RID* by weighting each model f's realized variable importance score (evaluated on each bootstrapped dataset) by the proportion of the bootstrapped datasets for which f is in the Rashomon set *and* the size of each Rashomon set in which f appears. Specifically, let $\mathcal{D}_b^{(n)}$ represent the dataset sampled with replacement from $\mathcal{D}^{(n)}$ in iteration $b = 1, \ldots, B$ of the bootstrap procedure. For each dataset $\mathcal{D}_b^{(n)}$, we find the Rashomon set $\mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}$. Finally, we compute an **empirical estimate** $\widehat{\textit{RID}}_j$ of $\textit{RID}_j$ by computing:

$$\widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) = \frac{1}{B} \sum_{b=1}^{B} \left(\frac{|\{f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}} : \phi_j(f, \mathcal{D}_b^{(n)}) \leq k\}|}{|\mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}|} \right).$$

Under Assumption 1, we can directly connect our estimate $\widehat{\textit{RID}}(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)$ to the DGP's variable importance distribution $\textit{RID}(k; \ell_{\text{max}}, \{g^*\}, \ell, \mathcal{P}_n, \lambda)$, which Theorem 1 formalizes.

Theorem 1. *Let Assumption 1 hold for distributional distance $\rho(A_1, A_2)$ between distributions A_1 and A_2. For any $t > 0$, $j \in \{0, \ldots, p\}$ as $\rho\left(LD^*(\cdot; \ell, n, \lambda), \textit{RLD}(\cdot; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)\right) \to 0$ and $B \to \infty$,*

$$\mathbb{P}\left(\left| \widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) - \textit{RID}_j(k; \varepsilon, \{g^*\}, \ell, \mathcal{P}_n, \lambda) \right| \geq t \right) \to 0.$$

For a set of models that performs sufficiently well in terms of loss, $\widehat{\textit{RID}}_j$ thus recovers the CDF of variable importance for the true model across all reasonable perturbations. Further, we can provide a finite sample bound for the estimation of a marginal distribution between $\widehat{\textit{RID}}_j$ and $\textit{RID}_j$ for the model class $\mathcal{F}$, as stated in Theorem 2. Note that this result does not require Assumption 1.

Theorem 2. *Let $t > 0$ and $\delta \in (0, 1)$ be some pre-specified values. Then, with probability at least $1 - \delta$ with respect to bootstrap samples of size n,*

$$\left| \widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) - \textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) \right| \leq t \tag{3}$$

with number of bootstrap samples $B \geq \frac{1}{2t^2} \ln\left(\frac{2}{\delta}\right)$ for any $k \in [\phi_{\min}, \phi_{\max}]$.

Because we use a bootstrap procedure, we can control the number of bootstrap iterations to ensure that the difference between $\textit{RID}_j$ and $\widehat{\textit{RID}}_j$ is within some pre-specified error. (As defined earlier, $\textit{RID}_j$ is the expectation over infinite bootstraps, whereas $\widehat{\textit{RID}}_j$ is the empirical average over B bootstraps.) For example, after 471 bootstrap iterations, we find that $\widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)$ is within 0.075 of $\textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)$ for any given k with 90% confidence. It also follows that as B tends to infinity, the estimated $\textit{RIV}_j$ will converge to the true value.

Since we stably estimate the entire distribution of variable importance values, we can create (1) stable point estimates of variable importance (e.g., expected variable importance) that account for the Rashomon effect, (2) interquantile ranges of variable importance, and (3) confidence regions that characterize uncertainty around a point estimate of variable importance. We prove exponential rates of convergence for these statistics estimated using our framework in Section B of the supplement.

Because our estimand and our estimation strategy (1) enable us to manage instability, (2) account for the Rashomon effect, and (3) are completely model-agnostic and flexibly work with most existing variable importance metrics, *RID* is a valuable quantification of variable importance.

4 Experiments With Known Data Generation Processes

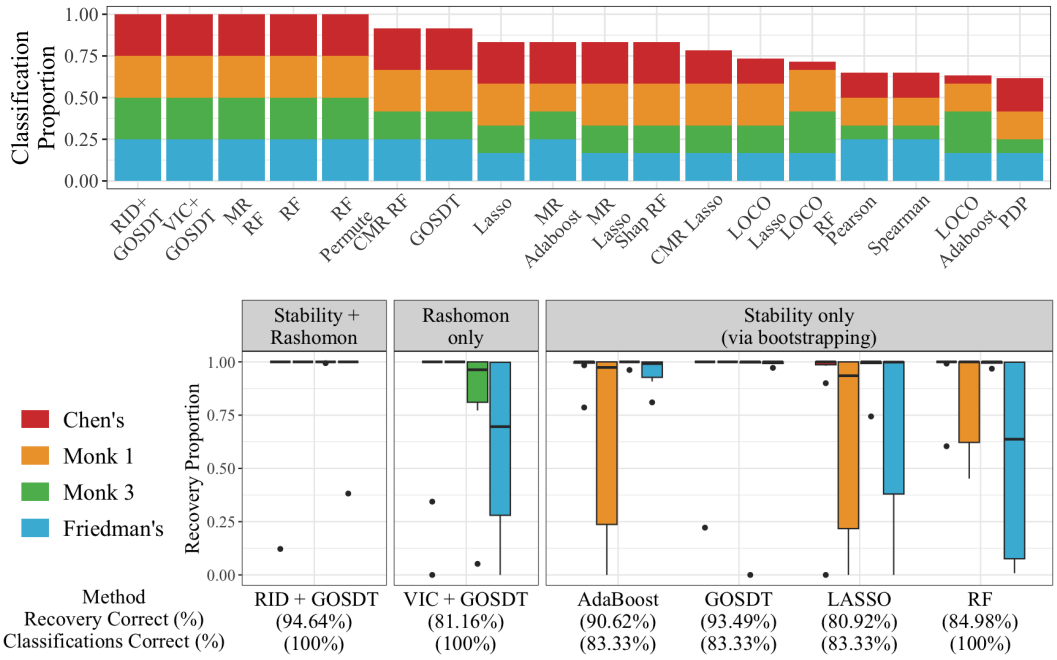


Figure 3: (Top) The proportion of features ranked correctly by each method on each data set represented as a *stacked* barplot. The figures are ordered by method performance across the four simulation setups. (Bottom) The proportion of independent DGP $\phi^{(sub)}$ calculations on *500 new datasets from the DGP* that were contained within the box-and-whiskers range computed using a single training set (with bootstrapping in all methods except VIC) for each method and variable in each simulation. Underneath each method's label, the first row shows the percentage of times across all 500 independently generated datasets and variables that the DGP's variable importance was inside of that method's box-and-whiskers interval. The second row shows the percentage of pairwise rankings correct for each method (from the top plot). Higher is better.

4.1 *RID* Distinguishes Important Variables from Extraneous Variables

There is no generally accepted ground truth measure for variable importance, so we first evaluate whether a variety of variable importance methods can correctly distinguish between variables used to generate the outcome (in a known data generation process) versus those that are not. We consider the following four data generation processes (DGPs). **Chen's** DGP [9]: $Y = \mathbb{1}[-2\sin(X_1)+\max(X_2,0)+X_3+\exp(-X_4)+\varepsilon \geq 2.048]$, where $X_1,\ldots,X_{10},\varepsilon \sim \mathcal{N}(0,1)$. Here, only $X_1,\ldots,X_4$ are relevant. **Friedman's** DGP [17]: $Y = \mathbb{1}[10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon \geq 15]$, where $X_1,\ldots,X_6 \sim \mathcal{U}(0,1), \varepsilon \sim \mathcal{N}(0,1)$. Here, only $X_1,\ldots,X_5$ are relevant. The **Monk 1** DGP [42]: $Y = \max\left(\mathbb{1}[X_1 = X_2], \mathbb{1}[X_5 = 1]\right)$, where the variables $X_1,\ldots,X_6$ have domains of 2, 3, or 4 unique integer values. Only X_1, X_2, X_5 are important. The **Monk 3** DGP [42]: $Y = \max\left(\mathbb{1}[X_5 = 3 \text{ and } X_4 = 1], \mathbb{1}[X_5 \neq 4 \text{ and } X_2 \neq 3]\right)$ for the same covariates in Monk 1. Also, 5% label noise is added. Here, X_2, X_4, and X_5 are relevant.

We compare the ability of *RID* to identify extraneous variables with that of the following baseline methods, whose details are provided in Section D of the supplement: subtractive model reliance ϕ^{sub} of a random forest (RF) [6], LASSO [20], boosted decision trees [16], and generalized optimal sparse decision trees (GOSDT) [26]; conditional model reliance (CMR) [15]; the impurity based model reliance metric for RF from [7]; the LOCO algorithm reliance [25] for RF and Lasso; the Pearson and Spearman correlation between each feature and the outcome; the mean of the partial dependency plot (PDP) [19] for each feature; the SHAP value [28] for RF; and mean of variable importance clouds (VIC) [12] for the Rashomon set of GOSDTs [49]. If we do not account for instability and simply learn a model and calculate variable importance, baseline models generally perform poorly, as shown

in Section E of the supplement. Thus, we chose to account for instability in a way that benefits the baselines. We evaluate each baseline method for each variable across 500 bootstrap samples and compute the *median VI across bootstraps*, with the exception of VIC — for VIC, we take the *median VI value across the Rashomon set* for the original dataset, as VIC accounts for Rashomon uncertainty. Here, we aim to see whether we can identify extraneous (i.e., unimportant variables). For a DGP with C extraneous variables, we classify the C variables with the C smallest median variable importance values as extraneous. We repeat this experiment with different values for the Rashomon threshold ε in Section E of the supplement.

Figure 3 (top) reports the proportion of variables that are correctly classified for each simulation setup as a stacked barplot. ***RID* identifies all important and unimportant variables** for these complex simulations. Note that four other baseline methods – MR RF, RF Impurity, RF Permute, and VIC – also differentiated all important from unimportant variables. Motivated by this finding, we next explore how well methods recover the true value for subtractive model reliance on the DGP, allowing us to distinguish between the best performing methods on the classification task.

4.2 *RID* Captures Model Reliance for the True Data Generation Process

RID allows us to quantify uncertainty in variable importance due to *both* the Rashomon effect and instability. We perform an ablation study investigating how accounting for both stability and the Rashomon effect compares to having one without the other. We evaluate what proportion of subtractive model reliances calculated for the DGP on 500 test sets are contained within uncertainty intervals generated using only one training dataset. This experiment tells us whether the intervals created on a single dataset will generalize.

To create the uncertainty interval on the training dataset and for each method, we first find the subtractive model reliance $\phi^{(sub)}$ across 500 bootstrap iterations of a given dataset for the four algorithms shown in Figure 3 (bottom) (baseline results without bootstrapping are in Section E of the supplementary material). Additionally, we find the VIC for the Rashomon set of GOSDTs on the original dataset. We summarize these model reliances (500 bootstraps $\times$ 28 variables across datasets $\times$ 4 algorithms + 8,247 models in VIC's + 10,840,535 total models across Rsets $\times$ 28 variables from *RID*) by computing their box-and-whisker ranges (1.5 $\times$ Interquartile range [46]). To compare with "ground truth," we sample 500 test datasets from the DGP and calculate $\phi^{(sub)}$ for the DGP for that dataset. For example, assume the DGP is $Y = X^2 + \varepsilon$. We would then use $f(X) = X^2$ as our predictive model and evaluate $\phi^{(sub)}(f, \mathcal{D}^{(n)})$ on f for each of the 500 test sets. We then check if the box-and-whisker range of each method's interval constructed on the training set contains the computed $\phi^{(sub)}$ for the DGP for each test dataset. Doing this allows us to understand whether our interval contains the *true* $\phi^{(sub)}$ for each test set.

Figure 3 (bottom) illustrates the proportion of times that the test variable importance values fell within the uncertainty intervals from training. These baselines fail to capture the test $\phi^{(sub)}$ *entirely* for at least one variable ($< 0.05\%$ recovery proportion). **Only *RID both* recovers important/unimportant classifications perfectly and achieves a strong recovery proportion at 95%**.

4.3 *RID* is Stable

Our final experiment investigates the stability of VIC and MCR (which capture only Rashomon uncertainty but not stability) to *RID*, which naturally considers data perturbations. We generate 50 independent datasets from each DGP and compute the box-and-whisker ranges (BWR) of each uncertainty metric for each dataset; for every pair of BWRs for a given method, we then calculate the Jaccard similarity between BWR's. For each generated dataset, we then average the Jaccard similarity across variables. Figure 14 shows these intervals for each non-extraneous variable from Chen's DGP. Supplement E.4 presents a similar figure for each DGP, showing that only *RID*'s intervals overlap across all generations for all datasets.

Figure 5 displays the similarity scores between the box and whisker ranges of MCR, VIC, and *RID* across the 50 datasets for each DGP. Note that Monk 1 has no noise added, so instability should not be a concern for any method. For datasets including noise, **MCR and VIC achieve median similarity below 0.55; *RID*'s median similarity is 0.69; it is much more stable**.

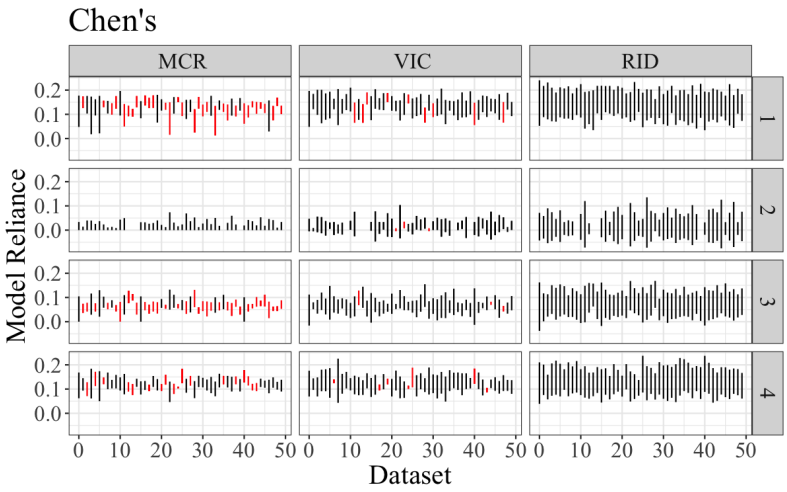


Figure 4: We generate 50 independent datasets from Chen's DGP and calculate MCR, BWRs for VIC, and BWRs for RID. The above plot shows the interval for each dataset for each non-null variable in Chen's DGP. All red-colored intervals do not overlap with at least one of the remaining 49 intervals.

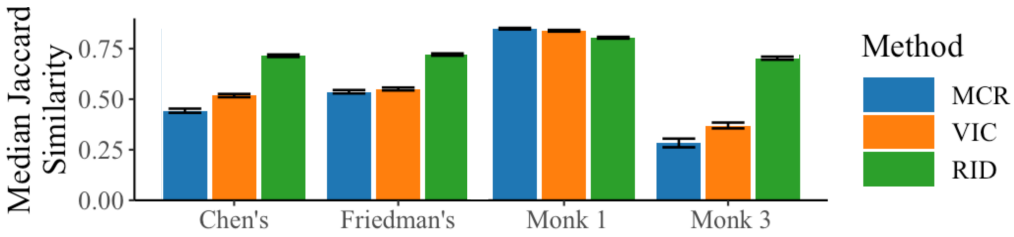


Figure 5: Median Jaccard similarity scores across 50 independently generated MCR, VIC, and *RID* box and whisker ranges for each DGP; 1 is perfect similarity. Error bars show 95% confidence interval around the median.

5 Case Study

Having validated *RID* on synthetic datasets, we demonstrate its utility in a real world application: studying which host cell transcripts and chromatin patterns are associated with high expression of Human Immunodeficiency Virus (HIV) RNA. We analyzed a dataset that combined single cell RNAseq/ATACseq profiles for 74,031 individual HIV infected cells from two different donors in the aims of finding new cellular cofactors for HIV expression that could be targeted to reactivate the latent HIV reservoir in people with HIV (PWH). A longer description of the data is in [29]. Finding results on this type of data allows us to create new hypotheses for which genes are important for HIV load prediction and might generalize to large portions of the population.

To identify which genes are stably importance across good models, we evaluated this dataset using *RID* over the model class of sparse decision trees using subtractive model reliance. We selected 14,614 samples (all 7,307 high HIV load samples and 7,307 random low HIV load samples) from the overall dataset in order to balance labels, and filtered the complete profiles down to the top 100 variables by individual AUC. We consider the binary classification problem of predicting high versus low HIV load. For full experimental details, see Section D of the supplement. Section E.5 of the supplement contains timing experiments for *RID* using this dataset.

Figure 6 illustrates the probability that *RID* is greater than 0 for the 10 highest probability variables (0 is when the variable is not important at all). **We find that LINC00486 – a less explored gene – is the most important variable**, with $1 - RID_{LINC00486}\,(0) = 78.4\%$. LINC00486 is a long non-coding RNA (i.e., it functions as an RNA molecule but does not encode a protein like most genes), and there is no literature on this gene and HIV, making this association a novel one. However, recent work [45] has shown that LINC00486 can enhance EBV (Epstein–Barr virus) infection by activating NF-κB.

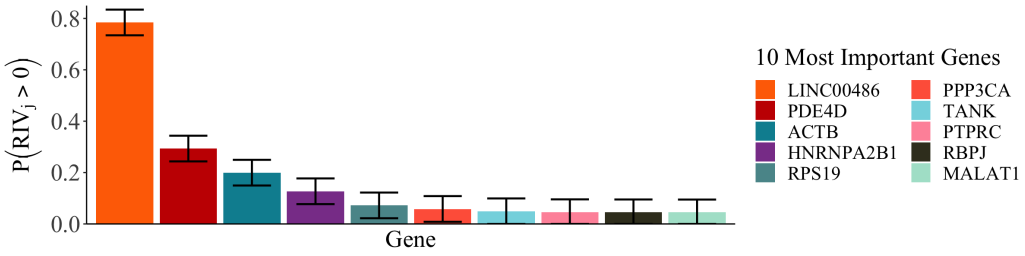


Figure 6: Probability of each gene's model reliance being greater than 0 across Rashomon sets across bootstrapped datasets for the ten genes with the highest $\mathbb{P}(RIV_j > 0)$. We ran 738 bootstrap iterations to ensure that $\mathbb{P}(\widehat{RIV}_j > 0)$ is within 0.05 of $\mathbb{P}(RIV_j > 0)$ with 95% confidence (from Theorem 2).

It is well established that NF-κB can regulate HIV expression [32, 23, 4], suggesting a possible mechanism and supporting future study. Notably, *RID* also highlighted PDE4D, which interacts with the Tat protein and thereby HIV transcription [39]; HNRNPA2B1, which promotes HIV expression by altering the structure of the viral promoter [38]; and MALAT1, which has recently been shown to be an important regulator of HIV expression [35]. These three findings validate prior work and show that *RID* can uncover variables that are known to interact with HIV.

Note that previous methods – even those that account for the Rashomon effect – could not produce this result. MCR and VIC do not account for instability. For example, after computing MCR for 738 bootstrap iterations, we find that the MCR for the LINC00486 gene has overlap with 0 in 96.2% of bootstrapped datasets, meaning MCR would not allow us to distinguish whether LINC00486 is important or not 96.2% of the time. Without *RID*, we would not have strong evidence that LINC00486 is necessary for good models. By explicitly accounting for instability, we increase trust in our analyses.

Critically, *RID* also found *very low* importance for the majority of variables, allowing researchers to dramatically reduce the number of possible directions for future experiments designed to test a gene's functional role. Such experiments are time consuming and cost tens of thousands of dollars *per donor*, so narrowing possible future directions to a small set of genes is of the utmost importance. **Our analysis provides a manageable set of clear directions for future work studying the functional roles of these genes in HIV.**

6 Conclusion and Limitations

We introduced *RID*, a method for recovering the importance of variables in a way that accounts for both instability and the Rashomon effect. We showed that *RID* distinguishes between important and extraneous variables, and that *RID* better captures the true variable importance for the DGP than prior methods. We showed through a case study in HIV load prediction that *RID* can provide insight into complicated real world problems. Our framework overcomes instability and the Rashomon effect, moving beyond variable importance for a single model and increasing reproducibility.

RID can be directly computed for any model class for which the Rashomon set can be found – at the time of publishing, decision trees, linear models, and GLMs. A limitation is that currently, there are relatively few model classes for which the Rashomon set can be computed. Therefore, future work should aim to compute and store the Rashomon set of a wider variety of model classes. Future work may investigate incorporating Rashomon sets that may be well-approximated (e.g., GAMs, [10]), but not computed exactly, into the *RID* approach. Nonetheless, sparse trees are highly flexible, and using them with *RID* improves the trustworthiness and transparency of variable importance measures, enabling researchers to uncover important, reproducible relationships about complex processes without being misled by the Rashomon effect.

7 Acknowledgements

We gratefully acknowledge support from NIH/NIDA R01DA054994, NIH/NIAID R01AI143381, DOE DE-SC0023194, NSF IIS-2147061, and NSF IIS-2130250.

Supplemental Material

A Proofs

First, recall the following assumption from the main paper:

Assumption 1. *If*

$$\rho\left(\textit{RLD}(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda),\textit{LD}^*(\cdot;\ell,n,\mathcal{P}_n,\lambda)\right)\leq\gamma \text{ then}$$
$$\rho\left(\textit{RID}_j(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda),\textit{RID}_j(\cdot;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right)\leq d(\gamma)$$

for a function $d:[0,\ell_{\max}-\ell_{\min}]\to[0,\phi_{\max}-\phi_{\min}]$ *such that* $\lim_{\gamma\to 0}d(\gamma)=0$. *Here,* ρ *represents any distributional distance metric (e.g., 1-Wasserstein).*

Theorem 1. *Let Assumption 1 hold for distributional distance* $\rho(A_1,A_2)$ *between distributions* A_1 *and* A_2. *For any* $t>0$, $j\in\{0,\ldots,p\}$ *as* $\rho\left(\textit{LD}^*(\cdot;\ell,n,\lambda),\textit{RLD}(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right)\to 0$ *and* $B\to\infty$,

$$\mathbb{P}\left(\left|\widehat{\textit{RID}}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right|\geq t\right)\to 0.$$

Proof. Let $\mathcal{D}^{(n)}$ be a dataset of n (x_i,y_i) tuples independently and identically distributed according to the empirical distribution $\mathcal{P}_n$. Let $k\in[\phi_{\min},\phi_{\max}]$.

Then, we know that

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\textit{RID}}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right|\geq t\right)\\
=&\mathbb{P}\Big(\Big|\widehat{\textit{RID}}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\\
&\quad+\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\Big|\geq t\Big)\\
&\quad\text{(by adding 0)}\\
\leq&\mathbb{P}\Big(\Big|\widehat{\textit{RID}}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\Big|\\
&\quad+\left|\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right|\geq t\Big)\\
&\quad\text{(by the triangle inequality)}\\
\leq&\mathbb{P}\left(\left|\widehat{\textit{RID}}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right|\geq\frac{t}{2}\right)\\
&\quad+\mathbb{P}\left(\left|\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right|\geq\frac{t}{2}\right)\\
&\quad\text{(by union bound).}
\end{aligned}$$

Recall that, in the theorem statement, we have assumed $\rho\left(\textit{LD}^*(\cdot;\ell,n,\lambda),\textit{RLD}(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right)\to 0$. Therefore, by Assumption 1,

$$\mathbb{P}\left(\left|\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right|\geq\frac{t}{2}\right)\to 0.$$

Additionally, we will show in Corollary 1 that as $B\to\infty$,

$$\mathbb{P}\left(\left|\widehat{\textit{RID}}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)-\textit{RID}_j(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right|\geq\frac{t}{2}\right)\to 0.$$

Therefore, as $B\to\infty$ and $\rho\left(\textit{LD}^*(\cdot;\ell,n,\lambda),\textit{RLD}(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right)\to 0$, the *estimated* Rashomon importance distribution for model class $\mathcal{F}$ converges to the true Rashomon importance distribution for the DGP g^*. □

Theorem 2. *Let* $\mathcal{D}^{(n)}$ *be a dataset of* n (x_i,y_i) *tuples independently and identically distributed according to the empirical distribution* $\mathcal{P}_n$. *Let* $k\in[\phi_{\min},\phi_{\max}]$. *Then, with probability* $1-\delta$, *with* $B\geq\frac{1}{2t^2}\ln\left(\frac{2}{\delta}\right)$ *bootstrap replications,*

$$\left|\widehat{\textit{RID}}_j(k)-\textit{RID}_j(k)\right|<t.$$

*Jon Donnelly and Srikar Katta contributed equally to this work.

Proof. First, let us restate the definition of $\textit{RID}_j$ and $\widehat{\textit{RID}}_j$. Let $n \in \mathbb{N}$. Let ε be the Rashomon threshold, and let the Rashomon set for some dataset $\mathcal{D}^{(n)}$ and some fixed model class $\mathcal{F}$ be denoted as $\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}$. Without loss of generality, assume $\mathcal{F}$ is a finite model class. Then, for a given $k \in [\phi_{\min}, \phi_{\max}]$,

$$\textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) = \mathbb{E}_{\mathcal{D}^{(n)} \sim \mathcal{P}_n} \left[\frac{\sum_{f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}} \mathbb{1}[\phi_j(f, \mathcal{D}^{(n)}) \leq k]}{|\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}|} \right].$$

Note that the expectation is over all datasets of size n sampled with replacement from the originally observed dataset, represented by $\mathcal{P}_n$; we are taking the expectation over bootstrap samples.

We then sample datasets of size n with replacement from the *empirical* CDF $\mathcal{P}_n$, find the Rashomon set for the replicate dataset, and compute the variable importance metric for each model in the discovered Rashomon set. For the same $k \in [\phi_{\min}, \phi_{\max}]$,

$$\widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) = \frac{1}{B} \sum_{\mathcal{D}_b^{(n)} \sim \mathcal{P}_n} \left[\frac{\sum_{f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}} \mathbb{1}[\phi_j(f, \mathcal{D}_b^{(n)}) \leq k]}{|\mathcal{R}^{\varepsilon}_{\mathcal{D}_b^{(n)}}|} \right],$$

where B represents the number of size n datasets sampled from $\mathcal{P}_n$.

Notice that

$$0 \leq \frac{\sum_{f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}} \mathbb{1}[\phi_j(f, \mathcal{D}^{(n)}) \leq k]}{|\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}|} \leq 1. \tag{4}$$

Because $\widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)$ is an Euclidean average of the quantity in Equation (4) and $\textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)$ is the expectation of the quantity in Equation (4), we can use Hoeffding's inequality to show that

$$\begin{aligned} &\mathbb{P}\left(\left| \widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) - \textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) \right| > t \right) \\ \leq &2 \exp\left(-2Bt^2\right) \end{aligned}$$

for some $t > 0$.

Now, we can manipulate Hoeffding's inequality to discover a finite sample bound. Instead of setting B and t, we will now find the B necessary to guarantee that

$$\mathbb{P}\left(\left| \widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) - \textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) \right| \geq t \right) \leq \delta \tag{5}$$

for some $\delta, t > 0$.

Let $\delta > 0$. From Hoeffding's inequality, we see that if we choose B such that $2 \exp\left(-2Bt^2\right) \leq \delta$, then

$$\mathbb{P}\left(\left| \widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) - \textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) \right| \geq t \right) \leq 2 \exp\left(-2Bt^2\right) \leq \delta.$$

Notice that $2 \exp\left(-2Bt^2\right) \leq \delta$ if and only if $B \geq \frac{1}{2t^2} \ln\left(\frac{2}{\delta}\right)$.

Therefore, with probability $1 - \delta$,

$$\left| \widehat{\textit{RID}}_j(k) - \textit{RID}_j(k) \right| \leq t$$

with $B \geq \frac{1}{2t^2} \ln\left(\frac{2}{\delta}\right)$ bootstrap iterations. □

Corollary 1. *Let $t > 0$, $k \in [\phi_{\min}, \phi_{\max}]$, and assume that $\mathcal{D}^{(n)} \sim \mathcal{P}_n$. As $B \to \infty$,*

$$\mathbb{P}\left(\left| \widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) - \textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) \right| \geq t \right) \to 0.$$

Proof. Recall the results of Theorem 2:

$$\begin{aligned} &\mathbb{P}\left(\left| \widehat{\textit{RID}}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) - \textit{RID}_j(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) \right| \geq t \right) \\ &\leq 2 \exp\left(-2Bt^2\right) \\ &\to 0 \text{ as B } \to \infty. \end{aligned}$$

□

B Statistics Derived From *RID*

Corollary 2. *Let $\varepsilon, B > 0$. Then,*

$$\mathbb{P}\left(\left|\mathbb{E}[\textit{RID}_j] - \mathbb{E}[\widehat{\textit{RID}}_j]\right| \geq \varepsilon_E\right) \leq 2\exp\left(\frac{-2B\varepsilon_E^2}{(\phi_{\max} - \phi_{\min})^2}\right). \tag{6}$$

Therefore, the expectation of $\widehat{\textit{RIV}}_j$ converges exponentially quickly to the expectation of $\textit{RIV}_j$. The notation $\mathbb{E}[\textit{RID}_j]$ denotes the expectation of the random variable distributed according to $\textit{RID}_j$.

Proof. Let $\phi_{\min}, \phi_{\max}$ represent the bounds of the variable importance metric ϕ. Assume that $0 \leq \phi_{\min} \leq \phi_{\max} < \infty$. If $\phi_{\min} < 0$, then we can modify the variable importance metric to be strictly positive; for example, if ϕ is Pearson correlation – which has a range between -1 and 1 – we can define a new variable importance metric that is the absolute value of the Pearson correlation *or* define another metric that is the Pearson correlation plus 1 so that the range is now bounded below by 0.

Now, recall that for any random variable X whose support is strictly greater than 0, we can calculate its expectation as $\mathbb{E}_X[X] = \int_0^\infty (1 - \mathbb{P}(X \leq x))dx$. Because $\phi_{\min} \geq 0$, we know that

$$\begin{aligned}
&\mathbb{E}[\textit{RID}_j] \\
&= \int_{\phi_{\min}}^{\phi_{\max}} (1 - \mathbb{P}(\textit{RIV}_j \leq k))\, dk \\
&= \int_{\phi_{\min}}^{\phi_{\max}} \left(1 - \mathbb{E}_{D^{(n)}}\left[\sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]}\right]\right) dk \\
&= \int_{\phi_{\min}}^{\phi_{\max}} dk - \int_{\phi_{\min}}^{\phi_{\max}} \mathbb{E}_{D^{(n)}}\left[\sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]}\right] dk \\
&= (\phi_{\max} - \phi_{\min}) - \mathbb{E}_{D^{(n)}}\left[\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right] \text{ by Fubini's theorem.}
\end{aligned}$$

Using similar logic we can show that

$$\begin{aligned}
\mathbb{E}[\widehat{\textit{RID}}_j] &= \int_{\phi_{\min}}^{\phi_{\max}} \left(1 - \frac{1}{B}\sum_{b=1}^{B}\sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D_b^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]}\right) dk \\
&= \int_{\phi_{\min}}^{\phi_{\max}} dk - \int_{\phi_{\min}}^{\phi_{\max}} \frac{1}{B}\sum_{b=1}^{B}\sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D_b^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk \\
&= (\phi_{\max} - \phi_{\min}) - \frac{1}{B}\sum_{b=1}^{B}\left(\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D_b^{(n)}, m) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right).
\end{aligned}$$

We can then rewrite $\left|\mathbb{E}[\textit{RID}_j] - \mathbb{E}[\widehat{\textit{RID}}_j]\right|$ using the calculations above:

$$\begin{aligned}
&\left|\mathbb{E}[\textit{RID}_j] - \mathbb{E}[\widehat{\textit{RID}}_j]\right| \\
&= \Bigg| (\phi_{\max} - \phi_{\min}) - \mathbb{E}_{D^{(n)}\sim\mathcal{P}_n}\left[\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi)j(f, D_b^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right] \\
&\quad - \left((\phi_{\max} - \phi_{\min}) - \frac{1}{B}\sum_{b=1}^{B}\left(\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D_b^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right)\right) \Bigg| \\
&= \Bigg| - \mathbb{E}_{D^{(n)}\sim\mathcal{P}_n}\left[\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D_b^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right] \\
&\quad + \frac{1}{B}\sum_{b=1}^{B}\left(\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f, D_b^{(n)}) \leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f \in \mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right) \Bigg|.
\end{aligned}$$

Because $0 \leq \mathbb{P}(\mathit{RIV}_j \leq k), \mathbb{P}(\widehat{\mathit{RIV}_j} \leq k) \leq 1$ for all $k \in \mathbb{R}$,

$$\int_{\phi_{\min}}^{\phi_{\max}} 0dk \leq \int_{\phi_{\min}}^{\phi_{\max}} \mathbb{P}(\mathit{RIV}_j \leq k)dk, \int_{\phi_{\min}}^{\phi_{\max}} \mathbb{P}(\widehat{\mathit{RIV}_j} \leq k)dk \leq \int_{\phi_{\min}}^{\phi_{\max}} 1dk$$
$$0 \leq \int_{\phi_{\min}}^{\phi_{\max}} \mathbb{P}(\mathit{RIV}_j \leq k)dk, \int_{\phi_{\min}}^{\phi_{\max}} \mathbb{P}(\widehat{\mathit{RIV}_j} \leq k)dk \leq (\phi_{\max} - \phi_{\min}),$$

suggesting that $\left(\int_{\phi_{\min}}^{\phi_{\max}} \frac{\sum_{f\in\mathcal{F}} \mathbb{1}[f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f,D_b^{(n)})\leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right)$ is bounded.

Then, by Hoeffding's inequality, we know that

$$\begin{aligned}
&\mathbb{P}\left(\left|\mathbb{E}[\mathit{RIV}_j] - \mathbb{E}[\widehat{\mathit{RIV}_j}]\right| > \varepsilon_E\right) \\
=&\mathbb{P}\left(\left|\mathbb{E}_{D^{(n)}\sim\mathcal{P}_n}\left[\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f,D_b^{(n)})\leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right]\right.\right. \\
&\left.\left. - \frac{1}{B}\sum_{b=1}^{B}\left(\int_{\phi_{\min}}^{\phi_{\max}} \sum_{f\in\mathcal{F}} \frac{\mathbb{1}[f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]\mathbb{1}[\phi_j(f,D_b^{(n)})\leq k]}{\sum_{f\in\mathcal{F}} \mathbb{1}[f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}]} dk\right)\right| > \varepsilon_E\right) \\
\leq&2\exp\left(\frac{-2B\varepsilon_E^2}{(\phi_{\max}-\phi_{\min})^2}\right).
\end{aligned}$$

□

Corollary 3. *Assume $\widehat{\mathit{RID}}_j(k)$ and $\mathit{RID}_j(k)$ are strictly increasing in $k \in [\phi_{\min}, \phi_{\max}]$. Then, the interquantile range (IQR) of $\widehat{\mathit{RID}}_j$ will converge in probability to the IQR of RID_j.*

Proof. Let $k_{0.25}$ be the k such that $\mathit{RID}_j(k_{0.25}) = 0.25$. And let $k_{0.75}$ be the k such that $\mathit{RID}_j(k_{0.75}) = 0.75$. Similarly, let $\hat{k}_{0.25}$ be the k such that $\widehat{\mathit{RID}}_j(\hat{k}_{0.25}) = 0.25$. And let $\hat{k}_{0.75}$ be the k such that $\widehat{\mathit{RID}}_j(\hat{k}_{0.75}) = 0.75$. The IQR of $\widehat{\mathit{RID}}_j$ converges to the IQR of RID_j if $\hat{k}_{0.25} \to k_{0.25}$ and $\hat{k}_{0.75} \to k_{0.75}$.

Because $\widehat{\mathit{RID}}_j(k)$ and $\mathit{RID}_j(k)$ are increasing in k, we know that if $\mathbb{P}\left(\widehat{\mathit{RIV}_j} \leq k_{0.25}\right) = 0.25$, then $\hat{k}_{0.25} = k_{0.25}$. An analogous statement holds for $\hat{k}_{0.75}$.

So, we will bound how far $\widehat{\mathit{RID}}_j(k_{0.25})$ is from $0.25 = \mathit{RID}_j(k_{0.25})$ and how far $\widehat{\mathit{RID}}_j(k_{0.75})$ is from $0.75 = \mathit{RID}_j(k_{0.75})$.

Let $t > 0$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\mathit{RID}}_j(k_{0.25}) - \mathit{RID}_j(k_{0.25})\right| + \left|\widehat{\mathit{RID}}_j(k_{0.75}) - \mathit{RID}_j(k_{0.75})\right| > t\right) \\
\leq&\mathbb{P}\left(\left\{\left|\widehat{\mathit{RID}}_j(k_{0.25}) - \mathit{RID}_j(k_{0.25})\right| > \frac{t}{2}\right\} \cup \left\{\left|\widehat{\mathit{RID}}_j(k_{0.75}) - \mathit{RID}_j(k_{0.75})\right| > \frac{t}{2}\right\}\right) \\
\leq&\mathbb{P}\left(\left\{\left|\widehat{\mathit{RID}}_j(k_{0.25}) - \mathit{RID}_j(k_{0.25})\right| > \frac{t}{2}\right\}\right) \\
&+ \mathbb{P}\left(\left\{\left|\widehat{\mathit{RID}}_j(k_{0.75}) - \mathit{RID}_j(k_{0.75})\right| > \frac{t}{2}\right\}\right) \text{ by Union bound.}
\end{aligned}$$

Then, by Theorem 2,

$$\mathbb{P}\left(\left|\widehat{\mathit{RID}}_j(k_{0.25}) - \mathit{RID}_j(k_{0.25})\right| > \frac{t}{2}\right) \leq 2\exp\left(-2B\frac{t^2}{4}\right).$$

So,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\textit{RID}}_j(k_{0.25}) - \textit{RID}_j(k_{0.25})\right| + \left|\widehat{\textit{RID}}_j(k_{0.75}) - \textit{RID}_j(k_{0.75})\right| > t\right) \\
\leq &\mathbb{P}\left(\left\{\left|\widehat{\textit{RID}}_j(k_{0.25}) - \textit{RID}_j(k_{0.25})\right|\right\} > \frac{t}{2}\right) \\
&+ \mathbb{P}\left(\left\{\left|\widehat{\textit{RID}}_j(k_{0.75}) - \textit{RID}_j(k_{0.75})\right|\right\} > \frac{t}{2}\right) \\
\leq &2\exp\left(-2B\frac{t^2}{4}\right) + 2\exp\left(-2B\frac{t^2}{4}\right) \\
= &4\exp\left(-2B\frac{t^2}{4}\right).
\end{aligned}
$$

So, as $B \to \infty$, $\mathbb{P}\left(\left|\widehat{\textit{RID}}_j(k_{0.25}) - \textit{RID}_j(k_{0.25})\right| + \left|\widehat{\textit{RID}}_j(k_{0.75}) - \textit{RID}_j(k_{0.75})\right| > t\right)$ ultimately converging to 0.

Therefore, the IQR of $\widehat{\textit{RID}}_j$ converges to the IQR of *RID*.

□

C Example Model Classes for Which *RID*Converges

First, recall the following assumption from the main paper:

Assumption 1. *If*

$$\rho\left(\textit{RLD}(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda),\textit{LD}^*(\cdot;\ell,n,\mathcal{P}_n,\lambda)\right)\leq\gamma \textit{ then}$$
$$\rho\left(\textit{RID}_j(\cdot;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda),\textit{RID}_j(\cdot;\varepsilon,\{g^*\},\ell,\mathcal{P}_n,\lambda)\right)\leq d(\gamma)$$

for a function $d:[0,\ell_{\max}-\ell_{\min}]\to[0,\phi_{\max}-\phi_{\min}]$ *such that* $\lim_{\gamma\to 0}d(\gamma)=0$. *Here,* ρ *represents any distributional distance metric (e.g., 1-Wasserstein).*

In this section, we highlight two simple examples of model classes and model reliance metrics for which Assumption 1 holds. First we show that Assumption 1 holds for the class of linear regression models with the model reliance metric being the coefficient assigned to each variable in Proposition 1; Proposition 2 presents a similar result for generalized additive models. We begin by presenting two lemmas which will help prove Proposition 1:

Lemma 1. *Let* ℓ *be unregularized mean square error, used as the objective for estimating optimal models in some class of continuous models* $\mathcal{F}$. *Assume that the DGP's noise* ϵ *is centered at 0:* $\mathbb{E}[\epsilon]=0$. *Define the function* $m:[0,\ell_{\max}]\to[0,1]$ *as:*

$$m(\varepsilon):=\lim_{n\to\infty}\int_{\ell_{\min}}^{\ell_{\max}}\left|\textit{LD}^*(k;\ell,n,\mathcal{P}_n,\lambda)-\textit{RLD}(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right|dk.$$

The function m *is a strictly increasing function of* ε; m *simply measures the integrated absolute error between the CDF of* g^**'s loss distribution and the CDF of the Rashomon set's loss distribution. Then, if* $g^*\in\mathcal{F}$, *then* $m(0)=0$.

Proof. Let ℓ be unregularized mean square error, used as the objective for estimating optimal models in some class of continuous models $\mathcal{F}$. Let g^* denote the unknown DGP. Throughout this proof, we consider the setting with $n\to\infty$, although we often omit this notation for simplicity.

First, we restate the definition of *RLD* and *LD** for reference:

$$\textit{RLD}(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda):=\mathbb{E}_{\mathcal{D}^{(n)}\sim\mathcal{P}_n}\left[\frac{\nu(\{f\in\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}}:\ell(f,\mathcal{D}^{(n)})\leq k\})}{\nu(\mathcal{R}^{\varepsilon}_{\mathcal{D}^{(n)}})}\right]$$

and

$$\textit{LD}^*(k;\ell,n,\mathcal{P}_n,\lambda):=\mathbb{E}_{\mathcal{D}^{(n)}\sim\mathcal{P}_n}\left[\mathbb{1}[\ell(g^*,\mathcal{D}^{(n)})\leq k]\right].$$

Because g^* is the DGP, we know that its expected loss should be lower than the expected loss for any other model in the model class: $\mathbb{E}_{\mathcal{D}^{(n)}\sim\mathcal{P}_n}[\ell(g^*,\mathcal{D}^{(n)})]\leq\mathbb{E}_{\mathcal{D}^{(n)}\sim\mathcal{P}_n}[\ell(f,\mathcal{D}^{(n)})]$ for any $f\in\mathcal{F}$ such that $f\neq g^*$, as we have assumed that any noise has expectation 0. For simplicity, we denote $\mathbb{E}_{\mathcal{D}^{(n)}\sim\mathcal{P}_n}[\ell(g^*,\mathcal{D}^{(n)})]$ by ℓ^*. We first show that m is monotonically increasing in ε by showing that, for any $\varepsilon>\varepsilon'\geq 0$:

$$\begin{aligned}&\lim_{n\to\infty}\int_{\ell_{\min}}^{\ell_{\max}}\left|\textit{LD}^*(k;\ell,n,\mathcal{P}_n,\lambda)-\textit{RLD}(k;\varepsilon,\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right|dk\\ >&\lim_{n\to\infty}\int_{\ell_{\min}}^{\ell_{\max}}\left|\textit{LD}^*(k;\ell,n,\mathcal{P}_n,\lambda)-\textit{RLD}(k;\varepsilon',\mathcal{F},\ell,\mathcal{P}_n,\lambda)\right|dk\end{aligned}$$

by demonstrating that the inequality holds for each individual value of k. First, note that:

$$\textit{LD}^*(k;\ell,n,\mathcal{P}_n,\lambda)=\mathbb{E}_{\mathcal{D}^{(n)}\sim\mathcal{P}_n}\left[\mathbb{1}[\ell(g^*,\mathcal{D}^{(n)})\leq k]\right].$$

As $n\to\infty$, this quantity approaches

$$\mathbb{E}_{\mathcal{D}^{(n)}\sim\mathcal{P}_n}\left[\mathbb{1}[\ell(g^*,\mathcal{D}^{(n)})\leq k]\right]=\mathbb{1}[\ell(g^*,\mathcal{D}^{(n)})\leq k].$$

We will consider three cases: first, we consider $\ell^*>k_1\geq 0$, followed by $\varepsilon'+\ell^*>k_2\geq\ell^*$, and finally $k_3\geq\varepsilon'+\ell^*$. Figure 7 provides a visual overview of these three cases and the broad idea within each case.

Case 1: $\ell^*>k_1\geq 0$

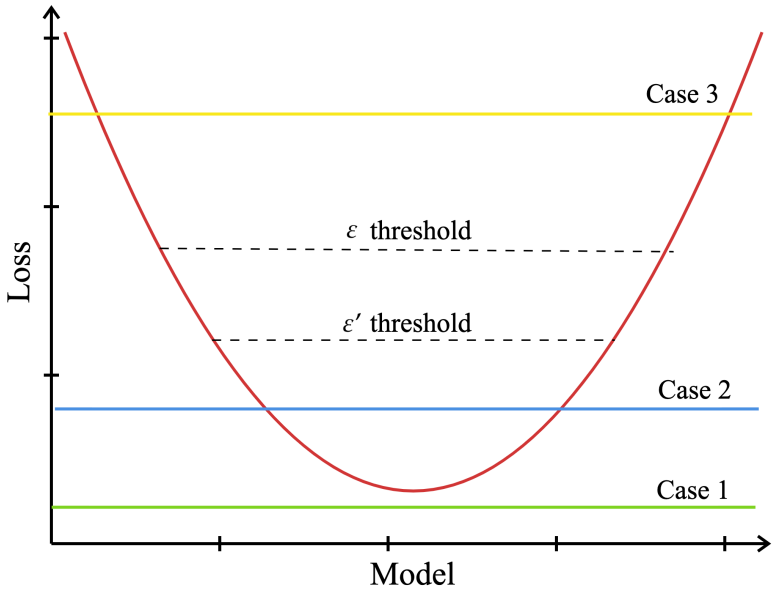


Figure 7: A visual overview of the proof of Lemma 1. In **Case 1**, we consider loss values that are achieved by no models in the model class, so each loss distribution has 0 mass below k in this case. **Case 2** covers each value of k such that k is larger than ℓ^*, so $LD^*(k) = 1$. The *RLD* for the ε' Rashomon set is closer to 1 than the ε Rashomon set because a larger proportion of this set falls below k. Under **Case 3**, all models in the ε' Rashomon set fall below k.

For any k_1 such that $\ell^* > k_1 \geq 0$, it holds that

$$\mathbb{1}[\ell(g^*, \mathcal{D}^{(n)}) \leq k_1] = 0,$$

since $\ell^* > k_1$ by definition. Further, because $\ell(g^*, \mathcal{D}^{(n)}) \leq \ell(f, \mathcal{D}^{(n)})$ for mean squared error in the infinite data setting,

$$\textit{RLD}(k_1; \varepsilon', \mathcal{F}, \ell, \mathcal{P}_n, \lambda) = \textit{RLD}(k_1; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda) = 0$$

Case 2: $\varepsilon' + \ell^* \geq k_2 \geq \ell^*$

For any k_2 such that $\varepsilon' + \ell^* \geq k_2 \geq \ell^*$,

$$\mathbb{1}[\ell(g^*, \mathcal{D}^{(n)}) \leq k_2] = 1,$$

since $\ell(g^*, \mathcal{D}^{(n)}) \leq k_2$ by the definition of k_2. Let ν denote a volume function over the target model class. Recalling that $\varepsilon > \varepsilon'$, we know that:

$$\begin{aligned}
\nu(\mathcal{R}^{\varepsilon}) > \nu(\mathcal{R}^{\varepsilon'}) &\iff \frac{1}{\nu(\mathcal{R}^{\varepsilon})} < \frac{1}{\nu(\mathcal{R}^{\varepsilon'})} \\
&\iff \frac{\nu(\{f \in \mathcal{R}^{\varepsilon} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon})} < \frac{\nu(\{f \in \mathcal{R}^{\varepsilon} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon'})} \\
&\iff \frac{\nu(\{f \in \mathcal{R}^{\varepsilon} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon})} < \frac{\nu(\{f \in \mathcal{R}^{\varepsilon'} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon'})},
\end{aligned}$$

since the set of models in the ε Rashomon set with loss less than k_2 is the same set as set of models in the ε' Rashomon set with loss less than k_2 for $k_2 \leq \varepsilon' + \ell^*$. We can further manipulate this quantity to show:

$$\begin{aligned}
&\frac{\nu(\{f \in \mathcal{R}^{\varepsilon} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon})} < \frac{\nu(\{f \in \mathcal{R}^{\varepsilon'} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon'})} \\
\iff & 1 - \frac{\nu(\{f \in \mathcal{R}^{\varepsilon} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon})} > 1 - \frac{\nu(\{f \in \mathcal{R}^{\varepsilon'} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon'})} \\
\iff & \left|1 - \frac{\nu(\{f \in \mathcal{R}^{\varepsilon} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon})}\right| > \left|1 - \frac{\nu(\{f \in \mathcal{R}^{\varepsilon'} : \ell(f, \mathcal{D}^{(n)}) \leq k_2\})}{\nu(\mathcal{R}^{\varepsilon'})}\right| \\
\iff & |1 - \textit{RLD}(k_2; \varepsilon, \mathcal{F}, \ell)| > \left|1 - \textit{RLD}(k_2; \varepsilon', \mathcal{F}, \ell)\right| \\
\iff & |\textit{LD}^*(k_2) - \textit{RLD}(k_2; \varepsilon, \mathcal{F}, \ell)| \\
& \quad > \left|\textit{LD}^*(k_2) - \textit{RLD}(k_2; \varepsilon', \mathcal{F}, \ell)\right|,
\end{aligned}$$

because $\textit{LD}^*(k_2) = 1$.

Case 3: $k_3 > \varepsilon' + \ell^*$

For any $k_3 > \varepsilon' + \ell^*$, we have

$$\begin{aligned}
\textit{RLD}(k_3; \varepsilon', \mathcal{F}, \ell) &= \frac{\nu(\{f \in \mathcal{R}^{\varepsilon'} : \ell(f, \mathcal{D}^{(n)}) \leq k_3\})}{\nu(\mathcal{R}^{\varepsilon'})} \\
&= \frac{\nu(\mathcal{R}^{\varepsilon'})}{\nu(\mathcal{R}^{\varepsilon'})} && \text{because } k_3 > \varepsilon' + \ell^* \\
&= 1.
\end{aligned}$$

This immediately gives that

$$\begin{aligned}
\left|\textit{LD}^*(k_3) - \textit{RLD}(k_3; \varepsilon', \mathcal{F}, \ell)\right| &= |1 - 1| \\
&= 0,
\end{aligned}$$

the minimum possible value for this quantity. We can then use the fact that the absolute value is greater than or equal to zero to show that

$$\begin{aligned}
&|\textit{LD}^*(k_3) - \textit{RLD}(k_3; \varepsilon, \mathcal{F}, \ell)| \\
&\geq 0 = \left|\textit{LD}^*(k_3) - \textit{RLD}(k_3; \varepsilon', \mathcal{F}, \ell)\right|
\end{aligned}$$

In summary, under cases 1 and 3,

$$\begin{aligned}
&|\textit{LD}^*(k; \ell, n, \mathcal{P}_n, \lambda) - \textit{RLD}(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)| \\
&\geq \left|\textit{LD}^*(k; \ell, n, \mathcal{P}_n, \lambda) - \textit{RLD}(k; \varepsilon', \mathcal{F}, \ell, \mathcal{P}_n, \lambda)\right|;
\end{aligned}$$

under case 2,

$$\begin{aligned}
&|\textit{LD}^*(k_2; \ell, n, \mathcal{P}_n, \lambda) - \textit{RLD}(k_2; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)| \\
&> \left|\textit{LD}^*(k_2; \ell, n, \mathcal{P}_n, \lambda) - \textit{RLD}(k_2; \varepsilon', \mathcal{F}, \ell, \mathcal{P}_n, \lambda)\right|.
\end{aligned}$$

Since there is some range of values $k \in [\ell^*, \varepsilon' + \ell^*)$ for which the inequality above is strict, it follows that

$$\begin{aligned}
&\int_{\ell_{\min}}^{\ell_{\max}} |\textit{LD}^*(k; \ell, n, \mathcal{P}_n, \lambda) - \textit{RLD}(k; \varepsilon, \mathcal{F}, \ell, \mathcal{P}_n, \lambda)|\, dk \\
&> \int_{\ell_{\min}}^{\ell_{\max}} \left|\textit{LD}^*(k; \ell, n, \mathcal{P}_n, \lambda) - \textit{RLD}(k; \varepsilon', \mathcal{F}, \ell, \mathcal{P}_n, \lambda)\right| dk,
\end{aligned}$$

showing that $\varepsilon > \varepsilon'$ is a *sufficient* condition for $m(\varepsilon) > m(\varepsilon')$. Observe that, for a loss function with no regularization and a fixed model class, *RLD* is a function of only ε. As such, varying ε is the only way to vary *RLD*, making $\varepsilon > \varepsilon'$ a *necessary* condition for the above. Therefore, we have shown that $\varepsilon > \varepsilon' \iff m(\varepsilon) > m(\varepsilon')$, i.e. m is strictly increasing.

Further, if $g^* \in \mathcal{F}$, the Rashomon set with $\varepsilon = 0$ will contain only g^* as n approaches infinity, immediately yielding that

$$m(0) = \int_{\ell_{\min}}^{\ell_{\max}} |\textit{LD}^*(k; \ell, n, \mathcal{P}_n, \lambda) - \textit{RLD}(k; 0, \mathcal{F}, \ell)|\, dk = 0.$$

□

Lemma 1 provides a mechanism through which *RLV* will approach *LD** in the infinite data setting. The following lemma states that each level set of the quadratic loss surface is a hyper-ellipsoid, providing another useful tool for the propositions given in this section.

Lemma 2. *The level set of the quadratic loss at ε is a hyper-ellipsoid defined by:*

$$(\theta - \theta^*)^T X^T X(\theta - \theta^*) = \varepsilon - c,$$

which is centered at θ^* and of constant shape in terms of ε.

Proof. Recall that the quadratic loss for some parameter vector θ is given by:

$$\ell(\theta) = \|y - X\theta\|^2$$

and that the optimal vector θ^* is given by:

$$\theta^* = (X^T X)^{-1} X^T y$$
$$\iff X^T X\theta^* = X^T y$$

With these facts, we show that the level set for the quadratic loss at some fixed value ε takes on the standard form for a hyper-ellipsoid. This is shown as:

$$\begin{aligned}
\ell(\theta) &= \|y - X\theta\|^2 \\
&= \|y - X\theta\|^2 \underbrace{-y^T(y - X\theta^*) + y^T(y - X\theta^*)}_{\text{add 0}} \\
&= \underbrace{y^T y - 2y^T X\theta + \theta^T X^T X\theta}_{\text{expand quadratic}} \underbrace{-y^T y - y^T X\theta^*}_{\text{distribute } y^T} + y^T(y - X\theta^*) \\
&= y^T y - 2y^T X\theta + \theta^T X^T X\theta - y^T y - \underbrace{(X^T y)^T \theta^*}_{\text{pull out transpose}} + y^T(y - X\theta^*) \\
&= y^T y - 2y^T X\theta + \theta^T X^T X\theta - y^T y - \underbrace{\theta^{*T} X^T X\theta^*}_{\text{because } X^T y = X^T X\theta^*} + y^T(y - X\theta^*) \\
&= y^T y - \underbrace{2(X^T y)^T \theta}_{\text{pull out transpose}} + \theta^T X^T X\theta - y^T y - \theta^{*T} X^T X\theta^* + y^T(y - X\theta^*) \\
&= y^T y - \underbrace{2\theta^* X^T X\theta}_{\text{because } X^T y = X^T X\theta^*} + \theta^T X^T X\theta - y^T y - \theta^{*T} X^T X\theta^* + y^T(y - X\theta^*) \\
&= \theta^T X^T X\theta - 2\theta^* X^T X\theta - \theta^{*T} X^T X\theta^* + y^T(y - X\theta^*) \text{ because } y^T y \text{ terms cancel out} \\
&= (\theta - \theta^*)^T X^T X(\theta - \theta^*) + y^T(y - X\theta^*) \text{ by factorization.}
\end{aligned}$$

Noting that the term $y^T(y - X\theta^*)$ is constant in terms of θ, so we can simplify this expression to $\ell(\theta) = (\theta - \theta^*)^T X^T X(\theta - \theta^*) + c$ where $c = y^T(y - X\theta^*)$. If we are interested in the level set at $\ell(\theta) = c + \varepsilon$ — that is, with loss ε greater than the optimal loss — this is exactly:

$$(\theta - \theta^*)^T X^T X(\theta - \theta^*) + c = c + \varepsilon$$
$$\iff (\theta - \theta^*)^T X^T X(\theta - \theta^*) = \varepsilon.$$

That is, the set of parameters θ yielding loss value $c + \varepsilon$ is a hyper-ellipsoid centered at θ^* according to the positive semi-definite matrix $X^T X$. □

Proposition 1. *If the DGP is a linear regression model, Assumption 1 is guaranteed to hold for the function class of linear models (i.e., $g^* \in \mathcal{F}$) as $n \to \infty$.*

Proof. We now turn our attention to *RID*. Let our variable importance metric $\phi_j := \theta_j$, the coefficient of a linear model, and let p denote the number of variables in the dataset such that $\boldsymbol{\theta} \in \mathbb{R}^p$. As in Lemma 1, we restrict ourselves to the setting in which $n \to \infty$, although we often omit this notation. Define the function $r_j : [0, \ell_{\max}] \to [0, 1]$ to be:

$$r_j(\varepsilon) := \int_{\phi_{\min}}^{\phi_{\max}} |\textit{RID}_j(k; \{g^*\}, 0) - \textit{RID}_j(k; \mathcal{F}, \varepsilon)|\, dk$$

We show that r_j is a monotonic function of ε, for any $j \in \{1, 2, \ldots, p\}$. In other words, as ε gets smaller, the value of $r_j(\varepsilon)$ gets smaller. We do so by showing that the following holds for this VI metric:

$$\begin{aligned}
&\int_{\phi_{\min}}^{\phi_{\max}} |\textit{RID}_j(k; \{g^*\}, 0) - \textit{RID}_j(k; \mathcal{F}, \varepsilon)|\, dk \\
&\geq \int_{\phi_{\min}}^{\phi_{\max}} \left|\textit{RID}_j(k; \{g^*\}, 0) - \textit{RID}_j(k; \mathcal{F}, \varepsilon')\right| dk
\end{aligned}$$

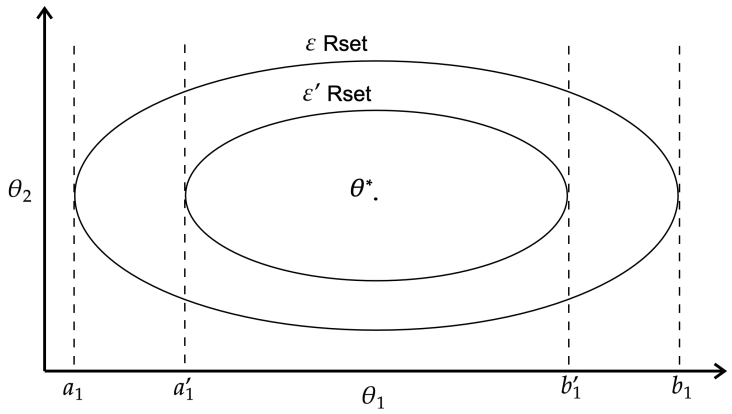


Figure 8: A visualization of the ε and ε' Rashomon sets for linear regression with two input features. We highlight the extrema of each Rashomon set along axis 1 (a_1 and b_1 for the ε Rashomon set, a'_1 and b'_1 for the ε' Rashomon set).

if and only if $\varepsilon > \varepsilon'$ by showing that, for any k,

$$\begin{aligned}&|\textit{RID}_j(k; \{g^*\}, 0) - \textit{RID}_j(k; \mathcal{F}, \varepsilon)| \\ &\geq \left|\textit{RID}_j(k; \{g^*\}, 0) - \textit{RID}_j(k; \mathcal{F}, \varepsilon')\right|.\end{aligned}$$

For simplicity of notation, we denote the linear regression model parameterized by some coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^p$ simply as $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^* \in \mathbb{R}^p$ denote the coefficient vector for the optimal model. Additionally, we define the following quantities to represent the most extreme values for θ_j (i.e., the coefficient along the j-th axis) for each Rashomon set. Let a_j and b_j be the two values defined as:

$$\begin{aligned}a_j &:= \min_{\mathbf{v} \in \mathbb{R}^p} (\boldsymbol{\theta}^* + \mathbf{v})_j \text{ s.t. } \ell(\boldsymbol{\theta}^* + \mathbf{v}, \mathcal{D}^{(n)}) = \ell^* + \varepsilon \\ b_j &:= \max_{\mathbf{v} \in \mathbb{R}^p} (\boldsymbol{\theta}^* + \mathbf{v})_j \text{ s.t. } \ell(\boldsymbol{\theta}^* + \mathbf{v}, \mathcal{D}^{(n)}) = \ell^* + \varepsilon.\end{aligned}$$

Similarly, let a'_j and b'_j be the two values defined as:

$$\begin{aligned}a'_j &:= \min_{\mathbf{v} \in \mathbb{R}^p} (\boldsymbol{\theta}^* + \mathbf{v})_j \text{ s.t. } \ell(\boldsymbol{\theta}^* + \mathbf{v}, \mathcal{D}^{(n)}) = \ell^* + \varepsilon' \\ b'_j &:= \max_{\mathbf{v} \in \mathbb{R}^p} (\boldsymbol{\theta}^* + \mathbf{v})_j \text{ s.t. } \ell(\boldsymbol{\theta}^* + \mathbf{v}, \mathcal{D}^{(n)}) = \ell^* + \varepsilon'.\end{aligned}$$

Intuitively, these values represent the most extreme values of $\boldsymbol{\theta}$ along dimension j that are still included in their respective Rashomon sets. Figure 8 provides a visual explanation of each of these quantities. Finally, recall that:

$$\textit{RID}_j(k; \{g^*\}, 0) = \begin{cases} 1 & \text{if } \theta^*_j \leq k \\ 0 & \text{otherwise,} \end{cases}$$

since $\boldsymbol{\theta}^*$ is a deterministic quantity given infinite data.

Without loss of generality, we will consider two cases:

1. The case where $\theta^*_j \leq k$,
2. The case where $k < \theta^*_j$.

Figures 9 and 10 give an intuitive overview of the mechanics of this proof. As depicted in Figure 9, we will show that the proportion of the volume of the ε'-Rashomon set with ϕ_j below k is closer to 1 than that of the ε-Rashomon set under case 1. We will than show that the opposite holds under case 2, as depicted in Fugre 10.

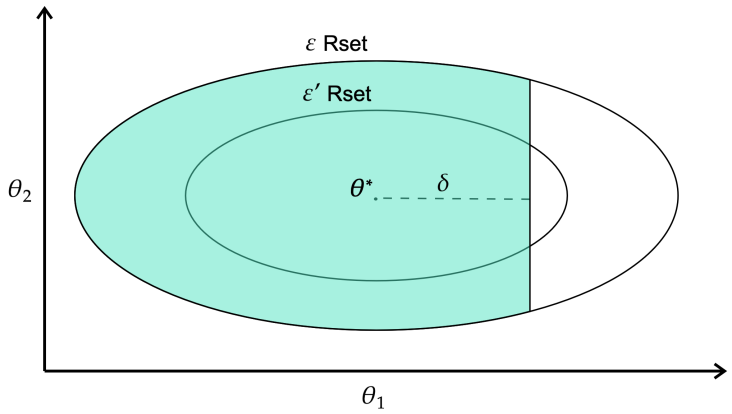


Figure 9: A simple illustration of the key idea in case 1 of the proof of Proposition 1. For two concentric ellipsoids of the same shape, the proportion of each ellipsoid's volume falling below some point greater than the center along axis j is greater for the smaller ellipsoid than for the larger ellipsoid.

Case 1: $\theta_j^* \leq k$

Define two functions $h : [a_j, b_j] \to [0, 1]$ and $h' : [a'_j, b'_j] \to [0, 1]$ as:

$$h(c) = \frac{c - a_j}{b_j - a_j}$$
$$h'(c) = \frac{c - a'_j}{b'_j - a'_j}.$$

These functions map each value c in the original space of θ_j to its *relative position* along each axis of the ε-Rashomon set and the ε'-Rashomon set respectively, with $h(b_j) = h'(b'_j) = 1$ and $h(a_j) = h'(a'_j) = 0$.

Define $\delta \in [0, b_j - \theta_j^*]$ to be the value such that $k = \theta_j^* + \delta$. Since in this case $\theta_j^* \leq k$, it follows that $\delta \geq 0$. As such, we can then quantify the proportion of the ε-Rashomon set along the j-th axis such that $\theta_j^* \leq \theta_j \leq k$ as:

$$\begin{aligned} h(\theta_j^* + \delta) - h(\theta_j^*) &= \frac{(\theta_j^* + \delta) - a_j}{b_j - a_j} - \frac{(\theta_j^* - a_j)}{(b_j - a_j)} \\ &= \frac{\theta_j^* + \delta - a_j - \theta_j^* + a_j}{b_j - a_j} \\ &= \frac{\delta}{b_j - a_j} \end{aligned}$$

Similarly, we can quantify the proportion of the ε'-Rashomon set along the j-th axis with θ_j between k and θ_j^* as:

$$\begin{aligned} h'(\delta + \theta_j^*) - h'(\theta_j^*) &= \frac{\theta_j^* + \delta - a'_j - \theta_j^* + a'_j}{b'_j - a'_j} \\ &= \frac{\delta}{b'_j - a'_j}. \end{aligned}$$

Recalling that, by definition, $a_j < a'_j < b'_j < b_j$, as well as the fact that $\delta \geq 0$ we can see that:

$$\begin{aligned} b_j - a_j > b'_j - a'_j &\iff \frac{1}{b_j - a_j} < \frac{1}{b'_j - a'_j} \\ &\iff \frac{\delta}{b_j - a_j} \leq \frac{\delta}{b'_j - a'_j} \\ &\iff h(\theta_j^* + \delta) - h(\theta_j^*) \leq h'(\theta_j^* + \delta) - h'(\theta_j^*) \\ &\iff h(k) - h(\theta_j^*) \leq h'(k) - h'(\theta_j^*). \end{aligned}$$

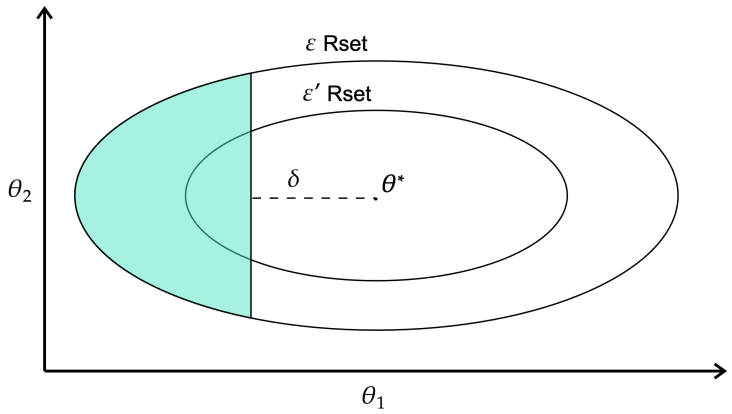


Figure 10: A simple illustration of the key idea in case 2 of the proof of Proposition 1. For two concentric ellipsoids of the same shape, the proportion of each ellipsoid's volume falling below some point less than the center along axis j is smaller for the smaller ellipsoid than for the larger ellipsoid.

That is, the proportion of the ε-Rashomon set along the j-th axis with θ_j between k and θ_j^* is *less than or equal to* the proportion of the ε'-Rashomon set along the j-th axis with θ_j between k and θ_j^*. By Lemma 2, recall that the ε-Rashomon set and the ε'-Rashomon set are concentric (centered at θ^*) and similar (with shape defined by X^TX). Let ν denote the volume function for some subsection of a hyper-ellipsoid. We then have

$$
\begin{aligned}
& h(k) - h(\theta_j^*) \leq h'(k) - h'(\theta_j^*) \\
& \iff \frac{\nu(\{\boldsymbol{\theta} \in \mathcal{R}^{\varepsilon} : \theta_j^* \leq \theta_j \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon}\})} \leq \frac{\nu(\{\boldsymbol{\theta}' \in \mathcal{R}^{\varepsilon'} : \theta_j^* \leq \theta_j' \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon'}\})} \\
& \iff \frac{1}{2} + \frac{\nu(\{\boldsymbol{\theta} \in \mathcal{R}^{\varepsilon} : \theta_j^* \leq \theta_j \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon}\})} \leq \frac{1}{2} + \frac{\nu(\{\boldsymbol{\theta}' \in \mathcal{R}^{\varepsilon'} : \theta_j^* \leq \theta_j' \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon'}\})} \\
& \iff \frac{\nu(\{\boldsymbol{\theta} \in \mathcal{R}^{\varepsilon} : \theta_j \leq \theta_j^*\})}{\nu(\{\mathcal{R}^{\varepsilon}\})} + \frac{\nu(\{\boldsymbol{\theta} \in \mathcal{R}^{\varepsilon} : \theta_j^* \leq \theta_j \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon}\})} \\
& \qquad \leq \frac{\nu(\{\boldsymbol{\theta}' \in \mathcal{R}^{\varepsilon'} : \theta_j' \leq \theta_j^*\})}{\nu(\{\mathcal{R}^{\varepsilon'}\})} + \frac{\nu(\{\boldsymbol{\theta}' \in \mathcal{R}^{\varepsilon'} : \theta_j^* \leq \theta_j' \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon'}\})} \\
& \iff \frac{\nu(\{\boldsymbol{\theta} \in \mathcal{R}^{\varepsilon} : \theta_j \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon}\})} \leq \frac{\nu(\{\boldsymbol{\theta}' \in \mathcal{R}^{\varepsilon'} : \theta_j' \leq k\})}{\nu(\{\mathcal{R}^{\varepsilon'}\})}.
\end{aligned}
$$

Recalling that, by definition, $\mathit{RID}_j(k; \mathcal{F}, \varepsilon') = \frac{\nu(\{\theta' \in \mathcal{R}^{\varepsilon'} : \theta_j' \leq k\})}{\nu(\mathcal{R}^{\varepsilon'})}$, it follows that:

$$
\begin{aligned}
& \mathit{RID}_j(k; \mathcal{F}, \varepsilon) \leq \mathit{RID}_j(k; \mathcal{F}, \varepsilon') \\
& \iff 1 - \mathit{RID}_j(k; \mathcal{F}, \varepsilon) \geq 1 - \mathit{RID}_j(k; \mathcal{F}, \varepsilon') \\
& \iff |1 - \mathit{RID}_j(k; \mathcal{F}, \varepsilon)| \geq |1 - \mathit{RID}_j(k; \mathcal{F}, \varepsilon')|.
\end{aligned}
$$

Recalling that $\mathit{RID}_j(k; \{g^*\}, 0) = 1$, since $k \geq \theta_j^*$, the above gives:

$$
\begin{aligned}
& |1 - \mathit{RID}_j(k; \mathcal{F}, \varepsilon)| \geq |1 - \mathit{RID}_j(k; \mathcal{F}, \varepsilon')| \\
& \iff |\mathit{RID}_j(k; \{g^*\}, \varepsilon) - \mathit{RID}_j(k; \mathcal{F}, \varepsilon)| \\
& \qquad \geq |\mathit{RID}_j(k; \{g^*\}, \varepsilon) - \mathit{RID}_j(k; \mathcal{F}, \varepsilon')|
\end{aligned}
$$

for all $\theta_j^* \leq k$.

Case 2: $k < \theta_j^*$

Let h and h' be defined as in Case 1. Define $\delta \in [a_j - \theta_j^*, 0]$ to be the quantity such that $k = \theta_j^* + \delta$. In this case, $k < \theta_j^*$, so it follows that $\delta < 0$. Repeating the derivation from Case 1, we then have:

$$\begin{aligned} b_j - a_j > b'_j - a'_j &\iff \frac{1}{b_j - a_j} < \frac{1}{b'_j - a'_j} \\ &\iff \frac{\delta}{b_j - a_j} > \frac{\delta}{b'_j - a'_j} \\ &\iff h(\theta_j^* + \delta) - h(\theta_j^*) > h'(\theta_j^* + \delta) - h'(\theta_j^*) \\ &\iff h(k) - h(\theta_j^*) > h'(k) - h'(\theta_j^*). \end{aligned}$$

That is, the proportion of the ε-Rashomon set along the j-th axis with θ_j between k and θ_j^* is *greater than* the proportion of the $\varepsilon'-$Rashomon set along the j-th axis with θ_j between k and θ_j^*. By similar reasoning as in Case 1, it follows that:

$$\begin{aligned} &\textit{RID}_j(k; \mathcal{F}, \varepsilon) > \textit{RID}_j(k; \mathcal{F}, \varepsilon') \\ \iff &|\textit{RID}_j(k; \mathcal{F}, \varepsilon) - 0| > |\textit{RID}_j(k; \mathcal{F}, \varepsilon') - 0| \end{aligned}$$

Recalling that $\textit{RID}_j(k; \{g^*\}, \varepsilon) = 0$, since $k < \theta_j^*$, the above gives:

$$\begin{aligned} |\mathbb{P}_{\mathcal{D}^{(n)} \sim \mathcal{P}_n}(\textit{RIV}_j(\mathcal{F}, \varepsilon) \leq k) - 0| &> |\textit{RID}_j(k; \mathcal{F}, \varepsilon') - 0| \\ \iff |\textit{RID}_j(k; \mathcal{F}, \varepsilon) - \textit{RID}_j(k; \{g^*\}, 0)| &> |\textit{RID}_j(k; \mathcal{F}, \varepsilon') - \textit{RID}_j(k; \{g^*\}, 0)| \end{aligned}$$

for all $a_j \leq k < \theta_j^*$. As such, for any k, we have that:

$$\begin{aligned} &|\textit{RID}_j(k; \{g^*\}, 0) - \textit{RID}_j(k; \mathcal{F}, \varepsilon)| \\ &\geq \left|\textit{RID}_j(k; \{g^*\}, 0) - \textit{RID}_j(k; \mathcal{F}, \varepsilon')\right|, \end{aligned}$$

showing that $\varepsilon > \varepsilon'$ is a *sufficient* condition for the above. Since *RID* is a function of only ε, varying ε is the only way to vary *RID*, making $\varepsilon > \varepsilon'$ a *necessary* condition for the above, yielding that $r_j(\varepsilon) > r_j(\varepsilon') \iff \varepsilon > \varepsilon'$ and r_j is monotonically increasing.

Let m be defined as in Lemma 1, and let γ be some value such that $m(\varepsilon) \leq \gamma$. Define the function $d := r_j \circ m^{-1}$ (note that m^{-1}, the inverse of m, is guaranteed to exist and be strictly increasing because m is strictly increasing). The function d is monotonically increasing as the composition of two monotonically increasing functions, and:

$$\begin{aligned} &m(\varepsilon) \leq \gamma \\ \iff &\varepsilon \leq m^{-1}(\gamma) \\ \iff &r_j(\varepsilon) \leq d(\gamma) \end{aligned}$$

as required.

Further, Lemma 1 states that $m(0) = 0$ if $g^* \in \mathcal{F}$. Note also that the Rashomon set with $\varepsilon = 0$ contains only g^*, and as such $r_j(0) = d(m^{-1}(0)) = 0$, meaning $d(0) = 0$. Therefore $\lim_{\gamma \to 0} d(\gamma) = 0$.

□

Proposition 2. *Assume the DGP is a generalized additive model (GAM). Then, Assumption 1 is guaranteed to hold for the function class of GAM's where our variable importance metric is the coefficient on each bin.*

Proof. Recall from Proposition 1 that Assumption 1 holds for the class of linear regression models with the model reliance metric $\phi_j = \theta_j$. A generalized additive model (GAM) [**?**] over p variables is generally represented as:

$$g(\mathbb{E}[Y]) = \omega + f_1(x_1) + \ldots + f_p(x_p),$$

where g is some link function, ω is a bias term, and $f_1, \ldots, f_p$ denote the shape functions associated with each of the variables. In practice, each shape function f_j generally takes the form of a linear function over binned variables [**?**]:

$$f_j(x_i) = \sum_{j'=0}^{\beta_j - 1} \theta_{j'} \mathbb{1}[b_{j'} \leq x_{ij} \leq b_{j'+1}],$$

where β_j denotes the number of possible bins associated with variable X_j, $b_{j'}$ denotes the j'-th cuttoff point associated with X_j, and $\theta_{j'}$ denotes the weight associated with the j'-th bin on variable X_j. With the above shape function, a GAM is a linear regression over a binned dataset; as such, for the variable importance metric $\phi_{j'} = \theta_{j'}$ on the complete, binned dataset, Assumption 1 holds by the same reasoning as Proposition 1. □

D Detailed Experimental Setup

In this work, we considered the following four simulation frameworks:

- Chen's [9]: $Y = \mathbb{1}[-2\sin(X_1) + \max(X_2, 0) + X_3 + \exp(-X_4) + \varepsilon \geq 2.048]$, where $X_1, \ldots, X_{10}, \varepsilon \sim \mathcal{N}(0, 1)$. Here, only $X_1, \ldots, X_4$ are relevant.
- Friedman's [17]: $Y = \mathbb{1}[10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon \geq 15]$, where $X_1, \ldots, X_6 \sim \mathcal{U}(0, 1), \varepsilon \sim \mathcal{N}(0, 1)$. Here, only $X_1, \ldots, X_5$ are relevant.
- Monk 1 [42]: $Y = \max\left(\mathbb{1}[X_1 = X_2], \mathbb{1}[X_5 = 1]\right)$, where the variables $X_1, \ldots, X_6$ have domains of 2, 3, or 4 unique integer values. Only X_1, X_2, X_5 are important.
- Monk 3 [42]: $Y = \max\left(\mathbb{1}[X_5 = 3 \text{ and } X_4 = 1], \mathbb{1}[X_5 \neq 4 \text{ and } X_2 \neq 3]\right)$ for the same covariates in Monk 1. Here, X_2, X_4, and X_5 are relevant, and 5% label noise is added.

DGP	Num Samples	Num Features	Num Extraneous Features
Chen's	1,000	10	6
Friedman's	200	6	1
HIV	14,742	100	Unknown
Monk 1	124	6	3
Monk 3	124	6	3

Table 1: Overview of the size of each dataset considered (or generated from a DGP) in this paper.

For our experiments in Sections 4.1 and 4.2 of the main paper, we trained and evaluated all models using the standard training set provided by [42] for Monk 1 and Monk 3. We generated 200 samples following the above process for Friedman's DGP, and 1000 samples following the above process for Chen's DGP.

In Section 5 of the main paper, we evaluated *RID*on a dataset studying which host cell transcripts and chromatin patterns are associated with high expression of Human Immunodeficiency Virus (HIV) RNA. We used the model class of sparse decision trees and subtractive model reliance. The dataset combined single cell RNAseq/ATACseq profiles for 74,031 individual HIV infected cells from two different donors in the aims of finding new cellular cofactors for HIV expression that could be targeted to reactivate the latent HIV reservoir in people with HIV (PWH). A longer description of the data is in [29].

We consider the binary classification problem of predicting high versus low HIV load, where high HIV load means an HIV load in the top 10% of observed values. We selected 14,614 samples (all 7,307 high HIV load samples and 7,307 random low HIV load samples) from the overall dataset in order to balance labels, and filtered the complete profiles down to the top 100 variables by individual AUC in order to accelerate the runtime of *RID*.

Table 1 summarizes the size of each dataset we considered. In all cases, we used random seed 0 for dataset generation, model training, and evaluation unless otherwise specified.

We compared the rankings produced by *RID* with the following baseline methods:

- Subtractive model reliance ϕ^{sub} of a random forest (RF) [6] using scikit-learn's implementation [**?**] of RF
- Subtractive model reliance ϕ^{sub} of an L1 regularized logistic regression model (Lasso) using scikit-learn's implementation [**?**] of Lasso
- Subtractive model reliance ϕ^{sub} of boosted decision trees [16] using scikit-learn's implementation [**?**] of AdaBoost
- Subtractive model reliance ϕ^{sub} of a generalized optimal sparse decision tree (GOSDT) [26] using the implementation from [49]
- Subtractive conditional model reliance (CMR) [15] – a metric designed to capture only the unique information of a variable – of RF using scikit-learn's implementation [**?**] of RF
- Subtractive conditional model reliance (CMR) [15] of Lasso using scikit-learn's implementation [**?**] of Lasso
- The impurity based model reliance metric for RF from [7] using scikit-learn's implementation [**?**] of RF
- The LOCO algorithm reliance [25] value for RF and for Lasso using scikit-learn's implementation [**?**] of both models
- The Pearson correlation between each feature and the outcome
- The Spearman correlation between each feature and the outcome

Dataset	Rashomon Threshold ε	Regularization Weight λ	Depth Bound
Chen's	0.01	0.01	5
Friedman's	0.025	0.02	6
HIV	0.075	0.005	3
Monk 1	0.1	0.03	5
Monk 3	0.05	0.025	7

Table 2: The parameters used for *RID*, VIC, and GOSDT by data generation process.

- The mean of the partial dependency plot (PDP) [19] for each feature using scikit-learn's implementation [**?**]
- The SHAP value [28] for RF using scikit-learn's implementation [**?**] of RF
- The mean of variable importance clouds (VIC) [12] for the Rashomon set of sparse decision trees, computed using TreeFarms [49].

We used the default parameters in scikit-learn's implementation [**?**] of each baseline model. The parameters used for *RID*, VIC, and GOSDT for each dataset are summarized in Table 2. In all cases, we constructed each of *RID*, VIC, and GOSDT using the code from [49].

D.1 Computational Resources

All experiments for this work were performed on an academic institution's cluster computer. We used up to 40 machines in parallel, selected from the specifications below:

- 2 Dell R610's with 2 E5540 Xeon Processors (16 cores)
- 10 Dell R730's with 2 Intel Xeon E5-2640 Processors (40 cores)
- 10 Dell R610's with 2 E5640 Xeon Processors (16 cores)
- 10 Dell R620's with 2 Xeon(R) CPU E5-2695 v2's (48 cores)
- 8 Dell R610's with 2 E5540 Xeon Processors (16 cores)

We did not use GPU acceleration for this work.

E Additional Experiments

E.1 Recovering MR without Bootstrapping Baseline Methods

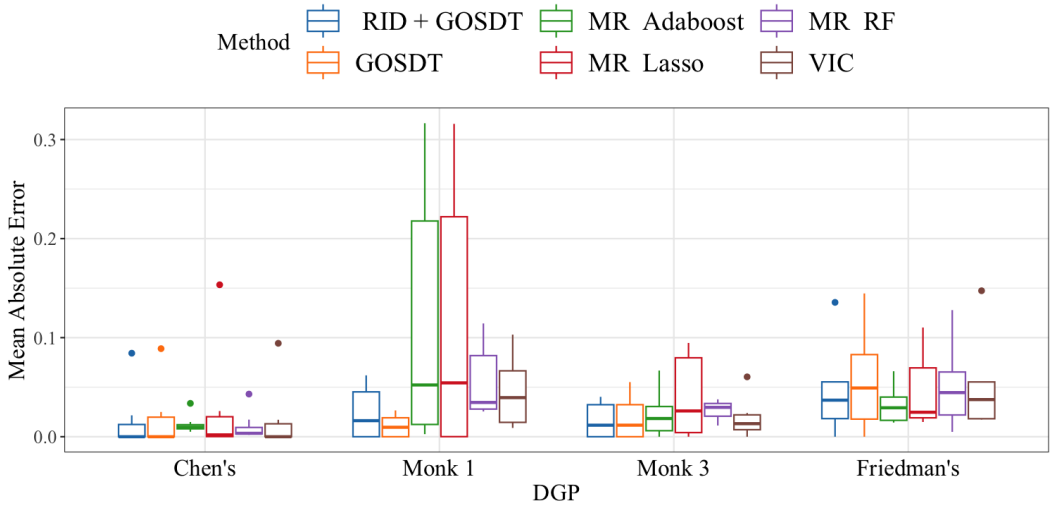


Figure 11: Boxplot over variables of the mean absolute error over test sets between the MR value produced by each method without bootstrapping (except *RID*) and the model reliance of the DGP for 500 test sets.

In this section, we evaluate the ability of each baseline method to recover the value of subtractive model reliance for the data generation process *without bootstrapping*. For this comparison, we use one training set to find the model reliance of each variable for each of the following algorithms: GOSDT, AdaBoost, Lasso, and Random Forest. Because *RID*and VIC produce distributions/samples, we instead estimate the *median* model reliance across *RID*and VIC's model reliance distributions.

We then sample 500 test sets independently for each DGP. We then calculate the model reliance for each test set using the DGP as if it were a predictive model (that is, if the DGP were $Y = X + \varepsilon$ for some Gaussian noise ε, our predictive model would simply be $f(X) = X$). Finally, we calculate the mean absolute error between the test model reliance values for the DGP and the train model reliance values for each algorithm.

Figure 11 shows the results of this experiment. As Figure 11 illustrates, *RID* produces more accurate point estimates than baseline methods even though this is not the goal of *RID*– the goal of *RID* is to produce *the entire distribution* of model reliance across good models over bootstrap datasets, not a single point estimate.

E.2 Width of Box and Whisker Ranges

When evaluating whether the box and whisker range (BWR) for each method captures the MR value for the DGP across test sets, a natural question is whether *RID*outperforms other methods simply because it produces wider BWR's. Figure 12 demonstrates the width of the BWR produced by each evaluated method across variables and datasets. As shown in Figure 12, *RID* consistently produces BWR widths on par with baseline methods.

E.3 The Performance of *RID* is Stable Across Reasonable Values for ε

The parameter ε controls what the maximum possible loss a model in the Rashomon set could be. We investigate whether this choice of ε significantly alters the performance of *RID*. In order to investigate this question, we repeat the coverage experiment from Section 4.2 of the main paper for three different values of ε for each dataset on VIC and *RID*(the two methods effected by ε). In particular, we construct the BWR over 100 bootstrap iterations for *RID*and over models for VIC for three different values of ε on each training dataset. These values are chosen as $0.75\varepsilon^*$, ε^*, and $1.25\varepsilon^*$, where ε^* denotes the value of ε used in the experiments presented in the main paper. We then generate 500 test datasets for each DGP and evaluate the subtractive model reliance for the DGP on each variable; we then measure what proportion of these test model reliance values are contained in each BWR. We refer to this proportion as the "recovery percentage".

Figure 13 illustrates that ***RID* is almost entirely invariant to reasonable choices of** ε: the recovery proportion for *RID* ranges from 90.38% to 90.64% on Chen's DGP, 100% to 100% on Monk 1, 99.43% to 99.93% on

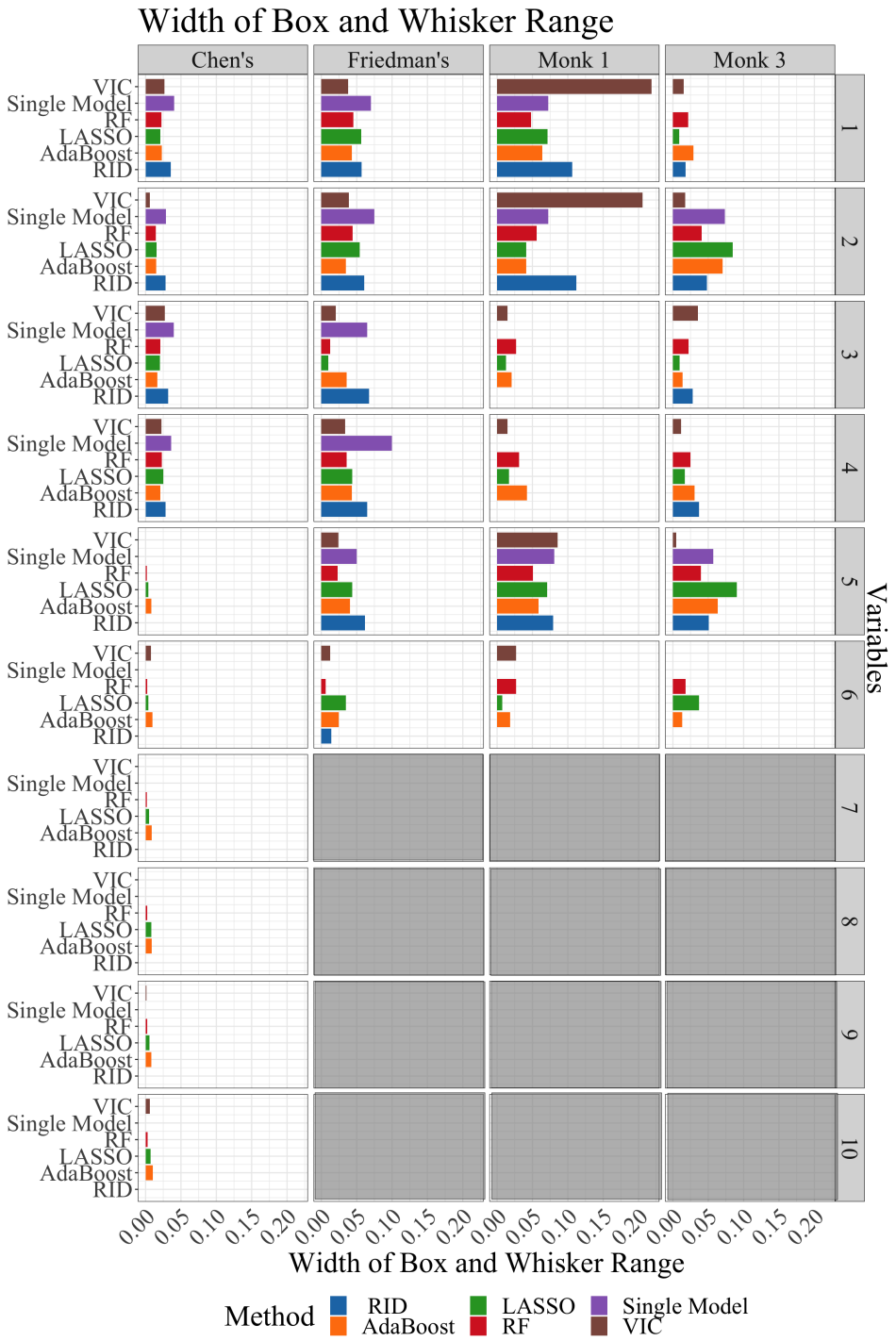


Figure 12: Width of the box and whisker range produced by each baseline method by dataset and variable. Gray subplots represent DGPs for which such a variable does not exist. Friedman's, Monk 1, and Monk 3 only have six variables.

Monk 3 DGP, and from 87.23% to 88.8% on Friedman's DGP. We find that VIC is somewhat more sensitive to choices of ε: the recovery proportion for VIC ranges from 83.44% to 89.62% on Chen's DGP, 100% to 100% on Monk 1, 75.30% to 79.17% on Monk 3 DGP, and from 60.53% to 75.57% on Friedman's DGP.

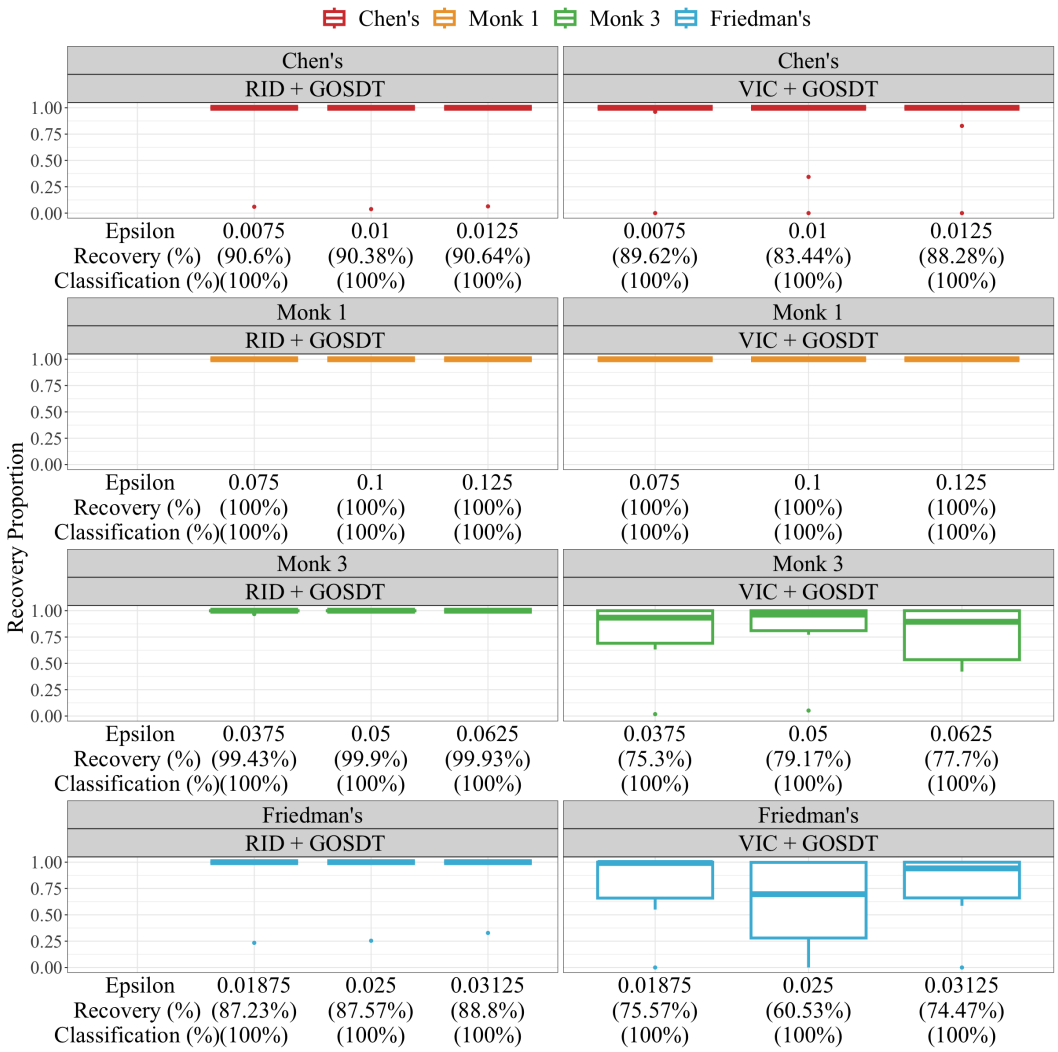


Figure 13: Box and whiskers plot over variables of the proportion test MR values for the DGP captured by the BWR range for *RID* and VIC at different loss thresholds ε. We find that the performance of *RID* is invariant to reasonable changes in ε.

E.4 Full Stability Results

In this section, we demonstrate each interval produced by MCR, the BWR of VIC, and the BWR of *RID* over 50 datasets generated from each DGP. We construct *RID* using 50 bootstraps from each of the 50 generated datasets.

Figures 14, 15, 16, and 17 illustrate the 50 resulting intervals produced by each method for each non-extraneous variable on each DGP. If a method produces generalizable results, we would expect it to produce overlapping intervals across datasets drawn from the same DGP. As shown in Figures 14, 16, and 17, both MCR and the BWR for VIC produced completely non-overlapping intervals between datasets for at least one variable on each of Chen's DGP, Monk 3, and Friedman's DGP, which means their results are not generalizable. In contrast, **the BWR range for *RID never* has zero overlap between the ranges produced for different datasets**. This highlights that *RID* is more likely to generalize than existing Rashomon-based methods.

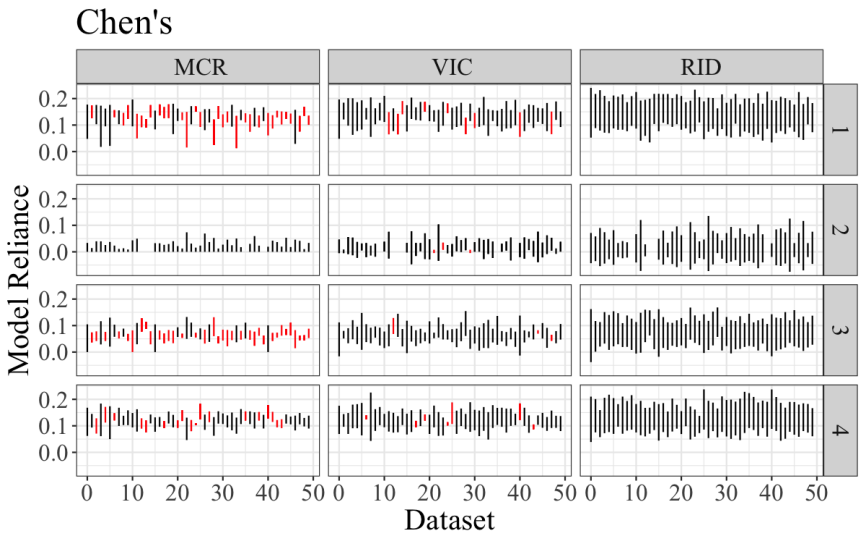


Figure 14: We generate 50 independent datasets from Chen's DGP and calculate MCR, BWRs for VIC, and BWRs for RID. The above plot shows the interval for each dataset for each non-null variable in Chen's DGP. All red-colored intervals do not overlap with at least one of the remaining 49 intervals.

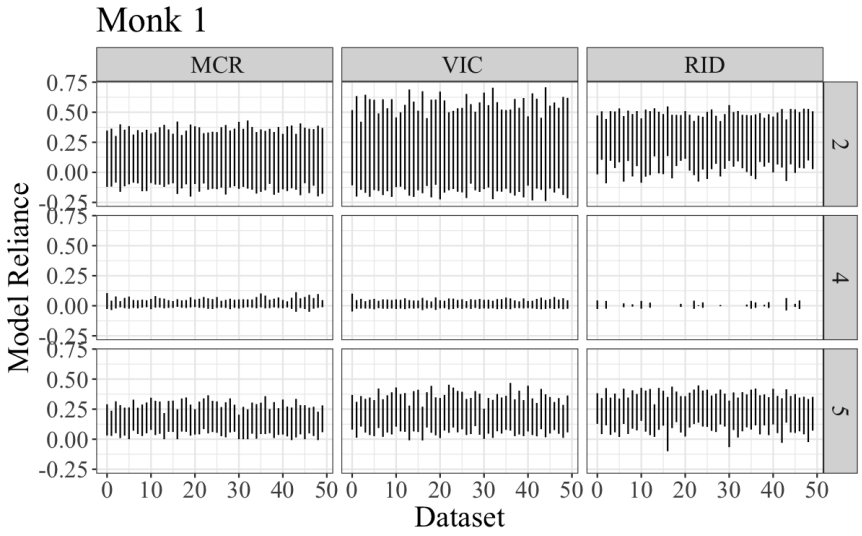


Figure 15: We generate 50 independent datasets from the Monk 1 DGP and calculate MCR, BWRs for VIC, and BWRs for RID. The above plot shows the interval for each dataset for each non-null variable in Monk 1 DGP. All red-colored intervals (there are none in this plot) do not overlap with at least one of the remaining 49 intervals.

E.5 Timing Experiments

Finally, we perform an experiment studying how well the runtime of *RID* scales with respect to the number of samples and the number of features in the input dataset using the HIV dataset [29]. The complete dataset used for the main paper consists of 14,742 samples measuring 100 features each. We compute *RID* using 30 bootstrap iterations for each combination of the following sample and feature subset sizes: 14,742 samples, 7,371 samples, and 3,686 samples; 100 features, 50 features, and 25 features.

Note that, in our implementation of *RID*, any number of bootstrap datasets may be handled in parallel; as such, we report the mean runtime per bootstrap iteration in Table 3, as this quantity is independent of how many machines are in use. As shown in Table 3, *RID* scales fairly well in the number of samples included, and

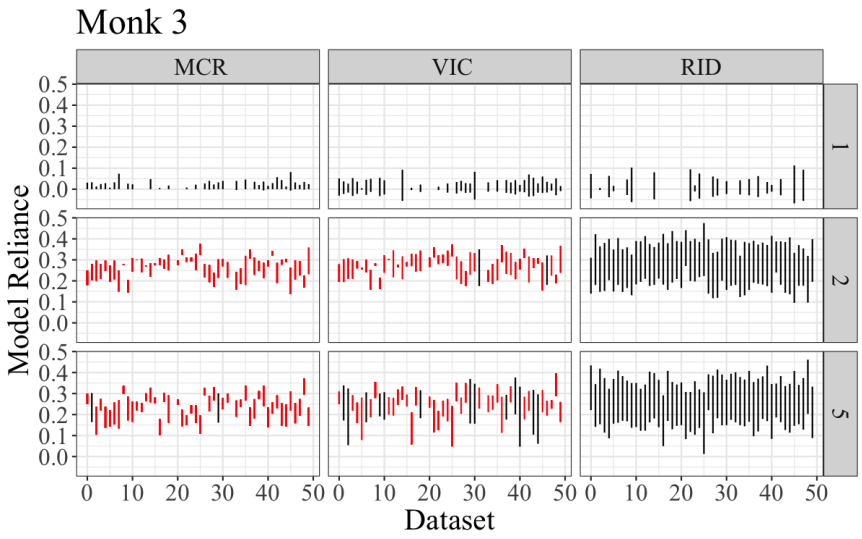


Figure 16: We generate 50 independent datasets from the Monk 3 DGP and calculate MCR, BWRs for VIC, and BWRs for RID. The above plot shows the interval for each dataset for each non-null variable in the Monk 3 DGP. All red-colored intervals do not overlap with at least one of the remaining 49 intervals.

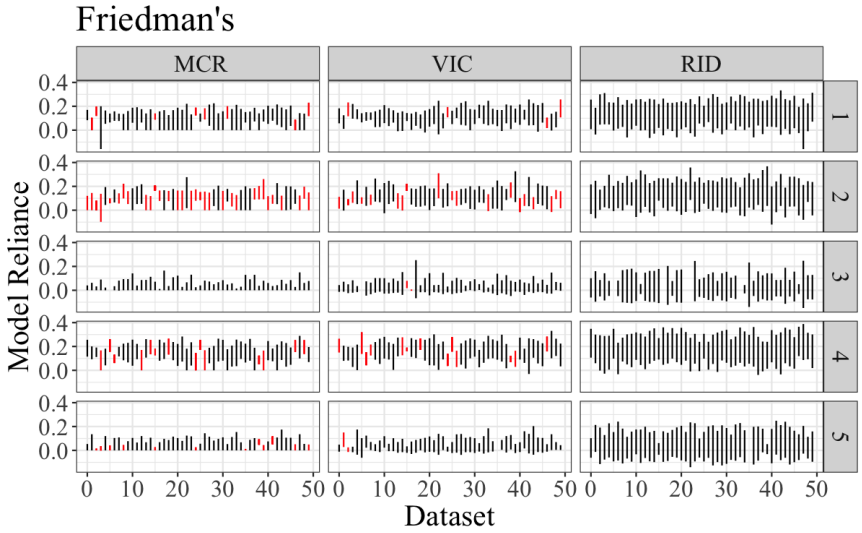


Figure 17: We generate 50 independent datasets from Friedmanś DGP and calculate MCR, BWRs for VIC, and BWRs for RID. The above plot shows the interval for each dataset for each non-null variable in Friedman's DGP. All red-colored intervals do not overlap with at least one of the remaining 49 intervals.

somewhat less well in the number of features. This is because the number of possible decision trees grows rapidly with the number of input features, making finding the Rashomon set a more difficult problem and leading to larger Rashomon sets. Nonetheless, even for a large number of samples and features, *RID* can be computed in a tractable amount of time: with 100 features and 14,742 samples, we found an average time per bootstrap of about 52 minutes.

References

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

Samples \ Variables	25	50	100
3,686	19.3 (0.9)	64.2 (6.2)	164.0 (14.6)
7,371	40.5 (2.5)	177.7 (18.8)	723.1 (106.4)
14,742	92.9 (6.8)	431.4 (39.9)	3128.7 (281.9)

Table 3: Average runtime in seconds per bootstrap for *RID* as a function of the number of variables and number of samples included from the HIV dataset. The standard error about each average is reported in parentheses.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[3] Massimo Aufiero and Lucas Janson. Surrogate-based global sensitivity analysis with statistical guarantees via floodgate. *arXiv preprint arXiv:2208.05885*, 2022.

[4] F Bachelerie, J Alcami, F Arenzana-Seisdedos, and JL Virelizier. HIV enhancer activity perpetuated by NF-κB induction on infection of monocytes. *Nature*, 350(6320):709–712, 1991.

[5] Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018.

[6] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[7] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[8] Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[9] Jianbo Chen, Mitchell Stern, Martin J Wainwright, and Michael I Jordan. Kernel feature selection via conditional covariance minimization. *Advances in Neural Information Processing Systems*, 30, 2017.

[10] Zhi Chen, Chudi Zhong, Margo Seltzer, and Cynthia Rudin. Understanding and exploring the whole set of good sparse generalized additive models. *Advances in Neural Information Processing Systems*, 36, 2023.

[11] Beau Coker, Cynthia Rudin, and Gary King. A theory of statistical inference for ensuring the robustness of scientific results. *Management Science*, 67(10):6174–6197, 2021.

[12] Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

[13] James Duncan, Rush Kapoor, Abhineet Agarwal, Chandan Singh, and Bin Yu. Veridicalflow: a python package for building trustworthy data science pipelines with PCS. *Journal of Open Source Software*, 7(69): 3895, 2022.

[14] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 2020.

[15] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[16] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[17] Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[18] Yves Grandvalet. Stability of bagged decision trees. In *Proceedings of the XLIII Scientific Meeting of the Italian Statistical Society*, pages 221–230. CLEUP, 2006.

[19] Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

[20] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[21] Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):1–16, 2021.

[22] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

[23] Marcus Kretzschmar, Michael Meisterernst, Claus Scheidereit, Gen Li, and Robert G Roeder. Transcriptional regulation of the HIV-1 promoter by NF-κ B in vitro. *Genes & Development*, 6(5):761–774, 1992.

[24] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, May 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[25] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[26] Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, pages 6150–6160. PMLR, 2020.

[27] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

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

[29] Ashokkumar Manickam, Jackson J Peterson, Yuriko Harigaya, David M Murdoch, David M Margolis, Alex Oesterling, Zhicheng Guo, Cynthia D Rudin, Yuchao Jiang, and Edward P Browne. Integrated single-cell multiomic analysis of HIV latency reversal reveals novel regulators of viral reactivation. *bioRxiv*, pages 2022–07, 2022.

[30] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

[31] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022.

[32] Christopher C Nixon, Maud Mavigner, Gavin C Sampey, Alyssa D Brooks, Rae Ann Spagnuolo, David M Irlbeck, Cameron Mattingly, Phong T Ho, Nils Schoof, Corinne G Cammon, et al. Systemic HIV and SIV latency reversal via non-canonical NF-κB signalling in vivo. *Nature*, 578(7793):160–165, 2020.

[33] Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *The Journal of Machine Learning Research*, 18(1):6345–6398, 2017.

[34] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, pages 1–13, 2022.

[35] Di Qu, Wei-Wei Sun, Li Li, Li Ma, Li Sun, Xia Jin, Taisheng Li, Wei Hou, and Jian-Hua Wang. Long noncoding RNA MALAT1 releases epigenetic silencing of HIV-1 replication by displacing the polycomb repressive complex 2 from binding to the LTR promoter. *Nucleic Acids Research*, 47(6):3013–3027, 2019.

[36] Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438 – 3469, 2019. doi: 10.1214/18-AOS1784.

[37] Cynthia Rudin and Joanna Radin. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2):10–1162, 2019.

[38] Matteo Scalabrin, Ilaria Frasson, Emanuela Ruggiero, Rosalba Perrone, Elena Tosoni, Sara Lago, Martina Tassinari, Giorgio Palù, and Sara N Richter. The cellular protein hnRNP A2/B1 enhances HIV-1 transcription by unfolding LTR promoter G-quadruplexes. *Scientific Reports*, 7(1):1–13, 2017.

[39] Paola Secchiero, Davide Zella, Sabrina Curreli, Prisco Mirandola, Silvano Capitani, Robert C Gallo, and Giorgio Zauli. Pivotal role of cyclic nucleoside phosphodiesterase 4 in tat-mediated CD4+ T cell hyperactivation and HIV type 1 replication. *Proceedings of the National Academy of Sciences*, 97(26):14620–14625, 2000.

[40] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

[41] Gavin Smith, Roberto Mansilla, and James Goulding. Model class reliance for random forests. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22305–22315. Curran Associates, Inc., 2020.

[42] S. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R.S. Michalski, T. Mitchell, P. Pachowicz, B. Roger, H. Vafaie, W. Van de Velde, W. Wenzel, J. Wnek, and J. Zhang. The MONK's problems: A performance comparison of different learning algorithms. Technical Report CMU-CS-91-197, Carnegie Mellon University, Computer Science Department, Pittsburgh, PA, 1991.

[43] Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *ECML/PKDD 2014*, page 121, 2014.

[44] Fang Wang, Shujia Huang, Rongsui Gao, Yuwen Zhou, Changxiang Lai, Zhichao Li, Wenjie Xian, Xiaobo Qian, Zhiyu Li, Yushan Huang, et al. Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility. *Cell Discovery*, 6(1):83, 2020.

[45] Wen-Fang Wang, Hui-Juan Zhong, Shu Cheng, Di Fu, Yan Zhao, Hua-Man Cai, Jie Xiong, and Wei-Li Zhao. A nuclear NKRF interacting long noncoding RNA controls EBV eradication and suppresses tumor progression in natural killer/T-cell lymphoma. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, page 166722, 2023.

[46] Wikipedia contributors. Box plot, 2023. URL `https://en.wikipedia.org/wiki/Box_plot`. [Online; accessed 14-May-2023].

[47] Brian Williamson and Jean Feng. Efficient nonparametric statistical inference on population feature importance using shapley values. In *International Conference on Machine Learning*, pages 10282–10291. PMLR, 2020.

[48] Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, pages 1–14, 2021.

[49] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. *Advances in Neural Information Processing Systems*, 35: 22305–22315, 2022.

[50] Bin Yu. Stability. *Bernoulli*, 19(4):1484 – 1500, 2013. doi: 10.3150/13-BEJSP14.

[51] Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117 (8):3920–3929, 2020. doi: 10.1073/pnas.1901326117.

[52] Lu Zhang and Lucas Janson. Floodgate: inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*, 2020.