Econometric Theory, **24**, 2008, 472–492. Printed in the United States of America.
DOI: 10.1017/S0266466608080195

LOCALIZED MODEL SELECTION FOR REGRESSION

YUHONG YANG
University of Minnesota

Research on model/procedure selection has focused on selecting a single model globally. In many applications, especially for high-dimensional or complex data, however, the relative performance of the candidate procedures typically depends on the location, and the globally best procedure can often be improved when selection of a model is allowed to depend on location. We consider localized model selection methods and derive their theoretical properties.

1. INTRODUCTION

In statistical modeling, usually a number of candidate models are considered. Traditional model selection theory and practice focus on the search for a single candidate that is treated as the true or best model. There are good reasons for this: (1) comparing the candidates in terms of global performance is a good starting point; (2) in simple situations, the relative performance (in ranking) of the candidate models may not depend on x (the vector of predictors); (3) if one of the candidate models is believed to describe the data very well, one does want to find it for optimal prediction and interpretation. However, for high-dimensional or complex data, global selection may be suboptimal.

Consider two example situations.

1. For univariate regression, if the mean function is infinite-dimensional (with respect to the candidate models), there does not seem to be a strong reason to believe that one model works better than the others at all x values. Thus a global ranking and selection may not be the best thing to do. This will be demonstrated later.
2. In many current statistical applications, the number of predictors is very large or even much larger than the number of observations. In such cases, there can be substantial uncertainty in modeling. For instance, when select-

This research was supported by U.S. National Science Foundation CAREER Grant DMS0094323. We thank three referees and the editors for helpful comments on improving the paper. Address correspondence to Yuhong Yang, School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, MN 55455, USA; e-mail: yyang@stat.umn.edu.

 © 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

ing important genes for explaining a response variable out of 5,000 genes based on 100 observations, any variable selection method is exploratory in nature, and it seems clear that one cannot expect the selected model to have really captured the relationship between the response and the predictors. When one considers different types of models, they often perform the best in different regions.

The preceding situations motivate the consideration of localized model selection. With the ever increasing computing power, selecting procedures locally becomes feasible in implementation.

Given the direction of localized model selection, one may wonder if it is better to take a local estimation approach in the first place and put the effort into building a single good estimator. In our opinion, this does not work in general. First, global considerations are important for avoiding overfit, and for high-dimensional situations, local estimation without global constraints is often not possible. Second, one has many different ways to do local estimation, and then one is back to the problem of procedure selection.

Now let us set up the problem mathematically. Let (X_i, Y_i), $i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) observations with X_i taking values in $\mathcal{X}$, a measurable set in R^d for some $d \geq 1$. Let $Y_i = f(X_i) + \varepsilon_i$, where f is the regression function to be estimated under squared error loss and the error ε_i has mean zero and variance σ^2. Unless stated otherwise, ε_i is assumed to be independent of X_i. Suppose $\delta_j, j \in J$ are a finite collection of statistical procedures for estimating the regression function, each producing an estimator of f based on a given sample. We will focus on the case with two procedures, although similar results hold more generally.

The rest of the paper is organized as follows. In Section 2, we give an illustration to motivate localized model selection. In Section 3, we mention three approaches to localized model selection. In Section 4, we provide a result that characterizes performance of a localized cross validation (CV) method. In Section 5, we study preference-region-based localized model selection. Concluding remarks are given in Section 6. Proofs of the theorems appear in the Appendix.

2. A MOTIVATING ILLUSTRATION

Consider estimating a regression function on $[-1, 1]$ based on 100 observations. The x values are uniformly spaced. The true regression function is

$$f(x) = 0.5x + \frac{0.8}{\sqrt{2\pi}} \exp(-200(x + 0.25)^2) - \frac{0.8}{\sqrt{2\pi}} \exp(-200(x - 0.25)^2),$$

and the error is from $N(0, 0.3^2)$.

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

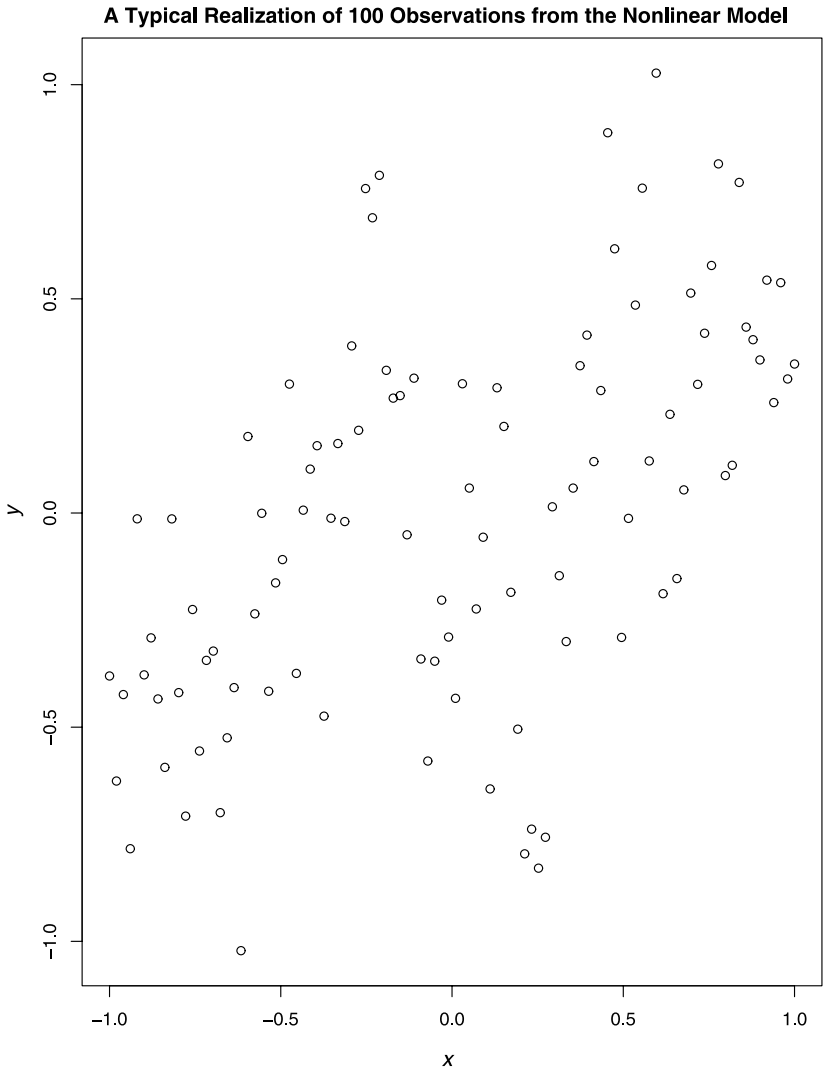


FIGURE 1. Scatter plot of a typical realization of 100 observations.

A typical realization of the data is given in Figure 1. The linear trend is obvious, but one also sees possible nonlinearities. Naturally, one may consider a simple linear model thinking that the somewhat unusual pattern in the middle might be caused by a few outliers or the deviation from linearity may not be serious enough to pursue. Alternatively, one may consider a nonparametric method. We choose a smoothing spline (SS) method that is provided in **R** (with the default choices of the control parameters).

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

2.1. Advantage of Localized Selection with a Certain Form of Preference Region

We compare a CV method for global selection between the linear regression and the SS procedures with a selection method that recognizes that the two estimators may perform well in different regions. For the CV method, we randomly split the data into two equally sized parts, find the linear and the SS estimates using the first part, and compute the prediction squared error on the second part. We repeat this 50 times and choose the one with smaller median prediction error over the random splittings. Let $\hat{f}^G(x)$ be the resulting estimator.

From the scatter plot, one may suspect that the linear estimate may perform poorly around 0, whereas the nonparametric estimate may be undersmooth when x is away from 0. Thus, for the competing nonglobal selection method, we consider estimators of the form

$$\hat{f}(x;c) = \hat{f}_L(x)I_{\{|x|\geq c\}} + \hat{f}_{SS}(x)I_{\{|x|<c\}},$$

where $\hat{f}_L$ is the estimator of f based on the linear model and $\hat{f}_{SS}$ is the SS estimator. We use CV similarly as before to choose c in the range of $[0,1]$ at a grid of width 0.01. Let $\hat{f}^{NG}(x)$ be the resulting estimator. Note that when $c = 0$, the linear estimator is selected for all x and when $c = 1$, the nonparametric estimator is selected for all x, but for c between, we use SS only when $|x|$ is no bigger than c.

At a given x_0, compute the squared error losses of $\hat{f}^G(x_0)$ and $\hat{f}^{NG}(x_0)$, respectively. Generate 200 independent data sets from the true model to simulate the risks of $\hat{f}^G(x_0)$ and $\hat{f}^{NG}(x_0)$. The results are presented in Figure 2 at a number of x_0 values in the range of $(-1,1)$.

The figure clearly shows that the nonglobal selection performs much better than the global selection except in a very small neighborhood around zero, demonstrating the potentially great advantage of considering nonglobal selection among candidate procedures.

It is probably fair to say that one may not necessarily choose the form of $\hat{f}(x;c)$ from inspecting the scatter plot. In Section 2.2, we use logistic regression to find the preference region automatically.

2.2. Classification for Estimating the Preference Region

We continue with the earlier setting except that the sample size is now 200 (so that the contrast is more clearly seen in the next figure). We focus on one typical realization of the data. A scatter plot of the data with the linear and the SS fits is given in the upper left panel of Figure 3. We use logistic regression as the classification method to find the region where the linear estimator performs better than the SS estimator. With a random splitting of the data into two parts of equal size, we fit a straight-line model and the SS using the first 100 observations and obtain the binary variable that indicates which method has a smaller

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

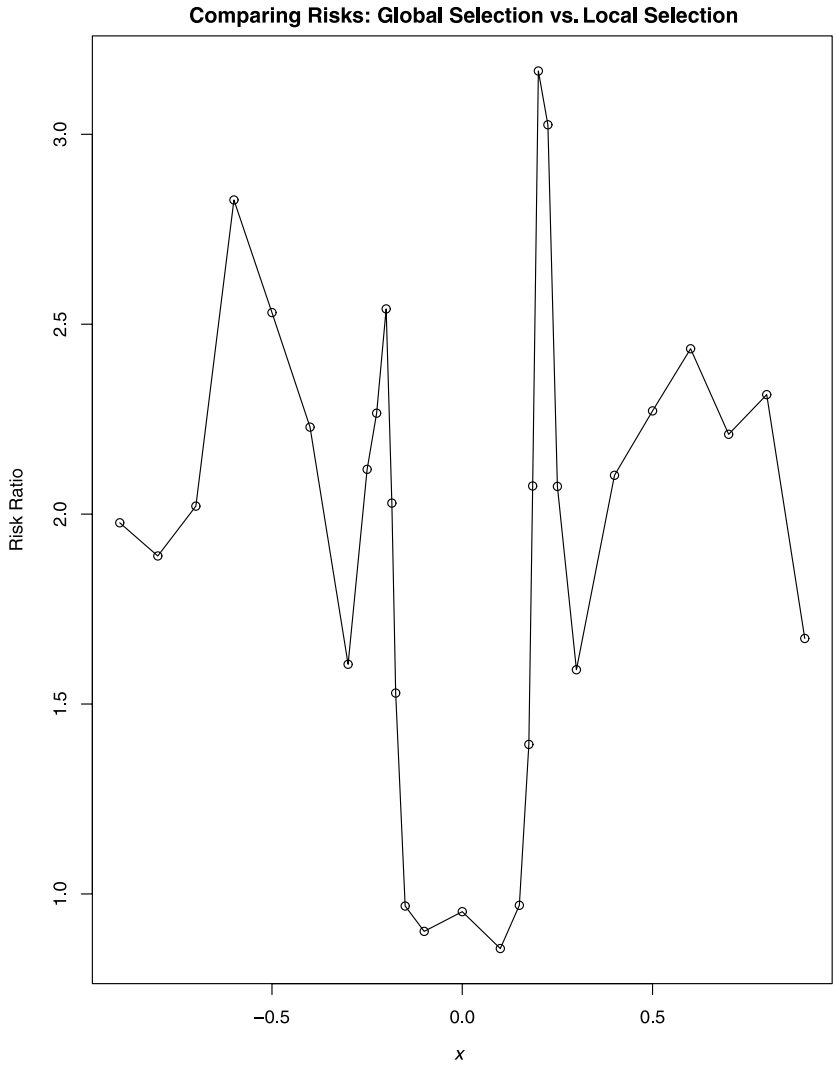


FIGURE 2. Risk ratio of a global CV vs. a nonglobal selection method.

prediction error on the second 100 observations. A typical outcome is in the upper right panel of Figure 3. One may get the impression that SS tends to do better in the middle and less well at the ends. We fit a logistic regression model with three terms: 1, x, and x^2. The estimated probability that the linear model performs better at x is

$$\hat{p}(x) = \frac{1}{1 + \exp(0.379 - 0.025x - 1.497x^2)}.$$

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

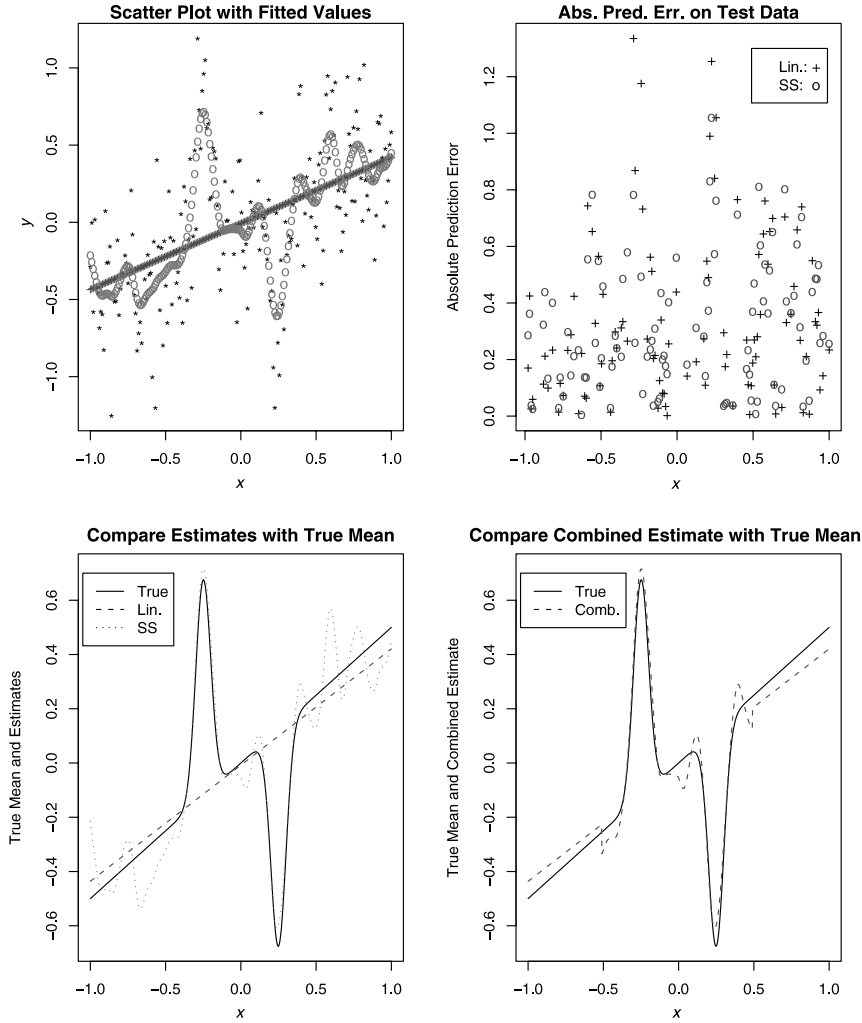


FIGURE 3. An example of using logistic regression for nonglobal selection.

Note that $\hat{p}(x) > 0.5$ corresponds to $x < (-0.51)$ or $x > 0.49$, which is very sensible judging from our knowledge of the true mean function (despite that the upper right panel of the figure does not seem to be very visually informative for relating the relative performance of the two estimators and x). This yields the following combined estimate of the regression function:

$$\hat{f}(x) = \begin{cases} \hat{f}_L(x) & \text{if } x < (-0.51) \text{ or } x > 0.49 \\ \hat{f}_{SS}(x) & \text{otherwise,} \end{cases}$$

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

where $\hat{f}_L(x)$ and $\hat{f}_{SS}(x)$ are the linear and SS estimates of f based on the full data. This estimate is compared with the true function in the lower right panel of Figure 3.

From the lower left panel of Figure 3, we see that the linear estimate is very poor around -0.25 and 0.25 (not surprisingly); in contrast, the SS estimate is quite good there, but it pays a price for being flexible in terms of accuracy at other locations. Globally, the SS is much better than the linear estimate. Indeed, the ratio of their integrated squared errors is over 6.23. Thus from the global perspective, the SS estimate is the clear winner. However, with the use of logistic regression, we properly figured out that when x is far away from zero by a certain degree, the linear estimate is better. Although the linear estimate is slightly biased even in the linear parts, its use in the final combined estimate is very helpful as seen in the figure. Numerically, the integrated squared error of the combined estimate is only 35% of that of the SS estimate. Therefore a globally poor method can still make a good contribution if it is properly combined with a globally much better estimator.

3. THREE APPROACHES TO NONGLOBAL MODEL SELECTION

Consider two procedures δ_1 and δ_2 with risks $R(\delta_1;x;n)$ and $R(\delta_2;x;n)$ at a given x value based on a sample of size n. Let $A^* = \{x : R(\delta_1;x;n) \le R(\delta_2;x;n)\}$ be the set of x at which the first procedure performs no worse than the second one. It is the preference region of δ_1 (relative to δ_2). Ideally, for selection, one would use δ_1 on A^* and δ_2 on $(A^*)^c$. In reality, one may have little prior knowledge of the form of A^*, and to deal with the issue, one may consider graphical inspections when x is of a low dimension or consider A from a class of sets with proper complexity in the hope that one member is close to A^*. One can consider various sets of A of different degrees of locality. We consider three approaches here.

3.1. Neighborhood-Based Selection

Instead of directly estimating A^*, at each given x_0, one considers a local neighborhood around x_0 and tries to find the candidate that performs better in the local area. This approach can be very helpful when A^* cannot be described well by simple sets. It will be studied in detail in Section 4.

3.2. Empirical Risk Minimization

One considers a collection of sets of a certain mathematical form and tries to identify the one with the best performance. Here the collection may depend on the sample size. The size of the collection can be predetermined or adaptively chosen. This approach will be explored briefly in Section 5.

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

3.3. Classification-Based Selection

Sometimes, one is less confident to go with any specific collection of sets as the preference region. This is especially true when the input dimension is high, where neighborhood-based selection may perform poorly because of the curse of dimensionality. In such a case, one can take advantage of classification methods to conduct localized selection as in Section 2.2. In general, one splits the data into two parts. The first part is used to obtain the estimates from the candidate procedures and then to make predictions for the second part. Based on the relative predictive performance in each case, we create a new variable that simply indicates which estimate is the better one. Then we can apply a sensible classification method to relate the performance indication variable to the covariates. If a classifier performs well, it indicates that the candidate procedures' relative performance depends on the covariates, and we can do better than globally selecting one of the procedures.

4. LOCALIZED CROSS VALIDATION SELECTION

CV (see, e.g., Allen, 1974; Stone, 1974; Geisser, 1975) has been widely used for model selection. In this section, we consider a localized CV method for selecting a candidate procedure locally. For a given x_0, consider the ball centered at x_0 with radius r_n for some $r_n > 0$ under the euclidean distance. We randomly split the data into a training set of size n_1 and a test set of size n_2, and then we use the training set to do estimation by each regression procedure. For evaluation, consider only the data points in the test set that are in the given neighborhood of x_0. Let $\hat{j}(x_0) = \hat{j}_n(x_0)$ be the procedure that has the smaller average squared prediction error. This process is repeated with a number of random splittings of the observations to avoid the splitting bias. The procedure that wins more frequently over the permutations is the final winner. We call this a localized cross validation (L-CV) at x_0 of r_n-neighborhood.

For the local selection problem, we are interested in the question of under which conditions the locally better estimator will be selected with high probability. More precisely, assuming that when n is large enough, one procedure has risk at x_0 smaller than that of the other one, we want to select the better one with probability approaching 1. If a selection method has this property, we say it is locally consistent in selection at x_0. For related results on global CV, see, e.g., Li (1987), Burman (1989), Zhang (1993), Shao (1997), Wegkamp (2003), and Yang (2007).

Given x_0 and $\eta > 0$, let $L(\hat{f}; x_0; \eta) = \int_{x \in B(x_0;\eta)} (\hat{f}(x) - f(x))^2 P_{\bar{X}}(dx)$ be a local loss of an estimate $\hat{f}$ around x_0, where $P_{\bar{X}}$ denotes the distribution of X conditional on that X takes value in the neighborhood $B(x_0; \eta)$. Let $\|\cdot\|_{s,x_0;\eta}$ denote the L_s norm around x_0 with radius η, i.e., $\|g\|_{s,x_0;\eta} = (\int_{x \in B(x_0;\eta)} |g(x)|^s P_{\bar{X}}(dx))^{1/s}$.

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

DEFINITION 1. *Procedure δ_1 (or $\hat{f}_{1,n}$) is asymptotically better than δ_2 (or $\hat{f}_{2,n}$) under the squared loss at η_n-neighborhood of x_0, denoted $\delta_1 \lhd \delta_2$ at $(x_0;\eta_n)$, if for each nonincreasing sequence $\tilde{\eta}_n$ with $0 < \tilde{\eta}_n \le \eta_n$ and every $0 < \epsilon < 1$, there exists a constant $c_\epsilon > 0$ such that when n is large enough,*

$$P(L(\hat{f}_{2,n};x_0;\tilde{\eta}_n) \ge (1 + c_\epsilon)L(\hat{f}_{1,n};x_0;\tilde{\eta}_n)) \ge 1 - \epsilon. \quad \textbf{(1)}$$

When one of the procedures is asymptotically better than the other at a neighborhood of x_0, we say that δ_1 and δ_2 are ordered at x_0. One may wonder if it is better to require the condition in (1) for η_n only. The answer is no because it is possible that $L(\hat{f}_{2,n};x_0;\eta_n)$ is smaller than $L(\hat{f}_{1,n};x_0;\eta_n)$ with high probability for one sequence of η_n yet the opposite holds for a smaller sequence $0 < \eta_n' \le \eta_n$. This can happen, e.g., when δ_1 is asymptotically better than δ_2 globally (i.e., with no restriction on η_n) but worse locally. In general, the space $\mathcal{X}$ can be decomposed into three regions: those x_0 at which δ_1 is asymptotically better than δ_2, those at which δ_2 is asymptotically better than δ_1, and the rest of x_0 at which δ_1 and δ_2 cannot be compared according to Definition 1. Local selection can be hoped to be successful only for the first two regions.

DEFINITION 2. *A procedure δ (or $\{\hat{f}_n\}_{n=1}^{\infty}$) is said to converge exactly at rate $\{a_n\}$ in probability at η_n-neighborhood of x_0 if for each nonincreasing sequence $\tilde{\eta}_n$ with $0 < \tilde{\eta}_n \le \eta_n$, $L(\hat{f};x_0;\tilde{\eta}_n) = O_p(a_n)$, and for every $0 < \epsilon < 1$, there exists $c_\epsilon > 0$ such that when n is large enough, $P(L(\hat{f};x_0;\tilde{\eta}_n) \ge c_\epsilon a_n) \ge 1 - \epsilon$.*

DEFINITION 3. *A selection method is said to be locally consistent in selection at x_0 if δ_1 and δ_2 are ordered at x_0 and the asymptotically better procedure at x_0 is selected with probability going to 1. If the selection method is locally consistent at every x_0 at which the procedures are ordered, it is said to be locally consistent.*

In Definition 2, the second condition simply indicates that the loss does not converge faster than a_n in an appropriate sense. For the following result, we assume that $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ converge exactly at rate $\{p_n\}$ and $\{q_n\}$ in probability at η_n-neighborhood of x_0, respectively, where p_n and q_n are two sequences of nonincreasing positive numbers.

Some technical conditions are needed.

Condition 1. There exists a sequence of positive numbers A_n such that for $j = 1, 2$,

$$\|f - \hat{f}_{j,n}\|_\infty = O_p(A_n).$$

Condition 2. There exists $\eta_n > 0$ such that $\delta_1 \lhd \delta_2$ at $(x_0;\eta_n)$ or $\delta_2 \lhd \delta_1$ at $(x_0;\eta_n)$. Let $j^*(x_0)$ denote the better procedure.

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

Condition 3. There exists a sequence of positive numbers $\{M_n\}$ such that for $j = 1, 2$, for each nonincreasing sequence $0 \le \tilde{\eta}_n \le \eta_n$, we have

$$\frac{\|f - \hat{f}_{j,n}\|_{4,x_0;\tilde{\eta}_n}}{\|f - \hat{f}_{j,n}\|_{2,x_0;\tilde{\eta}_n}} = O_p(M_n).$$

THEOREM 1. *Under Conditions 1–3, as long as n_1, n_2, and r_n are chosen to satisfy (1): $n_2 r_n^d M_{n_1}^{-4} \to \infty$; (2) $\sqrt{n_2 r_n^d}\max(p_{n_1}, q_{n_1})/(1 + A_{n_1}) \to \infty$; (3) $r_n \le \eta_n$, we have that with probability going to 1, the better procedure $\hat{f}_{j^*(x_0),n}$ will be selected.*

From the result, several factors affect the ability of the L-CV to identify the best estimator at a given x_0. The larger η_n, the larger the window for choosing the neighbor size r_n for L-CV. Intuitively, if η_n is very close to zero, which occurs when the two procedures constantly switch position in terms of local performance, it may not be feasible to identify the locally better one at all. The constants M_n and A_n come into the picture more for a technical reason. For simplicity, for the following discussion, we assume that M_n and A_n are both of order 1. Then the requirements on the choice of data splitting ratio and the neighbor size become $\sqrt{n_2 r_n^d}\max(p_{n_1}, q_{n_1}) \to \infty$ and $r_n \le \eta_n$. Consider, e.g., $\eta_n = (\log n)^{-1}$. When at least one of the procedures is nonparametric with $\max(p_n, q_n)$ converging more slowly than the parametric rate $n^{-1/2}$, Yang (2007) showed that for selecting the globally better procedure, it suffices to take n_2 at least of the same order as n_1 (which is not enough for comparing parametric models as shown in Shao, 1993). With local selection, however, we need to be more careful so as to satisfy

$$\frac{1}{n_2 \max(p_{n_1}^2, q_{n_1}^2)} \ll r_n^d \le (\log n_1)^{-d}.$$

If $\max(p_n, q_n)$ is of order $n^{-1/3}$, then with $n_1 = n_2$, the condition is simply $n^{-1/(3d)} \ll r_n \le (\log n_1)^{-1}$.

When the number of observations is moderately large, one can use a data dependent approach to choose r_n. For example, taking r_n to be the same for all x, one may consider another level of data splitting for empirically evaluating the performance of each choice of r_n. Theoretical results are possible, but we will not pursue them in this work.

5. PREFERENCE REGION SELECTION

Let $\hat{f}_1(x) = \hat{f}_{1,n_1}(x)$ and $\hat{f}_2(x) = \hat{f}_{2,n_1}(x)$ be the estimates by the two regression procedures based on the first part of the data of size n_1. For a measurable set A, define

$$\hat{f}(x;A) = \hat{f}_1(x) I_{\{x \in A\}} + \hat{f}_2(x) I_{\{x \notin A\}}.$$

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

Because

$$E(\hat{f}(x;A)-f(x))^2 = E((\hat{f}_1(x)-f(x))I_{\{x\in A\}} + (\hat{f}_2(x)-f(x))I_{\{x\notin A\}})^2$$
$$= E(\hat{f}_1(x)-f(x))^2 I_{\{x\in A\}} + E(\hat{f}_2(x)-f(x))^2 I_{\{x\notin A\}},$$

it follows that the preference region $A^* = \{x : E(\hat{f}_1(x)-f(x))^2 \le E(\hat{f}_2(x)-f(x))^2\}$ is the best choice of A. Obviously, we have

$$E\|\hat{f}(;A^*)-f\|^2 \le \min(E\|\hat{f}_1-f\|^2, \|\hat{f}_2-f\|^2);$$

i.e., if it is possible to identify A^*, the estimator $\hat{f}(x;A^*)$ is better (or at least no worse) than the global winner between the two candidate procedures. The ideal local selection can be arbitrarily better than the global selection in terms of their risk ratio.

We consider a collection of sets of manageable complexity and try to find the best set A within the class. Let $\mathcal{A}_n$ be a class of sets and let A_n^* be the set that minimizes $E\|\hat{f}(;A)-f\|^2$ over $\mathcal{A}_n$. Then

$$E\|\hat{f}(;A_n^*)-f\|^2 - E\|\hat{f}(;A^*)-f\|^2$$

is the approximation error because of the use of a set in $\mathcal{A}_n$, which may not contain A^*. Of course, A_n^* is also unknown and needs to be estimated.

Let $\mu(B \triangle A)$ denote the probability (under the distribution of X_i) of the symmetric difference of A and B. Assume that $\mathcal{A}_n$ has metric entropy bounded above by $H_n(\epsilon)$ under the distance $d(A,B) = \mu(B \triangle A)$ (for concept and related results involving metric entropy, see, e.g., Kolmogorov and Tihomirov, 1959; Yatracos, 1985; Birgé, 1986; van de Geer, 1993; Yang and Barron, 1999). Let $\mathcal{A}_{n,0}$ be a discretized collection of sets that serves as an ϵ_n-net for $\mathcal{A}_n$ under d. Let $\hat{A}_n$ be the minimizer of the empirical prediction error over $\mathcal{A}_{n,0}$, i.e.,

$$\hat{A}_n = \arg\min_{A\in\mathcal{A}_{n,0}} \frac{1}{n_2}\sum_{i=n_1+1}^{n}(Y_i-\hat{f}(x_i;A))^2.$$

THEOREM 2. *Assume that the errors are normally distributed with mean zero and variance $\sigma^2 > 0$ and $\|\hat{f}_{j,n}-f\|_\infty \le \bar{C}$ a.s. for some constant $\bar{C} < \infty$ for both procedures. Then the final estimator $\hat{f}(;\hat{A}_n)$ satisfies*

$$E\|\hat{f}(;\hat{A}_n)-f\|^2$$
$$\le C_1\left(E\|\hat{f}(;A_n^*)-f\|^2 + \epsilon_n + \frac{H_n(\epsilon_n)}{n_2}\right)$$
$$\le C_2\left(E\|f(;A^*)-f\|^2 + E\|\hat{f}(;A_n^*)-f(;A^*)\|^2 + \epsilon_n + \frac{H_n(\epsilon_n)}{n_2}\right),$$

where the constants C_1 and C_2 depend only on $\bar{C}$ and σ^2.

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

To optimize the risk bound given previously, with $\mathcal{A}_n$ given, we need to balance $(H_n(\epsilon_n))/n_2$ and ϵ_n. The issue becomes more complicated when one needs to choose $\mathcal{A}_n$. Clearly, a large choice of $\mathcal{A}_n$ reduces the potential approximation error but at the same time increases the estimation error because of searching over a larger class of sets A. One approach to handling the approximation error is to assume that A^* is in a given collection of sets $\mathcal{B}$ and then to characterize the uniform approximation error of sets in $\mathcal{B}$ by sets in $\mathcal{A}_n$. Define $\gamma_n = \sup_{B\in\mathcal{B}} \inf_{A\in\mathcal{A}_n} \mu(B \triangle A)$. Under proper assumptions on $\mathcal{B}$ and $\mathcal{A}_n$, the rate of convergence of γ_n can be derived and then used for obtaining the convergence rate of the final estimator from the localized selection.

Example

Let $\mathcal{A}_k$ consist of all possible unions of at most k cubes in $\mathcal{X} = [0,1]^d$. Let $\mathcal{B}$ be the collection of all the sets that each can be well approximated by a set in $\mathcal{A}_k$ in the sense that for each $B \in \mathcal{B}$, there exists a set A in $\mathcal{A}_k$ such that $\mu(B \triangle A) \le ck^{-\tau}$ for some constants $c, \tau > 0$. For simplicity, we assume that X_i has a uniform distribution on $[0,1]^d$, although a similar result holds if X_i has a uniformly upper bounded Lebesgue density.

To obtain an ε-net in $\mathcal{A}_k$ for $\mathcal{B}$, we first choose $k = k_\varepsilon$ such that $ck^{-\tau} \le \varepsilon/2$ (k_ε is then of order $\varepsilon^{-1/\tau}$). Then for any $B \in \mathcal{B}$, there exists a set A in $\mathcal{A}_{k_\varepsilon}$ such that $\mu(B \triangle A) \le \varepsilon/2$. Consequently, an $\varepsilon/2$-net in $\mathcal{A}_{k_\varepsilon}$ will be an ε-net for $\mathcal{B}$. Now if we have a grid on $[0,1]^d$ of width $\varepsilon/(2k)$ for each coordinate, then the collection of all cubes with vertices on the grid form an $\varepsilon/(2k)$-net for $\mathcal{A}_{k_\varepsilon}$ and thus also an ε-net for $\mathcal{B}$. The number of possible unions of k such cubes is of order

$$\left(\frac{k_\varepsilon}{\varepsilon}\right)^{k_\varepsilon d} = \varepsilon^{-(1+1/\tau)\varepsilon^{-1/\tau} d}.$$

It follows that the metric entropy of $\mathcal{B}$ is of order $O(\varepsilon^{-1/\tau}\log(\varepsilon^{-1}))$.

If we select A over the ε-net in $\mathcal{A}_k$ by CV with a proper discretization, by Theorem 2, with a choice of n_1 and n_2 both of order n, the combined estimator satisfies

$$E\|\hat{f}(;\hat{A}_n) - f\|^2 = O\left(E\|f(;A^*) - f\|^2 + \epsilon_n + \frac{\log \epsilon_n^{-1}}{n\epsilon_n^{1/\tau}}\right).$$

Balancing the last two terms in the preceding expression, the optimal choice of ϵ_n is $(\log n/n)^{\tau/(\tau+1)}$, and consequently the risk bound becomes

$$E\|\hat{f}(;\hat{A}_n) - f\|^2 = O(E\|f(;A^*) - f\|^2 + (\log n/n)^{\tau/(\tau+1)}).$$

Thus, when τ is large, the additional term $(\log n/n)^{\tau/(\tau+1)}$ is close to $\log n/n$, which is typically negligible for nonparametric regression. Then we achieve

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

the performance of the ideal localized selection up to a relatively small discrepancy term.

6. CONCLUDING REMARKS

Overall performance of a candidate model (procedure) has been the dominating measure used for model selection. If one candidate model is thought to be "true" or "optimal," it is then rational to try to identify it. However, in many applications, this practice is suboptimal because the globally best candidate procedure, even if it is much better than others, may still have unsatisfactory local behaviors in certain regions, which can be remedied well with helps from other candidates that are globally inferior.

We took two directions in localized model selection: a localized CV that selects a model/procedure at each x value based on performance assessment in a neighborhood of x and a preference region selection from a collection of sets that tries to estimate the regions where each candidate performs better. For the localized CV, as long as the neighborhood size and the data splitting ratio are chosen properly, the locally better estimator will be selected with high probability. For preference region selection, when the complexity of the collection of the candidate sets of the preference region is properly controlled, the final procedure based on the empirically selected preference region behaves well in risk.

Besides global selection and localized selection of a model, sometimes it is advantageous to consider global combination of the candidates with weights globally determined or localized combination where weights depend on the x value (see, e.g., Pan, Xiao, and Huang, 2006). For high-dimensional or complex data, these alternatives can provide flexibility needed to further improve performance of the candidate procedures.

REFERENCES

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

Birgé, L. (1986) On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields* 71, 271–291.

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

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

Kolmogorov, A.N. & V.M. Tihomirov (1959) ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspehi Mat. Nauk* 14, 3–86.

Li, K.-C. (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* 15, 958–975.

Pan, W., G. Xiao, & X. Huang (2006) Using input dependent weights for model combination and model selection with multiple sources of data. *Statistics Sinica* 16, 523–540.

Pollard, D. (1984) *Convergence of Stochastic Processes*. Springer.

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

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

Shao, J. (1997) An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* 7, 221–242.
Stone, M. (1974) Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* 36, 111–147.
van de Geer, S. (1993) Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics* 21, 14–44.
Wegkamp, M.H. (2003). Model selection in nonparametric regression. *Annals of Statistics* 31, 252–273.
Yang, Y. (2007) Consistency of cross validation for comparing regression procedures. *Annals of Statistics*, forthcoming.
Yang, Y. & A.R. Barron (1999) Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 27, 1564–1599.
Yatracos, Y.G. (1985) Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics* 13, 768–774.
Zhang, P. (1993) Model selection via multifold cross validation. *Annals of Statistics* 21, 299–313.

APPENDIX

Proof of Theorem 1. Much of the proof follows from Yang (2007), except that localized selection is considered in this work. We first focus on the analysis without multiple data splittings. Without loss of generality, assume that δ_1 is asymptotically better than δ_2 at x_0. Let $I = I_{x_0, r_n} = \{i : n_1 + 1 \le i \le n \text{ and } X_i \in B(x_0; r_n)\}$ denote the observations in the evaluation set with X_i close to x_0 and let $\tilde{n}_2$ be the size of I. Because

$$\begin{aligned} LCV(\hat{f}_{j,n_1}) &= \sum_{i\in I} (Y_i - \hat{f}_{j,n_1}(X_i))^2 \\ &= \sum_{i\in I} (f(X_i) - \hat{f}_{j,n_1}(X_i) + \varepsilon_i)^2 \\ &= \sum_{i\in I} \varepsilon_i^2 + \sum_{i\in I} (f(X_i) - \hat{f}_{j,n_1}(X_i))^2 + 2\sum_{i\in I} \varepsilon_i (f(X_i) - \hat{f}_{j,n_1}(X_i)), \end{aligned}$$

$LCV(\hat{f}_{1,n_1}) \le LCV(\hat{f}_{2,n_1})$ is equivalent to

$$2\sum_{i\in I} \varepsilon_i (\hat{f}_{2,n_1}(X_i) - \hat{f}_{1,n_1}(X_i)) \le \sum_{i\in I} (f(X_i) - \hat{f}_{2,n_1}(X_i))^2 - \sum_{i\in I} (f(X_i) - \hat{f}_{1,n_1}(X_i))^2.$$

Conditional on $Z^1 = (X_i, Y_i)_{i=1}^{n_1}$ and $X^2 = (X_{n_1+1}, \ldots, X_n)$, assuming $\sum_{i\in I}(f(X_i) - \hat{f}_{2,n_1}(X_i))^2$ is larger than $\sum_{i\in I}(f(X_i) - \hat{f}_{1,n_1}(X_i))^2$, by Chebyshev's inequality, we have

$$\begin{aligned} &P(LCV(\hat{f}_{1,n_1}) > LCV(\hat{f}_{2,n_1}) | Z^1, X^2) \\ &\le \min\left(1, \frac{4\sigma^2 \sum_{i\in I} (\hat{f}_{2,n_1}(X_i) - \hat{f}_{1,n_1}(X_i))^2}{\left(\sum_{i\in I} (f(X_i) - \hat{f}_{2,n_1}(X_i))^2 - \sum_{i\in I} (f(X_i) - \hat{f}_{1,n_1}(X_i))^2\right)^2}\right). \end{aligned}$$

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

Let Q_n denote the ratio in the upper bound in the preceding inequality and let S_n be the event of

$$\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2 > \sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2.$$

Then because

$$\begin{aligned}
&P(LCV(\hat{f}_{1,n_1}) > LCV(\hat{f}_{2,n_1}))\\
&\quad = P(\{LCV(\hat{f}_{1,n_1}) > LCV(\hat{f}_{2,n_1})\}\cap S_n) + P(\{CV(\hat{f}_{1,n_1}) > CV(\hat{f}_{2,n_1})\}\cap S_n^c)\\
&\quad \le E(P(CV(\hat{f}_{1,n_1}) > CV(\hat{f}_{2,n_1})|Z^1,X^2)I_{S_n}) + P(S_n^c)\\
&\quad \le E\min(1,Q_n) + P(S_n^c),
\end{aligned}$$

for consistency, it suffices to show that $P(S_n^c)\to 0$ and $Q_n\to 0$ in probability. Suppose we can show that for each $\epsilon>0$, there exists $\alpha_\epsilon>0$ such that when n is large enough,

$$P\left(\frac{\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2}{\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2}\ge 1+\alpha_\epsilon\right)\ge 1-\epsilon. \tag{A.1}$$

Then $P(S_n)\ge 1-\epsilon$ and thus $P(S_n^c)\to 0$ as $n\to\infty$. Because

$$\sum_{i\in I}(\hat{f}_{2,n_1}(X_i)-\hat{f}_{1,n_1}(X_i))^2 \le 2\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 + 2\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2,$$

with probability no less than $1-\epsilon$, we have

$$\begin{aligned}
Q_n &\le \frac{8\sigma^2\left(\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 + \sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2\right)}{\left(\left(1-\frac{1}{1+\alpha_\epsilon}\right)\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2\right)^2}\\
&\le \frac{8\sigma^2\left(1+\frac{1}{1+\alpha_\epsilon}\right)}{\left(1-\frac{1}{1+\alpha_\epsilon}\right)^2\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2}.
\end{aligned} \tag{A.2}$$

From (A.1) and (A.2), to show $P(S_n^c)\to 0$ and $Q_n\to 0$ in probability, it suffices to show (A.1) and

$$\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2\to\infty \quad \text{in probability.} \tag{A.3}$$

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

Suppose a slight relaxation of Condition 1 holds: for every $\epsilon > 0$, there exists $A_{n_1,\epsilon}$ such that for $j = 1, 2$, when n_1 is large enough,

$$P(\|f - \hat{f}_{j,n_1}\|_\infty \geq A_{n_1,\epsilon}) \leq \epsilon.$$

Let H_{n_1} denote the event $\{\max(\|f - \hat{f}_{1,n_1}\|_\infty, \|f - \hat{f}_{2,n_1}\|_\infty) \leq A_{n_1,\epsilon}\}$. Then on H_{n_1}, we have $W_i = (f(X_i) - \hat{f}_{j,n_1}(X_i))^2 - \|f - \hat{f}_{j,n_1}\|_2^2$ with $X_i \in B(x_0; r_n)$ is bounded between $-(A_{n_1,\epsilon})^2$ and $(A_{n_1,\epsilon})^2$. Conditional on Z^1 and H_{n_1}, $\mathrm{Var}_{Z^1}(W_{n_1+1}) \leq E_{Z^1}(f(X_{n_1+1}) - \hat{f}_{j,n_1}(X_{n_1+1}))^4 = \|f - \hat{f}_{j,n_1}\|_{4,x_0;r_n}^4$, where the subscript Z^1 in Var_{Z^1} and E_{Z^1} is used to denote the conditional expectation given Z^1. Thus conditional on Z^1, on H_{n_1}, by Bernstein's inequality (see, e.g., Pollard, 1984, p. 193), for each $x > 0$,

$$P_{Z^1}\left(\sum_{i\in I}(f(X_i) - \hat{f}_{1,n_1}(X_i))^2 - \tilde{n}_2\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2 \geq x\right)$$
$$\leq \exp\left(-\frac{1}{2}\frac{x^2}{\tilde{n}_2\|f - \hat{f}_{1,n_1}\|_{4,x_0;r_n}^4 + \dfrac{2(A_{n_1,\epsilon})^2 x}{3}}\right).$$

Taking $x = \beta_n\tilde{n}_2\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2$, the preceding inequality becomes

$$P_{Z^1}\left(\sum_{i\in I}(f(X_i) - \hat{f}_{1,n_1}(X_i))^2 \geq (1+\beta_n)\tilde{n}_2\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2\right)$$
$$\leq \exp\left(-\frac{1}{2}\frac{\beta_n^2\tilde{n}_2\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^4}{\|f - \hat{f}_{1,n_1}\|_{4,x_0;r_n}^4 + \dfrac{2(A_{n_1,\epsilon})^2\beta_n}{3}\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2}\right).$$

Under Condition 2, for every $\epsilon > 0$, there exists $\alpha'_\epsilon > 0$ such that when n is large enough,

$$P\left(\frac{\|f - \hat{f}_{2,n_1}\|_{2,x_0;r_n}^2}{\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2} \leq 1 + \alpha'_\epsilon\right) \leq \epsilon.$$

Take β_n such that

$$1 + \beta_n = \frac{\|f - \hat{f}_{2,n_1}\|_{2,x_0;r_n}^2}{(1 + \alpha'_\epsilon/2)\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2}.$$

Then with probability at least $1 - \epsilon$, $\beta_n \geq (\alpha'_\epsilon/2)/(1 + \alpha'_\epsilon/2)$. Let D_n denote this event. Then on D_n, we have

$$\beta_n = \frac{\|f - \hat{f}_{2,n_1}\|_{2,x_0;r_n}^2 - (1 + \alpha'_\epsilon/2)\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2}{(1 + \alpha'_\epsilon/2)\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2}$$
$$\geq \frac{\|f - \hat{f}_{2,n_1}\|_{2,x_0;r_n}^2 - \dfrac{(1 + \alpha'_\epsilon/2)\|f - \hat{f}_{2,n_1}\|_{2,x_0;r_n}^2}{1 + \alpha'_\epsilon}}{(1 + \alpha'_\epsilon/2)\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2}$$
$$= \frac{\alpha'_\epsilon\|f - \hat{f}_{2,n_1}\|_{2,x_0;r_n}^2}{2(1 + \alpha'_\epsilon)(1 + \alpha'_\epsilon/2)\|f - \hat{f}_{1,n_1}\|_{2,x_0;r_n}^2}$$

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

and

$$P_{Z^1}\left(\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 \ge (1+\beta_n)\tilde{n}_2\|f-\hat{f}_{1,n_1}\|^2_{2,x_0;r_n}\right)$$

$$= P_{Z^1}\left(\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 \ge \frac{\tilde{n}_2}{1+\alpha'_\epsilon/2}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\right)$$

$$\le P_{Z^1}\left(\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 \right.$$

$$\left. \ge \left(1+\frac{\alpha'_\epsilon\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}}{2(1+\alpha'_\epsilon)(1+\alpha'_\epsilon/2)\|f-\hat{f}_{1,n_1}\|^2_{2,x_0;r_n}}\right)\tilde{n}_2\|f-\hat{f}_{1,n_1}\|^2_{2,x_0;r_n}\right)$$

$$\le \exp\left(-\frac{(\alpha'_\epsilon)^2}{8(1+\alpha'_\epsilon)^2(1+\alpha'_\epsilon/2)^2}\right.$$

$$\left. \times \frac{\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^4_{2,x_0;r_n}}{\|f-\hat{f}_{1,n_1}\|^4_{4,x_0;r_n}+\frac{\alpha'_\epsilon(A_{n_1,\epsilon})^2}{3(1+\alpha'_\epsilon)(1+\alpha'_\epsilon/2)}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}}\right).$$

If we have

$$\frac{\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^4_{2,x_0;r_n}}{\|f-\hat{f}_{1,n_1}\|^4_{4,x_0;r_n}} \to \infty \quad \text{in probability,} \tag{A.4}$$

$$\frac{\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}}{(A_{n_1,\epsilon})^2} \to \infty \quad \text{in probability,} \tag{A.5}$$

then the upper bound in the last inequality in the expression preceding (A.4) converges to zero in probability. From these pieces, we can conclude that

$$P\left(\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 \ge \frac{\tilde{n}_2}{1+\alpha'_\epsilon/2}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\right) \le 3\epsilon+\Delta(\epsilon,n), \tag{A.6}$$

for some $\Delta(\epsilon,n)\to 0$ as $n\to\infty$. Indeed, for every given $\epsilon>0$, when n is large enough,

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

$$P\left(\frac{1}{\tilde{n}_2}\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 \geq \frac{1}{1+\alpha'_\epsilon/2}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\right)$$

$$\leq P(H^c_{n_1})+P(D^c_n)$$

$$+P\left(H_{n_1}\cap D_n\cap\left\{\frac{1}{\tilde{n}_2}\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 \geq \frac{1}{1+\alpha'_\epsilon/2}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\right\}\right)$$

$$\leq 3\epsilon + EP\left(H_{n_1}\cap D_n\cap\left\{\frac{1}{\tilde{n}_2}\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2 \geq \frac{1}{1+\alpha'_\epsilon/2}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\right\}\middle| Z^1\right)$$

$$\leq 3\epsilon + E\exp\left(-\frac{(\alpha'_\epsilon)^2}{8(1+\alpha'_\epsilon)^2(1+\alpha'_\epsilon/2)^2}\right.$$

$$\left.\times\frac{\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^4_{2,x_0;r_n}}{\|f-\hat{f}_{1,n_1}\|^4_{4,x_0;r_n}+\dfrac{\alpha'_\epsilon(A_{n_1,\epsilon})^2}{3(1+\alpha'_\epsilon)(1+\alpha'_\epsilon/2)}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}}\right)$$

$$\triangleq 3\epsilon+\Delta(\epsilon,n),$$

where the expectation in the upper bound of the last inequality (i.e., $\Delta(\epsilon,n)$) converges to zero because of the convergence in probability to zero of the random variables of the exponential expression and their uniform integrability (because they are bounded above by 1), provided that (A.4) and (A.5) hold. The assertion of (A.6) then follows.

For the other estimator, similarly, for $0<\tilde{\beta}_n<1$, we have

$$P_{Z^1}\left(\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2 \leq (1-\tilde{\beta}_n)\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\right)$$

$$\leq \exp\left(-\frac{1}{2}\frac{\tilde{n}_2\tilde{\beta}_n^2\|f-\hat{f}_{2,n_1}\|^4_{2,x_0;r_n}}{\|f-\hat{f}_{2,n_1}\|^4_{4,x_0;r_n}+\dfrac{2(A_{n_1,\epsilon})^2\tilde{\beta}_n}{3}\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}}\right).$$

If we have

$$\frac{\tilde{n}_2\tilde{\beta}_n^2\|f-\hat{f}_{2,n_1}\|^4_{2,x_0;r_n}}{\|f-\hat{f}_{2,n_1}\|^4_{4,x_0;r_n}}\to\infty \quad \text{in probability,} \tag{A.7}$$

$$\frac{\tilde{n}_2\tilde{\beta}_n\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}}{(A_{n_1,\epsilon})^2}\to\infty \quad \text{in probability,} \tag{A.8}$$

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

then following a similar argument used for $\hat{f}_{1,n_1}$, we have

$$P\left(\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2 \le (1-\tilde{\beta}_n)\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\right)\to 0. \tag{A.9}$$

From this, if $\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}\to\infty$ in probability and $\tilde{\beta}_n$ is bounded away from 1, then (A.3) holds. If in addition, we can choose $\tilde{\beta}_n\to 0$, then for each given ϵ, we have $(1-\tilde{\beta}_n)\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n} > (1+\alpha_\epsilon)/(1+\alpha'_\epsilon/2)\|f-\hat{f}_{2,n_1}\|^2_{2,x_0;r_n}$ for some small $\alpha_\epsilon>0$ when n_1 is large enough. Now for every $\tilde{\epsilon}>0$, we can find $\epsilon>0$ such that $3\epsilon\le\tilde{\epsilon}/3$ and there exists an integer n_0 such that when $n\ge n_0$ the probability in (A.9) is upper bounded by $\tilde{\epsilon}/3$ and $\Delta(\epsilon,n)\le\tilde{\epsilon}/3$. Consequently when $n\ge n_0$,

$$P\left(\frac{\sum_{i\in I}(f(X_i)-\hat{f}_{2,n_1}(X_i))^2}{\sum_{i\in I}(f(X_i)-\hat{f}_{1,n_1}(X_i))^2}\ge 1+\alpha_\epsilon\right)\ge 1-\tilde{\epsilon}.$$

Recall that we needed the conditions (A.4), (A.5), (A.7), and (A.8) for (A.1) to hold. Under Condition 3, $\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^4_{2,x_0;r_n}/\|f-\hat{f}_{1,n_1}\|^4_{4,x_0;r_n}$ is lower bounded in order in probability by $\tilde{n}_2\|f-\hat{f}_{2,n_1}\|^4_{2,x_0;r_n}/M^4_{n_1}\|f-\hat{f}_{1,n_1}\|^4_{2,x_0;r_n}$. From all that precedes, because $\hat{f}_{1,n_1}$ and $\hat{f}_{2,n_1}$ converge exactly at rates p_n and q_n, respectively, under the L_2 loss, we know that for the conclusion of Theorem 1 to hold, it suffices to have the following requirements: for every $\epsilon>0$, for some $\tilde{\beta}_n\to 0$,

$$\tilde{n}_2\tilde{\beta}_n^2 M_{n_1}^{-4}\to\infty,$$

$$\tilde{n}_2(q_{n_1}/p_{n_1})^4 M_{n_1}^{-4}\to\infty,$$

$$\tilde{n}_2\tilde{\beta}_n q_{n_1}^2(A_{n_1,\epsilon})^{-2}\to\infty,$$

$$\tilde{n}_2 q_{n_1}^2(A_{n_1,\epsilon})^{-2}\to\infty,$$

$$\tilde{n}_2 q_{n_1}^2\to\infty.$$

Note that $\tilde{n}_2$ is a random variable that has a binomial distribution with n_2 trials and success probability $P(X_i\in B(x_0;r_n))$. Under the assumption that the density of X is bounded away from zero in a neighborhood of x_0, $P(X_i\in B(x_0;r_n))$ is of order r_n^d. We need to lower bound $\tilde{n}_2$ in probability. For $1\le j\le n_2$, let D_j be independent Bernoulli random variables with success probability β. Then applying Bernstein's inequality (see, e.g., Pollard, 1984, p. 193), we have

$$P\left(\sum_{j=1}^{n_2} D_j\le n_2\beta/2\right)\le\exp\left(-\frac{3n_2\beta}{28}\right). \tag{A.10}$$

For our setting, β is of order r_n^d. Thus $\tilde{n}_2$ is at least of order $n_2 r_n^d$ in probability if $n_2 r_n^d\to\infty$. Consequently, the earlier requirements on the splitting ratio become that for every $\epsilon>0$, for some $\tilde{\beta}_n\to 0$,

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

$$n_2 r_n^d \tilde{\beta}_n^2 M_{n_1}^{-4} \to \infty,$$

$$n_2 r_n^d (q_{n_1}/p_{n_1})^4 M_{n_1}^{-4} \to \infty,$$

$$n_2 r_n^d \tilde{\beta}_n q_{n_1}^2 (A_{n_1,\epsilon})^{-2} \to \infty,$$

$$n_2 r_n^d q_{n_1}^2 (A_{n_1,\epsilon})^{-2} \to \infty,$$

$$n_2 r_n^d q_{n_1}^2 \to \infty.$$

Under Condition 1, for every $\epsilon > 0$, there exists a constant $B_\epsilon > 0$ such that $P(\|f - \hat{f}_{j,n_1}\|_\infty \geq B_\epsilon A_{n_1}) \leq \epsilon$ when n_1 is large enough. That is, for a given $\epsilon > 0$, we can take $A_{n_1,\epsilon} = O(A_{n_1})$. Therefore if we have $n_2 r_n^d M_{n_1}^{-4} \to \infty$ and $n_2 r_n^d q_{n_1}^2/(1 + A_{n_1}) \to \infty$, then we can find $\tilde{\beta}_n \to 0$ such that the preceding five displayed requirements are all satisfied.

Let π denote a permutation of the observations and let Π denote a set of such permutations. Let $LCV_\pi(\hat{f}_{j,n_1})$ be the L-CV criterion value under the permutation π. If $LCV_\pi(\hat{f}_{1,n_1}) \leq LCV_\pi(\hat{f}_{2,n_1})$, then let $\tau_\pi = 1$ and otherwise let $\tau_\pi = 0$. Let W denote the values of $(X_1, Y_1), \ldots, (X_n, Y_n)$ (ignoring the orders). Under the i.i.d. assumption on the observations, obviously, conditional on W, every ordering of these values has exactly the same probability, and thus

$$P(LCV(\hat{f}_{1,n_1}) \leq LCV(\hat{f}_{2,n_1})) = EP(LCV(\hat{f}_{1,n_1}) \leq LCV(\hat{f}_{2,n_1})|W)$$

$$= E\left(\frac{\sum_{\pi\in\Pi} \tau_\pi}{|\Pi|}\right).$$

From the earlier analysis, we know that $P(LCV(\hat{f}_{1,n_1}) \leq LCV(\hat{f}_{2,n_1})) \to 1$. Thus $E(\sum_{\pi\in\Pi} \tau_\pi/|\Pi|) \to 1$. Because $\sum_{\pi\in\Pi} \tau_\pi/|\Pi|$ is between 0 and 1, for its expectation to converge to 1, we must have $\sum_{\pi\in\Pi} \tau_\pi/|\Pi| \to 1$ in probability. This completes the proof of Theorem 1. ■

Proof of Theorem 2. Let $\tilde{A}_n^*$ be the minimizer of $E\|f(;A) - f\|^2$ over $A \in \mathcal{A}_{n,0}$. By the results of Wegkamp (2003, Sect. 2), in particular Theorem 2.1, we have

$$E\|\hat{f}(;\hat{A}_n) - f\|^2 \leq 2\left(E\|\hat{f}(;\tilde{A}_n^*) - f\|^2 + \frac{1}{n_2}\right)$$

$$+ \frac{6\bar{C}^2 \log(4|\mathcal{A}_{n,0}|e^{-1/(6\bar{C}^2)}) + 16\sigma^2(2 + \log(|\mathcal{A}_{n,0}|e^{-1/(16\sigma^2)}))}{n}$$

$$\leq 2E\|\hat{f}(;\tilde{A}_n^*) - f\|^2 + \frac{B_1}{n_2} + \frac{B_2 H_n(\epsilon_n)}{n_2},$$

where B_1 and B_2 are constants that depend only on $\bar{C}$ and σ^2.

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press

Note that

$$\begin{aligned}
&\int (\hat{f}(x;A) - \hat{f}(x;B))^2 \, d\mu \\
&\quad \le 2 \int ((\hat{f}_1(x) - f(x))(I_{\{x\in A\}} - I_{\{x\in B\}}))^2 \, d\mu \\
&\qquad + 2 \int ((\hat{f}_2(x) - f(x))(I_{\{x\in A\}} - I_{\{x\in B\}}))^2 \, d\mu \\
&\quad = 2 \int_{B\triangle A} (\hat{f}_1(x) - f(x))^2 \, d\mu + \int_{B^c\triangle A^c} (\hat{f}_2(x) - f(x))^2 \, d\mu \\
&\quad = 2 \int_{B\triangle A} ((\hat{f}_1(x) - f(x))^2 \, d\mu + (\hat{f}_2(x) - f(x))^2) \, d\mu \\
&\quad \le 4\bar{C}^2 \mu(B \triangle A).
\end{aligned}$$

Together with $\mathcal{A}_{n,0}$ being an ϵ_n-net under d, we have $E\|\hat{f}(;\tilde{A}_n^*) - f\|^2 \le 2E\|\hat{f}(;A_n^*) - f\|^2 + 8\bar{C}^2\epsilon_n$. The conclusions then follow. This completes the proof of Theorem 2. ■

https://doi.org/10.1017/S0266466608080195 Published online by Cambridge University Press