Journal of Artificial Intelligence Research 78 (2023) 491-526 Submitted 06/2022; published 11/2023

Asymptotics of K-Fold Cross Validation

Jessie Li JEQLI@UCSC.EDU
Department of Economics
University of California, Santa Cruz
1156 High Street, Santa Cruz, CA 95064, USA

Abstract

This paper investigates the asymptotic distribution of the K-fold cross validation error in an i.i.d. setting. As the number of observations n goes to infinity while keeping the number of folds K fixed, the K-fold cross validation error is $\sqrt{n}$-consistent for the expected out-of-sample error and has an asymptotically normal distribution. A consistent estimate of the asymptotic variance is derived and used to construct asymptotically valid confidence intervals for the expected out-of-sample error. A hypothesis test is developed for comparing two estimators' expected out-of-sample errors and a subsampling procedure is used to obtain critical values. Monte Carlo simulations demonstrate the asymptotic validity of our confidence intervals for the expected out-of-sample error and investigate the size and power properties of our test. In our empirical application, we use our estimator selection test to compare the out-of-sample predictive performance of OLS, Neural Networks, and Random Forests for predicting the sale price of a domain name in a GoDaddy expiry auction.

1. Introduction

This paper studies the asymptotics of K-fold cross validation in an i.i.d. setting as a way to approximate an estimator's expected out-of-sample error, which is the expected predictive loss evaluated at a new observation drawn independently but from the same distribution as the data. We first derive the asymptotic distribution of the K-fold cross validation error centered around the expected out-of-sample error as the number of observations n goes to infinity but the number of folds K is fixed. The assumption of fixed K is reasonable given that researchers typically use $K = 5$ or $K = 10$ in practice. We note that the assumption of fixed K rules out leave-one-out cross validation, which effectively sets $K = n$. We find that the asymptotic distribution of the K-fold cross validation error does not depend on K, so for sufficiently large n, the choice of K should not be a first order concern. We also provide a consistent estimate of the asymptotic variance, which gives researchers a simple way to construct asymptotically valid standard errors for the K-fold cross validation error. The researcher can use these standard errors to form confidence intervals for the expected out-of-sample error. Our results are derived for estimators with asymptotically linear representations and well-defined probability limits. The cross validation loss functions are three times continuously differentiable at this probability limit with derivatives having bounded second moments. Additionally, we allow for the objective function used to compute the estimators to be nondifferentiable; for example, we allow for quantile regression and ℓ_2-norm Support Vector Machine regression.

It is sometimes the case that researchers would like to have a formal statistical test for comparing the predictive performance of two different estimators. To aid them on this front, we formulate a hypothesis test where the null hypothesis is that the two estimators have the same expected out-of-sample error. Our null hypothesis is different from traditional model selection tests because we are not trying to discern the true functional form of the conditional mean function or some other feature of interest. Instead, we are comparing the out-of-sample predictive performance of different estimators. One motivation for doing such an exercise is that sometimes different estimators can be consistent for the same underlying conditional mean function; so traditional model selection tests would view them as equally good, and yet they can still differ in how well they predict the outcome out-of-sample. For instance, we can compare a kernel regression estimator to a local linear regression estimator using the same set of covariates, even if both estimators are consistent for the same underlying conditional mean function. Our test also allows for our estimators to be consistent for different conditional mean functions and the estimators can have different rates of convergence. For example, in the empirical application, we compare the predictive performance of OLS, Neural Networks, and Random Forests for predicting the sale price of a domain name in a GoDaddy expiry auction.

Section 2 provides a literature review. Section 3 contains the theoretical results on the asymptotic normality of the K-fold cross validation error and consistent estimation of the asymptotic variance. Section 4 contains the theoretical results of the hypothesis testing procedure. Section 5 contains Monte Carlo simulations examining the empirical coverage frequencies and rejection frequencies for a class of linear models. Section 6 contains an empirical application of our estimator selection test conducted pairwise and size-adjusted for multiple testing using a Bonferroni correction. Section 7 concludes. The appendix contains proofs of the main results. Some notation that will be used in this paper are as follows: $X_n = O_p(1)$ means for any $\epsilon > 0$, there exists a finite $M > 0$ and a finite $N > 0$ such that $P(|X_n| > M) < \epsilon \quad \forall n > N$, $X_n = o_p(1)$ means for any $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n| \geq \epsilon) = 0$, and plim denotes probability limit.

2. Literature Review

An early paper that discusses K-fold cross validation is (Burman, 1989), which examines the bias and variance of K-fold cross validation as an estimator for the expected prediction error (measured using squared error loss) conditional on the data. (Zhang, 1993) derives the probabilities of selecting each model when performing K-fold cross validation using linear models. (Shao, 1993) provides necessary and sufficient conditions for variable selection consistency for linear models using leave-n_v-out cross validation, where n_t observations are used for estimating the models using linear regression and n_v observations are used for validating the accuracy of the models' predictions. He demonstrates that necessary and sufficient conditions for variable selection consistency are $n_v/n_t \to \infty$ and $n_t \to \infty$. (Yang, 2007) considers an extension of (Shao, 1993)'s consistency results to nonparametric regression. If at least one model is estimated at a rate slower than $O(n_t^{-1/2})$, then leave-n_v-out cross validation is consistent for the best model when $n_t/n_v \to O(1)$ and sometimes even when $n_t/n_v \to \infty$. In the econometrics literature, (Li & Racine, 2004) and (Hall, Racine, & Li, 2004) examine the optimality properties of using cross validation to select the bandwidth for local linear regression and conditional probability density estimation, respectively, in the i.i.d. setting. (Racine, 2000) proposes an improved version of (Burman, Chow, & Nolan, 1994)'s h-block cross validation method for dependent data, and he demonstrates that the probability of selecting the model with the best predictive ability converges to 1 as the total number of observations approaches infinity. An excellent survey article on cross validation is (Arlot & Celisse, 2010), which is the paper that first motivated work on our paper and therefore contains similar notation as ours.

None of the above papers consider constructing confidence intervals for the expected out-of-sample errors or testing whether two models have the same expected out-of-sample error. A recent paper (Lei, 2019) considers testing whether a linear model has the lowest average predictive risk among a fixed number of candidate models in the classical linear regression setting. However, their notion of predictive risk (also called the conditional test error) is different from our notion of expected out-of-sample error because they condition on the training data when taking the expectation over the new observations, whereas we take the expectation over both the new observations and the training data. This distinction is important to consider because as noted in section 7.12 of (Hastie, Tibshirani, & Friedman, 2009) and also more recently, (Zhu & Timmermann, 2020) and (Bates, Hastie, & Tibshirani, 2023), K-fold cross validation only estimates well the expected out-of-sample error (also called the expected test error) rather than the conditional test error. Although (Bates et al., 2023) propose a resampling procedure to obtain confidence intervals for the expected out-of-sample error, they do not derive the asymptotic distribution of the K-fold cross validation error. Their procedure (called nested CV) is also more computationally intensive than our analytic standard errors because it involves performing cross validation repeatedly for many iterations. An important recent paper (Wager, 2020) derives the asymptotic distribution of K-fold cross validation for nonparametric estimators; we will show that our result is consistent with their result in the case of nonparametric estimators while also noting that our more general results allow for both parametric and nonparametric estimators. Another important recent paper (Bayle, Bayle, Janson, & Mackey, 2020) also derives the asymptotic distribution of K-fold cross validation but centered around the fold error rather than the expected out-of-sample error. They note at the end that further work could be done to extend their results to center around the expected-out-of-sample error. Another important recent paper (Austern & Zhou, 2020) does derive the asymptotic distribution centered around the expected out-of-sample error using different proof techniques based on Stein's method. However, they do not consider the problem of testing whether two estimators have the same expected out-of-sample error. Additionally, in their discussion of parametric M-estimators (see their Proposition 3 on page 12), they require the objective function to be strictly convex and twice differentiable in the parameter everywhere. We do not need the objective function to be differentiable or convex, which handles quantile regression and nonconvex maximum likelihood problems.

The literature on model selection tests also has many important papers; for example an early paper on model specification testing is (Bierens, 1982). He is testing the null hypothesis that the conditional expectation function can be specified as a parametric function, which is different from our null hypothesis comparing the expected out-of-sample errors of the two estimators. (Vuong, 1989) develops a likelihood ratio test for comparing models estimated by maximum likelihood. Later, (Rivers & Vuong, 2002) generalizes the framework to allow for models specified by moment conditions, but they require the estimator to be $\sqrt{n}$ consistent. (Lavergne & Vuong, 1996) propose a model selection test comparing two non-nested nonparametric models with different sets of regressors. (Lavergne, 2001) considers a test of the equality of two conditional expectation functions estimated using kernel regression. (Lavergne & Vuong, 2000) consider testing for the joint significance of a subset of continuous explanatory variables in nonparametric regression using kernel methods. (Chen, Hong, & Shum, 2007) use a likelihood ratio test to compare parametric likelihood models with moment-based models. (Hong & Preston, 2012) examine a more general class of estimators that include many non-likelihood based and semi-parametric estimators, but their procedure uses the Bayesian Information Criterion (BIC), which relies on knowing the effective model dimension, which can be difficult to determine for nonparametric estimators. More recently, (Shi, 2015) has developed a non-degenerate version of (Vuong, 1989)'s likelihood ratio test which exhibits uniform size control. (Liao & Shi, 2020) extend (Shi, 2015) to semi-parametric and nonparametric models estimated using a sieve M-estimator. Another parametric likelihood based model selection test that controls size uniformly over a large class of data generating processes is (Schennach & Wilhelm, 2017). The null hypotheses in these aforementioned papers are different than ours because they are testing properties about the true underlying model whereas we are interested in evaluating the out-of-sample predictive performance of our estimators. In this regard, our test might be more similar to tests of predictive ability, such as the Diebold-Mariano-West tests ((Diebold & Mariano, 1995), (Diebold, 2015), and (West, 1996)), the (Giacomini & White, 2006) test of conditional predictive ability, and the (Clark & McCracken, 2015) bootstrap-based test of out-of-sample forecast accuracy. However, in contrast to these papers, we allow for a wider class of estimators with different convergence rates and our focus is on i.i.d. cross-sectional data. The focus on i.i.d. cross-sectional data is because K-fold cross validation is primarily used in this context; however, it is not difficult to extend our method to balanced panel data.

3. Theoretical Properties of K-Fold Cross Validation

Consider an i.i.d. sample of data $\Xi \equiv \{\xi_i\}_{i=1}^n \equiv \{(x_i, y_i)\}_{i=1}^n$ from some unknown distribution P. Let $\mathbb{X} \subseteq \mathbb{R}^d$ be the support of x_i. The criterion for comparing a set of candidate estimators $\{\hat{s}_M : \mathbb{X} \mapsto \mathbb{R}, \text{ for } M \in \mathcal{M}\}$ is the expected out-of-sample error $EPE_{OUT,n}(M) \equiv \mathbb{E}[\gamma(\hat{s}_M; \tilde{\xi_i})]$, which is the expected loss if we were to apply estimator M's $\hat{s}_M(\cdot)$ computed over n observations on a new $\tilde{\xi_i}$ drawn independently from the same distribution as $\Xi \sim P$. The expectation $\mathbb{E}[\gamma(\hat{s}_M; \tilde{\xi_i})]$ is taken with respect to *both* the training data Ξ used to form $\hat{s}_M(\cdot)$ and the new observation $\tilde{\xi_i}$. For this reason, $EPE_{OUT,n}(M)$ can be viewed as a measure of unconditional expected predictive accuracy.

We will use K-fold cross validation to estimate the expected out-of-sample error. Formally, the K-fold cross validation procedure selects the estimator among the set of candidate estimators $\mathcal{M}$ which minimizes the average prediction error for observations ξ_i in the k^{th} validation fold, averaged over $k = 1, ..., K$.

$$\hat{\mathcal{L}}_n^{CV}(M) \equiv \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_v} \sum_{i \in I_k^{(v)}} \gamma(\hat{s}_M^{(-k)}; \xi_i)$$

$I_k^{(v)}$ are the indices of the observations in the k^{th} validation fold and $n_v \equiv \frac{n}{K}$ is the number of observations in each validation fold, assuming WLOG that n is divisible by K. Later, we will use $n_t \equiv \frac{K-1}{K}n$ to denote the number of observations in each training fold. $\hat{s}_M^{(-k)}(\cdot)$ is estimator M's estimator computed using the observations not in the k^{th} validation fold. $\gamma(\cdot;\cdot)$ is a loss function such as the squared error loss $\gamma(\hat{s}_M;\xi_i) = (y_i - \hat{s}_M(x_i))^2$.

While we do not discuss the details in this paper, it is not difficult to extend our results to balanced panel data where we have data on n individuals over a fixed time period T by redefining the K-fold Cross Validation error as

$$\hat{\mathcal{L}}_{n,T}^{CV}(M) \equiv \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_v}\sum_{i\in I_k^{(v)}}\frac{1}{T}\sum_{t=1}^{T}\gamma(\hat{s}_M^{(-k)};\xi_{it})$$

where $I_k^{(v)}$ are the indices of the individuals in the k^{th} validation fold, $n_v \equiv \left\lceil \frac{n}{K}\right\rceil$ is the number of individuals in each validation fold, and $\hat{s}_M^{(-k)}(\cdot)$ is model M's estimator computed using the individuals not in the k^{th} validation fold.

We now present the assumptions that we will need throughout the paper. The first assumption says that $s_M^*(x_i)$ is the probability limit of $\hat{s}_M(x_i)$ for all observations x_i and all estimators M in a finite set $\mathcal{M}$. We emphasize that different estimators may have different s_M^*'s, and all of the s_M^*'s can be different from the true feature s_0 corresponding to the underlying data generating process. Therefore, our first assumption allows for estimators that are inconsistent for the true conditional mean function or other features of the data generating process.

Assumption 3.1 *The set of candidate estimators $\mathcal{M}$ is finite, and*

$$\max_{M\in\mathcal{M}}\max_{1\le i\le n}|\hat{s}_M(x_i) - s_M^*(x_i)| = o_p(1)$$

The next assumption says the loss function is three times continuously differentiable with zero third derivative.

Assumption 3.2 *$\gamma(s;\xi_i)$ is a three times continuously differentiable function of $s(x_i)$ with $\mathbb{E}\left[\left(\frac{\partial\gamma(s;\xi_i)}{\partial s}\right)^2\right] < \infty$, $\mathbb{E}\left[\left(\frac{\partial^2\gamma(s;\xi_i)}{\partial s^2}\right)^2\right] < \infty$, and zero third derivative $\frac{\partial^3\gamma(s;\xi_i)}{\partial s^3} = 0$ for all i.*

We will use $\frac{\partial\gamma(s;\xi_i)}{\partial s}$ to denote the derivative. Although the notation may suggest otherwise, we are taking the derivative with respect to a particular point $s(x_i)$ rather than the whole function $s(\cdot)$ itself. An example of a loss function that satisfies Assumption 3.2 is the squared error loss. Note that even though our loss function is three times continuously differentiable, the objective function used when computing the estimator $\hat{s}_M(x_i)$ does not need to be differentiable.

The next assumption says the estimators have an asymptotically linear representation.

Assumption 3.3 *For each estimator $M \in \mathcal{M}$, there exists $\tau_n \to \infty$ satisfying $\tau_n/n^{1/4} \to \infty$ and $\tau_n/\sqrt{n} = O(1)$, and there exists a function $\phi_M(\xi_j, x_i; s_M^*)$ satisfying $\mathbb{E}\left[\phi_M(\xi_j, x_i; s_M^*)\middle| x_i\right] = 0$ and $\mathbb{E}\left[\left(\phi_M(\xi_j, x_i; s_M^*)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\right)^2\right] < \infty$ when $\tau_n = \sqrt{n}$, and*

$Var\left[\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\middle|\,\xi_i\right]\right] \to 0$ and $Var\left[\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\middle|\,\xi_j\right]\right] \to 0$ when $\tau_n \ll \sqrt{n}$, such that for each $x_i \in \mathcal{X}$,

$$\hat{s}_M(x_i) - s_M^*(x_i) = \frac{1}{n}\sum_{j=1}^{n}\phi_M\left(\xi_j, x_i; s_M^*\right) + R_{n,i}$$

where $\max\limits_{1\le i\le n} R_{n,i} = o_p\left(\frac{1}{\tau_n}\right)$.

Assumption 3.3 states that the difference between $\hat{s}_M(x_i)$ and its probability limit $s_M^*(x_i)$ at each validation data point x_i is asymptotically equivalent to a scaled average of terms $\phi_M\left(\xi_j, x_i; s_M^*\right)$ that depend only on the individual training observation ξ_j, the fixed point of evaluation x_i, the function s_M^*, and possibly some other nonrandom quantities. For $\sqrt{n}$ consistent estimators where $\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\middle|\, x_i\right] = 0$, the function $\phi_M(\cdot,\cdot;\cdot)$ is called the influence function (see e.g. (Van der Vaart, 2000) or (Newey & McFadden, 1994)). For example, OLS regression has an influence function $\phi_M\left(\xi_j, x_i; \beta_M^*\right) = x_i'\mathbb{E}[x_j x_j']^{-1}x_j(y_j - x_j'\beta_M^*)$ which satisfies Assumption 3.3 under the weak exogeneity assumption $\mathbb{E}\left[x_j(y_j - x_j'\beta_M^*)\right] = 0$. For nonparametric estimators, typically $\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\middle|\, x_i\right] \neq 0$ because of the bias, but $Var\left[\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\middle|\,\xi_i\right]\right] \to 0$ and $Var\left[\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\middle|\,\xi_j\right]\right] \to 0$. For nonparametric kernel estimators, the bias is of order h^2 (where h is the bandwidth), so $\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\middle|\,\xi_i\right] = h^2\lambda(x_i)$ and $\mathbb{E}\left[\phi_M\left(\xi_j, x_i; s_M^*\right)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\middle|\,\xi_j\right] = h^2\nu(x_j)$, where $h \to 0$ as $n \to \infty$. Under weak regularity conditions, it is possible to show that $Var\left[h^2\lambda(x_i)\right] \to 0$ and $Var\left[h^2\nu(x_i)\right] \to 0$.

The $\phi_M(\cdot,\cdot;\cdot)$ function can be obtained from the first order Taylor expansion of the first order condition. For example, if $\hat{s}_M(x_i) = F(x_i, \hat{\beta}_M)$, $s_M^*(x_i) = F(x_i, \beta_M^*)$, where $F(\cdot,\cdot)$ is known, $\hat{\beta}_M = \arg\min\limits_{\beta}\left\{Q_n^M(\beta) = \frac{1}{n}\sum_{j=1}^{n} q_j^M(\beta)\right\}$, and $\beta_M^* = \arg\min\limits_{\beta}\left\{Q^M(\beta) = \mathbb{E}\left[q_j^M(\beta)\right]\right\}$, then the first order Taylor expansion of the first order condition for $\hat{\beta}_M$ is $0 = \frac{\partial Q_n^M(\hat{\beta}_M)}{\partial\beta} = \frac{\partial Q_n^M(\beta_M^*)}{\partial\beta} + \left(\hat{\beta}_M - \beta_M^*\right)'\frac{\partial^2 Q_n^M(\beta_M^*)}{\partial\beta\partial\beta'} + O_p\left(\left\|\hat{\beta}_M - \beta_M^*\right\|^2\right)$, which implies

$$\begin{aligned}\hat{s}_M(x_i) - s_M^*(x_i) &= F(x_i, \hat{\beta}_M) - F(x_i, \beta_M^*)\\ &= \frac{\partial F(x_i, \beta_M^*)}{\partial\beta'}\left(\hat{\beta}_M - \beta_M^*\right) + O_p\left(\left\|\hat{\beta}_M - \beta_M^*\right\|^2\right)\\ &= \frac{1}{n}\sum_{j=1}^{n}\underbrace{-\frac{\partial F(x_i, \beta_M^*)}{\partial\beta'}\text{plim}\left(\frac{\partial^2 Q_n^M(\beta_M^*)}{\partial\beta\partial\beta'}\right)^{-1}\frac{\partial q_j^M(\beta_M^*)}{\partial\beta}}_{\phi_M\left(\xi_j, x_i; \beta_M^*\right)} + R_{n,i}\end{aligned}$$

For nonparametric estimators where $\hat{s}_M(x_i) = \arg\min\limits_{s}\{Q_n^M(s, x_i) = \frac{1}{n}\sum_{j=1}^{n} q_j^M(s, x_i)\}$, the first order Taylor expansion of the first order condition is $0 = \frac{\partial Q_n^M(\hat{s}_M)}{\partial s} = \frac{\partial Q_n^M(s_M^*)}{\partial s} +$

$(\hat{s}_M(x_i) - s^*_M(x_i))\frac{\partial^2 Q_n^M(s^*_M)}{\partial s^2} + O_P\left(|\hat{s}_M(x_i) - s^*_M(x_i)|^2\right)$, which implies

$$\begin{aligned}\hat{s}_M(x_i) - s^*_M(x_i) &= -\left(\frac{\partial^2 Q_n^M(s^*_M)}{\partial s^2}\right)^{-1}\frac{1}{n}\sum_{j=1}^n \frac{\partial q_j^M(s^*_M, x_i)}{\partial s} + O_P\left(|\hat{s}_M(x_i) - s^*_M(x_i)|^2\right)\\ &= \frac{1}{n}\sum_{j=1}^n \underbrace{-\text{plim}\left(\frac{\partial^2 Q_n^M(s^*_M)}{\partial s^2}\right)^{-1}\frac{\partial q_j^M(s^*_M, x_i)}{\partial s}}_{\phi_M(\xi_j, x_i; s^*_M)} + R_{n,i}\end{aligned}$$

Examples of $\phi_M(\cdot,\cdot;\cdot)$ for some well known estimators are the following:

1. Ordinary Least Squares: For $\hat{s}_M(x_i) = x_i'\hat{\beta}_M$,

$$\begin{aligned}&\hat{s}_M(x_i) = x_i'\hat{\beta}_M = x_i'(X'X)^{-1}X'Y\\ &\beta^*_M = \mathbb{E}[x_j x_j']^{-1}\mathbb{E}[x_j y_j']\\ &\phi_M(\xi_j, x_i; \beta^*_M) = x_i'\mathbb{E}[x_j x_j']^{-1}x_j(y_j - x_j'\beta^*_M)\end{aligned}$$

2. Nonlinear Least Squares: For known $F(\cdot,\cdot)$, $\hat{s}_M(x_i) = F(x_i, \hat{\beta}_M)$,

$$\begin{aligned}&\hat{\beta}_M = \arg\min_\beta \frac{1}{n}\sum_{j=1}^n (y_j - F(x_j, \beta))^2\\ &\beta^*_M = \arg\min_\beta \mathbb{E}\left[(y_j - F(x_j, \beta))^2\right]\\ &\phi_M(\xi_j, x_i; \beta^*_M)\\ &= \frac{\partial F(x_i, \beta^*_M)}{\partial \beta'}\mathbb{E}\left[\frac{\partial F(x_j, \beta^*_M)}{\partial \beta}\frac{\partial F(x_j, \beta^*_M)}{\partial \beta'}\right]^{-1}\frac{\partial F(x_j, \beta^*_M)}{\partial \beta}(y_j - F(x_j, \beta^*_M))\end{aligned}$$

3. Ridge Regression: For $\hat{s}_M(x_i) = x_i'\hat{\beta}_M$,

$$\begin{aligned}&\hat{\beta}_M = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n \left(y_i - x_i'\beta\right)^2 + \lambda\beta'\beta\\ &\beta^*_M = \arg\min_\beta \mathbb{E}\left[\left(y_i - x_i'\beta\right)^2\right] + \lambda\beta'\beta\\ &\hat{s}_M(x_i) = x_i'(X'X + \lambda I)^{-1}X'Y\\ &\phi_M(\xi_j, x_i; \beta^*_M) = x_i'\mathbb{E}\left[x_j x_j' + \lambda I\right]^{-1}\left(x_j(y_j - x_j'\beta^*_M) - \lambda\beta^*_M\right)\end{aligned}$$

4. Kernel Regression: For $h \to 0$, $nh^d \to \infty$, $K_h(x_i - x_j) = \frac{1}{h^d}K\left(\frac{x_i - x_j}{h}\right)$, $f(\cdot)$ the density function of x_i, and $s^*_M(x_i) = \mathbb{E}[y_i|x_i]$,

$$\begin{aligned}&\hat{s}_M(x_i) = \arg\min_s \frac{1}{n}\sum_{j=1}^n (y_j - s)^2 K_h(x_i - x_j) = \frac{\sum_{j=1}^n K_h(x_i - x_j)y_j}{\sum_{j=1}^n K_h(x_i - x_j)}\\ &\phi_M(\xi_j, x_i; s^*_M) = f(x_i)^{-1}K_h(x_i - x_j)(y_j - s^*_M(x_i))\end{aligned}$$

5. Local Linear Regression: For $h \to 0$, $nh^d \to \infty$, $K_h(x_i - x_j) = \frac{1}{h^d} K\left(\frac{x_i - x_j}{h}\right)$, $f(\cdot)$ the density function of x_i, and $s^*_M(x_i) = \mathbb{E}[y_i|x_i]$,

$$(\hat{s}_M(x_i), \hat{\beta}_M) = \arg\min_{s,\beta} \frac{1}{n} \sum_{j=1}^n (y_j - s - (x_i - x_j)'\beta)^2 K_h(x_i - x_j)$$
$$\phi_M\left(\xi_j, x_i; s^*_M, \beta^*_M\right) = f(x_i)^{-1} K_h(x_i - x_j)(y_j - s^*_M(x_i) - (x_i - x_j)'\beta^*_M)$$

6. Quantile Regression: For $\hat{s}_M(x_i) = x_i'\hat{\beta}_M$, $\rho_\tau(u) = \{(1-\tau)\,\mathbf{1}\,(u \le 0) + \tau\mathbf{1}\,(u > 0)\}\,|u|$, and $f_{y_j|x_j}(\cdot)$ the conditional density function of y_j given x_j,

$$\hat{\beta}_M = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - x_i'\beta)$$
$$\beta^*_M = \arg\min_{\beta}\ \mathbb{E}\left[\rho_\tau(y_i - x_i'\beta)\right]$$
$$\phi_M\left(\xi_j, x_i; \beta^*_M\right) = x_i'\mathbb{E}\left[x_j x_j' f_{y_j|x_j}\left(x_j'\beta^*_M\right)\right]^{-1} x_j\left(\tau - \mathbf{1}\left(y_j \le x_j'\beta^*_M\right)\right)$$

7. Local Linear Quantile Regression: For $h \to 0$, $nh^d \to \infty$, and $s^*_M(x_i) = Q_\tau[y_i|x_i]$ the conditional τth quantile of y_i given x_i,

$$(\hat{s}_M(x_i), \hat{\beta}_M) = \arg\min_{s,\beta} \frac{1}{n} \sum_{j=1}^n \rho_\tau(y_j - s - (x_i - x_j)'\beta) K_h(x_i - x_j)$$
$$\phi_M\left(\xi_j, x_i; s^*_M, \beta^*_M\right) = \left(\operatorname{plim} \frac{1}{n} \sum_{j=1}^n K_h(x_i - x_j)\left(f_{y_j|x_j}\left(s^*_M(x_i) + (x_i - x_j)'\beta^*_M\right)\right)\right)^{-1}$$
$$K_h(x_i - x_j)\left(\tau - \mathbf{1}\left(y_j \le s^*_M(x_i) + (x_i - x_j)'\beta^*_M\right)\right)$$

8. ℓ_2-norm SVM regression: For $x^+ \equiv \max(x, 0)$,

$$\hat{\beta}_M = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^n \left(\rho_\tau\left(y_i - x_i'\beta\right) - \kappa\right)^+ + \lambda\beta'\beta$$
$$\beta^*_M = \arg\min_{\beta}\ \mathbb{E}\left[\left(\rho_\tau\left(y_i - x_i'\beta\right) - \kappa\right)^+ + \lambda\beta'\beta\right]$$
$$\phi_M\left(\xi_j, x_i; \beta^*_M\right)$$
$$= x_i' H^{-1}\left(x_j\left(\tau\mathbf{1}\left(y_j \ge x_j'\beta^*_M + \frac{\kappa}{\tau}\right) - (1-\tau)\,\mathbf{1}\left(y_j \le x_j'\beta^*_M - \frac{\kappa}{1-\tau}\right)\right) - \lambda\beta^*_M\right)$$
$$H = \mathbb{E}\left[x_j x_j'\left(\tau f_{y_j|x_j}\left(x_j'\beta^*_M + \frac{\kappa}{\tau}\right) + (1-\tau)\, f_{y_j|x_j}\left(x_j'\beta^*_M - \frac{\kappa}{1-\tau}\right)\right)\right]$$

9. Generalized Method of Moments: For known $F(\cdot,\cdot)$, $\hat{s}_M(x_i) = F(x_i, \hat{\beta}_M)$, $s^*_M(x_i) = F(x_i, \beta^*_M)$, moment functions $\mathbb{E}\left[g(\cdot, \beta)\right]$ which might not equal zero at β^*_M, and $G(\beta^*_M) \equiv$

$\mathbb{E}\left[\frac{\partial g(\xi_j,\beta_M^*)}{\partial\beta}\right]$,

$$\hat{\beta}_M = \arg\min_\beta \left(\frac{1}{n}\sum_{j=1}^n g(\xi_j,\beta)\right)' W \left(\frac{1}{n}\sum_{j=1}^n g(\xi_j,\beta)\right)$$

$$\beta_M^* = \arg\min_\beta \mathbb{E}\left[g(\xi_j,\beta)\right]' W \mathbb{E}\left[g(\xi_j,\beta)\right]$$

$$\phi_M\left(\xi_j, x_i; \beta_M^*\right) = -\frac{\partial F(x_i,\beta_M^*)}{\partial\beta'}\left(G(\beta_M^*)' W G(\beta_M^*)\right)^{-1} G(\beta_M^*)' W g(\xi_j,\beta_M^*)$$

10. Maximum Likelihood Estimation: For known $F(\cdot,\cdot)$, $\hat{s}_M(x_i) = F(x_i,\hat{\beta}_M)$, $s_M^*(x_i) = F(x_i,\beta_M^*)$, and a density function $f(\cdot,\cdot)$ which might not be the true density of ξ_i,

$$\hat{\beta}_M = \arg\max_\beta \frac{1}{n}\sum_{j=1}^n \log f(\xi_j,\beta)$$

$$\beta_M^* = \arg\max_\beta \mathbb{E}\left[\log f(\xi_j,\beta)\right]$$

$$\phi_M\left(\xi_j, x_i; \beta_M^*\right) = -\frac{\partial F(x_i,\beta_M^*)}{\partial\beta'}\mathbb{E}\left[\frac{\partial^2 \log f(\xi_j,\beta_M^*)}{\partial\beta\partial\beta'}\right]^{-1}\frac{\partial \log f(\xi_j,\beta_M^*)}{\partial\beta}$$

Remark 3.1 *We would like to note that several other well-known machine learning estimators besides Ridge and ℓ_2-norm SVM regression also have an influence function representation. For example, (Wager & Athey, 2018) derived the asymptotic normality of honest Random Forest estimators using Hajek projections, which indicates the existence of an influence function representation (see Lemma 3.3 in (Wager & Athey, 2018) for details). Similarly, (White, 1989), (Chen & White, 1999), and (White & Racine, 2001) demonstrated asymptotic normality of single-hidden layer Neural Networks by formulating them as nonlinear least squares estimators $\hat{\beta}_n = \arg\min \frac{1}{n}\sum_{i=1}^n \left(y_i - \Lambda\left(\sum_{j=1}^q \psi\left(x_i'\gamma_j\right)\theta_j\right)\right)^2$, where $\Lambda(\cdot)$ is a smooth increasing output function, $\psi(\cdot)$ is a smooth increasing activation function such as the sigmoidal function and $\beta \equiv (\gamma', \theta')$ are the network parameters such as the weights on the inputs and hidden units. The influence function is given by example 2 above after replacing $F(x_i,\beta)$ with $\Lambda\left(\sum_{j=1}^q \psi\left(x_i'\gamma_j\right)\theta_j\right)$.*

However, there are also estimators that do not have an influence function representation generally. For example, Lasso typically does not have an influence function representation unless additional assumptions are imposed to guarantee asymptotic normality. In the finite-dimensional case, if $\sqrt{n}\lambda_n \to 0$, then $\hat{\beta}_M = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\beta)^2 + \lambda_n\|\beta\|_1$ will have an influence function representation equal to that for least squares because Lasso's non-standard asymptotic distribution will reduce down to a normal distribution. For more details, see the last line of the proof of Theorem 2 of (Knight & Fu, 2000).

3.1 Asymptotic Normality of K-Fold Cross Validation Criterion

In this section, we derive the asymptotic normality of $\sqrt{n}$ times the difference between the K-Fold cross validation criterion and the expected out-of-sample error for each estimator M

in a set of candidate estimator $\mathcal{M}$. Recall that $n_t \equiv \frac{K-1}{K}n$ is the number of observations in each training fold. We will use τ_{n_t} to denote the analog of τ_n evaluated using n_t observations.

Theorem 3.1 *Asymptotic Normality: Suppose Assumptions 3.1-3.3 are satisfied. Define for each estimator $M \in \mathcal{M}$*

$$\begin{aligned} \psi(\xi_j, \xi_i) &\equiv \phi_M\left(\xi_j, x_i; s_M^*\right) \frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} \\ \sigma_\gamma^2 &\equiv Var\left[\gamma(s_M^*; \xi_i)\right] \\ \sigma_{01,\psi}^2 &\equiv Var\left[\mathbb{E}\left[\psi(\xi_j, \xi_i) \middle| \xi_j\right]\right] \\ \lambda &\equiv Cov\left[\gamma(s_M^*; \xi_j), \mathbb{E}\left[\psi(\xi_j, \xi_i) | \xi_j\right]\right] \end{aligned}$$

As $n \to \infty$ while keeping K fixed,

$$\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M) - EPE_{OUT,n}(M)\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

$$\sigma^2 = \begin{cases} \sigma_\gamma^2 & , \tau_n \ll \sqrt{n} \\ \sigma_\gamma^2 + \sigma_{01,\psi}^2 + 2\lambda & , \tau_n = \sqrt{n} \end{cases}$$

Remark 3.2 *Note that for estimators that converge at a slower than $\sqrt{n}$ rate (i.e. nonparametric estimators), the K-fold cross validation error still converges at the $\sqrt{n}$ rate because the residual variance $\mathbb{E}[\gamma(s_M^*; \xi_i)]$ component of the expected out-of-sample error can be consistently estimated at the $\sqrt{n}$ rate. This result for nonparametric estimators is consistent with proposition 1 of (Wager, 2020).*

Remark 3.3 *Note that the rate of convergence and asymptotic variance of the K-fold cross validation error do not depend on the number of folds K, assuming that K is fixed. This suggests that the choice of which fixed K to use is not first-order important in large samples.*

3.2 Consistent Estimate of Asymptotic Variance

The benefit of deriving the asymptotic distribution of the K-fold cross validation error is that we can use a consistent estimate of the asymptotic variance to construct confidence intervals for the expected out-of-sample error. For nonparametric estimators, we will need to compute our estimator once at each data point $\hat{s}_M(x_i)$ for $i = 1...n$ when forming our estimate of $\hat{\sigma}_\gamma^2$. For parametric estimators, we will need to compute $\hat{\beta}_M$ once and then estimate the influence functions $\hat{\phi}_M\left(\xi_j, x_i; \hat{s}_M\right)$ and first derivatives of $\gamma\left(\hat{s}_M; \xi_i\right)$ at each data point. The following theorem demonstrates consistency of our estimate of the asymptotic variance.

Theorem 3.2 *Consistent Estimate of Asymptotic Variance: Define*

$$\hat{\sigma}_n^2 = \begin{cases} \hat{\sigma}_\gamma^2 & , \tau_n \ll \sqrt{n} \\ \hat{\sigma}_\gamma^2 + \hat{\sigma}_{01,\psi}^2 + 2\hat{\lambda} & , \tau_n = \sqrt{n} \end{cases}$$

where

$$\hat{\psi}\left(\xi_j, \xi_i; \hat{s}_M\right) = \hat{\phi}_M\left(\xi_j, x_i; \hat{s}_M\right) \frac{\partial \gamma\left(\hat{s}_M; \xi_i\right)}{\partial \hat{s}_M}$$

$$\hat{\sigma}_{\gamma}^{2} = \frac{1}{n}\sum_{i=1}^{n}\left(\gamma(\hat{s}_M;\xi_i) - \frac{1}{n}\sum_{k=1}^{n}\gamma(\hat{s}_M;\xi_k)\right)^2$$

$$\hat{\sigma}_{01,\psi}^{2} = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{h}_M(\xi_j) - \frac{1}{n}\sum_{k=1}^{n}\hat{h}_M(\xi_k)\right)^2$$

$$\hat{h}_M(\xi_j) = \frac{1}{n-1}\sum_{\{i|1\le i\le n, i\neq j\}}\hat{\psi}\left(\xi_j,\xi_i;\hat{s}_M\right)$$

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{h}_M(\xi_i) - \frac{1}{n}\sum_{k=1}^{n}\hat{h}_M(\xi_k)\right)\left(\gamma(\hat{s}_M;\xi_i) - \frac{1}{n}\sum_{k=1}^{n}\gamma(\hat{s}_M;\xi_k)\right)$$

Suppose Assumptions 3.1-3.3 are satisfied in addition to

$$\frac{1}{n}\sum_{j=1}^{n}\left|\frac{1}{n-1}\sum_{\{i|1\le i\le n, i\neq j\}}\hat{\psi}\left(\xi_j,\xi_i;\hat{s}_M\right) - \frac{1}{n-1}\sum_{\{i|1\le i\le n, i\neq j\}}\psi\left(\xi_j,\xi_i;s_M^*\right)\right|^2 \xrightarrow{p} 0$$

$$\frac{1}{n}\sum_{i=1}^{n}\left|\gamma\left(\hat{s}_M;\xi_i\right) - \gamma\left(s_M^*;\xi_i\right)\right|^2 \xrightarrow{p} 0$$

Then $\hat{\sigma}_n^2$ converges in probability to σ^2.

The additional two assumptions at the end are necessary to ensure that using the estimated influence function $\hat{\psi}\left(\cdot,\cdot;\cdot\right)$ and estimated feature $\hat{s}_M$ instead of their probability limits does not impact the consistency of our estimates $\hat{\sigma}_{\gamma}^2$, $\hat{\sigma}_{01,\psi}^2$, and $\hat{\lambda}$. Similar assumptions are assumed in Lemma 8.3 of (Newey & McFadden, 1994) when showing consistency of estimators of the asymptotic variance of two step estimators.

4. Estimator Selection Test

We now formulate a hypothesis test to compare the predictive performance of two estimators using their expected out-of-sample errors $EPE_{OUT,n}(M) \equiv \mathbb{E}\left[\gamma(\hat{s}_M;\tilde{\xi}_i)\right]$, where the expectation is taken with respect to both the data used to estimate $\hat{s}_M$ and the new observation $\tilde{\xi}_i$. The two estimators could be estimating either the same feature $s_{M_1}^* = s_{M_2}^*$ or different features $s_{M_1}^* \neq s_{M_2}^*$. The null hypothesis states that the expected out-of-sample errors of the two estimators are the same and the alternatives are that one is higher than the other:

$$\begin{aligned}
H_0 &: EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1) = 0\\
H_1 &: EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1) > 0\\
H_2 &: EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1) < 0
\end{aligned}$$

4.1 Asymptotic Distribution of Test Statistic in the Absence of First Order Degeneracy

When the difference between the expected out-of-sample errors of the two estimators can be consistently estimated at the $\sqrt{n}$ rate, we can use the following test statistic:

$$T_n = \frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1)\right)$$

$\hat{\tilde{\sigma}}_n^2$ is a consistent estimate of the difference in the K-fold cross validation errors' asymptotic variances, which differs for parametric and nonparametric estimators:

$$\tilde{\sigma}^2 = \begin{cases} \tilde{\sigma}_\gamma^2 + \sigma_{01,\psi_{M_1}}^2 + 2\lambda_{M_1} & ,\tau_{n,2} \ll \tau_{n,1} = \sqrt{n} \\ \tilde{\sigma}_\gamma^2 + \sigma_{01,\psi_{M_2}}^2 + 2\lambda_{M_2} & ,\tau_{n,1} \ll \tau_{n,2} = \sqrt{n} \\ \tilde{\sigma}_\gamma^2 & ,\tau_{n,2}, \tau_{n,1} \ll \sqrt{n} \\ \tilde{\sigma}_\gamma^2 + \tilde{\sigma}_{01,\psi}^2 + 2\tilde{\lambda} & ,\tau_{n,2} = \tau_{n,1} = \sqrt{n} \end{cases}$$

$$\tilde{\psi}(\xi_j, \xi_i) \equiv \left(\phi_{M_2}\left(\xi_j, x_i; s_{M_2}^*\right)\frac{\partial\gamma\left(s_{M_2}^*; \xi_i\right)}{\partial s_{M_2}^*} - \phi_{M_1}\left(\xi_j, x_i; s_{M_1}^*\right)\frac{\partial\gamma\left(s_{M_1}^*; \xi_i\right)}{\partial s_{M_1}^*}\right)$$

$$\tilde{\gamma}(\xi_i) \equiv \gamma\left(s_{M_2}^*; \xi_i\right) - \gamma\left(s_{M_1}^*; \xi_i\right), \quad \tilde{\sigma}_\gamma^2 \equiv Var\left[\tilde{\gamma}(\xi_i)\right]$$

$$\tilde{\sigma}_{01,\psi}^2 \equiv Var\left[\mathbb{E}\left[\tilde{\psi}(\xi_j, \xi_i)\middle|\xi_j\right]\right], \quad \tilde{\lambda} \equiv Cov\left[\tilde{\gamma}(\xi_j), \mathbb{E}\left[\tilde{\psi}(\xi_j, \xi_i)|\xi_j\right]\right]$$

The following theorem shows that when $0 < \tilde{\sigma}^2 < \infty$, T_n has a standard normal asymptotic distribution under H_0. It follows that the test which rejects H_0 when $|T_n| > Z_{1-\alpha/2}$, where Z_α is the α-th percentile of the standard normal, is pointwise consistent in level.

Theorem 4.1 *Suppose Assumptions 3.1-3.3 are satisfied for M_1 and M_2, and $0 < \tilde{\sigma}^2 < \infty$. Then $T_n \xrightarrow{d} N(0,1)$ under H_0 and under alternatives of the form $\sqrt{n}\left(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)\right) \to 0$. $T_n \xrightarrow{d} N(c/\tilde{\sigma}, 1)$ under alternatives of the form $\sqrt{n}\left(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)\right) \to c$, for some constant c. $T_n \xrightarrow{p} \infty$ under alternatives of the form $\sqrt{n}\left(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)\right) \to \infty$. $T_n \xrightarrow{p} -\infty$ under alternatives of the form $\sqrt{n}\left(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)\right) \to -\infty$.*

4.2 Consistent Estimation of Asymptotic Variance

We now provide a consistent estimate of $\tilde{\sigma}^2$.

Theorem 4.2 *Consistent Estimate of Asymptotic Variance: Define*

$$\hat{\tilde{\sigma}}_n^2 = \begin{cases} \hat{\tilde{\sigma}}_\gamma^2 + \hat{\sigma}_{01,\psi_{M_1}}^2 + 2\hat{\lambda}_{M_1} & ,\tau_{n,2} \ll \tau_{n,1} = \sqrt{n} \\ \hat{\tilde{\sigma}}_\gamma^2 + \hat{\sigma}_{01,\psi_{M_2}}^2 + 2\hat{\lambda}_{M_2} & ,\tau_{n,1} \ll \tau_{n,2} = \sqrt{n} \\ \hat{\tilde{\sigma}}_\gamma^2 & ,\tau_{n,2}, \tau_{n,1} \ll \sqrt{n} \\ \hat{\tilde{\sigma}}_\gamma^2 + \hat{\tilde{\sigma}}_{01,\psi}^2 + 2\hat{\tilde{\lambda}} & ,\tau_{n,2} = \tau_{n,1} = \sqrt{n} \end{cases}$$

where

$$\hat{\hat{\sigma}}_{\gamma}^{2} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\hat{\gamma}}(\xi_i) - \frac{1}{n}\sum_{k=1}^{n}\hat{\hat{\gamma}}(\xi_k)\right)^2$$

$$\hat{\hat{\gamma}}(\xi_i) \equiv \gamma\left(\hat{s}_{M_2};\xi_i\right) - \gamma\left(\hat{s}_{M_1};\xi_i\right)$$

$$\hat{\hat{\psi}}\left(\xi_j,\xi_i;\hat{s}_{M_1},\hat{s}_{M_2}\right) = \hat{\phi}_{M_2}\left(\xi_j,x_i;\hat{s}_{M_2}\right)\frac{\partial\gamma\left(\hat{s}_{M_2};\xi_i\right)}{\partial\hat{s}_{M_2}} - \hat{\phi}_{M_1}\left(\xi_j,x_i;\hat{s}_{M_1}\right)\frac{\partial\gamma\left(\hat{s}_{M_1};\xi_i\right)}{\partial\hat{s}_{M_1}}$$

$$\hat{\hat{\sigma}}_{01}^{2} = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{\hat{h}}(\xi_j) - \frac{1}{n}\sum_{k=1}^{n}\hat{\hat{h}}(\xi_k)\right)^2$$

$$\hat{\hat{h}}(\xi_j) = \frac{1}{n-1}\sum_{\{i|1\leq i\leq n, i\neq j\}}\hat{\hat{\psi}}\left(\xi_j,\xi_i;\hat{s}_{M_1},\hat{s}_{M_2}\right)$$

$$\hat{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\hat{\gamma}}(\xi_i) - \frac{1}{n}\sum_{k=1}^{n}\hat{\hat{\gamma}}(\xi_k)\right)\left(\hat{\hat{h}}(\xi_i) - \frac{1}{n}\sum_{k=1}^{n}\hat{\hat{h}}(\xi_k)\right)$$

If all of the conditions in theorem 3.2 are satisfied for M_1 and M_2, then $\hat{\hat{\sigma}}_n^2$ converges in probability to $\tilde{\sigma}^2$.

4.3 Testing Procedure Accounting for First Order Degeneracy

If $\tilde{\sigma}^2 = 0$, which can arise for example when two nonparametric estimators are estimating the same conditional mean function in which case $Var\left[\tilde{\gamma}\left(\xi\right)\right] = 0$, then the asymptotic distribution of T_n under H_0 becomes degenerate. In order to have a nondegenerate test statistic, we need to scale by n instead of $\sqrt{n}$ (see the last part of the proof of Theorem 4.1 for an explanation of this result). Although we do not explicitly characterize the asymptotic distribution of our test statistic under first order degeneracy, we can consistently estimate this distribution using subsampling. If we knew first order degeneracy were present, we could use $n\left(\hat{\mathcal{L}}_n^{CV}\left(M_2\right) - \hat{\mathcal{L}}_n^{CV}\left(M_1\right)\right)$ as the test statistic. However, it is often very difficult to know if first order degeneracy is present, so instead, we first obtain an estimate $n^{\hat{\delta}}$ of the rate of convergence of $\hat{\mathcal{L}}_n^{CV}\left(M_2\right) - \hat{\mathcal{L}}_n^{CV}\left(M_1\right)$ using the following algorithm (described in Theorem 8.2.2 in (Politis, Romano, & Wolf, 1999)).

1. For $r = 1...R$ replications, draw I different subsamples of sizes b_i for $i = 1...I$ where $b_i \to \infty$ and $b_i/n \to 0$ and compute $\Delta_{b_i,r}^{*} \equiv \hat{\mathcal{L}}_{b,i}^{CV,*}\left(M_2\right) - \hat{\mathcal{L}}_{b,i}^{CV,*}\left(M_1\right) - \left(\hat{\mathcal{L}}_n^{CV}\left(M_2\right) - \hat{\mathcal{L}}_n^{CV}\left(M_1\right)\right)$.

2. For each $i = 1...I$ and $j = 1...J$, define $c_{\tau_{ij}}$ and $c_{\rho_{ij}}$ as the τ_jth and ρ_jth percentiles of $\Delta_{b_i,r}^{*}$ for $i = 1...I$, where τ_j and ρ_j are the jth elements of

   $\tau = \left\{70, 70 + \frac{20}{J-1}, 70 + \frac{40}{J-1}, ..., 90\right\}$ and $\rho = \left\{10, 10 + \frac{20}{J-1}, 10 + \frac{40}{J-1}, ..., 30\right\}$.

3. Define $y_{ij} = log\left(c_{\tau_{ij}} - c_{\rho_{ij}}\right)$, $\bar{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_{ij}$, $\bar{y} = \frac{1}{I}\sum_{i=1}^{I}\bar{y}_i$, $\overline{log\left(b\right)} = \frac{1}{I}\sum_{i=1}^{I} log\left(b_i\right)$.

4. The estimated rate of convergence is $n^{\hat{\delta}}$ for

$$\hat{\delta} \equiv -\frac{\sum_{i=1}^{I} \left(\bar{y}_i - \bar{y}\right) \left(log\left(b_i\right) - \overline{log\left(b\right)}\right)}{\sum_{i=1}^{I} \left(log\left(b_i\right) - \overline{log\left(b\right)}\right)^2}$$

We then use the test statistic $\tilde{T}_n \equiv n^{\hat{\delta}} \left(\hat{\mathcal{L}}_n^{CV}\left(M_2\right) - \hat{\mathcal{L}}_n^{CV}\left(M_1\right)\right)$, and the percentiles of $\tilde{\hat{L}}_{n,b}\left(x\right) = \frac{1}{B}\sum_{i=1}^{B} 1\left(b^{\hat{\delta}}\left(\hat{\mathcal{L}}_{b,i}^{CV,*}\left(M_2\right) - \hat{\mathcal{L}}_{b,i}^{CV,*}\left(M_1\right) - \left(\hat{\mathcal{L}}_n^{CV}\left(M_2\right) - \hat{\mathcal{L}}_n^{CV}\left(M_1\right)\right)\right) \leq x\right)$ as the critical values, where the subsample size b satisfies $b \to \infty$ and $b/n \to 0$, and $i = 1...B$ indexes a random sample of the $\binom{n}{b}$ possible subsets of size b.

This procedure requires computing the difference in the cross validation errors $(I+1)B+1$ times, where I is the number of different subsample sizes we use to estimate the rate of convergence δ using the procedure in Theorem 8.2.2 of (Politis et al., 1999). We acknowledge that the computational cost of this procedure can be high due to the repeated computation of the cross validation errors, but we cannot avoid this cost unless we knew what the rate of convergence of our test statistic is, which is difficult in practice. The next theorem states that our test using the subsampling procedure is pointwise consistent in level.

Theorem 4.3 *Pointwise Consistency in Level: Suppose Assumptions 3.1-3.3 are satisfied. Let $\tilde{c}_{\alpha}^{*}$ be the α-th percentile of $\tilde{\hat{L}}_{n,b}\left(x\right)$. For $\tilde{\phi}_n = 1\left(\tilde{T}_n > \tilde{c}_{1-\alpha/2}^{*} \bigcup \tilde{T}_n < \tilde{c}_{\alpha/2}^{*}\right)$,*

$$\limsup_{n\to\infty} \mathbb{E}_{H_0}\left[\tilde{\phi}_n\right] \leq \alpha$$

Proof: Result follows from the consistency of subsampling for statistics with an estimated rate of convergence (e.g. Theorem 8.3.1 in (Politis et al., 1999)).

Remark 4.1 *In order to illustrate when first order degeneracy may arise, take the simple example of linear regression with nested regressors:*

$$\begin{aligned} M_1 :& Y_i = \alpha_1 W_i + \eta_i \\ M_2 :& Y_i = \alpha_2 W_i + \beta_2 X_i + \epsilon_i \end{aligned}$$

Define $Z_i' = [W_i, X_i]$. The influence functions for the OLS estimates are

$$\begin{aligned} \phi_{M_1}\left(\xi_j, W_i; \alpha_1\right) &= W_i' \mathbb{E}\left[W_j W_j'\right]^{-1} W_j \left(Y_j - \alpha_1 W_j\right) \\ \phi_{M_2}\left(\xi_j, Z_i; \alpha_2, \beta_2\right) &= Z_i' \mathbb{E}\left[Z_j Z_j'\right]^{-1} Z_j \left(Y_j - \alpha_2 W_j - \beta_2 X_j\right) \end{aligned}$$

The loss function is squared error loss.

$$\begin{aligned}
\gamma\left(s^*_{M_1};\xi_i\right) &= \left(Y_i-\alpha_1 W_i\right)^2\\
\gamma\left(s^*_{M_2};\xi_i\right) &= \left(Y_i-\alpha_2 W_i-\beta_2 X_i\right)^2\\
\tilde{\psi}(\xi_j,\xi_i) &\equiv \left(\phi_{M_2}\left(\xi_j,Z_i;\alpha_2,\beta_2\right)'\frac{\partial\gamma\left(s^*_{M_2};\xi_i\right)}{\partial s^*_{M_2}}-\phi_{M_1}\left(\xi_j,W_i;\alpha_1\right)'\frac{\partial\gamma\left(s^*_{M_1};\xi_i\right)}{\partial s^*_{M_1}}\right)\\
&=-2\left(Y_j-\alpha_2W_j-\beta_2X_j\right)Z_j'\mathbb{E}\left[Z_jZ_j'\right]^{-1}Z_i\left(Y_i-\alpha_2W_i-\beta_2X_i\right)\\
&\quad+2\left(Y_j-\alpha_1W_j\right)W_j'\mathbb{E}\left[W_jW_j'\right]^{-1}W_i\left(Y_i-\alpha_1W_i\right)\\
\mathbb{E}\left[\tilde{\psi}\left(\xi_j,\xi_i\right)\middle|\,\xi_j\right]&=-2\left(Y_j-\alpha_2W_j-\beta_2X_j\right)Z_j'\mathbb{E}\left[Z_jZ_j'\right]^{-1}\mathbb{E}\left[Z_i\left(Y_i-\alpha_2W_i-\beta_2X_i\right)\right]\\
&\quad+2\left(Y_j-\alpha_1W_j\right)W_j'\mathbb{E}\left[W_jW_j'\right]^{-1}\mathbb{E}\left[W_i\left(Y_i-\alpha_1W_i\right)\right]
\end{aligned}$$

If it were the case that $\mathbb{E}\left[Z_i\left(Y_i-\alpha_2W_i-\beta_2X_i\right)\right]=0$, $\mathbb{E}\left[W_i\left(Y_i-\alpha_1W_i\right)\right]=0$, *and* $Var\left[\epsilon_i^2-\eta_i^2\right]=0$, *then we have first order degeneracy since* $\tilde{\sigma}^2\equiv\tilde{\sigma}^2_{\gamma}+\tilde{\sigma}^2_{01,\psi}+2\tilde{\lambda}=0$. *This can happen if model 1 were the true model with* $\mathbb{E}\left[W_i\left(Y_i-\alpha_1W_i\right)\right]=0$, *which would imply that* $\beta_2=0$.

Remark 4.2 *For multiple model comparisons, if we are interested in constructing a data-dependent set of estimators that contain the best estimator(s) with a certain probability, we can construct a confidence set using a procedure similar to Definition 2 of (Hansen, Lunde, & Nason, 2011) which does not require a benchmark model to be selected as in (White, 2000). To construct the model confidence set, we first test the null hypotheses* $H_{0,\mathcal{M}}: EPE_{OUT,n}\left(M_i\right)-EPE_{OUT,n}\left(M_j\right)=0$ *for all* $M_i,M_j\in\mathcal{M}$ *using the test statistic* $T_{\mathcal{M},n}=\max_{i,j}\left|\hat{\mathcal{L}}_n^{CV}\left(M_i\right)-\hat{\mathcal{L}}_n^{CV}\left(M_j\right)\right|$, *which is the maximum of the absolute difference of the K-fold cross validation errors for all pairs of estimators in* $\mathcal{M}$. *If* $H_{0,\mathcal{M}}$ *is rejected, an elimination rule is used to eliminate from* $\mathcal{M}$ *an estimator with worst performance. We repeat the test using the smaller model set until it is accepted and the model confidence set is defined as the set of remaining models. Since the asymptotics of the test statistic* $T_{\mathcal{M},n}$ *are nonstandard, we will have to resort to subsampling or other consistent resampling methods to obtain critical values. If we are instead interested in obtaining pairwise rankings, then we can use the multiple testing procedures discussed in (Lehmann & Romano, 2005) and (Romano, Shaikh, & Wolf, 2010) and references therein. For example, we can use the Bonferroni correction, as we do in our empirical application.*

5. Monte Carlo

In this section, we use Monte Carlo simulations to study the coverage of our K-fold cross validation confidence intervals and the rejection frequencies of our estimator selection test.

5.1 Confidence Intervals for expected out-of-sample error

We demonstrate how we can use the asymptotic distribution of the K-fold cross validation error to construct confidence intervals for the expected out-of-sample error. We investigate the empirical coverage frequencies for these confidence intervals for three data generating processes (DGPs).

5.1.1 Linear DGP

For each Monte Carlo simulation $r = 1...R$, we generate the test data $(\tilde{x}_{ri}, \tilde{z}_{ri}, \tilde{y}_{ri})_{i=1}^{n_{Test}}$ and the training data $(x_{ri}, z_{ri}, y_{ri})_{i=1}^{n}$ independently of each other using the linear DGP:

$$y_i = \beta_0 x_i + z_i' \gamma_0 + \epsilon_i, \quad \epsilon_i \sim N(0,1), \quad \epsilon_i \perp x_i, z_i$$

$$\begin{pmatrix} x_i \\ z_i \end{pmatrix} \sim MVN\left(\begin{pmatrix} 1 \\ \mu \end{pmatrix}, 0.5 \left(I_{10} + \iota_{10} \iota_{10}' \right) \right)$$

where $\mu' = \begin{pmatrix} 2 & 3 & 4 & 5 & 4 & 5 & 6 & 7 & 8 \end{pmatrix}$, I_{10} is the 10×10 identity matrix, and ι_{10} is the 10×1 vector of all ones.
Define $w_i' \equiv [x_i, z_i']$. Our candidate models include linear models for different values of $p \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$:

$$y_i = \sum_{j=1}^{p} \theta_j w_{ij} + \eta_i$$

We also look at two additional linear models with group effects α_g, which are 50 dummy variables generated independently of ϵ, x, and z.

$$y_i = \theta_1 x_i + \alpha_g + \nu_i, \quad y_i = w_i' \theta + \alpha_g + \zeta_i$$

All of our linear models are estimated using OLS. In addition, we also consider a univariate nonparametric model that is estimated using kernel (local-constant) regression of y on x using a Gaussian kernel $K_{h_n}(x) = K(x/h_n)$, $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$, and bandwidth $h_n = (4/3)^{1/5} n^{-1/5}$. The 5-fold cross validation error using squared error loss is

$$\hat{\mathcal{L}}_n^{CV} = \frac{1}{5} \sum_{k=1}^{5} \frac{1}{n_v} \sum_{i \in I_k^{(v)}} \left(y_i - \hat{s}_{n_t}^{(-k)}(w_i) \right)^2$$

where $\hat{s}_{n_t}^{(-k)}$ is the estimate of the conditional mean of y given w computed using the n_t training observations, which are the observations not in the k-th fold. For OLS, $\hat{s}_{n_t}^{(-k)}(w_i) = w_i' \left(\frac{1}{n_t} \sum_{j \notin I_k^{(v)}} w_j w_j' \right)^{-1} \left(\frac{1}{n_t} \sum_{j \notin I_k^{(v)}} w_j y_j \right)$. For the kernel estimator, $\hat{s}_{n_t}^{(-k)}(x_i) = \frac{\sum_{j \notin I_k^{(v)}} K_{h_{n_t}}(x_i - x_j) y_j}{\sum_{j \notin I_k^{(v)}} K_{h_{n_t}}(x_i - x_j)}$.

We estimate the expected out-of-sample error using the test error averaged across R Monte Carlo simulations:

$$\widehat{EPE}_{OUT,n} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{n_{Test}} \sum_{i=1}^{n_{Test}} \left(\tilde{y}_{ri} - \hat{s}_{n,r}(\tilde{w}_{ri}) \right)^2$$

where $\hat{s}_{n,r}$ is the estimate of the conditional mean of y given w computed using the training data on the r-th simulation. For OLS, $\hat{s}_{n,r}(\tilde{w}_{ri}) = \tilde{w}_{ri}' \left(\frac{1}{n} \sum_{j=1}^{n} w_{rj} w_{rj}' \right)^{-1} \left(\frac{1}{n} \sum_{j=1}^{n} w_{rj} y_{rj} \right)$. For the kernel estimator, $\hat{s}_{n,r}(\tilde{x}_{ri}) = \frac{\sum_{j=1}^{n} K_{h_n}(\tilde{x}_{ri} - x_{rj}) y_j}{\sum_{j=1}^{n} K_{h_n}(\tilde{x}_{ri} - x_{rj})}$.

Table 1 shows the empirical coverage frequencies for $\widehat{EPE}_{OUT,n}$ of the nominal 95% equal-tailed two-sided confidence intervals $\left[\hat{\mathcal{L}}_n^{CV} \pm 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$, where $\hat{\sigma}_n$ is given in Theorem 3.2. Note that we do not need to compute the influence function for our nonparametric estimator because $\hat{\sigma}_n^2 = \hat{\sigma}_\gamma^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{\gamma}(\xi_i) - \frac{1}{n}\sum_{k=1}^n \hat{\gamma}(\xi_k)\right)^2$, where $\hat{\gamma}(\xi_i) = (y_i - \hat{s}_n(x_i))^2$. For the OLS estimators, we use $\hat{\phi}_M\left(\xi_j, w_i; \hat{\theta}\right) = w_i'\left(\frac{1}{n}\sum_{j=1}^n w_j w_j'\right)^{-1} w_j \left(y_j - w_j'\hat{\theta}\right)$ as our estimated influence functions. We consider three different values for γ_0 while keeping β_0 at 0.5. p refers to the number of right hand size variables, so the $p = 51$ and $p = 60$ columns refer to the models with group effects.

p	1	2	3	4	5	6	7	8	9	10	51	60	kernel
$\gamma_{0i} = 0.5$	0.950	0.948	0.952	0.951	0.954	0.949	0.947	0.947	0.959	0.956	0.950	0.954	0.948
$\gamma_{0i} = \frac{1}{\sqrt{n}}$	0.956	0.950	0.956	0.951	0.952	0.952	0.954	0.956	0.957	0.958	0.954	0.956	0.948
$\gamma_{0i} = \frac{1}{n}$	0.958	0.956	0.956	0.958	0.957	0.957	0.958	0.958	0.959	0.958	0.952	0.949	0.953

Table 1: Empirical Coverage Frequencies, Linear DGP, $n = 2000$, $n_{Test} = 4000$, $R = 2000$

The empirical coverage frequencies for all estimators are close to the nominal level, which supports the use of our asymptotic theory to construct asymptotically valid confidence intervals for the expected out-of-sample error. Note that the probability limits of our estimators $\hat{s}_M$, except for the one with $p = 10$, are all different from the true conditional mean function of the underlying data generating process. We do not require our estimators to be consistent for some true feature of the underlying DGP when deriving the asymptotic distribution of the K-fold cross validation error.

5.1.2 Nonlinear DGP

We now repeat the same exercise as in the main text, but with a nonlinear DGP:

$$y_i = \frac{\exp(\beta_0 x_i)}{1 + \exp(\beta_0 x_i)} + z_i'\gamma_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2), \quad \epsilon_i \perp x_i, z_i$$

Table 2 shows the empirical coverage frequencies for $\widehat{EPE}_{OUT,n}$ of the nominal 95% equal-tailed two-sided confidence intervals $\left[\hat{\mathcal{L}}_n^{CV} \pm 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$, where $\hat{\sigma}_n$ is given in Theorem 3.2. We see again that the coverage is close to the nominal level, even though our estimators are not consistent for the true underlying conditional mean function.

p	1	2	3	4	5	6	7	8	9	10	51	60	kernel
$\gamma_{0i} = 0.5$	0.950	0.945	0.952	0.952	0.951	0.949	0.949	0.939	0.959	0.951	0.948	0.951	0.946
$\gamma_{0i} = \frac{1}{\sqrt{n}}$	0.943	0.949	0.953	0.953	0.951	0.954	0.956	0.952	0.955	0.953	0.946	0.948	0.951
$\gamma_{0i} = \frac{1}{n}$	0.952	0.949	0.952	0.950	0.953	0.952	0.951	0.954	0.952	0.953	0.953	0.954	0.947

Table 2: Empirical Coverage Frequencies, Nonlinear DGP, $n = 2000$, $n_{Test} = 4000$, $R = 2000$

5.1.3 Gaussian Mixture Model

We generate data according to the following model: for $x \sim N(1,1)$,

$$y_1|x \sim N\left(x'\mu_1, \sigma_1\right), y_2|x \sim N\left(x'\mu_2, \sigma_2\right), d|x \sim Bernoulli\left(\pi\right), y = \left(1-d\right)y_1 + dy_2$$

For $\beta_0 = (\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$, the density of Y is

$$f\left(y|x;\beta_0\right) = \left(1-\pi\right)\frac{1}{\sigma_1}\phi\left(\frac{y-x'\mu_1}{\sigma_1}\right) + \pi\frac{1}{\sigma_2}\phi\left(\frac{y-x'\mu_2}{\sigma_2}\right)$$

and the conditional mean function is $F\left(x,\beta_0\right) \equiv \mathbb{E}\left[y|x\right] = \left(1-\pi\right)x'\mu_1 + \pi x'\mu_2$. Using the information matrix equality, the influence function is given by

$$\begin{aligned} \phi_M\left(\xi_j, x_i;\beta_0\right) &= -\frac{\partial F\left(x_i,\beta_0\right)}{\partial\beta'}\mathbb{E}\left[\frac{\partial^2\log f\left(y_j|x_j;\beta_0\right)}{\partial\beta\partial\beta'}\right]^{-1}\frac{\partial\log f\left(y_j|x_j;\beta_0\right)}{\partial\beta} \\ &= \frac{\partial F\left(x_i,\beta_0\right)}{\partial\beta'}\mathbb{E}\left[\frac{\partial\log f\left(y_j|x_j;\beta_0\right)}{\partial\beta}\frac{\partial\log f\left(y_j|x_j;\beta_0\right)}{\partial\beta'}\right]^{-1}\frac{\partial\log f\left(y_j|x_j;\beta_0\right)}{\partial\beta} \end{aligned}$$

$$\frac{\partial F\left(x_i,\beta_0\right)}{\partial\beta'} = \left[\begin{array}{ccccc} x'\left(\mu_2-\mu_1\right) & \left(1-\pi\right)x' & \pi x' & 0 & 0 \end{array}\right]$$

$$\begin{aligned} \frac{\partial\log f\left(y|x;\beta_0\right)}{\partial\beta} &= \left[\begin{array}{c} \frac{1}{f(y|x;\beta_0)}\left(\frac{1}{\sigma_2}\phi\left(\frac{y-x'\mu_2}{\sigma_2}\right) - \frac{1}{\sigma_1}\phi\left(\frac{y-x'\mu_1}{\sigma_1}\right)\right) \\ -\frac{1}{f(y|x;\beta_0)}\left(1-\pi\right)\frac{x}{\sigma_1^2}\phi'\left(\frac{y-x'\mu_1}{\sigma_1}\right) \\ -\frac{1}{f(y|x;\beta_0)}\pi\frac{x}{\sigma_2^2}\phi'\left(\frac{y-x'\mu_2}{\sigma_2}\right) \\ -\frac{1}{f(y|x;\beta_0)}\left(1-\pi\right)\left\{\left(\frac{y-x'\mu_1}{\sigma_1^3}\right)\phi'\left(\frac{y-x'\mu_1}{\sigma_1}\right) + \frac{1}{\sigma_1^2}\phi\left(\frac{y-x'\mu_1}{\sigma_1}\right)\right\} \\ -\frac{1}{f(y|x;\beta_0)}\pi\left\{\left(\frac{y-x'\mu_2}{\sigma_2^3}\right)\phi'\left(\frac{y-x'\mu_2}{\sigma_2}\right) + \frac{1}{\sigma_2^2}\phi\left(\frac{y-x'\mu_2}{\sigma_2}\right)\right\} \end{array}\right] \\ &= \left[\begin{array}{c} \frac{1}{f(y|x;\beta_0)}\left(\frac{1}{\sigma_2}\phi\left(\frac{y-x'\mu_2}{\sigma_2}\right) - \frac{1}{\sigma_1}\phi\left(\frac{y-x'\mu_1}{\sigma_1}\right)\right) \\ \frac{1}{f(y|x;\beta_0)}\left(1-\pi\right)\frac{x}{\sigma_1^2}\left(\frac{y-x'\mu_1}{\sigma_1}\right)\phi\left(\frac{y-x'\mu_1}{\sigma_1}\right) \\ \frac{1}{f(y|x;\beta_0)}\pi\frac{x}{\sigma_2^2}\left(\frac{y-x'\mu_2}{\sigma_2}\right)\phi\left(\frac{y-x'\mu_2}{\sigma_2}\right) \\ \frac{1}{f(y|x;\beta_0)}\frac{1-\pi}{\sigma_1^2}\left(\left(\frac{y-x'\mu_1}{\sigma_1}\right)^2 - 1\right)\phi\left(\frac{y-x'\mu_1}{\sigma_1}\right) \\ \frac{1}{f(y|x;\beta_0)}\frac{\pi}{\sigma_2^2}\left(\left(\frac{y-x'\mu_2}{\sigma_2}\right)^2 - 1\right)\phi\left(\frac{y-x'\mu_2}{\sigma_2}\right) \end{array}\right] \end{aligned}$$

We set the true values of the mean parameters to $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$. We estimate 9 different Gaussian mixture regression models corresponding to $\pi \in \{0.1, 0.2, 0.3, ..., 0.9\}$. The K-fold cross validation error is

$$\hat{\mathcal{L}}_n^{CV} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_v}\sum_{i\in I_v^{(k)}}\left(y_i - F\left(x_i, \hat{\beta}^{(-k)}\right)\right)^2$$

We estimate the expected out-of-sample error using the test error averaged across R Monte Carlo simulations:

$$\widehat{EPE}_{OUT,n} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{Test}}\sum_{i=1}^{n_{Test}}\left(\tilde{y}_{ri} - F\left(\tilde{x}_{ri}, \hat{\beta}_{n,r}\right)\right)^2$$

We construct nominal 95% equal-tailed two-sided confidence intervals $\left[\hat{\mathcal{L}}_n^{CV} \pm 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$, where $\hat{\sigma}_n$ is given in Theorem 3.2, using two different ways of estimating the influence function. The first way uses the information matrix equality formulation of the influence function.

$$\hat{\phi}_M\left(\xi_j, x_i; \hat{\beta}\right) = \frac{\partial F\left(x_i, \hat{\beta}\right)}{\partial \beta'}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\partial \log f\left(y_j|x_j;\hat{\beta}\right)}{\partial \beta}\frac{\partial \log f\left(y_j|x_j;\hat{\beta}\right)}{\partial \beta'}\right)^{-1}\frac{\partial \log f\left(y_j|x_j;\hat{\beta}\right)}{\partial \beta}$$

The second way uses the estimated Hessian given as an output of the fmincon solver.

$$\hat{\phi}_M\left(\xi_j, x_i; \hat{\beta}\right) = -\frac{\partial F\left(x_i, \hat{\beta}\right)}{\partial \beta'}\hat{H}^{-1}\frac{\partial \log f\left(y_j|x_j;\hat{\beta}\right)}{\partial \beta}$$

The empirical coverage frequencies for $\widehat{EPE}_{OUT,n}$ are in Table 3, and we can see that they are all close to 95%.

	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$	$\pi = 0.4$	$\pi = 0.5$	$\pi = 0.6$	$\pi = 0.7$	$\pi = 0.8$	$\pi = 0.9$
Info. Matrix Eq.	0.948	0.951	0.950	0.952	0.954	0.952	0.951	0.950	0.946
Est. Hessian	0.948	0.951	0.949	0.952	0.954	0.952	0.950	0.950	0.946

Table 3: Empirical Coverage Frequencies, Gaussian Mixture Model DGP, $n = 2000$, $n_{Test} = 4000$, $R = 5000$

5.2 Estimator Selection Test

We examine the rejection frequencies of our estimator selection test for testing the equivalence of the expected out-of-sample errors. Consider the following simple data generating process:

$$Y_i = \alpha_0 W_i + \beta_0 X_i + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0,1), \quad \epsilon_i \perp \begin{pmatrix} W_i \\ X_i \end{pmatrix} \overset{i.i.d.}{\sim} N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}\right).$$

We compare two linear models estimated by OLS, one of which is nested inside the other:

$$\begin{aligned} M_1 &: Y_i = \alpha_1 W_i + \eta_i \\ M_2 &: Y_i = \alpha_2 W_i + \beta_2 X_i + \epsilon_i \end{aligned}$$

We examine the empirical frequencies of failing to reject the null that both models are equally good in terms of out-of-sample predictive accuracy, rejecting in favor of model 1, and rejecting in favor of model 2 under six different choices of $\beta_0 \in \left\{0, \frac{1}{n}, \frac{1}{\sqrt{n}}, n^{-1/4}, n^{-1/6}, 1\right\}$. Table 4 shows the empirical rejection frequencies for nominal 5% and 10% tests using $n = 5000$ observations, $R = 5000$ Monte Carlo simulations, and $B = 5000$ subsampling replications with a subsample size of $b = \sqrt{n}$. In all cases we use 5-fold cross validation to form the test statistic $\tilde{T}_n \equiv n^{\hat{\delta}}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1)\right)$, and estimate $\hat{\delta}$ using $I =$

10 different values of the subsample size $n^{(0.5:0.05:0.95)}$. We also report the $\hat{\delta}$ estimates averaged over the $R = 5000$ Monte Carlo simulations. The average $\hat{\delta}$ is fairly close to 1 for $\beta_0 \in \left\{0, \frac{1}{n}, \frac{1}{\sqrt{n}}\right\}$ and close to 0.5 for $\beta_0 \in \left\{n^{-1/4}, n^{-1/6}, 1\right\}$. This suggests that first order degeneracy is a problem for values of β_0 drifting very quickly to zero, but not an issue for the slower drifting values. For $\beta_0 \in \left\{0, \frac{1}{n}, \frac{1}{\sqrt{n}}\right\}$, the two models are sufficiently similar such that our test fails to reject the null of equal expected out-of-sample error for a majority of the simulations. For $\beta_0 \in \left\{n^{-1/4}, n^{-1/6}, 1\right\}$, the two models are different enough so that the probability of rejecting the null in favor of model M_2 is very close to 1.

	$\beta_0 = 0$	$\beta_0 = n^{-1}$	$\beta_0 = n^{-1/2}$	$\beta_0 = n^{-1/4}$	$\beta_0 = n^{-1/6}$	$\beta_0 = 1$
			nominal 5%			
Fail to Reject	0.9852	0.9852	0.9922	0.0004	0	0
Favor M_1	0.0148	0.0148	0.0078	0	0	0
Favor M_2	0	0	0	0.9996	1	1
			nominal 10%			
Fail to Reject	0.9626	0.9624	0.9602	0	0	0
Favor M_1	0.0344	0.0344	0.0214	0	0	0
Favor M_2	0.0030	0.0032	0.0184	1	1	1
Average $\hat{\delta}$	0.972	0.972	0.925	0.527	0.513	0.506

Table 4: Empirical Rejection/Non-rejection Frequencies, $n = 5000$, $B = 5000$, $R = 5000$

6. Empirical Application

The data come from GoDaddy, a domain name registrar responsible for managing sales of internet domain names. Each observation is a particular domain name listed on a GoDaddy expiry auction between May 12th, 2017 and July 11th, 2017. The domains are auctioned off in an open-bid English auction with a minimum bid of \$12 and a duration of approximately 10 days. If the domain is still not sold after the 10 days are over, there is a 5 day closeout Dutch auction. One interesting fact about these auctions is that the majority of participants are speculators who have no intrinsic use of the domain except turning a profit when they resell the name in an aftermarket. Another interesting fact is that very few of the English auctions result in sale, partly due to the sheer volume of domains that are listed for sale. For example, of the 2178187 auctions with a start time on or after May 12th, 2017 and before July 11th, 2017, only 28448 auctions met the minimum bid requirement. Starting on May 12th, 2017, GoDaddy implemented a simple randomized experiment where some domain names would receive a valuation metric provided by a machine learning algorithm using deep learning. The idea was to provide auction participants with a better sense of the value of a domain name.

Our goal is to compare three different estimators for predicting the sale price for those 28448 auctions which met the minimum bid requirement using the following nine features: dummy for whether the domain has a valuation assigned by the deep learning algorithm, the domain valuation assigned by the deep learning algorithm (coded as 0 for those domains without a valuation), the number of characters in the domain name, three dummies for whether the domain is a .com, .org, or .net, and three dummies for whether the domain contains a word in the English dictionary, a number, or a vowel.

The first estimator is OLS using all independent variables. The second estimator is an honest Random Forest estimator using all independent variables and computed using the `grf` R package's `regression_forest` command with the default options. The third estimator is a single-hidden layer Neural Network with a sigmoidal activation function and 5 hidden units using all independent variables and computed using the `nn.train` command in the `deepnet` R package.

We perform three pairwise nominal (5/3)%-level 5-fold cross validation with squared error loss estimator selection tests using $B = 1000$ subsampling iterations. We divide by 3 because we are using the Bonferroni correction to control the Family Wise Error Rate at 5%; this is just one of many ways that we could have corrected the multiple testing issue. The 5-fold cross validation errors are 68323.73 for Random Forests, 94291.50 for Neural Networks, and 94444.18 for OLS. For all three tests, we reject the null of equal expected out-of-sample error for the two estimators. When we compare Random Forests to Neural Networks, we find Random Forests has greater predictive accuracy than Neural Networks. When we compare Random Forests to OLS, we find Random Forests has greater predictive accuracy than OLS. When we compare Neural Networks to OLS, we find Neural Networks has greater predictive accuracy than OLS. The results are the same across a range of different values of the subsample size $b = \lfloor n^{\kappa} \rfloor$, where $\kappa \in \{0.4, 0.5, 0.6, 0.7, 0.8\}$.

7. Conclusion

We have demonstrated asymptotic normality of the K-fold cross validation error as the number of observations n goes to infinity keeping the number of folds K fixed. The rate of convergence of the K-fold cross validation error to the expected out-of-sample error is $\sqrt{n}$ for both parametric and nonparametric estimators, and the asymptotic variance does not depend on K. We have constructed an analytic estimate of the asymptotic variance, which can be used to construct confidence intervals for the expected out-of-sample error. We have also developed a hypothesis test for comparing two estimators' expected out-of-sample errors by looking at the difference in their K-fold cross validation errors. In the absence of first-order degeneracy, the test statistic is asymptotically normal under the null and can be benchmarked against the standard normal critical values. In the presence of first-order degeneracy, the test statistic is not asymptotically normal but its distribution under the null is consistently estimable using subsampling.

Appendix A. Appendix

A.1 Proof of Theorem 3.1

Denote τ_{n_t} as the analog of τ_n using $n_t = n(K-1)/K$ observations, and $I_k^{(t)}$ as the indices of the observations in the k^{th} training fold. Taking second order Taylor expansions of $\gamma(\hat{s}_M^{(-k)}; \xi_i)$ and $\gamma(\hat{s}_M; \tilde{\xi}_i)$ around $\gamma(s_M^*; \xi_i)$ and $\gamma(s_M^*; \tilde{\xi}_i)$,

$$\begin{aligned}
&\sqrt{n}(\hat{\mathcal{L}}_n^{CV}(M) - EPE_{OUT,n}(M)) \\
=&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_v}} \sum_{i \in I_k^{(v)}} \Bigg\{ \gamma(\hat{s}_M^{(-k)}; \xi_i) - \mathbb{E}[\gamma(s_M^*; \tilde{\xi}_i)] - \mathbb{E}\left[(\hat{s}_M(\tilde{x}_i) - s_M^*(\tilde{x}_i)) \frac{\partial \gamma(s_M^*; \tilde{\xi}_i)}{\partial s_M^*} \right] \\
&\qquad\qquad - \mathbb{E}\left[(\hat{s}_M(\tilde{x}_i) - s_M^*(\tilde{x}_i))^2 \frac{1}{2} \frac{\partial^2 \gamma(s_M^*; \tilde{\xi}_i)}{\partial {s_M^*}^2} \right] \Bigg\} \\
=&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_v}} \sum_{i \in I_k^{(v)}} \left\{ \gamma(s_M^*; \xi_i) - \mathbb{E}[\gamma(s_M^*; \tilde{\xi}_i)] \right\} \\
+&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_v}} \sum_{i \in I_k^{(v)}} \left\{ \left(\hat{s}_M^{(-k)}(x_i) - s_M^*(x_i) \right) \frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} - \mathbb{E}\left[(\hat{s}_M(\tilde{x}_i) - s_M^*(\tilde{x}_i)) \frac{\partial \gamma(s_M^*; \tilde{\xi}_i)}{\partial s_M^*} \right] \right\} \\
+&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_v}} \sum_{i \in I_k^{(v)}} \left\{ \left(\hat{s}_M^{(-k)}(x_i) - s_M^*(x_i) \right)^2 \frac{\partial^2 \gamma(s_M^*; \xi_i)}{\partial {s_M^*}^2} - \mathbb{E}\left[(\hat{s}_M(\tilde{x}_i) - s_M^*(\tilde{x}_i))^2 \frac{\partial^2 \gamma(s_M^*; \tilde{\xi}_i)}{\partial {s_M^*}^2} \right] \right\}
\end{aligned}$$

Using Assumption 3.3 and the fact that $\tilde{x}_i$ is drawn from the same distribution as x_i,

$$\begin{aligned}
&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_v}} \sum_{i \in I_k^{(v)}} \left\{ (\hat{s}_M^{(-k)}(x_i) - s_M^*(x_i)) \frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} - \mathbb{E}\left[(\hat{s}_{M,n_t}(\tilde{x}_i) - s_M^*(\tilde{x}_i)) \frac{\partial \gamma(s_M^*; \tilde{\xi}_i)}{\partial s_M^*} \right] \right\} \\
=&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_v}} \sum_{i \in I_k^{(v)}} \frac{1}{n_t} \sum_{j \in I_k^{(t)}} \left\{ \phi_M(\xi_j, x_i) \frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} - \mathbb{E}\left[\phi_M(\xi_j, x_i) \frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} \right] \right\} \\
+&o_p\left(\frac{1}{\tau_{n_t}} \right) \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n_v}} \sum_{i \in I_k^{(v)}} \left(\frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} - \mathbb{E}\left[\frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} \right] \right) \\
=&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sqrt{n_v} \frac{1}{n_v} \frac{1}{n_t} \sum_{i \in I_k^{(v)}} \sum_{j \in I_k^{(t)}} \underbrace{\left\{ \phi_M(\xi_j, x_i) \frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} - \mathbb{E}\left[\phi_M(\xi_j, x_i) \frac{\partial \gamma(s_M^*; \xi_i)}{\partial s_M^*} \right] \right\}}_{\psi(\xi_j, \xi_i) - \mathbb{E}\left[\psi(\xi_j, \xi_i)\right]} + o_p\left(\frac{1}{\tau_{n_t}} \right) O_P(1) \\
\equiv&\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sqrt{n_v} \underbrace{\frac{1}{n_v} \frac{1}{n_t} \sum_{i \in I_k^{(v)}} \sum_{j \in I_k^{(t)}} \left\{ \psi(\xi_j, \xi_i) - \mathbb{E}\left[\psi(\xi_j, \xi_i) \right] \right\}}_{U_{1k}} + o_p\left(\frac{1}{\tau_{n_t}} \right)
\end{aligned}$$

Furthermore,

$$
\begin{aligned}
&\frac{1}{2}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left\{\left(\hat{s}_M^{(-k)}(x_i)-s_M^*(x_i)\right)^2\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}-\mathbb{E}\left[\left(\hat{s}_{M,n_t}(\tilde{x}_i)-s_M^*(\tilde{x}_i)\right)^2\frac{\partial^2\gamma(s_M^*;\tilde{\xi}_i)}{\partial {s_M^*}^2}\right]\right\}\\
&=\frac{1}{2}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left\{\frac{1}{n_t^2}\sum_{j_1\in I_k^{(t)}}\sum_{j_2\in I_k^{(t)}}\phi_M(\xi_{j_1},x_i)\phi_M(\xi_{j_2},x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right.\\
&\left.-\frac{1}{n_t}\mathbb{E}\left[\phi_M(\xi_j,x_i)^2\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right]-\frac{n_t-1}{n_t}\mathbb{E}\left[\phi_M(\xi_{j_1},x_i)\phi_M(\xi_{j_2},x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right]\right\}\\
&+o_p\left(\frac{1}{\tau_{n_t}}\right)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\frac{1}{n_t}\sum_{j\in I_k^{(t)}}\left(\phi_M(\xi_j,x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}-\mathbb{E}\left[\phi_M(\xi_j,x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right]\right)\\
&+o_p\left(\frac{1}{\tau_{n_t}^2}\right)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left(\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}-\mathbb{E}\left[\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right]\right)\\
&=\frac{1}{2}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{\sqrt{n_v}}{n_t}\underbrace{\frac{1}{n_v}\frac{1}{n_t}\sum_{i\in I_k^{(v)}}\sum_{j\in I_k^{(t)}}\left\{\phi_M(\xi_j,x_i)^2\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}-\mathbb{E}\left[\phi_M(\xi_j,x_i)^2\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right]\right\}}_{R_{1k}}\\
&+\frac{\sqrt{n_v}}{\sqrt{K}}\sum_{k=1}^{K}\underbrace{\frac{2}{n_vn_t(n_t-1)}\sum_{i\in I_k^{(v)}}\sum_{j_1\in I_k^{(t)}}\sum_{j_2>j_1}\left\{\underbrace{\frac{1}{2}\phi_M(\xi_{j_1},x_i)\phi_M(\xi_{j_2},x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}}_{\nu(\xi_{j_1},\xi_{j_2},\xi_i)}-\mathbb{E}\left[\nu(\xi_{j_1},\xi_{j_2},\xi_i)\right]\right\}}_{U_{2k}}\\
&-\frac{1}{n_t}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_v}U_{2k}\\
&+o_p\left(\frac{1}{\tau_{n_t}}\right)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_v}\underbrace{\frac{1}{n_v}\frac{1}{n_t}\sum_{i\in I_k^{(v)}}\sum_{j\in I_k^{(t)}}\left(\phi_M(\xi_j,x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}-\mathbb{E}\left[\phi_M(\xi_j,x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right]\right)}_{R_{2k}}\\
&+o_p\left(\frac{1}{\tau_{n_t}^2}\right)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left(\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}-\mathbb{E}\left[\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\right]\right)
\end{aligned}
$$

For each $k = 1...K$, U_{1k} is a two-sample U-statistic of degree (1,1) with kernel $\psi(\xi_j,\xi_i) = \phi_M(\xi_j,x_i)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}$, U_{2k} is a two-sample U-statistic of degree (2,1) with kernel $\nu(\xi_{j_1},\xi_{j_2},\xi_i) = \frac{1}{2}\phi_M(\xi_{j_1},x_i)\phi_M(\xi_{j_2},x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}$, R_{1k} is a two-sample U-statistic of degree (1,1) with kernel $\kappa(\xi_j,\xi_i) = \phi_M(\xi_j,x_i)^2\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}$, and R_{2k} is a two-sample U-statistic of degree (1,1) with kernel $\eta(\xi_j,\xi_i) = \phi_M(\xi_j,x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}$. Results for two-sample U-statistics in e.g. (Van der Vaart, 2000) show that R_{1k} and R_{2k} are $O_p\left(\frac{1}{\sqrt{n_v+n_t}}\right)$ for all k, which imply they do not contribute to the asymptotic distribution of $\hat{\mathcal{L}}_n^{CV}(M)$.

For $\sqrt{n}$-consistent estimators, $\mathbb{E}\left[\phi_M(\xi_j,x_i)\middle|\, x_i\right] = 0$ implies $\mathbb{E}\left[\psi(\xi_j,\xi_i)\middle|\,\xi_i\right] = 0$,

$\mathbb{E}\left[\nu(\xi_{j_1}, \xi_{j_2}, \xi_i) \middle| \xi_i\right] = 0$, $\mathbb{E}\left[\nu(\xi_{j_1}, \xi_h, \xi_i) \middle| \xi_h\right] = 0$, and $\mathbb{E}\left[\nu(\xi_h, \xi_{j_2}, \xi_i) \middle| \xi_h\right] = 0$ because

$$\begin{aligned}
&\mathbb{E}\left[\phi_M(\xi_j, x_i)\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*}\middle|\xi_i\right] = \mathbb{E}\left[\phi_M(\xi_j, x_i)\middle| x_i\right]\frac{\partial\gamma(s_M^*;\xi_i)}{\partial s_M^*} = 0\\
&\mathbb{E}\left[\phi_M(\xi_{j_1}, x_i)\phi_M(\xi_{j_2}, x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\middle|\xi_i\right] = \mathbb{E}\left[\phi_M(\xi_{j_1}, x_i)\middle| x_i\right]\mathbb{E}\left[\phi_M(\xi_{j_2}, x_i)\middle| x_i\right]\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2} = 0\\
&\mathbb{E}\left[\phi_M(\xi_{j_1}, x_i)\phi_M(\xi_{j_2}, x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\middle|\xi_{j_1}\right] = \mathbb{E}\left[\phi_M(\xi_{j_1}, x_i)\mathbb{E}\left[\phi_M(\xi_{j_2}, x_i)\middle| x_i\right]\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\middle|\xi_{j_1}\right] = 0\\
&\mathbb{E}\left[\phi_M(\xi_{j_1}, x_i)\phi_M(\xi_{j_2}, x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\middle|\xi_{j_2}\right] = \mathbb{E}\left[\mathbb{E}\left[\phi_M(\xi_{j_1}, x_i)\middle| x_i\right]\phi_M(\xi_{j_2}, x_i)\frac{\partial^2\gamma(s_M^*;\xi_i)}{\partial {s_M^*}^2}\middle|\xi_{j_2}\right] = 0
\end{aligned}$$

This implies that U_{2k} is a degenerate two-sample U-statistic of order $O_p\left(\frac{1}{\sqrt{n_v n_t}}\right)$, and only U_{1k} contributes to the asymptotic distribution. Additionally, $\mathbb{E}\left[\psi(\xi_j, \xi_i)\middle| \xi_i\right] = 0$ implies that one of the projection terms disappears in the Hoeffding decomposition for U_{1k} (see e.g. (Lehmann, 1951), (Van der Vaart, 2000), or (Neumeyer, 2004)).

$$\begin{aligned}
\sqrt{n_v}U_{1k} =& \frac{\sqrt{n_v}}{\sqrt{n_t}}\frac{1}{\sqrt{n_t}}\sum_{j\in I_k^{(t)}}\mathbb{E}\left[\psi(\xi_j, \xi_i) - \mathbb{E}\left[\psi(\xi_j, \xi_i)\right]\middle| \xi_j\right]\\
&+ \underbrace{\frac{\sqrt{n_v}}{\sqrt{n_v}}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\mathbb{E}\left[\psi(\xi_j, \xi_i) - \mathbb{E}\left[\psi(\xi_j, \xi_i)\right]\middle| \xi_i\right]}_{=0}\\
&+ \sqrt{n_v}\Delta_{1k}
\end{aligned}$$

where Δ_{1k} is a degenerate two-sample U-statistic of order $O_p\left(\frac{1}{\sqrt{n_v n_t}}\right)$.

Because each observation ξ_j appears in $K-1$ training sets,

$$\sum_{k=1}^K\sum_{j\in I_k^{(t)}}\mathbb{E}\left[\psi(\xi_j, \xi_i) - \mathbb{E}\left[\psi(\xi_j, \xi_i)\right]\middle| \xi_j\right] = (K-1)\sum_{j=1}^n\mathbb{E}\left[\psi(\xi_j, \xi_i) - \mathbb{E}\left[\psi(\xi_j, \xi_i)\right]\middle| \xi_j\right]$$

Since $n_t = (K-1)n_v$ and $n_t = n(K-1)/K$,

$$\frac{1}{\sqrt{K}}\sum_{k=1}^K\sqrt{n_v}U_{1k} = \frac{1}{\sqrt{n}}\sum_{j=1}^n\mathbb{E}\left[\psi(\xi_j, \xi_i) - \mathbb{E}\left[\psi(\xi_j, \xi_i)\right]\middle| \xi_j\right] + \frac{1}{\sqrt{K}}\sum_{k=1}^K\sqrt{n_v}\Delta_{1k}$$

$$\begin{aligned}
\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M) - EPE_{OUT,n}(M)\right) =& \frac{1}{\sqrt{K}}\sum_{k=1}^K\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left\{\gamma(s_M^*;\xi_i) - \mathbb{E}[\gamma(s_M^*;\xi_i)]\right\}\\
&+ \frac{1}{\sqrt{K}}\sum_{k=1}^K\sqrt{n_v}U_{1k} + o_p(1)
\end{aligned}$$

By the Lindeberg-Levy Central Limit Theorem,

$$\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M) - EPE_{OUT,n}(M)\right) \xrightarrow{d} N\left(0, \sigma_\gamma^2 + \sigma_{01,\psi}^2 + 2\lambda\right)$$

where $\sigma_\gamma^2 \equiv Var\left[\gamma(s_M^*;\xi_i)\right]$, $\sigma_{01,\psi}^2 \equiv Var\left[\mathbb{E}\left[\psi(\xi_j,\xi_i)|\,\xi_j\right]\right]$, and $\lambda \equiv Cov\left[\gamma(s_M^*;\xi_j),\mathbb{E}\left[\psi(\xi_j,\xi_i)|\xi_j\right]\right]$.

For nonparametric estimators, note that our assumptions of $Var\left[\mathbb{E}\left[\psi\left(\xi_j,\xi_i\right)|\,\xi_i\right]\right] \to 0$ and $Var\left[\mathbb{E}\left[\psi\left(\xi_j,\xi_i\right)|\,\xi_j\right]\right] \to 0$ imply U_{1k} is a degenerate two-sample U-statistic, so only the first term contributes to the asymptotic distribution:

$$\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M) - EPE_{OUT,n}(M)\right) = \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left\{\gamma(s_M^*;\xi_i) - \mathbb{E}[\gamma(s_M^*;\xi_i)]\right\} + o_p(1)$$

The asymptotic variance will be $\sigma^2 = \sigma_\gamma^2$.

□

A.2 Proof of Theorem 3.2

Define $\hat{\gamma}_M\left(\xi_i\right) = \gamma\left(\hat{s}_M;\xi_i\right)$ and $\gamma_{0M}\left(\xi_i\right) = \gamma\left(s_M^*;\xi_i\right)$. We assumed $\frac{1}{n}\sum_{i=1}^n\left|\hat{\gamma}_M\left(\xi_i\right) - \gamma_{0M}\left(\xi_i\right)\right|^2 = o_p(1)$, and by the law of large numbers, $\frac{1}{n}\sum_{i=1}^n\gamma_{0M}\left(\xi_i\right) \xrightarrow{p} \mathbb{E}\left[\gamma_{0M}\left(\xi_i\right)\right]$. Therefore,

$$\left|\frac{1}{n}\sum_{i=1}^n\hat{\gamma}_M\left(\xi_i\right) - \mathbb{E}\left[\gamma_{0M}\left(\xi_i\right)\right]\right| \le \frac{1}{n}\sum_{i=1}^n\left|\hat{\gamma}_M\left(\xi_i\right) - \gamma_{0M}\left(\xi_i\right)\right| + \left|\frac{1}{n}\sum_{i=1}^n\gamma_{0M}\left(\xi_i\right) - \mathbb{E}\left[\gamma_{0M}\left(\xi_i\right)\right]\right| \xrightarrow{p} 0$$

$$\left|\frac{1}{n}\sum_{i=1}^n\hat{\gamma}_M^2\left(\xi_i\right) - \mathbb{E}\left[\gamma_{0M}^2\left(\xi_i\right)\right]\right|$$

$$\le \frac{1}{n}\sum_{i=1}^n\left|\hat{\gamma}_M\left(\xi_i\right) - \gamma_{0M}\left(\xi_i\right)\right|^2 + 2\frac{1}{n}\sum_{i=1}^n\left|\gamma_{0M}\left(\xi_i\right)\right|\left|\hat{\gamma}_M\left(\xi_i\right) - \gamma_{0M}\left(\xi_i\right)\right|$$

$$+ \left|\frac{1}{n}\sum_{i=1}^n\gamma_{0M}^2\left(\xi_i\right) - \mathbb{E}\left[\gamma_{0M}^2\left(\xi_i\right)\right]\right|$$

$$\le \frac{1}{n}\sum_{i=1}^n\left|\hat{\gamma}_M\left(\xi_i\right) - \gamma_{0M}\left(\xi_i\right)\right|^2 + 2\sqrt{\frac{1}{n}\sum_{i=1}^n\left|\gamma_{0M}\left(\xi_i\right)\right|^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\left|\hat{\gamma}_M\left(\xi_i\right) - \gamma_{0M}\left(\xi_i\right)\right|^2} + o_p(1) \xrightarrow{p} 0$$

Therefore, $\hat{\sigma}_\gamma^2 = \frac{1}{n}\sum_{i=1}^n\hat{\gamma}_M^2\left(\xi_i\right) - \left(\frac{1}{n}\sum_{i=1}^n\hat{\gamma}_M\left(\xi_i\right)\right)^2 \xrightarrow{p} Var\left[\gamma_{0M}(\xi_i)\right]$. We have shown that we can consistently estimate the asymptotic variance of the K-fold cross validation error for nonparametric estimators.

Now we consider the case of $\sqrt{n}$-consistent estimators. Define $h_{0M}\left(\xi_j\right) = \mathbb{E}\left[\psi(\xi_j,\xi_i;s_M^*)|\,\xi_j\right]$, $\tilde{h}_M\left(\xi_j\right) = \frac{1}{n-1}\sum_{i\neq j}\psi\left(\xi_j,\xi_i;s_M^*\right)$, and $\hat{h}_M\left(\xi_j\right) = \frac{1}{n-1}\sum_{i\neq j}\hat{\psi}\left(\xi_j,\xi_i;\hat{s}_M\right)$. Notice that $\frac{1}{n}\sum_{j=1}^n\tilde{h}_M\left(\xi_j\right) = \frac{1}{n-1}\frac{1}{n}\sum_{j=1}^n\sum_{i\neq j}\psi\left(\xi_j,\xi_i;s_M^*\right)$ is a U-statistic of order 2. Notice that $\frac{1}{n}\sum_{j=1}^n h_{0M}\left(\xi_j\right)$ is one of the projection terms in the Hoeffding decomposition for U-statistics, and the other projection term is zero since $\mathbb{E}\left[\psi\left(\xi_j,\xi_i;s_M^*\right)|\,\xi_i\right] = 0$. Since the remainder term in the Hoeffding decomposition must converge in probability to zero,

$$\left|\frac{1}{n}\sum_{j=1}^n\tilde{h}_M\left(\xi_j\right) - \frac{1}{n}\sum_{j=1}^n h_{0M}\left(\xi_j\right)\right| = \left|\frac{1}{n-1}\frac{1}{n}\sum_{j=1}^n\sum_{i\neq j}\psi\left(\xi_j,\xi_i;s_M^*\right) - \frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\psi\left(\xi_j,\xi_i;s_M^*\right)|\,\xi_j\right]\right| \xrightarrow{p} 0$$

Similarly, notice that $\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_M\left(\xi_j\right)^2=\frac{1}{(n-1)^2}\frac{1}{n}\sum_{j=1}^{n}\sum_{i_1\neq j}\sum_{i_2\neq j}\psi\left(\xi_j,\xi_{i_1};s_M^*\right)\psi\left(\xi_j,\xi_{i_2};s_M^*\right)$ is asymptotically equivalent to a U-statistic of order 3. Using the Hoeffding decomposition and $\mathbb{E}\left[\psi\left(\xi_j,\xi_{i_1};s_M^*\right)\psi\left(\xi_j,\xi_{i_2};s_M^*\right)\middle|\,\xi_j\right]=\mathbb{E}\left[\psi\left(\xi_j,\xi_i;s_M^*\right)\middle|\,\xi_j\right]^2$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_M\left(\xi_j\right)^2-\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)^2\right|\\
=&\left|\frac{1}{n\left(n-1\right)\left(n-2\right)}\sum_{j=1}^{n}\sum_{i_1\neq j}\sum_{i_2\neq i_1}\psi\left(\xi_j,\xi_{i_1};s_M^*\right)\psi\left(\xi_j,\xi_{i_2};s_M^*\right)-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\psi\left(\xi_j,\xi_i;s_M^*\right)\middle|\,\xi_j\right]^2\right|+o_p(1)\\
=&o_p(1)
\end{aligned}$$

Since we assumed $\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_M\left(\xi_j\right)-\tilde{h}_M\left(\xi_j\right)\right|^2\xrightarrow{p}0$, it follows that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)^2-\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)^2\right|\\
&\leq\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_M\left(\xi_j\right)-\tilde{h}_M\left(\xi_j\right)\right|^2+2\frac{1}{n}\sum_{j=1}^{n}\left|\tilde{h}_M\left(\xi_j\right)\right|\left|\hat{h}_M\left(\xi_j\right)-\tilde{h}_M\left(\xi_j\right)\right|\\
&+\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_M\left(\xi_j\right)^2-\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)^2\right|\\
&\leq\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_M\left(\xi_j\right)-\tilde{h}_M\left(\xi_j\right)\right|^2+2\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left|\tilde{h}_M\left(\xi_j\right)\right|^2}\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_M\left(\xi_j\right)-\tilde{h}_M\left(\xi_j\right)\right|^2}+o_p(1)\\
&\xrightarrow{p}0
\end{aligned}$$

$$\begin{aligned}
\left|\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right|&\leq\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_M\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right|+\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_M\left(\xi_j\right)-\tilde{h}_M\left(\xi_j\right)\right|\\
&\xrightarrow{p}0
\end{aligned}$$

which implies that

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right)^2\\
\leq&\left|\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)+\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right|\left|\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right|\\
=&\,O_p(1)o_p(1)=o_p(1)
\end{aligned}$$

and therefore,

$$\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right)^2\right)\xrightarrow{p}0$$

Since $\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right)^2\xrightarrow{p}Var\left(h_{0M}\left(\xi_j\right)\right)$, it follows that $\hat{\sigma}^2_{01,\psi}=\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)\right)^2\xrightarrow{p}Var\left(h_{0M}\left(\xi_j\right)\right)$.

Note that $\frac{1}{n}\sum_{j=1}^{n}\gamma_{0M}\left(\xi_j\right)\tilde{h}_M\left(\xi_j\right)$ is a U-statistic of order 2. By the Hoeffding decomposition,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{j=1}^{n}\gamma_{0M}\left(\xi_j\right)\tilde{h}_M\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}\gamma_{0M}\left(\xi_j\right)h_{0M}\left(\xi_j\right)\right|\\&=\left|\frac{1}{n-1}\frac{1}{n}\sum_{j=1}^{n}\sum_{i\neq j}\gamma\left(s_M^*;\xi_j\right)\psi\left(\xi_j,\xi_i;s_M^*\right)-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\gamma\left(s_M^*;\xi_j\right)\psi\left(\xi_j,\xi_i;s_M^*\right)\middle|\,\xi_j\right]\right|\\&\xrightarrow{p}0\end{aligned}$$

Our assumptions and results so far imply

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_M\left(\xi_i\right)\hat{h}_M\left(\xi_i\right)-\frac{1}{n}\sum_{i=1}^{n}\gamma_{0M}\left(\xi_i\right)h_{0M}\left(\xi_i\right)\right|\\&\leq\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_M\left(\xi_i\right)-\gamma_{0M}\left(\xi_i\right)\right)\hat{h}_M\left(\xi_i\right)\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{h}_M\left(\xi_i\right)-\tilde{h}_M\left(\xi_i\right)\right)\gamma_{0M}\left(\xi_i\right)\right|\\&+\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_{0M}\left(\xi_i\right)\tilde{h}_M\left(\xi_i\right)-\frac{1}{n}\sum_{i=1}^{n}\gamma_{0M}\left(\xi_i\right)h_{0M}\left(\xi_i\right)\right|\\&\leq\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\gamma}_M\left(\xi_i\right)-\gamma_{0M}\left(\xi_i\right)\right|\left|\hat{h}_M\left(\xi_i\right)\right|+\frac{1}{n}\sum_{i=1}^{n}\left|\hat{h}_M\left(\xi_i\right)-\tilde{h}_M\left(\xi_i\right)\right|\left|\gamma_{0M}\left(\xi_i\right)\right|+o_p(1)\\&\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\gamma}_M\left(\xi_i\right)-\gamma_{0M}\left(\xi_i\right)\right|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{h}_M\left(\xi_i\right)\right|^2}\\&+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{h}_M\left(\xi_i\right)-\tilde{h}_M\left(\xi_i\right)\right|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\gamma_{0M}\left(\xi_i\right)\right|^2}+o_p(1)\\&=o_p\left(1\right)O_p\left(1\right)+o_p\left(1\right)O_p\left(1\right)=o_p\left(1\right)\end{aligned}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\left(\gamma_{0M}(\xi_i)h_{0M}(\xi_i)-E\left[\gamma_{0M}(\xi_i)h_{0M}(\xi_i)\right]\right)\xrightarrow{p}0$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_M(\xi_i)\hat{h}_M(\xi_i)-E\left[\gamma_{0M}(\xi_i)h_{0M}(\xi_i)\right]\right)\xrightarrow{p}0$$

Since also $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_M(\xi_i) - E\left[\gamma_{0M}(\xi_i)\right]\right) \xrightarrow{p} 0$ and $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{h}_M(\xi_i) - E\left[h_{0M}(\xi_i)\right]\right) \xrightarrow{p} 0$,

$$\hat{\lambda} \xrightarrow{p} \lambda \equiv Cov\left[\gamma_{0M}(\xi_i), h_{0M}(\xi_i)\right]$$

□

A.3 Proof of Theorem 4.1

The arguments are similar to the Proof of Theorem 3.1, except that we now have to account for the possibly different rates of convergence of the two estimators. Taking Taylor expansions of $\gamma(\hat{s}_M^{(-k)};\xi_i)$ and $\gamma(\hat{s}_M;\tilde{\xi}_i)$ around $\gamma(s_M^*;\xi_i)$ and $\gamma(s_M^*;\tilde{\xi}_i)$ for $M \in \{M_1, M_2\}$,

$$
\begin{aligned}
&\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - \left(\mathbb{E}[\gamma(\hat{s}_{M_2};\tilde{\xi}_i)] - \mathbb{E}[\gamma(\hat{s}_{M_1};\tilde{\xi}_i)]\right)\right)\\
&= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left\{\gamma(s_{M_2}^*;\xi_i) - \gamma(s_{M_1}^*;\xi_i) - \left(\mathbb{E}[\gamma(s_{M_2}^*;\tilde{\xi}_i)] - \mathbb{E}[\gamma(s_{M_1}^*;\tilde{\xi}_i)]\right)\right\}\\
&+ \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\left\{\left(\hat{s}_{M_2}^{(-k)}(x_i) - s_{M_2}^*(x_i)\right)\frac{\partial\gamma(s_{M_2}^*;\xi_i)}{\partial s_{M_2}^*} - \left(\hat{s}_{M_1}^{(-k)}(x_i) - s_{M_1}^*(x_i)\right)\frac{\partial\gamma(s_{M_1}^*;\xi_i)}{\partial s_{M_1}^*}\right.\\
&- \left.\left(\mathbb{E}\left[(\hat{s}_{M_2}(\tilde{x}_i) - s_{M_2}^*(\tilde{x}_i))\frac{\partial\gamma(s_{M_2}^*;\tilde{\xi}_i)}{\partial s_{M_2}^*}\right] - \mathbb{E}\left[(\hat{s}_{M_1}(\tilde{x}_i) - s_{M_1}^*(\tilde{x}_i))\frac{\partial\gamma(s_{M_1}^*;\tilde{\xi}_i)}{\partial s_{M_1}^*}\right]\right)\right\}\\
&+ \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in I_k^{(v)}}\left\{(\hat{s}_{M_2}^{(-k)}(x_i) - s_{M_2}^*(x_i))^2\frac{\partial^2\gamma(s_{M_2}^*;\xi_i)}{\partial {s_{M_2}^*}^2} - \mathbb{E}\left[(\hat{s}_{M_2}(\tilde{x}_i) - s_{M_2}^*(\tilde{x}_i))^2\frac{\partial^2\gamma(s_{M_2}^*;\tilde{\xi}_i)}{\partial {s_{M_2}^*}^2}\right]\right\}\\
&- \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in I_k^{(v)}}\left\{(\hat{s}_{M_1}^{(-k)}(x_i) - s_{M_1}^*(x_i))^2\frac{\partial^2\gamma(s_{M_1}^*;\xi_i)}{\partial {s_{M_1}^*}^2} - \mathbb{E}\left[(\hat{s}_{M_1}(\tilde{x}_i) - s_{M_1}^*(\tilde{x}_i))^2\frac{\partial^2\gamma(s_{M_1}^*;\tilde{\xi}_i)}{\partial {s_{M_1}^*}^2}\right]\right\}\\
&= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\underbrace{\{\gamma(s_{M_2}^*;\xi_i) - \gamma(s_{M_1}^*;\xi_i) - (\mathbb{E}[\gamma(s_{M_2}^*;\xi_i)] - \mathbb{E}[\gamma(s_{M_1}^*;\xi_i)])\}}_{\tilde{\gamma}(\xi_i)-\mathbb{E}[\tilde{\gamma}(\xi_i)]}\\
&+ \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_v}\frac{1}{n_v}\frac{1}{n_t}\sum_{i\in I_k^{(v)}}\sum_{j\in I_k^{(t)}}\left\{\underbrace{\phi_{M_2}(\xi_j, x_i)\frac{\partial\gamma(s_{M_2}^*;\xi_i)}{\partial s_{M_2}^*}}_{\psi_{M_2}(\xi_j,\xi_i)} - \mathbb{E}\left[\psi_{M_2}(\xi_j,\xi_i)\right]\right\}\\
&- \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_v}\frac{1}{n_v}\frac{1}{n_t}\sum_{i\in I_k^{(v)}}\sum_{j\in I_k^{(t)}}\left\{\underbrace{\phi_{M_1}(\xi_j, x_i)\frac{\partial\gamma(s_{M_1}^*;\xi_i)}{\partial s_{M_1}^*}}_{\psi_{M_1}(\xi_j,\xi_i)} - \mathbb{E}\left[\psi_{M_1}(\xi_j,\xi_i)\right]\right\} + o_p\left(1\right)\\
&\equiv \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_v}}\sum_{i\in I_k^{(v)}}\{\tilde{\gamma}(\xi_i) - \mathbb{E}[\tilde{\gamma}(\xi_i)]\} + \left\{\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_v}U_{1,M_2,k} - \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_v}U_{1,M_1,k}\right\} + o_p(1)
\end{aligned}
$$

Note that $\tilde{U}_{1k} \equiv U_{1,M_2,k} - U_{1,M_1,k}$ is a two-sample U-statistic with kernel function $\tilde{\psi}(\xi_j,\xi_i) \equiv \psi_{M_2}(\xi_j,\xi_i) - \psi_{M_1}(\xi_j,\xi_i)$. For $\sqrt{n}$-consistent estimators $\tau_{n,1} = \tau_{n,2} = \sqrt{n}$, $\mathbb{E}\left[\phi_M(\xi_j, x_i)\middle|\, x_i\right] = 0$ for $M \in \{M_1, M_2\}$ implies $\mathbb{E}\left[\tilde{\psi}(\xi_j,\xi_i)\middle|\,\xi_i\right] = 0$. Using similar Hoeffding decomposition arguments as in the proof of Theorem 3.1,

$$\begin{aligned}&\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)\\&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\tilde{\gamma}(\xi_i) - \mathbb{E}[\tilde{\gamma}(\xi_i)]\} + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbb{E}\left[\tilde{\psi}(\xi_j,\xi_i) - \mathbb{E}\left[\tilde{\psi}(\xi_j,\xi_i)\right]\middle|\,\xi_j\right] + o_p(1)\end{aligned}$$

The asymptotic variance is

$$\tilde{\sigma}^2 = \tilde{\sigma}_\gamma^2 + \tilde{\sigma}_{01,\psi}^2 + 2\tilde{\lambda}$$

If $\tau_{n,1} \ll \tau_{n,2} = \sqrt{n}$, since $Var\left[\mathbb{E}\left[\psi_{M_1}(\xi_j,\xi_i)\middle|\,\xi_i\right]\right] \to 0$ and $Var\left[\mathbb{E}\left[\psi_{M_1}(\xi_j,\xi_i)\middle|\,\xi_j\right]\right] \to 0$, $U_{1,M_1,k}$ is a degenerate two-sample U-statistic, so

$$\begin{aligned}&\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)\\&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\tilde{\gamma}(\xi_i) - \mathbb{E}[\tilde{\gamma}(\xi_i)]\} + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbb{E}\left[\psi_{M_2}(\xi_j,\xi_i) - \mathbb{E}\left[\psi_{M_2}(\xi_j,\xi_i)\right]\middle|\,\xi_j\right] + o_p(1)\end{aligned}$$

The asymptotic variance is

$$\tilde{\sigma}^2 = \tilde{\sigma}_\gamma^2 + \sigma_{01,\psi_{M_2}}^2 + 2\lambda_{M_2}$$

If $\tau_{n,2} \ll \tau_{n,1} = \sqrt{n}$, since $Var\left[\mathbb{E}\left[\psi_{M_2}(\xi_j,\xi_i)\middle|\,\xi_i\right]\right] \to 0$ and $Var\left[\mathbb{E}\left[\psi_{M_2}(\xi_j,\xi_i)\middle|\,\xi_j\right]\right] \to 0$, $U_{1,M_2,k}$ is a degenerate two-sample U-statistic, so

$$\begin{aligned}&\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)\\&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\tilde{\gamma}(\xi_i) - \mathbb{E}[\tilde{\gamma}(\xi_i)]\} - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbb{E}\left[\psi_{M_1}(\xi_j,\xi_i) - \mathbb{E}\left[\psi_{M_1}(\xi_j,\xi_i)\right]\middle|\,\xi_j\right] + o_p(1)\end{aligned}$$

The asymptotic variance is

$$\tilde{\sigma}^2 = \tilde{\sigma}_\gamma^2 + \sigma_{01,\psi_{M_1}}^2 + 2\lambda_{M_1}$$

If $\tau_{n,1}, \tau_{n,2} \ll \sqrt{n}$, $Var\left[\mathbb{E}\left[\psi_{M_1}(\xi_j,\xi_i)\middle|\,\xi_i\right]\right] \to 0$, $Var\left[\mathbb{E}\left[\psi_{M_1}(\xi_j,\xi_i)\middle|\,\xi_j\right]\right] \to 0$, $Var\left[\mathbb{E}\left[\psi_{M_2}(\xi_j,\xi_i)\middle|\,\xi_i\right]\right] \to 0$, and $Var\left[\mathbb{E}\left[\psi_{M_2}(\xi_j,\xi_i)\middle|\,\xi_j\right]\right] \to 0$ imply $U_{1,M_1,k}$ and $U_{1,M_2,k}$ are degenerate two-sample U-statistics. Then only the residual variance term contributes to the asymptotic distribution:

$$\begin{aligned}&\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)\\&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\tilde{\gamma}(\xi_i) - \mathbb{E}[\tilde{\gamma}(\tilde{\xi}_i)]\right\} + o_p(1)\end{aligned}$$

The asymptotic variance will be $\tilde{\sigma}^2 = \tilde{\sigma}_\gamma^2$.

Under H_0 and alternatives of the form $\sqrt{n}\,(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)) \to 0$, $\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1)\right)$ has the same asymptotic distribution as $\sqrt{n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)$. Since $\hat{\tilde{\sigma}}_n^2 \xrightarrow{p} \tilde{\sigma}^2$,

$$T_n = \frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1)\right) \xrightarrow{d} N(0,1)$$

Under alternatives of the form $\sqrt{n}\,(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)) \to c$, we have

$$\begin{aligned} T_n =& \underbrace{\frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)}_{\xrightarrow{d} N(0,1)} \\ &+ \underbrace{\frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))}_{\to c/\tilde{\sigma}} \xrightarrow{d} N(c/\tilde{\sigma}, 1). \end{aligned}$$

Under alternatives of the form $\sqrt{n}\,(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)) \to \infty$, it follows that

$$\begin{aligned} T_n =& \underbrace{\frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)}_{\xrightarrow{d} N(0,1)} \\ &+ \underbrace{\frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))}_{\to \infty} \to \infty. \end{aligned}$$

Under alternatives of the form $\sqrt{n}\,(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1)) \to -\infty$, it follows that

$$\begin{aligned} T_n =& \underbrace{\frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right)}_{\xrightarrow{d} N(0,1)} \\ &+ \underbrace{\frac{\sqrt{n}}{\hat{\tilde{\sigma}}_n}(EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))}_{\to -\infty} \to -\infty. \end{aligned}$$

Up to this point we have examined the case when the rate of convergence of $\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1)$ is $\sqrt{n}$, but it can happen that the rate is n when $\tilde{\sigma}^2 = 0$, which can occur when $Var\left[\tilde{\gamma}(\xi)\right] = 0$ and $\tilde{U}_{1k}$ is a degenerate two-sample U-statistic of degree (1,1). There are results in the literature that characterize the asymptotic distribution of degenerate two-sample U-statistics of degree (1,1) (see e.g. (Neuhaus, 1977), (Eagleson, 1979), (Dewan & Rao, 2001)). They show that the rate of convergence is the square root of the product of the two sample sizes. In our case, this would mean that $\tilde{U}_{1k} = O_P\left(\frac{1}{\sqrt{n_v n_t}}\right)$.

$$\begin{aligned} &n\left(\hat{\mathcal{L}}_n^{CV}(M_2) - \hat{\mathcal{L}}_n^{CV}(M_1) - (EPE_{OUT,n}(M_2) - EPE_{OUT,n}(M_1))\right) \\ =&\sum_{k=1}^{K} n_v \tilde{U}_{1k} + o_p(1) = \sum_{k=1}^{K} O_P\left(\frac{n_v}{\sqrt{n_v n_t}}\right) + o_p(1) = O_P(1) \end{aligned}$$

Although there are results in the literature that characterize the asymptotic distribution of a single degenerate two-sample U-statistic of degree (1,1) (see e.g. (Neuhaus, 1977), (Eagleson, 1979), (Dewan & Rao, 2001)), deriving the asymptotic distribution of a sum of K degenerate two-sample U-statistics which are not independent of each other is still fairly complicated. We leave the details for future research and instead use subsampling to estimate the distribution in practice. □

A.4 Proof of Theorem 4.2

First, note that $\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\gamma}_M\left(\xi_i\right)-\gamma_{0M}\left(\xi_i\right)\right|^2\xrightarrow{p}0$ for $M\in\{M_1,M_2\}$ implies

$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|^2\xrightarrow{p}0$ and therefore $\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|\xrightarrow{p}0$ for $\hat{\tilde{\gamma}}\left(\xi_i\right)\equiv\hat{\gamma}_{M_2}\left(\xi_i\right)-\hat{\gamma}_{M_1}\left(\xi_i\right)$ and $\tilde{\gamma}_0\left(\xi_i\right)\equiv\gamma_{0M_2}\left(\xi_i\right)-\gamma_{0M_1}\left(\xi_i\right)$. Additionally, by the law of large numbers $\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0\left(\xi_i\right)\xrightarrow{p}\mathbb{E}\left[\tilde{\gamma}_0\left(\xi_i\right)\right]$. Therefore,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{\gamma}}\left(\xi_i\right)-\mathbb{E}\left[\tilde{\gamma}_0\left(\xi_i\right)\right]\right|\leq\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0\left(\xi_i\right)-\mathbb{E}\left[\tilde{\gamma}_0\left(\xi_i\right)\right]\right|\xrightarrow{p}0$$

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{\gamma}}^2\left(\xi_i\right)-\mathbb{E}\left[\tilde{\gamma}_0^2\left(\xi_i\right)\right]\right|\\&\leq\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|^2+2\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{\gamma}_0\left(\xi_i\right)\right|\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0^2\left(\xi_i\right)-\mathbb{E}\left[\tilde{\gamma}_0^2\left(\xi_i\right)\right]\right|\\&\leq\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|^2+2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{\gamma}_0\left(\xi_i\right)\right|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|^2}+o_p(1)\xrightarrow{p}0\end{aligned}$$

Therefore, $\hat{\tilde{\sigma}}_\gamma^2=\frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{\gamma}}^2\left(\xi_i\right)-\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{\gamma}}\left(\xi_i\right)\right)^2\xrightarrow{p}Var\left[\tilde{\gamma}_0(\xi_i)\right]$. Our assumption of $\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_M\left(\xi_j\right)-\tilde{h}_M\left(\xi_j\right)\right|^2\xrightarrow{p}0$ for $M\in\{M_1,M_2\}$ implies

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\left|\hat{\tilde{h}}\left(\xi_j\right)-\tilde{\tilde{h}}\left(\xi_j\right)\right|^2&=\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_{M_2}\left(\xi_j\right)-\tilde{h}_{M_2}\left(\xi_j\right)-\left(\hat{h}_{M_1}\left(\xi_j\right)-\tilde{h}_{M_1}\left(\xi_j\right)\right)\right|^2\\&\leq\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_{M_2}\left(\xi_j\right)-\tilde{h}_{M_2}\left(\xi_j\right)\right|^2+\frac{1}{n}\sum_{j=1}^{n}\left|\hat{h}_{M_1}\left(\xi_j\right)-\tilde{h}_{M_1}\left(\xi_j\right)\right|^2\\&=o_p(1)\end{aligned}$$

For $\tilde{\tilde{h}}\left(\xi_j\right)\equiv\tilde{h}_{M_2}\left(\xi_j\right)-\tilde{h}_{M_1}\left(\xi_j\right)$, $\frac{1}{n}\sum_{j=1}^{n}\tilde{\tilde{h}}\left(\xi_j\right)^2=\frac{1}{(n-1)^2}\frac{1}{n}\sum_{j=1}^{n}\sum_{i_1\neq j}\sum_{i_2\neq j}\tilde{\psi}\left(\xi_j,\xi_{i_1}\right)\tilde{\psi}\left(\xi_j,\xi_{i_2}\right)$ is asymptotically equivalent to a U-statistic of order 3. Using the Hoeffding decomposition and $\mathbb{E}\left[\tilde{\psi}\left(\xi_j,\xi_{i_1}\right)\tilde{\psi}\left(\xi_j,\xi_{i_2}\right)\middle|\xi_j\right]=\mathbb{E}\left[\tilde{\psi}\left(\xi_j,\xi_i\right)\middle|\xi_j\right]^2=\tilde{h}_0\left(\xi_j\right)$ for $\tilde{h}_0\left(\xi_j\right)\equiv h_{0M_2}\left(\xi_j\right)-$

$h_{0M_1}\left(\xi_j\right)$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{\tilde{h}}\left(\xi_j\right)^2-\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_0\left(\xi_j\right)^2\right|\\
=&\left|\frac{1}{n\left(n-1\right)\left(n-2\right)}\sum_{j=1}^{n}\sum_{i_1\neq j}\sum_{i_2\neq i_1}\tilde{\psi}\left(\xi_j,\xi_{i_1}\right)\tilde{\psi}\left(\xi_j,\xi_{i_2}\right)-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\tilde{\psi}\left(\xi_j,\xi_i\right)\middle|\,\xi_j\right]^2\right|+o_p(1)\\
=&o_p(1)
\end{aligned}$$

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j=1}^{n}\hat{\tilde{h}}\left(\xi_j\right)^2-\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_0\left(\xi_j\right)^2\right|\\
&\leq\frac{1}{n}\sum_{j=1}^{n}\left|\hat{\tilde{h}}\left(\xi_j\right)-\tilde{\tilde{h}}\left(\xi_j\right)\right|^2+2\frac{1}{n}\sum_{j=1}^{n}\left|\tilde{\tilde{h}}\left(\xi_j\right)\right|\left|\hat{\tilde{h}}\left(\xi_j\right)-\tilde{\tilde{h}}\left(\xi_j\right)\right|+\left|\frac{1}{n}\sum_{j=1}^{n}\tilde{\tilde{h}}\left(\xi_j\right)^2-\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_0\left(\xi_j\right)^2\right|\\
&\leq\frac{1}{n}\sum_{j=1}^{n}\left|\hat{\tilde{h}}\left(\xi_j\right)-\tilde{\tilde{h}}\left(\xi_j\right)\right|^2+2\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left|\tilde{\tilde{h}}\left(\xi_j\right)\right|^2}\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left|\hat{\tilde{h}}\left(\xi_j\right)-\tilde{\tilde{h}}\left(\xi_j\right)\right|^2}+o_p(1)\\
&\xrightarrow{p}0
\end{aligned}$$

Since we showed in Theorem 3.2 that $\left|\frac{1}{n}\sum_{j=1}^{n}\hat{h}_M\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}h_{0M}\left(\xi_j\right)\right|=o_p(1)$ for $M\in\{M_1,M_2\}$,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{j=1}^{n}\hat{\tilde{h}}\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_0\left(\xi_j\right)\right|\leq&\left|\frac{1}{n}\sum_{j=1}^{n}\hat{h}_{M_2}\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}h_{0M_2}\left(\xi_j\right)\right|\\
&+\left|\frac{1}{n}\sum_{j=1}^{n}\hat{h}_{M_1}\left(\xi_j\right)-\frac{1}{n}\sum_{j=1}^{n}h_{0M_1}\left(\xi_j\right)\right|\\
&\xrightarrow{p}0
\end{aligned}$$

Since $\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_0\left(\xi_j\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_0\left(\xi_j\right)\right)^2\xrightarrow{p}Var\left(\tilde{h}_0\left(\xi_j\right)\right)$, it follows that $\hat{\tilde{\sigma}}_{01,\psi}^2=\frac{1}{n}\sum_{j=1}^{n}\hat{\tilde{h}}\left(\xi_j\right)^2-\left(\frac{1}{n}\sum_{j=1}^{n}\hat{\tilde{h}}\left(\xi_j\right)\right)^2\xrightarrow{p}Var\left(\tilde{h}_0\left(\xi_j\right)\right)$.

Next, $\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0\left(\xi_i\right)\tilde{\tilde{h}}\left(\xi_i\right)-\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0\left(\xi_i\right)\tilde{h}_0\left(\xi_i\right)\right|\xrightarrow{p}0$ since for $M,M'\in\{M_1,M_2\}$, $\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_{0M}\left(\xi_i\right)\tilde{h}_{M'}\left(\xi_i\right)-\frac{1}{n}\sum_{i=1}^{n}\gamma_{0M}\left(\xi_i\right)h_{0M'}\left(\xi_i\right)\right|\xrightarrow{p}0$. Since also $\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{h}}\left(\xi_i\right)-\tilde{\tilde{h}}\left(\xi_i\right)\right|\xrightarrow{p}$

0 and $\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right| \xrightarrow{p} 0$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{\gamma}}\left(\xi_i\right)\hat{\tilde{h}}\left(\xi_i\right)-\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0\left(\xi_i\right)\tilde{h}_0\left(\xi_i\right)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right)\hat{\tilde{h}}\left(\xi_i\right)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tilde{h}}\left(\xi_i\right)-\tilde{\tilde{h}}\left(\xi_i\right)\right)\tilde{\gamma}_0\left(\xi_i\right)\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0\left(\xi_i\right)\tilde{\tilde{h}}\left(\xi_i\right)-\frac{1}{n}\sum_{i=1}^{n}\tilde{\gamma}_0\left(\xi_i\right)\tilde{h}_0\left(\xi_i\right)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|\left|\hat{\tilde{h}}\left(\xi_i\right)\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{h}}\left(\xi_i\right)-\tilde{\tilde{h}}\left(\xi_i\right)\right|\left|\tilde{\gamma}_0\left(\xi_i\right)\right| + o_p(1) \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{\gamma}}\left(\xi_i\right)-\tilde{\gamma}_0\left(\xi_i\right)\right|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{h}}\left(\xi_i\right)\right|^2} \\
&+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\tilde{h}}\left(\xi_i\right)-\tilde{\tilde{h}}\left(\xi_i\right)\right|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{\gamma}_0\left(\xi_i\right)\right|^2} + o_p(1) \\
&= o_p\left(1\right)O_p\left(1\right) + o_p\left(1\right)O_p\left(1\right) = o_p\left(1\right)
\end{aligned}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\gamma}_0(\xi_i)\tilde{h}_0(\xi_i)-E\left[\tilde{\gamma}_0(\xi_i)\tilde{h}_0(\xi_i)\right]\right) \xrightarrow{p} 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tilde{\gamma}}(\xi_i)\hat{\tilde{h}}(\xi_i)-E\left[\tilde{\gamma}_0(\xi_i)\tilde{h}_0(\xi_i)\right]\right) \xrightarrow{p} 0$$

Since also $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tilde{\gamma}}(\xi_i)-E\left[\tilde{\gamma}_0(\xi_i)\right]\right) \xrightarrow{p} 0$ and $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tilde{h}}(\xi_i)-E\left[\tilde{h}_0(\xi_i)\right]\right) \xrightarrow{p} 0$,

$$\hat{\tilde{\lambda}} \xrightarrow{p} \tilde{\lambda} \equiv Cov\left[\tilde{\gamma}_0(\xi_i),\tilde{h}_0(\xi_i)\right]$$

□

References

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Austern, M., & Zhou, W. (2020). Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111.*

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*, 1–12.

Bayle, P., Bayle, A., Janson, L., & Mackey, L. (2020). Cross-validation confidence intervals for test error. *34th Conference on Neural Information Processing Systems (NeurIPS 2020).*

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, *20*(1), 105–134.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, *76*(3), 503–514.

Burman, P., Chow, E., & Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, *81*(2), 351–358.

Chen, X., Hong, H., & Shum, M. (2007). Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. *Journal of Econometrics*, *141*(1), 109–140.

Chen, X., & White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, *45*(2), 682–691.

Clark, T. E., & McCracken, M. W. (2015). Nested forecast model comparisons: a new approach to testing equal accuracy. *Journal of Econometrics*, *186*(1), 160–177.

Dewan, I., & Rao, B. P. (2001). Asymptotic normality for u-statistics of associated random variables. *Journal of Statistical Planning and Inference*, *97*(2), 201–225.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, *33*(1), 1–1.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Eagleson, G. (1979). Orthogonal expansions and u-statistics. *Australian Journal of Statistics*, *21*(3), 221–237.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Hall, P., Racine, J., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, *99*(468), 1015–1026.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*, Vol. 2. Springer.

Hong, H., & Preston, B. (2012). Bayesian averaging, prediction and nonnested model selection. *Journal of Econometrics*, *167*(2), 358–369.

Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, *28*(5), 1356–1378.

Lavergne, P. (2001). An equality test across nonparametric regressions. *Journal of Econometrics*, *103*(1-2), 307–344.

Lavergne, P., & Vuong, Q. (2000). Nonparametric significance testing. *Econometric Theory*, *16*(4), 576–601.

Lavergne, P., & Vuong, Q. H. (1996). Nonparametric selection of regressors: The nonnested case. *Econometrica: Journal of the Econometric Society*, 207–219.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, 165–179.

Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses (Springer Texts in Statistics)*. Springer.

Lei, J. (2019). Cross-validation with confidence. *Journal of the American Statistical Association*, 1–20.

Li, Q., & Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, 485–512.

Liao, Z., & Shi, X. (2020). A nondegenerate vuong test and post selection confidence intervals for semi/nonparametric models. *Quantitative Economics*, *11*(3), 983–1017.

Neuhaus, G. (1977). Functional limit theorems for u-statistics in the degenerate case. *Journal of Multivariate Analysis*, *7*(3), 424–439.

Neumeyer, N. (2004). A central limit theorem for two-sample u-processes. *Statistics and Probability Letters*, *67*, 73–85.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Politis, D., Romano, J., & Wolf, M. (1999). *Subsampling.* Springer Series in Statistics.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, *99*, 39–61.

Rivers, D., & Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *The Econometrics Journal*, *5*(1), 1–39.

Romano, J. P., Shaikh, A. M., & Wolf, M. (2010). Hypothesis testing in econometrics. *Annual Review of Economics*, *2*(1), 75–104.

Schennach, S. M., & Wilhelm, D. (2017). A simple parametric model selection test. *Journal of the American Statistical Association*, *112*(520), 1663–1674.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Shi, X. (2015). A non-degenerate vuong test. *Quantitative Economics*, 85–121.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Vol. 3. Cambridge university press.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333.

Wager, S. (2020). Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani. *Journal of the American Statistical Association*, *115*(529), 157–160.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, 1067–1084.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, *84*(408), 1003–1013.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.

White, H., & Racine, J. (2001). Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates. *IEEE Transactions on Neural Networks*, *12*(4), 657–673.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, *35*(6), 2450–2473.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 299–313.

Zhu, Y., & Timmermann, A. (2020). Can two forecasts have the same conditional expected accuracy?. *arXiv preprint arXiv:2006.03238*.