The Annals of Statistics
1991, Vol. 19, No. 4, 1869–1882

VARIABLE SELECTION IN NONPARAMETRIC REGRESSION WITH CONTINUOUS COVARIATES[1]

BY PING ZHANG

University of Pennsylvania

In a nonparametric regression setup where the covariates are continuous, the problem of estimating the number of covariates will be discussed in this paper. The kernel method is used to estimate the regression function and the selection criterion is based on minimizing the cross-validation estimate of the mean squared prediction error. We consider choosing both the bandwidth and the number of covariates based on the data. Unlike the case of linear regression, it turns out that the selection is consistent and efficient even when the true model has only a finite number of covariates. In addition, we also observe the curse of dimensionality at work.

1. Introduction. In this paper, we consider the estimation of the number of covariates in a nonparametric regression model. Although much of the theoretical aspect of variable selection methods is yet to be explored, the concept has been well accepted by most statisticians and some criteria are being routinely used in practice. Current literature, however, has been focusing on linear regression and as far as we know, there has been no published study on model selection methods under a nonparametric setup.

Throughout this paper, it is assumed that all the covariates are continuous. Under a linear regression setup, it has been pointed out be many authors that when the true model has finitely many covariates, criteria such as AIC tend to choose overfitted models with positive probability. Although concepts like AIC and maximum likelihood do not carry over to the nonparametric situation in a straightforward fashion, it does make sense to talk about prediction error and cross-validation in the general framework. The equivalence of the AIC and the cross-validation (CV) criterion was observed by Stone [14]. One naturally expects the behavior of the CV criterion in nonparametric regression to be similar to that of AIC in linear regression. Namely, it does a good job only when the true model has infinitely many covariates (see Shibata [13]). While this has been shown to be true when the covariates are categorical variables, we will show momentarily that in the continuous case, it is not true. To understand the difference, we have to realize that it is a misconception, often adopted in the literature, to mix up the *number of covariates* and the *model dimension*. The latter should be interpreted as the number of parameters needed to describe the model, and in the earlier studies (linear regression or nonparametric regression with categorical covariates), these two quantities

Received February 1990, revised December 1990.
[1]Prepared with partial support of NSF Grant DMS-87-01426.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62J99.
Key words and phrases. Cross-validation, kernel estimate, model selection.

happen to coincide. That is, a regression model with a finite number of covariates can indeed be described by a finite number of parameters. However, this is no longer the case for the model considered in this paper. Even though there is only one covariate, the model dimension is always infinity, or goes to infinity depending on how one looks at the problem (see Buja, Hastie and Tibshirani [2]). In conclusion, it seems that for consistency in variable selection, what matters is the model dimension rather than the number of covariates. If the true model is finite dimensional, cross-validation sometimes chooses overfitted models. Otherwise, it can be consistent. It is also illuminating to notice that a similar classification also exists in the asymptotic distribution of the mean squared error. For models with finite model dimension, it usually tends to a mixture of χ^2's. Otherwise, it tends to a normal distribution (see Hall [7]).

Formally speaking, assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are iid random vectors with $(X_1, Y_1) \in R^\infty \times R^1$. When a new observation X arrives, we want to predict the corresponding Y. The best prediction under squared error loss would be $m(x) = \mathbf{E}(Y|X = x)$ and we are going to estimate this function nonparametrically (using a kernel estimate in particular). Given the observations, the object now is to find a subset of covariates which best predict the response. Let $X = (X(1), X(2), \ldots) \in R^\infty$. We assume, as usual, that $X(i)$, $i = 1, 2, \ldots,$ are preordered according to their importance so that it will be only necessary to select the number of covariates rather than searching among all possible subsets. Generally speaking, let the d-dimensional prediction function be

$$m_d(x) = \mathbf{E}(Y|X(1) = x(1), \ldots, X(d) = x(d)).$$

We use the so-called Watson–Nadaraya estimate

$$\hat{m}_d(x) = \frac{1}{nh^d \hat{f}(x)} \sum Y_i K\left(\frac{X_i - x}{h}\right),$$

where

$$\hat{f}(x) = \frac{1}{nh^d} \sum K\left(\frac{X_i - x}{h}\right) + \frac{1}{nh^d}.$$

Some notation and conventions are warranted before further arguments can take place. Whenever x appears as an argument of a function, it is regarded as if it had a subscript d to indicate that $x \in R^d$. This rule is followed throughout this paper. Let $S_d \subset R^d$ be a compact set and w_d: $R^d \to R$ be a weight function supported on S_d and $w_d(x) \leq C$ for some constant C not depending on d. Let $f_d(x)$ be the joint density function of the first d coordinates of X. We define

$$\Delta_n(d, h) = \mathbf{E}(Y - \hat{m}_d(X))^2 w_d(X),$$

$$M_n(d, h) = \mathbf{E}\left[(Y - \hat{m}_d(X))^2 w_d(X) \middle| (X_1, Y_1), \ldots, (X_n, Y_n)\right]$$

and their cross-validation approximation

$$\hat{\Delta}_n(d,h) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \hat{m}_{d(-i)}(X_i)\big)^2 w_d(X_i),$$

where $\hat{m}_{d(-i)}$ is the kernel estimate of the regression function using all but the ith observation.

The rest of the paper is organized in the following way. Section 2 shows the asymptotic optimally of the cross-validation criterion. In particular, unlike the parametric case, the CV criterion is consistent. In Section 3, we show that our results hold generally under mild restrictions. Topics such as selection range and boundary effects are discussed. Section 4 discusses the potentials and limitations of the CV criterion. And finally, lengthy proofs are deferred to the Appendix.

2. Optimality of cross-validation. In the following, $m''_d(x)$ stands for a $d \times d$ matrix with the (i,j)'s element $\partial^2 m_d(x)/\partial x(i)\partial x(j)$. And $m'_d(x)f'(x)$ denotes the $d \times d$ matrix with the (i,j)'s element $\partial m_d(x)/\partial x(i) \times \partial f(x)/\partial x(j)$. Also, define

$$B_d(x) = \int u^\tau\big(m'_d(x)\,f'(x)/f(x) + m''_d(x)/2\big)uK(u)\,du.$$

It should be stressed that throughout this paper, we use C to denote a generic constant not depending on n, d and h. The C's appearing at different places need not be the same. Here below, we use $\|\cdot\|$ (or $\|\cdot\|_d$ to indicate the dimension of its argument) for the Euclidean norm and when applied to a matrix, it would be the corresponding induced norm. The notation $\langle\cdot\rangle$ represents the usual inner products.

We make the following assumptions.

(A) For any integer l, $\mathbf{E}Y^{2l} < \infty$ and $\mathbf{E}[Y^{2l}|X = x] \le C^{2l}$ for $x \in S_d^\varepsilon$.

(B) $f(x)$ is twice differentiable, and there exist constants $A_d > 1$ and $\alpha < 1/2$ such that for all $x, y \in S_d^\varepsilon$, the following hold:

$$f(x) > 0, \qquad \big\|f'(x)/f(x)\big\| \le A_d, \qquad \big\|f''(x)/f(x)\big\| \le A_d,$$

$$f(y)/f(x) = 1 + \langle f'(x)/f(x), y - x\rangle + (y-x)^\tau\big(f''(x)/2f(x)\big)(y-x) + O\big(A_d\|y-x\|^{2+\alpha}\big).$$

(C) $m_d(x)$ is twice differentiable, and there exist constants $A_d > 1$ and $\alpha < 1/2$ such that for all $x, y \in S_d^\varepsilon$, the following hold:

$$\|m'_d(x)\| \le A_d, \qquad \|m''_d(x)\| \le A_d, \qquad \|m''_d(x) - m''_d(y)\| \le A_d\|x-y\|^\alpha.$$

(D) There exist positive constants a, b such that $a^d \le f(x) \le b^d$ for $x \in S_d^\varepsilon$.

In the above, $S_d^\varepsilon = \{y \in R^d\colon \|y - x\| \le \varepsilon;\ x \in S_d\}$, and when there is no confusion, the subscript d is often dropped for convenience of presentation.

For the kernel function K, we assume:

(E) $K(u)$ is a continuously differentiable density function supported on $\|u\|_d < C$, where $\|\cdot\|_d$ is the Euclidean norm in R^d.

Condition (E) implies that $K(u) \le C$ and $|\langle u, K'(u)\rangle| \le C$. The restrictions on kernel function are for convenience in carrying out calculations rather than necessary. It is, however, well known, that the choice of kernel function is not as important as the choice of bandwidth (see Rosenblatt [12]).

Defining

$$V(d,h) = \frac{\int K^2(u)\,du}{nh^d}\mathbf{E}\left[\frac{(Y - m_d(X))^2 w_d(X)}{f(X)}\right] + \mathbf{E}\left[w_d(X)B_d^2(X)\right]h^4,$$

let $H(\delta_n) = \{h > 0\colon A_d h^\alpha \le \delta_n\}$ and $\mathbf{L}_p(C) = \{X\colon X \in \mathbf{L}_p;\ \mathbf{E}|X|^p \le C\}$, where C does not depend on p. Then we have the following lemma.

LEMMA 1. *Suppose that $h \in H(\delta_n)$ for some $\delta_n \to 0$, and assumptions* (A) *to* (E) *hold. Then for any integer l, there exists a random variable $\hat{\theta} \in \mathbf{L}_{2l}(C)$ such that*

$$\begin{aligned}\Delta_n(d,h) &= \mathbf{E}(Y - m_d(X))^2 w_d(X) + V(d,h)\\ &\quad + C^d O\left(A_d(nh^d)^{-3/2} + A_d^2 h^\alpha/nh^d + A_d^4 h^{4+\alpha}\right),\end{aligned} \tag{2.1}$$

$$\begin{aligned}M_n(d,h) &= \mathbf{E}(Y - m_d(X))^2 w_d(X) + V(d,h)\\ &\quad + \hat{\theta}(n,d,h)C^d\left(A_d(nh^d)^{-3/2} + A_d^2 h^\alpha/nh^d + A_d^4 h^{4+\alpha}\right)\end{aligned} \tag{2.2}$$

and

$$\begin{aligned}\hat{\Delta}_n(d,h) &= \frac{1}{n}\sum \varepsilon_{id}^2 w_d(X_i) + V(d,h)\\ &\quad + \hat{\theta}(n,d,h)C^d\left(A_d^2(nh^d)^{-3/2} + A_d^2 h^\alpha/nh^d + A_d^4 h^{4+\alpha}\right).\end{aligned} \tag{2.3}$$

PROOF. See Zhang [15]. □

It is assumed throughout this section that $Y_i = m(X_i) + \varepsilon_i$, $i = 1, 2, \ldots, n$, where ε_i's are iid with mean 0 and variance σ^2, and ε_i is independent of X_i. Let $m(x)$ be some smooth function satisfying assumption (C). By saying that the true model has d_0 (not necessarily finite) covariates, we mean that the value of $m(x)$, $x \in R^\infty$, depends only on the first d_0 coordinates of x. We further introduce the following assumptions:

(F) $\int K^2(u)\,du = \int K^2(u_1, \ldots, u_d)\,du_1 \cdots du_d \ge C^d$ for some constant C.

(G) $\mathbf{E}w_d(X)$ does not depend on d; $\lim_{d\to\infty} w_d(X) = w_\infty(X)$ exists and $\int w_\infty(x)\,dx_1\,dx_2 \cdots \ne 0$.

THE SELECTION RULE. Let $\Omega_n = \{(d, h)$: d is an integer, $d^2/\log(n) \le \delta_n$; $C_1 n^{-1/(d+4)} \le h \le C_2 n^{-1/(d+4)}\}$, where $\delta_n \to 0$. Let $\mathrm{PE}(d, h)$ be any measure of prediction error. The selection procedure then goes as follows. Given d, select an $h = h(d)$ which minimizes $\mathrm{PE}(d, h)$, then select d to minimize $\mathrm{PE}(d, h(d))$. All the selections are done with the constraint $(d, h) \in \Omega_n$. In the next section, we show that this constraint is very reasonable in the sense that the *global* minimizer actually falls in this region.

THEOREM 1. *Suppose that assumptions* (A) *to* (G) *hold and* $A_d \le C^d$ *for some constant* C. *Suppose the true model has* d_0 $(\le \infty)$ *covariates, and* $d(n)$ *is the minimizer of* $\Delta_d(d, h(d))$. *Then* $\lim_{n\to\infty} d(n) = d_0$.

PROOF. See the Appendix. □

LEMMA 2. *Let* $\hat{\theta}(n, d, h) \in \mathbf{L}_{2l}(C)$ *be the random variables in Lemma* 1. *Under the assumptions of Theorem* 1, *it can be shown that for any* $t > 0$,

$$\sup_{(d,h)\in\Omega_n} \left|\hat{\theta}(n, d, h)\right| C^d (nh^d)^{-t} \Rightarrow_{\mathbf{P}} 0.$$

PROOF. See Zhang [15]. □

Here the notation $\Rightarrow_{\mathbf{P}}$ represents convergence in probability. Importantly, this lemma implies that the remainder terms in various expansions in Lemma 1 are *uniformly* negligible.

THEOREM 2. *Let the assumptions of Theorem* 1 *hold. Suppose* $\bar{d}(n)$ *is the minimizer of* $M_n(d, h(d))$. *Then* $\bar{d}(n) \Rightarrow_{\mathbf{P}} d_0$.

PROOF. First, from the same argument leading to Theorem 1, one can show that underfitting is impossible. That is, $\lim \mathbf{P}[\bar{d}(n) < d_0] = 0$. To eliminate the possibility of overfitting, from Lemma 1, when $d \ge d_0$, we can write

$$M_n(d, h) = \sigma^2 \mathbf{E} w_\infty(X) + V(d, h)[1 + r_n],$$

where from Lemma 2, by appropriately choosing t, it is easy to check that

$$\sup_{(d,h)\in\Omega_n} r_n \Rightarrow_{\mathbf{P}} 0.$$

In other words, when $d \ge d_0$,

$$M_n(d, h) = \sigma^2 \mathbf{E} w_\infty(X) + V(d, h)\left[1 + o_p(1)\right]. \tag{2.4}$$

The same argument leading to Theorem 1 would then eliminate the possibility of $\limsup \bar{d}(n)$ being greater than d_0 with positive probability. Consequently, $\bar{d}(n) \Rightarrow_{\mathbf{P}} d_0$. □

To deal with $\hat{\Delta}_n(d, h)$, note from Lemma 1 that the first term in the expansion yields an error of the order $O_p(n^{-1/2})$. Since the optimal bandwidth

h is of order $n^{-1/(4+d)}$, $n^{-1/2}$ is asymptotically negligible in comparison with the term $V(d, h)$ (which is of the order $n^{-4/(d+4)}$) only when $d > 4$. In other words, our criterion would fail when $d < 4$. A simple remedy is as follows. Assume

(H) $\quad w_d(x) = w_4(x) = w_4(x(1), x(2), x(3), x(4)) \quad$ for $d \leq 4$.

For $d \leq 4$, instead of letting $w_d(x)$ be a function of the first d coordinates, we allow it to depend on the first four coordinates. By doing so, we make the leading term in the expansion of $\hat{\Delta}_n(d, h)$ independent of d for $d \leq 4$. Obviously, this is only a technical device and has nothing to do with the problem itself. Let Ω_n be as before. We have the following theorem.

THEOREM 3. *In addition to the assumptions of Theorem* 2, *assume* (H) *holds. Suppose* $\hat{d}(n)$ *is the minimizer of* $\hat{\Delta}_n(d, h(d))$. *Then* $\hat{d}(n) \Rightarrow_{\mathbf{P}} d_0$.

PROOF. We follow the same line of arguments as before. It is easy to show that underfitting is asymptotically impossible. From the argument in previous theorems, if we can show for $d \geq d_0$ that

$$\hat{\Delta}_n(d, h) = \frac{1}{n} \sum \varepsilon_i^2 w_4(X_i) + V(d, h)\big[1 + o_p(1)\big], \tag{2.5}$$

then overfitting is also impossible. To prove (2.5), we have from Lemma 1 and the argument in the previous theorems that

$$\hat{\Delta}_n(d, h) = \frac{1}{n} \sum \varepsilon_{id}^2 w_d(X_i) + V(d, h)\big[1 + o_p(1)\big].$$

Next, by assumptions (F), (G) and (H), we have for $d \geq d_0$, that

$$\begin{aligned} \frac{1}{n} \sum \varepsilon_{id}^2 w_d(X_i) &= \frac{1}{n} \sum \varepsilon_i^2 w_d(X_i) \\ &= \frac{1}{n} \sum \varepsilon_i^2 w_4(X_i) + O_p(n^{-1/2}) \\ &= \frac{1}{n} \sum \varepsilon_i^2 w_4(X_i) + o_p(V(d, h)). \end{aligned}$$

Hence follow (2.5) and the conclusion. □

Acknowledging the previous results, we can easily state the efficiency (or the optimality, as in Breiman and Freedman [1]) of the cross-validation criterion under the same assumptions. This is summarized as follows.

THEOREM 4. *Suppose that the assumptions in Theorem* 3 *hold. Let* $(d(n), h(d(n)))$, $(\hat{d}(n), h(\hat{d}(n)))$ *and* $(\bar{d}(n), h(\bar{d}(n)))$ *be the minimizers of*

$\Delta_n(d, h)$, $\hat{\Delta}_n(d, h)$ *and* $M_n(d, h)$, *respectively. Then*

$$\text{(i)} \qquad \frac{\left[M_n\left(\bar{d}(n), h\left(\bar{d}(n)\right)\right) - \sigma^2 \mathbf{E} w_\infty(X)\right]}{\left[\Delta_n(d(n), h(d(n))) - \sigma^2 \mathbf{E} w_\infty(X)\right]} \Rightarrow_{\mathbf{P}} 1$$

and

$$\text{(ii)} \qquad \frac{\left[M_n\left(\hat{d}(n), h\left(\hat{d}(n)\right)\right) - \sigma^2 \mathbf{E} w_\infty(X)\right]}{\left[\Delta_n(d(n), h(d(n))) - \sigma^2 \mathbf{E} w_\infty(X)\right]} \Rightarrow_{\mathbf{P}} 1.$$

PROOF. (i) Let $f(d, h) = \Delta_n(d, h) - \sigma^2 \mathbf{E} w_\infty(X)$, $g(d, h) = M_n(d, h) - \sigma^2 \mathbf{E} w_\infty(X)$. Then (A.2) and (2.4) imply that uniformly for all $(d, h) \in \Omega_n$ and $d \le d_0$,

$$f(d, h) = V(d, h)[1 + o(1)] \tag{2.6}$$

and

$$g(d, h) = V(d, h)\left[1 + o_p(1)\right]. \tag{2.7}$$

Now minimizing $\Delta_n(d, h)$ and $M_n(d, h)$ is the same as minimizing $f(d, h)$ and $g(d, h)$. Notice the consistency results proved previously, we get asymptotically that

$$g\left(\bar{d}(n), h\left(\bar{d}(n)\right)\right) \le g(d(n), h(d(n))) = f(d(n), h(d(n)))\left[1 + o_p(1)\right]$$

and

$$\begin{aligned} g\left(\bar{d}(n), h\left(\bar{d}(n)\right)\right) &= f\left(\bar{d}(n), h\left(\bar{d}(n)\right)\right)\left[1 + o_p(1)\right] \\ &\le f(d(n), h(d(n)))\left[1 + o_p(1)\right]. \end{aligned}$$

The above two equations imply (i). Part (ii) can be derived through a very similar argument by letting $\hat{f}(d, h) = \hat{\Delta}_n(d, h) - n^{-1}\Sigma \varepsilon_i w_4(X_i)$. □

3. Some further results. For practical reasons, efforts have been made in bandwidth selection literature to allow larger selection ranges (see Härdle and Marron [8]). It is therefore arguable that our selection range $(d, h) \in \Omega_n$ might eventually lead to a local minimum. The next result will to some extent ease such concerns by showing that in some sense, the global minimizer actually falls into Ω_n.

THEOREM 5. *In addition to assumptions* (A) *to* (H), *assume that* $\mathbf{E} w_d(X) B_d^2(X) \ge C^d$ *for some* C. *Let* $\Omega_n^* = \{(d, h)$: d *is an integer*, $n^{1/d}h \ge \delta_n^{-1}$; $h^{1/d} \le \delta_n\}$, $\delta_n \to 0$. *Suppose* (d^*, h^*) *is the minimizer of* $\Delta_n(d, h)$ *within* Ω_n^* *according to the selection rule. Then* $(d^*, h^*) \in \Omega_n$ *for some* $\delta_n \to 0$ *and* $C_2 > C_1 > 0$.

PROOF. Clearly, when $(d, h) \in \Omega_n^*$, (2.1) holds. By the assumptions, it is easy to show that for some C,

$$\Delta_n(d, h) \geq \mathbf{E}(Y - m_d(X))^2 w_d(X) + C^d(h^4 + n^{-1}h^{-d}).$$

If (d^*, h^*) is the minimizer, then

$$\Delta_n(d, Cn^{-1/(d+4)}) \geq \Delta_n(d^*, h^*).$$

This implies, first of all, that there exist constants C_1, C_2, such that

$$C_1 n^{-1/(d^*+4)} \leq h^* \leq C_2 n^{-1/(d^*+4)},$$

and second, that $d^{*2}/\log(n) \to 0$ since $h^{*1/d^*} \to 0$. The proof is complete. □

For $(d, h) \in \Omega_n^*$, we have $n^{1/d}h \to \infty$ and $h^{1/d} \to 0$. Given any d, these are the minimal requirements for the consistency of the kernel regression estimate. As a result, we can reasonably regard (d^*, h^*) as the *global* minimizer.

Before the selection procedure starts, for any dimension d, there are two key elements that need to be determined, a kernel function $K(u)$ and a weight function $w_d(x)$. It is generally believed that the choice of kernel is not essential. The weight function, however, ties in closely with the properties of underlying covariate distribution $f(x)$, and therefore cannot be chosen arbitrarily.

One of the main constraints about the weight function is that $\mathbf{E}w_d(X)$ does not depend on d. Let G be the distribution function of $\|X\|$. If the joint density $f(x)$ is known, the task becomes trivial. An obvious candidate would be the indicator function on the set $S_d = \{x: \|X\| \leq G^{-1}(\mu)\}$ for some μ close to 1. The fixed design case (X not random) can be handled in a similar fashion. It is known that for fixed design problems, $m(x)$ cannot be properly estimated when x is near the boundary of the design set (see Eubank [4]). Our S_d should eliminate such points. Generally, we choose S_d to satisfy the following conditions:

(i) $\mathbf{P}(S_d) \geq 1 - \varepsilon$, ε prespecified;
(ii) $a^d \leq f_d(x) \leq b^d$, $x \in S_d^\varepsilon$ for some $a, b > 0$;
(iii) diameter of $S_d = O(C^d)$ for some constant C.

For $f(x)$ unknown, we turn to the data for help. In other words, we construct a random weight function by somehow estimating the above indicator function. Some other types of random weight functions have been considered by Hall [7].

If $\tilde{w}_d(x) = \tilde{w}_d(x; X_1, \ldots, X_n)$ is such a random weight function, we can modify the cross-validation criterion so that

$$\tilde{\Delta}_n(d, h) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \hat{m}_{d(-i)}(X_i)\right)^2 \tilde{w}_{d(-i)}(X_i),$$

where $\tilde{w}_{d(-i)}$ is the corresponding version of $\tilde{w}_d$ with X_i being removed.

Let $\|X\|_{(r)}$ denote the rth-order statistic of $\|X_1\|, \ldots, \|X_n\|$, where $r = [n(1 - \varepsilon)]$ for some prespecified $\varepsilon > 0$. Let $\mu = r/(n + 1)$. We define

$$\tilde{w}_d(x) = I\big(\|x\| \leq \|X\|_{(r)}\big). \tag{3.1}$$

Since $\|X\|_{(r)}$ is the natural estimator of $G^{-1}(\mu)$, it is expected that $\tilde{w}_d(x)$ would approximate $I(\|x\| \leq G^{-1}(\mu))$ well enough so that when substituted, the modified cross-validation criterion still gives optimal selection. We have the following result.

THEOREM 6. *Suppose that* $S_d = \{x: \|x\| \leq G^{-1}(\mu)\}$ *satisfies conditions* (i) *to* (iii) *above and assumptions* (A) *to* (H) *hold with* $w_d(x) = I(x \in S_d)$. *Then the results of Theorems* 3 *and* 4 *still hold when replacing the role of* $\hat{\Delta}_n(d, h)$ *with* $\tilde{\Delta}_n(d, h)$.

PROOF. See the Appendix. □

Theorem 6 provides us with a fully adaptive variable selection procedure which is equivalent to CV, thus asymptotically optimal. To verify conditions (i) to (iii), we give the following examples [note that (i) is trivial for the chosen S_d].

EXAMPLE 1. If the covariates $X(1), X(2), \ldots$ are iid with density $f_0(x)$ which is bounded above and $f_0(x) \geq C_1 \exp(-C_2 x^2)$ for some positive constants C_1 and C_2, then conditions (ii) and (iii) are satisfied. To see this, since $\|X\| = O_p(\sqrt{d})$ from the law of large numbers, we have $G^{-1}(\mu) = O(\sqrt{d})$, hence condition (iii). This further implies condition (ii) since $f_0(x) \geq C_1 \exp(-C_2 x^2)$.

The above claim can be extended to the situation when $f_0(x) \geq C_1 \exp(-C_2|x|^\alpha)$ for some $\alpha > 0$. In this case, we need to use the $\mathbf{L}_\alpha$ norm instead of the $\mathbf{L}_2$ norm in the definition of S_d and everything else follows by exact analogy.

EXAMPLE 2. Suppose that there is a sequence of random variables $\tilde{X}(1), \tilde{X}(2), \ldots$ satisfying conditions (ii) and (iii). If our covariates are such that for any d, $(X(1), \ldots, X(d)) = (\tilde{X}(1), \ldots, \tilde{X}(d))\Gamma_d$, where Γ_d is a $d \times d$ matrix and the eigenvalues of $\Gamma_d'\Gamma_d$ are bounded below and above, then the sequence $X(1), X(2), \ldots$ also satisfies conditions (ii) and (iii). The proof is obvious when writing out the density of x in terms of the density of $\tilde{X}$ and noticing that $C_1\|\tilde{X}\| \leq \|X\| \leq C_2\|\tilde{X}\|$.

Roughly speaking, the above two examples indicate that when the covariates are not heavily interdependent and the tail of the joint density $f_d(x)$ is not too light, the cross-validation criterion is always going to work. Let us conclude this section with a specific example.

EXAMPLE 3. If $X(1), X(2), \ldots$ consists of a stationary Gaussian process and the corresponding spectral density $h(\omega)$ is bounded below and above, then conditions (ii) and (iii) are satisfied. In particular, this includes the usual ARMA stationary processes. In light of the previous two examples, we only need to show that the eigenvalues of the autocovariance matrix are bounded below and above. This can be seen by the following argument (see Priestly [9]). Using the spectral representation $X(t) = \int_{-\pi}^{\pi} \exp(t\omega)\, dZ(\omega)$, we can write the autocovariance matrix as $A = \int_{-\pi}^{\pi} B(\omega)h(\omega)\, d\omega$, where $B(\omega)$ is the nonnegative definite hermitian matrix with the (j, k)'s element $\exp(i(j - k)\omega)$. If $h(\omega) \geq \delta > 0$, then $A \geq \delta \int_{-\pi}^{\pi} B(\omega)\, d\omega = 2\pi\delta I_d$, where I_d is the identity matrix of order d. This implies that the smallest eigenvalue of A is no less than $2\pi\delta$. The same argument leads to an upper bound.

4. Discussion. The kernel method is used mainly for analytical convenience. Fan [5] has pointed out that theoretically, the Watson–Nadaraya estimator can be rather deficient compared with other kernel methods. The main trouble remains, however, that general kernel methods, although of substantial theoretical merits, all require the choice of a bandwidth, which makes the task of variable selection unnecessarily fussy. Less sophisticated methods such as GCV or the nearest-neighbor methods might be better off in this respect simply because more efficient algorithms can be developed (see Eubank [4] and Cleveland and Devlin [3]). We wish to report work in this area elsewhere.

We have assumed that the covariates are preordered according to their importance. Notice that the CV criterion also applies to the general all subsets selection problem. We conjecture that all the previous results can be generalized to this case. However, all subsets selection is simply too cumbersome to be used in practice. When an ordering of covariates is not given a priori, people have tried various data-driven methods to order the covariates. The issue goes beyond the scope of this paper and to our knowledge, no satisfactory resolution has been found.

Most of the work on nonparametric regression deals with a single covariate, and the parallel multiple regression case is often regarded as a straightforward generalization. Not until recently did people start to realize the practical difficulty in dealing with high dimensional data. The problem is termed by some the *curse of dimensionality*. See Friedman and Stuetzle [6]. We encountered the same problem from the perspective of variable selection. Technically, we observed this by treating the number of covariates as an unknown variable and all the expansions must be adjusted accordingly. This results in assumptions like $d^2/\log(n) \leq \delta_n$, which, according to Theorem 5, is almost necessary. Thus, in principle, even with very large sample size, we are only able to deal with models with very few covariates. On the one hand, this has handicapped the procedure from being potentially a useful practical method in data analysis. On the other hand, it also shows that the curse of dimensionality is inherent and any attempt to tackle multidimensional data the same way as dealing with univariate case would be unrealistic. For recent developments in

more powerful nonparametric regression techniques, see Buja, Hastie and Tibshirani [2].

APPENDIX

PROOF OF THEOREM 1. Since $A_d \le C^d$, it is seen that

$$A_d h^\alpha \le C^d n^{-\alpha/(d+4)} = \left[Cn^{-\alpha/d(d+4)}\right]^d.$$

Note that the quantity in the brackets tends to 0 uniformly in Ω_n, so $\sup_{(d,h)\in\Omega_n} A_d h^\alpha \to 0$. Hence Lemma 1 holds uniformly for $(d, h) \in \Omega_n$. This in turn implies that

$$\text{(A.1)} \quad \limsup_{n\to\infty} \Delta_n(d(n), h(d(n))) \le \mathbf{E}(Y - m(X))^2 w_\infty(X) = \sigma^2 \mathbf{E} w_\infty(X).$$

Suppose the claim were not true, say, without losing generality, that $\lim_{n\to\infty} d(n) \le d' < d_0$ for some d'. Then from Lemma 1, it is easy to see that

$$\begin{aligned}\Delta_n(d(n), h(d(n))) &= \mathbf{E}\big(Y - m_{d(n)}(X)\big)^2 w_{d(n)}(X) + o(1)\\ &\ge \sigma^2 \mathbf{E} w_\infty(X) + \mathbf{E}\big(m_{d'}(X) - m_{d_0}(X)\big)^2 w_{d(n)}(X)\\ &\quad + o(1).\end{aligned}$$

Hence

$$\lim_{n\to\infty} \Delta_n(d(n), h(d(n))) > \sigma^2 \mathbf{E} w_\infty(X),$$

which contradicts (A.1). Consequently, we must have $\liminf_{n\to\infty} d(n) \ge d_0$.

To eliminate the possibility of overfit, suppose on the contrary that $d(n) \ge d' > d_0$ for n large enough (this only happens when $d_0 < \infty$). By Lemma 1, for all $(d, h) \in \Omega_n$, we can write

$$\Delta_n(d, h) = \sigma^2 \mathbf{E} w_\infty(X) + \left[\sigma_d^2 + V(d, h)\right][1 + r_n],$$

where $\sigma_d^2 = \mathbf{E}(m(X) - m_d(X))^2 w_d(X)$. It is also easy to check that

$$\sup_{(d,h)\in\Omega_n} |r_n| \to 0.$$

Thus uniformly for $d \ge d_0$, we have

$$\text{(A.2)} \qquad \Delta_n(d, h) = \sigma^2 \mathbf{E} w_\infty(X) + V(d, h)[1 + o(1)].$$

Under assumptions (F) and (G), (2.2) combined with Lemma 1 would show that

$$\lim_{n\to\infty} \frac{\left[\Delta_n(d(n), h(d(n))) - \sigma^2 \mathbf{E} w_\infty(X)\right]}{\left[\Delta_n(d_0, h(d_0)) - \sigma^2 \mathbf{E} w_\infty(X)\right]} = \infty.$$

The above could not be true since $d(n)$ is the minimizer of $\Delta_n(d, h(d)) - \sigma^2 \mathbf{E} w_\infty(X)$ (the ratio should be less than 1). The contradiction proves that asymptotically, in the range $d \ge d_0$, $\Delta_n(d, h(d))$ is minimized at $d = d_0$. An

earlier argument says that underfitting is also impossible, so it has to be that $\lim_{n\to\infty} d(n) = d_0$. □

PROOF OF THEOREM 6. It is enough to show that $\tilde{\Delta}_n(d,h)$ and $\hat{\Delta}_n(d,h)$ have the same expansion when $(d,h) \in \Omega_n$, namely,

$$\tilde{\Delta}_n(d,h) = \frac{1}{n}\sum \varepsilon_{id}^2 w_d(X_i) + V(d,h)\big(1+o_p(1)\big), \tag{A.3}$$

where $o_p(1)$ is uniform over $(d,h) \in \Omega_n$.

Clearly, we can write

$$\tilde{w}_d(x) = w_d(x) + \tilde{w}_d^{(1)}(x) + \tilde{w}_d^{(2)}(x),$$

where

$$\tilde{w}_d^{(1)}(x) = \big(\tilde{w}_d(x) - w_d(x)\big) I\big(\|x\|_{(r)} \le G^{-1}(\mu+\varepsilon/2)\big)$$

and

$$\tilde{w}_d^{(2)}(x) = \big(\tilde{w}_d(x) - w_d(x)\big) I\big(\|x\|_{(r)} > G^{-1}(\mu+\varepsilon/2)\big).$$

Correspondingly,

$$\tilde{\Delta}_n(d,h) = \hat{\Delta}_n(d,h) + \Delta_1 + \Delta_2. \tag{A.4}$$

We have shown previously that

$$\hat{\Delta}_n(d,h) = \frac{1}{n}\sum \varepsilon_{id}^2 w_d(X_i) + V(d,h)\big(1+o_p(1)\big). \tag{A.5}$$

Next, since

$$\begin{aligned}
\Delta_1 &= \frac{1}{n}\sum \big(Y_i - \hat{m}_{d(-i)}(X_i)\big)^2 \tilde{w}_{d(-i)}^{(1)}(X_i) \\
&\le \frac{2}{n}\sum \varepsilon_{id}^2 \tilde{w}_{d(-i)}^{(1)}(X_i) + \frac{2}{n}\sum \big(\hat{m}_{d(-i)}(X_i) - m_d(X_i)\big)^2 \tilde{w}_{d(-i)}^{(1)}(X_i) \\
&= J_1 + J_2
\end{aligned}$$

and

$$\begin{aligned}
\tilde{w}_d^{(1)}(X) \le \big[& I\big(\|X\|_{(r)} < \|X\| \le G^{-1}(\mu)\big) + I\big(G^{-1}(\mu) < \|X\| \le \|X\|_{(r)}\big)\big] \\
& \times I\big(\|X\|_{(r)} \le G^{-1}(\mu+\varepsilon/2)\big).
\end{aligned}$$

We get by Hölder's inequality that

$$\begin{aligned}
\mathbf{E}J_1 &\le 2\mathbf{E}\varepsilon_{1d}^2 \tilde{w}_{d(-1)}^{(1)}(X_1) \le 2\big[\mathbf{E}\varepsilon_{1d}^{2p}\big]^{1/p}\Big[\big(\tilde{w}_{d(-1)}^{(1)}(X_1)\big)^q\Big]^{1/q} \\
&\le C\Big\{\mathbf{E}\big[I\big(\|X\|_{(r)} < \|X\| \le G^{-1}(\mu)\big) + I\big(G^{-1}(\mu) < \|X\| \le \|X\|_{(r)}\big)\big] \\
&\qquad\qquad \times I\big(\|X\|_{(r)} \le G^{-1}(\mu+\varepsilon/2)\big)\Big\}^{1/q} \\
&\le C\big\{\mathbf{E}|U_{(r)} - \mu|\big\}^{1/q} \le Cn^{-1/2q},
\end{aligned}$$

where $U_{(r)}$ is the corresponding order statistic from standard uniform distribution and the last step is from a result in Reiss [11], Chapter 3. This implies that $J_1 = O_p(n^{-1/2q})$ for any $q > 1$. Furthermore, we have

$$J_2 \le 2\sqrt{\frac{1}{n}\sum \left(\hat{m}_{d(-i)}(X_i) - m_d(X_i)\right)^4 I\left(\|X_i\| \le G^{-1}(\mu + \varepsilon/2)\right)}$$

$$\times \sqrt{\frac{1}{n}\sum \left[\tilde{w}^{(1)}_{d(-i)}(X_i)\right]^2} = 2J_1'J_2'.$$

From an argument similar to that leading to Lemma 1, we can show that $J_1' = O_p(V(d,h))$. Also, since $\mathbf{E}J_2' \le \sqrt{\mathbf{E}\left[\tilde{w}^{(1)}_{d(-1)}(X_1)\right]^2} \le Cn^{-1/4}$, we have $J_2' = o_p(1)$. Hence $J_2 = o_p(V(d,h))$. The above arguments yield that for any $q > 1$,

$$\Delta_1 = O_p(n^{-1/2q}) + o_p(V(d,h)). \tag{A.6}$$

For the Δ_2 in (A.4), since $\hat{m}_d(x) \le C\sum |Y_i|$, we have

$$\mathbf{E}\sup_{(d,h)\in\Omega_n} \Delta_2 = \mathbf{E}\sup \frac{1}{n}\sum \left(Y_i - \hat{m}_{d(-i)}(X_i)\right)^2 \tilde{w}^{(2)}_{d(-i)}(X_i)$$

$$\le C\mathbf{E}\left[|Y| + \sum |Y_i|\right]^2 I\left(\|X\|_{(r)} > G^{-1}(\mu + \varepsilon/2)\right)$$

$$\le Cn^2\mathbf{P}^{1/2}\left[|U_{(r)} - \mu| > \varepsilon/2\right] \le C\exp(-\gamma n).$$

Again, $U_{(r)}$ is the order statistics from uniform distribution, γ is some constant and we used the Cauchy–Schwarz inequality and the exponential bound of order statistics deviation (see Reiss [11]). In other words, we have just shown that

$$\Delta_2 = O_p(\exp(-\gamma n)). \tag{A.7}$$

Finally, putting together (A.5) to (A.7), we get (A.3) and the proof is complete. □

Acknowledgments. Many thanks go to Professor Peter Bickel for his encouragement and numerous helpful discussions during the course of this study. The author is also grateful to the comments of two referees which have led to substantial improvements in the manuscript.

REFERENCES

[1] BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136.

[2] BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

[3] CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

[4] EUBANK, R. (1988). *Spline Smoothing and Nonparametric Regression*. North-Holland, Amsterdam.

[5] FAN, J. Q. (1990). A remedy to regression estimators and nonparametric minimax efficiency. Unpublished manuscript.
[6] FRIEDMAN, J. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.
[7] HALL, P. (1984). Integrated square error properties of kernel estimators of regression functions. *Ann. Statist.* **12** 241–260.
[8] HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.
[9] PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series* **1**. Academic, New York.
[10] PRAKASA RAO, B. L. S. (1983). *Nonparametric Function Estimation*. Academic, New York.
[11] REISS, R. D. (1989). *Approximate Distributions of Order Statistics*. Springer, New York.
[12] ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.
[13] SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.
[14] STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47.
[15] ZHANG, P. (1990). Variable selection in nonparametric regression. Ph.D dissertation, Dept. Statistics, Univ. California, Berkeley.

DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104