JOURNAL OF

TIME SERIES ANALYSIS

A JOURNAL SPONSORED BY THE BERNOULLI SOCIETY FOR
MATHEMATICAL STATISTICS AND PROBABILITY

VOLUME 13, No. 3, 1992
ISSN 0143-9782

BLACKWELL
Publishers

DATA-DEPENDENT ESTIMATION OF PREDICTION FUNCTIONS

BY P. BURMAN AND D. NOLAN

University of California

First version received March 1990

Abstract. The technique of cross-validation for model selection where the observations have martingale-like structure is developed. It is argued that cross-validation works, unaltered, in this more general setting. The specific example of the stationary Markov process is considered in detail. An estimate of the one-step prediction function of this process is selected from a collection of splines by minimizing the cross-validatory version of the prediction error. Asymptotic optimality of the estimate is established.

Keywords. Cross-validation; B-splines; Markov chain; asymptotic optimality; prediction error; autoregressive process.

1. INTRODUCTION

Cross-validation (Stone, 1974; Geisser, 1975) is a technique acclaimed for its model selection ability. This is especially true in the nonparametric setting. It is our belief that the cross-validation folklore implicitly claims its success from the independence of the training sample and the test sample. The aim of this paper is to establish a much broader setting in which cross-validation successfully selects a model. This broader setting includes observations with a martingale-like structure where the criterion function to be cross-validated is of quadratic form.

To best illustrate our claim, consider the regression setting. Suppose one observes pairs (X_i, Y_i) where

$$Y_i = m(X_i) + \varepsilon_i,$$

for some unknown function m and error ε. The prediction error assesses the predictive ability of an estimate $\hat{m}$ of m:

$$\mathrm{PE}(\hat{m}) = E\{Y - \hat{m}(X)\}^2. \tag{1.1}$$

The expection in (1.1) is over a new independent observation (X, Y), conditional on the data. Leave-one-out cross-validation estimates the expected value in $\mathrm{PE}(\hat{m})$ by forming an expectation with respect to the empirical distribution based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. To correct for the double use of the data, the ith pair (X_i, Y_i) is removed from the estimate $\hat{m}$ in the evaluation $\{Y_i - \hat{m}(X_i)\}^2$. This leave-out-one estimate is denoted by $\hat{m}_{-i}$. The exclusion of (X_i, Y_i) enables $\widehat{\mathrm{PE}}(\hat{m})$ to mimic $\mathrm{PE}(\hat{m})$, where

0143-9782/92/03 189–20 \$02.50/0

$$\hat{\mathrm{PE}}(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{-i}(X_i)\}^2. \tag{1.2}$$

Here the training sample is $\{(X_j, Y_j): j \neq i\}$ and the test sample is $\{(X_i, Y_i)\}$. A heuristic argument shows that it is not so much the independence of these two samples that makes cross-validation work, but the quadratic form in (X_i, Y_i), (X_j, Y_j) found in $\hat{\mathrm{PE}}$. More specifically, the zero conditional expectation of the quadratic form for the $i \neq j$ terms drives the successful approximation of PE by $\hat{\mathrm{PE}}$. This can be seen in the approximation below.

$$\begin{aligned}\mathrm{PE}(\hat{m}) &= \int(\hat{m} - m)^2 dF + \int\sigma^2\, dF\\ &= \int(\hat{m} - \bar{m})^2\, dF + \int(\bar{m} - m)^2\, dF + \int\sigma^2\, dF\\ &\approx \sum_{i,j=1}^{n} \varepsilon_i\varepsilon_j W(X_i, X_j) + \int(\bar{m} - m)^2\, dF + \int\sigma^2\, dF,\end{aligned}$$

where $\bar{m}(x) = E\{\hat{m}(x)\}$, $\sigma^2(x) = \mathrm{var}\,(Y_1|X_1 = x)$ and W is a deterministic weight function. Typically, $\hat{\mathrm{PE}}$ approximates PE well if its expectation matches the conditional expectation

$$\sum_{i=1}^{n}\sigma^2(X_i)W(X_i, X_i) + \int(\bar{m} - m)^2\, dF + \int\sigma^2\, dF.$$

This happens when the errors are such that

$$E(\varepsilon_i\varepsilon_j|X_1, \ldots, X_j) = 0 \text{ for } i < j. \tag{1.3}$$

That is, the technique of cross-validation can carry over, without modification, to the dependent setting.

In this paper, we formally establish this assertion for a specific example, where $X_1, \ldots, X_{n+1}$ form a stationary Markov process with one-step prediction function m, that is,

$$Y_i = X_{i+1} = m(X_i) + \varepsilon_i.$$

Section 2 contains a detailed description of this set-up. It is shown that, when nonparametric techniques are used to estimate m, cross-validation of the prediction error can be used successfully to choose the estimate. In Akaike's (1970) study of autoregressive processes he considers the choice of $\hat{m}$ as a prediction problem with prediction error $E\{X_2' - \hat{m}(X_1')\}^2$, for $X_1', X_2', \ldots$, a stationary Markov process with the same distribution as and independent of the original process. We show here that if one cross-validates this prediction error then, just as in the independent case, the estimate $\hat{m}$ that minimizes $\hat{\mathrm{PE}}$ is asymptotically optimal in the sense of Härdle and Marron (1985).

It should be noted that, in dependent settings where (1.3) does not hold, the technique of cross-validation does not work. Instead, it yields a biased estimate of the prediction error. Burman *et al.* (1990), Chu (1989) and Györfi *et al*. (1990) consider modifications of the leave-one-out technique to handle these cases.

Cross-validation, unaltered, can be a tool for model selection in the dependent setting because it capitalizes on the quadratic form of the prediction error. The recent contributions of nonparametric techniques to the arena of dependent data (see Bierens, 1983; Robinson, 1983; Collomb and Härdle, 1986; Truong and Stone, 1989) should find this observation especially welcome. In addition, functionals other than squared error loss may be successfully cross-validated, provided they are smooth enough to be approximated by a quadratic form. This general case is beyond the scope of this paper.

The remainder of the paper is organized as follows. Section 2 formally introduces the notation and main result of the paper. A simulation comparing the cross-validation technique with two simple rules-of-thumb for estimating the prediction error appears in Section 3, along with a numerical example. Section 4 contains some preliminary technical details useful in the proof of the main result, which appears in Section 5 of the paper.

2. THE MAIN RESULT

Let $X_1, \ldots, X_{n+1}$ be the first $n+1$ observations from a stationary Markov process which can be represented as

$$X_{i+1} = m(X_i) + \varepsilon_i,$$

where it is assumed that $E(\varepsilon_i|X_i) = 0$. It is also assumed that the sequence $\{X_i: i = 1, 2, \ldots\}$ satisfies the mixing condition below.

There exists a strictly decreasing function φ such that

(a) $|P(A \cap B) - P(A)P(B)| \leqslant \varphi(t)P(A)$, for all $A \in \mathcal{B}_i$ and $B \in \overline{\mathcal{B}}_{i+t}$, for any i and t, and

(b) $\sum_{j \geqslant 1} \varphi(j) < \infty$,

where $\mathcal{B}_i$ and $\overline{\mathcal{B}}_i$ are the σ-fields generated by $\{X_1, \ldots, X_i\}$ and $\{X_i, X_{i+1}, \ldots\}$, respectively.

Use G to denote the joint distribution function of (X_1, X_2) and g to denote its density. Let F be the marginal distribution function of X_1 and f be its corresponding density. Following Stone (1985) and Burman (1989), we consider spline estimates of the one-step prediction function m restricted to the interval $[0, 1]$.

For positive integers k and v, let $\mathcal{S}_{k,v}$ represent the class of all functions $\{s\}$ on $[0, 1]$ that satisfy the following two properties.

(i) s is a polynomial of degree v on $[(t-1)k^{-1}, tk^{-1}]$, $t = 1, \ldots, k$.

(ii) s is $v-1$ times continuously differentiable on $[0, 1]$.

That is, $\mathcal{S}_{k,v}$ is the class of splines of degree v with k equispaced knots. It is well known that $\mathcal{S}_{k,v}$ has a basis consisting of $k+v$ normalized B-splines $\{B_{k,j}\colon j = 1, \ldots, k+v\}$ (see de Boor, 1978). Our interest is in the spline

$$\hat{m}_k(x) = \sum_{t=1}^{k+v} \hat{\theta}_t B_{kt}(x),$$

where $\hat{\boldsymbol{\theta}}$ is chosen to minimize

$$n^{-1}\sum_{i=1}^{n}\left\{X_{i+1} - \sum_{t=1}^{k+v}\theta_t B_{kt}(X_i)\right\}^2 I(X_i). \tag{2.1}$$

The function $I(\cdot)$ is the indicator function for the interval $[0, 1]$.

As in Section 1, denote by $\hat{m}_{ki}$ the spline constructed from all the observations apart from (X_i, Y_i). Then take $\hat{k}$ to minimize the cross-validated version of the prediction error:

$$\hat{\mathrm{PE}}_k = n^{-1}\sum_{i=1}^{n}\{X_{i+1} - \hat{m}_{ki}(X_i)\}^2 I(X_i),$$

where k ranges from 1 to $K_n = n^{1-\alpha}$ for some arbitrary small positive α. The minimization of $\hat{\mathrm{PE}}(k)$ is almost as good as the minimization of $\mathrm{PE}(k)$. This is formally stated in the following theorem.

THEOREM 2.1. *Suppose the following conditions hold for $X_1, \ldots, X_{n+1}$, a stationary Markov process satisfying the mixing condition stated above*:

(i) $0 < f(x) < c$, *for all* $x \in [0, 1]$, *for some constant* c;
(ii) *the conditional densities* $\{f_{X_i|X_j}\}$ *are uniformly bounded on* $[0, 1]^2$;
(iii) *the prediction function* m *is not a spline of degree* v *for any* k; (iv) $\sup_{0\leq x\leq 1} E\{|X_2|^r|X_1 = x\} < \infty$, *for all* $r > 0$.

Then:

$$\frac{\int(\hat{m}_{\hat{k}} - m)^2 I dF}{\inf_{k\leq K_n}\int(\hat{m}_k - m)^2 I dF} \to^{\mathrm{P}} 1.$$

REMARKS

1. It is possible to extend this theorem to hold for a sequence $\{X_j\}$ that is nonstationary but has stationary transition probabilities. In this case, if the joint distribution function of X_i and X_{i+1} is denoted by $G_{i,i+1}$, then PE_k is taken to be
$$\int\{y - \hat{m}_k(x)\}^2 I(x)\, d\bar{G}_n(x, y)$$
where $\bar{G}_n(x, y) = n^{-1}\sum_{1\leq i\leq n} G_{i,i+1}(x, y)$.
2. It is assumed here that the knots of the splines are equispaced. In practice, however, it may be desirable to place the knots at the sample quantiles. The main result should carry over to this case, but we do not prove it here.
3. We believe that Theorem 2.1 can be extended to the pth-order Markov chain where the problem is to estimate $m(x_1, \ldots, x_p) = E(X_i|X_{i-1} = x_1, \ldots, X_{i-p} = x_p)$. To do so, we would use tensor products of one-dimensional B-splines to create multivariate spline estimates (see Burman

and Chen, 1989). Also, we believe our arguments are valid for other types of estimators such as those based on kernal or nearest-neighbor techniques.

4. The mixing condition given above is satisfied if
$$\sup_x \int |f_{X_{n+1}|X_1}(y|x) - f(y)|dy \leqslant \varphi(n),$$
for a decreasing sequence $\{\varphi(\cdot)\}$ converging to zero. In fact, if
$$\sup_x \int |f_{X_2|X_1}(y|x) - f(y)|dy = \rho < 1$$
then take $\varphi(n) = \rho^n$ to meet this condition. The model used in the simulation in the next section is such an example.
5. Condition (iv) in Theorem 2.1 can be weakened to hold for $r \leqslant 8$. This requires excessive calculation in the proof of the theorem, however, for we can no longer use part (g) of Lemma 4.1 (see later).

3. SOME NUMERICAL RESULTS

3.1. *A simulation*

In this section we present a simulation in support of our theoretical results. Observations are generated from a stationary Markov process with Uniform(0, 1) marginal density and joint density (of X_1 and X_2)
$$g(x_1, x_2) = \begin{cases} 1 + 0.9\sin(2\pi x_1)(2x_2 - 1) & 0 \leqslant x_1, x_2 \leqslant 1 \\ 0 & \text{otherwise.} \end{cases}$$

The one-step prediction function is then
$$m(x) = 0.5 + 0.15\sin(2\pi x).$$

To estimate m, fit a quadratic spline with knots placed at the sample quantiles, choosing the number of knots by the cross-validation method as described in Section 2. Compare the cross-validatory choice of the number of knots with the *ad hoc* rule that takes the number of knots to be a fixed fraction of the sample, say 10% or 20%. To measure the performance of these methods compute the ratio
$$\frac{\inf_k \int (\hat{m}_k - m)^2 dF}{\int (\hat{m}_{k''} - m)^2 dF}$$
where $\hat{m}_k$ is the quadratic spline estimate of m with k knots and k'' stands for the number of knots chosen according to one of the competing rules. Table I presents the mean, median and standard deviations of these ratios for $n = 50$ and $n = 100$. All are based on 400 repeats. The closer the ratio is to 1, the better is the method. Table I shows that the median ratio for the cross-validatory choice (denoted by CV in the table) is close to 1 for both sample sizes.

TABLE I

THE MEAN, MEDIAN AND STANDARD DEVIATION OF THE THREE PERFORMANCE RATIOS

Knot selection	$n = 50$			$n = 100$		
	CV	10% of n	20% of n	CV	10% of n	20% of n
Mean	0.74	0.53	0.25	0.78	0.30	0.13
Median	0.94	0.27	0.19	0.998	0.28	0.10
Standard deviation	0.34	0.54	0.21	0.32	0.18	0.11

3.2. *A numerical example*

We close this section with an illustration of the use of a cross-validatory choice of model in the Markov chain context. One hundred pairs (X_i, X_{i+1}) were generated according to the following specifications:

$$X_{i+1} = 0.5 + 0.25\sin(2\pi X_i) + U_i,\ i = 1, \ldots, 100,$$

where $\{U_i\}$ are independent observations from the Uniform distribution on $[-0.25, 0.25]$. The prediction error (1.1) was then cross-validated (1.2) and minimized over the collection of quadratic splines on the unit interval with knots located at the sample quantiles. In this example, a quadratic spline with two knots was chosen. The spline estimate is plotted in Figure 1 along with the true prediction function and the observations. The two curves appear to be remarkably close to one another, except possibly near the endpoints of the interval of estimation.

4. PRELIMINARIES

In this section we present some preliminary results to be used in establishing asymptotic optimality of $\hat{k}$. For the vector $\boldsymbol{u}$ in R^t, write $\|\boldsymbol{u}\|$ to denote the usual Euclidean norm of $\boldsymbol{u}$ and for a $t \times t$ matrix H, write its matrix norm as $\|H\|$. The empirical distribution function constructed from $(X_1, X_2), \ldots, (X_n, X_{n+1})$ is denoted by G_n and that of $X_1, \ldots, X_n$ is denoted by F_n. For any two sequences of random variables $\{\xi_{n,\rho}\colon \rho \in D_n\}$ and $\{\eta_{n,\rho}\colon \rho \in D_n\}$, where D_n is an index set, write $\xi_{n,\rho} = \mathrm{O_p}(\eta_{n,\rho})$ to mean that $\sup_{\rho\in D_n} |\xi_{n,\rho}/\eta_{n,\rho}| = \mathrm{O_p}(1)$. Similarly define $\xi_{n,\rho} = \mathrm{o_p}(\eta_{n,\rho})$.

Define

$$A_k = \int \boldsymbol{B}_k(x)\boldsymbol{B}'_k(x)I(x)dF(x) \quad \boldsymbol{b}_k = \int m(x)\boldsymbol{B}_k(x)I(x)dF(x)$$

$$\hat{A}_k = \int \boldsymbol{B}_k(x)\boldsymbol{B}'_k(x)I(x)dF_n(x) \quad \tilde{\boldsymbol{b}}_k = \int m(x)\boldsymbol{B}_k(x)I(x)dF_n(x)$$

$$\hat{\boldsymbol{b}}_k = n^{-1}\sum_{i=1}^{n} X_{i+1}\boldsymbol{B}_k(X_i)I(X_i) \quad \boldsymbol{\theta}_k = A_k^{-1}\boldsymbol{b}_k.$$

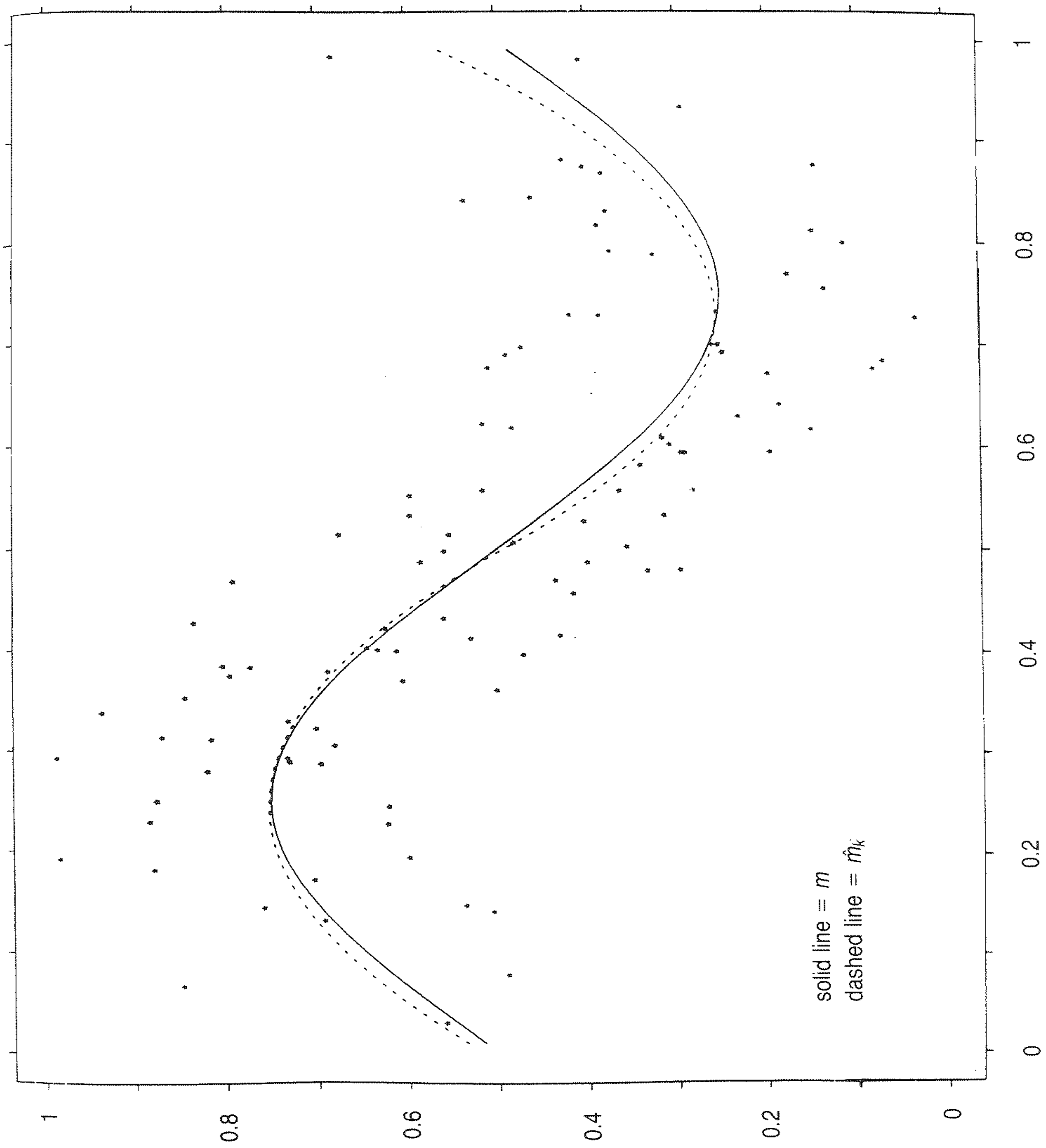

FIGURE 1. Cross-validatory quadratic-spline approximation to $m(x) = 0.5 + 0.25\sin(2\pi x)$.

Let $\delta_{nk} = 1/(kn^{1-\delta})^{1/2}$ for some $0 < \delta < \alpha/2$ where α is the constant that appears in the upper bound K_n on k. The proof of the following result is contained in Lemma 3.6 of Burman (1989); it uses properties of splines.

LEMMA 4.1. *Under conditions (i), (ii) and (iv) of Theorem 2.1:*

(a) The eigenvalues of A_k are bounded by $c_1 k^{-1}$ and $c_2 k$, for $0 < c_1 < c_2$;
(b) $\|\widehat{A}_k - A_k\| = o_P(\delta_{nk})$;
(c) $\|\widehat{A}_k\| = O_P(k^{-1})$;
(d) $\widehat{A}_k^{-1}$ exists except on an event whose probability goes to zero as $n \to \infty$;
(e) $\|\widehat{A}_k^{-1} - A_k^{-1}\| = o_P(k^2\delta_{nk})$;
(f) $\|\widehat{A}_k^{-1}\| = O_P(k)$;
(g) $\|\widehat{\boldsymbol{b}}_k - \widetilde{\boldsymbol{b}}_k\| = o_P(k^{1/2}\delta_{nk})$;
(h) $\|\widetilde{\boldsymbol{b}}_k - \boldsymbol{b}_k\| = o_P(k^{1/2}\delta_{nk})$.

LEMMA 4.2. *Suppose that q is a function on $[0, 1]$ such that $|q| \leq 1$ and $\int q\,dF = 0$. Assume that*

$$\sup_{0 \leq x \leq 1} \int \frac{g(x, y)}{f(x)f(y)}\, dF(y) < \infty.$$

Then for any positive integer s

$$E\left\{\sum_{i=1}^{n} q(X_i)\right\}^{2s} \leq c(s)\sum_{t=1}^{s}\left\{n\int q^2\,dF\right\}^{t},$$

where $c(s)$ is a constant that depends only on s.

Finally, re-express

$$\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k = A_k^{-1}\boldsymbol{e}_k + \boldsymbol{r}_k, \tag{4.1}$$

where

$$\boldsymbol{e}_k = \widehat{\boldsymbol{b}}_k - \widetilde{\boldsymbol{b}}_k = n^{-1}\sum_{i=1}^{n}\varepsilon_i \boldsymbol{B}_k(X_i)I(X_i),$$

$$\varepsilon_i = X_{i+1} - m(X_i)$$

and

$$\boldsymbol{r}_k = (\widehat{A}_k^{1} - A_k^{-1})\boldsymbol{e}_k - \widehat{A}_k^{-1}\int \boldsymbol{B}_k(x)\{m_k(x) - m(x)\}I(x)d(F_n - F)(x).$$

LEMMA 4.3. *Under conditions (i), (ii) and (iv) of Theorem 2.1:*

(a) $\|\int \boldsymbol{B}_k(x)\{m_k(x) - m(x)\}I(x)d(F_n - F)(x)\|^2$
$= o_P(n^{2\delta})\{kn^{-2} + n^{-1}\|m_k - m\|^2\}$,
(b) $\|\boldsymbol{r}_k\| = o_P(n^{\delta})\{k^{3/2}n^{-1} + k^{1/2}n^{-1/2}\|m_k - m\|\}$,

where for any function u, $\|u\| = \{\int u^2(x)I(x)dF(x)\}^{1/2}$.

The proofs of Lemmas 4.2 and 4.3 appear at the end of Section 5.

5. THE PROOFS

Throughout the proofs c will denote an arbitrary positive constant that changes from one line to the next.

Define

$$L_n(k) = \int\{\hat{m}_k(x) - m_k(x)\}^2 I(x)dF(x) + \int\{m_k(x) - m(x)\}^2 I(x)dF(x),$$

$$\hat{L}_n(k) = \hat{\mathrm{PE}}_k - \int\sigma^2(x)dF(x).$$

$$V_n(k) = n^{-1}\int\sigma^2(x)\boldsymbol{B}'_k(x)A_k^{-1}\boldsymbol{B}_k(x)I(x)dF(x)$$
$$+ \int\{m_k(x) - m(x)\}^2 I(x)dF(x).$$

Note that

$$\|\hat{m}_k - m\|^2 = L_n(k),$$
$$\mathrm{PE}_k = \int\sigma^2(x)dF(x) + L_n(k),$$

and

$$V_n(k) \geqslant c_4(k/n) + \|m_k - m\|^2. \tag{5.1}$$

Theorem 2.1 follows easily from Lemmas 5.1 and 5.2 below. These lemmas are typical of proofs of asymptotic optimality for cross-validatory choice of smoothing parameters in nonparametric function estimation (in the independent setting). The main difference, however, is the approximation

$$\sum_{i,j=1}^{n} \varepsilon_i\varepsilon_j W(X_i, X_j) \approx \sum_{i=1}^{n}\sigma^2(X_i)W(X_i, X_i),$$

where $W(x, y) = \boldsymbol{B}'_k(x)A_k^{-1}\boldsymbol{B}_k(y)I(x)I(y)$. This approximation is key to the proofs of Lemmas 5.1 and 5.2. It is presented as a separate lemma below.

LEMMA 5.1. *As $n \to \infty$ we have that*

$$\sup_{k \leqslant K_n} \frac{|L_n(k) - V_n(k)|}{V_n(k)} \to^{\mathrm{P}} 0.$$

LEMMA 5.2. *If k^* minimizes V_n over all $k \leqslant K_n$, then*

$$\sup_{k \leqslant K_n} \frac{|\{\hat{L}_n(k) - L_n(k)\} - \{\hat{L}_n(k^*) - L_n(k^*)\}|}{V_n(k)} \to^{\mathrm{P}} 0.$$

LEMMA 5.3. *There exists a constant $c > 0$ such that*

$$E\left\{\sum_{i,j=1}^{n} \varepsilon_i\varepsilon_j W(X_i, X_j) - \sum_{i=1}^{n}\sigma^2(X_i)W(X_i, X_i)\right\}^4 \leqslant ck^2n^4.$$

PROOF OF LEMMA 5.1. Note that

$$L_n(k) - V_n(k) = \|\widehat{m}_k - m_k\|^2 - n^{-1}\int \sigma^2(x) W(x, x)\, dF(x). \tag{5.2}$$

By Lemmas 4.1 and 4.3

$$\begin{aligned}\big|\|\widehat{m}_k - m_k\|^2 - \|\boldsymbol{e}_k A_k^{-1} \boldsymbol{B}_k\|^2\big| &\leqslant \int (\boldsymbol{r}'_k \boldsymbol{B}_k)^2 dF + 2\left|\int (\boldsymbol{r}'_k \boldsymbol{B}_k)(\boldsymbol{e}'_k A_k^{-1} \boldsymbol{B}_k)\, dF\right| \\ &\leqslant \|\boldsymbol{r}_k\|^2 \|A_k\| + 2\|\boldsymbol{r}_k\| \|A_k\| \|A_k^{-1}\| \|\boldsymbol{e}_k\| \\ &= o_P(n^{2\delta})\{k^3 n^{-2} + k n^{-1} \|m_k - m\|^2\} O(k^{-1}) \\ &\quad + o_P(n^{\delta})\{k^{3/2} n^{-1} + k^{1/2} n^{-1/2} \|m_k - m\|\} O(k^{-1}) O(k) o_P(k^{1/2} \delta_{nk}) \\ &= o_P(V_n(k)).\end{aligned} \tag{5.3}$$

Substitute (5.3) into (5.2) to get

$$\begin{aligned}L_n(k) - V_n(k) = \int \{\boldsymbol{e}'_k A_k^{-1} \boldsymbol{B}_k(x)\}^2 I(x)\, dF(x) \\ - n^{-1}\int \sigma^2(x) W(x, x)\, dF(x) + o_P(V_n(k)).\end{aligned}$$

Lemma 4.3 implies that the first two terms together are $o_P(V_n(k))$. This completes the proof of Lemma 5.1.

PROOF OF LEMMA 5.2. Re-express $\widehat{\mathrm{PE}}(k)$ (Cook and Weisberg (1982), p. 33, Equation 2.2.23) as

$$\sum_{i=1}^{n} \frac{\{X_{i+1} - \widehat{m}_k(X_i)\}^2 I(X_i)}{(1 - h_i/n)^2}$$

where

$$h_i = \boldsymbol{B}'_k(X_i) \widehat{A}_k^{-1} \boldsymbol{B}_k(X_i).$$

Then

$$\widehat{L}_n(k) = n^{-1} \sum_{i=1}^{n} \frac{\{X_{i+1} - \widehat{m}_k(X_i)\}^2 I(X_i)}{1 + 2n^{-1} h_i} + o_P(V_n(k)) - \int \sigma^2 dF \tag{5.4}$$

The stochastic order term follows from Lemma 4.1 applied to the inequality

$$|h_i|^2 \leqslant \|\widehat{A}_k^{-1}\|^2 = O_P(k^2),$$

and from the fact that

$$n^{-1} \sum_{i=1}^{n} \{X_{i+1} - \widehat{m}_k(X_i)\}^2 I(X_i) = O_P(1).$$

Ignore the stochastic order term in (5.4) and rewrite the difference $\widehat{L}_n(k) - L_n(k)$ as

$$
\begin{aligned}
&\Big\{n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 I(X_i) - \int\sigma^2 IdF\Big\} + \Big\{\int(\hat{m}_k - m)^2 Id(F_n - F)\Big\}\\
&\quad -2\Big[n^{-1}\sum_{i+1}^{n}\varepsilon_i\{\hat{m}_k(X_i) - m(X_i)\}I(X_i) - n^{-2}\sum_{i=1}^{n}\varepsilon_i^2 I(X_i)h_i\Big]\\
&\quad + 2\Big[n^{-2}\sum_{i=1}^{n}\{\hat{m}_k(X_i) - m(X_i)\}^2 I(X_i)h_i\\
&\quad + 4n^{-2}\sum_{i=1}^{n}\varepsilon_i\{\hat{m}_k(X_i) - m(X_i)\}I(X_i)h_i\Big]\\
&\quad - 4\Big[n^{-2}\sum_{i=1}^{n}\varepsilon_i\{\hat{m}_k(X_i) - m(X_i)\}I(X_i)h_i\Big]\\
&\quad = T_1 + T_2(k) - 2T_3(k) + 2T_4(k) - 4T_5(k), \text{ say.} \qquad (5.5)
\end{aligned}
$$

The first term T_1 does not depend on k. Therefore it does not enter into the difference $\{\hat{L}_n(k) - L_n(k)\} - \{\hat{L}_n(k^*) - L_n(k^*)\}$. The remaining terms are handled one by one, beginning with T_2.

$$
\begin{aligned}
T_2(k) &= \int(\hat{m}_k - m)^2 Id(F_n - F)\\
&= \int(\hat{m}_k - m_k)^2 Id(F_n - F) + \int(m_k - m)^2 Id(F_n - F)\\
&\quad + 2\int(\hat{m}_k - m_k)(m_k - m)Id(F_n - F)\\
&= T_{21}(k) + T_{22}(k) + 2T_{23}(k).
\end{aligned}
$$

Lemmas 4.1 and 4.3 imply

$$
\begin{aligned}
|T_{21}(k)| &= \Big|\int(\hat{m}_k - m_k)^2 Id(F_n - F)\Big|\\
&= \Big|\int[\boldsymbol{e}_k' A_k^{-1}\boldsymbol{B}_k(x)\boldsymbol{B}_k'(x)A_k^{-1}\boldsymbol{e}_k + \boldsymbol{r}_k'\boldsymbol{B}_k(x)\boldsymbol{B}_k'(x)\boldsymbol{r}_k\\
&\qquad - 2\boldsymbol{e}_k' A_k^{-1}\boldsymbol{B}_k(x)\boldsymbol{B}_k'(x)\boldsymbol{r}_k]I(x)d(F_n - F)(x)\Big|\\
&= |\boldsymbol{e}_k' A_k^{-1}(\hat{A}_k - A_k)A_k^{-1}\boldsymbol{e}_k + \boldsymbol{r}_k'(\hat{A}_k - A_k)\boldsymbol{r}_k\\
&\qquad - 2\boldsymbol{e}_k' A_k^{-1}(\hat{A}_k - A_k)\boldsymbol{r}_k|\\
&\leqslant \|A_k^{-1}\|^2\|\hat{A}_k - A_k\|\|\boldsymbol{e}_k\|^2 + \|\boldsymbol{r}_k\|^2\|\hat{A}_k - A_k\|\\
&\qquad + 2\|\boldsymbol{e}_k\|\|A_k^{-1}\|\|\hat{A}_k - A_k\|\|\boldsymbol{r}_k\|\\
&= \mathrm{O_P}(k^2)\mathrm{o_P}(\delta_{nk})\mathrm{o_P}(k\delta_{nk}^2) + \mathrm{o_P}(n^{2\delta})(k^3 n^{-2} + kn^{-1}\|m_k - m\|^2)\mathrm{o_P}(\delta_{nk})\\
&\qquad + \mathrm{o_P}(k^{1/2}\delta_{nk})\mathrm{O}(k)\mathrm{o_P}(\delta_{nk})\mathrm{o_P}(n^{\delta})(k^{3/2}n^{-1} + k^{1/2}n^{-1/2}\|m_k - m\|)\\
&= \mathrm{o_P}(V_n(k)). \qquad (5.6)
\end{aligned}
$$

Apply Lemma 4.2 to find

$$E|T_{22}(k)|^4 = \mathrm{O}(1)n^{-4}\left[\left\{n\int(m_k - m)^2 IdF\right\} + \left\{n\int(m_k - m)^2 IdF\right\}^2\right]$$

which shows that $T_{22}(k) = \mathrm{o_P}(V_n(k))$. Then for T_{23} Lemmas 4.1 and 4.3 and a little algebra imply

$$\begin{aligned}|T_{23}(k)| &= \left|\int\{\boldsymbol{e}'_k A_k^{-1}\boldsymbol{B}_k(x) + \boldsymbol{r}'_k\boldsymbol{B}_k(x)\}\{m_k(x) - m(x)\}I(x)d(F_n - F)(x)\right| \\ &\leqslant (\|\boldsymbol{e}_k\|\|A_k^{-1}\| + \|\boldsymbol{r}_k\|)\|\boldsymbol{B}_k(x)\{m_k(x) - m(x)\}I(x)d(F_n - F)(x)\| \\ &= \{\mathrm{o_P}(k^{1/2}\delta_{nk})\mathrm{O}(k) + \mathrm{o_p}(n^{\delta})(k^{3/2}n^{-1} + k^{1/2}n^{-1/2}\|m_k - m\|)\} \\ &\quad \times\{\mathrm{o_P}(n^{\delta})(k^{1/2}n^{-1} + n^{-1/2}\|m_k - m\|)\} \\ &= \mathrm{o_P}(n^{2\delta})(k^{3/2}n^{-1} + k^{1/2}n^{-1/2}\|m_k - m\|)(k^{1/2}n^{-1} + n^{-1/2}\|m_k - m\|) \\ &= \mathrm{o_P}(n^{2\delta})(k^{3/2}n^{-2} + k^{3/2}n^{-3/2}\|m_k - m\| + k^{1/2}n^{-1}\|m_k - m\|^2) \\ &= \mathrm{o_P}(V_n(k)).\end{aligned}$$

The term T_2 has been dealt with and we now turn to T_3. Denote $\bar{h}_i = W(X_i, X_i)$. Break T_3 into five subterms as follows.

$$\begin{aligned}T_3(k) = &\{n^{-1}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{e}'_k A_k^{-1}\boldsymbol{B}_k(X_i)I(X_i) - n^{-1}\int\sigma^2(x)W(x, x)dF(x)\} \\ &+ \left\{n^{-1}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{r}'_k\boldsymbol{B}_k(X_i)\right\} - \left[n^2\sum_{i=1}^{n}\{\varepsilon_i^2 - \sigma^2(X_i)\}I(X_i)(h_i - \bar{h}_i)\right] \\ &- \left[n^{-2}\sum_{i=1}^{n}\{\varepsilon_i^2 - \sigma^2(X_i)\}I(X_i)\bar{h}_i\right] \\ &- \left\{n^{-1}\int\sigma^2(x)W(x, x)d(F_n - F)(x)\right\} \\ = &\ T_{31}(k) + T_{32}(k) - T_{33}(k) - T_{34}(k) - T_{35}(k).\end{aligned}$$

We show that $T_{31}, \ldots, T_{35}$ are each $\mathrm{o_P}(V_n(k))$. For T_{31} , apply Lemma 5.3. As for the second term, it is a candidate for Lemma 4.3 because $T_{32}(k) = \boldsymbol{e}_k\boldsymbol{r}_k$. To deal with T_{33} use the bound implied by Lemma 4.1,

$$\sup_{1\leqslant i\leqslant n}|h_i - \bar{h}_i| \leqslant \|\hat{A}_k^{-1} - A_k^{-1}\| = \mathrm{o_P}(k^2\delta_{nk}),$$

to find

$$|T_{33}| \leqslant n^{-1}\mathrm{o_P}(k^2\delta_{nk})\left\{n^{-1}\sum_{i=1}^{n}|\varepsilon_i^2 - \sigma^2(X_i)|I(X_i)\right\} = \mathrm{o_P}(n^{-1}k^2\delta_{nk})\mathrm{O_P}(1).$$

Next, for T_{34}, use the fact that $\{[\varepsilon_i^2 - \sigma^2(X_i)]I(X_i)\bar{h}_i\}$ is a sequence of martingale differences with respect to the σ-fields $\{\mathcal{B}_i\}$ to show

$$\begin{aligned}E|T_{34}(k)|^2 &= n^{-4}\sum_{i=1}^{n}E[\{\varepsilon_i^2 - \sigma^2(X_i)\}I(X_i)\bar{h}_i]^2 \\ &= \mathrm{O}(1)k^2n^{-3}\end{aligned}$$

since $|\bar{h}_i| = O(k)$. Then, for T_{35}, by Lemma 4.1,

$$\sup_{0\leqslant x\leqslant 1} W(x, x) \leqslant \|A_k^{-1}\| \sup_{0\leqslant x\leqslant 1} \|\boldsymbol{B}_k(x)\|^2 \leqslant \|A_k^{-1}\| = O(k).$$

This observation along with Lemma 3.2 gives us

$$E|T_{35}(k)|^4 = O(1)n^{-4}k^2.$$

Therefore $T_3 = o_P(V_n(k))$.

Equation (5.6) implies that T_4 is also $o_P(V_n(k))$. Finally,

$$\begin{aligned}|T_5(k)| &= O_P\left(\frac{k}{n}\right)\left(n^{-1}\sum_{i=1}^{n}\varepsilon_i^2\right)^{1/2}\left\{\int(\hat{m}_k - m)^2 I dF_n\right\}^{1/2}\\ &= o_P(V_n(k)).\end{aligned}$$

The proof of Lemma 5.2 is complete.

Proof of lemma 5.3. Note that

$$\begin{aligned}&\sum_{i,j=1}^{n}\varepsilon_i\varepsilon_j W(X_i, X_j) - \sum_{i=1}^{n}\sigma^2(X_i)W(X_i, X_i)\\ &\qquad= \sum_{i=1}^{n}\{\varepsilon_i^2 - \sigma^2(X_i)\}W(X_i, X_i) + 2\sum_{i=2}^{n}\varepsilon_i\sum_{j<i}\varepsilon_j W(X_i, X_j)\\ &\qquad= T_6 + 2T_7, \text{ say.}\end{aligned}$$

Note that $\{[\varepsilon_i^2 - \sigma^2(X_i)]W(X_i, X_j)\}$ is a sequence of martingale differences with respect to the σ-fields $\{\mathcal{B}_i\}$. Using Burkholder's inequality (Hall and Heyde (1980), p. 23, Theorem 2.10) we obtain

$$E|T_6|^4 \leqslant cE\left|\sum_{i=1}^{n}\{\varepsilon_i^2 - \sigma^2(X_i)\}^2 W^2(X_i, X_i)\right|^2.$$

Note that $\|\boldsymbol{B}_k(x)\|^2 = \sum_{t=1}^{k+v}\boldsymbol{B}_{kt}^2(x) \leqslant 1$ for all x. Consequently,

$$|W(x, y)| \leqslant \|\boldsymbol{B}_k(x)\|\|\boldsymbol{B}_k(y)\|\|A_k^{-1}\| \leqslant ck. \tag{5.7}$$

It then follows from (5.7) and condition (iv) of Theorem 2.1 that

$$E|T_6|^4 \leqslant cn^2k^3.$$

Turn to T_7. The sequence $\{\varepsilon_i\sum_{j<i}\varepsilon_j W(X_i, X_j)\}$ is also a sequence of martingale differences with respect to the σ-fields $\{\mathcal{B}_i\}$. Once again, apply Burkholder's inequality followed by Jensen's inequality.

$$\begin{aligned}E|T_7|^4 &\leqslant cE\left|\sum_{i=2}^{n}\left\{\varepsilon_i\sum_{j<i}\varepsilon_j W(X_i, X_j)\right\}^2\right|^2\\ &\leqslant cn\sum_{i=2}^{n}E\left\{\varepsilon_i\sum_{j<i}\varepsilon_j W(X_i, X_j)\right\}^4\\ &= cn\sum_{i=2}^{n}E\left\{\sum_{j<i}\varepsilon_j W(X_i, X_j)\right\}^4.\end{aligned}$$

Then, provided that

$$E\left\{\sum_{j<i}\varepsilon_j W(X_i,\, X_j)\right\}^4 \leqslant c(k^4 + ik^3 + i^2k^2), \tag{5.8}$$

we have

$$\begin{aligned} E|T_7|^4 &\leqslant cn\sum_{i=2}^{n}(k^4 + ik^3 + i^2k^2) \\ &\leqslant c(n^2k^4 + n^3k^3 + n^4k^2). \end{aligned}$$

The proof will be complete once we establish (5.8).

$$\begin{aligned} \sum_{j<i}\varepsilon_j W(X_i,\, X_j) &= \varepsilon_{i-1}W(X_i,\, X_{i-1}) + \sum_{j<i-1}\varepsilon_j W(X_i,\, X_j) \\ &= T_8 + T_9. \end{aligned}$$

Clearly,

$$E|T_8|^4 \leqslant ck^4. \tag{5.9}$$

Let the elements of A_k^{-1} be denoted by $a^{s,t}$. For notational simplicity we will write $\{B_{k,t}\}$ as $\{B_t\}$. Then

$$W(x,\, y) = \sum_{s,t=1}^{k+v} a^{s,t}B_s(x)B_t(y).$$

Express T_9 as

$$\sum_{s,t=1}^{k+v} a^{s,t}B_s(X_i)\sum_{j<i-1}\varepsilon_j B_t(X_j) = \sum_{s,t=1}^{k+v} B_s(X_i)\mu_s(\boldsymbol{Z}_i),$$

where $\boldsymbol{Z}_i = (X_1,\, \ldots,\, X_{i-1})$ and

$$\mu_s(\boldsymbol{Z}_i) = \sum_{t=1}^{k+v} a^{s,t}\sum_{j<i-1}\varepsilon_j B_t(X_j).$$

Then

$$E|T_9|^4 \leqslant \sum_{s_1,\ldots,s_4=1}^{k+v}\ \sum_{u_1+\ldots+u_4=4} E\left[\prod_{\tau=1}^{4}\{B_{s_\tau}(X_i)\mu_{s_\tau}(\boldsymbol{Z}_i)^{u_\tau}\right] \tag{5.10}$$

The second summation in (5.10) is over all nonnegative integers $u_1,\, \ldots,\, u_4$ such that $u_1 + \ldots + u_4 = 4$. The B-splines are zero outside an interval of length $(v+1)/k$ and the product $B_{S_1}(x) \cdots B_{s_4}(x)$ is zero except when $|s_i - s_j| \leqslant v + 1$ for $1 \leqslant i, j \leqslant 4$. Bound the nonzero summands in (5.10) using assumption (ii) of Theorem 2.1 and two applications of the Cauchy–Schwarz inequality.

$$\begin{aligned} \left|E\prod_{\tau=1}^{4}\{B_{s_\tau}(X_i)\mu_{s_\tau}(\boldsymbol{Z}_i)\}^{u_\tau}\right| &\leqslant E\left|\prod_{\tau=1}^{4}\{\mu_{s_\tau}(\boldsymbol{Z}_i)\}^{u_\tau}E\left[\prod_{\tau=1}^{4}\{B_{s_\tau}(X_i)\}^{u_\tau}\middle|\mathcal{B}_{i-1}\right]\right| \\ &\leqslant ck^{-1}E\prod_{\tau=1}^{4}|\mu_{s_\tau}(\boldsymbol{Z}_i)|^{u_\tau} \end{aligned}$$

$$\leqslant ck^{-1}\prod_{\tau=1}^{4}[E\{\mu_{s_\tau}(Z_i)\}^4]^{u_\tau/4}$$

$$\leqslant ck^{-1}\sup_{1\leqslant\tau\leqslant 4} E\{\mu_s(Z_i)\}^4. \tag{5.11}$$

Substitute (5.11) into (5.10) for the nonzero summands to get

$$E|T_9|^4 \leqslant c\sup_{1\leqslant\tau\leqslant 4} E\{\mu_s(Z_i)\}^4.$$

By a result of Demko (1977).

$$|a^{s,t}| \leqslant ck\rho^{|s-t|}$$

for some $0<\rho<1$. Apply Minkowski's inequality to get

$$E|T_9|^4 \leqslant c\sup_{1\leqslant s\leqslant k+v}\left|\sum_{t=1}^{k+v}|a^{s,t}|\left[E\left\{\sum_{j<i-1}\varepsilon_j B_t(X_j)\right\}^4\right]^{1/4}\right|^4$$

$$\leqslant ck^4\sup_{1\leqslant t\leqslant k+v} E\left\{\sum_{j<i-1}\varepsilon_j B_t(X_j)\right\}^4 \tag{5.12}$$

Since $\{\varepsilon_j B_t(X_j)\}$ is a sequence of martingale differences with respect to the σ-fields $\{\mathcal{B}_j\}$. Burkholder's inequality implies

$$E\left\{\sum_{j<i-1}\varepsilon_j B_t(X_j)\right\}^4 \leqslant E\left[\sum_{j<i-1}\{\varepsilon_j B_t(X_j)\}^2\right]^2$$

$$= \sum_{j<i-1}E\{\varepsilon_j B_t(X_j)\}^4$$

$$+ 2\sum_{j_1<j_2<i-1}E\{\varepsilon_{j_1}B_t(X_{j_1})\}^2\{\varepsilon_{j_2}B_t(X_{j_2})\}^2$$

$$\leqslant c(ik^{-1}+i^2k^{-2}).$$

Substituting this last inequality into (5.12) we obtain

$$E|T_9|^4 \leqslant c(ik^3+i^2k^2). \tag{5.13}$$

Complete the proof by noting that (5.8) follows from (5.9) and (5.13).

Proof of lemma 4.2. Rewrite $\sum_{1\leqslant i\leqslant n}q(X_i)$ as

$$\sum_{i=1}^{n}q(X_i) = \sum_{i=1}^{n}[q(X_i) - E\{q(X_i)|X_{i-1}\}] + \sum_{i=1}^{n}E\{q(X_i)|X_{i-1}\}$$

$$= T_{10} + T_{11}.$$

Denote $E\{q(X_i)|X_{i-1}=x\}$ by $q_1(x)$. Then

$$\sup_{0\leqslant x\leqslant 1}|q_1(x)| \leqslant \|q\|\sup_{0\leqslant x\leqslant 1}\left|\int\left\{\frac{g(x,y)}{f(x)f(y)}\right\}^2 dF(y)\right|$$

$$\leqslant c\|q\|.$$

According to Theorem 3.1 in Roussas (1988), if $\{Z_i\}$ is a stationary stochastic process satisfying the mixing condition in Section 2, and $Y_{ni} = \rho_n(Z_i)$ for some function ρ_n with $E(Y_{ni}) = 0$ and $|Y_{ni}| \leqslant 1$ then

$$E\left\{\sum_{i=1}^{n} Y_{ni}\right\}^{2s} \leqslant c(s)n^s. \tag{5.14}$$

When $\{Z_i\}$ is a Markov process, it is possible to extend (5.14) to the nonstationary case. Using (5.14) we obtain

$$E|T_{11}|^{2s} = E\left|\sum_{i=1}^{n} q_1(X_{i-1})\right|^{2s} \leqslant cn^s\|q\|^{2s}. \tag{5.15}$$

Since $\{q(X_i) - E[q(X_i)|X_{i-1}]\}$ is a sequence of martingale differences with respect to $\{\mathcal{B}_i\}$. Burkholder's inequality and (5.15) imply

$$\begin{aligned} E|T_{10}|^{2s} &\leqslant E\left(\sum_{i=1}^{n}[q(X_i) - E\{q(X_i)|X_{i-1}\}]^2\right)^s \\ &\leqslant 2^{s-1}\left[E\left|\sum_{i=1}^{n} q^2(X_i)\right|^s + E\left|\sum_{i=1}^{n} E\{q(X_i)|X_{i-1}\}^2\right|^s\right] \\ &\leqslant 2^{s-1}\left(E\left|\sum_{i=1}^{n} q^2(X_i)\right|^s + n^{s-1}\sum_{i=1}^{n} E[E\{q(X_i)|X_{i-1}\}]^{2s}\right) \\ &\leqslant 2^{s-1}\left\{E\left|\sum_{i=1}^{n} q^2(X_i)\right|^s + cn^s\|q\|^{2s}\right\}. \end{aligned}$$

Therefore it is enough to show that

$$E\left|\sum_{i=1}^{n} q^2(X_i)\right|^s \leqslant c\sum_{u=1}^{s} n^u\|q\|^{2s}.$$

Since $|q| \leqslant 1$. It is easy to see that

$$\begin{aligned} E\left|\sum_{i=1}^{n} q^2(X_i)\right|^s &\leqslant c\sum_{u=1}^{s}\ \sum_{\tau_1+\ldots+\tau_u=s}\ \sum_{1\leqslant i_1<\ldots<i_u\leqslant n} E\{q^{2\tau_1}(X_{i_1}) \cdots q^{2\tau_u}(X_{iu})\} \\ &\leqslant c\sum_{u=1}^{s}\ \sum_{1\leqslant i_1<\ldots<i_u\leqslant n} E\{q^2(X_{i_1}) \cdots q^2(X_{i_u})\}. \end{aligned}$$

The second sum in the middle inequality is over positive integers $\tau_1 \ldots, \tau_u$ such that $\tau_1 + \ldots + \tau_u = s$. The proof of this lemma will be complete once we show that, for any postive integer u, there exists a constant $c(u)$ such that

$$\sum_{1\leqslant i_1<\ldots<i_u\leqslant n} E\{q^2(X_{i_1}) \cdots q^2(X_{i_u})\} \leqslant c(u)\sum_{l=1}^{u} n^l\|q\|^{2l}. \tag{5.16}$$

We prove (5.16) by induction. This bound is obvious for $u = 1$. Now we show that it is true for $u + 1$ if it is true for u.

$$\sum_{1\leqslant i_1<\ldots<i_{u+1}\leqslant n} E\{q^2(X_{i_1}) \cdots q^2(X_{i_{u+1}})\}$$

$$\begin{aligned}
&\leqslant \sum_{1\leqslant i_1<\ldots<i_{u+1}\leqslant n} |E\{q^2(X_{i_1})\cdots q^2(X_{i_{u+1}})\} - E\{q^2(X_{i_1})\cdots \\
&\qquad q^2(X_{i_u})\}E\{q^2(X_{i_{u+1}})\}| \\
&\qquad + \sum_{1\leqslant i_1<\ldots<i_{u+1}\leqslant n} E\{q^2(X_{i_1})\cdots q^2(X_{i_u})\}E\{q^2(X_{i_{u+1}})\} \\
&\leqslant \sum_{1\leqslant i_1<\ldots<i_{u+1}\leqslant n} \varphi(i_{u+1}-i_u)E\{q^2(X_{i_1})\cdots q^2(X_{i_u})\} \\
&\quad + n\|q\|^2 \sum_{1\leqslant i_1<\ldots<i_u\leqslant n} E\{q^2(X_{i_1})\cdots q^2(X_{i_u})\} \\
&\leqslant c\{1+n\|q\|^2\} \sum_{1\leqslant i_1<\ldots<i_u\leqslant n} E\{q^2(X_{i_1})\cdots q^2(X_{i_u})\} \text{ (since } \sum_{i\geqslant 1}\varphi(i)<\infty) \\
&\leqslant c\{1+n\|q\|^2\}c(u)\sum_{l=1}^{u} n^l\|q\|^{2l}.
\end{aligned}$$

The proof of this lemma is now complete.

Proof of lemma 4.3. For part (a) let

$$\xi_{nkt} = \left\{n^{-1}\int (B_{kt}|m_k - m|I)^2\,dF + n^{-2}\right\}^{1/2}.$$

Then by Lemma 3.2

$$\begin{aligned}
&\mathrm{P}\left\{\sup_{k<K_n}\ \sup_{1\leqslant t\leqslant k+v} \frac{|\int B_{kt}(m_k - m)Id(F_n - F)|}{\xi_{nkt}} > n^{\delta}\right\} \\
&\qquad\qquad \leqslant \sum_{k<K_n}\sum_{t=1}^{k+v} n^{-2s\delta}(n\xi_{nkt})^{-2s}\sum_{u=1}^{s}\left\{n\int (B_{kt}|m_k - m|)^2\,IdF\right\}^u \\
&\qquad\qquad \leqslant s\sum_{k\leqslant K_n}\sum_{t=1}^{k+v} n^{-2s\delta} \\
&\qquad\qquad \leqslant cn^{2(1-\alpha)}n^{-2s\delta}.
\end{aligned}$$

The last expression converges to zero for s larger than $(1-\alpha)/\delta$. In this case,

$$\begin{aligned}
\|\boldsymbol{B}_k(m_k - m)Id(F_n - F)\|^2 &= \sum_{t=1}^{k+v}\left\{\int B_{kt}(m_k - m)Id(F_n - F)\right\}^2 \\
&= \mathrm{o_P}(1)\sum_{t=1}^{k+v}(n^{\delta}\xi_{knt})^2 \\
&= \mathrm{o_P}(1)n^{2\delta}\sum_{t=1}^{k+v}\left\{n^{-1}\int (B_{kt}|m_k - m|I)^2\,dF + n^{-2}\right\} \\
&= \mathrm{o_P}(1)n^{2\delta}\left(n^{-1}\int |m_k - m|^2\,IdF + kn^{-2}\right) \\
&\qquad\qquad \text{(since } \sum_{t=1}^{k+v} B_{kt}^2 \leqslant 1)
\end{aligned}$$

$$= \mathrm{o_P}(1)n^{2\delta}(n^{-1}\|m_k - m\|^2 + kn^{-2}).$$

Part (a) is established.

To prove part (b) apply Lemma 4.1.

$$\begin{aligned}\|\boldsymbol{r}_k\| &\leqslant \|\hat{A}_k^{-1} - A_k^{-1}\|\|\boldsymbol{e}_k\| + \|\hat{A}_k^{-1}\|\|\boldsymbol{B}_k(m_k - m)d(F_n - F)\|\\ &= \mathrm{o_P}(k^2\delta_{nk})\mathrm{o_P}(k^{1/2}\delta_{nk}) + \mathrm{O_P}(k)\|\boldsymbol{B}_k(m_k - m)d(F_n - F)\|\\ &= \mathrm{o_P}(k^{5/2}\delta_{nk}^2) + \mathrm{O_P}(k)\|\boldsymbol{B}_k(m_k - m)d(F_n - F)\|.\end{aligned}$$

The result now follows with an application of part (a) of this lemma.

ACKNOWLEDGEMENTS

DN was partly supported by NSF grant DMS-8701426.

REFERENCES

Akaike, H. (1970) Statistical predictor identification. *Ann. Inst. Statist. Math.* 22, 203–17.

Bierens, H. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Am. Statist. Assoc.* 78, 699–707.

Burman, P. (1989) Rate of convergence for the spline estimates for Markov chains. *Stat. Prob. Lett.* 8, 245–53.

——, Chow, E. and Nolan, D. (1990) A cross-validatory method for dependent data. Manuscript.

—— and Chen, K.-W. (1989) Nonparametric estimation of a regression function. *Ann. Statist* 17, 1567–97.

Chu, C. K. (1989) Some results in nonparametric regression. Ph.D. Dissertation, Department of Statistics, University of North Carolina, Chapel Hill.

Collomb, G. and Härdle, W. (1986) Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. *Stoch. Process Appl.* 23, 77–89.

Cook, D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.

Demko, S. (1977) Inverses of band matrices and local convergence of spline projections. *Siam J. Numer. Anal.* 14, 616–19.

de Boor, C. (1978) *A Practical Guide to Splines*. New York: Springer-Verlag.

Geisser, S. (1975) A predictive sample reuse method with applications. *J. Am. Statist. Assoc.* 70, 320–28.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989) *Nonparametric Curve Estimation from Time Series*. New York: Springer-Verlag.

Hall, P. and Heyde, C. C. (1980) *Martingale Limit Theory and its Applications*. New York: Academic Press.

Härdle, W. and Marron, J. S. (1985) Optimal bandwidth selection in non-parametric regression function estimation. *Ann. Statist.* 13, 1465–81.

Robinson, P. M. (1983) Nonparametric estimators for time series. *J. Time Series Anal.* 4, 185–207.

Roussas, G. (1988) A moment inequality of S_{n,k_n} for triangular arrays of random variables under mixing conditions, with applications. *Statistical Theory and Data Analysis II* (ed. K. Matsusita). Amsterdam: North-Holland.

Stone, C. (1985) Additive regression and other nonparametric models. *Ann. Statist.* 13, 689–705.

STONE, M. (1974) Cross-validatory choice and the assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc., Ser B*. 36, 111–33.

TRUONG, Y. and STONE, C. (1989) Nonparametric function estimation involving time series. Manuscript.

YAKOWITZ, S. (1985) Nonparametric density estimation and prediction for Markov sequences. *J. Am. Statist. Assoc.* 80, 215–21.