Doctor Rashomon and the UNIVERSE of Madness: Variable Importance with Unobserved Confounding and the Rashomon Effect

Jon Donnelly* Duke University

Srikar Katta* Duke University

Emanuele Borgonovo Bocconi University

Cynthia Rudin Duke University

Abstract

Variable importance (VI) methods are often used for hypothesis generation, feature selection, and scientific validation. In the standard VI pipeline, an analyst estimates VI for a *single* predictive model with only the *observed* features. However, the importance of a feature depends heavily on which other variables are included in the model, and essential variables are often omitted from observational datasets. Moreover, the VI estimated for one model is often not the same as the VI estimated for another equally-good model – a phenomenon known as the *Rashomon Effect*. We address these gaps by introducing UNobservables and Inference for Variable importancE using Rashomon SEts (UNIVERSE). Our approach adapts Rashomon sets – the sets of near-optimal models in a dataset – to produce bounds on the true VI even with missing features. We theoretically guarantee the robustness of our approach, show strong performance on semi-synthetic simulations, and demonstrate its utility in a credit risk task.

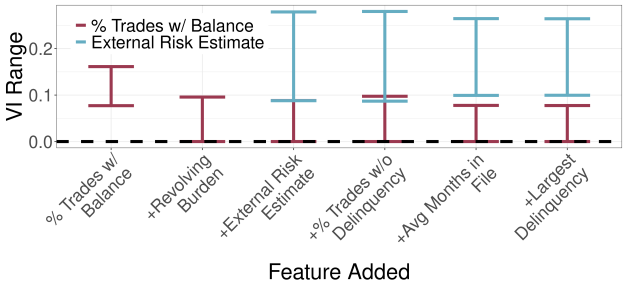

Figure 1: **Variable importance intervals for two variables from the FICO dataset, as more variables are observed.** When we estimate the range of variable importance (quantified by subtractive model reliance of Fisher et al., 2019) with only one observed feature – `% Trades w/ Balance` – this feature seems important, with a confidence interval that does not overlap with 0. However, as we introduce more variables into the analysis pipeline, this variable's confidence intervals overlap with 0, no longer yielding significant results. In contrast, `External Risk Estimate` – the third feature incorporated into the analysis – remains significant even after controlling for other observed variables. Thus, observing additional information changes the conclusion a practitioner would draw about `% Trades w/ Balance`.

1 INTRODUCTION

Variable importance (VI) methods are used throughout science to study the strength of different risk factors, generate new hypotheses, and understand which features are worth collecting for future predictions. In the standard variable importance pipeline, a scientist calculates the importance of a variable for a *single* predictive model using only the *observed* features. However, this approach may lead to misleading insights because multiple models may explain a dataset equally well – a phenomenon known as the *Rashomon effect* (Breiman, 2001b) – and a variable deemed important for one model may not be important for another. To overcome the Rashomon effect, recent research has developed algorithms to compute the importance of variables across the *Rashomon set*, the set of all near-optimal models for a given dataset (Fisher et al., 2019; Dong and Rudin, 2020; Donnelly et al., 2023; Xin et al., 2022; Chen et al., 2023; Babbar et al., 2025). These methods are able to discover the true importance of variables in experiments Donnelly et al. (2023) but are not valid when there may be important unobserved variables, which is often the case in practice.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

*Jon Donnelly and Srikar Katta contributed equally to this work

Figure 1 illustrates this problem: we present the range of variable importance for two variables across the Rashomon set as we incorporate more features into the model to predict loan default risk. When only a single feature is observed, the `%Trades w/ Balance` feature appears statistically significant for predicting loan default. But as we incorporate features that we know are important, this variable no longer seems important. This inconsistency shows how omitted variables can undermine decision-making.

In this work, we introduce UNobservables and Inference for Variable importancE using Rashomon SEts (UNIVERSE) to address both challenges. We take advantage of Rashomon sets to derive bounds on the true variable importance under the possibility of unobserved variables and adapt our Rashomon sets to account for finite sample uncertainty. We prove theoretically that our bounds contain the true variable importance for the true conditional mean function under unobserved confounding and have valid Type-1 error control. Via semi-synthetic experiments, we demonstrate that our bounds are tight enough to draw useful conclusions in practice, even in the presence of unmeasured variables and the Rashomon effect. The code for this work is available at https://github.com/jdonnelly36/UNIVERSE.

2 RELATED WORK

We review the literature on (i) variable importance measures, (ii) the Rashomon effect, and (iii) unobserved confounding. To our knowledge, our work is the first to consider all three components simultaneously.

Variable Importance The most classical examples of variable importance metrics include the parameters of a linear model and the permutation importance from a decision tree (Louppe et al., 2013; Kazemitabar et al., 2017). Because these metrics may be restrictive, researchers have also introduced perturbation-based methods that can be applied to any model, which quantify how much a specified model's performance changes as variables are perturbed. Such methods include Shapley additive explanations (SHAP) (Lundberg and Lee, 2017), local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), model reliance (Breiman, 2001a; Fisher et al., 2019), and conditional model reliance (Fisher et al., 2019). However, these approaches only estimate the importance of a feature for the *specified* model and do not consider the *Rashomon Effect*.

The Rashomon Effect The Rashomon Effect – coined by Breiman (2001b) – describes the phenomenon in which multiple, possibly distinct, models describe a given dataset equally well. The Rashomon effect is ubiquitous (Paes et al., 2023) and has been observed in high-stakes domains, making the development of tools that are robust to the Rashomon effect essential for trustworthy decision making (Rudin et al., 2024). The most common approach to ensure robustness to the Rashomon effect is to estimate a Rashomon *set* – the set of all near-optimal models in a model class. Recent advancements have led to the estimation/approximation of Rashomon sets for decision trees (Xin et al., 2022; Babbar et al., 2025), risk scores (Liu et al., 2022), generalized additive models (Chen et al., 2023), and prototypical part neural networks (Donnelly et al., 2025). Analysts can then find the range, point cloud, or distribution of variable importance across Rashomon sets (Fisher et al., 2019; Dong and Rudin, 2020; Donnelly et al., 2023), or identify variables are consistently more important than others across the Rashomon set Laberge et al. (2023).

Several approaches also offer a related type of robustness – to model misspecification stemming from finite sample model uncertainty. Some methods are built for specific variable importance metrics, such as Leave One Covariate Out (LOCO) or SHapley Additive Explanations (Williamson et al., 2021; Williamson and Feng, 2020; Zhang and Janson, 2020; Aufiero and Janson, 2025; Verdinelli and Wasserman, 2024). While other methods should guarantee robustness due to finite-sample model misspecification (Lei et al., 2018), these approaches are not robust to unobserved confounding.

Unobserved confounding in causal inference Unobserved features pose a pervasive problem in causal inference: when some important confounders are unobserved, treatment effects estimated on the observed data will be biased. To overcome this issue, analysts often design sensitivity analyses to identify a *set* of viable causal effects under a reasonable level of possible confounding (Rosenbaum, 1987, 2007; Manski, 2003). We take a similar approach: we identify upper and lower bounds on variable importance under a reasonable level of unobserved confounding. Unlike causal sensitivity analyses that require a new approach for each new causal estimand, our approach is generic and can handle *any* model-based variable importance measure. While some approaches have relatively mild assumptions for causal estimands, these are too restrictive for variable importance metrics. For example, the most flexible sensitivity analysis framework by Chernozhukov et al. (2022) still requires that the causal estimand is a linear functional of the data but common variable importance metrics like SHAP are *quadratic* functionals (Verdinelli and Wasserman, 2024). Our approach overcomes these restrictive assumptions and is useful for a wide class of VI metrics.

3 METHODS

Let $\mathcal{D}^{(n)} = \{(X_i, Y_i)\}_{i=1}^n$ represent a dataset consisting of n independent and identically distributed (i.i.d.) tuples drawn from some distribution $\mathcal{P}_{XY}$, where $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is the prediction target of interest and $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$ is a vector of p covariates.

Consider a function class $\mathcal{F}$ consisting of models that output a prediction given the observed variables X_i (e.g., the space of all possible sparse decision trees constructed using X_i). Let ϕ_j denote a function that quantifies the importance of variable j to some model $f \in \mathcal{F}$ for some observation (X_i, Y_i); this is a *local* variable importance quantity, such as local SHAP. We describe how some common variable importance metrics, including permutation importance and SHAP, can be expressed in these terms in Appendix D. In our experiments, we focus on subtractive model reliance (Fisher et al., 2019) because it is simple to interpret and easy to compute. Let $\Phi_j(f, \mathcal{P}_{XY}) := \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\phi_j(f, (X, Y))]$ represent the *average* importance of variable j over the population data distribution $\mathcal{P}_{XY}$; for example, when our target quantity is global SHAP for feature j and the model f, $\phi_j(f, (X_i, Y_i))$ measures the local SHAP for subject i, and $\Phi_j(f, \mathcal{P}_{XY})$ measures global SHAP as the average over subject-level SHAP. We refer to the average importance of variable j over the *empirical* data distribution as $\Phi_j(f, \mathcal{D}^{(n)}) := \frac{1}{n}\sum_{i=1}^n \phi_j(f, (x_i, y_i))$.

We consider the setting in which p_U key variables $U \in \mathcal{U} \subseteq \mathbb{R}^{p_U}$ are not observed. Our goal is to quantify variable importance for the true conditional mean function that observes both X and U: $g^*(x, u) := \mathbb{E}_{Y|X=x,U=u}[Y \mid x, u]$. We define our estimand as $\Phi_j(g^*, \mathcal{P}_U) := \mathbb{E}_{(X,U,Y)\sim\mathcal{P}_U}[\phi_j(g^*, (X, U, Y))]$ where $\mathcal{P}_U$ describes the true distribution from which (X, U, Y) tuples are sampled. Because $\mathcal{P}_U \neq \mathcal{P}_{XY}$ in many contexts, the VI of the population risk minimizer defined over the *observed* variables is not necessarily the same as the VI for the optimal model defined over both observed *and* unobserved variables.

If we fix the unobserved variables to a specific quantity, then the only input to g^* that varies is X. We define this *conditional sub-model*, conditioned on the unobserved variables being $U = u$, as $f_u(x) := \mathbb{E}_{Y|X=x,U=u}[Y \mid X = x, U = u]$. We will construct bounds on the true variable importance for the optimal model using the *set* of *all* conditional sub-models, defined as $S^* := \{f_u \mid u \in \mathcal{U}\}$. Throughout this work, we assume that $\mathcal{U}$, S^*, and the model class $\mathcal{F}$, are finite sets for simplicity of notation, although these results can easily be extended to infinite sets. Figure 2 outlines our framework.

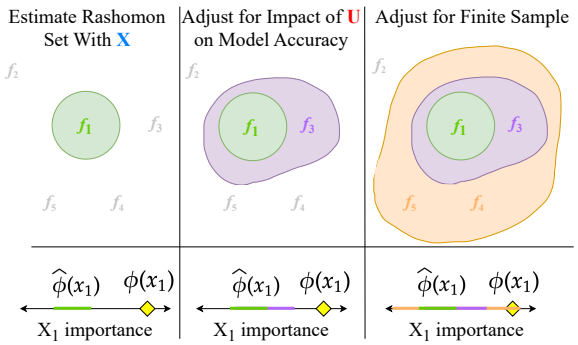

Figure 2: An overview of our framework showing the expansion of the Rashomon set as we account for more sources of error. The ♦ represents the true importance of variable X_1 having observed all necessary features. We consider three key factors: the Rashomon Effect, unobserved variables U, and finite sample errors. When adjusting for U, we expand which models are considered as part of the Rashomon set, potentially widening our variable importance interval. Adjusting for finite sample considerations both expands the Rashomon set and expands the range of variable importance values for each model, further widening our intervals.

3.1 Estimating Distribution-Invariant Quantities Over S^*

Some target functions Φ, such as the coefficient vector of linear models, are **distribution-invariant**, which means $\Phi(f, \mathcal{P}) = \Phi(f, \mathcal{P}')$ for all $f \in \mathcal{F}$, even if data distributions differ. We first discuss how we can construct bounds for distribution-invariant target functions when we have omitted variables and then generalize our analysis to consider **distribution-dependent** functions (e.g., permutation importance) in Section 3.2.

We capture S^* by leveraging the *Rashomon set*: the set of *all good models* for some objective. We define the *population, ϵ-threshold Rashomon set* as

$$\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda, \mathcal{P}_{XY}) := \{f \in \mathcal{F} : \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)] + \lambda(f) \leq \epsilon\}. \quad (1)$$

The population Rashomon set is the set of all models f from a model class $\mathcal{F}$ with expected loss $\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with respect to the distribution $\mathcal{P}_{XY}$ plus model-level regularization penalty $\lambda : \mathcal{F} \to \mathbb{R}$ below some specified threshold ϵ. When regularization is not applied, we use the "null regularization," which we define as $\lambda_0(f) := 0$ $\forall f \in \mathcal{F}$. For notational convenience, we drop notation for conditioning on $\mathcal{F}, \ell, \mathcal{P}_{XY}$ unless necessary. We apply the following assumption to connect S^* to $\mathcal{R}$:

Definition 1. *Define $\epsilon_{unobs} \geq 0$ such that,*

$$\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f_u, X, Y)] \leq \epsilon_{unobs} \quad \forall f_u \in S^*. \quad (2)$$

The value of ϵ_{unobs} bounds the loss of each conditional sub-model (f_u) over observed features $(X, Y) \sim \mathcal{P}_{XY}$. In practice, we leave it to practitioners to provide a reasonable upper bound on ϵ_{unobs}. This quantity measures the *heterogeneity* of the true conditional mean function across subgroups defined by the unobserved features. If there was no heterogeneity between unobserved groups, then f_u would predict Y perfectly given X and $\epsilon_{unobs} = 0$, so a small value for ϵ_{unobs} reflects the belief that most of the important information for this predictive task has been measured. Given Definition 1, we can use ϵ_{unobs} to connect S^* to a quantity dependent only on observed features:

Proposition 1. *Given Definition 1 and $S^* \subseteq \mathcal{F}$, we know that $S^* \subseteq \mathcal{R}(\epsilon_{unobs}; \lambda_0)$.*

This proposition simplifies our goal: rather than finding S^* – which depends on unobservables – we can instead estimate $\mathcal{R}(\epsilon_{unobs})$, a superset of S^* that depends only on the distribution of observed quantities, given knowledge of an upper bound on ϵ_{unobs}. However, $\mathcal{R}(\epsilon_{unobs})$ is a population-level quantity, and we only observe a finite sample of data. We define the *empirical, ϵ-threshold Rashomon set* over a set of n samples as:

$$\hat{\mathcal{R}}^{(n)}(\epsilon; \mathcal{F}, \ell, \lambda) := \left\{ f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} \ell(f, x_i, y_i) + \lambda(f) \leq \epsilon \right\}. \quad (3)$$

The empirical Rashomon set contains models whose *empirical loss* is less than some specified threshold ϵ. We can directly compute the empirical Rashomon set for several model classes including decision trees (Xin et al., 2022) and kernel ridge regression (Fisher et al., 2019). In general, we cannot guarantee that $\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda_0) = \hat{\mathcal{R}}^{(n)}(\epsilon; \mathcal{F}, \ell, \lambda)$ because of sampling uncertainty and regularization bias. The following theorem provides probabilistic, finite sample bounds connecting $\hat{\mathcal{R}}^{(n)}$ to $\mathcal{R}$ by correcting for these issues:

Theorem 1. *For any loss function ℓ bounded between $\ell_{\min}$ and $\ell_{\max}$ and $\epsilon \in [\ell_{\min}, \ell_{\max}]$, it holds that*

$$\mathbb{P}\left(\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \subseteq \hat{\mathcal{R}}^{(n)}\left(\varepsilon_n + \epsilon + \lambda_{\sup}; \mathcal{F}, \ell, \lambda\right)\right) \geq 1 - \delta,$$

$$\text{where } \varepsilon_n = \sqrt{\frac{(\ell_{\max} - \ell_{\min})^2 \ln\left(\frac{C}{\delta}\right)}{2n}}$$

for sample size n, for any value $C \geq |\mathcal{R}(\epsilon_{unobs}; \mathcal{F}, \ell, \lambda)|$ (e.g., $C := |\mathcal{F}|$), regularization penalty λ, and regularization upper bound $\lambda_{\sup} = \sup_{f \in \mathcal{F}} \lambda(f)$. Moreover,

$$\mathbb{P}\left(\hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda) \subseteq \mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda)\right) \geq 1 - |\mathcal{F}| \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{max} - \ell_{min})^2}\right\}.$$

Theorem 1 guarantees, with high-probability, that the estimated Rashomon set with Rashomon threshold $\varepsilon_n + \epsilon + \lambda_{\sup}$ is a *finite sample* superset of the population ϵ-threshold Rashomon set. Although our population Rashomon set in Proposition 1 considers an unregularized objective (i.e., $\lambda = \lambda_0$), existing algorithms for estimating Rashomon sets *require* non-zero regularization penalties (Xin et al., 2022; Chen et al., 2023); as such, we also include the worst-case regularization $\lambda_{\sup}$ in our finite-sample correction. This ensures that a model that would be in the unregularized empirical Rashomon set is not excluded due to regularization bias. For example, if the regularization function is 0.01 times the number of leaves in a tree and f_u is a tree with two leaves, we would have $\ell(f_u, X, Y; \lambda) = 0 + \lambda(f_u) = 0.02$ for loss over $\mathcal{P}_{XY|U=u}$. Even though f_u predicts perfectly, it has loss 0.02, and could be excluded from Rashomon sets if we do not adjust for regularization bias.

The second part of Theorem 1 guarantees that our empirical Rashomon set is not a vacuous superset. Specifically, all models that are not in the population Rashomon set are also excluded from our empirical Rashomon set with high probability asymptotically. As our sample size grows, our empirical Rashomon sets converges to the true Rashomon set with high probability, offering a guarantee akin to asymptotic type-II error control but in the context of our Rashomon sets.

Theorem 1 is of independent interest to researchers using Rashomon sets in other contexts, like fairness (Marx et al., 2020). **Theorem 1 is the first bound connecting empirical and population Rashomon sets that controls type-1 *and* type-2 error by accounting for *both* finite sample *and* regularization biases.**

Because we can construct high-probability supersets for the population Rashomon set, which is a superset of S^*, we know that the empirical Rashomon set with threshold specified in Theorem 1 contains all models in S^* with high probability:

Corollary 1. *Let ϵ_{unobs} be defined as in Definition 1, and assume that $S^* \subseteq \mathcal{F}$. For any loss function ℓ bounded between $\ell_{\min}$ and $\ell_{\max}$, it holds that*

$$\mathbb{P}\left(S^* \subseteq \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup}; \mathcal{F}, \ell, \lambda)\right) \geq 1 - \delta,$$

for a sample of size n and regularization penalty λ, with ε_n defined as in Theorem 1.

Note that both Theorem 1 and Corollary 1 define ε_n using an upper bound on the size of the population Rashomon set, i.e. $C \geq |\mathcal{R}(\epsilon_{unobs}; \mathcal{F}, \ell, \lambda)|$. It can be difficult to calculate a tight C, since in practice we do not know $|\mathcal{R}(\epsilon_{unobs}; \mathcal{F}, \ell, \lambda)|$. We can guarantee these

bounds hold by setting $C := |\mathcal{F}|$, but this can yield large empirical Rashomon sets for large model classes.

In practice, we can often provide tighter bounds by first estimating the size of the population Rashomon set using a separate data split. The following proposition shows that the size of the empirical Rashomon set quickly converges to the size of the population Rashomon set with the same parameters, meaning that we can guarantee *asymptotic*, type-1 error control:

Proposition 2. *The size of the estimated Rashomon set with threshold $\epsilon' > \epsilon$ is an upper bound on the size of the population Rashomon set with threshold ϵ with high probability:*

$$\mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(|\hat{\mathcal{R}}^{(n)}(\epsilon')| > |\mathcal{R}(\epsilon)|\right) = 1 - O(n^{-1}).$$

In the appendix, Corollary 2 uses Proposition 2 to show how using an estimated Rashomon set-size will yield a superset of S^* with high probability asymptotically. Based on this theory, in practice, we use an estimate of the size of the Rashomon set for C.

We now use the tools necessary to recover the models from S^* to guarantee coverage of distribution-invariant variable importance functions across S^*. We present a general result that provides probabilistic coverage over the variable importance with respect to the entire *observed* data distribution for each model in S^*:

Theorem 2. *Let α be a value such that, for all $f \in S^*$,*

$$\mathbb{P}_{\mathcal{D}^{(n)}}\left(\Phi_j(f,\mathcal{P}_{XY}) \in \left[\Phi_j(f,\mathcal{D}^{(n)}) \pm \alpha\right]\right) \geq 1-\gamma,$$

where $0 \leq \gamma \leq 1$ and $\mathbb{P}_{\mathcal{D}^{(n)}}$ denotes the probability of drawing the observed n samples from $\mathcal{P}_{XY}$. It follows that, for all $f \in S^*$, with probability at least $1-(\delta+\gamma)$,

$$\{\Phi_j(f,\mathcal{P}_{XY}) \mid f \in S^*\} \subseteq \left[\begin{array}{l}\inf_{f'\in\hat{\mathcal{R}}^{(n)}(\varepsilon_n+\epsilon_{unobs}+\lambda_{\sup})}\Phi_j(f',\mathcal{D}^{(n)}) - \alpha, \\ \sup_{f'\in\hat{\mathcal{R}}^{(n)}(\varepsilon_n+\epsilon_{unobs}+\lambda_{\sup})}\Phi_j(f',\mathcal{D}^{(n)}) + \alpha\end{array}\right],$$

where ε_n and δ are defined as in Corollary 1.

Theorem 2 uses Corollary 1 to show that, given a finite sample bound for the estimation of the variable importance Φ_j, we can compute an interval that contains the importance of a variable simultaneously for every model in S^* with high probability. In our experiments, we apply established finite sample bounds for subtractive model reliance (Fisher et al., 2019).

3.2 Connecting to $\Phi(g^*, \mathcal{P}_U)$

Because many variable importance functions depend heavily on the input distribution, it is not necessarily true that the variable importance for the optimal model will be contained within the bounds on variable importance for conditional submodels evaluated on the *observed* data. This inequality can be seen through a simple application of the law of iterated expectation:

$$\begin{aligned}
&\Phi_j(g^*,\mathcal{P}_U) \\
&= \mathbb{E}_{(X,U,Y)\sim\mathcal{P}_U}[\phi_j(g^*,(X,Y))] \text{ (by definition)} \\
&= \sum_{u\in\mathcal{U}} \mathbb{P}(U=u)\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY|U=u}}[\phi_j(g^*,(X,Y))] \\
&\quad \text{(by law of iterated expectation)} \\
&= \sum_{u\in\mathcal{U}} \mathbb{P}(U=u)\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY|U=u}}[\phi_j(f_u,(X,Y))] \\
&\quad \text{(by definition, } g^* = f_u \text{ when } U=u) \\
&\neq \sum_{u\in\mathcal{U}} \mathbb{P}(U=u)\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\phi_j(f_u,(X,Y))] \\
&\quad \text{(because } \mathcal{P}_{XY|U=u} \neq \mathcal{P}_{XY}).
\end{aligned}$$

That is, the importance of variable j to g^* can be expressed in terms of the variable importance across every f_u *with respect to* $\mathcal{P}_{XY|U=u}$, the distribution of observed data *conditional on the unobserved features.* This presents a problem, because we do not know which conditional distribution $\mathcal{P}_{XY|U=u}$ each sample in our observed dataset is drawn from because we do not know u. We overcome this challenge by bounding how sensitive our variable importance metric is to this kind of distribution shift:

Assumption 1. *Assume that there exists a known τ_j such that, for all $u \in \mathcal{U}$, $\left|\Phi_j(f_u,\mathcal{P}_{XY|U=u}) - \Phi_j(f_u,\mathcal{P}_{XY|U\neq u})\right| \leq \tau_j$.*

We refer to this quantity τ_j as VI-drift. VI-drift is large if the data distribution conditioned on $U=u$ is very different from that conditioned on $U \neq u$ for some u, and if the variable importance metric Φ_j is sensitive to this difference. We elaborate on this assumption in Appendix C. Under Assumption 1, we can now build on Theorem 2 to provide finite sample bounds for $\Phi_j(g^*,\mathcal{P}_U)$:

Theorem 3. *Let α and γ be defined as in Theorem 2, ε_n and δ be defined as in Theorem 1, and τ_j be defined as in Assumption 1. With probability at least $1-(\delta+\gamma)$, $\Phi_j(g^*,\mathcal{P}_U) \in$*

$$\left[\begin{array}{l}\inf_{f\in\hat{\mathcal{R}}^{(n)}(\varepsilon_n+\epsilon_{unobs}+\lambda_{\sup})}\Phi_j(f,\mathcal{D}^{(n)}) - \tau_j - \alpha, \\ \sup_{f\in\hat{\mathcal{R}}^{(n)}(\varepsilon_n+\epsilon_{unobs}+\lambda_{\sup})}\Phi_j(f,\mathcal{D}^{(n)}) + \tau_j + \alpha\end{array}\right].$$

In Theorem 3, the scalar τ_j considers corrections to the model-level variable importance estimate due to differences in distribution. If there exists no difference between the distributions (implying that $\tau_j = 0$), then the variable importance computed for f_u on the observed data will be a consistent estimate for $\Phi(f_u,\mathcal{P}_{XY}) =$

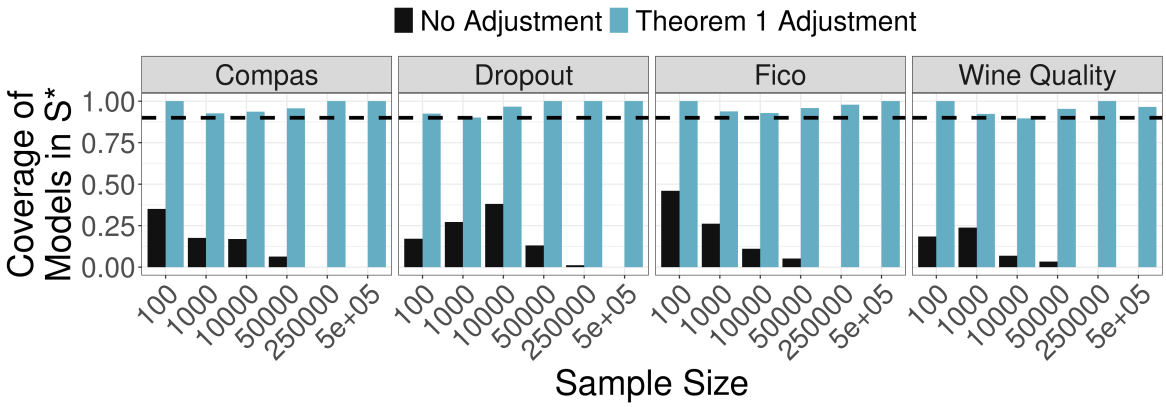

Figure 3: Verifying Theorem 1 in finite sample datasets. We compute the proportion of 100 random draws of the each dataset in which Rashomon sets estimated with the Rashomon threshold adjusting for finite sample biases as in Theorem 1 (in blue) and without any adjustment (in black) captures each f_u for each setting. The target coverage rate is ≥ 0.9, with $\delta = 0.1$. Across all sample sizes and datasets, omitting finite sample adjustments yields Rashomon sets that leave out necessary models. In contrast, our adjustment yields the target coverage rate, verifying the theorem holds. We use the estimated Rashomon set size as our upper bound on the size of S^*.

$\Phi(f_u, \mathcal{P}_{XY|U=u})$ because $\mathcal{P}_{XY} = \mathcal{P}_{XY|U=u}$, and we do not need to perform any correction; if the distributions are very different, $\Phi(f_u, \mathcal{P}_{XY})$ could be very different. Theorem 3 combines the finite-sample Rashomon set estimation correction from Theorem 1, the finite-sample variable importance estimation correction from Theorem 2, and finally introduces a distribution shift correction. In doing so, **we guarantee that our bounds contain the true variable importance even with omitted variables with high probability.** Figure 6 (discussed later) contains an example of how this analysis may work in practice.

4 EXPERIMENTS

4.1 Semi-synthetic Experiments

We now turn to evaluate each of our primary theoretical claims empirically. Because we cannot know the true variable importance in real observational data, we use "semi-synthetic" datasets generated as follows. Given a tabular dataset $\mathcal{D}^{(n)}$, we first split the dataset into K equally sized partitions $\mathcal{D}_1^{(n/K)}, \mathcal{D}_2^{(n/K)}, \ldots \mathcal{D}_K^{(n/K)}$. We then fit a decision tree classifier on each partition sequentially, yielding K distinct models $f_1, f_2, \ldots f_K$. In order to ensure that these models are distinct, when fitting the k-th tree, we set all entries of each feature used by the previously fitted models $f_1, f_2, \ldots f_{k-1}$ equal to 0. For each partition, we create a semi-synthetic dataset where each label is replaced by the prediction of the corresponding model, i.e., $\bar{\mathcal{D}}_k^{(n/K)} := \{(X_i, f_k(X_i))\}_{i=(k-1)n/K}^{kn/K}$. Finally, we combine the semi-synthetic datasets back into one. In doing so, we treat the partition to which each sample was assigned as an important unobserved feature $U \in \mathcal{U} := \{1, 2, \ldots, K\}$; if the i-th sample was assigned to partition k, we say $u_i = k$. This yields a known true conditional mean function $g^*(x_i, u_i) := f_{u_i}(x_i), u_i \in \{1, 2, \ldots, K\}$. This setup allows us to know each ground truth quantity of interest (g^* and S^* with their corresponding variable importance values, ϵ_{unobs}) while working with a fairly realistic distribution for X.

In the following experiments, we apply this procedure to four disparate real-world datasets with $K = 2$: (1) Compas (Larson et al., 2016), which concerns criminal recidivism prediction for 6,907 individuals; (2) Dropout (Martins et al., 2021), which concerns predicting whether students will drop out of college for 4,424 individuals; (3) FICO (FICO et al., 2018), which concerns predicting whether an individual will repay line of credit within 2 years for 10,459 individuals; (4) Wine Quality (Cortez et al., 2009), which concerns predicting wine quality ratings over 6,497 wines.

For each dataset, we evaluate our framework using random draws with replacement of sample sizes 100; 1,000; 10,000; 50,000; 250,000; and 500,000 from $\bar{\mathcal{D}}^{(n)}$, each over 100 iterations. We consider only the first 8 features (except Compas, which has 7 features) from each dataset to enable fast computation of the Rashomon set. In each setting, we consider the model class $\mathcal{F}$ of depth 3 or less decision trees found using TreeFarms Xin et al. (2022), and similarly restrict each f_k to have a depth of at most 3. We repeat each of these experiments using an approximation of the Rashomon set of random forests (Laberge et al., 2023) in Appendix G. To faithfully evaluate the theoretical claims above, we

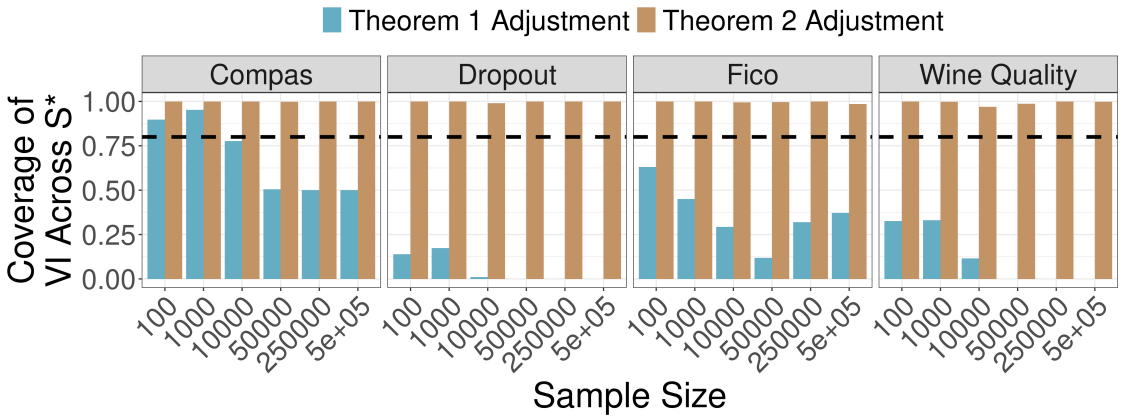

Figure 4: Verifying Theorem 2. We achieve the specified coverage rate of ≥ 0.8 only when adjusting for (i) model uncertainty via Theorem 1 and (ii) variable importance estimation uncertainty (in gold). Adjusting for model uncertainty alone (in blue) is not sufficient. For each setting, we compute the proportion of 100 experiments where our variable importance bounds capture the true variable importance for all submodels $f_u \in S^*$, averaged over variables. We use an estimate of the true Rashomon set size as our upper bound on the size of S^*.

use the true value for ϵ_{unobs} and τ_j in each experiment.

Finding each model in S^* Corollary 1 demonstrates how we may select a value ε_n such that, with high probability, the empirical Rashomon set contains S^*. We first evaluate the validity of this Corollary with a target confidence of $1 - \delta = 0.9$.

Figure 3 presents the results of this evaluation. We find that, across all six sample sizes and all four reference datasets, omitting finite sample and regularization adjustments yields Rashomon sets that leave out necessary models. In contrast, **our empirical Rashomon set based on Corollary 1 contains every model in S^* at above the specified rate**. These results demonstrate that adjusting for finite sample and regularization biases is necessary for constructing high-probability, finite-sample covers of the population Rashomon sets. Even at large sample sizes, not adjusting for these factors yields *under* coverage because of the regularization that existing algorithms necessitate to compute Rashomon sets. Specifically, ϵ_{unobs} is defined as the maximum expected *unregularized* loss for any model in S^*. As such, the expected *regularized* loss for the highest-loss model in S^* will exceed ϵ_{unobs} and the highest-loss model is therefore consistently omitted from the empirical Rashomon set.

Recovering variable importance for each model in S^* We now turn to evaluate Theorem 2: given a finite sample bound for the estimation of our variable importance metric for a specified model, we can cover the importance of each variable to each model in S^* with at least a specified probability. In this experiment, we set our target probability to $1 - (\delta + \gamma) = 0.8$, and apply the finite sample bound for subtractive model reliance (MR) from Equation B.26 of Fisher et al. (2019). Figure 4 demonstrates that adjusting only for model uncertainty as in Corollary 1 yields invalid bounds on model-level variable importance. In fact, for sample sizes larger than 10,000 observations, omitting variable importance uncertainty quantification yields intervals that *never* contain the true variable importance for all conditional submodels. In contrast, **across all four datasets and at all sample sizes, our approach achieves nominal coverage.**

Recovering variable importance for the g^* with unobserved features Finally, we evaluate Theorem 3, which provides high probability bounds on variable importance to the true model g^*, even with unobserved variables. We again set our target coverage rate to $1 - (\delta + \gamma) = 0.8$ and apply the finite sample bound for subtractive MR Fisher et al. (2019). This reflects applying the entire UNIVERSE framework.

Figure 5 demonstrates that **we consistently cover the true importance to g^* in more than the specified 80% of cases across all four datasets**, even though g^* depends on unobserved variables. Moreover, Figure 5 demonstrates that each component of Theorem 3 is necessary to achieve this coverage rate. Without these adjustments, our bounds would substantially undercover the true variable importance, as highlighted on the Compas dataset.

4.2 Case Study

We use our framework to study credit risk assessment using a dataset developed by the Fair Isaac Corpora-

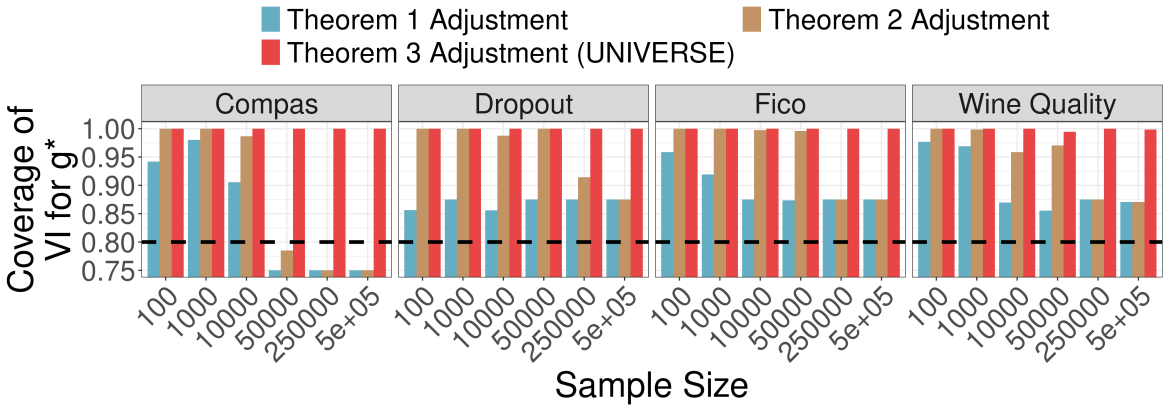

Figure 5: Verifying Theorem 3. We consistently achieve the specified coverage rate of ≥ 0.8 only when we account for (i) model uncertainty, (ii) variable importance uncertainty, and (iii) VI drift. Each bar measures the proportion of 100 experiments in which our bounds capture the true variable importance for the true model g^*. Plots are colored such that blue only accounts for finite sample model uncertainty as in Theorem 1, gold also adjusts for uncertainty in estimating subtractive model reliance (MR) at the model-level as in Theorem 2, and red adjusts for the previous two *and* distribution shifts induced by omitted variables 3. All three adjustments are necessary to achieve the target coverage rate of ≥ 0.8, with $\delta = \gamma = 0.1$.

tion (FICO), focusing on the most powerful predictive feature: `External Risk Estimate` (ERE). ERE is a risk score developed by external agencies like FICO for which banks often pay when evaluating credit applications. After controlling for other observed features, ERE remains important to *every* nearly-optimal model, as shown in Figure 1. However, the FICO dataset does not share information about features that are likely predictive of loan default, like income. If these unmeasured features could easily explain the information in ERE, banks could collect alternative features that would better serve customers. UNIVERSE allows analysts to answer the following: how much signal would another feature need to capture to replace ERE?

We use the Rashomon set of sparse decision trees with 0-1 loss and the same hyperparameters as in Section 4 to compute UNIVERSE. This is an appropriate model class because sparse decision trees have achieved similar performance to more complex model classes like boosted trees on this dataset (Rudin et al., 2024).

Figure 6 displays the results of our analyses. The x-axis displays the Rashomon threshold we used for estimating Rashomon sets (adjusting for finite sample and regularization biases as in Theorem 1), with larger x-values reflecting a large belief about the value of ϵ_{unobs}. The y-axis displays the lower bound on subtractive model reliance accounting for finite sample uncertainty and distribution shifts. Each color displays results under different amounts of assumed VI drift, τ_j.

Setting $\tau_j = 0$ means that conditioning on an unobserved variable does not change the distributions of

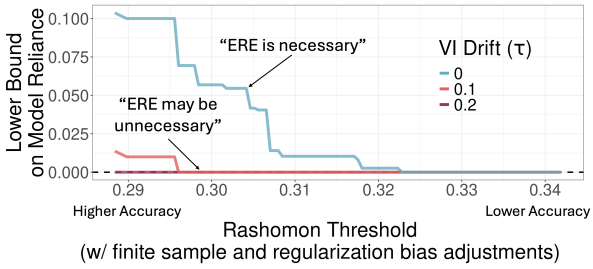

Figure 6: An example of how our tool may be used in practice. The x-axis displays the Rashomon threshold used for estimating Rashomon sets (adjusting for finite sample and regularization biases as in Theorem 1). The y-axis displays the range of subtractive model reliance (MR) we compute adjusting for model uncertainty, variable importance estimation uncertainty, and distribution shifts as in Theorem 3 as we vary the Rashomon threshold. Each color shows the range as the assumed amount of variable importance drift from distribution shift τ_j increases. The x-axis begins at the lowest achievable loss in the dataset.

ERE or outcome in a way that affects variable importance. Even under this assumption, we find that we would only need to identify a feature whose conditional sub-models achieve a 0-1 loss of $\epsilon_{unobs} = 0.32$ (which corresponds to an accuracy of 68%) to find a model that does not depend on ERE; our current best model achieves a loss of 0.29 (accuracy of 71%). This accuracy requirement becomes even smaller as we allow τ_j to increase, suggesting that ERE's importance is not robust to a moderate degree of unobserved confounding. If

we depended only on the observed features, an analyst may conclude that ERE is really important and must always be considered to predict credit risk. However, our analysis suggests that other, more interpretable features could easily replace ERE, enabling the bank to give more meaningful feedback to loan applicants.

5 CONCLUSION

In this work, we introduced UNIVERSE, the first variable importance method to account for finite sample concerns, the Rashomon effect, and unobserved variables. We proved both theoretically and empirically that UNIVERSE can recover the true variable importance to the true underlying conditional mean function, even in the presence of these sources of error.

The UNIVERSE framework is general and applicable to *any* model class. However, finding the complete Rashomon set is often computationally intensive. Future work should extend our framework to settings in which we *approximate* the complete Rashomon set of more complex model classes like neural networks (Donnelly et al., 2025) or deep decision trees (Babbar et al., 2025). Additionally, we develop UNIVERSE for classification problems, but many problems require modeling continuous or survival outcomes. Applied practitioners may benefit from generalizations of UNIVERSE for these outcomes. Nonetheless, UNIVERSE represents a substantial step in using VI in challenging, realistic settings with unobserved confounders.

Acknowledgments

Support from Bocconi Senior Researchers' Grant, 2024, 603020 is gratefully acknowledged by Emanuele Borgonovo. Srikar Katta is supported by the Apple Scholar in AI/ML Fellowship. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant number DGE 2139754, as well as National Institutes of Health/NIDA grant number R01DA054994. We also gratefully acknowledge support from the National Science Foundation FAI under grant number 2147061 and an Amazon FAI Gift. Thank you to the anonymous reviewers whose feedback greatly improved this manuscript.

References

Aufiero, M. and Janson, L. (2025). Surrogate-Based Global Sensitivity Analysis With Statistical Guarantees via Floodgate. *SIAM/ASA Journal on Uncertainty Quantification*, 13(2):563–590.

Babbar, V., McTavish, H., Rudin, C., and Seltzer, M. (2025). Near-Optimal Decision Trees in a SPLIT Second. In *Forty-second International Conference on Machine Learning*.

Breiman, L. (2001a). Random Forests. *Machine Learning*, 45:5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (With Comments and a Rejoinder by the Author). *Statistical Science*, 16(3):199–231.

Chen, Z., Zhong, C., Seltzer, M., and Rudin, C. (2023). Understanding and Exploring the Whole Set of Good Sparse Generalized Additive Models. *Advances in Neural Information Processing Systems*, 36.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. (2022). Long Story Short: Omitted Variable Bias in Causal Machine Learning. Technical report, National Bureau of Economic Research.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling Wine Preferences by Data Mining From Physicochemical Properties. *Decision Support Systems*, 47(4):547–553.

Dong, J. and Rudin, C. (2020). Exploring the Cloud of Variable Importance for the Set of All Good Models. *Nature Machine Intelligence*, 2(12):810–824.

Donnelly, J., Guo, Z., Barnett, A. J., McTavish, H., Chen, C., and Rudin, C. (2025). Rashomon Sets for Prototypical-Part Networks: Editing Interpretable Models in Real-Time. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4528–4538.

Donnelly, J., Katta, S., Rudin, C., and Browne, E. (2023). The Rashomon Importance Distribution: Getting Rid of Unstable, Single Model-Based Variable Importance. *Advances in Neural Information Processing Systems*, 36:6267–6279.

FICO, Google, Imperial College London, MIT, University of Oxford, UC Irvine, and UC Berkeley (2018). Explainable Machine Learning Challenge. `https://community.fico.com/s/explainable-machine-learning-challenge`.

Fisher, A., Rudin, C., and Dominici, F. (2019). 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.

Kazemitabar, J., Amini, A., Bloniarz, A., and Talwalkar, A. S. (2017). Variable Importance Using Decision Trees. In *Advances in Neural Information Processing Aystems*, volume 30.

Laberge, G., Pequignot, Y., Mathieu, A., Khomh, F., and Marchand, M. (2023). Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set. *Journal of Machine Learning Research*, 24(364):1–50.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analyzed the Compas Recidivism Algorithm. *ProPublica*.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Liu, J., Zhong, C., Li, B., Seltzer, M., and Rudin, C. (2022). FasterRisk: Fast and Accurate Interpretable Risk Scores. *Advances in Neural Information Processing Systems*, 35:17760–17773.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding Variable Importances in Forests of Randomized Trees. In *Advances in Neural Information Processing Systems*, volume 26.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer.

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., and Realinho, V. (2021). Early Prediction of Student's Performance in Higher Education: A Case Study. In *World Conference on Information Systems and Technologies*, pages 166–175. Springer.

Marx, C., Calmon, F., and Ustun, B. (2020). Predictive Multiplicity in Classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR.

Paes, L. M., Cruz, R., Calmon, F. P., and Diaz, M. (2023). On the inevitability of the rashomon effect. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 549–554. IEEE.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Rosenbaum, P. R. (1987). Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2007). Sensitivity Analysis for M-Estimates, Tests, and Confidence Intervals in Matched Observational Studies. *Biometrics*, 63(2):456–464.

Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., Katta, S., Donnelly, J., Chen, H., and Boner, Z. (2024). Position: Amazing things come from having many good models. In *Forty-first International Conference on Machine Learning*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Verdinelli, I. and Wasserman, L. (2024). Feature Importance: A Closer Look at Shapley Values and Loco. *Statistical Science*, 39(4):623–636.

Williamson, B. and Feng, J. (2020). Efficient Nonparametric Statistical Inference on Population Feature Importance Using Shapley Values. In *International Conference on Machine Learning*, pages 10282–10291. PMLR.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2021). A General Framework for Inference on Algorithm-Agnostic Variable Importance. *Journal of the American Statistical Association*, pages 1–14.

Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M., and Rudin, C. (2022). Exploring the Whole Rashomon Set of Sparse Decision Trees. *Advances in Neural Information Processing Systems*, 35:22305–22315.

Zhang, L. and Janson, L. (2020). Floodgate: Inference for Model-Free Variable Importance. *arXiv Preprint arXiv:2007.01283*.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes** – we explicitly state the assumptions behind our theory.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Not Applicable** – our framework inherits the runtime of the underlying Rashomon set computation algorithm, and as such is not interesting to provide in general.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes**
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **Yes** – these are provided in Appendix E.
   (c) Clear explanations of any assumptions. **Yes** – we contextualize each assumption with illustrative examples.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes** – these are available in the GitHub link.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes** – hyperparameters are described in the text, and data splits can be found through the code.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes**
   (b) The license information of the assets, if applicable. **Not Applicable** – public code packages and datasets were used as existing assets, and as such each have license information available.
   (c) New assets either in the supplemental material or as a URL, if applicable. **Yes**
   (d) Information about consent from data providers/curators. **Not Applicable** – only public data was used.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

Doctor Rashomon and the UNIVERSE of Madness: Variable Importance with Unobserved Confounding and the Rashomon Effect (Appendix)

A NOTATION/DEFINITIONS

Notation	Definition
$\mathcal{D}^{(n)}$	The observed dataset of n samples
x	Observed covariates for a single observation
u	Unobserved covariates for a single observation
y	The outcome of interest
g^*	The conditional mean function of y given *all* covariates
f_u	The conditional mean function of *observed* variables given the fixed values of the *unobserved* variables to u
S^*	The set of all ground-truth conditional submodels
$\mathcal{R}$	The population rashomon set
$\hat{\mathcal{R}}^{(n)}$	The rashomon set estimated from n observations
$\mathcal{X}$	The space of *observed* covariates
$\mathcal{U}$	The space of *unobserved* covariates
$\mathcal{Y}$	The space of outcomes
ϵ_{unobs}	The (assumed) maximum expected loss over conditional submodels when applied to the full population
δ	The allowed probability that a model in the population Rashomon set is not in our estimated Rashomon set
ε_n	Finite sample epsilon correction to insure our estimated Rashmon Set contains the population Rashomon set with high probability
γ	The allowed probability that the estimated variable importance for a model is outisde our predicted bounds on variable importance
λ_{sup}	Worst case regularization possible from model class
τ	The maximum shift in variable importance over conditional submodels due to conditioning on a given value of the unobserved covariates
τ_j	The maximum shift in variable importance over conditional submodels due to conditioning on a given value of the unobserved covariates
ϕ_j	The variable importance for variable j, with respect to a single sample
Φ_j	The expectation of ϕ_j over samples from the data distribution

Table 1: A brief summary of the notation used in this work.

B MOTIVATIONAL EXAMPLE

In this section, we provide a brief example where unobserved confounding leads to misleading results. Let $X_1, X_2, X_3, U \sim \text{Bernoulli}(p = 0.5)$, and our label be generated as:

$$Y = \begin{cases} X_1 \text{ XOR } X_2 & \text{if } U = 0 \\ X_1 \text{ XOR } X_3 & \text{otherwise} \end{cases}$$

Let our model class of interest $\mathcal{F}$ consist of universal approximators, and the algorithm considered $\mathcal{A}$ always produce an optimal model for the given data. We will consider 0-1 loss (or, equivalently in this case, MSE).

If we observe X_1, X_2, X_3, and allow $n \to \infty$, note that there are four settings of X_1, X_2, X_3 which contain two distinct values for Y. Thus, the Bayes' optimal accuracy over this data is 75%, and any model that achieves this accuracy is optimal. Two such models are:

$$f_1(X_1, X_2, X_3) = X_1 \text{ XOR } X_2$$
$$f_2(X_1, X_2, X_3) = X_1 \text{ XOR } X_3$$

Thus, $\mathcal{A}(X_1, X_2, X_3)$ may produce either of f_1, f_2. We therefore have that

$$\begin{aligned} LOCO(X_2) &= \ell(\mathcal{A}(X_1, X_3, Y)) - \ell(\mathcal{A}(X_1, X_2, X_3, Y)) \\ &= \ell(f_2) - \ell(f_1) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} LOCO(X_3) &= \ell(\mathcal{A}(X_1, X_2, Y)) - \ell(\mathcal{A}(X_1, X_2, X_3, Y)) \\ &= \ell(f_1) - \ell(f_2) \\ &= 0, \end{aligned}$$

even though X_2 and X_3 *are* used by the DGP. In fact, a wide class of measures of statistical association demonstrate a similar behavior because *our outcome is marginally independent of each variable.* By the definition of conditional probability, we have:

$$\begin{aligned} P(Y = y | X_1 = x_1) &= \frac{P(Y = y, X_1 = x)}{P(X_1 = x)} \\ &= \frac{4/16}{8/16} \\ &= P(Y = y) \\ P(Y = y | X_2 = x_2) &= \frac{P(Y = y, X_2 = x_2)}{P(X_2 = x_2)} \\ &= \frac{4/16}{8/16} \\ &= P(Y = y) \\ P(Y = y | X_3 = x_3) &= \frac{P(Y = y, X_3 = x_3)}{P(X_3 = x)} \\ &= \frac{4/16}{8/16} \\ &= P(Y = y) \end{aligned}$$

Consider an arbitrary measure of statistical association M satisfying the "zero-independence" property (i.e., $M(X, Y) = 0$ if and only if $X \perp Y$). Because $X_2 \perp Y$, $X_2 \perp Y$, and $X_3 \perp Y$, we have that $M(X_1, Y) = M(X_2, Y) = M(X_3, Y) = 0$.

C ELABORATION ON ASSUMPTION 1

In the main paper, we introduced the following assumption:

Assumption 2. *Assume that there exists a known τ_j such that, for all $u \in \mathcal{U}$,*

$$\left|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u})\right| \leq \tau_j.$$

In this section, we elaborate on the meaning of the term $|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u})|$ through an example. Consider the data distribution reflected in Table 2 with each row of the table having equal probability of being drawn. In Table 2, the first four rows have $U = 0$, and as such $Y = f_0(X) := X_1$. In the second four, we have $U = 1$ and $Y = f_1(X) := X_1 \neq X_2$. In this setting, U is unobserved. Because there are only two values for U to take, we will compute quantities for $\mathcal{P}_{XY|U=0}$ using the first four rows, and $\mathcal{P}_{XY|U\neq 0}$ using the second four.

We will step through computing the subtractive model reliance and 0-1 loss for f_0. First, by definition, we know

$$\Phi_1(f_0, \mathcal{P}_{XY|U=u}) = \mathbb{E}_{\mathcal{P}_{XY|U=u}}[\mathbb{E}_{X_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(X_i, X_{i'}), Y_i)]] - \mathbb{E}_{\mathcal{P}_{XY|U=u}}[\ell(f_0, X_i, Y_i)],$$

where $swap_j(A, B)$ is a function replacing the j-th entry of A with the j-th entry of B. Because we make perfect predictions for the subpopulation with $U = 0$ using the model f_0, we know that $\mathbb{E}_{\mathcal{P}_{XY|U=u}}[\ell(f_0, X_i, Y_i)] = 0$ and $\Phi_1(f_0, \mathcal{P}_{XY|U=u}) = \mathbb{E}_{\mathcal{P}_{XY|U=u}}[\mathbb{E}_{X_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(X_i, X_{i'}), Y_i)]]$.

The term $\mathbb{E}_{\mathcal{P}_{XY|U=u}}[\mathbb{E}_{X_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(X_i, X_{i'}), Y_i)]]$ is slightly onerous to compute by hand, but can be intuited from the construction of the problem. Within $\mathcal{P}_{XY|U=u}$, there is equal probability that $X_1 = 0$ and $X_1 = 1$. Because f_0 simply returns the value of X_1, drawing a random value X_1' from the marginal distribution of X_1 will have equal likelihood of producing $f_0(X') = 0$ and $f_0(X') = 1$. Thus, given the fact that $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 0.5$, we have $\mathbb{E}_{\mathcal{P}_{XY|U=u}}[\mathbb{E}_{X_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(X_i, X_{i'}), Y_i)]] = 0.5$, yielding $\Phi_1(f_0, \mathcal{P}_{XY|U=u}) = 0.5$

We now turn to compute $\Phi_j(f_u, \mathcal{P}_{XY|U\neq u})$. In this case, we consider the bottom four rows of Table 2, yielding:

$$\begin{aligned}\Phi_j(f_u, \mathcal{P}_{XY|U\neq u}) &= \mathbb{E}_{\mathcal{P}_{XY|U\neq u}}[\mathbb{E}_{X_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(X_i, X_{i'}), Y_i)]] - \mathbb{E}_{\mathcal{P}_{XY|U\neq u}}[\ell(f_0, X_i, Y_i)] \\ &= \mathbb{E}_{\mathcal{P}_{XY|U\neq u}}[\mathbb{E}_{X_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(X_i, X_{i'}), Y_i)]] - 0.5\end{aligned}$$

The term $\mathbb{E}_{\mathcal{P}_{XY|U\neq u}}[\mathbb{E}_{X_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(X_i, X_{i'}), Y_i)]]$ is similarly annoying to compute by hand, but can be intuited from the construction of the problem by the exact same reasoning. Within $\mathcal{P}_{XY|U\neq u}$, there is equal probability that $X_1 = 0$ and $X_1 = 1$. Because f_0 simply returns the value of X_1, drawing a random value X_1' from the marginal distribution of X_1 will have equal likelihood of producing $f_0(X') = 0$ and $f_0(X') = 1$. Thus, given the fact that $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 0.5$, we have $\mathbb{E}_{\mathcal{P}_{XY|U\neq u}}[\mathbb{E}_{x_{i'}\sim\mathcal{P}_X}[\ell(f_0, swap_1(x_i, x_{i'}), y_i)]] = 0.5$. However, in this case the initial loss for f_0 was poorer; as such, we have $\Phi_1(f_0, \mathcal{P}_{XY|U\neq u}) = 0.5 - 0.5 = 0$.

The shift in the distribution of Y given $U = 1$ vs $U = 0$ changes the importance of X_1 for the conditional submodel f_0. When this effect happens, τ_j is large, because it measures the change in variable importance for conditional submodels across different subgroups of unobserved features. Similarly, if the distribution of X_1 changed substantially between conditions, we would expect τ_j to be larger.

X_1	X_2	U	$f_0(X)$	$f_1(X)$	Y
0	0	0	0	0	0
1	0	0	1	1	1
0	1	0	0	1	0
1	1	0	1	0	1
0	0	1	0	0	0
1	0	1	1	1	1
0	1	1	0	1	1
1	1	1	1	0	0

Table 2: The full distribution of data used for the example below. Each row in this table has equal probability of occurring. In this example, $f_0(X) := X_1$ and $f_1(X) := X_1 \neq X_2$

D APPLICABLE VARIABLE IMPORTANCE METRICS

Our framework primarily considers variable importance metrics that can be expressed as the expectation of some sample-wise variable importance quantity. Here, we list two example variable metrics that fit this definition.

Subtractive Model Reliance is defined by Fisher et al. (2019) as

$$\Phi_j(f, \mathcal{P}_{XY}) = \mathbb{E}_{\mathcal{P}_{XY}}[\mathbb{E}_{x_{i'} \sim \mathcal{P}_X}[\ell(f, swap_j(x_i, x_{i'}), y_i)]] - \mathbb{E}_{\mathcal{P}_{XY}}[\ell(f, x_i, y_i)]$$

where

$$swap_j(x_i, x_{i'}) := \begin{bmatrix} x_{i,1} & x_{i,2} & \dots x_{i',j} & \dots x_{i,p} \end{bmatrix}^T$$

is a function that replaces the $j-$th entry of x_i with that of $x_{i'}$. We can express this as the expectation of a sample-wise variable importance quantity as follows

$$\begin{aligned} \Phi_j(f, \mathcal{P}_{XY}) &= \mathbb{E}_{\mathcal{P}_{XY}}[\mathbb{E}_{x_{i'} \sim \mathcal{P}_X}[\ell(f, swap_j(x_i, x_{i'}), y_i)]] - \mathbb{E}_{\mathcal{P}_{XY}}[\ell(f, x_i, y_i)] \\ &= \mathbb{E}_{\mathcal{P}_{XY}}[\mathbb{E}_{x_{i'} \sim \mathcal{P}_X}[\ell(f, swap_j(x_i, x_{i'}), y_i)] - \ell(f, x_i, y_i)] \\ &= \mathbb{E}_{\mathcal{P}_{XY}}\left[\phi_j(f, x_i, y_i)\right] \end{aligned}$$

for $\phi_j(f, x_i, y_i) := \mathbb{E}_{x_{i'} \sim \mathcal{P}_X}[\ell(f, swap_j(x_i, x_{i'}), y_i)] - \ell(f, x_i, y_i)$.

SHAP is defined by Lundberg and Lee (2017) as a sample-wise variable importance metric. Let $\phi_j^{SHAP}(f, x_i, y_i)$ denote the SHAP importance for feature j on sample i. In the SHAP package, global importance values are obtained by taking the mean absolute value over i; that is, $\Phi_j(f, \mathcal{D}) := \mathbb{E}[|\phi_j^{SHAP}(f, x_i, y_i)|]$. Similar reasoning applies to a wide range of extensions of SHAP.

Average Integrated Gradients Suppose our function class is composed of only differentiable functions $\mathcal{F}$. Choose some $f \in \mathcal{F}$. As defined in Sundararajan et al. (2017), for a given sample (x_i, y_i), its integrated gradients is computed as

$$\phi_j(f, (x_i, y_i)) = (x_i - x_i')) \int_0^1 \frac{\partial f(x_i + \alpha(x_i - x_i')}{dx_i} d\alpha$$

for some alternative covariate profile x_i'. To measure the importance of only feature j, we can define $x_{i,j}' = x_{i,j} + \Delta$ for some fixed Δ and keep all other features $x_{i,j'}' = x_{i,j}$ fixed. We can now measure the average importance of feature j as an expectation of sample-level integrated gradients over the empirical or population distribution.

E PROOFS

Proposition 1. *Given Definition 1, we know that* $S^* \subseteq \mathcal{R}(\epsilon_{unobs}; \lambda_0)$.

Proof. Definition 1 states that, for some $\epsilon_{unobs} > 0$,

$$\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f_u, X, Y)] \leq \epsilon_{unobs} \quad \forall f_u \in S^*. \tag{4}$$

The Rashomon set with regularization λ_0 is defined as

$$\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda_0, \mathcal{P}_{XY}) := \{f \in \mathcal{F} : \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda) + 0 \leq \epsilon]\}.$$

Thus, $S^* \subseteq \mathcal{R}(\epsilon_{unobs}; \lambda_0)$ by the definition of $\mathcal{R}(\epsilon_{unobs}; \lambda_0)$. □

Theorem 1 (Recovering the Population Rashomon Set)**.** *For any loss function ℓ bounded between $\ell_{\min}$ and $\ell_{\max}$ and $\epsilon \in [\ell_{\min}, \ell_{\max}]$, it holds that*

$$\mathbb{P}\left(\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \subseteq \hat{\mathcal{R}}^{(n)}\left(\varepsilon_n + \epsilon + \lambda_{\sup}; \mathcal{F}, \ell, \lambda\right)\right) \geq 1 - \delta, \text{ where } \varepsilon_n = \sqrt{\frac{(\ell_{\max} - \ell_{\min})^2 \ln\left(\frac{C}{\delta}\right)}{2n}}$$

for sample size n, for any value $C \geq |\mathcal{R}(\epsilon_{unobs}; \mathcal{F}, \ell, \lambda)|$ (e.g., $C := |\mathcal{F}|$), regularization penalty λ, and regularization upper bound $\lambda_{\sup} = \sup_{f\in\mathcal{F}} \lambda(f)$.

Moreover,

$$\mathbb{P}\left(\hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda) \subseteq \mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda)\right) \geq 1 - |\mathcal{F}| \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{max} - \ell_{min})^2}\right\}.$$

Proof. Throughout this proof, we denote the expected loss $\mathbb{E}_{X,Y\sim\mathcal{P}_{XY}}\ell(f, X, Y)$ simply as $\ell(f)$ and the empirical loss $\frac{1}{n}\sum_{i=1}^n \ell(f, x_i, y_i)$ as $\hat{\ell}(f)$. All probabilities are with respect to $X, Y \sim \mathcal{P}_{XY}$. First, recall that we can form the following uniform bound:

$$\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \text{ s.t. } f \notin \hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda)\right) \\
&= \mathbb{P}\left(\exists f \in \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) + \lambda(f) > \epsilon + \varepsilon_n + \lambda_{sup}\right) \text{ by definition of Rashomon set exclusion} \\
&\leq \mathbb{P}\left(\exists f \in \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) > \epsilon + \varepsilon_n\right) \text{ because } \lambda_{sup} - \lambda(f) \geq 0 \text{ by definition of } \lambda_{sup} \\
&= \mathbb{P}\left(\exists f \in \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) - \ell(f) > \epsilon + \varepsilon_n - \ell(f)\right) \text{ by subtracting } \ell(f) \text{ from both sides} \\
&\leq \mathbb{P}\left(\exists f \in \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) - \ell(f) > \varepsilon_n\right) \text{ because } \epsilon - \ell(f) \geq 0 \text{ for } f \in \mathcal{R}(\epsilon) \\
&\leq \sum_{f\in\mathcal{R}(\epsilon)} \mathbb{P}\left(\hat{\ell}(f) - \ell(f) > \varepsilon_n\right) \text{ by the Union bound} \\
&\leq \sum_{f\in\mathcal{R}(\epsilon)} \exp\left\{\frac{-2n\varepsilon_n^2}{(b-a)^2}\right\} \text{ by Hoeffding's inequality} \\
&= |\mathcal{R}(\epsilon)| \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{\max} - \ell_{\min})^2}\right\} \\
&\leq C \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{\max} - \ell_{\min})^2}\right\}
\end{aligned}$$

for any $C \geq |\mathcal{R}(\epsilon)|$.

We can form the complementary bound as follows

$$
\begin{aligned}
&\mathbb{P}\left(\exists f \notin \mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \text{ s.t. } f \in \hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda)\right) \\
&= \mathbb{P}\left(\exists f \notin \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) + \lambda(f) \leq \epsilon + \varepsilon_n + \lambda_{sup}\right) \text{ by definition of Rashomon set exclusion} \\
&\geq \mathbb{P}\left(\exists f \notin \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) \leq \epsilon + \varepsilon_n\right) \text{ because } \lambda_{sup} - \lambda(f) \geq 0 \text{ by definition of } \lambda_{sup} \\
&= \mathbb{P}\left(\exists f \notin \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) - \ell(f) \leq \epsilon + \varepsilon_n - \ell(f)\right) \text{ by subtracting } \ell(f) \text{ from both sides} \\
&\geq \mathbb{P}\left(\exists f \notin \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) - \ell(f) \leq \varepsilon_n\right) \text{ because } \epsilon - \ell(f) \geq 0 \text{ for } f \notin \mathcal{R}(\epsilon) \\
&= 1 - \mathbb{P}\left(\forall f \notin \mathcal{R}(\epsilon) \text{ s.t. } \hat{\ell}(f) - \ell(f) > \varepsilon_n\right) \\
&\geq 1 - \sum_{f \in \mathcal{F}} \mathbb{P}\left(\hat{\ell}(f) - \ell(f) > \varepsilon_n\right) \text{ by the Union bound} \\
&\geq 1 - \sum_{f \in \mathcal{F}(\epsilon)} \exp\left\{\frac{-2n\varepsilon_n^2}{(b-a)^2}\right\} \text{ by Hoeffding's inequality} \\
&= 1 - |\mathcal{F}| \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{\max} - \ell_{\min})^2}\right\}.
\end{aligned}
$$

Thus, for a given ε_n, we have that the probability of a type-1 error is **upper** bounded by $C \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{\max}-\ell_{\min})^2}\right\}$, and the probability of a type-2 error is **lower** bounded by $1 - |\mathcal{F}| \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{\max}-\ell_{\min})^2}\right\}$.

Define $\delta := C \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{\max}-\ell_{\min})^2}\right\}$. We can rearrange the upper bound to express ε_n in terms of a desired probability δ as follows:

$$
\begin{aligned}
&C \exp\left\{\frac{-2n\varepsilon_n^2}{(\ell_{\max} - \ell_{\min})^2}\right\} = \delta \\
&\iff \frac{-2n\varepsilon_n^2}{(\ell_{\max} - \ell_{\min})^2} = \ln(\frac{\delta}{C}) \\
&\iff \varepsilon_n^2 = \frac{(\ell_{\max} - \ell_{\min})^2 \ln(\frac{\delta}{C})}{-2n} \\
&\iff \varepsilon_n = \sqrt{\frac{(\ell_{\max} - \ell_{\min})^2 \ln(\frac{\delta}{C})}{-2n}} \\
&\iff \varepsilon_n = \sqrt{\frac{(\ell_{\max} - \ell_{\min})^2 \ln(\frac{C}{\delta})}{2n}}
\end{aligned}
$$

Thus, using this value of ε_n, we have

$$
\begin{aligned}
\varepsilon_n &= \sqrt{\frac{(\ell_{\max} - \ell_{\min})^2 \ln(\frac{C}{\delta})}{2n}} \\
&\implies \mathbb{P}\left(\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \not\subseteq \hat{\mathcal{R}}^{(n)}\left(\varepsilon_n + \epsilon + \lambda_{\sup}; \mathcal{F}, \ell, \lambda\right)\right) \leq \delta \\
&\iff \mathbb{P}\left(\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \subseteq \hat{\mathcal{R}}^{(n)}\left(\varepsilon_n + \epsilon + \lambda_{\sup}; \mathcal{F}, \ell, \lambda\right)\right) \geq 1 - \delta
\end{aligned}
$$

as required. □

Corollary 1 (Capturing Each Unobserved Submodel)**.** *Let ϵ_{unobs} be defined as in Definition 1, and assume that $S^* \subseteq \mathcal{F}$. For any loss function ℓ bounded between $\ell_{\min}$ and $\ell_{\max}$, it holds that*

$$
\mathbb{P}\left(S^* \subseteq \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup}; \mathcal{F}, \ell, \lambda)\right) \geq 1 - \delta,
$$

for a sample of size n and regularization penalty λ, with ε_n defined as in Theorem 1.

Proof. We have that:

$$\begin{aligned}
&\mathbb{P}\left(S^* \subseteq \hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda)\right) \\
&= 1 - \mathbb{P}\left(\exists f \in S^* \text{ s.t. } f \notin \hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda)\right) \\
&\geq 1 - \mathbb{P}\left(\exists f \in \mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \notin \hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda)\right) \text{ because } \ell(f) \leq \epsilon \forall f \in S^* \\
&= \mathbb{P}\left(\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) \subseteq \hat{\mathcal{R}}^{(n)}(\epsilon + \varepsilon_n + \lambda_{sup}; \mathcal{F}, \ell, \lambda)\right) \\
&\geq 1 - \delta
\end{aligned}$$

for $\varepsilon_n = \sqrt{\frac{(\ell_{\max} - \ell_{\min})^2 \ln\left(\frac{C}{\delta}\right)}{2n}}$ by Theorem 1.

□

Proposition 2. *The size of the estimated Rashomon set with threshold $\epsilon' > \epsilon$ is an upper bound on the size of the population Rashomon set with threshold ϵ with high probability:*

$$\mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(|\hat{\mathcal{R}}^{(n)}(\epsilon')| > |\mathcal{R}(\epsilon)|\right) = 1 - O(n^{-1}).$$

Proof. Recall how we define our estimated and population Rashomon sets:

$$\mathcal{R}(\epsilon; \mathcal{F}, \ell, \lambda) = \{f \in \mathcal{F} : \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda) + \lambda(f)] \leq \epsilon\}$$
$$\hat{\mathcal{R}}^{(n)}(\epsilon'; \mathcal{F}, \ell, \lambda) = \left\{f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} \ell(f, x_i, y_i) + \lambda(f) \leq \epsilon'\right\}.$$

Before moving further, recognize that if all models in $\mathcal{R}(\epsilon)$ are members of the estimated Rashomon set $\hat{\mathcal{R}}^{(n)}(\epsilon')$ (i.e., $\mathcal{R}(\epsilon) \subseteq \hat{\mathcal{R}}^{(n)}(\epsilon')$, then $|\hat{\mathcal{R}}^{(n)}(\epsilon')| > |\mathcal{R}(\epsilon)|$). Therefore, we will find a lower bound on the probability of $|\hat{\mathcal{R}}^{(n)}(\epsilon')| > |\mathcal{R}(\epsilon)|$ and show that this goes to 1 asymptotically.

$$\begin{aligned}
&\mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(|\hat{\mathcal{R}}^{(n)}(\epsilon')| > |\mathcal{R}(\epsilon)|\right) \\
&\geq \mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(\mathcal{R}(\epsilon) \subseteq \hat{\mathcal{R}}^{(n)}(\epsilon')\right) \text{ (from above discussion)} \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}(\forall f \in \mathcal{R}(\epsilon), f \in \hat{\mathcal{R}}^{(n)}(\epsilon')) \text{ (by definition of } \subseteq) \\
&= 1 - \mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}(\exists f \in \mathcal{R}(\epsilon), f \notin \hat{\mathcal{R}}^{(n)}(\epsilon')) \text{ (by law of total probability)} \\
&\geq 1 - \sum_{f\in\mathcal{R}(\epsilon)} \mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}(f \notin \hat{\mathcal{R}}^{(n)}(\epsilon')) \text{ (by Union bound).}
\end{aligned}$$

Let us work on bounding this model-level probability. First, recall when a model is in the estimated and population Rashomon sets: when the empirical average or expectation over sample-level losses is below the specified threshold. So, we now want to bound for a given model f in the population ϵ-threshold Rashomon set,

$$\begin{aligned}
&\mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(f \notin \hat{\mathcal{R}}^{(n)}(\epsilon')\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(f, X_i, Y_i; \lambda) + \lambda(f) > \epsilon'\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(f, X_i, Y_i; \lambda) - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)] > \epsilon' - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)]\right) \\
&\leq \frac{\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \ell(f, X_i, Y_i; \lambda) - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)]\right)^2\right]}{\left(\epsilon' - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)]\right)^2} \text{ (by Chebyshev's inequality)} \\
&= \frac{\sum_{i=1}^{n} \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}\left[\left(\ell(f, X_i, Y_i; \lambda) - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)]\right)^2\right]}{n^2\left(\epsilon' - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)]\right)^2} \text{ (because data are IID)} \\
&\leq \frac{\sum_{i=1}^{n} (\ell_{\max} - \ell_{\min})^2}{4n^2\left(\epsilon' - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)]\right)^2} \text{ (because range of loss is bounded)} \\
&= \frac{(\ell_{\max} - \ell_{\min})^2}{4n\left(\epsilon' - \mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY}}[\ell(f, X, Y; \lambda)]\right)^2} \text{ (by canceling out constant factors)} \\
&\leq \frac{(\ell_{\max} - \ell_{\min})^2}{4n\left(\epsilon' - \epsilon\right)^2} \text{ (because expected loss of } f \text{ is smaller than } \epsilon).
\end{aligned}$$

Plugging this last result back into the Union bound across all models, we can see then that

$$\begin{aligned}\mathbb{P}_{(X,Y)\sim\mathcal{P}_{XY}}\left(|\hat{\mathcal{R}}^{(n)}(\epsilon')| > |\mathcal{R}(\epsilon)|\right) &\geq 1 - \sum_{f\in\mathcal{R}(\epsilon)} \frac{(\ell_{\max}-\ell_{\min})^2}{4n\left(\epsilon'-\epsilon\right)^2} \\ &= 1 - |\mathcal{R}(\epsilon)|\frac{(\ell_{\max}-\ell_{\min})^2}{4n\left(\epsilon'-\epsilon\right)^2} \\ &= 1 - O(n^{-1}) \text{ (because all other factors are constants)}\end{aligned}$$

□

Corollary 2. *Let ϵ_{unobs} be defined as in Definition 1, and assume that $S^* \subseteq \mathcal{F}$. Define*

$$\hat{\varepsilon}_n = \sqrt{\frac{(\ell_{\max}-\ell_{\min})^2 \ln\left(\frac{|\hat{\mathcal{R}}^{(n)}(\epsilon'+\lambda_{\sup})|}{\delta}\right)}{2n}},$$

using the size of an estimated Rashomon set with threshold $\epsilon' > \epsilon_{unobs}$. Then, for any loss function ℓ bounded between $\ell_{\min}$ and $\ell_{\max}$, it holds that

$$\mathbb{P}\left(S^* \subset \hat{\mathcal{R}}^{(n)}(\hat{\varepsilon}_n + \epsilon_{unobs} + \lambda_{\sup}; \mathcal{F}, \ell, \lambda)\right) \geq 1 - \delta + O(n^{-1}),$$

given a sample of size n and regularization penalty λ.

Proof. The proof connects Corollary 1 with Proposition 2:

$$\begin{aligned}&\mathbb{P}\left(S^* \subset \hat{\mathcal{R}}^{(n)}(\hat{\varepsilon}_n + \epsilon_{unobs} + \lambda_{\sup}; \mathcal{F}, \ell, \lambda)\right) \\ &\geq \mathbb{P}\left(S^* \subset \hat{\mathcal{R}}^{(n)}(\hat{\varepsilon}_n + \epsilon_{unobs} + \lambda_{\sup}; \mathcal{F}, \ell, \lambda) \,\middle|\, |\hat{\mathcal{R}}^{(n)}(\epsilon'+\lambda_{\sup})| > |\mathcal{R}(\epsilon_{unobs})|\right) \mathbb{P}\left(|\hat{\mathcal{R}}^{(n)}(\epsilon'+\lambda_{\sup})| > |\mathcal{R}(\epsilon_{unobs})|\right) \\ &\geq (1-\delta)\left(1 - O(n^{-1})\right) \text{ (from Theorem 1 and Proposition 2)} \\ &= 1 - \delta - (1-\delta)O(n^{-1}) \\ &= 1 - \delta - O(n^{-1}).\end{aligned}$$

□

Theorem 2 (Capturing Variable Importance for Each Submodel)**.** *Let $\gamma \in (0,1)$ and α be a value such that,*

$$\mathbb{P}_{\mathcal{D}^{(n)}}\left(\forall f \in S^*, \Phi_j(f, \mathcal{P}_{XY}) \in \left[\Phi_j(f, \mathcal{D}^{(n)}) \pm \alpha\right]\right) \geq 1 - \gamma,$$

where $\mathbb{P}_{\mathcal{D}^{(n)}}$ denotes the probability of drawing the observed n samples from $\mathcal{P}_{XY}$. It follows that, for all $f \in S^*$, with probability at least $1 - (\delta + \gamma)$,

$$\{\Phi_j(f, \mathcal{P}_{XY}) \mid f \in S^*\} \subseteq \left[\begin{array}{l} \inf_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) - \alpha, \\ \sup_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) + \alpha \end{array}\right],$$

where ε_n and δ are defined as in Corollary 1.

Proof.

$$\mathbb{P}\left(\{\Phi_j(f, \mathcal{P}_{XY}) \mid f \in S^*\} \subseteq \left[\inf_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) - \alpha, \sup_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) + \alpha\right]\right)$$

$$\geq \mathbb{P}\Big(\left(S^* \subseteq \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})\right) \cap \left(\Phi_j(f_u, \mathcal{P}_{XY}) \in \left[\Phi_j(f_u, \mathcal{D}^{(n)}) - \alpha, \Phi_j(f_u, \mathcal{D}^{(n)}) + \alpha\right] \forall f_u \in S^*\right)\Big)$$

Because $S^* \subseteq \hat{\mathcal{R}}^{(n)}$ and the true VI for each f_u contained in the interval is a sufficient condition for above

$$= 1 - \mathbb{P}\left(\left(S^* \not\subseteq \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})\right) \cup \left(\exists f_u \in S^* \text{ s.t. } \Phi_j(f_u, \mathcal{P}_{XY}) \notin \left[\Phi_j(f_u, \mathcal{D}^{(n)}) - \alpha, \Phi_j(f_u, \mathcal{D}^{(n)}) + \alpha\right]\right)\right)$$

By the law of total probability

$$\geq 1 - \mathbb{P}\left(S^* \not\subseteq \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})\right) - \mathbb{P}\left(\exists f_u \in S^* \text{ s.t. } \Phi_j(f_u, \mathcal{P}_{XY}) \notin \left[\Phi_j(f_u, \mathcal{D}^{(n)}) - \alpha, \Phi_j(f_u, \mathcal{D}^{(n)}) + \alpha\right]\right)$$

By the Union Bound

$$= \mathbb{P}\left(S^* \subseteq \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})\right) - \mathbb{P}\left(\exists f_u \in S^* \text{ s.t. } \Phi_j(f_u, \mathcal{P}_{XY}) \notin \left[\Phi_j(f_u, \mathcal{D}^{(n)}) - \alpha, \Phi_j(f_u, \mathcal{D}^{(n)}) + \alpha\right]\right)$$

$$\geq 1 - \delta - \mathbb{P}\left(\exists f_u \in S^* \text{ s.t. } \Phi_j(f_u, \mathcal{P}_{XY}) \notin \left[\Phi_j(f_u, \mathcal{D}^{(n)}) - \alpha, \Phi_j(f_u, \mathcal{D}^{(n)}) + \alpha\right]\right) \text{ By Corollary 1}$$

$$\geq 1 - \delta - \gamma \text{ By definition of } \alpha$$

That is,

$$\mathbb{P}\left(\{\Phi_j(f, \mathcal{P}_{XY}) \mid f \in S^*\} \subseteq \left[\sup_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) - \alpha, \inf_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) + \alpha\right]\right)$$

$$\geq 1 - (\delta + \gamma)$$

as required.

□

Lemma 1. *Let τ_j be defined as in Assumption 1. It holds that*

$$\Phi_j(g^*, \mathcal{P}_U) \in \left[\min_{f_u} \Phi_j(f_u, \mathcal{P}_{XY}) - \tau_j, \max_{f_u} \Phi_j(f_u, \mathcal{P}_{XY}) + \tau_j\right]$$

Proof. First, realize that the variable importance for g^* on the complete data distribution–which includes X, U, Y– is actually a convex combination of the variable importance for each f on its corresponding conditional data distribution:

$$\begin{aligned}
\Phi_j(g^*, \mathcal{P}_U) &= \mathbb{E}_{(X,U,Y)\sim\mathcal{P}_U}\left[\phi_j(g^*, (X, U, Y))\right] \text{ by definition of } \Phi_j \\
&= \sum_{u\in\mathcal{U}} \mathbb{P}(U = u)\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY|U=u}}\left[\phi_j(g^*, (X, u, Y)) \mid U = u\right] \text{ by the law of iterated expectation.}
\end{aligned}$$

Now, because we have conditioned on $U = u$, we know that $g^* = f_u$. This equality means that we can rewrite

$$\begin{aligned}
&\sum_{u\in\mathcal{U}} \mathbb{P}(U = u)\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY|U=u}}\left[\phi_j(g^*, (X, u, Y)) \mid U = u\right] \\
&= \sum_{u\in\mathcal{U}} \mathbb{P}(U = u)\mathbb{E}_{(X,Y)\sim\mathcal{P}_{XY|U=u}}\left[\phi_j(f_u, (X, Y)) \mid U = u\right] \\
&= \sum_{u\in\mathcal{U}} \mathbb{P}(U = u)\Phi_j(f_u, \mathcal{P}_{XY|U=u}) \text{ by definition of } \Phi_j.
\end{aligned}$$

In other words, $\Phi_j(g^*, \mathcal{P}_U)$ is a convex combination of $\Phi_j(f_u, \mathcal{P}_{XY|U=u})$ across all $u \in \mathcal{U}$. Because of this convexity, we then know that $\Phi_j(g^*, \mathcal{P}_U)$ is upper and lower bounded by the conditional submodel importances:

$$\min_u \Phi_j(f_u, \mathcal{P}_{XY|U=u}) \leq \Phi_j(g^*, \mathcal{P}_U) \leq \max_u \Phi_j(f_u, \mathcal{P}_{XY|U=u}).$$

Unfortunately, these upper and lower bounds are still in terms of non-identifiable distributions because we never know when $U = u$ or $U \neq u$. However, we will next show how we can construct upper and lower bounds on $\Phi_j(f_u, \mathcal{P}_{XY|U=u})$ in terms of the observable distribution $\mathcal{P}_{XY}$ and our parameter τ_j.

Choose some $u \in \mathcal{U}$. Recognize that we can compute the bias in our variable importance estimate by examining the difference between $\Phi_j(f_u, \mathcal{P}_{XY|U=u})$ and $\Phi_j(f_u, \mathcal{P}_{XY})$. First, recognize by the law of iterated expectation that

$$\Phi_j(f_u, \mathcal{P}_{XY}) = \mathbb{P}(U = u)\Phi_j(f_u, \mathcal{P}_{XY|U=u}) + \mathbb{P}(U \neq u)\Phi_j(f_u, \mathcal{P}_{XY|U\neq u}).$$

So, we can rewrite the difference between conditional and marginal VI for f_u as

$$\begin{aligned}
&\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY}) \\
&= \Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \mathbb{P}(U = u)\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \mathbb{P}(U \neq u)\Phi_j(f_u, \mathcal{P}_{XY|U\neq u})) \\
&= (1 - \mathbb{P}(U = u))\,\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \mathbb{P}(U \neq u)\Phi_j(f_u, \mathcal{P}_{XY|U\neq u})) \\
&= \mathbb{P}(U \neq u)\left[\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right]
\end{aligned}$$

where the last two lines come the law of total probability and factoring out the $\mathbb{P}(U \neq u)$ term. In other words, the distance between the conditional VI and marginal VI for f_u is broken into two parts: what are the chances that an observation is in another unobserved subgroup, and what is the distance between the VI for conditional subgroups. Let us use this decomposition to first find an upper bound on $\Phi_j(f_u, \mathcal{P}_{XY|U=u})$:

$$\begin{aligned}
\Phi_j(f_u, \mathcal{P}_{XY|U=u}) &= \Phi_j(f_u, \mathcal{P}_{XY}) + \mathbb{P}(U \neq u)\left[\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right] \\
&\leq \Phi_j(f_u, \mathcal{P}_{XY}) + \mathbb{P}(U \neq u)\left|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right| \\
&\qquad \text{by definition of absolute value} \\
&\leq \Phi_j(f_u, \mathcal{P}_{XY}) + \left|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right| \\
&\qquad \text{because } \mathbb{P}(U \neq u) \leq 1 \\
&\leq \Phi_j(f_u, \mathcal{P}_{XY}) + \max_{u\in\mathcal{U}}\left|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right| \\
&\qquad \text{by definition of max} \\
&= \Phi_j(f_u, \mathcal{P}_{XY}) + \tau_j \text{ by definition of } \tau_j \\
&\leq \max_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY}) + \tau_j \text{ by definition of max}
\end{aligned}$$

Our upper bound shows that for any conditional submodel, the VI for its corresponding conditional subgroup $\mathcal{P}_{XY|U=u}$ is bounded by the largest conditional submodel VI applied to the observed distribution $\mathcal{P}_{XY}$ plus some correction factor τ_j. This correction factor measures how different is the VI for any conditional submodel between different subgroups.

Now, we will derive a lower bound. These steps are very similar to the steps for deriving the upper bound are included for completeness:

$$\begin{aligned}
\Phi_j(f_u, \mathcal{P}_{XY|U=u}) &= \Phi_j(f_u, \mathcal{P}_{XY}) + \mathbb{P}(U \neq u)\left[\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right] \\
&\geq \Phi_j(f_u, \mathcal{P}_{XY}) - \mathbb{P}(U \neq u)\left|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right| \\
&\quad \text{by definition of absolute value} \\
&\geq \Phi_j(f_u, \mathcal{P}_{XY}) - \left|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right| \\
&\quad \text{because } \mathbb{P}(U \neq u) \geq 0 \\
&\geq \Phi_j(f_u, \mathcal{P}_{XY}) - \max_{u\in\mathcal{U}}\left|\Phi_j(f_u, \mathcal{P}_{XY|U=u}) - \Phi_j(f_u, \mathcal{P}_{XY|U\neq u}))\right| \\
&\quad \text{by definition of } -\max \\
&= \Phi_j(f_u, \mathcal{P}_{XY}) - \tau_j \text{ by definition of } \tau_j \\
&\geq \min_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY}) - \tau_j \text{ by definition of min.}
\end{aligned}$$

Now, we will use these upper and lower bounds for any conditional submodel to bound $\Phi_j(g^*, \mathcal{P}_U)$.

We will first focus on the lower bound. From the convex combination argument from earlier, we know that

$$\Phi_j(g^*, \mathcal{P}_U) \geq \min_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY|U=u}).$$

From our derived lower bound at the conditional submodel level, we know that

$$\min_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY|U=u}) \geq \min_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY}) - \tau_j.$$

Therefore,

$$\Phi_j(g^*, \mathcal{P}_U) \geq \min_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY}) - \tau_j.$$

Because we similarly know that

$$\max_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY|U=u}) \leq \max_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY}) + \tau_j,$$

we can also conclude that

$$\Phi_j(g^*, \mathcal{P}_U) \leq \min_{u\in\mathcal{U}} \Phi_j(f_u, \mathcal{P}_{XY}) + \tau_j.$$

That is,

$$\Phi_j(g^*, \mathcal{P}_U) \in \left[\min_{f_u} \Phi_j(f_u, \mathcal{P}_{XY}) - \tau_j, \max_{f_u} \Phi_j(f_u, \mathcal{P}_{XY}) + \tau_j\right]$$

as required. □

Theorem 3. *Let α and γ be defined as in Theorem 2, ε_n and δ be defined as in Theorem 1, and τ_j be defined as in Assumption 1.*

With probability at least $1-(\delta+\gamma)$,

$$\Phi_j(g^*, \mathcal{P}_U) \in \left[\inf_{f \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f, \mathcal{D}^{(n)}) - \tau_j - \alpha, \sup_{f \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f, \mathcal{D}^{(n)}) + \tau_j + \alpha \right]$$

Proof. Theorem 2 states that, with probability at least $1-(\delta+\gamma)$,

$$\{\Phi_j(f, \mathcal{P}_{XY}) \mid f \in S^*\} \subseteq \left[\inf_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) - \alpha, \sup_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) + \alpha \right].$$

This implies that

$$\begin{aligned}
& \inf_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) - \alpha \leq \inf_{f_u \in S^*} \Phi_j(f_u, \mathcal{P}_{XY}) \\
\iff & \inf_{f' \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f', \mathcal{D}^{(n)}) - \alpha - \tau_j \leq \inf_{f_u \in S^*} \Phi_j(f_u, \mathcal{P}_{XY}) - \tau_j \\
& \qquad \leq \Phi_j(g^*, \mathcal{P}_U) \qquad \text{By Lemma 1}
\end{aligned}$$

A symmetric argument applies for the upper bound, yielding that, with probability at least $1-(\delta+\gamma)$,

$$\Phi_j(g^*, \mathcal{P}_U) \in \left[\inf_{f \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f, \mathcal{D}^{(n)}) - \tau_j - \alpha, \sup_{f \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\sup})} \Phi_j(f, \mathcal{D}^{(n)}) + \tau_j + \alpha \right]$$

as required.

□

F EXPERIMENTAL DETAILS

All experiments were run using TreeFarms Xin et al. (2022) to compute the Rashomon set of decision trees, with a depth bound of 3 and a regularization value of 0.001. For computational efficiency, we dropped all but the first 8 variables (before binarization) from each dataset in our semi-synthetic experiments (except for the Compas dataset, which only has 7 features); during binarization, each variable was processed into 3 binary variables as evenly spaced quantiles over the distribution of the original input variable. We used 80% of all data in each setting to compute the Rashomon set, and the remaining 20% to estimate variable importance over this set.

Note that wine quality Cortez et al. (2009) reports the rating for each sample on a 0 to 10 scale. We converted this into binary classification problem where the goal was to predict whether the rating was greater than or equal to 5.

All experiments for this work were performed on an academic institution's cluster computer. We used up to 10 machines in parallel, each with a Dell R730's with 2 Intel Xeon E5-2640 Processors (40 cores).

G EXPERIMENTS WITH APPROXIMATE RANDOM FOREST RASHOMON SETS

The experimental results in the main body of this work used the Rashomon set of decision trees, calculated using TreeFarms (Xin et al., 2022). However, the UNIVERSE framework directly applies to any model class for which the Rashomon set can be computed or well-approximated. In this section, we consider the approximate bounds on variable importance across the random forest Rashomon set introduced by Laberge et al. (2023). This approach does not explicitly enumerate the Rashomon set, but allows us to estimate $\inf_{f \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\text{sup}})} \Phi_j(f, \mathcal{D}^{(n)})$ and $\sup_{f \in \hat{\mathcal{R}}^{(n)}(\varepsilon_n + \epsilon_{unobs} + \lambda_{\text{sup}})} \Phi_j(f, \mathcal{D}^{(n)})$, which are all that we require for UNIVERSE.

We repeat each of the semi-synthetic experiments from Section 4 using these bounds. Given a tabular dataset $\mathcal{D}^{(n)}$, we first split the dataset into K equally sized partitions $\mathcal{D}_1^{(n/K)}, \mathcal{D}_2^{(n/K)}, \dots \mathcal{D}_K^{(n/K)}$. We then fit a decision tree classifier on each partition sequentially, yielding K distinct models $f_1, f_2, \dots f_K$. For this set of experiments, we increase the complexity of the true data generation process by fitting full depth 6 decision trees. In order to ensure that these models are distinct, when fitting the k-th tree, we set all entries of each feature used by the previously fitted models $f_1, f_2, \dots f_{k-1}$ equal to 0. For each partition, we create a semi-synthetic dataset where each label is replaced by the prediction of the corresponding model, i.e., $\bar{\mathcal{D}}_k^{(n/K)} := \{(X_i, f_k(X_i))\}_{i=(k-1)n/K}^{kn/K}$. Finally, we combine the semi-synthetic datasets back into one. In doing so, we treat the partition to which each sample was assigned as an important unobserved feature $U \in \mathcal{U} := \{1, 2, \dots, K\}$; if the i-th sample was assigned to partition k, we say $u_i = k$. This yields a known true conditional mean function $g^*(x_i, u_i) := f_{u_i}(x_i), u_i \in \{1, 2, \dots, K\}$. This setup allows us to know each ground truth quantity of interest (g^* and S^* with their corresponding variable importance values, ϵ_{unobs}) while working with a fairly realistic distribution for X.

We again apply this procedure to four disparate real-world datasets with $K = 2$:

1. Compas (Larson et al., 2016), which concerns criminal recidivism prediction for 6,907 individuals;
2. Dropout (Martins et al., 2021), which concerns predicting whether students will drop out of college for 4,424 individuals;
3. FICO (FICO et al., 2018), which concerns predicting whether an individual will repay line of credit within 2 years over 10,459 individuals;
4. Wine Quality (Cortez et al., 2009), which concerns predicting whether a wine will be highly rated over 6,497 wines.

For each dataset, we evaluate our framework using random draws with replacement of sample sizes 100; 1,000; 10,000; 50,000; 250,000; and 500,000 from $\bar{\mathcal{D}}^{(n)}$, each over 100 iterations. For these experiments, we include every feature from each dataset. In this section, we consider the model class $\mathcal{F}$ of random forests, and restrict each f_k to have a depth of at most 6. To faithfully evaluate the theoretical claims of Section 3, we again use the true value for ϵ_{unobs} and τ_j in each experiment.

G.1 Finding each model in S^*

Corollary 1 demonstrates how we may select a value ε_n such that, with high probability, the empirical Rashomon set contains S^*. We first evaluate the validity of this Corollary with a target confidence of $1 - \delta = 0.9$.

Figure 7 presents the results of this evaluation for the random forest Rashomon set. We again find that, across all six sample sizes and all four reference datasets, omitting the finite sample adjustment yields Rashomon sets that leave out necessary models. In contrast, **our empirical Rashomon set based on Corollary 1 contains every model in S^* at above the specified rate**. These results demonstrate that adjusting for finite sample bias is necessary for constructing high-probability, finite-sample covers of the population Rashomon sets.

G.2 Recovering variable importance for each model in S^*

We now turn to evaluate Theorem 2 for the random forest Rashomon set. Given a finite sample bound for the estimation of our variable importance metric for a specified model, we can cover the importance of each

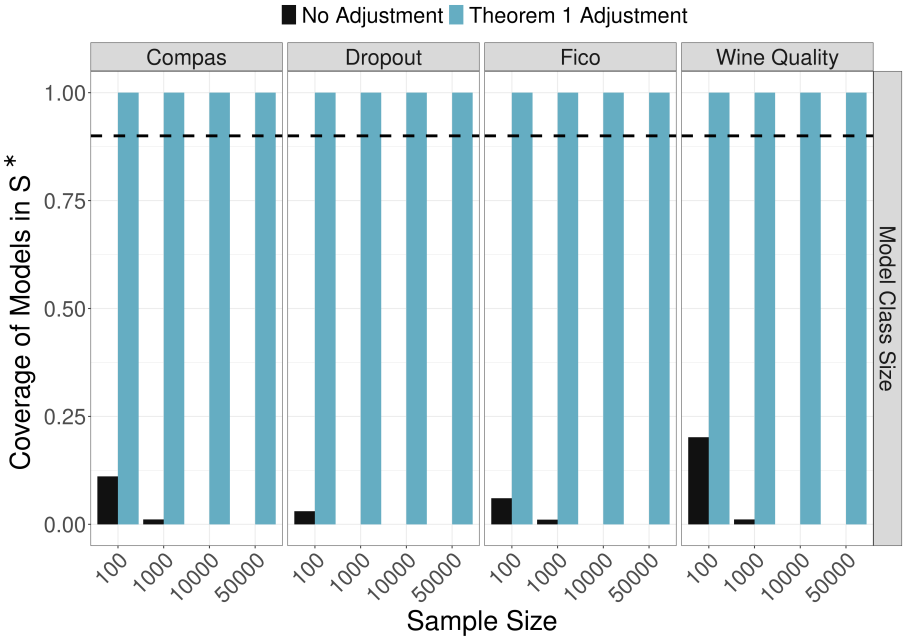

Figure 7: Verifying Theorem 1 in finite sample datasets using random forests. We compute the proportion of 100 random draws of the each dataset in which Rashomon sets estimated with the Rashomon threshold adjusting for finite sample biases as in Theorem 1 (in blue) and without any adjustment (in black) captures each f_u for each setting. The target coverage rate is ≥ 0.9, with $\delta = 0.1$. Across all sample sizes and datasets, omitting finite sample adjustments yields Rashomon sets that leave out necessary models. In contrast, our adjustment yields the target coverage rate, verifying the theorem holds.

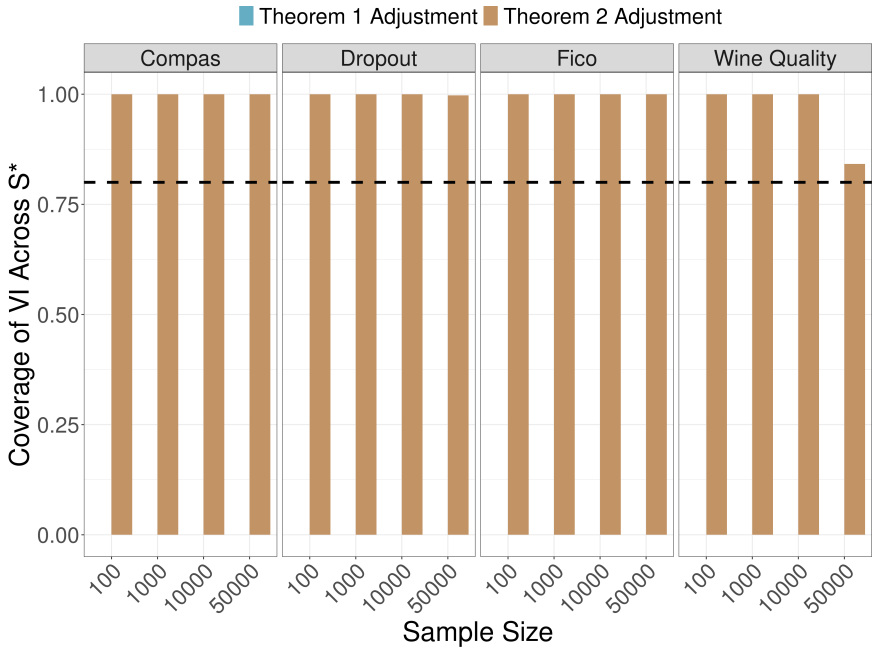

Figure 8: Verifying Theorem 2 using random forests. We achieve the specified coverage rate of ≥ 0.8 only when adjusting for (i) model uncertainty via Theorem 1 and (ii) variable importance estimation uncertainty (in gold). Adjusting for model uncertainty alone (in blue; here, always 0) is not sufficient. For each setting, we compute the proportion of 100 experiments where our variable importance bounds capture the true variable importance for all submodels $f_u \in S^*$, averaged over variables.

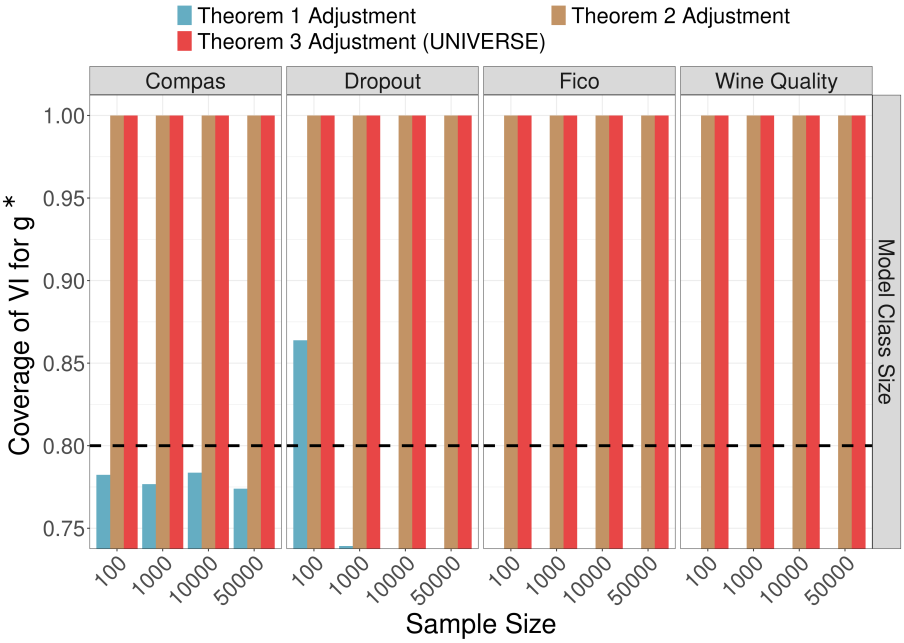

Figure 9: Verifying Theorem 3 using random forests. We consistently achieve the specified coverage rate of ≥ 0.8 when we account for (i) model uncertainty, (ii) variable importance uncertainty, and (iii) VI drift. Each bar measures the proportion of 100 experiments in which our bounds capture the true variable importance for the true model g^*. Plots are colored such that blue only accounts for finite sample model uncertainty as in Theorem 1, gold also adjusts for uncertainty in estimating subtractive model reliance (MR) at the model-level as in Theorem 2, and red adjusts for the previous two *and* distribution shifts induced by omitted variables 3. Applying al three adjustments yields the target coverage rate of ≥ 0.8, with $\delta = \gamma = 0.1$.

variable to each model in S^* with at least a specified probability. In this experiment, we set our target probability to $1 - (\delta + \gamma) = 0.8$, and apply the finite sample bound for subtractive model reliance (MR) from Equation B.26 of Fisher et al. (2019). Figure 8 demonstrates that adjusting only for model uncertainty as in Corollary 1 yields invalid bounds on model-level variable importance. In fact, omitting variable importance uncertainty quantification yields intervals that *never* contain the true variable importance for all conditional submodels. In contrast, **across all four datasets and at all sample sizes, our approach achieves nominal coverage.**

G.3 Recovering variable importance for the g^* with unobserved features

Finally, we evaluate Theorem 3 for the random forest Rashomon set, which provides high probability bounds on variable importance to the true model g^*, even with unobserved variables. We again set our target coverage rate to $1 - (\delta + \gamma) = 0.8$ and apply the finite sample bound for subtractive MR Fisher et al. (2019). This reflects applying the entire UNIVERSE framework.

Figure 9 demonstrates that **we consistently cover the true importance to g^* in more than the specified** 80% **of cases across all four datasets**, even though g^* depends on unobserved variables.

In contrast to our experiments with the Rashomon set of sparse decision trees in Figure 5, we find that only adjusting for uncertainty in variable importance (Theorem 2 Adjustment) tends to yield the nominal coverage rate. This is because the model class of random forests is much more expressive than that of sparse decision trees, which results in more varied Rashomon sets and ultimately looser bounds on variable importance. This makes sense when the chosen model class is viewed as a prior; if a practitioner believes a smaller, less expressive model class is sufficient to model capture the data generation process, this reflects a stronger prior than using a large, expressive model class, and results in tighter bounds.

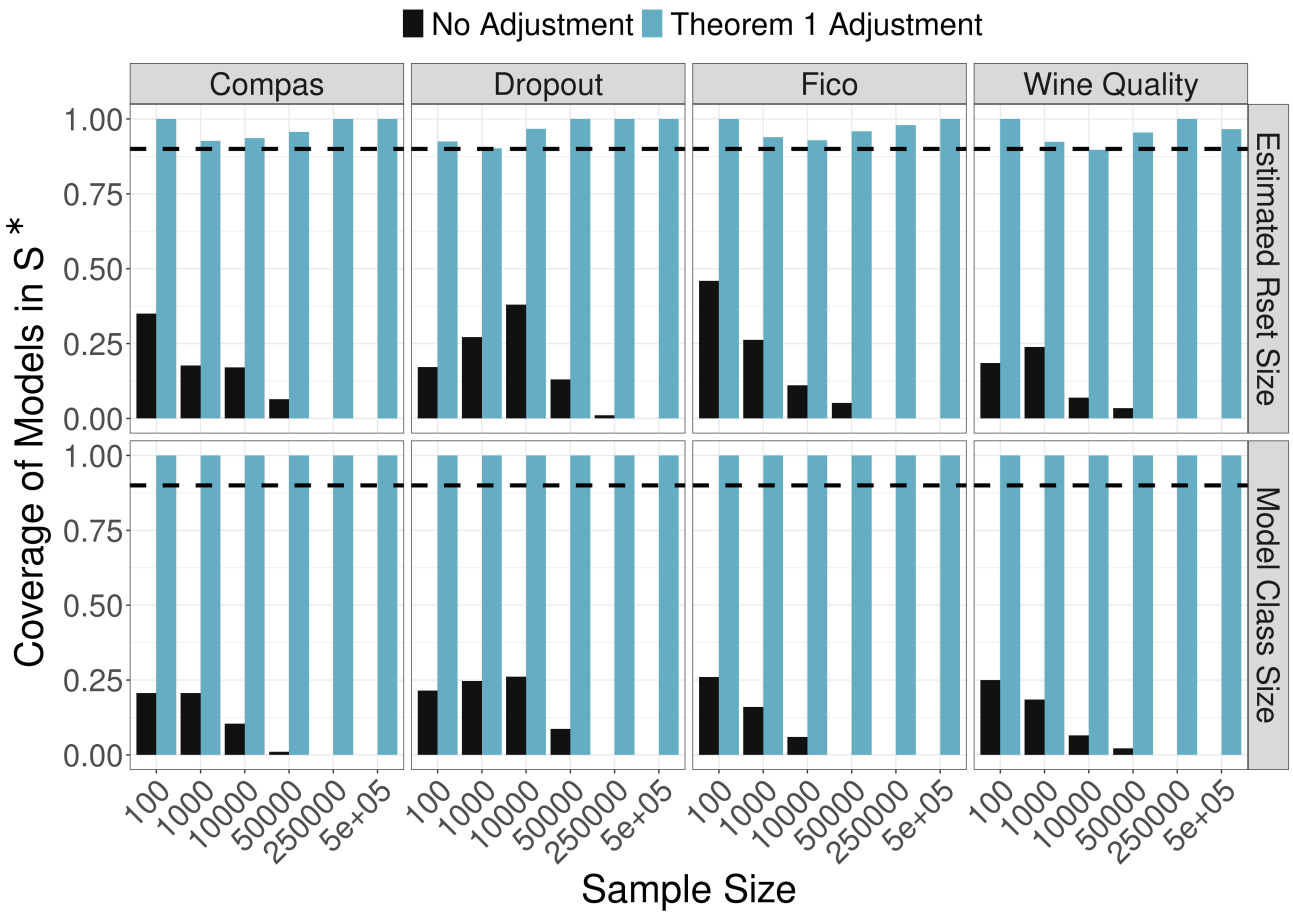

Figure 10: Verifying Theorem 1 in finite sample datasets. We compute the proportion of 100 random draws of the each dataset in which Rashomon sets estimated with the Rashomon threshold adjusting for finite sample biases as in Theorem 1 (in blue) and without any adjustment (in black) captures each f_u for each setting. The target coverage rate is ≥ 0.9, with $\delta = 0.1$. Across all sample sizes and datasets, omitting finite sample adjustments yields Rashomon sets that leave out necessary models. In contrast, our adjustment yields the target coverage rate, verifying the theorem holds. We use the estimated Rashomon set size (top row) and the model class size (bottom row) as our upper bounds on the size of S^*.

H EXPERIMENTS WITH MODEL CLASS SIZE

Section 4 evaluates each of our primary theoretical claims empirically by first estimating the size of the Rashomon set. Because we are using an *estimate* of the Rashomon set size, these bounds are not necessarily guaranteed to be finite-sample valid. For scientists worried about finite-sample validity, we present an alternative strategy where we use the size of the model class as our upper bound on the size of S^*.

Figure 10 displays the results verifying Theorem 1. The top row displays the main paper results again while the bottom row displays the new results; in the bottom row, not adjusting for finite sample or regularization behavior leads to overly conservative Rashomon sets with severe undercoverage of the models in S^*. In contrast, our approach with an adjusted Rashomon threshold guarantees coverage at all sample sizes. However, this conservative correction is overly conservative, leading to Rashomon sets that contains S^* at *all* sample sizes. In contrast, using an estimate of the Rashomon set size leads to Rashomon sets that are probabilistically more valid.

Figure 11 displays the results verifying Theorem 2. The top row displays the main paper results again while the bottom row displays the new results; in the bottom row, adjusting only for model uncertainty and not VI-estimation uncertainty yields bounds that undercover the variable importance for models in S^*. In contrast, our approach that accounts for VI estimation uncertainty yields bounds that achieve the specified error rate.

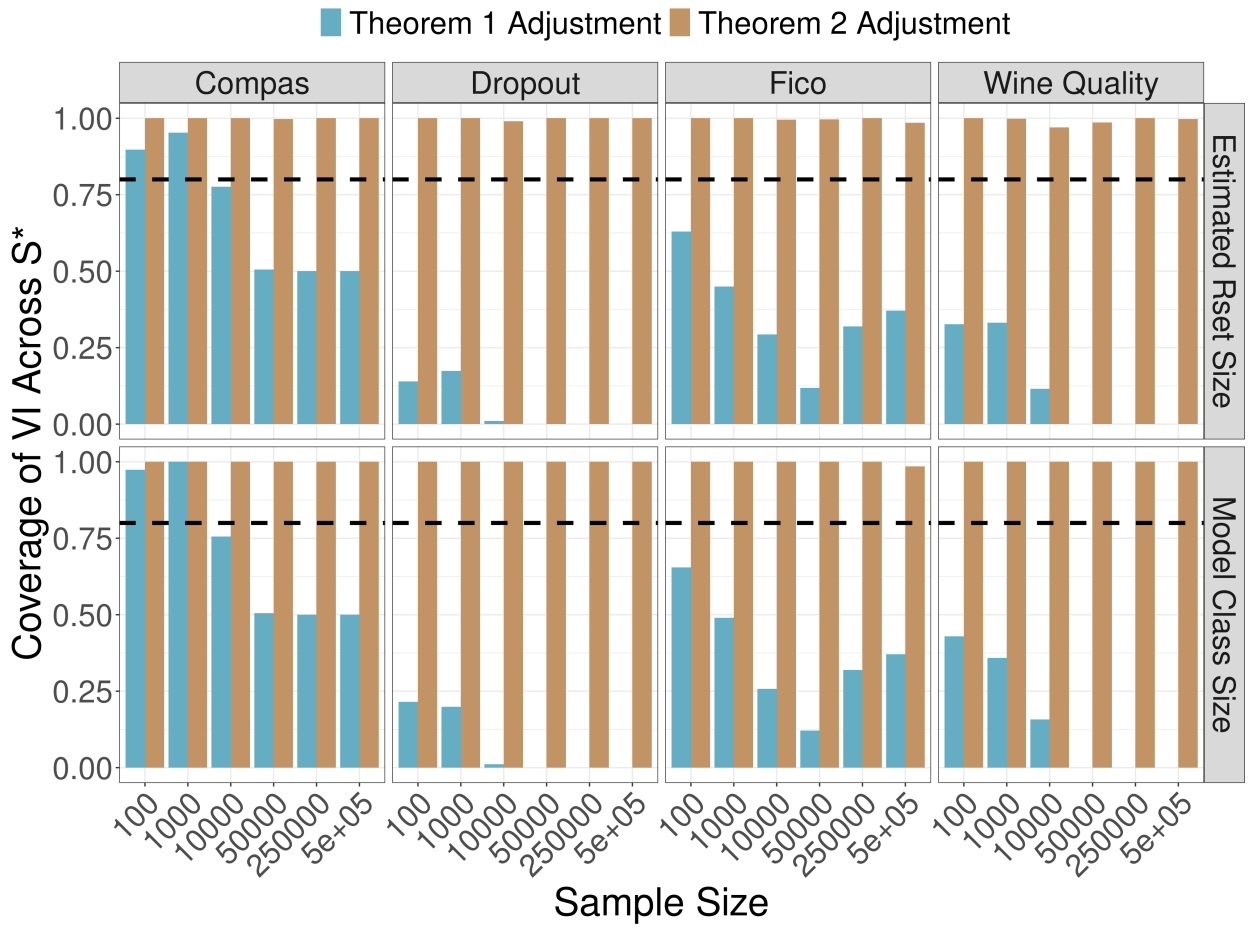

Figure 11: Verifying Theorem 2. We achieve the specified coverage rate of ≥ 0.8 only when adjusting for (i) model uncertainty via Theorem 1 and (ii) variable importance estimation uncertainty (in gold). Adjusting for model uncertainty alone (in blue) is not sufficient. For each setting, we compute the proportion of 100 experiments where our variable importance bounds capture the true variable importance for all submodels $f_u \in S^*$, averaged over variables. We use the estimated Rashomon set size (top row) and the model class size (bottom row) as our upper bounds on the size of S^*.

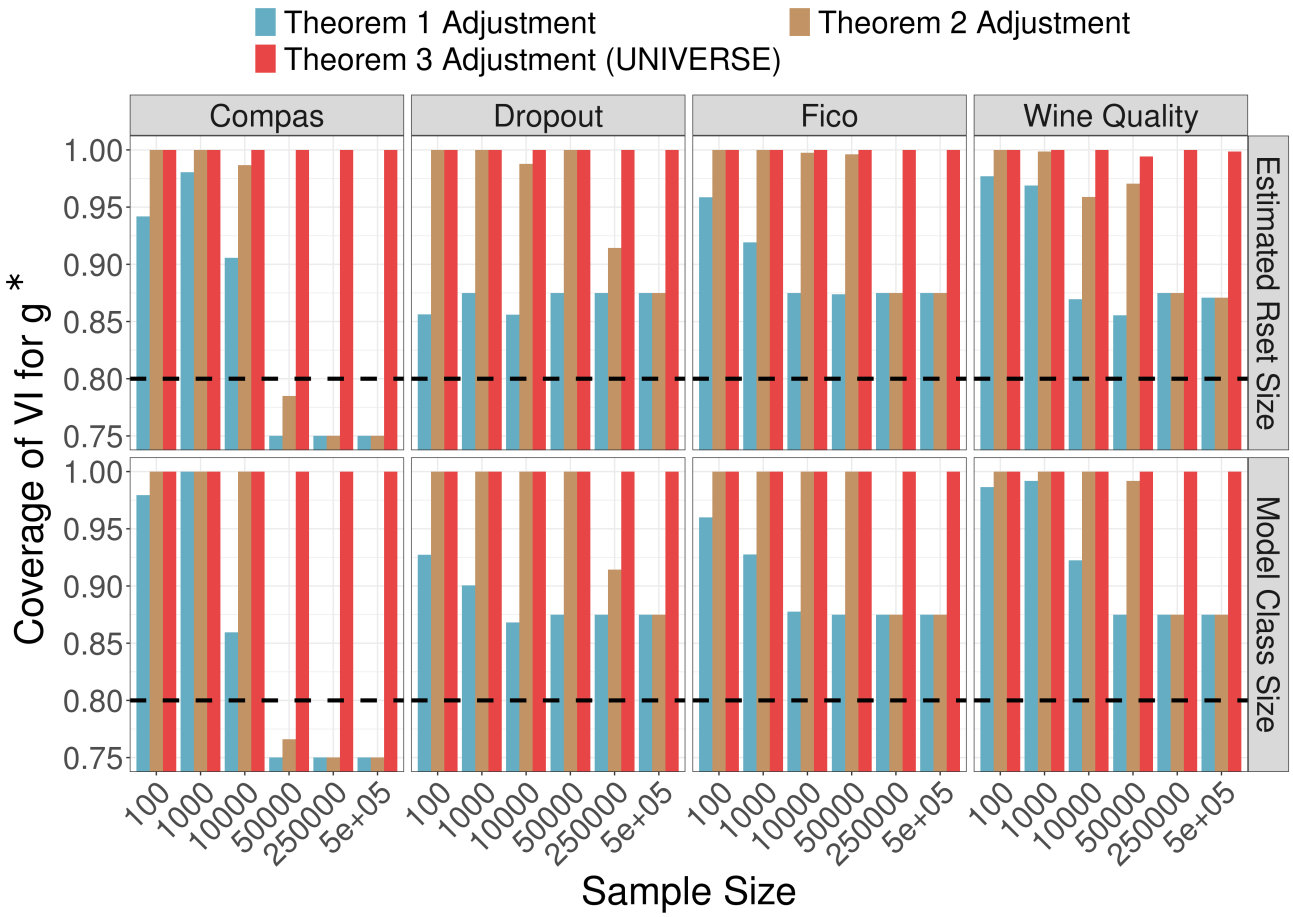

Figure 12: Verifying Theorem 3. We consistently achieve the specified coverage rate of ≥ 0.8 only when we account for (i) model uncertainty, (ii) variable importance uncertainty, and (iii) VI drift. Each bar measures the proportion of 100 experiments in which our bounds capture the true variable importance for the true model g^*. Plots are colored such that blue only accounts for finite sample model uncertainty as in Theorem 1, gold also adjusts for uncertainty in estimating subtractive model reliance (MR) at the model-level as in Theorem 2, and red adjusts for the previous two *and* distribution shifts induced by omitted variables 3. All three adjustments are necessary to achieve the target coverage rate of ≥ 0.8, with $\delta = \gamma = 0.1$. We use the estimated Rashomon set size (top row) and the model class size (bottom row) as our upper bounds on the size of S^*.

However, this conservative correction is overly conservative, leading to bounds that contain the true VI for all models in S^* always. In contrast, using an estimate of the Rashomon set size leads to bounds that are slightly less conservative; for example, the Wine Quality dataset at 10,000 and 50,000 samples creates bounds that have coverage $< 100\%$.

Finally, Figure 12 displays the results verifying Theorem 3. The top row displays the main paper results again while the bottom row displays the new results; in the bottom row, not adjusting for VI drift yields bounds that undercover the variable importance for the true model g^*. For example, the blue and gold bars that represent only performing adjustments from Theorem 1 and Theorem 2 respectively achieve a coverage of only 0.76 on the Compas dataset even at extremely large sample sizes of 500,000 even when using the conservative model class size adjustment (bottom row). In contrast, our approach remains valid. However, using the conservative model class size is overly conservative with bounds that contain the true VI at all sample sizes for all datasets. In contrast, using the estimated Rashomon set size yields slightly tighter intervals that can achieve $< 100\%$ coverage (e.g., Wine Quality dataset with sample size 50,000).

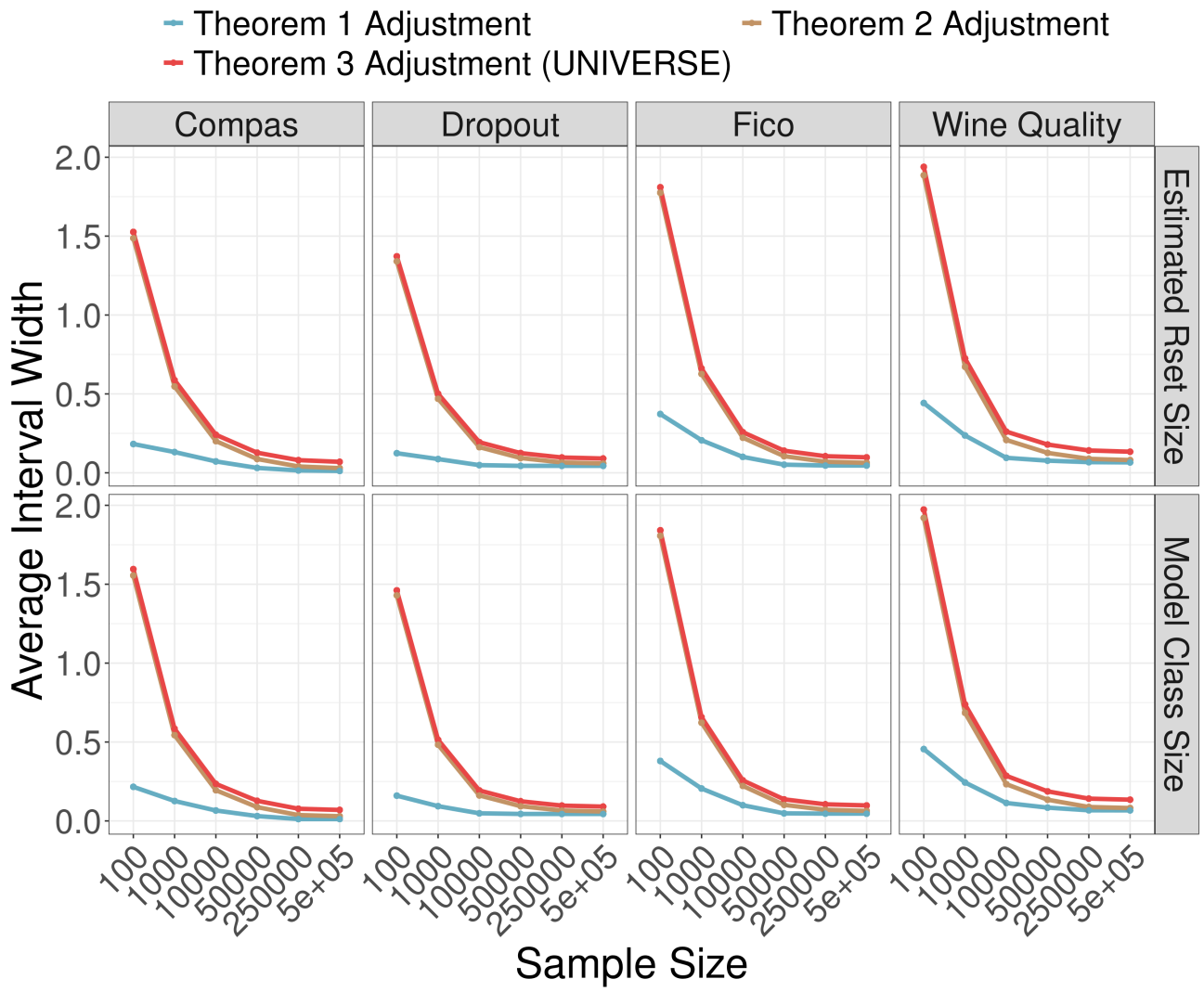

Figure 13: Evaluating the width of our intervals. We display the average interval width for each experiment conducted to evaluate coverage in Section 4. As the sample size increases, our intervals become much tighter– regardless of whether we estimate the Rashomon set size or use the model class size as our upper bound on the size of S^*.

I INTERVAL WIDTHS

In this section, we evaluate the width of our intervals. We display the average interval width for experiment conducted to evaluate coverage in Section 4. As the sample size increases, our intervals become much tighter– regardless of whether we estimate the Rashomon set size or use the model class size as our upper bound on the size of S^*. Importantly, each additional adjustment increases interval widths very slightly but contribute to coverage at the specified rate. Our adjustments are therefore a small cost to pay for improved Type-1 error control, especially at larger sample sizes.