Bernoulli

Targeted Cross-Validation

JIAWEI ZHANG* JIE DING** YUHONG YANG†

School of Statistics, University of Minnesota, MN 55455, USA.

E-mail:*zhan4362@umn.edu, **dingj@umn.edu, †yangx374@umn.edu

In many applications, we have access to the complete dataset but are only interested in the prediction of a particular region of predictor variables. A standard approach is to find the globally best modeling method from a set of candidate methods. However, it is perhaps rare in reality that one candidate method is uniformly better than the others. A natural approach for this scenario is to apply a weighted L_2 loss in performance assessment to reflect the region-specific interest. We propose a targeted cross-validation (TCV) to select models or procedures based on a general weighted L_2 loss. We show that the TCV is consistent in selecting the best performing candidate under the weighted L_2 loss. Experimental studies are used to demonstrate the use of TCV and its potential advantage over the global CV or the approach of using only local data for modeling a local region.

Previous investigations on CV have relied on the condition that when the sample size is large enough, the ranking of two candidates stays the same. However, in many applications with the setup of changing data-generating processes or highly adaptive modeling methods, the relative performance of the methods is not static as the sample size varies. Even with a fixed data-generating process, it is possible that the ranking of two methods switches infinitely many times. In this work, we broaden the concept of the selection consistency by allowing the best candidate to switch as the sample size varies, and then establish the consistency of the TCV. This flexible framework can be applied to high-dimensional and complex machine learning scenarios where the relative performances of modeling procedures are dynamic.

Keywords: Consistency; Cross-validation; Model selection; Regression

1. Introduction

Cross-validation (CV) is one of the most powerful tools for selecting models or procedures. Different CV methods include leave-one-out [1, 15, 29], leave-p-out [25, 36], V-fold [15], repeated learning testing [7, 8, 36], Monte-Carlo CV [23], and generalized CV [11].

There are various works related to the asymptotic properties of CV. For model selection, the consistency of CV in both linear regression and time series models has been studied, e.g., [19], [25, 26], and [24]. For a broader setting of selecting general models or modeling procedures, [31, 32] provided conditions for consistency of CV in the context of regression and classification. The application of CV to selecting a model selection procedure in a high-dimensional setting was studied in [37]. Apart from CV, the methods from [5] and [6] can also be used to select general modeling procedures for function estimation. It has been shown that CV can select tuning parameters for optimal nonparametric estimations such as the Nadaraya-Watson estimator [30], smoothing spline [11, 27], and nearest neighbor method [18]. A comprehensive summary about CV-related works can be found in [2, 12]. More recently, [3] designed a CV-based method to detect change points in the heteroscedastic framework. A closed-form expression of risks from the leave-p-out CV, which provides insights into the choice of p in both estimation and identification problems, was derived in [10]. A non-asymptotic oracle inequality for the V-fold CV was obtained in [4] , and it shows that the V-fold CV is asymptotically optimal when $V \to \infty$ in a nonparametric setting. A classification procedure that combines CV with aggregation was introduced in [20]. A consistent CV procedure for selecting high-dimensional generalized linear models was studied in [14]. A CV-based method that selects a subset of candidate models containing the best one with high probability was introduced in [17].

In various applications, we may want to select a candidate method with the best performance in a small region of interest. One may attempt to build a model for the region alone. However, this may not be a good solution because the local data are often limited for efficient estimation, and the full data may help a good candidate model or procedure achieve optimal or near-optimal performance. The current studies about CV focus on selecting the candidate method with the best global performance and do not apply to finding the best one for a specific region. An effort was made in [33], where the issue of selecting the best candidate at a single point has been studied, and the data-generating model is considered fixed.

In this work, we propose a method named targeted CV (TCV) for the above problem of selecting an optimal candidate method for any particular region of interest. Our method allows for the consideration of flexible high-dimensional regression methods as candidates. More generally, it incorporates a weight function to reflect where the comparison of the candidate models or procedures is of most interest. A straightforward way for the weight is to assign a 0/1 value according to whether an observation is in the region. Compared with CV without weight, which selects the best global model, the TCV may work better than the regular CV when the globally best candidate does not perform uniformly the best. Additionally, the TCV can be used to compare the methods applied to the complete dataset with those based on the local data only. On the one hand, candidate methods based on the complete dataset have the advantage of a larger sample size. On the other hand, the data outside the local region may introduce undesirable biases. Fortunately, the TCV provides a data-adaptive way to compare them.

For the intended application of our TCV, the candidates are allowed to include non-model-based procedures in addition to models. We have found an apparently ignored but important aspect in previous theoretical developments on selection consistency when comparing general learning procedures. The issue is that the existing results assume that one procedure stays the best as the sample size approaches infinity. However, this view ignores the fact that in many applications, the comparison of the competing procedures is dynamic. Specifically, the performance ranking of the candidate procedures may keep changing as the sample size varies, especially when the candidate methods are highly adaptive and evolving with the sample size. We will provide an example to show that even with a fixed data-generating process, the ranking of two sensible models changes infinitely many times. To accommodate for this inevitable complication in reality, we enable the TCV to work for a triangular array setup where the best candidate method may not be fixed. Under this broad setting, the goal is to find out the best candidate method under the current sample size. To define the best candidate in an asymptotic sense, we introduce a new concept of performance comparison elaborated in Section 3.

The outline of the paper is given below. Section 2 defines the problem. Section 3 introduces new concepts for performance comparison. Section 4 introduces our TCV and shows its consistency, Section 5 presents the numerical studies. Section 6 concludes the paper. The appendixes include the proofs.

2. Problem

We consider the random design regression model $Y_i = f(\boldsymbol{X}_i) + \varepsilon_i$, where $1 \le i \le n$. Predictors $\boldsymbol{X}_i = (X_i^{(1)}, \cdots X_i^{(p)})$ are independent and identically distributed p-dimensional random variables. For each $k \in \{1, 2, \cdots, p\}$, $X_i^{(k)}$ can be either continuous or discrete. Let $P_{\boldsymbol{X}}$ denote the joint probability distribution of the predictors and $\mathcal{S}$ denote the domain of $\boldsymbol{X}$. Let $\varepsilon_1, \cdots, \varepsilon_n$ be independent random errors with $E(\varepsilon_i|\boldsymbol{X}_i) = 0$ and $E(\varepsilon_i^2|\boldsymbol{X}_i) < \infty$ almost surely. Let $W_n(\boldsymbol{x})$ be a nonnegative weight function that satisfies

$$\int_{\mathcal{S}} W_n(\boldsymbol{x}) P_{\boldsymbol{X}}(d\boldsymbol{x}) = 1. \tag{2.1}$$

We define the weighted L_q norm $\|f\|_{q,W_n} = \left(\int_{\mathcal{S}} W_n(\boldsymbol{x}) \cdot |f(\boldsymbol{x})|^q P_{\boldsymbol{X}}(d\boldsymbol{x})\right)^{1/q}$, where $0 < q < \infty$. We require that $\|f\|_{2,W_n} < \infty$ and ess-sup$|W_n| < \infty$.

We have a set of candidate regression procedures $\delta_j \in \mathcal{M}$, where $j \in \mathcal{J} = \{1, 2, \ldots, m\}$, and each candidate is based on a regression model or a general procedure. The set $\mathcal{M}$ may possibly change with n, but the number of candidate methods m is upper bounded by a fixed constant. Let $\widehat{f}_n^{(j)}$ be the fitted regression function from δ_j with training data size n. We want to find the candidate δ_{j^*} with $j^* = \arg\min_{j \in \mathcal{J}} \|f - \widehat{f}_n^{(j)}\|_{2,W_n}$. Three examples of the weight function are as follows.

Example 1. (Region-based weight) Suppose we are only interested in the performances of the candidate methods in a fixed region A. We can take an indicator weight $W_n(\boldsymbol{x}) = C^{-1}\mathbb{1}(\boldsymbol{x} \in A)$ where C equals the probability of $\boldsymbol{X} \in A$ and $\mathbb{1}(\cdot)$ is the indicator function.

Example 2. (Conditional variance-based weight) Suppose we know that the conditional variance of the response satisfies $Var(Y_i|\boldsymbol{X}_i) = Var(\epsilon_i) = \sigma^2(\boldsymbol{X}_i)$, where σ^2 is a positive function of predictors such that $\int_{\mathcal{S}} 1/\sigma^2(\boldsymbol{x}) P_{\boldsymbol{X}}(d\boldsymbol{x}) < \infty$. We can adjust the L_2 loss according to the conditional variance by

$$W_n(\boldsymbol{x}) = \frac{1/\sigma^2(\boldsymbol{x})}{\int_{\mathcal{S}} 1/\sigma^2(\boldsymbol{x}) P_{\boldsymbol{X}}(d\boldsymbol{x})}.$$

Example 3. (Single point-based weight) Suppose we are interested in modeling $f(\boldsymbol{x})$ evaluated at a single point $\boldsymbol{x} = \boldsymbol{x}^*$. We can define the weight by a positive function centered at $\boldsymbol{x}^*$ that shrinks toward that point as the sample size n goes to infinity. For instance,

$$W_n(\boldsymbol{x}) = \frac{\exp\left(-\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \cdot n\right)}{\int_{\mathcal{S}} \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \cdot n\right) P_{\boldsymbol{X}}(d\boldsymbol{x})}.$$

3. A Neglected Aspect in Selection Consistency Theories for CV

In Section 3.1, we address the need to extend the scope of the CV to a more flexible framework that is based on the triangular array setting with possibly changing data-generating distributions. We also present the definition of the best candidate method under the extended setting. In Section 3.2, we exemplify the choice of splitting ratios needed to identify the best candidate method.

3.1. A triangular array setting

The selection consistency of CV has been established both for parametric model selection and for procedure selection. For the former, the candidates are assumed to be fixed linear models [25]. As the sample size increases to infinity, there is no ambiguity in terms of which model is the best. For the latter, results that allow the inclusion of general regression procedures are given in [32] and [37]. A major limitation of these theoretical results is that they assume a static ordering of the candidate procedures in terms of performance. In many applications, however, the relative performances among the candidate procedures may be dynamically changing with the sample size even if the true data-generating process stays fixed (see Section 4.3). Moreover, modeling methods for high-dimensional data usually assume changing true sparsity or true coefficients in deriving theoretical properties [13, 35], which implicitly indicate a triangular array setup.

Now, consider two variable selection methods, and the goal is to choose between them. Suppose they are known to perform optimally under severe or moderate sparsity, respectively (e.g., the number of non-zero coefficients being of order $\log n$ and $n^{1/10}$, respectively). In this context, for an interesting theoretical investigation, it makes most sense to allow the unknown data generating model and the best candidate to change according to the sample size. This is just one example of situations where the relative ranking is not static due to the fact that the modeling procedures may react quite differently with more or fewer observations. The present CV framework unfortunately cannot handle this reality. Thus, it is essential to explicitly set up a flexible framework that gives each candidate method a chance to work better and confront the reality of possibly changing relative performances. Otherwise, a fixed truth or a rigid triangular array setup may lead to the conclusion that one of the two methods would always be preferred when n is large enough, which is detached from many real applications. The above reasons motivate us to study the consistency of the TCV under the following triangular array framework, where data are represented in a triangular array.

$$
\begin{aligned}
&(\boldsymbol{X}_{1,1}, Y_{1,1}) \\
&(\boldsymbol{X}_{2,1}, Y_{2,1}), (\boldsymbol{X}_{2,2}, Y_{2,2}) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&(\boldsymbol{X}_{n,1}, Y_{n,1}), (\boldsymbol{X}_{n,2}, Y_{n,2}), \cdots, (\boldsymbol{X}_{n,n}, Y_{n,n}) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots .
\end{aligned}
$$

In each row, the random pairs are i.i.d. For each n, $\boldsymbol{X}_{n,1}, \cdots \boldsymbol{X}_{n,n}$ follow the distribution $P_{\boldsymbol{X}_n}$, and the weighted L_2 norm is defined according to $P_{\boldsymbol{X}_n}$. Under this setting, the goal is to find out the best candidate method given the current sample size.

The fitted regression function $\widehat{f}_{n_1}^{(j)}$ with $j \in \mathcal{J}$ was obtained by first sampling (without replacement) n_1 observations as the training set and then applying the candidate methods δ_j with $j \in \mathcal{J}$ to this dataset. Let $c_{(n_1,n)}$ be a sequence of positive numbers. Let l_n be a sequence of positive integers such that $l_n < n$, and $l_n \to \infty$ as $n \to \infty$. For illustrative purposes, we first focus on the case with two candidate methods. Let g_n and b_n be two sequences that take values from $\mathcal{J} = \{1, 2\}$ and satisfy $g_n + b_n = 3$, where g_n stands for "good" and b_n for "bad".

Definition 1 (($W_n, l_n, c_{(n_1,n)}$)-better)**.** *The candidate method δ_{g_n} is said to be the $(W_n, l_n, c_{(n_1,n)})$-better one asymptotically out of the two candidates δ_1, δ_2 at sample size n if for all $l_n \le n_1 < n$, we have*

$$
P(\|f - \widehat{f}_{n_1}^{(b_n)}\|_{2,W_n} \ge (1 + c_{(n_1,n)})\|f - \widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}) \to 1, \tag{3.1}
$$

as $n \to \infty$.

The quantities involved in the above definition capture the key aspects in the TCV comparison. The sequence $c_{(n_1,n)}$ characterizes the difference between the losses of the two candidate methods. It is related to n_1 in the sense that the convergence rate of the fitted regression function depends on the training data size. It may also depend on n since the data-generating process, weight W_n, and better candidate g_n may change with n. In the high-dimensional setting with the number of predictor variables increasing to infinity, $c_{(n_1,n)}$ may need to go to 0 when the two candidate methods are very close. An example concerns the choice of $c_{(n_1,n)}$ for comparing the underlying true model with an over-fitting model with one additional term. As will be seen in our main theorem, the TCV will require a higher

portion of the test data to handle the challenge of a decreasing performance difference in such a case. If l_n is much smaller than n, then for a wide range of choices of n_1 for the TCV, the comparison result of δ_{g_n} and δ_{b_n} at the reduced sample size n_1 matches that at the full sample size. In contrast, suppose for instance, δ_{g_n} is $(W_n, l_n, c_{(n_1,n)})$-better than δ_{b_n}, but δ_{g_n} is not better even at a slightly reduced sample size $n_1 < n$, then one may not be able to tell which candidate is better at the full sample size n when data splitting is done. In this case, it is hard to get a consistent selection for the TCV, as expected.

In the general case where $\mathcal{M}$ may contain more than two candidate methods, we let g_n denote the to-be-defined best candidate and

$$\mathcal{J}_b \triangleq \{j \in \mathcal{J} : j \neq g_n\}. \tag{3.2}$$

We define the following extension of Definition 1.

Definition 2 ($(W_n, l_n, c_{(n_1,n)})$-best)**.** *A candidate method δ_{g_n} from $\mathcal{M}$ is said to be the $(W_n, l_n, c_{(n_1,n)})$-best one asymptotically if there exist $0 < l_n < n$ and $c_{(n_1,n)} > 0$, such that the method δ_{g_n} is $(W_n, l_n, c_{(n_1,n)})$-better than δ_i for each $i \in \mathcal{J}_b$.*

3.2. An example of the changing l_n

Recall that under the $(W_n, l_n, c_{(n_1,n)})$-better condition, δ_{g_n} is better than δ_{b_n} as long as the training set size n_1 is larger than or equal to a lower bound l_n. In this subsection, we provide a toy example of l_n, which shows that the increasing speed of n_1 cannot be too slow compared with n, in order to guarantee that the performance ranking of the candidate methods at the sample size n_1 remains the same as that of n.

We consider the data-generating process $Y_i = f(X_i) + \varepsilon_i$, where $f(x) = x^2$, ε_i's are i.i.d. from $N(0, \sigma^2)$, and X_i's are i.i.d. from $U(0, 1)$. We are interested in estimating $f(x)$ where x is close to 0. We consider

$$W_n = \begin{cases} C^{-1} & \text{if } 0 \le x \le n^{-\frac{1}{8}}, \\ 0 & \text{otherwise,} \end{cases}$$

where the normalizing constant $C = \int_0^{n^{-1/8}} dx = n^{-1/8}$. The candidate models are

$$\text{Model 1: } \widehat{f}_{n_1}^{(1)}(x) = x^2 + \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - x_i^2) = f(x) + \frac{1}{n_1}\sum_{i=1}^{n_1}\varepsilon_i,$$

$$\text{Model 2: } \widehat{f}_{n_1}^{(2)}(x) \equiv 0.$$

For model 1, we have

$$\|f - \widehat{f}_{n_1}^{(1)}\|_{2,W_n}^2 = \int_0^{n^{-\frac{1}{8}}} \left(f(x) - f(x) - \frac{1}{n_1}\sum_{i=1}^{n_1}\varepsilon_i\right)^2 \cdot C^{-1} dx = \left(\frac{1}{n_1}\sum_{i=1}^{n_1}\varepsilon_i\right)^2.$$

Therefore, $\|f - \widehat{f}_{n_1}^{(1)}\|_{2,W_n}^2$ converges at the rate n_1^{-1}. For model 2,

$$\|f - \widehat{f}_{n_1}^{(2)}\|_{2,W_n}^2 = \int_0^{n^{-\frac{1}{8}}} (f(x) - 0)^2 \cdot C^{-1} dx = \int_0^{n^{-\frac{1}{8}}} x^4 dx \cdot n^{\frac{1}{8}} = \frac{1}{5} \cdot n^{-\frac{1}{2}}.$$

Thus, $\|f-\widehat{f}_{n_1}^{(2)}\|_{2,W_n}^2$ converges exactly at rate $n^{-\frac{1}{2}}$, regardless of n_1.

For the estimators $\widehat{f}_n^{(1)}$ and $\widehat{f}_n^{(2)}$ using the complete dataset, the convergence rates of $\|f-\widehat{f}_n^{(1)}\|_{2,W_n}^2$ and $\|f-\widehat{f}_n^{(2)}\|_{2,W_n}^2$ are n^{-1} and $n^{-\frac{1}{2}}$, respectively. Therefore, the weighted L_2 loss of $\widehat{f}_n^{(1)}$ converges faster than the weighted L_2 loss of $\widehat{f}_n^{(2)}$. However, in order to get the same ranking of $\widehat{f}_{n_1}^{(1)}$ and $\widehat{f}_{n_1}^{(2)}$ based on the training data with size n_1, we need $n_1/\sqrt{n}\to\infty$ as $n\to\infty$.

The above example is meant for a quick illustration. In reality, suppose that δ_1 and δ_2 denote two teams of data scientists participating in an online data competition. It is conceivable that the two teams may try various learning tools with validation feedback on their performances, and their relative ranking may not be static. In this case, to fairly evaluate their performances at a reduced sample size $n_1<n$, the lower bound l_n needs to be carefully chosen.

4. Method and Main Result

In Section 4.1, we present the TCV method. In Section 4.2, we introduce the theoretical result that shows the model selection consistency of the TCV. In Section 4.3, we present a nonparametric regression example with alternating best candidate method and verify the requirements for the property of the TCV. In Section 4.4, we extend the theoretical result of the TCV from a single splitting to multiple splittings that aim to stabilize the selection result.

4.1. Targeted cross-validation

Recall that we have randomly partitioned the dataset into a training set with size n_1 and a test set with size $n_2=n-n_2$. We define our weighted squared prediction error by

$$TCV_{W_n}(\widehat{f}_{n_1}^{(j)})=\sum_{i=n_1+1}^{n}(Y_i-\widehat{f}_{n_1}^{(j)}(\boldsymbol{X}_i))^2\cdot W_n(\boldsymbol{X}_i), \tag{4.1}$$

where $j\in\mathcal{J}$ and the summation is taken over the test set, and for notational simplicity, $(\boldsymbol{X_i},Y_i)$ denote $(\boldsymbol{X_{n,i}},Y_{n,i})$. The TCV selects the candidate $\hat{j}=\underset{j\in\mathcal{J}}{\arg\min}\, TCV_{W_n}(\widehat{f}_{n_1}^{(j)})$.

4.2. The main theorem

We first introduce some necessary definitions and conditions.

Definition 3 (W_n-consistent selection)**.** *We assume that there exists a candidate method δ_{g_n} that is the $(W_n,l_n,c_{(n_1,n)})$-best out of $\mathcal{M}$ at sample size n. A selection rule is called W_n-consistent if its probability of selecting δ_{g_n} goes to 1 as $n\to\infty$.*

Definition 4 (Lower bound of the rate of the weighted L_2 loss)**.** *A fitted regression function $\widehat{f}_{n_1}$ is said to have (W_n,l_n)-convergence rate lower bounded by $a_{(n_1,n)}$ under the weighted L_2 loss if for each $0<\epsilon<1$, there exist $c_\epsilon>0, N\in\mathbb{Z}^+$ such that for all $n\geq N$ and $l_n\leq n_1\leq n$,*

$$P\left(\|f-\widehat{f}_{n_1}\|_{2,W_n}\geq c_\epsilon a_{(n_1,n)}\right)\geq 1-\epsilon. \tag{4.2}$$

Condition 1 (Error variances)**.** *The error variances $E(\varepsilon_i^2|\boldsymbol{X}_i)$ are upper bounded by a constant $\overline{\sigma}^2 > 0$ almost surely for all $i \geq 1$.*

Condition 2 (Relating weighted L_4 and L_2 losses)**.** *There exists a sequence of positive numbers M_n such that for all $l_n \leq n_1 < n$,*

$$\sup_{j\in\mathcal{J}} \big(\|f - \widehat{f}_{n_1}^{(j)}\|_{4,W_n} / \|f - \widehat{f}_{n_1}^{(j)}\|_{2,W_n}\big) = O_p(M_n),$$

as $n \to \infty$.

Condition 3 (Lower bound of the convergence rates)**.** *There exists a sequence $q_{(n_1,n)}$ such that for each $j \in \mathcal{J}_b$, $\widehat{f}_{n_1}^{(j)}$ has (W_n, l_n)-converge rate lower bounded by $q_{(n_1,n)}$ under the weighted L_2 loss.*

Condition 1 is a mild requirement that is satisfied when, e.g., the random errors have the same finite variance. For Condition 2, it has been shown that for some familiar function classes, the estimators may have the same rates for both L_4 and L_2 losses. For instance, Lipschitz class [Section 1.2, 22], Hölder class (see, e.g., Section 1.3 of [22] and [28]). and Sobolev class [Section 2.1&2.2, 22]. Also, an example that compares AIC-based and BIC-based selection procedures with $M_n \equiv 1$ can be found in Section 3 of [37]. For Condition 3, similar to $c_{(n_1,n)}$, both n_1 and n are involved in determining the rate $q_{(n_1,n)}$. For instance, in the example from Section 3.2, when $n_1/\sqrt{n} \to \infty$ as $n \to \infty$, we have that $q_{(n_1,n)} = 1/\sqrt{n}$.

Let S_{W_n} denote $\operatorname*{ess\text{-}sup}_{\boldsymbol{x}\in\mathcal{S}}(W_n(\boldsymbol{x}))$. Our main theorem is as follows.

Theorem 1 (W_n-consistency of the TCV)**.** *Assume that Conditions 1-3 hold, and the data splitting is such that for all $l_n \leq n_1 < n$, we have*

(i). $n_2 \cdot c_{(n_1,n)}^2 / (S_{W_n} M_n^4) \to \infty$,

(ii). $n_2 \cdot (c_{(n_1,n)} \cdot q_{(n_1,n)})^2 / S_{W_n} \to \infty$,

as $n \to \infty$. Then, the TCV is W_n-consistent.

The detailed proof can be found in Appendix 7. The requirements ***(i)*** and ***(ii)*** indicate that the TCV needs the test size n_2 to be sufficiently large. In the case of comparing nested high-dimensional regression models with a fixed number of additional predictors, under some mild conditions from Section 4.2 of [37], we have $(c_{(n_1,n)} \cdot q_{(n_1,n)})^2 = O_p(1/n_1)$. Then, if $M_n = 1$ and both S_{W_n} and $q_{(n_1,n)})$ are bounded above by fixed constants, the requirements (***i***) and (***ii***) can be simplified by $n_2 \to \infty$ and $n_2/n_1 \to \infty$. This requirement on the splitting ratio is in accordance with those from Theorem 1 of [25] for classical linear regression and Theorem 3.3 from [10] for density estimation. It is interesting to note that this splitting ratio direction may be opposite to that for risk prediction or asymptotically optimal estimation in some contexts, which need $n_2/n_1 \to 0$ (see, e.g., [4, 8, 9]). When $W_n(x)$ is chosen to focus on a region R_n with decreasing probability, which implies that S_{W_n} goes to infinity, the two conditions (***i***) and (***ii***) require n_2 to be larger compared with the global CV. In such cases, it is crucial to have enough evaluation data points in order to separate close competitors on a small region.

Compared with former related works, e.g., [32] and [37], our theoretical results differ in three major aspects. First, our theory takes the effect of the weight into account. It applies to a broader range of CV applications such as illustrated by Examples 1 to 3 in Section 2, and our result highlights the need to adjust the data splitting ratio accordingly. Second, our method works in a more general and realistic

framework with possibly changing relative performances of the candidate methods. Third, with an improved derivation, we have removed the requirements on the upper bound of the sup-norm loss and exact rate of the L_2 loss as those in Condition 1 and Definition 3 of [32] and Conditions 1 and 4 of [37].

4.3. An illustrative example with an alternating better candidate method

In this subsection, we present an example where we observe the alternating relative performances between two candidate methods. This example is particularly interesting in that the data-generating process is fixed. We also verify the conditions for the TCV and provide a valid range for the required data splitting ratio.

We consider the i.i.d. data-generating process $y = f(x) + \epsilon$, where $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$, $x \sim U(0,1)$, the random error ϵ, assumed to be independent of x, has mean zero and variance one, and for convenience, $\phi_j(x) = \sqrt{2}\sin 4^j \pi x$. Let $\mathbb{S} = \{2^{(12\times 3^{q-1})} : q \in \mathbb{N}\}$, where $\mathbb{N}$ denotes the set of natural numbers. Let $\mathbb{S}_1$ denote the subset of $\mathbb{S}$ with odd q and $\mathbb{S}_2$ denote those with even q. The data-generating coefficients are

$$\beta_j = \begin{cases} 0 & \text{when } j \in \cup_{\nu\in\mathbb{S}_1}[\lfloor \nu^{1/12}\rfloor, \lfloor \nu^{1/4}\rfloor], \\ \frac{1}{j^2} & \text{otherwise,} \end{cases}$$

where $\lfloor x \rfloor$ stands for the largest integer no bigger than x. Our candidate models with training set size n_1 are

$$\text{Model 1:} \quad \hat{f}^{(1)}_{n_1} = \sum_{j=1}^{p^{(1)}_{n_1}} \hat{\beta}^{(n_1)}_j \phi_j(x), \quad \text{Model 2:} \quad \hat{f}^{(2)}_{n_1} = \sum_{j=1}^{p^{(2)}_{n_1}} \hat{\beta}^{(n_1)}_j \phi_j(x),$$

where $\hat{\beta}^{(n_1)}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i \phi_j(x_i)$, $p^{(1)}_{n_1} = \lfloor n_1^{1/4} \rfloor - 1$, and $p^{(2)}_{n_1} = \lfloor n_1^{1/4} \rfloor$. This definition is valid if we consider $n_1 \geq 16$. The coefficient $\beta_{p^{(2)}_{n_1}}$, which corresponds to the additional variable in model 2 not in model 1, has an alternating pattern as shown in Figure 1. We want to find out the better candidate model when $n \in \mathbb{S}$.

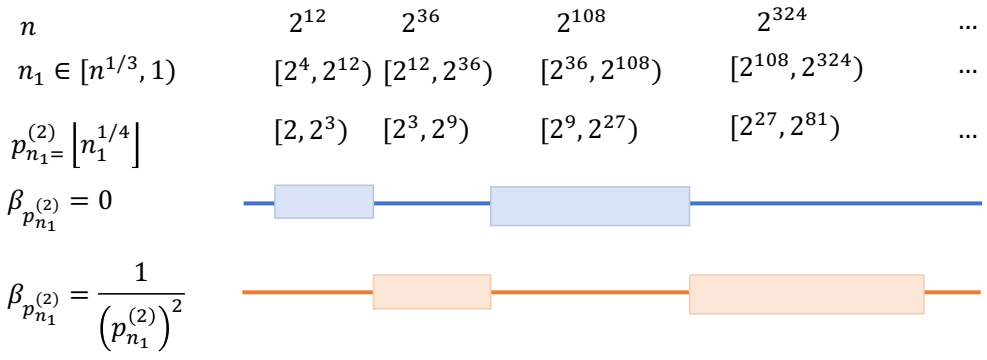

Figure 1: An illustration of $\beta_{p^{(2)}_{n_1}}$ when $n_1 \in [n^{1/3}, n)$ and $n \in \mathbb{S}$.

Proposition 1 (Alternating relative performance)**.** *Take* $l_n = n^{1/3}$, $n_1 \in [l_n, n)$, *and* $W_n(\boldsymbol{x}) \equiv 1$. *Then, when* $n \in \mathbb{S}_1$, *model 1 is* $(W_n, l_n, n_1^{-1/3}/3)$*-better than model 2, and when* $n \in \mathbb{S}_2$, *model 2 is* $(W_n, l_n, n_1^{-1/3}/3)$*-better than model 1.*

Proposition 2 (Verifying the requirements for the TCV)**.** *Assume that* $n_2/n_1^{17/12} \to \infty$ *as* $n \to \infty$ *With* $c_{(n_1,n)} = n_1^{-1/3}/3$ *and* $W_n(\boldsymbol{x}) \equiv 1$, *we have*

i. $n_2 \cdot c_{(n_1,n)}^2 / M_n^4 \to \infty$,
ii. $n_2 \cdot (c_{(n_1,n)} \cdot q_{n_1,n})^2 \to \infty$,

as $n \to \infty$. *That is, the requirements for consistency of the TCV in this example are satisfied.*

The proofs are in Appendices 8 and 9.

4.4. Multiple data splittings

In practice, we apply the TCV with multiple data splittings in order to lower the variability of the selection result. The need of a number of data splittings to achieve stability in the outcome of selection is numerically demonstrated in [34]. The training set and test set in each single splitting are assumed to be independent. With any multiple-splitting method, we obtain a set of MSEs for each candidate. Next, we introduce two ways to combine the results. Let $\widehat{f}_{n_1,k}^{(j)}$ with $j \in \mathcal{J}$ be the estimators from the candidate procedures based on the training set in the k-th splitting. Let K denote the total number of splittings. We define

$$MTCV_{W_n}^a(\delta_j) = \frac{1}{K} \sum_{k=1}^{K} TCV_{W_n,k}(\widehat{f}_{n_1,k}^{(j)}), \tag{4.3}$$

$$MTCV_{W_n}^v(\delta_j) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(j = \arg\min_{i \in \mathcal{J}} TCV_{W_n,k}(\widehat{f}_{n_1,k}^{(i)})), \tag{4.4}$$

for the multiple splitting TCV by averaging and by voting, respectively. The two multiple splitting TCV methods will select $\delta_j \in \mathcal{M}$ that minimizes $MTCV_{W_n}^a(\delta_j)$ and maximizes $MTCV_{W_n}^v(\delta_j)$, respectively. The following corollary shows that the TCV with multiple splittings is also W_n-consistent.

Corollary 1 (Consistency of the TCV with multiple splitting)**.** *Under the same conditions as in Theorem 1, if the number of splittings* K *is independent of the data, the multiple splitting TCV by voting is* W_n*-consistent. If additionally,* K *is upper bounded by a fixed constant (independent of* n*), the multiple splitting TCV by averaging is also* W_n*-consistent.*

5. Simulations and Real Data Example

The goal for the simulation and real data examples is to investigate if the TCV can indeed improve over (regular) CV when one's focus is not on the global performance. We also include local-dataset-based methods as candidates in some cases. The examples are chosen to highlight the differences among the competing methods. For multiple splittings, we apply the scheme of Monte-Carlo CV and aggregate the results by averaging as shown in Equation (4.3).

5.1. Simulation 1: a simple model versus a comprehensive model

Consider an example that involves the comparison between a simple model and a comprehensive model.

Let the data-generating process with predictors $X^{(0)}, \ldots, X^{(100)}$ and $\mathcal{I}$ be

$$Y = X^{(0)} + (1-\mathcal{I})(X^{(1)} + \cdots + X^{(100)}) + \varepsilon,$$

where $X^{(0)}, \ldots, X^{(100)}$ follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} 20 & 0 & 0 & \ldots & 0 \\ 0 & 0.1 & 0.1 & \ldots & 0.1 \\ 0 & 0.1 & 0.1 & \ldots & 0.1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0.1 & 0.1 & \ldots & 0.1 \end{pmatrix},$$

$\mathcal{I}$ follows $Bernoulli(0.1)$ and is independent of $X^{(0)}, \ldots, X^{(100)}$, and the random error ε follows $N(0, \sigma^2)$. We consider error variances $\sigma = 25$ or 3. The goal is to find out the best candidate method for the local region

$$\{(x^{(0)}, x^{(1)}, x^{(2)} \cdots x^{(100)}, \mathcal{I})^{\mathrm{T}} : \mathcal{I} = 1, (x^{(0)}, x^{(1)}, x^{(2)} \cdots x^{(100)})^{\mathrm{T}} \in \mathbb{R}^{101}\}. \tag{5.1}$$

Our candidate models are:

δ_1: $Y = \beta_0 X^{(0)} + \varepsilon$, fitted on the complete dataset,

δ_2: $Y = \beta_0 X^{(0)} + \cdots + \beta_{100} X^{(100)} + \beta_{101} \mathcal{I} \cdot X^{(1)} + \cdots \beta_{200} \mathcal{I} \cdot X^{(100)} + \varepsilon$,

fitted on the complete dataset,

δ_3: $Y = \beta_0 X^{(0)} + \varepsilon$, fitted on the local dataset, where $\mathcal{I} = 1$.

The candidate models δ_1 and δ_3 are only correct in the local region in (5.1) and δ_2 is the overall correct model. Intuitively, δ_1 can be better than δ_2 in the local region since δ_2 has too many parameters to be estimated. As for δ_1 versus δ_3, it depends on whether the gain from the increased sample size outweighs the loss from the variability introduced by $X^{(1)}, \ldots, X^{(100)}$. Thus, the two error variances σ_1^2 and σ_2^2 are likely to result in different outcomes.

First, we randomly generate a dataset with $n = 800$. Next, we apply both the CV and TCV with $W_n(\boldsymbol{x}) = \mathbb{1}(\mathcal{I} = 1)$. The training set and test set sizes $n_1 = n_2 = 400$, and the number of splittings $K = 100$. To measure the performance of the candidate models and selection results, we independently generate an evaluation set from the data-generating model with size 5000 and calculate **i)** weighted MSE with $0/1$ weight on $\mathcal{I}$, which measures the local performance, and **ii)** MSE without weight, which measures the overall performance, for the three candidate models and the models selected by the CV and TCV. The MSEs and associated standard errors based on 500 replications are shown in Table 1.

It can be seen from Table 1 that δ_2 is not the best model for the local region. Therefore, in this case, the complex global model, while being correct, does not perform as well as the simple model or that based on the local region due to the relatively small sample size given the complexity of the whole model. The table also shows that for $\sigma = 25$, δ_1 outperforms δ_3 for the local region. This is because

Table 1. The MSEs from δ_1, δ_2, δ_3, TCV, and CV for both the local and overall performances. The standard errors of the MSEs are shown in parentheses. The bold numbers stand for the better one out of the CV and TCV.

		δ_1	δ_2	δ_3	TCV	CV
$\sigma = 25$	Local	62.8	65.6	63.3	**63.1**	65.6
		(0.2)	(0.2)	(0.2)	(0.2)	(0.2)
	Overall	1528.3	631.7	1533.1	1518.0	**631.7**
		(1.4)	(0.6)	(1.5)	(4.9)	(0.6)
$\sigma = 3$	Local	0.997	0.945	0.911	**0.914**	0.945
		(0.007)	(0.004)	(0.003)	(0.003)	(0.003)
	Overall	910.8	9.1	909.9	886.6	**9.1**
		(0.861)	(0.009)	(0.864)	(6.474)	(0.009)

with the use of more observations, δ_1 gains much in variance reduction and has a better performance than δ_3 for the local region. For $\sigma = 3$, however, the gain from the variance reduction is blown away by the bias introduced by the outside data (recall that $X^{(1)}, \ldots, X^{(100)}$ are not included in δ_1), resulting in the worse local performance. In both cases, δ_2 always has the best overall performance. For the selection results, the TCV has better performance than CV for the local region as desired, but worse for the overall performance.

To demonstrate the effect of the number of data splittings in TCV, we calculate the weighted cross validation MSEs (as defined in (4.3)) with $K = 1$ and $K = 100$ respectively. As shown in the left panel of Table 2, for the weighted cross validation MSEs of the candidate models, the respective standard deviations based on 500 replications decrease with the increased number of splittings. The same is true for the weighted squared L_2 loss (simulated based on 5000 independently generated predictor values) of the regression estimators from the models selected by TCV, as seen in the right panel of Table 2.

Table 2. Effect of the number of data splittings ($K = 1$ versus 100) on cross validation errors and regression estimation losses.

(a) Standard deviations of the weighted cross validation MSEs of the candidate models.

Candidate Model	σ	$K = 1$	$K = 100$
δ_1	25	15.07	10.61
	3	0.37	0.21
δ_2	25	8586.74	5724.64
	3	123.65	82.43
δ_3	25	7700.46	5368.10
	3	110.89	77.30

(b) Standard deviations of the weighted squared L_2 losses of the models selected by TCV.

	σ	$K = 1$	$K = 100$
Local	25	5.25	4.93
	3	0.089	0.073
Overall	25	323.25	86.60
	3	372.85	154.93

5.2. Simulation 2: a continuous weighting for TCV

Sometimes, it may be preferable to assign weights between 0 and 1, e.g., $W(\boldsymbol{x}) = a$ if $\boldsymbol{x}$ is in a certain region, where $0 < a < 1$ and $W(\boldsymbol{x}) = 1 - a$ otherwise. In this way, we aim to find out the candidate methods with excellent performance for the region with the larger weight value and acceptable performance on the remaining part.

We consider the data-generating process

$$Y = \begin{cases} 250(X + 0.1)^2 + \varepsilon, & \text{if } 0 < X \leq 0.1, \\ 100X + \varepsilon, & \text{if } 0.1 < X < 1, \end{cases}$$

where the predictor X is generated i.i.d. from $U(0, 1)$, and the random error ε's are i.i.d. from $N(0, 1)$. It can be seen that the underlying true regression function is quadratic on the left and linear on the right. Figure 2 is an example of a random sample of this model. Our candidate methods are the Nadaraya-

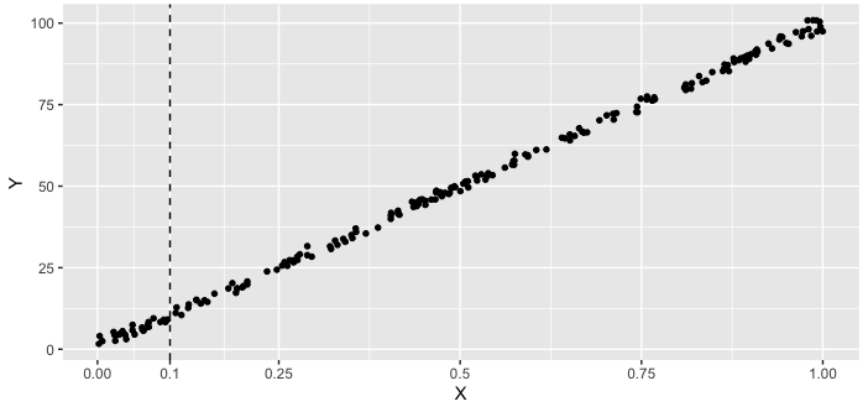

Figure 2: An example of simulated data points from the data-generating model, where the regression function is quadratic when $X \leq 0.1$ and linear otherwise.

Watson estimator [21] with Gaussian kernel and linear regression. The Nadaraya-Watson estimator is estimated by "npreg" from the R package "np" with bandwidth selected by least-squares cross-validation. Intuitively, the Nadaraya-Watson estimator may have better performance on the quadratic part and the linear regression may have better performance on the linear part. We compare four kinds of TCVs with weights:

$$W_{n,0.5}(X) = \begin{cases} 0.5, & X < 0.1, \\ 0.5, & X \geq 0.1, \end{cases} \quad W_{n,0.8}(X) = \begin{cases} 0.8, & X < 0.1, \\ 0.2, & X \geq 0.1, \end{cases}$$

$$W_{n,0.9}(X) = \begin{cases} 0.9, & X < 0.1, \\ 0.1, & X \geq 0.1, \end{cases} \quad W_{n,1}(X) = \begin{cases} 1, & X < 0.1, \\ 0, & X \geq 0.1. \end{cases}$$

We randomly generate a dataset with the size 200 and apply the TCVs with the different weight functions to the candidate methods with the training and the test sizes $n_1 = n_2 = 100$ and the number

of splittings $K = 100$. The evaluation procedure is the same as in Section 5.1. The weighted MSE for local performance, performance outside the local region, and the overall performance are considered. The MSEs and their standard errors based on 500 replications for the candidate methods and TCVs are shown in Table 3. The notations "NW" and "Linear" stand for the Nadaraya-Watson estimator and linear regression, respectively. The notation "TCV$_\alpha$" with $\alpha = 0.5, 0.8, 0.9, 0.1$ stand for the TCVs with the corresponding weight $W_{n,\alpha}$, respectively.

Table 3. MSEs from the Nadaraya-Watson estimator, linear regression, and TCVs with different weights. We consider their local region performances, outside local region performances, and overall performances. The standard errors of the MSEs are shown in parentheses.

	NW	Linear	TCV$_{0.5}$	TCV$_{0.8}$	TCV$_{0.9}$	TCV$_1$
Local	0.126	0.186	0.186	0.162	0.152	0.150
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
Outside	1.120	0.927	0.927	1.013	1.043	1.051
	(0.003)	(0.001)	(0.001)	(0.005)	(0.005)	(0.005)
Overall	1.245	1.112	1.112	1.176	1.195	1.201
	(0.003)	(0.001)	(0.001)	(0.004)	(0.004)	(0.004)

We see that the Nadaraya-Watson estimator performs better in the local region ($X < 0.1$), and linear regression performs better outside the local region ($X > 0.1$). Additionally, linear regression has better overall performance. With $W_{n,0.5}$, the TCV is equivalent to CV, and it prefers linear regression. As the weight in the local region increases, the performance of the TCV gets closer to that of the Nadaraya-Watson estimator, with improved local region performance. Nevertheless, as a tradeoff, it has worse global and outside the local region performances. The results show that the TCV can address the problems where we need a balance between local and overall performance.

5.3. Simulation 3: a high-dimensional case

In this example, we apply TCV in a high-dimensional setting. We consider the data-generating process

$$Y = 2\exp\big(-5{X^{(1)}}^2\big) + 2X^{(1)} + X^{(2)} + 0.5X^{(3)} + 0.1X^{(4)} + \varepsilon.$$

The 1000 predictors $X^{(1)}, \ldots, X^{(1000)}$ follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $V \in \mathbb{R}^{1000}$ with $V_{i,j} = 0.1^{|i-j|}$. The random error ε's are i.i.d. from $N(0,1)$. We want to find out the best candidate method for the local region

$$\{(x^{(1)}, x^{(2)} \cdots x^{(1000)})^{\mathrm{T}} : (x^{(1)}, x^{(2)})^{\mathrm{T}} \in (-0.5, 0.5)^2, (x^{(3)} \cdots x^{(1000)})^{\mathrm{T}} \in \mathbb{R}^{998}\}.$$

Four candidate models/procedures are considered. They are random forest and lasso regression fitted on the complete dataset, and their local versions fitted on the local region data.

The nonlinear term $\exp\big(-5{X^{(1)}}^2\big)$ has a large influence on the regression function when $X^{(1)}$ is near 0. According to this, the random forest may have the best performance in the local region centered at 0. However, lasso regression also has a chance to overperform the random forest if the influence

of the nonlinearity is relatively small compared with the advantage obtained from sparsity. The local candidate models are more specific to the local region compared with their global counterparts, but they have fewer data.

We randomly generate datasets with size 200, evaluate the candidate methods and CV selection results, and independently repeat this process 100 times. The other simulation settings are the same as previous subsections. The random forest is fitted by 500 trees, and it selects 32 variables each time. The tuning parameter of the lasso regression is selected by the 10-fold CV.

The MSEs and their standard errors are shown in Table 4. The notations "lasso" and "RF" stand for lasso regression and random forest based on the complete dataset, respectively, and "lasso_local" and "RF_local" stand for the corresponding local versions. It can be seen that the random forest based on the complete dataset has the best local performance, and lasso regression based on the complete dataset has the best overall performance. The TCV selects the complete dataset-based random forest for most of the times, and CV often selects the complete dataset-based lasso regression. Consequently, the TCV has better local performance and the CV has better overall performance, as expected.

Table 4. MSEs from lasso regression and random forest built on all or local data, TCV, and CV. We consider their local region performances and overall performances. The standard errors of the MSEs are shown in parentheses. The bold numbers stand for the better one out of the CV and TCV. A t-test at the significance level of 0.05 shows that there is no significant difference between the local performances of 'lasso'(with MSE 2.02) and 'lasso_local'(with MSE 2.00).

	lasso	RF	lasso_local	RF_local	TCV	CV
Local	2.02 (0.02)	1.47 (0.01)	2.00 (0.02)	1.90 (0.01)	**1.62** (0.03)	2.02 (0.02)
Overall	1.74 (0.01)	2.78 (0.02)	7.48 (0.13)	7.86 (0.05)	3.85 (0.23)	**1.75** (0.02)

5.4. Boston housing data

We demonstrate the application of the TCV using the Boston housing data from [16] with 506 observations. They fitted a model for the median value of owner-occupied homes based on the model

$$\begin{aligned}\log(MV) = {} & a_1 + a_2 RM^2 + a_3 AGE + a_4 \log(DIS) + a_5 \log(RAD) + \\ & a_6 TAX + a_7 PTRATIO + a_8 B + a_9 \log(LSTAT) + a_{10} CRIM + \\ & a_{11} ZN + a_{12} INDUS + a_{13} CHAS + a_{14} NOX^2 + \varepsilon.\end{aligned} \tag{5.2}$$

Instead of fitting a model of the overall house price like (5.2), we consider the house prices of relatively new buildings. Since the age of a house is one of the most important factors that a home buyer typically considers, it will definitely make a difference in terms of the price. According to this fact, other candidate models may perform better than the global model (5.2).

The predictor AGE is the proportion of owner-occupied homes built prior to 1940, which measures the overall age of the houses in a census tract. Here, we are particularly interested in the medium home

value in relatively new areas with less than 50% of the houses built before 1940 ($AGE < 50$). There are 147 observations with $AGE < 50$ in the data. We consider the following three natural candidate models:

1. δ_1 uses the model (5.2) with all the available data.
2. δ_2 performs an additive regression, using the model

$$\begin{aligned}\log(MV) = a_1 + a_2 f(RM^2) + a_3 f(AGE) + a_4 f(\log(DIS)) + \\ a_5 f(\log(RAD)) + a_6 f(TAX) + a_7 f(PTRATIO) + \\ a_8 f(B) + \ a_9 f(\log(LSTAT)) + a_{10} f(CRIM) + \\ a_{11} f(ZN) + a_{12} f(INDUS) + a_{13} CHAS + \\ a_{14} f(NOX^2) + \varepsilon,\end{aligned} \tag{5.3}$$

   with all available data, where each f represents a smoothing spline with three degrees of freedom
3. δ_3 is from model (5.2) but only based on the local data with $AGE < 50$.

Note that the local version of δ_2 is not considered as a candidate since we do not have enough local data to support the nonparametric estimation.

We first randomly set aside 20% of the observations as the evaluation set and apply half-half splitting 100 times on the rest of the observations for the CV and TCV. To avoid not having enough local data in the randomly splitted data, we apply a stratified sampling. The weight for the TCV is $W_n(X) = \mathbb{1}(AGE < 50)$. We calculate the weighted MSE based on the local region and the overall MSE for the three candidate models together with the TCV and CV on the evaluation set. The above procedure is independently repeated 500 times. The MSEs, together with their standard errors, for the candidate models, the TCV, and CV are shown in Table 5.

Table 5. MSEs of δ_1, δ_2, δ_3, TCV, and CV. We consider both their local region performances and overall performances. The standard errors of the MSEs are shown in parentheses. The bold numbers stand for the better one out of the CV and TCV.

	δ_1	δ_2	δ_3	TCV	CV
Local	0.0049 (6.7×10^{-5})	0.0037 (6.4×10^{-5})	0.0018 (2.4×10^{-5})	**0.0018** (2.5×10^{-5})	0.0037 (6.4×10^{-5})
Overall	0.0360 (3.3×10^{-4})	0.0303 (3.1×10^{-4})	0.1710 (0.007)	0.1707 (0.007)	**0.0304** (3.1×10^{-4})

It can be seen that δ_3 has the best performance in the local region where $AGE < 50$. For the overall performance, δ_2 is the best. The TCV always selected δ_3 and the CV always selected δ_2. Clearly, the TCV has excellent performance since it selected the candidate with the best performance for the region of interest.

6. Concluding Remarks

We have proposed the TCV for selecting the best candidate regression procedure defined via weighted L_2 losses. An application of the TCV is to find a candidate method with the best performance for a local region. With a proper data splitting, our method can consistently identify the best-performing candidate that possibly varies with the sample size. Simulation and real data examples have illustrated that the TCV can outperform the regular CV when the candidate methods do not rank uniformly over the local region and outside that region.

With the availability of large numbers of observations, often with high input dimensions, highly adaptive non-traditional methods can better approximate complicated regression functions. In this background, the traditional framework of a fixed best model or procedure waiting to be identified may be overly simplistic in real applications. In this paper, we utilize a triangular array setup to facilitate the needed flexibility for selecting the dynamically best candidate. This more dynamic framework may be generally helpful or even necessary in establishing adaptive learning theories for high-dimensional and big data when data splittings are involved.

It is interesting to observe that when a weighting function is used to define the loss of interest, it may have a major impact on how we need to split the data. Indeed, when we care most about a small region, for instance, the task of finding out the best candidate procedure among close competitors in that region becomes harder, compared with that of identifying the globally best. Then, we need to shift the data splitting ratio more towards the evaluation part. Our main theorem gives sufficient conditions to enable consistent selection in terms of the natures of the weighting function and the convergence rates of the candidate procedures. We have also observed that the TCV may exhibit poor performance or high variability in terms of the global performance. Therefore, a proper CV method should be chosen according to the interest of the application.

In this work, the number of candidate procedures to be compared is essentially not allowed to grow as the sample size increases. One interesting future direction is to handle the situation where the list of candidate procedures expands when more observations become available, which provides more flexibility in regression modeling.

7. Proof of the Main Theorem

Since the number of candidate methods in $\mathcal{M}$ is upper bounded, it suffices to show that the TCV is W_n-consistent when comparing the $(W_n, l_n, c_{(n_1,n)})$-best candidate method δ_{g_n} and every other candidate method from $\mathcal{M}$. We take an arbitrary candidate method that is not δ_{g_n} from $\mathcal{M}$ and denote it by δ_{b_n}.

In this proof, we will show that when $l_n \le n_1 < n$, the probability of selecting the $(W_n, l_n, c_{(n_1,n)})$-better candidate

$$P\left(TCV_{W_n}(\widehat{f}_{n_1}^{(b_n)}) > TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)})\right) \to 1, \tag{7.1}$$

as $n \to \infty$. Since $Y_{n,i} = f(\boldsymbol{X}_{n,i}) + \varepsilon_{n,i}$, we have

$$TCV_{W_n}(\widehat{f}_{n_1}^{(j)}) = \sum_{i=n_1+1}^{n} \varepsilon_{n,i}^2 \cdot W_n(\boldsymbol{X}_{n,i}) + \sum_{i=n_1+1}^{n} (f(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(j)}(\boldsymbol{X}_{n,i}))^2 \cdot$$

$$W_n(\boldsymbol{X}_{n,i}) + 2 \sum_{i=n_1+1}^{n} \varepsilon_{n,i}(f(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(j)}(\boldsymbol{X}_{n,i})) \cdot W_n(\boldsymbol{X}_{n,i}),$$

where $j = 1$ or 2. We have

$$\begin{aligned} & TCV_{W_n}(\widehat{f}_{n_1}^{(b_n)}) - TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)}) \leq 0 \\ \iff & 2\sum_{i=n_1+1}^{n} \varepsilon_{n,i}(\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i})) \cdot W_n(\boldsymbol{X}_{n,i}) \geq \mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)}, \end{aligned} \tag{7.2}$$

where

$$\mathcal{K}^{(g_n)} = \sum_{i=n_1+1}^{n} (f(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i}))^2 \cdot W_n(\boldsymbol{X}_{n,i}), \tag{7.3}$$

$$\mathcal{K}^{(b_n)} = \sum_{i=n_1+1}^{n} (f(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}))^2 \cdot W_n(\boldsymbol{X}_{n,i}). \tag{7.4}$$

Denote the event $\{\mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)} > 0\}$ by S_n and the training data by Z_{n_1}. Conditional on Z_{n_1}, the predictor data $\boldsymbol{X}_n^{(2)}$ from the test set, and given S_n holds, by the result from (7.2) and Chebyshev's inequality, we have

$$P\left(TCV_{W_n}(\widehat{f}_{n_1}^{(b_n)}) \leq TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)}) | Z_{n_1}, \boldsymbol{X}_n^{(2)}, S_n\right) \tag{7.5}$$

$$= P\left(2\sum_{i=n_1+1}^{n} \epsilon_i \left(\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i})\right) \cdot W_n(\boldsymbol{X}_{n,i}) \geq \right. \tag{7.6}$$

$$\left. \mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)} | Z_{n_1}, \boldsymbol{X}_n^{(2)}, S_n\right)$$

$$\leq \frac{Var(2\sum_{i=n_1+1}^{n} \epsilon_i (\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i})) \cdot W_n(\boldsymbol{X}_{n,i}) | Z_{n_1}, \boldsymbol{X}_n^{(2)}, S_n)}{\left(\mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)}\right)^2}$$

$$\leq \frac{4\sum_{n_1+1}^{n} (\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i}))^2 \cdot (W_n(\boldsymbol{X}_{n,i}))^2 \cdot \overline{\sigma}^2}{\left(\mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)}\right)^2} \leq Q_n,$$

almost surely, where $Q_n = \frac{4\overline{\sigma}^2 \sum_{n_1+1}^{n} (\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i}))^2 \cdot W_n(\boldsymbol{X}_{n,i}) \cdot S_{W_n}}{\left(\mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)}\right)^2}$ (recall that S_{W_n} denotes ess-sup$_{\boldsymbol{x}}(W_n(\boldsymbol{x}))$). Therefore, (7.5) is upper bounded by $\min(1, Q_n)$, Thus, for the unconditional probability we have

$$\begin{aligned} & P\left(TCV_{W_n}(\widehat{f}_{n_1}^{(b_n)}) < TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)})\right) \\ & = E\left[P\left(\{TCV_{W_n}(\widehat{f}_{n_1}^{(b_n)}) < TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)})\} \cap S_n | Z_{n_1}, \boldsymbol{X}_n^{(2)}\right)\right] + \\ & \quad E\left[P\left(\{TCV_{W_n}(\widehat{f}_{n_1}^{(b_n)}) < TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)})\} \cap S_n^c | Z_{n_1}, \boldsymbol{X}_n^{(2)}\right)\right] \\ & \leq E\left[P\left(\{TCV_{W_n}(\widehat{f}_{n_1}^{(b_n)}) < TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)})\} | Z_{n_1}, \boldsymbol{X}_n^{(2)}, S_n\right)\right] + E\left[P\left(S_n^c | Z_{n_1}, \boldsymbol{X}_n^{(2)}\right)\right] \\ & \leq E\min(1, Q_n) + P(S_n^c). \end{aligned}$$

For the bound of $P(S_n^c)$, we first assume that for $l_n \leq n_1 < n$, there exists an upper bounded positive sequence α_{n_1}, such that

$$P\left(\frac{\mathcal{K}^{(b_n)}}{\mathcal{K}^{(g_n)}} \geq 1 + \alpha_{n_1}\right) \to 1, \tag{7.7}$$

as $n \to \infty$. The above inequality implies that $P(S_n) \to 1$ as $n \to \infty$. For the bound of $E\min(1, Q_n)$, we have

$$\begin{aligned}
&P\left(\frac{\mathcal{K}^{(g_n)}}{\mathcal{K}^{(b_n)}} \leq \frac{1}{1+\alpha_{n_1}}\right) \\
&= P\left(\frac{\mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)}}{\mathcal{K}^{(b_n)}} \geq 1 - \frac{1}{1+\alpha_{n_1}}\right) \\
&\leq P\left(\left(\frac{\mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)}}{(1-1/(1+\alpha_{n_1}))\mathcal{K}^{(b_n)}}\right)^2 \geq 1\right) \\
&= P\left(\left(\frac{\mathcal{K}^{(b_n)} - \mathcal{K}^{(g_n)}}{(1-1/(1+\alpha_{n_1}))\mathcal{K}^{(b_n)}}\right)^2 \cdot Q_n \geq Q_n\right) \\
&= P\left(Q_n \leq \frac{4\overline{\sigma}^2 \sum_{n_1+1}^{n} \left(\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i})\right)^2 W_n(\boldsymbol{X}_{n,i}) \cdot S_{W_n}}{\left((1-1/(1+\alpha_{n_1}))\mathcal{K}^{(b_n)}\right)^2}\right).
\end{aligned}$$

Combining the above result and inequality (7.7), we have

$$P\left(Q_n \leq \frac{4\overline{\sigma}^2 \sum_{n_1+1}^{n} \left(\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i})\right)^2 W_n(\boldsymbol{X}_{n,i}) \cdot S_{W_n}}{\left((1-1/(1+\alpha_{n_1}))\mathcal{K}^{(b_n)}\right)^2}\right) \to 1, \tag{7.8}$$

as $n \to \infty$. By the fact that $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\sum_{n_1+1}^{n} \left(\widehat{f}_{n_1}^{(b_n)}(\boldsymbol{X}_{n,i}) - \widehat{f}_{n_1}^{(g_n)}(\boldsymbol{X}_{n,i})\right)^2 \cdot W_n(\boldsymbol{X}_{n,i}) \leq 2\big(\mathcal{K}^{(b_n)} + \mathcal{K}^{(g_n)}\big). \tag{7.9}$$

Given (7.7), by combining (7.8) and (7.9), we obtain

$$P\left(Q_n \leq \frac{8\overline{\sigma}^2(\mathcal{K}^{(b_n)} + \mathcal{K}^{(g_n)}) \cdot S_{W_n}}{\left((1-1/(1+\alpha_{n_1}))\mathcal{K}^{(b_n)}\right)^2}\right) \to 1, \tag{7.10}$$

as $n \to \infty$. When (7.7) holds, we also have

$$P\left(\frac{\mathcal{K}^{(b_n)} + \mathcal{K}^{(g_n)}}{\mathcal{K}^{(b_n)}} \leq 1 + \frac{1}{1+\alpha_{n_1}}\right)$$

$$= P\left(\frac{8\overline{\sigma}^2(\mathcal{K}^{(b_n)}+\mathcal{K}^{(g_n)})\cdot S_{W_n}}{\left((1-1/(1+\alpha_{n_1}))\mathcal{K}^{(b_n)}\right)^2} \le \frac{8\overline{\sigma}^2(1+1/(1+\alpha_{n_1}))\cdot S_{W_n}}{(1-1/(1+\alpha_{n_1}))^2\mathcal{K}^{(b_n)}}\right) \to 1, \tag{7.11}$$

as $n \to \infty$. It follows from (7.10) and (7.11) that

$$P\left(Q_n \le \frac{8\overline{\sigma}^2(1+1/(1+\alpha_{n_1}))\cdot S_{W_n}}{(1-1/(1+\alpha_{n_1}))^2\mathcal{K}^{(b_n)}}\right) \to 1, \tag{7.12}$$

as $n \to \infty$. Therefore, by the fact that $Q_n \ge 0$ and the dominated convergence theorem, to control the upper bound of $E\min(1, Q_n)$, it suffices to show

$$\alpha_{n_1}^2 \cdot \mathcal{K}^{(b_n)}/S_{W_n} \xrightarrow{p} \infty. \tag{7.13}$$

According to the above results, to prove Thereom 1, it suffices to show (7.7) and (7.13). Next, we will establish one inequality between $\mathcal{K}^{(g_n)}$ and $\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2$ and another inequality between $\mathcal{K}^{(b_n)}$ and $\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2$. We first derive the inequality between $\mathcal{K}^{(g_n)}$ and $\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2$. We define $\mathcal{D}_{n,i,j} \triangleq (f(\boldsymbol{X}_{n,i})-\widehat{f}_{n_1}^{(j)}(\boldsymbol{X}_{n,i}))^2\cdot W_n(\boldsymbol{X}_{n,i}) - \|f-\widehat{f}_{n_1}^{(j)}\|_{2,W_n}^2$. Let $E_{Z_{n_1}}(\cdot)$ and $Var_{Z_{n_1}}(\cdot)$ denote the expectation and variance conditional on Z_{n_1}, respectively. We have

$$\begin{aligned} Var_{Z_{n_1}}(\mathcal{D}_{n,n_1+1,j}) &\le E_{Z_{n_1}}\left((f(\boldsymbol{X}_{n,n_1+1})-\widehat{f}_{n_1,j}(\boldsymbol{X}_{n,n_1+1}))^4\cdot\left(W_n(\boldsymbol{X}_{n,n_1+1})\right)^2\right) \\ &\le \|f-\widehat{f}_{n_1}^{(j)}\|_{4,W_n}^4\cdot S_{W_n}. \end{aligned} \tag{7.14}$$

The above inequality holds for both the $(W_n, l_n, c_{(n_1,n)})$-better candidate δ_{b_n} and worse candidate δ_{g_n}. Therefore, by (7.14) and Chebyshev's inequality, we obtain that for each $a > 0$,

$$\begin{aligned} P_{Z_{n_1}}\left(\mathcal{K}^{(g_n)} - n_2\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2 \ge a\right) &\le \frac{n_2 Var_{Z_{n_1}}(\mathcal{D}_{n,n_1+1,g_n})}{a^2} \\ &\le \frac{n_2\|f-\widehat{f}_{n_1}^{(g_n)}\|_{4,W_n}^4\cdot S_{W_n}}{a^2}, \end{aligned}$$

where $P_{Z_{n_1}}$ is the probability conditional on the training data Z_{n_1}. We take $a = \theta_{1,n_1}n_2\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2$, where $\theta_{1,n_1} > 0$ and will be determined latter. According to the above inequality, we have

$$P_{Z_{n_1}}\left(\mathcal{K}^{(g_n)} \ge (1+\theta_{1,n_1})n_2\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2\right) \le \frac{\|f-\widehat{f}_{n_1}^{(g_n)}\|_{4,W_n}^4\cdot S_{W_n}}{\theta_{1,n_1}^2 n_2\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^4}. \tag{7.15}$$

We denote event $\left\{\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2/\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2 \ge 1+c_{(n_1,n)}\right\}$ by D_n. Then, according to Definition 2, we have

$$P(D_n) \to 1, \tag{7.16}$$

as $n \to \infty$. Without losing generality, we require that the sequence $\{c_{(n_1,n)}\}$ is upper bounded (otherwise, we can always replace the original $\{c_{(n_1,n)}\}$ by a bounded sequence). To obtain the inequality between $\mathcal{K}^{(g_n)}$ and $\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2$, we take $\theta_{1,n_1} = \|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2/\left((1+c_{(n_1,n)}/2\right)\|f-$

$\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2) - 1$. Conditional on D_n, we have

$$\begin{aligned}\theta_{1,n_1} &= \frac{\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2 - \|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2(1+c_{(n_1,n)}/2)}{\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2(1+c_{(n_1,n)}/2)} \\ &\geq \frac{\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2 - \|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2\frac{1+c_{(n_1,n)}/2}{1+c_{(n_1,n)}}}{\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2(1+c_{(n_1,n)}/2)} \\ &= \frac{c_{(n_1,n)}\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2}{2(1+c_{(n_1,n)})(1+c_{(n_1,n)}/2)\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2}. \end{aligned} \tag{7.17}$$

According to (7.17) and (7.15), we obtain

$$\begin{aligned} & P\left(\mathcal{K}^{(g_n)} \geq \frac{n_2}{1+c_{(n_1,n)}/2}\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2\right) \\ \leq & P\left(\left\{\mathcal{K}^{(g_n)} \geq \frac{n_2}{1+c_{(n_1,n)}/2}\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2\right\}\cap D_n\right) + P(D_n^c), \\ \leq & P\left(\mathcal{K}^{(g_n)} \geq \frac{n_2}{1+c_{(n_1,n)}/2}\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2 \mid D_n\right) + P(D_n^c), \\ = & E\bigg(P_{Z_{n_1}}\left(\mathcal{K}^{(g_n)} \geq (1+\theta_{1,n_1})n_2\|f-\widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2\right) \mid D_n\bigg) + P(D_n^c) \\ \leq & E(Q_n^{(1)}) + P(D_n^c), \end{aligned} \tag{7.18}$$

where $Q_n^{(1)} \triangleq \min\left\{1, \frac{4(1+c_{(n_1,n)})^2(1+c_{(n_1,n)}/2)^2}{c_{(n_1,n)}^2}\cdot\frac{\|f-\widehat{f}_{n_1}^{(g_n)}\|_{4,W_n}^4\cdot S_{W_n}}{n_2\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^4}\right\}$.

Next, we derive the other inequality that compares $\mathcal{K}^{(b_n)}$ and $\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2$. Let $\{\theta_{2,n_1}\}$ be a positive sequence whose complete definition will be given latter. We require it to be bounded above by 1. By Chebyshev's inequality and (7.14), we have

$$\begin{aligned} & P(\mathcal{K}^{(b_n)} < (1-\theta_{2,n_1})n_2\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2) \\ & = P(-\mathcal{K}^{(b_n)} + n_2\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2 > \theta_{2,n_1}n_2\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2) \\ & = E\bigg(P_{Z_{n_1}}\left(-\mathcal{K}^{(b_n)} + n_2\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2 > \theta_{2,n_1}n_2\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2\right)\bigg) \\ & \leq E(Q_n^{(2)}), \end{aligned} \tag{7.19}$$

where $Q_n^{(2)} \triangleq \min\left\{1, \frac{\|f-\widehat{f}_{n_1}^{(b_n)}\|_{4,W_n}^4\cdot S_{W_n}}{\theta_{2,n_1}^2 n_2\|f-\widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^4}\right\}$.

To show (7.7), we take $\theta_{2,n_1} = 1 - \frac{1+c_{(n_1,n)}/4}{1+c_{(n_1,n)}/2}$ and $\alpha_{n_1} = c_{(n_1,n)}/4$. By combining (7.18) and (7.19), we have $P\left(\mathcal{K}^{(b_n)} \geq (1+\alpha_{n_1}) \cdot \mathcal{K}^{(g_n)}\right) \geq 1 - P(D_n^c) - E(Q_n^{(1)}) - E(Q_n^{(3)})$, where $Q_n^{(3)}$ is $Q_n^{(2)}$ with $\theta_{2,n_1} = 1 - \frac{1+c_{(n_1,n)}/4}{1+c_{(n_1,n)}/2}$. Recall that the sequence $\{c_{(n_1,n)}\}$ is upper bounded. Therefore, by the dominated convergence theorem, it remains to show that

$$\frac{1}{c_{(n_1,n)}^2} \cdot \frac{\|f - \widehat{f}_{n_1}^{(g_n)}\|_{4,W_n}^4 \cdot S_{W_n}}{n_2 \|f - \widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^4} < \frac{1}{c_{(n_1,n)}^2} \cdot \frac{\|f - \widehat{f}_{n_1}^{(g_n)}\|_{4,W_n}^4 \cdot S_{W_n}}{n_2 \|f - \widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^4} \to 0,$$

$$\frac{1}{c_{(n_1,n)}^2} \cdot \frac{\|f - \widehat{f}_{n_1}^{(b_n)}\|_{4,W_n}^4 \cdot S_{W_n}}{n_2 \|f - \widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^4} \to 0, \tag{7.20}$$

as $n \to \infty$ and $l_n \leq n_1 < n$. According to Condition 2, to obtain the above results, it suffices to show that $M_n^4 \cdot S_{W_n}/(c_{(n_1,n)}^2 \cdot n_2) \to 0$, as $n \to \infty$, and this is required by Theorem 1.

For (7.13), recall that we have $\alpha_{n_1} = c_{(n_1,n)}/4$. Therefore, it suffices to show that

$$c_{(n_1,n)}^2 \cdot \mathcal{K}^{(b_n)} / S_{W_n} \xrightarrow{p} \infty, \tag{7.21}$$

as $n \to \infty$. By (7.19), (7.20), and the fact that $1 - \theta_{2,n_1} = \frac{1+c_{(n_1,n)}/4}{1+c_{(n_1,n)}/2}$, which is bounded between $1/2$ and 1, we have that the rate of $K^{(b_n)}$ is lower bounded by $q_{(n_1,n)}$. Therefore, we obtain (7.21) by the requirement **(ii)** from Theorem 1. Thus, we obtain (7.7) and (7.13) and complete the proof.

8. Proof of Proposition 1

By the fact that $(\sqrt{1+n_1^{-1/3}} - 1)/n_1^{-1/3} \to 1/2$ as $n_1 \to \infty$, we have that when n is sufficiently large and $n_1 \in [l_n, n)$,

$$\begin{aligned} &\|f - \widehat{f}_{n_1}^{(b_n)}\|_{2,W_n}^2 \geq (1 + n_1^{-1/3}) \|f - \widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}^2 \\ \Rightarrow &\|f - \widehat{f}_{n_1}^{(b_n)}\|_{2,W_n} \geq \left(1 + n_1^{-1/3}/3\right) \|f - \widehat{f}_{n_1}^{(g_n)}\|_{2,W_n}. \end{aligned}$$

Therefore, it suffices to show that as $n \to \infty$,

$$\text{when } n \in \mathbb{S}_1,\ P\left(n_1^{-1/3} \cdot \frac{\|f - \hat{f}_{n_1}^{(1)}\|_2^2}{\|f - \hat{f}_{n_1}^{(2)}\|_2^2 - \|f - \hat{f}_{n_1}^{(1)}\|_2^2} < 1\right) \to 1, \tag{8.1}$$

$$\text{when } n \in \mathbb{S}_2,\ P\left(n_1^{-1/3} \cdot \frac{\|f - \hat{f}_{n_1}^{(2)}\|_2^2}{\|f - \hat{f}_{n_1}^{(1)}\|_2^2 - \|f - \hat{f}_{n_1}^{(2)}\|_2^2} < 1\right) \to 1. \tag{8.2}$$

Note that for $m = 1, 2$,

$$\|f - \hat{f}_{n_1}^{(m)}\|_2^2 = \sum_{j=1}^{p_{n_1}^{(m)}} (\beta_j - \hat{\beta}_j^{(n_1)})^2 + \sum_{j=p_{n_1}^{(m)}+1}^{\infty} \beta_j^2, \tag{8.3}$$

$$\|f-\hat{f}_{n_1}^{(1)}\|_2^2-\|f-\hat{f}_{n_1}^{(2)}\|_2^2=-\left(\beta_{p_{n_1}^{(2)}}-\hat{\beta}_{p_{n_1}^{(2)}}^{(n_1)}\right)^2+\beta_{p_{n_1}^{(2)}}^2.$$

For $n\in\mathbb{S}_1$, since $p_{n_1}^{(2)}=n_1^{1/4}\in[n^{1/12},n^{1/4}]$, we have $\beta_{p_{n_1}^{(2)}}^2=0$. Thus,

$$\begin{aligned}&\|f-\hat{f}_{n_1}^{(2)}\|_2^2-\|f-\hat{f}_{n_1}^{(1)}\|_2^2\\=&\left(\beta_{p_{n_1}^{(2)}}-\hat{\beta}_{p_{n_1}^{(2)}}^{(n_1)}\right)^2\\=&\left(\beta_{p_{n_1}^{(2)}}-\frac{1}{n_1}\sum_{i=1}^{n_1}f(x_i)\phi_{p_{n_1}^{(2)}}(x_i)-\frac{1}{n_1}\sum_{i=1}^{n_1}\epsilon_i\phi_{p_{n_1}^{(2)}}(x_i)\right)^2.\end{aligned}$$

Since β_j, $f(x_i)$, and $\phi_j(x_i)$ are all uniformly upper bounded by fixed constants for all i and j, we have that $E(\beta_j-f(x)\phi_j(x)-\epsilon\phi_j(x))^2$ and $E(\beta_j-f(x)\phi_j(x)-\epsilon\phi_j(x))^3$ are uniformly upper bounded. Also, by

$$\begin{aligned}E(\beta_j-f(x)\phi_j(x)-\epsilon\phi_j(x))^2=&E(\beta_j-f(x)\phi_j(x))^2+E\epsilon^2\phi_j(x)^2-\\&2E\epsilon\phi_j(x)(\beta_j-f(x)\phi_j(x))\\=&E(\beta_j-f(x)\phi_j(x))^2+E\epsilon^2\phi_j(x)^2-0\\\geq&E\epsilon^2\phi_j(x)^2\\=&1,\end{aligned}$$

we have that $E(\beta_j-f(x)\phi_j(x)-\epsilon\phi_j(x))^2$ is uniformly lower bounded away from 0. Therefore, by the Lyapunov CLT, we have

$$\frac{1}{\mathcal{V}_{p_{n_1}^{(2)}}}\cdot n_1\cdot\left(\beta_{p_{n_1}^{(2)}}-\hat{\beta}_{p_{n_1}^{(2)}}^{(n_1)}\right)\to_d N(0,1)\tag{8.4}$$

as $n\to\infty$, where $\mathcal{V}_{p_{n_1}^{(2)}}=\sqrt{\sum_{i=1}^{n_1}E(\beta_{p_{n_1}^{(2)}}-f(x_i)\phi_{p_{n_1}^{(2)}}(x_i)-\epsilon\phi_{p_{n_1}^{(2)}}(x_i))^2}$. Next, we denote the uniform upper bound of $E(\beta_j-f(x)\phi_j(x)-\epsilon\phi_j(x))^2$ by $\mathcal{C}$. Then,

$$\begin{aligned}E\big((\beta_j-\hat{\beta}_j^{(n_1)})^2\big)=&\frac{1}{n_1^2}E\left(\left(\sum_{i=1}^{n_1}(\beta_j-f(x_i)\phi_j(x_i)-\epsilon\phi_j(x_i))\right)^2\right)\\=&\frac{1}{n_1^2}E\left(\sum_{i=1}^{n_1}\big(\beta_j-f(x_i)\phi_j(x_i)-\epsilon\phi_j(x_i)\big)^2\right)\leq\mathcal{C}/n_1.\end{aligned}$$

By Markov's inequality, for each $t>0$,

$$P\left(\sum_{j=1}^{p_{n_1}^{(1)}}(\beta_j-\hat{\beta}_j^{(n_1)})^2>t\right)\leq\frac{E\left(\sum_{j=1}^{p_{n_1}^{(1)}}(\beta_j-\hat{\beta}_j^{(n_1)})^2\right)}{t}\leq\frac{p_{n_1}^{(1)}\mathcal{C}}{n_1t}.\tag{8.5}$$

Thus,

$$\sum_{j=1}^{p_{n_1}^{(1)}}(\beta_j - \hat{\beta}_j^{(n_1)})^2 = O_p\big(n_1^{-3/4}\big). \tag{8.6}$$

Also, by $p_{n_1}^{(2)} = p_{n_1}^{(1)} + 1 \in [n^{1/12}, n^{1/4}]$, we have that (note that when $n \in \mathbb{S}_1 \cup \mathbb{S}_2$, $n^{1/4}$ is an integer) $\sum_{j=p_{n_1}^{(1)}+1}^{\infty} \beta_j^2 = 0 + \sum_{j=n^{1/4}}^{\infty} \beta_j^2 \leq \int_{n^{1/4}-1}^{\infty} \frac{1}{j^4} dj = \frac{1}{3} \cdot \frac{1}{(n^{1/4}-1)^3}$. Therefore,

$$\sum_{j=p_{n_1}^{(1)}+1}^{\infty} \beta_j^2 = O_p(n^{-3/4}). \tag{8.7}$$

According to (8.6), (8.7), and the fact that $n_1 < n$, we have that

$$\|f - \hat{f}_{n_1}^{(1)}\|_2^2 = O_p(n_1^{-3/4}). \tag{8.8}$$

Therefore, by (8.4), (8.8), and the fact that $n_1^{-3/4} \cdot n_1 = n_1^{1/4}$, which goes to infinity at a slower rate than $n_1^{1/3}$, we obtain (8.1).

For $n \in \mathbb{S}_2$, we have $\|f - \hat{f}_{n_1}^{(1)}\|_2^2 - \|f - \hat{f}_{n_1}^{(2)}\|_2^2 = -\left(\beta_{p_{n_1}^{(2)}} - \hat{\beta}_{p_{n_1}^{(2)}}^{(n_1)}\right)^2 + \beta_{p_{n_1}^{(2)}}^2 = -\left(\beta_{p_{n_1}^{(2)}} - \hat{\beta}_{p_{n_1}^{(2)}}^{(n_1)}\right)^2 + n_1^{-1/2}$. According to the result from (8.4), $\big(\beta_{p_{n_1}^{(2)}} - \hat{\beta}_{p_{n_1}^{(2)}}^{(n_1)}\big)^2$ goes to 0 at the exact rate $1/n_1$, which is faster than that of $n_1^{-1/2}$. Therefore, $\|f - \hat{f}_{n_1}^{(1)}\|_2^2 - \|f - \hat{f}_{n_1}^{(2)}\|_2^2 \to 0$ as $n \to \infty$ with the rate lower bounded by $1/n_1$. With a similar derivation of (8.6), we have that

$$\sum_{j=1}^{p_{n_1}^{(2)}}(\beta_j - \hat{\beta}_j^{(n_1)})^2 = O_p\big(n_1^{-3/4}\big). \tag{8.9}$$

We also have $\sum_{j=p_{n_1}^{(2)}+1}^{\infty} \beta_j^2 = \sum_{j=n_1^{1/4}+1}^{\infty} \beta_j^2 \leq \int_{n_1^{1/4}}^{\infty} \frac{1}{j^4} dj = \frac{1}{3} \cdot n_1^{-3/4}$. Combining the above result with (8.9), we that $\|f - \hat{f}_{n_1}^{(2)}\|_2^2 = O_p(n_1^{-3/4})$. Therefore, according to the above results, we also obtain (8.2) and complete the proof.

9. Proof of Proposition 2

First, we show the existence of the n_1 such that $n_2/n_1^{17/12} \to \infty$ as $n \to \infty$. Since $n_2/n_1^{17/12} = (n - n_1)/n_1^{17/12}$, it suffices to have $n/n_1^{17/12} \to \infty$ as $n \to \infty$. According to $n^{12/17}/l_n = n^{19/51}$ and $n_1 \in [l_n, n)$, we have verified the existence of such an n_1.

Next, we verify the requirement **(ii)**. It can be seen from the derivation from Appendix 8 that both $\|f - \hat{f}_{n_1}^{(1)}\|_2^2$ and $\|f - \hat{f}_{n_1}^{(2)}\|_2^2$ converge at the exact rate $n_1^{-3/4}$. Thus, $q_{(n_1,n)}^2 = n_1^{-3/4}$. Since $c_{(n_1,n)}^2 = n_1^{-2/3}/9$, when $n_2 \cdot n_1^{-3/4} \cdot n_1^{-2/3} = n_2/n_1^{17/12} \to \infty$, the requirement **(ii)** is satisfied.

For the requirement **(i)**, it remains to show that $M_n = O_p(1)$. According to the proof of Theorem 1 and Corollary 3.1. from [37], it suffices to show that as $p \to \infty$, $\big\|\sum_{j=p}^{\infty} \beta_j \phi_j(x)\big\|_4^4 =$

$O\big(\|\sum_{j=p}^{\infty}\beta_j\phi_j(x)\|_2^4\big)$. First, we have

$$\|\sum_{j=p}^{\infty}\beta_j\phi_j(x)\|_2^4=\big(\sum_{j=p}^{\infty}\beta_j^2\big)^2=\sum_{j=p}^{\infty}\beta_j^4+2\sum_{p\le j<k}\beta_j^2\beta_k^2.$$

Second, for $a\le b\le c\le d$ from the set $\{4\pi x,4^2\pi x,4^3\pi x,\cdots\}$, we have $\sin(a)\sin(b)\sin(c)\sin(d)=\frac{1}{8}(\cos(a-b+c-d)+\cos(a-b-c+d)-\cos(a-b+c+d)-\cos(a-b-c-d)-\cos(a+b+c-d)-\cos(a+b-c+d)+\cos(a+b+c+d)+\cos(a+b-c-d))$. It can be seen that $a-b+c+d$, $a-b-c-d$, $a+b+c-d$, $a+b-c+d$, $a+b+c+d$ are all in the form of πx times a non-zero even factor. Therefore, $\int_0^1(\sin(a)\sin(b)\sin(c)\sin(d))dx=\frac{1}{8}\int_0^1(\cos(a-b+c-d)+\cos(a-b-c+d)+\cos(a+b-c-d))dx$. It can been seen that for cases with $a<b<c<d$, $a=b<c<d$, $a<b=c<d$, $a<b<c=d$, $a=b=c<d$, $a<b=c=d$, we have $a-b+c-d$, $a-b-c+d$, and $a+b-c-d$ are all in the form of πx times a non-zero even factor. Therefore, $\int_0^1(\sin(a)\sin(b)\sin(c)\sin(d))dx=0$ in these cases. For $a=b<c=d$, we have $\int_0^1(\sin(a)\sin(b)\sin(c)\sin(d))dx=\frac{1}{8}(1+1+0)=\frac{1}{4}$. For $a=b=c=d$, we have $\int_0^1(\sin(a)\sin(b)\sin(c)\sin(d))dx=\frac{1}{8}(1+1+1)=\frac{3}{8}$. Therefore,

$$\begin{aligned}\left\|\sum_{j=p}^{\infty}\beta_j\phi_j(x)\right\|_4^4=&\int_0^1\left(\sum_{j=p}^{\infty}(\beta_j\phi_j(x))^4+6\sum_{p\le j<k}(\beta_j\phi_j(x))^2(\beta_k\phi_k(x))^2\right)dx\\=&\frac{3}{8}\sum_{j=p}^{\infty}\beta_j^4+\frac{3}{2}\sum_{p\le j<k}\beta_j^2\beta_k^2.\end{aligned}$$

We have $\frac{\|\sum_{j=p}^{\infty}\beta_j\phi_j(x)\|_4^4}{\|\sum_{j=p}^{\infty}\beta_j\phi_j(x)\|_2^4}=\frac{3}{8}+\frac{\frac{3}{4}\sum_{p\le j<k}\beta_j^2\beta_k^2}{\sum_{j=p}^{\infty}\beta_j^4+2\sum_{i<k}\beta_j^2\beta_k^2}\le\frac{3}{8}+\frac{3}{8}=\frac{3}{4}$. Thus, we obtain $M_n=O_p(1)$ and complete the proof.

10. Proof of Corollary 1

For the multiple splitting TCV by voting, we have

$$\begin{aligned}E\big(MTCV_{W_n}^v(\delta_{g_n})\big)=&\frac{1}{K}\sum_{k=1}^{K}E(\mathbb{1}(j=\arg\min_{i\in\mathcal{J}}TCV_{W_n,k}(\widehat{f}_{n_1,k}^{(i)})))\\=&P\left(TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)})=\min_{i\in\mathcal{J}}TCV_{W_n}(\widehat{f}_{n_1}^{(i)})\right)\end{aligned}$$

Therefore, when the requirements in Theorem 1 are met, we have

$$E\big(MTCV_{W_n}^v(\delta_{g_n})\big)\to 1,$$

as $n\to\infty$. Since $MTCV_{W_n}^v(\delta_{g_n})\le 1$, we have that $MTCV_{W_n}^v(\delta_{g_n})\to_p 1$, which implies that the multiple splitting TCV by voting is W_n-consistent.

For the multiple splitting TCV by averaging, we have

$$P\left(MTCV_{W_n}^a(\delta_{g_n})=\min_{i\in\mathcal{J}}MTCV_{W_n}^a(\delta_i)\right)$$

$$\geq 1 - K \cdot P\left(TCV_{W_n}(\widehat{f}_{n_1}^{(g_n)}) \geq \min_{i \in \mathcal{J}} TCV_{W_n}(\widehat{f}_{n_1}^{(i)})\right).$$

Since K is upper bounded by a fixed constant, we also have the
W_n-consistency of the multiple splitting TCV by averaging given the requirements in Theorem 1.

Acknowledgement

The authors sincerely thank three anonymous reviewers and the Associate Editor for their constructive and very helpful comments, which led to a substantial improvement of the work.

References

[1] Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127.

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

[3] Arlot, S. and Celisse, A. (2011). Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, 21(4):613–632.

[4] Arlot, S. and Lerasle, M. (2016). Choice of V for V-fold cross-validation in least-squares density estimation. *The Journal of Machine Learning Research*, 17(1):7256–7305.

[5] Baraud, Y. (2011). Estimator selection with respect to hellinger-type risks. *Probability theory and related fields*, 151(1-2):353–401.

[6] Baraud, Y., Giraud, C., and Huet, S. (2014). Estimator selection in the Gaussian setting. 50(3):1092–1119.

[7] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

[8] 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.

[9] Burman, P. (1990). Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā: The Indian Journal of Statistics, Series A*, 52(3):314–345.

[10] Celisse, A. (2014). Optimal cross-validation in density estimation with the L^2-loss. *The Annals of Statistics*, 42(5):1879–1910.

[11] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403.

[12] Ding, J., Tarokh, V., and Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34.

[13] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

[14] Feng, Y. and Yu, Y. (2019). The restricted consistency property of leave-n_v-out cross-validation for high-dimensional variable selection. *Statistica Sinica*, 29(3):1607–1630.

[15] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

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

[17] Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997.

[18] Li, K.-C. (1984). Consistency for cross-validated nearest neighbor estimates in nonparametric regression. *The Annals of Statistics*, 12(1):230–240.

[19] Li, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 15(3):958–975.

[20] Maillard, G., Arlot, S., and Lerasle, M. (2021). Aggregated hold-out. *Journal of Machine Learning Research*, 22(20):1–55.

[21] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

[22] Nemirovski, A. (2000). Topics in non-parametric statistics. *Lectures on Probability Theory and Statistics (Saint-Flour 1998)*, 1738:85–277.

[23] Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583.

[24] Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61.

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

[26] Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, 7(2):221–242.

[27] Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, 13(3):970–983.

[28] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

[29] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society. Series B*, 36(2):111–147.

[30] Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *The Annals of Statistics*, 11(4):1136–1141.

[31] Yang, Y. (2006). Comparing learning methods for classification. *Statistica Sinica*, 16(2):635–657.

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

[33] Yang, Y. (2008). Localized model selection for regression. *Econometric Theory*, 24(2):472–492.

[34] Zhan, Z. and Yang, Y. (2022). Profile electoral college cross-validation. *Information Sciences*, 586:24–40.

[35] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

[36] Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 21(1):299–313.

[37] Zhang, Y. and Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112.