Model Class Selection

Ryan Cecil and Lucas Mentch
Department of Statistics, University of Pittsburgh

November 17, 2025

Abstract

Classical model selection seeks to find a single model within a particular class that optimizes some pre-specified criteria, such as maximizing a likelihood or minimizing a risk. More recently, there has been an increased interest in model set selection (MSS), where the aim is to identify a (confidence) set of near-optimal models. Here, we generalize the MSS framework further by introducing the idea of model class selection (MCS). In MCS, multiple model collections are evaluated, and all collections that contain at least one optimal model are sought for identification. Under mild conditions, data splitting based approaches are shown to provide general solutions for MCS. As a direct consequence, for particular datasets we are able to investigate formally whether classes of simpler and more interpretable statistical models are able to perform on par with more complex black-box machine learning models. A variety of simulated and real-data experiments are provided.

Keywords: Model Selection, Rashomon Effect, Model Confidence Set, Universal Inference

1 Introduction

In statistical and machine learning (ML) studies, it is standard to use model evaluation metrics such as AIC, BIC, step-wise selection, holdout, or cross-validation to identify a single empirically optimal model (Hastie et al. 2009). These methods are used to compare the predictive performances of various modeling algorithms often associated with different model classes. For example, suppose a data scientist wishes to construct a model that can predict if a loan applicant will default in less than a year based on the available data at their company. Given the data, they know strategies for fitting various types of models such as an additive linear model, random forest, and neural network. They are likely unsure, however, of which type of model will perform best. A common approach is thus to simply identify the model type that minimizes some empirical metric (e.g. cross-validation error) and re-fit on the full dataset to obtain a final model for inference or prediction.

Oftentimes, many well-performing models exist that optimize the chosen model selection criteria. This concept, called the Rashomon effect, may lead researchers with different datasets to construct entirely different models (Breiman 2001). In recognition of this effect, procedures for model *set* selection (MSS) have been growing in popularity. Such MSS methods generally seek to identify an entire collection of similarly optimal models (i.e. an approximate *Rashomon set*) rather than relying on only the single one identified via classical model selection methods (e.g. see Fisher et al. (2019), Wasserman et al. (2020), Kissel & Mentch (2024)). Frequently, these sets of models are computed such that they are guaranteed to contain the optimal model(s) of a given type with a desired level of confidence under certain conditions and sample sizes. The size of the Rashomon set can be seen as providing a measure of the uncertainty of the model selection problem. In addition, the computed model set can be used to explore the data and identify important predictors (Fisher et al. 2019).

Recently, ML methods have achieved significant breakthroughs in a variety of fields, surpassing benchmark performances once thought to be unattainable. These advances have been driven in large part by the modern growth of computational resources, allowing researchers to train large scale models with potentially billions of parameters. Compute at scale, however, often comes with a cost. Unlike classical statistical models that are interpretable, have quantifiable uncertainty, and are able to be used to conduct valid inference, the theoretical underpinnings of ML approaches are not fully developed (Hunter & Holmes 2023). Recent work in mathematics, statistics, and computer science have made significant strides towards addressing these issues (e.g. Abdar et al. (2021), Berner et al. (2022), Mentch & Hooker (2016), Rudin et al. (2022)). However, there still exists a significant trade-off in interpretability and theoretical guarantees when one moves from simple, well-studied classes of models (e.g. linear models) to complex classes (e.g. random forests or neural networks) to improve predictive accuracy. This is often an issue in studies focused on inference or causality and settings where the model is employed in high-stakes environments such as healthcare or criminal justice (Wexler 2017, Varshney & Alemzadeh 2017, Rudin 2019).

On the other hand, it is also somewhat common for practitioners to rule out classical models *a priori* whenever the data is assumed to be too large or complex. Recent research, however, suggests that this may be frequently misguided as there are numerous real-world settings in which simpler, interpretable models can perform on par with ML alternatives (Rudin 2019, Semenova et al. 2022, 2023, Boner et al. 2024). When this is the case, researchers may naturally prefer the simpler class in most settings. On the other hand, if complex ML models significantly outperform simpler alternatives, then there may be a preference for performance over interpretability. To our knowledge, little theoretical attention seems to have been given to formally detecting when these trade-offs exist. To address this, in this work, we introduce model class selection (MCS) methods that are able to determine with

high probability when no model within a simple class can match the performance gains realized by more complex ML models.

We formally investigate this problem in Section 2 by proposing a general MCS hypothesis testing framework. In Section 3, we extend data splitting based tools to this general MCS setting. In Sections 4 and 5, we consider alternative approaches to MCS and examine the power of our proposed methodology in simulated and real-world settings before concluding with a discussion in Section 6.

2 Constructing Methods for MCS

Denote the data as $\mathcal{D}^{(n)} = \{Z_1, \ldots, Z_n\}$ where $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ are independent and identically distributed with support over the set $\mathcal{Z}$. Let $\mathcal{H} \subseteq \{h \mid h : \mathcal{Z} \to \mathcal{Y}\}$ denote a model class, defined as a set of measurable functions mapping from the data space $\mathcal{Z}$ to a prediction space $\mathcal{Y}$. Define $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ to be a measurable function that specifies the error of a model $h \in \mathcal{H}$ on some data point $z \in \mathcal{Z}$. The goal of classical model selection is to choose a model $h \in \mathcal{H}$ that minimizes the risk $R(h) = \mathbb{E}_{Z_0}[\ell(h, Z_0)]$, defined as the expected loss of h at an independent test point $Z_0 \sim P$.

To formally establish the goals of MCS, for $j = 1, ..., d$ where $d \geq 2$, let T_j denote a non-empty index set corresponding to models $h_t \in \mathcal{H}$ for any $t \in T_j$. Define $\mathcal{H}_j := \{h_t : t \in T_j\}$ for $j = 1, ..., d$ to be the model classes of interest. Denoting $[d] := \{1, ..., d\}$, our goal is to form a confidence set for

$$\Theta = \Theta\left(P\right) := \underset{j \in [d]}{\arg\min} \, \underset{t \in T_j}{\inf} \, R\left(h_t\right)$$

which is the set of all indices corresponding to model classes containing at least one model that achieves the optimal risk. In other words, given $r \in [d]$, if $r \in \Theta$, then the model class $\mathcal{H}_r$ contains a model that performs as well or better than any other model from any class.

On the other hand, if $r \notin \Theta$, then the model class $\mathcal{H}_r$ should not be considered if the goal is to choose a model with the best possible performance. Although the problem above may look complex in nature, we will show that the process of constructing valid confidence sets may be simplified by focusing on hypothesis tests for MSS.

2.1 Hypothesis Tests

For simplicity, we consider a dual hypothesis testing problem. Given $r \in [d]$, we evaluate the hypotheses

$$H_{0,r} : r \in \Theta \quad \text{vs} \quad H_{A,r} : r \notin \Theta.$$

Let $\psi_r : \{Z_1, \ldots, Z_n\} \to \{0, 1\}$ denote a test function that rejects the null hypothesis when $\psi_r = 1$. Assume $P \in \mathcal{P}$ where $\mathcal{P}$ is a class of distributions. A test ψ_r with appropriate control of the type I error satisfies

$$\left(\overset{(as)}{\underset{n \to \infty}{\lim \sup}} \right) \sup_{P \in \mathcal{P}_{0,r}} P\left(\psi_r = 1\right) \leq \alpha \tag{1}$$

where $\mathcal{P}_{0,r} \subseteq \mathcal{P}$ denotes the distributions under which $H_{0,r}$ is true. The paranthetical limit (as) in (1) distinguishes two versions of the condition. A finite-sample condition would require that $\sup_{P \in \mathcal{P}_{0,r}} P\left(\psi_r = 1\right) \leq \alpha$, while an asymptotic version would require that $\lim \sup_{n \to \infty} \sup_{P \in \mathcal{P}_{0,r}} P\left(\psi_r = 1\right) \leq \alpha$. The latter is a weaker condition that only ensures the type I error is controlled in large sample sizes.

Given tests $\psi_1, \ldots, \psi_d$, we construct the dual confidence set $\hat{\Theta} := \{j \in [d] : \psi_j = 0\}$, which denotes the indices that were not rejected by the corresponding tests. Our goal will then be to show that $\hat{\Theta}$ satisfies point-wise coverage of Θ i.e.

$$\left(\overset{(as)}{\underset{n \to \infty}{\lim \inf}} \right) \inf_{P \in \mathcal{P}} \inf_{r \in \Theta(P)} P\left(r \in \hat{\Theta}\right) \geq 1 - \alpha. \tag{2}$$

5

2.2 From MSS to MCS: General Strategies for MCS

Interestingly, the MCS problem is strongly related to problems posed in the MSS literature. To see this, let $\mathcal{F}(\mathcal{H}) = \{f \mid f : \bigcup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}\}$ denote a collection of measurable algorithms that map from the space of sample data $\bigcup_{m=1}^{\infty} \mathcal{Z}^m$ to a model in $\mathcal{H}$ and let $f_1 \in \mathcal{F}(\mathcal{H}_1), \ldots, f_d \in \mathcal{F}(\mathcal{H}_d)$ represent pre-chosen model fitting algorithms that map to each model class.

We require a risk metric for $f \in \mathcal{F}(\mathcal{H})$ similar to that of $R(h)$ for $h \in \mathcal{H}$. To account for risk estimation methods that utilize an algorithm to fit multiple models (e.g. cross-validation), let $k_n > 0$ be a sequence of positive integers specifying the number of models we will fit for each class. Define $\left\{\mathcal{D}_j^{(n)}\right\}_{j=1}^{k_n}$ to be a collection of k_n training sets for fitting models where $\mathcal{D}_j^{(n)} \subset \mathcal{D}^{(n)}$ for $j = 1, ..., k_n$. For simplicity, assume that each of these subsets have the same number of observations and set $n_{tr} = |\mathcal{D}_1^{(n)}| = \cdots = |\mathcal{D}_{k_n}^{(n)}|$. For any algorithm $f \in \mathcal{F}(\mathcal{H})$, define the risk of f to be $R_n(f) = \sum_{j=1}^{k_n} R\left(f\left(\mathcal{D}_j^{(n)}\right)\right)$. In works on cross-validation, the quantity $R_n(f)$ has been referred to as the multi-fold test error and characterizes the performance of the ensemble of models fit across the k_n datasets (Bayle et al. 2020, Austern & Zhou 2025). The multi-fold test error is similar to the out-of-sample error, $\mathbb{E}[R_n(f)]$, which measures the expected performance of $f \in \mathcal{F}(\mathcal{H})$ across all possible training sets of size n_{tr}. Although it is possible to estimate the out-of sample error using cross-validation based approaches (Bates et al. 2023), such estimation is typically complex and requires strong stability conditions on the chosen loss function and algorithms (Luo & Barber 2024).

Previous works on MSS have devised strategies for constructing confidence sets for $\arg\min_{k \in [d]} R_n(f_k)$, which corresponds to the indices of algorithms $f_1, \ldots, f_d$ that achieve the minimal multi-fold test error. Procedures immediately applicable to estimating $\arg\min_{k \in [d]} R_n(f_k)$ are typically based on cross-validation (Lei 2020, Kissel & Lei 2023). Other methods could be used in cases where the algorithms $f_1, \ldots, f_d$ are constant and

only map to a single model (e.g. Fisher et al. (2019), Wasserman et al. (2020), Takatsu & Kuchibhotla (2025), Kim & Ramdas (2025)).

Let $\phi : \{Z_1, \ldots, Z_n\} \times \mathcal{F}(\mathcal{H}) \to \{0, 1\}$ denote a test function that rejects the null hypothesis $H_{0,n,r}^{MSS} : r \in \Theta_{\epsilon,n}^{MSS}$ when $\phi(f_r) = 1$ where $\Theta_{\epsilon,n}^{MSS} = \left\{ k \in [d] : R_n(f_k) \leq \min_{j \in [d]} R_n(f_j) + \epsilon \right\}$ denotes the indices of algorithms that are near-optimal with margin of error $\epsilon > 0$ in regards to the multi-fold test error. Then, ϕ appropriately controls the MSS type I error for $r \in [d]$ if

$$\left(\overset{(as)}{\underset{n \to \infty}{\lim \sup}} \right) \sup_{P \in \mathcal{P}} P \left(\phi(f_r) = 1 \cap r \in \Theta_{\epsilon,n}^{MSS} \right) \leq \alpha. \tag{3}$$

It should be emphasized that Θ is fixed, while $\Theta_{\epsilon,n}^{MSS}$ is a random quantity that depends on the sampled training data. The term $\epsilon > 0$ can be viewed as a user chosen amount the error has to improve for there to exist a valid trade-off in performance between algorithms or model classes.

Throughout the remainder of this work, we make the distinction between the finite-sample and asymptotic versions of conditions (1), (2), and (3) when necessary. In cases where we do not specify the exact form of the condition, this means that the result holds in either case. In other words, it holds both when treating (1), (2), and (3) as all finite-sample or all asymptotic. The following result shows that if ϕ satisfies (3) for all $r \in [d]$, then we may use ϕ to construct a valid test for MCS. For ease of notation, let $\phi(h)$ correspond to $\phi(f_h)$ where $f_h : \bigcup_{m=1}^{\infty} \mathcal{Z}^m \to \{h\}$ is a constant algorithm for all $h \in \mathcal{H}$.

Proposition 1 (Uniform MCS Test). *Define the test $\psi_r = I \{\inf_{t \in T_r} \phi(h_t) = 1\}$ for $r \in [d]$. Assume for some $r \in [d]$ that f_r satisfies*

$$R_n(f_r) \leq \inf_{f \in \mathcal{F}(\mathcal{H}_r)} R_n(f) + \epsilon \tag{4}$$

and ϕ satisfies (3). Then, ψ_r satisfies (1). Moreover, if (3) and (4) are satisfied for all $r \in [d]$, then the confidence set $\hat{\Theta}$ generated by $\psi_1, \ldots, \psi_d$ satisfies (2).

In other words, Proposition 1 shows that the test $\psi_r = I\{\inf_{t \in T_r} \phi(h_t) = 1\}$ for $r \in [d]$ satisfies the point-wise coverage property (2) if there exist algorithms $f_1, \ldots, f_d$ satisfying (4) for all $r \in [d]$. Notably, the computation of $\psi_1, \ldots, \psi_d$ does not require identifying $f_1, \ldots, f_d$; all that is required is that such algorithms satisfying (3) and (4) exist. The test $\psi_r = I\{\inf_{t \in T_r} \phi(h_t) = 1\}$ is equivalent to rejecting $\mathcal{H}_r$ if the class does not appear to contain any well-performing models based on the MSS test ϕ. This approach is synonymous with recently proposed methods for universal hypothesis testing which is equivalent to the problem of MCS when $d = 2$, $\mathcal{H}_1 \subseteq \mathcal{H}_2$, and the primary goal is to test $H_{0,1}$ (Wasserman et al. 2020, Dey et al. 2025).

The advantage of applying ϕ to the entire model class is that both finite-sample and asymptotic coverage guarantees hold. The computation of such a procedure, however, may be computationally expensive. In large sample size scenarios, one would hope that we could utilize the data to select a set of nearly optimal models from each class and use that to conduct the comparison instead. Under suitable assumptions on $f_1, \ldots, f_d$, it turns out that this is a feasible approach to obtain an asymptotic result similar to Proposition 1.

Proposition 2 (Selective MCS Test)**.** *Define the test* $\psi_r = \phi(f_r)$ *for all* $r \in [d]$. *Assume for some* $r \in [d]$ *that* f_r *satisfies*

$$R_n(f_r) - \inf_{t \in T_r} R(h_t) - \epsilon \leq o_p(1) \tag{5}$$

and ϕ *satisfies (3). Then,* ψ_r *satisfies the asymptotic version of (1). Moreover, if (3) and (5) are satisfied for all* $r \in [d]$, *then the confidence set* $\hat{\Theta}$ *generated by* $\psi_1, \ldots, \psi_d$ *satisfies the asymptotic version of (2).*

Although condition (5) is a strong requirement, it allows us to accurately underestimate the optimal risk $\inf_{t \in T_r} R(h_t)$ in large samples. The condition can generally be satisfied in learnable problems when $f_r\left(\mathcal{D}^{(n)}\right)$ computes an approximate empirical risk minimizer

(ERM) across $\mathcal{H}_r$ i.e.

$$f_r\left(\mathcal{D}^{(n)}\right) \in \left\{h \in \mathcal{H}_r \mid n^{-1} \sum_{Z \in \mathcal{D}^{(n)}} \ell\left(h, Z\right) \le \arg\min_{h \in \mathcal{H}_r} n^{-1} \sum_{Z \in \mathcal{D}^{(n)}} \ell\left(h, Z\right) + o\left(n^{-\beta}\right)\right\}$$

where $\beta > 0$ controls the rate of convergence. A problem is often considered to be learnable if the empirical risks of models in the model class converge to their population risk uniformly (Vapnik 2000, Shalev-Shwartz et al. 2010). Such uniform convergence can be written as

$$\sup_{t \in T_r} \left| R\left(h_t\right) - \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_t, Z_i\right) \right| \xrightarrow{p} 0 \tag{6}$$

under the data generating process P. It has been shown that (6) is implied for a model class $\mathcal{H}_r$ whenever $\mathcal{H}_r$ is finite, has finite VC dimension, or has finite fat-shattering dimension (Vapnik 2000). Recent work has also shown that under sufficient stability assumptions, there exist strategies besides almost-ERMs satisfying (5) (Shalev-Shwartz et al. 2010).

3   Proposed Methods

In Section 2, we showed that if there exists a MSS test ϕ satisfying (3), then by Propositions 1 or 2, we may use ϕ to a construct a MCS test for the model class $\mathcal{H}_r$. As a result, in this section, we detail and extend current data splitting based approaches for MSS that generally satisfy these conditions under mild assumptions.

To motivate the proposed methodology, for all $r \in [d]$, let $s_r \in [d]\backslash r$ be any other class index besides r. Let $\mathcal{D}_{-j}^{(n)} = \mathcal{D}^{(n)}\backslash\mathcal{D}_j^{(n)}$ denote the data held out from training for each data split $j = 1, .., k_n$. In addition, set $n_{te} = n - n_{tr}$ to be the number of held out observations.

Observe that the hypothesis $H_{0,n,r}^{MSS}$ implies $R_n\left(f_r\right) \le \min_{j \in [d]} R_n\left(f_j\right) + \epsilon$. So, we would hope to use the data to estimate $R_n(f_r) - \min_{j \in [d]} R_n\left(f_j\right) - \epsilon$. However, we do not have access to $\min_{j \in [d]} R_n\left(f_j\right)$. To alleviate this issue, note that when $H_{0,n,r}^{MSS}$ is true, $R_n\left(f_r\right) \le R_n\left(f_{s_r}\right) + \epsilon$ and so we can instead test if $R_n(f_r) > R_n(f_{s_r}) + \epsilon$. We are able to estimate the expected risk difference $R_n(f_r) - R_n(f_{s_r})$, by the sample average $\bar{R}_n\left(f_r, f_{s_r}\right)$ where

9

$$\bar{R}_n\left(f_r, f_{s_r}\right) = \frac{1}{k_n n_{te}} \sum_{j=1}^{k_n} \sum_{Z \in \mathcal{D}_{-j}^{(n)}} \nabla_{n,j}\left(f_r, f_{s_r}, Z\right)$$

and

$$\nabla_{n,j}(f_r, f_{s_r}, Z) = \ell\left(f_r\left(\mathcal{D}_j^{(n)}\right), Z\right) - \ell\left(f_{s_r}\left(\mathcal{D}_j^{(n)}\right), Z\right) \text{ for } j = 1, ..., k_n.$$

3.1 Studentization

A collection of recently proposed approaches for MSS are based on sample splitting and self normalization strategies (Takatsu & Kuchibhotla 2025, Kim & Ramdas 2025). In this section, we show that such studentized procedures also allow for the construction of valid methods for MCS.

To showcase the approach, let

$$\bar{\sigma}_n\left(f_r, f_{s_r}\right) = \frac{1}{k_n} \sum_{j=1}^{k_n} \frac{1}{n_{te}-1} \sum_{Z \in \mathcal{D}_{-j}^{(n)}} \left\{ \nabla_{n,j}\left(f_r, f_{s_r}, Z\right) - \frac{1}{n_{te}} \sum_{Z \in \mathcal{D}_{-j}^{(n)}} \nabla_{n,j}\left(f_r, f_{s_r}, Z\right) \right\}^2$$

be an average of the sample variances of the differences in loss between the fitted models. The following result shows that when there is a single data split ($k_n = 1$), a corresponding test dependent on central limit theory satisfies (3) in large samples.

Theorem 1 (Validity of ϕ_{CLT} when $k_n = 1$). *Set $k_n = 1$. Suppose n_{te} is an increasing sequence such that $n_{te} \to \infty$. Let $\mu_n = R_n(f_r) - R_n(f_{s_r})$ and $\sigma_n^2 = Var_{Z_0}\left[\nabla_{n,1}(f_r, f_{s_r}, Z_0)\right]$. So long as the sequence $(\nabla_{n,1}(f_r, f_{s_r}, Z_0) - \mu_n)^2 / \sigma_n^2$ is uniformly integrable, the test*

$$\phi_{CLT}\left(f_r\right) = I\left\{ \bar{R}_n\left(f_r, f_{s_r}\right) > k_n^{-1/2} n_{te}^{-1/2} \bar{\sigma}_n\left(f_r, f_{s_r}\right) \left(\Phi^{-1}(1-\alpha) + \epsilon\right) \right\} \tag{7}$$

satisfies the asymptotic version of (3).

The uniform integrability condition ensures that the differences in losses are well-behaved and hold if $\sup_n \mathbb{E}\left[|\nabla_{n,1}\left(f_r, f_{s_r}, Z_0\right)/\sigma_n|^\beta\right] < \infty$ for some $\beta > 2$ (which occurs under bounded loss) and does not converge to a degenerate distribution (Bayle et al. 2020).

10

Computation of $\inf_{t \in T_r} \phi_{CLT}(h_t)$ is difficult since we have to account for the sample standard deviation estimate $\bar{\sigma}_n(f_r, f_{s_r})$. This makes Proposition 2 a more feasible solution to extending the test ϕ_{CLT} to MCS than Proposition 1. Recall that to utilize Proposition 2, we would require $R_n(f_r) - \inf_{t \in T_r} R(h_t) - \epsilon \leq o_p(1)$. In other words, the multi-fold test error of f_r must eventually be ϵ-close to the optimal risk of models in the class $\mathcal{H}_r$. As noted in Section 2.2, taking f_r to be an (almost)-ERM is one way to satisfy this condition. For most stable algorithms, the size of the training sets n_{tr} typically controls the rate of convergence of $R_n(f_r)$. If the size of the training sets is small relative to the size of the test sets, then the event $R_n(f_r) - \inf_{t \in T_r} R(h_t) - \epsilon \leq 0$ may only be likely to occur in extremely large samples when $\epsilon > 0$ is very small. See Section B.1 of the Appendix for an example of such a setting and a way to conservatively modify ϕ_{CLT} to fix this potential issue. In practice with ϕ_{CLT} in Section 5, we set the size of the test set to be smaller than or equal to the size of the training set and $\epsilon \approx 0$.

Finally, it is worth noting that there is reason to expect Theorem 1 could be established under more mild conditions than uniform integrability (see, e.g. recent work from Takatsu & Kuchibhotla (2025) and Kim & Ramdas (2025)). For simplicity, we utilized the basic definition of uniform integrability and our results were developed irrespective of previous central limit theorems related to this approach.

3.2   Cross-Validation

In Theorem 1, we focused on the case when only a single data split is utilized ($k_n = 1$). Relying on only a single data split, however, may result in instability. The typical solution to reduce such variability is to take $k_n > 1$ and average across the results of multiple splits in the computation of $\bar{R}_n(f_r, f_{s_r})$. Such an average involves a sum of k_n dependent statistics. Consequently, we cannot directly apply the classical central limit theorems to asymptotically

control the type I error.

Cross-validation is widely recognized as a standard approach to achieving potentially more stable error estimates by partitioning the data into k_n folds and treating each as a hold-out set (Bates et al. 2023). Without loss of generality, for the remainder of this section, we will assume that $k_n > 1$ evenly divides n and define the folds sequentially. Let the jth test fold for $j \in [k_n]$ be defined as $\mathcal{D}_{-j}^{(n)} = \left\{ Z_{(j-1)n_{te}+1}, \ldots, Z_{jn_{te}} \right\}$ where $n_{te} = \frac{n}{k_n}$. Then, $\bar{R}_n \left(f_r, f_{s_r} \right)$ may be interpreted as the cross-validation estimate of $R_n \left(f_r \right) - R_n \left(f_{s_r} \right)$.

A collection of recent works have shown that $\bar{R}_n \left(f_r, f_{s_r} \right)$ will be asymptotically normal as long as the model fitting algorithms, f_r and f_{s_r}, satisfy sufficient notions of loss stability (Austern & Zhou 2025, Bayle et al. 2020). In this work, we consider a weak form of loss stability which ensures that the standardized cross-validation estimate converges asymptotically to a standard normal centered around $R_n \left(f_r \right) - R_n \left(f_{s_r} \right)$ (Bayle et al. 2020). To define this notion of loss stability, let $\mathcal{D}^{(n)\backslash i}$ represent $\mathcal{D}^{(n)}$ but with Z_i replaced by Z_0' where $Z_0' \overset{\text{i.i.d.}}{\sim} P$. For any function $g : \bigcup_{m=1}^{\infty} \mathcal{Z}^m \times \mathcal{Z} \to \mathbb{R}$, the loss stability of g for m training observations is defined as

$$\gamma_m^{loss}(g) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} \left[\left(g' \left(\mathcal{D}^{(m)}, Z_0 \right) - g' \left(\mathcal{D}^{(m)\backslash i}, Z_0 \right) \right)^2 \right]$$

where

$$g' \left(\mathcal{D}^{(m)}, Z_0 \right) = g \left(\mathcal{D}^{(m)}, Z_0 \right) - \mathbb{E} \left[g \left(\mathcal{D}^{(m)}, Z_0 \right) \middle| \mathcal{D}^{(m)} \right].$$

Beyond stability of the model fitting algorithms, we will also require integrability conditions similar to those of Theorem 1, but for the average loss difference across all possible training folds of size $n_{tr} = n - n_{te}$. Let

$$\bar{\nabla}_n(z) = \mathbb{E}_{\mathcal{D}_1^{(n)}} \left[\nabla_{n,1} \left(f_r, f_{s_r}, z \right) \right]$$

represent such an average. Under the assumption of loss stability, the result of Theorem 1 still holds even when averaging across $k_n < n$ dependent statistics.

Theorem 2 (Validity when $k_n < n$)**.** *Assume $k_n < n$. Let $\bar{\mu}_n = \mathbb{E}\left[R_n\left(f_r\right) - R_n\left(f_{s_r}\right)\right]$,
$\sigma_n^2 = Var\left[\bar{\nabla}_n\left(Z_0\right)\right]$, and $g\left(\mathcal{D}^{(m)}, z\right) = \ell\left(f_r\left(\mathcal{D}^{(m)}\right), z\right) - \ell\left(f_{s_r}\left(\mathcal{D}^{(m)}\right), z\right)$. Assume the
sequence $\left(\bar{\nabla}_n(Z_0) - \bar{\mu}_n\right)^2 / \sigma_n^2$ is uniformly integrable and $\gamma_{n_{tr}}^{loss}(g) = o\left(\sigma_n^2/n\right)$. Then, the test
(7) satisfies the asymptotic version of (3).*

Various learning algorithms are capable of satisfying the loss stability condition with suitable
hyper-parameter choices (Bayle et al. 2020). A few notable examples include stochastic
gradient descent, nearest neighbor, and ensemble based methods (Hardt et al. 2016, Devroye
& Wagner 1979, Elisseeff et al. 2005). Different choices of standardization or stability
conditions have also been previously considered in works on cross-validation and could be a
potential avenue for different results (Austern & Zhou 2025, Bayle et al. 2020).

3.3   Universal Infernce

Thus far, we have considered studentization strategies that yield asymptotic tests for MCS.
In this section, we propose an alternative approach to creating tests that appropriately
control the type I error in finite samples. The methodology is based on a MSS test referred
to as universal inference (UI) (Wasserman et al. 2020). The UI method was constructed
as a finite-sample alternative to the asymptotic likelihood ratio test for regular statistical
models and was later generalized to arbitrary loss functions and model classes (Dey et al.
2025). In the next result, we show that universal inference is a valid strategy satisfying (3)
under a type of strong central condition.

Theorem 3 (Validity of UI)**.** *For some $r \in [d]$, assume that there exists $\bar{\omega} > 0$ such that*

$$\mathbb{E}_{Z_0} \exp\left[\omega\left\{\nabla_{n,1}\left(f_r, f_{s_r}, Z_0\right)\right\}\right] \leq 1 \text{ almost surely for all } \omega \in [0, \bar{\omega}). \tag{8}$$

Then, the UI test

$$\phi_{UI}\left(f_r\right) = I\left\{\bar{R}_n^{\exp}(f_r, f_{s_r}, \omega) - \epsilon > \alpha^{-1}\right\} \tag{9}$$

13

where

$$\bar{R}_n^{\exp}(f_r, f_{s_r}, \omega) = k_n^{-1} \sum_{j=1}^{k_n} \exp\left\{\omega \sum_{Z \in \mathcal{D}_{-j}^{(n)}} \nabla_{n,j}(f_r, f_{s_r}, Z)\right\},$$

satisfies the finite-sample version of (3) for any $\omega \in [0, \bar{\omega})$.

The proof of Theorem 3 follows in a similar fashion to that of (Dey et al. 2025, Lemma 2) which is related to the MCS case when $d = 2$ and $k_n = 1$. Condition (8) is a generalized form of the strong central condition used by Dey et al. (2025) as it reduces to their definition when f_r maps to the risk minimizer of $\mathcal{H}_r$ and f_{s_r} selects models uniformly at random from $\mathcal{H}_{s_r}$. The primary benefit of condition (8) is that it does not require assuming a strong central condition on all models in $\mathcal{H}_{s_r}$, only the subset that may be selected by f_{s_r}. The strong central condition has been shown to imply fast rates of convergence for learning algorithms (Erven et al. 2015). A specific case where the strong central condition holds with $\bar{\omega} = 1$ is when $\mathcal{H}$ is composed of regular statistical models and the loss ℓ is the negative log likelihood. There are numerous other examples where this condition can be shown to hold (see (Dey et al. 2025, Section 3) or (Erven et al. 2015, Section 2.2)). In situations where the correct choice of ω is unknown, it may be possible to conservatively estimate it using general posterior calibration based on the nonparametric bootstrap (Dey et al. 2025).

Combining Theorem 3 and Proposition 1 shows that if $\inf_{t \in T_r} \phi_{UI}(h_t)$ is feasible to compute, then the UI procedure can exactly control the type I error rate of MCS in finite samples. On the other hand, if such optimization is not practical, then so long as we choose $f_r \in \mathcal{F}(\mathcal{H}_r)$ well enough such that (5) is satisfied, then Proposition 2 shows that UI will remain valid for large samples.

4 Alternative Approaches

In this section, we provide a brief overview of alternative methods from the MSS literature that may be applied to the problem of MCS. These methods yield both finite sample and asymptotic valid tests for (3). In contrast to those reported in Section 3, these approaches typically require more stringent assumptions on the model classes and algorithms of interest or are more computationally expensive.

MCS for classes of regular statistical models: When the model classes $\mathcal{H}_1, \ldots, \mathcal{H}_d$ are composed of regular statistical models and the loss function corresponds to the negative log-likelihood, the asymptotic likelihood ratio test may be used to conduct MCS. A likelihood ratio based methodology has already been proposed to conduct valid MCS with collections of appropriately nested model classes (Zheng et al. 2019, Li et al. 2019). Another line of work proposes a similar method for model classes composed of linear mixed models (Jiang et al. 2008). Furthermore, when the classes are composed of gaussian linear models, the likelihood ratio test is equivalent to conducting an F-test (Bickel & Doksum 2015, Proposition 6.1.2) and provides exact coverage. Similar methodology relying on the F-test applicable to MCS for nested gaussian linear model classes has also been suggested (Ferrari & Yang 2015). See Section B.2 of the Appendix for a more detailed account of how to construct a valid MCS test based on the likelihood ratio test.

Discrete argmin inference on algorithms and models: There exist MSS cross-validation based methods that satisfy the asymptotic version of (3) (Lei 2020, Kissel & Lei 2023). In contrast to the proposed methods of Section 3, however, these approaches are more computationally expensive and require sub-weibull stability assumptions as they investigate the estimated joint distribution of the differences in loss between every pair-wise combination of the algorithms $f_1, \ldots, f_d$. Another group of methods from the field of discrete argmin inference satisfy the asymptotic version of (3) when the algorithms $f_1, \ldots, f_d$ map

to a fixed model (Hansen et al. 2011, Kim & Ramdas 2025, Zhang et al. 2025). With additional modification, these methods are able to be applied to MCS. A simple example is the studentization method of Section 3.1, as it relies on the same data splitting and self normalization strategies of Kim & Ramdas (2025).

Concentration inequalities: Universal inference falls under a general class of approaches based on concentration inequalities. Takatsu & Kuchibhotla (2025) showcase a valid construction of another applicable concentration inequality using the one-sided empirical Bernstein inequality. Various forms of other inequalities that do not utilize data splitting can be found in the Probably Approximately Correct (PAC) bound literature (Valiant 1984, Vapnik 2000, Mohri et al. 2018, Elisseeff et al. 2005, Shalev-Shwartz et al. 2010, Feldman & Vondrak 2018, Dziugaite & Roy 2017). Generally, these methods could be used to conduct finite-sample inference for MCS when combined with Proposition 1. These approaches, however, typically require stringent assumptions on the complexity of the model classes or algorithms of interest, or exponential moment inequalities similar to the strong central condition. Under weaker but similar conditions, concentration inequalities can also be used to construct methods satisfying the asymptotic version of (3). For example, procedures for constructing empirical rashomon sets can be used for MCS if there is a constraint on the complexity of the model classes (Fisher et al. 2019). Further discussion of these methods is provided in Appendix B.3.

5   Numerical Experiments

In this section, we examine the power of the proposed methodology for MCS in simulated and real data scenarios. To conduct MCS, the studentized test ϕ_{CLT} is combined with the testing strategy outlined in Proposition 2. Such an approach directly utilizes $\phi_{CLT}(f_r)$ to test $H_{0,r} : r \in \Theta$ for any $r \in [d]$. We use $\phi_{CLT}^{Holdout}$ and ϕ_{CLT}^{CV} to distinguish between the

holdout ($k_n = 1$) and cross-validation ($k_n < n$) versions of the studentized test discussed in Sections 3.1 and 3.2, respectively. In the holdout case, a 50%/50% split of the data is used to construct the train/test sets. In the cross-validation case, the number of folds is set to $k_n = 10$. Any performance improvement will be considered significant, so we set $\epsilon \approx 0$.

Where applicable, the test ϕ_{CLT} is compared to the UI approach detailed in Section 3.3. The notation $\phi_{UI-\omega_{or}}$ corresponds to an application of UI to MCS by combining ϕ_{UI} with the testing procedure of Proposition 1. In other words, $\phi_{UI-\omega_{or}}$ represents the classical approach to hypothesis testing with UI where we reject $\mathcal{H}_r$ if $\inf_{t \in T_r} \phi_{UI}(h_t) = 1$ (Wasserman et al. 2020). Theorem 3 and Proposition 1 show that as long as $\omega > 0$ is chosen such that condition (8) holds, then $\phi_{UI-\omega_{or}}$ allows for valid inference in finite samples. In the following experiments, there is not a clear choice for $\omega \in (0, \bar{\omega}]$. A nonparametric bootstrap procedure was recently built for estimating $\bar{\omega}$ (Dey et al. 2025). However, this approach is computationally expensive to replicate across many simulations, has been shown to generally under-estimate $\bar{\omega}$, and requires that an ERM must be able to be computed for $\mathcal{H}_r$ (Dey et al. 2025). So, instead, to examine the potential power of UI in the following simulations, we use an oracle ω_{or} that is estimated to appropriately control type I error. For testing $H_{0,r}$, in an approach similar to (Dey et al. 2025, Algorithm 2), ω_{or} was chosen by drawing samples of size n from the data generating process under $H_{0,r}$, computing $\phi_{UI}(f_r)$ on each sample for varying $\omega \in [0, 1]$, then choosing ω_{or} to be the largest value for ω such that the type I error across the simulations was below α.

5.1 Nonlinear Regression

A primary motivation for MCS is to be able to compare simpler classes of models to complex alternatives. First, we investigate if the proposed methodology is able to properly identify classes capable of modeling nonlinear effects in a nonlinear setting. Assume that

$Z_i = (Y_i, X_i^\top) \in \mathbb{R} \times \mathbb{R}^5$ where $X_{i,j} \overset{\text{i.i.d.}}{\sim} U[0,1]$ for $i = 1, .., n$, $j = 1, ..., 5$, and $U[0,1]$ denotes the uniform distribution on $[0,1]$. Let

$$Y_i = 10\gamma \sin(\pi X_{i,1} X_{i,2}) + 20\gamma (X_{i,3} - 0.05)^2 + 10 X_{i,4} + 5 X_{i,5} + \epsilon_i$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 10)$ and $\gamma \geq 0$. We refer to this data generating process as the MARS regression model based on previous works that make use of it (Friedman 1991, Mentch & Zhou 2020). The term $\gamma \geq 0$ controls the degree of non-linearity and interactive components in the process. When $\gamma = 0$, the MARS model is the same as a gaussian linear model with two features. We consider a class of additive linear models, $\mathcal{H}_1$, and a class of random forest regressors $\mathcal{H}_2$. Specifically, a model $h \in \mathcal{H}_1$ makes predictions of the form $h(Z_i) = \beta^\top X_i$ for some $\beta \in \mathbb{R}^5$. In contrast, a model $h \in \mathcal{H}_2$ represents an ensemble of decision trees, with predictions for Z_i given by the average output of those trees when applied to the features X_i. The algorithms f_1 and f_2 correspond to the OLS estimator of $\mathcal{H}_1$ and the sklearn `randomforestregressor` algorithm with the argument $n_{estimators} = 30$, respectively. Lastly, the loss function, ℓ, is taken as squared error loss: $\ell(h, Z_i) = (h(Z_i) - Y_i)^2$. The goal of the experiment is to test whether or not an additive linear model is capable of achieving the optimal risk (i.e. in the notation of Section 2 the goal is to test $H_{0,1} : 1 \in \Theta$). In the tests, we compare f_1 to f_{s_1} where $s_1 = 2$. When $\gamma = 0$, $H_{0,1}$ is true. On the other hand, when $\gamma > 0$, $1 \notin \Theta$ due to the nonlinear effects. For the $\phi_{UI-\omega_{or}}, \phi_{CLT}^{Holdout}$, and ϕ_{CLT}^{CV} methods, the results are displayed in Figure 1 for $0 \leq \gamma \leq 6$ and sample sizes $n = 200, 500,$ and $1,000$. All three methods are capable of detecting the nonlinear signal for large $\gamma > 0$. Notably, the ϕ_{CLT}^{CV} approach attain higher power levels than UI with ω_{or} for most $\gamma > 0$.

5.2 Feature Selection

It is also possible to use MCS to compare nested model classes representing different combinations of features. In contrast to the previous section, the simpler classes in the
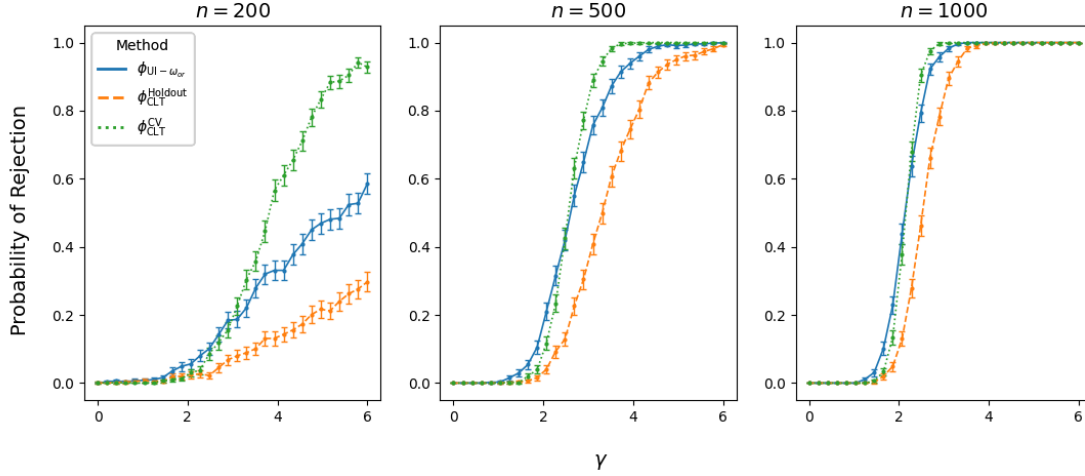
Figure 1: Rejection probabilities estimated across 1,000 simulations for various tests in the nonlinear regression setting with nominal level $\alpha = 0.05$. The solid, dashed, and dotted lines correspond to the tests $\phi_{UI-\omega_{or}}, \phi_{CLT}^{Holdout}$, and ϕ_{CLT}^{CV}, respectively. The bars corespond to estimated 95% confidence intervals for the power. We estimate the oracle choice of ω_{or} from the data generating process when $\gamma = 0$.

following experiment will be those that utilize a smaller number of the features. Consider the regression case where for $i = 1, ..., n$, $Z_i = (Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^6$, $X_i \overset{i.i.d.}{\sim} N(0_6, \Sigma)$, $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, and $Y_i = X_{i,1} + X_{i,3} + X_{i,5} + \epsilon_i$. The covariance matrix, $\Sigma \in \mathbb{R}^{6 \times 6}$, has entries $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.35$. In line with previous works, we choose the noise level σ^2 based on a corresponding signal to noise ratio level ν where $\sigma^2 = \frac{\beta^\top \Sigma \beta}{\nu}$ with $\beta^\top = (1, 0, 1, 0, 1, 0)$ (Hastie et al. 2020, Mentch & Zhou 2020). In this setting, there are $2^6 - 1 = 63$ combinations of covariates we could choose to include in a model. Let ℓ again correspond to the squared error loss. Let the model classes $\mathcal{H}_1, \ldots, \mathcal{H}_{63} \subset \mathcal{H}$ correspond to the 63 unique covariate inclusion combinations. Assume $\mathcal{H}_{63}$ corresponds to the class of models that includes all the features. Provided the simulated data and tests ϕ_r for $r \in [d]$, we construct the confidence set $\hat{\Theta} = \{r \in [62] : \phi_r = 0\}$. Let r^* correspond to the model class $\mathcal{H}_{r^*}$ that includes only the 1st, 3rd, and 5th covariates. In Figure 2, for $0.25 \leq \nu \leq 4$, we estimate the miss-coverage

19

probability of $r^* \notin \hat{\Theta}$, average size of $\hat{\Theta}$, and average uniform coverage rate of $\hat{\Theta} \supseteq \Theta$. We consider treating $\mathcal{H}_1, \ldots, \mathcal{H}_{63}$ as either classes of additive linear models or random forests with 30 estimated trees. For each model type and $r \in [d]$, let f_r correspond to fitting models in $\mathcal{H}_r$ using the sklearn `linearregression.fit` or `randomforestregressor.fit` algorithm, respectively. For each $r \in [d]$, we compare f_r to f_{s_r} where $s_r = 63$. In other words, we compare each nested class to the most complex class. In the case of the random forest model classes, only $\phi_{CLT}^{Holdout}$ and ϕ_{CLT}^{CV} are considered since $\inf_{t \in T_r} \phi_{UI}(h_t)$ is unable to be computed for any $r \in [d]$. The results in Figure 2 show that all procedures appear to attain similar size model class confidence sets $\hat{\Theta}$. For $\nu \leq 2$, the type I errors of $\phi_{CLT}^{Holdout}$ and ϕ_{CLT}^{CV} rise above the significance level in the random forest case. This is possible for both methods as the random forest(s) selected by f_r for $r \in [d]$ may not be near-optimal (i.e. lie in the Rashomon set) in small sample sizes.

5.3   MCS in the Presence of Noise

A recent line of works have shown that adding noise to the data generating process leads to implicit regularization in classification and regression problems (Semenova et al. 2022, 2023, Boner et al. 2024, Mentch & Zhou 2020). In other words, as more noise is added to the underlying process, these results suggest that simpler models will begin to perform on par with more complex alternatives. In MCS, such an effect would appear as a loss of power, since the optimal models in each class would perform as well as any other under sufficient additive noise. We experimentally investigate this claim using the $\phi_{CLT}^{Holdout}$ and ϕ_{CLT}^{CV} methods.

Similar to Section 5.1, consider the model classes $\mathcal{H}_1$ and $\mathcal{H}_2$ which represent a class of additive linear models and random forests, respectively. To investigate the effect of additive noise, a collection of real-world datasets from Section 5.4 where ϕ_{CLT}^{CV} rejects $H_{0,1} : 1 \in \Theta$
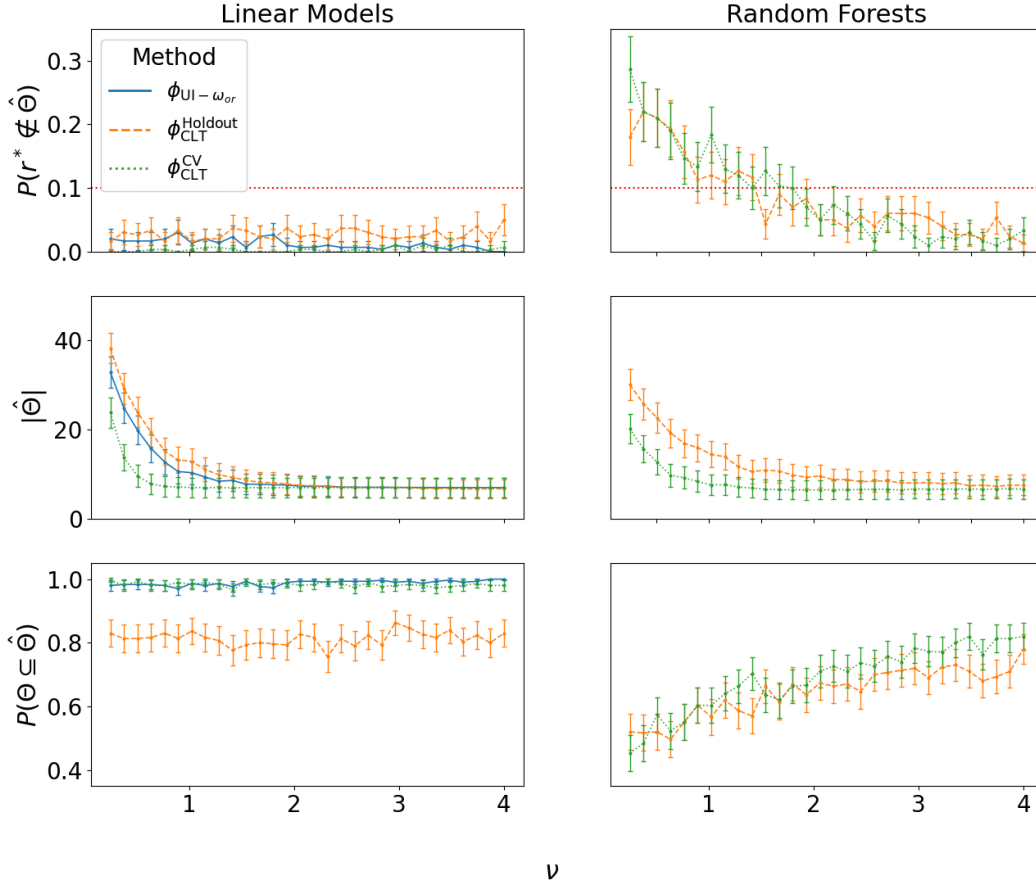
20

Figure 2: Feature selection performance of MCS with sample size $n = 200$ and $\alpha = 0.1$ estimated across 300 simulations. Each column corresponds to setting the classes $\mathcal{H}_1, \ldots, \mathcal{H}_{63}$ to different types of models. The first, second, and third row depicts the estimated miss-coverage probability of $r^* \notin \hat{\Theta}$, average size of $\hat{\Theta}$, and average uniform coverage rate of $\hat{\Theta} \supseteq \Theta$ for varying levels of ν. The red, dotted line represents the significance level. The bars correspond to estimated 95% confidence intervals. For each level ν, we estimate the oracle choice of ω_{or} for testing $H_{0,r^*} : r^* \in \Theta$.

are used. Two settings are considered: regression and classification. In the regression case, the algorithms f_1 and f_2 correspond to the OLS estimator of $\mathcal{H}_1$ and the sklearn `randomforestregressor` algorithm with the argument $n_{estimators} = 100$, respectively. In
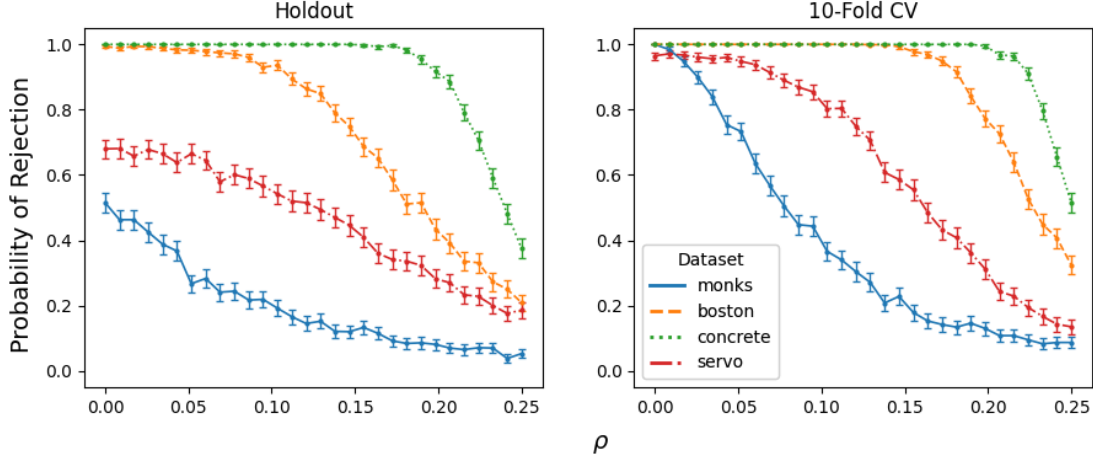
Figure 3: Effects of additive noise on ϕ_{CLT} at nominal level $\alpha = 0.1$ estimated across 1,000 simulations. The left and right plots correspond to the estimated probability that $\phi_{CLT}^{Holdout}(f_1) = 1$ and $\phi_{CLT}^{CV}(f_1) = 1$, respectively. The solid, dashed, dotted, and solid/dotted lines correspond to the datasets monks, boston, concrete, and servo, respectively, which are discussed further in Section 5.4.

the classification case, the same algorithms represent the sklearn `logisticregression` and `randomforestclassifier` algorithms with the argument $n_{estimators} = 1,000$. Given a dataset $\mathcal{D}^{(n)} = (Y_i, X_i)_{i=1}^n$, we generate a new dataset $\mathcal{D}_\rho^{(n)} = (Y_i^\rho, X_i)_{i=1}^n$ where for $i = 1, ..., n$, in the classification case, Y_i^ρ is flipped independently with probability ρ, and in the regression case, $Y_i^\rho = Y_i + 4\rho\epsilon_i$ where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \hat{\sigma})$ with $\hat{\sigma}$ being the sample variance of Y. Following similar experimental work, in the simulations that follow we consider $0 \leq \rho \leq .25$ (Semenova et al. 2023). In the regression setting, ρ is multiplied by 4 so that the noise level of each ϵ_i varies from 0 to $\hat{\sigma}$. In the classification and regression cases, we let ℓ correspond to the log loss and squared error loss, respectively. Figure 3 shows the results for the $\phi_{CLT}^{Holdout}$ and ϕ_{CLT}^{CV} methods. The results validate previous works as we do indeed see a substantial loss in power once the noise level is sufficiently high.

5.4 Real-World Data

In this section, ϕ_{CLT}^{CV} is applied to a diverse collection of real-world datasets. The goal is to test the same scenario as that of Section 5.3. Namely, if $H_{0,1} : 1 \in \Theta$ is true when $\mathcal{H}_1$ and $\mathcal{H}_2$ represent classes of additive linear models and random forests, respectively. Table 1 displays the results of the MCS test on each dataset.

Table 1 shows that when applied to real-world data, the ϕ_{CLT}^{CV} test produces a range of results on both regression and classification datasets. In Figure 5 of the Appendix, the cross-validation errors of models selected using f_1 and f_2 are compared. Notably, the corresponding statistic in Table 1 is large when the random forest cross-validation errors are on average much larger than that of the linear models. Table 1 and Figure 5 also demonstrate that these results do not appear to follow a standard pattern nor obviously correlate with dataset size or dimension.

6 Conclusion

This work introduces a general framework for constructing tests for MCS from simpler methods. The proposed framework incorporates current uniform testing approaches as well as more efficient selective versions. Tests based on studentization and UI are shown to satisfy the requirements of this framework under mild assumptions. Experimental results provide evidence that the proposed methods appropriately control the type I error rates in finite samples across a range of simulated MCS scenarios while also attaining notable degrees of power. Applications on real-world datasets demonstrate that the proposed framework is not only theoretically valid but also practically effective in distinguishing instances where simple models are similarly accurate from those in which more complex ML alternatives perform significantly better.

Dataset	Task	n	p	$\bar{R}_n(f_1, f_{s_1})$	Decision
Adult Income (Barry Becker 1996)	C	30162	96	-4.72	Fail
German Credit (Hofmann 1994)	C	1000	49	1.35	Reject
COMPAS (Larson et al. 2016)	C	6172	9	-7.87	Fail
Telco (Cognos Analytics 2018)	C	7043	8	-12.16	Fail
Monks (Wnek 1993)	C	124	12	4.71	Reject
Boston Housing (Harrison & Rubinfeld 1978)	R	506	13	6.06	Reject
Auto (Quinlan 1993b)	R	392	9	5.22	Reject
Bike Sharing (Fanaee-T & Gama 2014)	R	731	12	-0.17	Fail
Abalone Age (Waugh 1995)	R	4177	10	1.64	Reject
Concrete Strength (Yeh 1998)	R	1030	9	14.92	Reject
CPU (Ein-Dor & Feldmesser 1987)	R	209	36	14.00	Reject
CSM (Ahmed et al. 2015)	R	187	24	3.84	Reject
Facebook Metrics (Moro et al. 2016)	R	499	15	-3.65	Fail
Parkinsons (Athanasios Tsanas 2009)	R	5875	21	-20.96	Fail
Servo System (Quinlan 1993a)	R	167	11	3.43	Reject
Melting Point (Bergström et al. 2003)	R	274	1143	5.86	Reject

Table 1: Overview of results of ϕ_{CLT}^{CV} when applied to a collection real-world regression and classification datasets. In the task column, C and R stand for classification and regression, respectively. The columns n, p, and $\bar{R}_n(f_1, f_{s_1})$ represent the sample size, number of features, and test statistic.

It has become widely recognized that the application of models like random forests or neural networks often leads to increased power at the expense of interpretability. Until now, a rigorous method for detecting when such interpretability-performance trade-offs exist has not been proposed. This work introduces a methodology that enables users to formally identify these trade-offs, providing a more comprehensive understanding of the data. By adjusting the performance rejection threshold $\epsilon > 0$ of the proposed tests, users may also define how much the error has to significantly improve for there to exist a valid trade-off.

There are various MCS future directions to explore. While our focus was on point-wise coverage (2), the proposed methodology could be modified so that $\hat{\Theta}$ uniformly covers Θ with a desired level of confidence. See Section B.4 for more discussion on this topic. Additionally, we proposed sufficient conditions for constructing MCS tests based on MSS tests. It remains an open question whether milder or necessary conditions could be identified. Future research could also benefit from exploring and comparing alternative methodologies not examined in this study. Furthermore, conducting theoretical power analyses in simpler settings could provide valuable insights.

7 Acknowledgements

8 Data Availability and Reproducibility

Data sharing is not applicable to this article as no new data were created in this study. See https://github.com/ryanmcecil/model-class-selection to reproduce the results of Section 5.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V. & Nahavandi, S. (2021), 'A review of uncertainty quantification in deep learning: Techniques, applications and challenges', *Information Fusion* **76**, 243–297.

Ahmed, M., Jahangir, M., Afzal, H., Majeed, A. & Siddiqi, I. (2015), Using Crowd-Source Based Features from Social Media and Conventional Features to Predict the Movies Popularity, *in* '2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)', pp. 273–278.

Athanasios Tsanas, M. L. (2009), 'Parkinsons Telemonitoring'.

Austern, M. & Zhou, W. (2025), 'Asymptotics of cross-validation', *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **61**(4), 2804–2865. Publisher: Institut Henri Poincaré.

Barry Becker, R. K. (1996), 'Adult'.

Bates, S., Hastie, T. & Tibshirani, R. (2023), 'Cross-Validation: What Does It Estimate and How Well Does It Do It?', *Journal of the American Statistical Association* **0**(0), 1–12. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2023.2197686.

Bayle, P., Bayle, A., Janson, L. & Mackey, L. (2020), Cross-validation Confidence Intervals for Test Error, *in* 'Advances in Neural Information Processing Systems', Vol. 33, Curran Associates, Inc., pp. 16339–16350.

Bergström, C. A. S., Norinder, U., Luthman, K. & Artursson, P. (2003), 'Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs', *Journal of Chemical Information and Computer Sciences* **43**(4), 1177–1185. Publisher: American Chemical Society.

Berner, J., Grohs, P., Kutyniok, G. & Petersen, P. (2022), The Modern Mathematics of Deep Learning, *in* G. Kutyniok & P. Grohs, eds, 'Mathematical Aspects of Deep Learning', Cambridge University Press, Cambridge, pp. 1–111.

Bickel, P. J. & Doksum, K. A. (2015), *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*, 0 edn, Chapman and Hall/CRC.

Boner, Z., Chen, H., Semenova, L., Parr, R. & Rudin, C. (2024), 'Using Noise to Infer Aspects of Simplicity Without Learning', *Advances in Neural Information Processing Systems* **37**, 131824–131858.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.

Breiman, L. (2001), 'Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)', *Statistical Science* **16**(3), 199–231. Publisher: Institute of Mathematical Statistics.

Chafaï, D., Guédon, O., Lecué, G. & Pajor, A. (2012), *Interactions between compressed sensing random matrices and high dimensional geometry*, number no. 37 *in* 'Panoramas et synthèses', Société Mathématique de France, Paris. OCLC: ocn861119770.

Cognos Analytics, I. (2018), 'Telco Customer Churn'.

Devroye, L. & Wagner, T. (1979), 'Distribution-free inequalities for the deleted and holdout error estimates', *IEEE Transactions on Information Theory* **25**(2), 202–207. Conference Name: IEEE Transactions on Information Theory.

Dey, N., Martin, R. & Williams, J. P. (2025), 'Generalized universal inference on risk minimizers', *Journal of the Royal Statistical Society Series B: Statistical Methodology* p. qkaf065.

Dey, N., Martin, R. & Williams, J. P. (2026), 'Multiple testing in generalized universal inference', *Statistics & Probability Letters* **228**, 110559.

Dziugaite, G. K. & Roy, D. M. (2017), Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data., *in* 'Proceedings of the Conference on Uncertainty in Artificial Intelligence'.

Ein-Dor, P. & Feldmesser, J. (1987), 'Attributes of the performance of central processing units: a relative performance prediction model', *Commun. ACM* **30**(4), 308–317.

Elisseeff, A., Evgeniou, T. & Pontil, M. (2005), 'Stability of Randomized Learning Algorithms', *Journal of Machine Learning Research* **6**(3), 55–79.

Erven, T. v., Grünwald, P. D., Mehta, N. A., Reid, M. D. & Williamson, R. C. (2015), 'Fast Rates in Statistical and Online Learning', *Journal of Machine Learning Research* **16**(54), 1793–1861.

Fanaee-T, H. & Gama, J. (2014), 'Event labeling combining ensemble detectors and background knowledge', *Progress in Artificial Intelligence* **2**(2), 113–127.

Feldman, V. & Vondrak, J. (2018), Generalization Bounds for Uniformly Stable Algorithms, *in* 'Advances in Neural Information Processing Systems', Vol. 31, Curran Associates, Inc.

Ferrari, D. & Yang, Y. (2015), 'Confidence sets for model selection by F -testing', *Statistica Sinica* .

Fisher, A., Rudin, C. & 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.

Friedman, J. H. (1991), 'Multivariate Adaptive Regression Splines', *The Annals of Statistics* **19**(1), 1–67. Publisher: Institute of Mathematical Statistics.

Hansen, P. R., Lunde, A. & Nason, J. M. (2011), 'The Model Confidence Set', *Econometrica* **79**(2), 453–497. Publisher: [Wiley, Econometric Society].

Hara, S. & Ishihata, M. (2018), 'Approximate and Exact Enumeration of Rule Models', *Proceedings of the AAAI Conference on Artificial Intelligence* **32**(1). Number: 1.

Hardt, M., Recht, B. & Singer, Y. (2016), Train faster, generalize better: stability of stochastic gradient descent, *in* 'Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48', ICML'16, JMLR.org, New York, NY, USA, pp. 1225–1234.

Harrison, D. & Rubinfeld, D. L. (1978), 'Hedonic housing prices and the demand for clean air', *Journal of Environmental Economics and Management* **5**(1), 81–102.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York, New York, NY.

Hastie, T., Tibshirani, R. & Tibshirani, R. (2020), 'Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons', *Statistical Science* **35**(4), 579–592. Publisher: Institute of Mathematical Statistics.

Hofmann, H. (1994), 'Statlog (German Credit Data)'.

Hunter, D. J. & Holmes, C. (2023), 'Where Medical Statistics Meets Artificial Intelligence', *New England Journal of Medicine* **389**(13), 1211–1219.

Jiang, J., Rao, J. S., Gu, Z. & Nguyen, T. (2008), 'Fence methods for mixed model selection', *The Annals of Statistics* **36**(4), 1669–1692. Publisher: Institute of Mathematical Statistics.

Kim, I. & Ramdas, A. (2025), 'Locally minimax optimal and dimension-agnostic discrete argmin inference'. arXiv:2503.21639 [math] version: 1.

Kissel, N. & Lei, J. (2023), 'Black-Box Model Confidence Sets Using Cross-Validation with High-Dimensional Gaussian Comparison'. arXiv:2211.04958 [math, stat].

Kissel, N. & Mentch, L. (2024), 'Forward stability and model path selection', *Statistics and Computing* **34**(2), 82.

Laberge, G., Pequignot, Y., Mathieu, A., Khomh, F. & 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., Surya, M., Kirchner, L. & Angwin, J. (2016), 'How We Analyzed the COMPAS Recidivism Algorithm'.

Lehmann, E. L. & Romano, J. P. (2010), *Testing statistical hypotheses*, Springer texts in statistics, 3., rd ed. 2005. corr. 2nd printing. softcover version of original hardcover edition 2005 edn, Springer New York, New York, NY.

Lei, J. (2020), 'Cross-Validation With Confidence', *Journal of the American Statistical Association* **115**(532), 1978–1997. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2019.1672556.

Li, Y., Luo, Y., Ferrari, D., Hu, X. & Qin, Y. (2019), 'Model confidence bounds for variable selection', *Biometrics* **75**(2), 392–403. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13024.

Luo, Y. & Barber, R. F. (2024), 'The Limits of Assumption-free Tests for Algorithm Performance'. arXiv:2402.07388 [cs, math, stat].

Mata, K., Kanamori, K. & Arimura, H. (2022), 'Computing the Collection of Good Models for Rule Lists'. arXiv:2204.11285 [cs].

Mentch, L. & Hooker, G. (2016), 'Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests', *Journal of Machine Learning Research* **17**(26), 1–41.

Mentch, L. & Zhou, S. (2020), 'Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success', *Journal of Machine Learning Research* **21**(171), 1–36.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018), *Foundations of machine learning*, Adaptive computation and machine learning series, second edition edn, MIT Press, Cambridge, MA.

Moro, S., Rita, P. & Vala, B. (2016), 'Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach', *Journal of Business Research* **69**(9), 3341–3351.

Neyman, J., Pearson, E. S. & Pearson, K. (1997), 'IX. On the problem of the most efficient tests of statistical hypotheses', *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**(694-706), 289–337. Publisher: Royal Society.

Quinlan, J. R. (1993*a*), Combining instance-based and model-based learning, *in* 'Proceedings of the Tenth International Conference on International Conference on Machine Learning', ICML'93, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 236–243.

Quinlan, R. (1993*b*), 'Auto MPG'.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215. Publisher: Nature Publishing Group.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. & Zhong, C. (2022), 'Interpretable machine learning: Fundamental principles and 10 grand challenges', *Statistics Surveys*

16(none), 1–85. Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada.

Semenova, L., Chen, H., Parr, R. & Rudin, C. (2023), 'A Path to Simpler Models Starts With Noise', *Advances in Neural Information Processing Systems* **36**, 3362–3401.

Semenova, L., Rudin, C. & Parr, R. (2022), On the Existence of Simpler Machine Learning Models, *in* 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', FAccT '22, Association for Computing Machinery, New York, NY, USA, pp. 1827–1858.

Shalev-Shwartz, S., Shamir, O., Srebro, N. & Sridharan, K. (2010), 'Learnability, Stability and Uniform Convergence', *Journal of Machine Learning Research* **11**(90), 2635–2670.

Takatsu, K. & Kuchibhotla, A. K. (2025), 'Bridging Root-n and Non-standard Asymptotics: Adaptive Inference in M-Estimation'. arXiv:2501.07772 [math].

Valiant, L. G. (1984), 'A theory of the learnable', *Communications of the ACM* **27**(11), 1134–1142.

Vapnik, V. N. (2000), *The Nature of Statistical Learning Theory*, Springer New York, New York, NY.

Varshney, K. R. & Alemzadeh, H. (2017), 'On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products', *Big Data* **5**(3), 246–255. Publisher: Mary Ann Liebert, Inc., publishers.

Wasserman, L., Ramdas, A. & Balakrishnan, S. (2020), 'Universal inference', *Proceedings of the National Academy of Sciences* **117**(29), 16880–16890. Publisher: Proceedings of the National Academy of Sciences.

Waugh, S. G. (1995), Extending and benchmarking Cascade-Correlation : extensions to the

Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks, thesis, University of Tasmania.

Wexler, R. (2017), 'Opinion | When a Computer Program Keeps You in Jail', *The New York Times* .

Wilks, S. S. (1938), 'The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses', *The Annals of Mathematical Statistics* **9**(1), 60–62. Publisher: Institute of Mathematical Statistics.

Wnek, J. (1993), 'MONK's Problems'.

Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M. & Rudin, C. (2022), 'Exploring the Whole Rashomon Set of Sparse Decision Trees', *Advances in Neural Information Processing Systems* **35**, 14071–14084.

Yeh, I. C. (1998), 'Modeling of strength of high-performance concrete using artificial neural networks', *Cement and Concrete Research* **28**(12), 1797–1808.

Zhang, T., Lee, H. & Lei, J. (2025), 'Winners with Confidence: Discrete Argmin Inference with an Application to Model Selection'. arXiv:2408.02060 [math].

Zheng, C., Ferrari, D. & Yang, Y. (2019), 'Model Selection Confidence Sets by Likelihood Ratio Testing', *Statistica Sinica* **29**(2), 827–851. Publisher: Institute of Statistical Science, Academia Sinica.

Zhong, C., Chen, Z., Liu, J., Seltzer, M. & Rudin, C. (2023), 'Exploring and Interacting with the Set of Good Sparse Generalized Additive Models', *Advances in Neural Information Processing Systems* **36**, 56673–56699.

A   Proofs of Main Results

A.1   Proof of Proposition 1

Proof. We provide the proof steps without considering the parenthetical limit (as) because whether (as) is included or not, the steps are identical. First, assume for some $r \in [d]$ that (3) and (4) are satisfied. Then, by definition

$$\sup_{P \in \mathcal{P}_{0,r}} P \{\psi_r = 1\} = \sup_{P \in \mathcal{P}_{0,r}} P \left\{ \inf_{t \in T_r} \phi(h_t) = 1 \right\} \leq \sup_{P \in \mathcal{P}_{0,r}} P \{\phi(f_r) = 1\}.$$

Assuming $P \in \mathcal{P}_{0,r}$, by the assumption of (4), $R_n(f_r) \leq \arg \min_{k \in [d]} \inf_{t \in T_k} R(h_t) + \epsilon$ which implies $r \in \Theta_{\epsilon,n}^{MSS}$. Thus,

$$\sup_{P \in \mathcal{P}_{0,r}} P \{\phi(f_r) = 1\} = \sup_{P \in \mathcal{P}_{0,r}} P \left\{ \phi(f_r) = 1 \cap r \in \Theta_{n,\epsilon}^{MSS} \right\} \leq \alpha$$

which completes the first result. Next, assume for all $r \in [d]$ that (4) and (3) are satisfied. By the previous result,

$$\inf_{P \in \mathcal{P}} \inf_{r \in \Theta(P)} P \left\{ r \in \hat{\Theta} \right\} \geq 1 - \sup_{P \in \mathcal{P}} \sup_{r \in \Theta(P)} P \{\psi_r = 1\} \geq 1 - \alpha.$$

$\square$

A.2   Proof of Proposition 2

Proof. First, assume for some $r \in [d]$ that (3) and (5) are satisfied. Let $E_{n,r}$ be the event that $R_n(f_r) \leq \inf_{t \in T_r} R(h_t) + \epsilon$. Then, by definition

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{0,r}} P \{\psi_r = 1\} = \limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{0,r}} P \{\phi(f_r) = 1\}$$

$$= \limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{0,r}} \left[P \{\phi(f_r) = 1 \cap E_{n,r}\} + P \left\{ \phi(f_r) = 1 \cap E_{n,r}^C \right\} \right].$$

By (5), $P \left\{ E_{n,r}^C \right\} \to 0$ as $n \to \infty$. This implies that $\sup_{P \in \mathcal{P}_{0,r}} P \left\{ \phi(f_r) = 1 \cap E_{n,r}^C \right\} \to 0$ as $n \to \infty$. Furthermore, assuming $P \in \mathcal{P}_{0,r}$ and $E_{n,r}$ is true, then $R_n(f_r) \leq$

34

$\arg\min_{k\in[d]} \inf_{t\in T_k} R\left(h_t\right) + \epsilon$. Thus, in such a case, $r \in \Theta_{n,\epsilon}^{MSS}$. This implies that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{0,r}} P\left\{\phi\left(f_r\right) = 1 \cap E_{n,r}\right\} = \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{0,r}} P\left\{\phi\left(f_r\right) = 1 \cap E_{n,r} \cap r \in \Theta_{n,\epsilon}^{MSS}\right\}$$

$$\leq \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{0,r}} P\left\{\phi\left(f_r\right) = 1 \cap r \in \Theta_{n,\epsilon}^{MSS}\right\}$$

$$\leq \alpha$$

which completes the first result. The point-wise coverage result follows in the same manner as described in the Proof of Proposition 1.

$\square$

A.3 Proof of Theorem 1

Proof. Assume that $\left(\nabla_{n,1}(f_r, f_{s_r}, Z_0) - \mu_n\right)^2 / \sigma_n^2$ is uniformly integrable. Without loss of generality, assume that $\mathcal{D}_{-1}^{(n)} = \{Z_1, \ldots, Z_{n_{te}}\}$. For ease of notation, let $X_{n,i} = \nabla_{n,1}\left(f_r, f_{s_r}, Z_i\right)$. Then, note that since $k_n = 1$,

$$\frac{(n_{te} - 1)\bar{\sigma}_n^2\left(f_r, f_{s_r}\right)}{n_{te}} = \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} \left[(X_{n,i} - \mu_n)^2 - (\bar{R}_n(f_r, f_{s_r}) - \mu_n)^2\right].$$

Let $Y_{n,i} = \frac{X_{n,i} - \mu_n}{\sigma_n}$. By the uniform integrability assumption,

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[Y_{n,i}^2 I\left\{Y_{n,i}^2 > \lambda\right\}\right] = 0,$$

which implies

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[|Y_{n,i}| I\left\{|Y_{n,i}| > \lambda\right\}\right] = 0.$$

Thus, by (Lehmann & Romano 2010, Lemma 11.4.2),

$$\frac{\bar{R}_n(f_r, f_{s_r}) - \mu_n}{\sigma_n} = \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} Y_{n,i} \xrightarrow{p} 0.$$

By Lemma 11.4.3 of the same work,

$$\frac{\frac{1}{n_{te}}\sum_{i=1}^{n_{te}} (X_{n,i} - \mu_n)^2}{\sigma_n^2} = \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} Y_{n,i}^2 \xrightarrow{p} 1.$$

It then follows by the continuous mapping theorem that

$$\frac{\bar{\sigma}_n(f_r, f_{s_r})}{\sigma_n} \xrightarrow{p} 1.$$

Furthermore,

$$\frac{\sqrt{n_{te}} \left[\bar{R}_n(f_r, f_{s_r}) - \mu_n \right]}{\sigma_n} \xrightarrow{d} N(0,1)$$

by (Lehmann & Romano 2010, Lemma 11.4.1). Note that when $r \in \Theta_{\epsilon,n}^{MSS}$, $\mu_n - \epsilon \leq 0$. Thus, for any $P \in \mathcal{P}$, by Slutsky's Lemma and Polya's Theorem,

$$
\begin{aligned}
\limsup_{n \to \infty} P\{\phi_{CLT}(f_r) = 1 \cap E_{r,\epsilon,n}\} &= \limsup_{n \to \infty} P\left\{ \frac{\sqrt{n_{te}}}{\bar{\sigma}_n(f_r, f_{s_r})} \bar{R}_n(f_r, f_{s_r}) > \delta(\alpha) + \epsilon \cap E_{r,\epsilon,n} \right\} \\
&\leq \limsup_{n \to \infty} P\left\{ \frac{\sqrt{n_{te}}}{\bar{\sigma}_n(f_r, f_{s_r})} \left(\bar{R}_n(f_r, f_{s_r}) - \mu_n \right) > \delta(\alpha) \cap E_{r,\epsilon,n} \right\} \\
&\leq \limsup_{n \to \infty} P\left\{ \frac{\sqrt{n_{te}}}{\bar{\sigma}_n(f_r, f_{s_r})} \left(\bar{R}_n(f_r, f_{s_r}) - \mu_n \right) > \delta(\alpha) \right\} \\
&= \limsup_{n \to \infty} P\left\{ \frac{\sqrt{n_{te}}}{\sigma_n(f_r, f_{s_r})} \left(\bar{R}_n(f_r, f_{s_r}) - \mu_n \right) > \delta(\alpha) \right\} \\
&\leq \alpha
\end{aligned}
$$

where $\delta(\alpha) = \Phi^{-1}(1 - \alpha)$ and $E_{r,\epsilon,n}$ denotes the event that $r \in \Theta_{\epsilon,n}^{MSS}$. $\qquad\square$

A.4 Proof of Theorem 2

Proof. Assume the sequence $\left(\bar{\nabla}_n(Z_0) - \bar{\mu}_n \right)^2 / \sigma_n^2$ is uniformly integrable and $\gamma_{n_{tr}}^{loss}(g) = o(\sigma_n^2/n)$. Let $\mu_n = R_n(f_r) - R_n(f_{s_r})$. By (Bayle et al. 2020, Theorems 2 and 4),

$$\frac{\bar{\sigma}_n(f_r, f_{s_r})}{\sigma_n} \xrightarrow{p} 1 \text{ and } \frac{\sqrt{n}}{\sigma_n} \left(\bar{R}_n(f_r, f_{s_r}) - \mu_n \right) \xrightarrow{d} N(0,1).$$

Therefore, by the same steps as the end of the proof of Theorem 1, $\limsup_{n \to \infty} P\{\phi_{CLT}(f_r) = 1 \cap r \in \Theta_{n,\epsilon}^{MSS}\} \leq \alpha$ for any $P \in \mathcal{P}$. $\qquad\square$

A.5 Proof of Theorem 3

Proof. By the law of total expectation and the fact that each $Z \in \mathcal{D}^{(n)}$ are i.i.d.,

$$\mathbb{E}\left[\bar{R}_n^{\exp}\left(f_r, f_{s_r}, \omega\right)\right] = k_n^{-1} \sum_{j=1}^{k_n} \mathbb{E}\left[\exp\left\{\omega \sum_{Z \in \mathcal{D}_{-j}^{(n)}} \nabla_{n,j}\left(f_r, f_{s_r}, Z\right)\right\}\right]$$

$$= k_n^{-1} \sum_{j=1}^{k_n} \mathbb{E}\left[\prod_{Z \in \mathcal{D}_{-j}^{(n)}} \mathbb{E}_Z\left[\exp\left\{\omega \nabla_{n,j}\left(f_r, f_{s_r}, Z\right)\right\}\right]\right].$$

By the strong central condition, $\mathbb{E}_Z\left[\exp\left\{\omega \nabla_{n,j}\left(f_r, f_{s_r}, Z\right)\right\}\right] \leq 1$ almost surely for any $j \in [k_n]$ and $Z \in \mathcal{D}_{-j}^{(n)}$. Thus, $\mathbb{E}\left[\bar{R}_n^{\exp}\left(f_r, f_{s_r}, \omega\right)\right] \leq 1$. By Markov's Inequality, for any $P \in \mathcal{P}$,

$$P\left(\phi_{UI}\left(f_r\right) = 1 \cap r \in \Theta_{\epsilon,n}^{MSS}\right) \leq P\left(\bar{R}_n^{\exp}\left(f_r, f_{s_r}, \omega\right) > \alpha^{-1} \cap r \in \Theta_{\epsilon,n}^{MSS}\right)$$

$$\leq P\left(\bar{R}_n^{\exp}\left(f_r, f_{s_r}, \omega\right) > \alpha^{-1}\right)$$

$$\leq \alpha \mathbb{E}\left[\bar{R}_n^{\exp}\left(f_r, f_{s_r}, \omega\right)\right]$$

$$\leq \alpha.$$

$\square$

B Further Discussion and Results

B.1 A Conservative Version of ϕ_{CLT}

In Section 3.1, the studentized test ϕ_{CLT} was proposed. Recall that for ϕ_{CLT} to be combined with Proposition 2 to conduct valid MCS, we require condition (5). There, however, exist some settings in which $R_n\left(f_r\right) - \inf_{t \in T_r} R\left(h_t\right) - \epsilon \leq 0$ is only likely to occur in extremely large sample sizes. A simple example of such an atypical setting is when the training set size n_{tr} is much smaller than the test set size n_{te}. Figure 4 showcases a simulated regression scenario where the MCS test $\phi_{CLT}^{Holdout}$ as defined in Section 5, fails to appropriately control the type I error even when the sample size is very large.
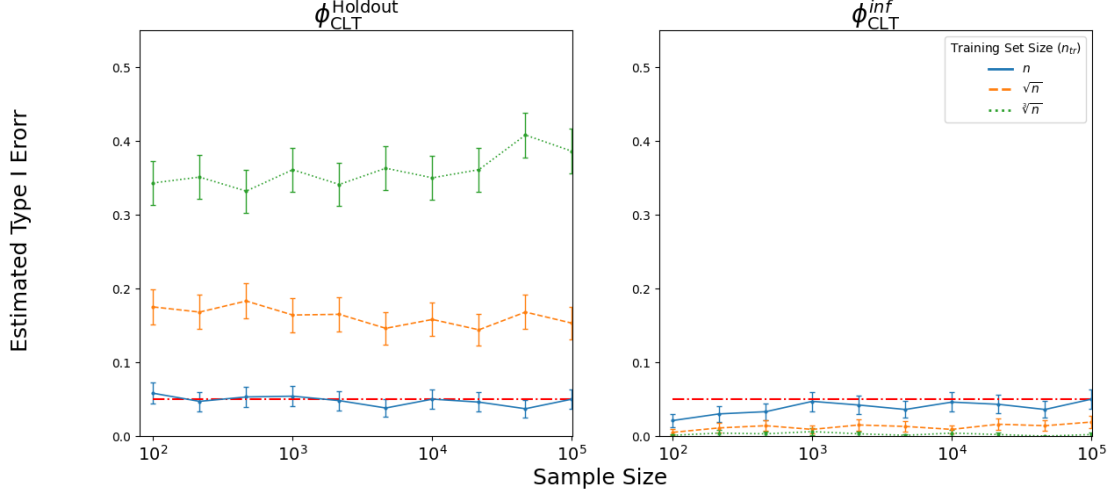
Figure 4: Example of a simulated regression scenario where $\phi_{CLT}^{Holdout}$ fails to control the type I error while ϕ_{CLT}^{inf} does. For ϕ_{CLT}^{inf}, we set $k_n = 1$ and use the same holdout split as $\phi_{CLT}^{Holdout}$. Let $Z_i = (Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^3$ where $Y_i = X_i\beta + \epsilon_i$, $X_i \overset{\text{i.i.d.}}{\sim} N(0_3, I_3)$, $\beta = (1, 1, 1)^\top$, and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$. $\mathcal{H}_1 = \{h_t \ : \ t \in \mathbb{R}^3, t_2 = 0\}$ and $\mathcal{H}_2 = \{h_t \ : \ t \in \mathbb{R}^3, t_3 = 0\}$ correspond to classes of additive linear models where the second and third features are not used, respectively. The loss, ℓ, is the squared error loss: $\ell(h_\theta, Z_i) = (X_i\theta - Y_i)^2$. The left and right plots correspond to using the tests $\phi_{CLT}^{Holdout}$ and ϕ_{CLT}^{inf} to test $H_{0,1} : 1 \in \Theta$. Note by definition that $H_{0,1}$ is true. We set f_1 and f_2 to output the OLS esimates for $\mathcal{H}_1$ and $\mathcal{H}_2$, $s_1 = 2$, and $\epsilon = 1 \times 10^{-6}$. The solid, dashed, and dotted lines correspond to the estimated type I error across 1,000 simulations for training set sizes $n/2$, $\sqrt{n}$, and $\sqrt[3]{n}$, respectively. The dashed/dotted horizontal line depicts the significance level $\alpha = .05$.

Define the test

$$\phi_{CLT}^{inf}(f_r) = I\left\{\inf_{h \in \mathcal{H}_r} \bar{R}_n(f_h, f_{s_r}) > k_n^{-1/2} n_{te}^{-1/2} \bar{\sigma}_n(f_r, f_{s_r})\left(\Phi^{-1}(1 - \alpha) + \epsilon\right)\right\}. \quad (10)$$

Note that $\inf_{h \in \mathcal{H}_r} \bar{R}_n(f_h, f_{s_r})$ is a conservative estimate for $\bar{R}_n(f_r, f_{s_r})$ when $k_n = 1$. The next result shows that ϕ_{CLT}^{inf} satisfies similar asymptotic properties to ϕ_{CLT}.

Corollary 1 (Validity of ϕ_{CLT}^{inf} when $k_n = 1$)**.** *Under the assumptions of Theorem 1, the*

38

test (10) satisfies the asymptotic version of (3).

Proof. Note that $\inf_{h \in H_r} \bar{R}_n(f_h, f_{s_r}) \leq \bar{R}_n(f_r, f_{s_r})$. Thus,

$$\limsup_{n \to \infty} P\left\{\phi_{CLT}^{inf}(f_r) = 1 \cap r \in \Theta_{n,\epsilon}^{MSS}\right\} \leq \limsup_{n \to \infty} P\left\{\phi_{CLT}(f_r) = 1 \cap r \in \Theta_{n,\epsilon}^{MSS}\right\}.$$

The result follows from Theorem 1. □

In contrast to ϕ_{CLT}, the test ϕ_{CLT}^{inf} appropriately controls the type I error in the situation displayed by Figure 4. A downside of ϕ_{CLT}^{inf} is that $\inf_{h \in \mathcal{H}_r} \bar{R}_n(f_h, f_{s_r})$ has to be computable which limits its application to only model classes where an empirical risk minimizer is able to be found.

B.2 Likelihood Ratio Test

Assume for $r \in [d]$ that $\mathcal{H}_r = \{P_\theta : \theta \in \Theta_r\}$ is a class of statistical models where each P_θ for $\theta \in \Theta_r \subseteq \Theta \subseteq \mathbb{R}^p$ denotes a unique distribution parameterized by the p-dimensional vector θ. Typically in such a setting, the goal is to correctly identify the parameter $\theta \in \Theta$ which is most likely to have generated the data $\{Z_1, \ldots, Z_n\}$. With this in mind, let the loss function be the negative log-likelihood function

$$\ell(P_\theta, z) = -\log(p_\theta(z))$$

where $p_\theta(z)$ is the likelihood of observating $z \in \mathcal{Z}$ under the distribution P_θ. Assuming the true data generating distribution P_{θ^*} exists such that $\theta^* \in \Theta$, then θ^* will be the minimizer of $R(P_\theta)$ across $\theta \in \Theta$ (Bickel & Doksum 2015).

Let

$$L_n(\theta) = \sum_{i=1}^n \log(p_\theta(Z_i)) = \log\left(\prod_{i=1}^n p_\theta(Z_i)\right)$$

be the empirical error for the parameter $\theta \in \Theta$. Set $\hat{\theta} = \arg\max_{\theta \in \Theta} \prod_{i=1}^n p_\theta(Z_i)$ to be the maximum likelihood estimate across Θ. Let $\chi_p^2(1-\alpha)$ represent the upper α-quantile of the

chi-squared distribution with p degrees of freedom. Assume $f_1, \ldots, f_d$ represent empirical risk minimization strategies for $\mathcal{H}_1, \ldots, \mathcal{H}_d$. Then, the likelihood ratio test

$$\phi\left(f_r\right) = I\left\{-2\left[L_n\left(\theta_r\right) - L_n\left(\hat{\theta}\right)\right] > \chi_p^2(1 - \alpha)\right\} \text{ where } P_{\theta_r} = f_r\left(\mathcal{D}^{(n)}\right) \qquad (11)$$

satisfies (3) as long as a few regularity conditions are satisfied (Wilks 1938). See (Bickel & Doksum 2015, Section 6.2) or (Wasserman et al. 2020, Section 3) for a discussion of the necessary regularity conditions which include statistical models being identifiable and differentiable in quadratic mean, a compact parameter space, and the log-likehood being a smooth function for any $\theta \in \Theta$. In certain statistical settings the likelihood ratio test been shown to be the the uniformly most powerful test (Neyman et al. 1997). Alternatives to the likelihood ratio statistic with similar forms of limiting distributions are the wald statistic and rao score test statistic (Bickel & Doksum 2015).

We are not the first to notice the potential application of the asymptotics of the likelihood ratio test statistic for MCS. A likelihood ratio based methodology has been proposed to conduct valid MCS with collections of appropriately nested model classes (Zheng et al. 2019, Li et al. 2019). Another line of work proposes a similar method for model classes composed of linear mixed models (Jiang et al. 2008). Furthermore, when $\mathcal{H}_1, \ldots, \mathcal{H}_d$ are composed of gaussian linear models, the likelihood ratio test is equivalent to conducting an F-test (Bickel & Doksum 2015, Proposition 6.1.2) and provides exact coverage. Similar methodology relying on the F-test applicable to MCS for nested gaussian linear model classes has also been suggested (Ferrari & Yang 2015). Lastly, it should be noted that the original UI methodology was inspired by the likelihood ratio test (Wasserman et al. 2020).

B.3   Concentration Inequality based Approaches

For simplicity, let $\mathcal{D}_0^{(n)} \subseteq \mathcal{D}^{(n)}$ and assume in the notation of Section 3 that $k_n = 1$. In the proof of Theorem 3, the strong central type condition, (8), was combined with Markov's

inequality to construct a concentration inequality similar to

$$\mathbb{P}\left\{\sum_{Z \in \mathcal{D}_0^{(n)}} \nabla_{n,1}\left(f_r, f_{s_r}, Z\right) - \left[R_n\left(f_r\right) - R_n\left(f_{s_r}\right)\right] \leq \delta_n(\alpha)\right\} \geq 1 - \alpha \qquad (12)$$

for some arbitrary choice of $\delta_n\left(\alpha\right)$. A result like that of Theorem 3 holds for this general

inequality as well. Define the test

$$\phi_{CI}\left(f_r\right) = I\left\{\sum_{Z \in \mathcal{D}_0^{(n)}} \nabla_{n,1}\left(f_r, f_{s_r}, Z\right) > \delta_n\left(\alpha\right) + \epsilon\right\}. \qquad (13)$$

Then, under the assumption of (12),

$$P\left\{\phi_{CI}\left(f_r\right) = 1 \cap r \in \Theta_{\epsilon,n}^{MSS}\right\} = P\left\{\sum_{Z \in \mathcal{D}_0^{(n)}} \nabla_{n,1}\left(f_r, f_{s_r}, Z\right) > \delta_n\left(\alpha\right) + \epsilon \cap r \in \Theta_{\epsilon,n}^{MSS}\right\}$$

$$\overset{(a)}{\leq} P\left\{\sum_{Z \in \mathcal{D}_0^{(n)}} \nabla_{n,1}\left(f_r, f_{s_r}, Z\right) - \left[R_n\left(f_r\right) - R_n\left(f_{s_r}\right)\right] > \delta_n\left(\alpha\right) \cap r \in \Theta_{\epsilon,n}^{MSS}\right\}$$

$$\leq P\left\{\sum_{Z \in \mathcal{D}_0^{(n)}} \nabla_{n,1}\left(f_r, f_{s_r}, Z\right) - \left[R_n\left(f_r\right) - R_n\left(f_{s_r}\right)\right] > \delta_n\left(\alpha\right)\right\}$$

$$\leq \alpha$$

where (a) follows from $r \in \Theta_{\epsilon,n}^{MSS}$ which implies that $R_n\left(f_r\right) - R_n\left(f_{s_r}\right) - \epsilon \leq 0$.

There are many methods for deriving concentration inequalities in the form of (12) (Boucheron et al. 2013). A recent work on MSS showcases a valid construction of (12) using a one-sided empirical Bernstein inequality (Takatsu & Kuchibhotla 2025). When $\mathcal{D}_0^{(n)}$ and $\mathcal{D}_1^{(n)}$ form a partition of $\mathcal{D}^{(n)}$, popular alternative approaches to creating bounds in the form (12) often utilize exponential moment inequalities. For instance, one could assume that $\nabla_{n,1}\left(f_r, f_{s_r}, Z_0\right) - \left[R_n\left(f_r\right) - R_n\left(f_{s_r}\right)\right]$ is sub-gaussian, sub-exponential, or sub-weibull. Notably, the sub-gaussian assumption holds whenever $\nabla_{n,1}\left(f_r, f_{s_r}, Z_0\right)$ is almost surely bounded.

PAC bounds are a popular concept in statistical learning theory that provide probabilistic guarantees on a learning algorithm's performance. The PAC framework also allows for

the computation of data-dependent, uniform bounds of the form of (12) (Valiant 1984). In general, if the model class $\mathcal{H}'_r \subseteq \mathcal{H}$ where $f_r, f_{s_r} \in \mathcal{F}(\mathcal{H}'_r)$ is limited in complexity (as measured by e.g. its cardinality, rademacher complexity, fat shattering dimension, covering number, or VC dimension) and the loss function is light-tailed, then a bound may be derived through an appropriate choice of concentration inequalities (Vapnik 2000, Mohri et al. 2018, Fisher et al. 2019). Typical assumptions to ensure that the loss is light-tailed are to either assume it is bounded or satisfies an exponential moment inequality (Chafaï et al. 2012, Boucheron et al. 2013). In cases where $\mathcal{H}'_r$ is very complex (which is typical of ML algorithms), the limiting assumption on the complexity of $\mathcal{H}'_r$ may be replaced by stability assumptions on f_r and f_{s_r} (Elisseeff et al. 2005, Shalev-Shwartz et al. 2010, Feldman & Vondrak 2018). Similar to the assumptions of Theorem 2, these stability assumptions typically force the difference between $\ell\left(f_r\left(\mathcal{D}^{(n)}\right), Z_0\right) - \ell\left(f_{s_r}\left(\mathcal{D}^{(n)}\right), Z_0\right)$ and $\ell\left(f_r\left(\mathcal{D}^{(n)\setminus i}\right), Z_0\right) - \ell\left(f_{s_r}\left(\mathcal{D}^{(n)\setminus i}\right), Z_0\right)$ to decrease towards 0 in probability at a rate dependent on n. Recently, non-vacuous bounds for complex machine learning models such as neural networks have been achieved through PAC bayesian approaches that require suitable choices of data dependent priors (Dziugaite & Roy 2017). Current work on PAC bounds focus on developing tighter bounds with milder assumptions. Although noteworthy, PAC bounds are typically not very tight and yield conservative tests.

Another line of work has shown that if the loss is bounded and $\delta_n(\alpha)$ is chosen dependent on the complexity of $\mathcal{H}'_r$ where $f_r, f_{s_r} \in \mathcal{F}(\mathcal{H}'_r)$, then a bound of the form of (12) can be achieved (Fisher et al. 2019, Lemma 23). Follow up works have constructed procedures that would allow for computing $\inf_{t \in T_r} \phi_{CI}(h_t)$ in such cases. This would permit an application of Proposition 1 to construct a valid MCS test. These methods are viable when $\mathcal{H}'_r$ is composed of simpler model types such as (regularized) linear models, linear models in a reproducing kernel Hilbert space, rule lists, sparse decision trees, kernel ridge regression

models, random forests with a fixed set of pre-trained trees, and sparse generalized additive models (Fisher et al. 2019, Hara & Ishihata 2018, Mata et al. 2022, Xin et al. 2022, Laberge et al. 2023, Zhong et al. 2023).

B.4 Uniform Coverage

Thus far, we have only discussed test construction strategies that yield a confidence set $\hat{\Theta}$ with point-wise coverage guarantees (2). We wish to point out that attaining stronger uniform coverage guarantees of the form

$$\left(\overset{(as)}{\underset{n \to \infty}{\lim\inf}} \right) \inf_{P \in \mathcal{P}} P\left(\Theta \subseteq \hat{\Theta} \right) \geq 1 - \alpha \tag{14}$$

would require controlling the family-wise error rate of the tests ψ_r for $r \in [d]$. The simplest approach would be to utilize methods from the multiple comparisons literature. For example, setting the significance level to α/d would appropriately result in uniform coverage by the union bound if (1) is satisfied. As another option, the Benjamini-Hochberg method is a multiple comparisons procedure that fixes some potential issues of the union bound. Notably, it has already been applied in MSS settings (Dey et al. 2026).

Alternatively, it has been shown that dimension agnostic uniform coverage is possible in some settings by including a data-dependent screening step (Kim & Ramdas 2025). Such methodology would require constructing an MCS confidence set from a subset of the data, then using the size of the confidence set to estimate an appropriate signifiance level at which to run the MCS procedure on the full sample. These methods, however, are beyond the scope of this paper, and we leave further investigation of the potential of these approaches to future work.
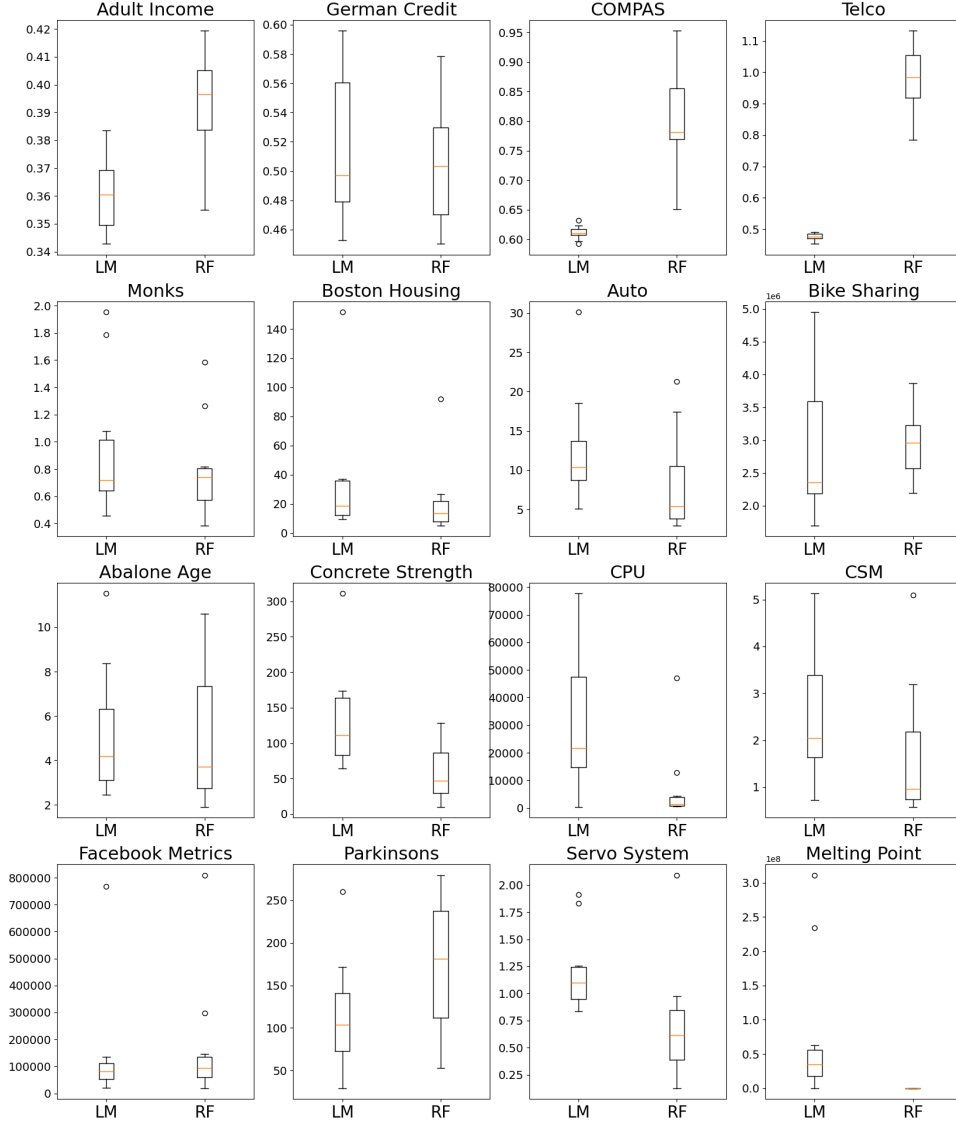
B.5 Box Plots for Real Data Experiments

Figure 5: Box plots representing the 10-fold cross-validation errors for f_1 and f_2 that are used in the computation of the test statistic $\bar{R}_n\left(f_1, f_{s_1}\right)$ when using the test ϕ_{CLT}^{CV}. LM and RF correspond to classes $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively.