The Annals of Statistics
1989, Vol. 17, No. 4, 1567–1596

NONPARAMETRIC ESTIMATION OF A REGRESSION FUNCTION

BY PRABIR BURMAN AND KEH-WEI CHEN

Rutgers University and University of North Carolina, Charlotte

A data dependent method of estimating a regression function is proposed here. The model is allowed to be heteroscedastic. Applications to piecewise polynomial, spline, orthogonal series, kernel and nearest neighbor methods are discussed.

1. Introduction. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. on $I \times R$ where I is a compact set in R^p. Let $m(x) = E(Y|X = x)$. Then m is called the regression function of Y on X. In this paper we are interested in estimating m and we allow the model to be heteroscedastic. In the literature one can find many nonparametric methods for estimating m, e.g., kernel, orthogonal series and spline, to name a few. Optimal convergence properties of these estimators have been studied in great detail, e.g., by Stone (1982). It is known that for each of these methods how good an estimator is depends on how smooth the underlying regression function is. Typically, optimal estimation involves an index which depends on the smoothness of m, e.g., the bandwidth for the kernel methods or the number of terms for the orthogonal series method. In practice, however, the amount of smoothness of the unknown regression function is never known. So the problem of finding a data dependent method of choosing the smoothness index is an important task. Fortunately, quite a few data dependent methods already exist in the literature. Chen (1983) used the final prediction error (FPE) method to investigate the properties of the piecewise polynomial estimator in the homoscedastic case. Properties of FPE, AIC, C_p, C_L, cross-validation and generalized cross-validation methods for the homoscedastic case with X_j's as constants have been studied by many authors, e.g., Shibata (1981) and Li (1987). Härdle and Marron (1985) studied the optimality properties for the cross-validated kernel estimator under the assumption that m and f (the marginal density of X) satisfy the Hölder condition. In this paper all we need to assume is that m and f are bounded.

In Section 2, we formulate the problem and state the main result. We discuss piecewise polynomial, spline, orthogonal series, kernel and nearest neighbor methods in Section 3. Section 4 contains the main results, assumptions and some technical lemmas. In Section 5 we present a numerical example and in Section 6 we prove the main results of this paper (i.e., the results of Sections 2 and 4). In Section 7, we prove the optimality results for different types of regression estimates described in Section 3.

2. General method and goal. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. on $I \times R$, where I is a compact subset of R^p. The regression function of Y on X is defined

Received February 1988; revised January 1989.
AMS 1980 *subject classifications.* Primary 62G05; secondary 62J05.
Key words and phrases. Nonparametric regression, data dependent method, asymptotic optimality.

by $m(x) = E(Y|X = x)$. The main object of this paper is to estimate m for the heteroscedastic regression model under the assumption that m is bounded. We will assume that $\sup_x E(Y^{4s}|X = x) < \infty$, for some positive integer $s \geq 2$, and the conditional variance function $\sigma^2(x) = \mathrm{Var}(Y|X = x)$ is bounded above by a positive constant. Let F be the marginal distribution function of X and let F_n be its empirical d.f. Let $\{W_{nk}: k = (k_1, \dots, k_p) \in D_n\}$ be a sequence of weight functions (depending on F) on $I \times I$, where D_n is an index set and $\mathrm{card}(D_n) = K_n$, $K_n/n^s \to 0$. From now on, whenever we write $k \leq K_n$ we mean that $k \in D_n$.

(A1) $\int W_{nk_n}(x, x', F)\, dF(x') \to 1$ for each x and for any sequence $k_n = (k_{in}, i = 1, \dots, p) \to \infty$, i.e., $k_{in} \to \infty$ for $i = 1, \dots, p$.

(A2) If $m_{nk} = \int m(x') W_{nk}(x, x', F)\, dF(x')$, then $\int (m_{nk_n} - m)^2\, dF \to 0$ for any $k_n \to \infty$.

(A3) $m_{nk} \neq m$ a.e. F for any k and n and $\sup_{n,k} \sup_x |m_{nk}(x)| < \infty$.

Notation and conventions. For functions g_1 and g_2 let $\langle g_1, g_2 \rangle = \int g_1 g_2\, dF$ and $\langle g_1, g_2 \rangle_n = \int g_1 g_2\, dF_n$ and $\|g_1\|^2 = \int g_1^2\, dF$, $\|g_1^2\|_n^2 = \int g_1^2\, dF_n$. Let $\{\xi_{\eta\alpha}\}$ be a class of random variables indexed by α and $\{\eta_{n\alpha}\}$ be a class of positive random variables or a sequence of positive constants. Then by $\xi_{\eta\alpha} = O_p(\eta_{n\alpha})$, we mean $\sup_\alpha |\xi_{\eta\alpha}/\eta_{n\alpha}| = O_p(1)$. Similarly we can define $\xi_{\eta\alpha} = o_p(\eta_{n\alpha})$. For notational convenience we will write $W_k(\cdot,\cdot,\cdot)$, $m_k(\cdot)$ instead of $W_{nk}(\cdot,\cdot,\cdot)$, $m_{nk}(\cdot)$. For $x = (x_1, \dots, x_t) \in R^t$, let us denote $|x| = |x_1| + \cdots + |x_t|$.

From the above sequence of weight functions we can construct a sequence of estimates $\hat{m}_k(x) = \sum Y_j W_k(x, X_j, F_n)/n$, $k \leq K_n$. How large K_n is depends on the type of weight function and hence we postpone this question to Section 3. The criterion for performance of our estimator is given by how good a predictor it is. Let (X_{n+1}, Y_{n+1}) be an independent copy of (X_i, Y_i), $i = 1, \dots, n$. Hence the prediction error is

$$E\Big(\big(Y_{n+1} - \hat{m}_k(X_{n+1})\big)^2 \Big| X_i, Y_i,\, i = 1, \dots, n\Big) = \int \sigma^2\, dF + L_n(k)$$

where $L_n(k) = \|\hat{m}_k - m\|^2$. This tells us that we should look for the index $\tilde{k}$ at which $L_n(k)$ attains its minimum value. Unfortunately, $\tilde{k}$ depends on the unknown distribution of (X, Y). So we try a different approach. Heuristically,

$$L_n(k) \simeq \|\hat{m}_k - m\|_n^2 = \|m\|_n^2 + \|\hat{m}_k\|_n^2 - 2\langle \hat{m}_k, m \rangle_n.$$

If we estimate $\langle \hat{m}_k, m \rangle_n$ by $n^{-1}\sum Y_j \hat{m}_k(X_j)$, the bias given $X_1, \dots, X_n$ is $n^{-1}\int \sigma^2(x) W_k(x, x, F_n)\, dF_n(x)$. This bias can be large and needs to be removed [e.g., it is $O_p(k/n)$ if a polynomial of degree k has been fitted]. An estimate of this is $n^{-2}\sum \hat{\varepsilon}_{kj}^2 W_k(X_j, X_j, F_n)$, where $\hat{\varepsilon}_{kj} = Y_j - \hat{m}_k(X_j)$. So an estimate of $\langle \hat{m}_k, m \rangle_n$ is given by $n^{-1}\sum Y_j \hat{m}_k(X_j) - n^{-2}\sum \hat{\varepsilon}_{kj}^2 W_k(X_j, X_j, F_n)$ (see Remark 2.4). Let

$$\begin{aligned}\hat{L}_n(k) &= \|m\|_n^2 + \|\hat{m}_k\|_n^2 - 2n^{-1}\sum Y_j \hat{m}_k(X_j) + 2n^{-2}\sum \hat{\varepsilon}_{kj}^2 W_k(X_j, X_j, F_n) \\ &= \|m\|_n^2 - n^{-1}\sum Y_j^2 + n^{-1}\sum \hat{\varepsilon}_{kj}^2\big[1 + 2n^{-1} W_k(X_j, X_j, F_n)\big].\end{aligned}$$

Then $\hat{L}_n(k)$ is a good estimate of $L_n(k)$. Let

$$\hat{T}_n(k) = n^{-2}\sum \hat{\varepsilon}_{kj}^2\left[1 + 2n^{-1}W_k(X_j, X_j, F_n)\right] \quad \text{and} \quad \hat{T}_n(\hat{k}) = \inf_{k \le K_n} \hat{T}_n(k).$$

Then we use $\hat{m}_{\hat{k}}$ as our estimate of the unknown regression function. The fact that $\hat{m}_{\hat{k}}$ is a good choice is shown by the results (see Section 4)

$$L_n(\hat{k})/L_n(\tilde{k}) \to_P 1 \quad \text{and} \quad V_n(\hat{k})/V_n(k^*) \to_P 1, \tag{2.1}$$

where

$$\begin{aligned} V_n(k) &= E(L_n(k)|X_1, \dots, X_n) \\ &= n^{-1}\int \sigma^2(x)W_k^2(x, x, F_n)\, dF_n(x) + \|\tilde{m}_k - m\|^2, \\ \tilde{m}_k(x) &= E(\hat{m}_k(x)|X_1, \dots, X_n) \text{ and } V_n(k^*) = \inf_{k \le K_n} V_n(k). \end{aligned}$$

REMARK 2.1. In some cases our selection rule may involve less computation than ordinary cross-validation, e.g., the nearest neighbor method (see Section 3).

REMARK 2.2. $L_n(k) = \|\hat{m}_k - m\|^2$ should be preferred to the loss $\Sigma(\hat{m}_k(X_i) - m(X_i))^2/n$ on the ground that the latter measures the discrepancy between $\hat{m}_k$ and m only on the points $X_1, \dots, X_n$, whereas, $L_n(k)$ is the overall loss. However, they can be shown to be equivalent in many cases [see Marron and Härdle (1986)].

REMARK 2.3. It is possible to choose our loss function to be $\int(\hat{m}_k - m)^2 h(x)\, dF(x)$ for some nonnegative weight function h. Indeed, such a loss function with an h vanishing at the boundaries will help us to get rid of the problem of boundary effects associated with the kernel and nearest neighbor methods. It is not hard to see what the corresponding $\hat{L}_n$ should be and all the results of this paper will remain valid. Since we can always choose a weight function h which is 1 on the compact set I and 0 otherwise, it is not necessary to assume that X is in a compact set if our goal is to estimate m on I.

REMARK 2.4. For the purpose of discussion here, let $w_{ij} = W_k(X_i, X_j, F_n)$, $\tilde{w}_{ij} = w_{ij}^2/(1 - 2n^{-1}w_{ii})$ and $w_{kj}^* = \{w_{jj} - n^{-2}\Sigma_{l=1}^n \tilde{w}_{lj} w_{ll}\}/(1 - 2n^{-1}w_{jj})$. Let us note that $n^{-2}\Sigma\hat{\varepsilon}_{kj}^2 w_{kj}^*$ is an estimate of $n^{-1}\int\sigma^2(x)W_k(x, x, F_n)\, dF_n(x)$ and it can be shown that this estimate has a smaller bias than the one introduced before. This gives rise to the following estimate of $L_n(k)$; $\hat{L}_n(k) = \|m\|_n^2 = n^{-1}\Sigma Y_j^2 + n^{-1}\Sigma\hat{\varepsilon}_{kj}^2[1 + 2n^{-1}w_{kj}^*]$. Another estimate of $n^{-1}\int\sigma^2(x)W_k(x, x, F_n)\, dF_n$ is given by $n^{-2}\Sigma Y_j \hat{\varepsilon}_{kj} W_k^\dagger(X_j, X_j, F_n)$, where $W_k^\dagger = W_k(1 + n^{-1}W_k)$. It is easy to see what the corresponding estimates of $\langle m_k, m\rangle_n$ and $\hat{L}_n(k)$ should be and the results of this paper remain valid.

3. Applications. For all the applications to be considered here we assume that the marginal density of X satisfies $0 < b \le f(x) \le B$ for all $x \in I$, for some $0 < b < B$ and $I = [0,1]^p$. Let us note that even if X is not assumed to be valued in a compact set, this assumption on the density can be made without loss of generality since we wish to estimate m only on compact sets (see Remark 2.3). Let us recall here that the criterion function is $T_n(k) = n^{-1}\Sigma\hat{\varepsilon}_{k_j}^2[1 + 2n^{-1}W_k(X_j, X_j, F_n)]$ and $\hat{k}$ is the index at which $T_n(k)$ is minimized, i.e., $T_n(\hat{k}) = \inf_{k \le K_n} T_n(k)$. We will call $\hat{m}_{\hat{k}}$ optimal if (2.1) holds. From now on, for notational simplicity we will write $W_k(x, x', F)$ and $W_k(x, x', F_n)$ as $W_k(x, x')$ and $\overline{W}_k(x, x')$, respectively.

3.1. *Least squares method.* Let $\psi_k(x)$ be a λ_k dimensional vector of functions. Here we get $\hat{m}_k$ as a linear combination of these λ_k functions by the least squares fitting. For all the applicatiions we have in mind, $c_1(k_1 \cdots k_p) \le \lambda_k \le c_2(k_1 \cdots k_p)$ for positive constants c_1 and c_2. Let K_n be such that $K_n/n^{1-\alpha} \to 0$ for some $\alpha > 0$. Here the index set is of the form $D_n = \{k\colon \lambda_k \le K_n\}$. Let

$$A_k = \int \psi_k\psi_k'\,dF, \qquad d_k = \int m\psi_k\,dF, \qquad W_k(x, x') = \psi_k'(x)A_k^{-1}\psi_k(x'), \tag{3.1a}$$

$$\hat{A}_k = \int \psi_k\psi_k'\,dF_n, \qquad \hat{d}_k = \sum Y_j\psi_k(X_j)/n, \qquad \overline{W}_k(x, x') = \psi_k'(x)\hat{A}_k^{-1}\psi_k(x'). \tag{3.1b}$$

The regression and the criterion functions are

$$\hat{m}_k(x) = \hat{d}_k'\hat{A}_k^{-1}\psi_k(x) \quad \text{and} \qquad T_n(k) = n^{-1}\sum\hat{\varepsilon}_{kj}^2\big[1 + 2n^{-1}\psi_k'(X_j)\hat{A}_k^{-1}\psi_k(X_j)\big]. \tag{3.2}$$

THEOREM 3.1. *Let us assume that* (A2) *and* (A3) *hold. Let* $\delta_{nk} = \lambda_k^{-1/2}n^{-(1-\delta)/2}$ *for some* $0 < \delta < \alpha/2$. *Then* $\hat{m}_{\hat{k}}$ *is optimal if:*

(i) *All the eigenvalues of* A_k *are between* $c_3\lambda_k^{-1}$ *and* $c_4\lambda_k^{-1}$ *for some positive constants* c_3 *and* c_4.
(ii) $\|\hat{A}_k - A_k\| = o_p(\delta_{nk})$, *where* $\|\cdot\|$ *is the matrix norm.*

Piecewise polynomials with fixed partitions [Chen (1983)]. For each $k = (k_1, \dots, k_p)$, let $\{C_{kt}\colon t = (t_1, \dots, t_p),\ 1 \le t_i \le k_i,\ i = 1, \dots, p\}$ be a collection of disjoint rectangles covering I such that lengths of the sides of each C_{kt} are $(k_1^{-1}, \dots, k_p^{-1})$. On each C_{kt} we fit a polynomial of a prescribed degree (say v) by the method of least squares, i.e., on each C_{kt} we fit a linear combination of the functions ϕ_{ktu} by the method of least squares, where $\phi_{ktu}(x) = \prod_{i=1}^p x_i^{u_i}$, $x = (x_1, \dots, x_p)$, $u = (u_1, \dots, u_p)$ and $|u| = u_1 + \cdots + u_p \le v$. Here, $\lambda_k = v_1k_1 \cdots k_p$ where $v_1 = \sum_{j=0}^{v}\binom{p-1+j}{j}$. Let $\phi_{kt}(x)$ be the vector of functions ϕ_{ktu}, $|u| \le v$, $I_{kt}(x)$ the indicator function of C_{kt} and

$$A_{kt} = \int I_{kt}(x)\phi_{kt}(x)\phi_{kt}'(x)\,dF(x), \qquad \hat{A}_{kt} = \int I_{kt}(x)\phi_{kt}(x)\phi_{kt}'(x)\,dF_n(x),$$

$$d_{kt} = \int I_{kt}m\phi_{kt}\,dF, \qquad \hat{d}_{kt} = \sum_j I_{kt}(X_j)\phi_{kt}(X_j)Y_j/n.$$

Here

$$W_k(x, x') = \sum_t I_{kt}(x) I_{kt}(x') \phi_{kt}(x) A_{kt}^{-1} \phi_{kt}(x'),$$

$$\overline{W}_k(x, x') = \sum_t I_{kt}(x) I_{kt}(x') \phi'_{kt}(x) \hat{A}_{kt}^{-1} \phi_{kt}(x').$$

The regression estimate and the criterion function are

$$\hat{m}_k(x) = \sum_t I_{kt}(x) \phi'_{kt}(x) \hat{A}_{kt}^{-1} \hat{d}_{kt} \quad \text{and}$$

$$T_n(k) = n^{-1} \sum \hat{\varepsilon}_{kj}^2 \left[1 + 2n^{-1} \sum_t \phi'_{kt}(X_j) \hat{A}_{kt}^{-1} \phi_{kt}(X_j) \right].$$

In this case, ψ_k is the vector of the functions $\phi_{ktu} I_{kt}$.

Piecewise polynomials with random partitions. From a practical point of view, the fixed partition case sometimes may have a serious problem. It may happen that one or more of the rectangles may have very few or no observations at all. Random partitioning avoids this problem and is carried out as follows.

For variable X_i, divide $[0,1]$ into k_i subintervals $[\zeta_{i,j-1}, \zeta_{i,j})$, where $\zeta_{i,0} = 0$, $\zeta_{i,k_i} = 1$ and ζ_{ij} is the (j/k_i)th sample quantile for the variable X_i. If this is done for $1 \le i \le p$, it gives rise to a partition $\{C_{kt}\}$ of $[0,1]^p$. The rest including the criterion function $T_n(k)$ remain the same as in the case of fixed partitioning.

THEOREM 3.2. *Let us assume that m is not piecewise polynomial. For the piecewise polynomial regression method (with fixed or random partitioning), $\hat{m}_{\hat{k}}$ is optimal if m is bounded.*

From now on, for technical reasons we will write

$$\phi_{ktu}(x) = \prod_{i=1}^{p} x_i^{u_i} (x_i - x_{kti})^{u_i},$$

where $x_{kt} = (x_{kt1}, \dots, x_{ktp})$ is the middle point of C_{kt}. We note that with these ϕ_{ktu}'s we get exactly the same estimates as before.

Spline regression: Equispaced knots. Fore each integer l, let Sp_l be the class of all functions s on $[0,1]$ such that for each $i = 1, \dots, l$, s is a polynomial of degree v on $[(i-1)l^{-1}, il^{-1}]$ and is $(v-1)$ times continuously differentiable on $[0,1]$. Sp_l is called the class of splines of degree v on $[0,1]$ with knot spacing l^{-1}. Let $\{B_{l,j}\colon j = 1, \dots, l+v\}$ be a basis of Sp_l consisting of normalized B-splines of degree v (with knot spacing l^{-1}) [see de Boor (1978)]. Let $\{B_{k,t}\colon t = (t_1, \dots, t_p),\ 1 \le t_i \le k_i + v\}$ be the class of multivariate normalized B-splines on I defined as the product of univariate normalized B-splines, i.e., $B_{k,t}(x) = \prod_{i=1}^{p} B_{k_i, t_i}(x_i)$. Here $\lambda_k = \prod_{i=1}^{p}(k_i + v)$. Let ψ_k be the vector of B_{kt}'s. The regression function $\hat{m}_k$ and the criterion function $T_n(k)$ are of the form given in (3.2).

Spline regression: Random knots. As in the case of piecewise polynomials, fitting spline regression with equispaced knots may also run into trouble. From a practical point of view it is advisable to place the knots at the quantiles. For variable X_i, we could place the knots at $0 = \zeta_{i0} < \zeta_{i1} < \zeta_{i2} < \cdots < \zeta_{ik_i} = 1$, where ζ_{ij} is the (j/k_i)th sample quantile for variable X_i. Apart from the placement of the knots, everything else [including the criterion function $T_n(k)$] remains the same as in the fixed knots.

THEOREM 3.3. *Let us assume that m is not a spline of degree v. For spline regression with equispaced or random knots, $\hat{m}_{\hat{k}}$ is optimal if m is bounded.*

REMARK 3.1. For the spline regression with random knots we should write

$$W_k(x, x') = \psi_k'(x) A_k^{-1} \psi_k(x') \quad \text{and} \quad \overline{W}_k(x, x') = \hat{\psi}_k'(x) \hat{A}_k^{-1} \hat{\psi}_k(x'),$$

where $\hat{\psi}_k'$ is the vector of B-splines with random knots. However, all the proofs go through with some modification in arguments. These same comments also apply for the piecewise polynomial method with random partitioning. We will discuss spline regression with random knots in a forthcoming paper.

Orthogonal series method. There are essentially two methods of constructing regression functions using the orthogonal series method. The first is a variant of the kernel method and hence no special treatment is given to it.

1. Let $\{\phi_t\}$ be an orthonormal system on I. For convenience we will write the system as $\{\phi_1, \phi_2, \dots\}$. The weight function here is defined as

$$W_k(x, x') = \sum_{j \le \lambda_k} \phi_j(x)\phi_j(x') \Big/ \sum_{j \le \lambda_k} \phi_j(x)\theta_j,$$

where $\theta_j = \int \phi_j \, dF$ or, in order to avoid possible zeros in the denominator, we could write the other form as

$$W_k(x, x') = \sum_{j \le \lambda_k} a_{kj}\phi_j(x)\phi_j(x') \Big/ \sum_{j \le \lambda_k} a_{kj}\phi_j(x)\theta_j,$$

where $\{a_{kj}\colon j \le \lambda_k\}$ is a sequence of constants chosen so that $\lim_k a_{kj} = 0$ for each j and $\sum_{j \le \lambda_k} a_{kj}\phi_j(x)\phi_j(x') \ge 0$ for all x and x'. If trigonometric polynomials are taken with $a_{k,j} = 1$ or a_{kj}'s as the Fejer weights, then $W_k(x, x')$ looks very much like a kernel.

2. As in method 1 we will write the orthonormal system as $\{\phi_1, \phi_2, \dots\}$. We will present this case mainly for $p = 1$ and trigonometric polynomials even though our results are written in a general form. Let us assume that ϕ_j's are uniformly bounded by a constant, say 1. Let $\psi_k(x)$ be the column vector of the functions $\phi_t(x)$, $t \le \lambda_k$ ($\lambda_k = 2k + 1$ for trigonometric polynomials and $\lambda_k = k$ for Legendre polynomials). The regression function and the criterion function are of the form given in (3.2). Different orthogonal series require different types of assumptions on m to ensure that $\sup_k \sup_x |\hat{m}_k(x)| < \infty$ [see Jackson (1930)].

THEOREM 3.4. *Let us assume that m cannot be written as a linear combination of the finite number of* ϕ_j's. $\hat{m}_{\hat{k}}$ *is optimal if m is*:

(i) *Absolutely continuous when trigonometric polynomials are used.*
(ii) *Differentiable when Legendre polynomials are used.*

For technical reasons, from now on we will denote the vector of functions $\lambda_k^{-1/2}\phi_j$, $j = 1,\dots,\lambda_k$, by ψ_k. We note this change does not affect the estimates given above.

3.2. *Kernel method.* Let w be a bounded nonnegative density on R^p vanishing outside a compact set containing the origin and satisfying a Hölder condition, i.e., $|w(z_1) - w(z_2)| \le T\|z_1 - z_2\|^\beta$ for all z_1 and z_2, for some $T > 0$ and $\beta > 0$. Let us assume that $\sup_x E(Y^{2u}|X = x) < \infty$ for all $u \ge 1$. Let $f_h(x) = \lambda\int w(h(x - x'))\, dF(x')$ and $\hat{f}_h(x) = \lambda\int w(h(x - x'))\, dF_n(x')$, where $h = (h_1,\dots,h_p)$, $\lambda = h_1 \cdots h_p$ and for any $z \in R^p$, $hz = (h_1z_1,\dots,h_pz_p)$. Here we allow the bandwidths (i.e., h_i's) to vary continuously in the interval $[n^\alpha, n^{1-\alpha}]$ for any small $\alpha > 0$, so that $n^\alpha \le \prod_{i=1}^p h_i \le n^{1-\alpha}$. Let $C = \{h = (h_1,\dots,h_p)$: $n^\alpha \le h_i \le n^{1-\alpha}$, $n^\alpha \le \prod_{i=1}^p h_i \le n^{1-\alpha}\}$. Let $W_h(x, x') = \lambda w(h\ (x - x'))/f_h(x)$, $\overline{W}_h(x, x') = \lambda w(h(x - x'))/\hat{f}_h(x)$ and $m_h(x) = \int m(x')W_h(x, x')\, dF(x')$. Here

$$\hat{m}_h(x) = (\lambda/n)\sum Y_j w\big(h(x - X_j)\big)/\hat{f}_h(x),$$

$$T_n(h) = n^{-1}\sum \hat{\varepsilon}_{kj}^2\big[1 + 2n^{-1}\lambda w(0)\hat{f}^{-1}(X_j)\big]$$

and $\inf_{h\in C} T_n(h) = T_n(\hat{h})$. As usual, $L_n(h) = \|\hat{m}_h - m\|^2$ and $V_n(h) = E(L_n(h)|X_1,\dots,X_n)$.

THEOREM 3.5. *Let us assume that* $m \ne m_h$ *a.e. for any h. If m is bounded, then*

$$L_n(\hat{h})/\inf_{h\in C} L_n(h) \to_P 1 \quad \textit{and} \quad V_n(\hat{h})/\inf_{h\in C} V_n(h) \to_P 1.$$

To use our main result of Section 4, we will prove Theorem 3.5 for a discrete choice of bandwidths with $h_{k_i} = n^\alpha + (k_i - 1)n^{-\beta^{-1}}$, $k_i = 1,\dots,K_n = n^{1-\alpha+\beta^{-1}}$, $i = 1,\dots,p$ and show that this discrete choice approximates the continuous case (Lemma 3.7).

Abusing some notation we will write W_{h_k}, m_{h_k}, $\hat{m}_{h_k}$, $L_n(h_k)$, etc., as W_k, m_k, $\hat{m}_k$, $L_n(k)$, etc. Lemma 3.6 will be needed very often. Let

$$\varepsilon_{nk} = (\log n)^{1/2+\delta}/\sqrt{\lambda_k n}\,,\ \overline{m}_k(x) = \int m(x')W_k(x, x')\, dF_n(x')$$

and

$$m_k(x) = \int m(x')W_k(x, x')\, dF(x'),$$

where $\lambda_k = h_{k_1} \cdots h_{k_p}$.

LEMMA 3.6. $\sup_x|\hat{f}_k(x) - f_k(x)| = o_p(\lambda_k \varepsilon_{nk})$ *and* $\sup_x|\overline{m}_k(x) - m_k(x)| = o_p(\lambda_k \varepsilon_{nk})$.

The next lemma tells us that the discrete case allows us continuous bandwidth selection. Let $C_k = \prod_{i=1}^{p}[h_{k_i - 1}, h_{k_i}]$, for $k = (k_1, \dots, k_p)$.

LEMMA 3.7.

(i) $$\sup_k \sup_{h \in C_k} |V_n(h) - V_n(k)|/V_n(k) \to_P 0.$$

(ii) $$\sup_k \sup_{h \in C_k} |L_n(h) - V_n(k)|/V_n(k) \to_P 0.$$

(iii) $$\sup_k \sup_{h \in C_k} |\hat{L}_n(h) - \hat{L}_n(k)|/V_n(k) \to_P 0.$$

REMARK 3.2. It is possible to relax the condition that w is nonnegative if we can choose the bandwidths in such a way that $\inf_h \inf_x f_h(x) > 0$, which is in fact enough for all the proofs to go through. Consequently, all we need to assume is that $w(0) > 0$.

REMARK 3.3. We note that the conditions imposed on w here are not true for the uniform kernel. However, calculations similar to the ones done here will hold for the case of the uniform kernel with slight modifications and hence no special treatment is given for it.

REMARK 3.4. We note that if we do not assume the existence of all the conditional moments of Y given X, α will depend on s in a complicated manner, where $4s$ is the largest moment of Y given X assumed to exist.

REMARK 3.5. Härdle and Marron (1985) assume that both m and f satisfy Hölder conditions. Here all we need is that m and f are bounded.

3.3. *Nearest neighbor method.* For an integer k, the nearest neighbor estimate at point x is defined as the weighted average of those Y_i's for which the X_i's are the k nearest observations to x. Now let us describe this method mathematically. Let w be a function on $[0,1]$, $w(0) > 0$ and $\int w(t)\,dt = 1$. For any two points x_1 and x_2 in R^p, let $\|x_1 - x_2\|$ be the usual Euclidean distance (or it could be the distance $\sum|x_{1i} - x_{2i}|$). Let $N(x, r) = \{x'\colon \|x - x'\| \le r\}$. For $1 \le k \le n^{1-\alpha}$, for some small $\alpha > 0$, let

$$\overline{W}_k(x, x') = \lambda_k w(\lambda_k G_n(x\colon \|x - x'\|))/\nu_k,$$

$$W_k(x, x') = \lambda_k w(\lambda_k G(x, \|x - x'\|))/\nu_k,$$

where $\lambda_k = n/k$, $G_n(x, r) = F_n(N(x, r))$, $G(x, r) = F(N(x, r))$ and $\nu_k = k^{-1}\sum_{j=1}^{k} w(j/k)$. The nearest neighbor regression estimate and the criterion function are

$$\hat{m}_k(x) = n^{-1}\sum Y_j \overline{W}_k(x, X_j) = k^{-1}\sum Y_j w\big(\lambda_k G_n(x, \|x - X_j\|)\big)/\nu_k,$$

$$T_n(k) = n^{-1}\sum \hat{\varepsilon}_{kj}^2\big[1 + 2k^{-1}w(k^{-1})/\nu_k\big].$$

We allow k to vary between 1 and $n^{1-\alpha}$. We will assume here that $E(Y^{4s})$ exists for all s. In order to make the proofs simple we will assume that either w is uniform or it is linear or it is differentiable on (0, 1) with the derivative satisfying a Hölder condition. Now let us write down the main result. Let $m_k(x) = \int m(x')W_k(x, x')\,dF(x')$

THEOREM 3.8. *Let us assume that $m_k \neq m$ a.e. for any k. For the nearest neighbor method, $\hat{m}_{\hat{k}}$ is optimal if m is bounded.*

The following lemma is very important.

LEMMA 3.9. *Let $\alpha_n = (\log n)^{1/2+\delta}/\sqrt{n}$ for some $\delta > 0$. Let $I(x, r)$ be the indicator function $I(G(x, r) \le \alpha_n)$. Then the following is true uniformly for all $x \in I$ and $r > 0$:*

$$G_n(x, r) - G(x, r) = o_p(\alpha_n^2)I(x, r) + o_p(\alpha_n)\sqrt{G(x, r)}\,(1 - I(x, r)).$$

The proof of this lemma follows from the results given in Breiman, Friedman, Olshen and Stone (1984).

REMARK 3.6. Let us note that the nearest neighbor case is slightly different from the others in the sense that we have only one smoothing parameter here. However the general theory presented in Sections 2 and 4 holds here without any modification.

REMARK 3.7. In order to be consistent with (A2) we should write $\lambda_k = n/(n + 1 - k)$. However, this change does not alter the proof in any way.

4. Main results.

4.1. *The main results under general assumptions.* The results of this section are very general. We will prove the main results under some general assumptions which will be given later in this section. In order to make the proofs simpler, we will assume that the variance function σ^2 is bounded away from zero.

Let us recall $\tilde{m}_k(x) = E(\hat{m}_k(x)|\underline{X}_n) = \int m(x')\overline{W}_k(x, x')\,dF_n(x')$ and $\overline{m}_k(x) = \int m(x')W_k(x, x')\,dF_n(x')$, where $\underline{X}_n = (X_1, \dots, X_n)$. Let us now state the main results.

LEMMA 4.1. *Under* (A1)–(A9),

$$\sup_{k \le K_n} \left| \frac{L_n(k)}{V_n(k)} - 1 \right| \to_P 0.$$

The following is an immediate consequence of Lemma 4.1.

COROLLARY 4.2. *Under* (A1)–(A9), $V_n(k^*)/L_n(\tilde{k}) \to_P 1$.

LEMMA 4.3. *Under* (A1)–(A9),

$$\sup_{k \le K_n} |\hat{L}_n(k) - V_n(k) - (\hat{L}_n(k^*) - V_n(k^*))|/V_n(k) \to_P 0.$$

Because of Lemma 4.3, we get for any $\varepsilon > 0$,

$$\hat{L}_n(k) - \hat{L}_n(k^*) \ge (1 - \varepsilon)V_n(k) - V_n(k^*)$$

for all $k \le K_n$ with probability tending to 1. Since $\hat{L}_n(\hat{k}) - \hat{L}_n(k^*) \le 0$, $V_n(k^*)/V_n(\hat{k}) \ge 1 - \varepsilon$ and hence $1 \ge V_n(k^*)/V_n(\hat{k}) \le 1 - \varepsilon$ with probability tending to 1. Since $\varepsilon > 0$ is arbitrary and because of Corollary 4.2 we conclude:

THEOREM 4.4. *Under* (A1)–(A9),

$$L_n(\hat{k})/L_n(\tilde{k}) \to_P 1 \quad \textit{and} \quad V_n(\hat{k})/V_n(k^*) \to_P 1.$$

REMARK 4.1. If it is assumed that $\sup_x E((Y - m(x))^{2u}|X = x) \le \zeta^u$ for all $u \ge 1$, for some $\zeta > 0$, then all the statements in Lemmas 4.1 and 4.3 and Theorem 4.4 could be proved "with probability 1" instead of "in probability" by repeated use of Borel–Cantelli lemma.

REMARK 4.2. A look at the proofs of Lemmas 4.1 and 4.3 will tell us that all the results of this paper are valid if X_i's are constants. In that case the loss is taken to be $n^{-1}\Sigma(\hat{m}_k(X_i) - m(X_i))^2$. We would like to point out that the case for random X_i's involves a lot more mathematical complications than the case for constant X_i's.

Now let us write down the conditions needed to prove Theorem 4.4. Let us recall that $\sup_x E(Y^{4s}|X = x) < \infty$ for some $s \ge 2$.

(A4) W_k's depend on h_{nk}'s, where $h_{nk} = (h_{nk_1}, \dots, h_{nk_p})$. Let $\lambda_{nk} = h_{nk_1} \cdots h_{nk_p}$. For each n, $\{h_{nj}\}$ is increasing in j and

(i) $\sup_n \sum_{j=1}^{\infty} h_{nj}^{-s} < \infty$,

(ii) $\sup_{k \le K_n} \lambda_{nk}/n \to 0$.

The sequences $\{h_{nj}\}$ and $\{\lambda_{nk}\}$ arise quite naturally (see Section 3).

(A5) (i) $\inf_{k \le K_n} \lambda_{nk}^{-1} \int \overline{W}_k^2(x, x')\, dF_n(x')\, dF(x)$ is bounded below and above by two positive constants (independent of k) with probability tending to 1.

(ii) $\sup_{k \le K_n} |\lambda_{nk}^{-1} \int \sigma^2(x') \overline{W}_k^2(x, x')\, dF_n(x')\, d(F_n - F)(x)| = o_p(1)$.

(iii) $\sup_{k \le K_n} \sup_x \lambda_{nk}^{-1} |\overline{W}_k(x, x)| = O_p(1)$.

Let $a_k(x_1, x_2) = \int \overline{W}_k(x_1, x) \overline{W}_k(x_2, x)\, dF_n(x)$.

(A6) (i) $\sup_{k \le K_n} \lambda_{nk}^{-1} \int a_k^2(x_1, x_2)\, dF(x_1)\, dF(x_2) = O_p(1)$.

(ii) $\sup_{k \le K_n} \lambda_{nk}^{-1} \int a_k^2(x_1, x_2)\, dF_n(x_1)\, dF_n(x_2) = O_p(1)$.

(A7) (i) $\sup_x \int |\overline{W}_k(x, x_1)|\, dF_n(x_1) = O_p(1)$.

(ii) $\sup_{x_1} \int |\overline{W}_k(x, x_1)|\, dF_n(x) = O_p(1)$.

(iii) $\sup_{x_1} \int |\overline{W}(x, x_1)|\, dF(x) = O_p(1)$.

Let

$$\overline{V}_n(k) = n^{-1} \int \sigma^2(x') \overline{W}_k^2(x, x')\, dF_n(x')\, dF(x) + \|\tilde{m}_k - m_k\|^2 + \|m_k - m\|^2.$$

(A8) $\sup_{k \le K_n} |V_n(k) - \overline{V}_n(k)| / \overline{V}_n(k) \to_P 0$.

(A9) $\int (\tilde{m}_k - m)^2\, d(F_n - F) = o_p(V_n(k))$.

4.2. *Some technical results.* In this section we will write down a few important technical results. From now on we will write h_j and λ_k instead of h_{nj} and λ_{nk}.

The following result is essentially a consequence of the assumption, $m_k \ne m$ a.e. F for any k [(A3)]. Similar results appear in Li (1987) and Shibata (1981).

LEMMA 4.5. *Under* (A3), (A4), (A5) *and* (A8),

(i) (a) $\inf_{k \le K_n} n\overline{V}_n(k) \to_P \infty$,

(b) $\inf_{k \le K_n} nV_n(k) \to_P \infty$.

(ii) (a) $\sum_{k \le K_n} \left(n\overline{V}_n(k)\right)^{-s} \to_P 0$,

(b) $\sum_{k \le K_n} \left(nV_n(k)\right)^{-s} \to_P 0$.

Let α_n be the same as in Lemma 3.9. The following useful result follows from Hoeffding's inequality. For a proof see Breiman, Friedman, Olshen and Stone (1984).

LEMMA 4.6. *Let* $\{\phi_{nu}\colon u \in D_n\}$ *be a collection of functions uniformly bounded in absolute value by* 1. *Let* t_n *be the cardinality of* D_n *and* $\sigma_{nu}^2 = \mathrm{Var}(\phi_{nu}(x))$. *If*

$t_n/n^a \to 0$ for some $a > 0$, then

$$\int \phi_{nu} d(F_n - F) = o_p(\alpha_n^2) I_{nu} + o_p(\alpha_n \sigma_{nu})(1 - I_{nu}),$$

where $I_{nu} = 1$ or 0 according as $\sigma_{nu} \le \alpha_n$ or $\sigma_{nu} > \alpha_n$.

The following result due to Whittle (1960) will be used repeatedly.

THEOREM 4.7. *Let Z_j's, $j = 1, \dots, n$, be independent random variables with zero means and $EZ_j^{4s} < \infty$ for $j = 1, \dots, n$. Let $\gamma_j(u) = \{E|Z_j|^u\}^{1/u}$. Then for any sequence of real numbers $\{b_j: 1 \le j \le n\}$ and $\{a_{ij}: 1 \le i, j \le n\}$, there exist constants $c_1(s) > 0$ and $c_2(s) > 0$ such that*

(i) $$E\left(\sum b_j Z_j\right)^{2s} \le c_1(s)\left(\sum b_j^2 \gamma_j^2(2s)\right)^s,$$

(ii) $$E\left(\sum a_{ij} Z_i Z_j - E\sum a_{ij} Z_i Z_j\right)^{2s} \le c_2(s)\left(\sum a_{ij}^2 \gamma_i^2(2s) \gamma_j^2(2s)\right)^s.$$

The following lemma is simple to prove, but harder to describe. For $l \le t$, l and t positive integers, let

$$Z(t, l) = \{(i_1, \dots, i_t): 1 \le i_1, \dots, i_t \le l,\ i_1, \dots, i_t \text{ are integers and } (i_1, \dots, i_t) \text{ has } l \text{ distinct values}\}$$

and, for any $\xi \in Z(t, l)$, let $a(\xi)$ equal the number of indices in ξ appearing only once.

LEMMA 4.8. *Let ϕ be a bounded function I^t. Then there exists a constant c (depending only on t) such that*

$$\left| E \int \phi(x_1, \dots, x_t) \prod_{i=1}^{t} d(F_n - F)(x_i) \right| \le c \sum_{l=1}^{t} n^{-t+l} \sum_{\xi \in Z(t, l)} n^{-[(a(\xi)+1)/2]} \int |\phi(x_\xi)|\, dF(x_1) \cdots dF(x_l),$$

where $[v]$ is the largest integer not exceeding v.

5. A numerical example. In order to show how our model selection method works we present a numerical example here. The data are taken from the book by Mardia, Kent and Bibby (1979). The data contain the analysis and statistics scores of 88 students. We treat the statistics score as the Y-variable and use the kernel method to fit a regression curve. The kernel we have used is $w(z) = 1.5(1 - |z|)^2$, $|z| \le 1$. Since the standard deviation of the x-values is 14.85, 3 sd = (3)(14.85) = 44.55. Taking $\alpha = 0.01$ we allow λ to vary between $n^{0.01}/44.55 = 0.02$ and $n^{0.99}/44.55 = 1.89$. Since the kernel w vanishes outside $[-1, 1]$, $\hat{m}_\lambda(X_i)$'s and $\hat{f}_\lambda(X_i)$'s remain the same for all $\lambda > 1$. Consequently, $T_n(\lambda)$ is an increasing function of λ on $(1, \infty)$ and this tells us it is enough to consider $T_n(\lambda)$ over $[0.02, 1]$. We calculate $T_n(\lambda)$ on a grid of 100 values of λ in the range 0.02 to 1 and $\hat{\lambda}$ (i.e., the value of λ at which T_n attains its minimum) turned out to be about 0.10. Indeed, the minimum of $T_n(\lambda)$ over $(0, 1]$ is attained

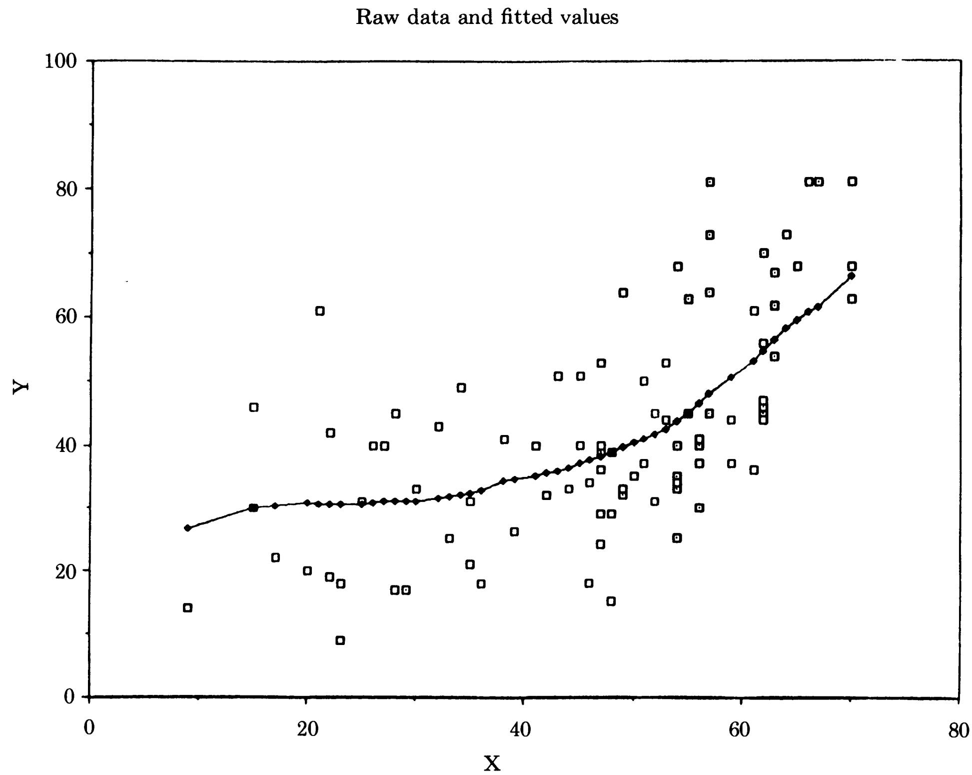

FIG. 1. *Scatter diagram showing the analysis* (X) *and statistics* (Y) *scores of* 88 *students along with the graph of the fitted kernel estimate of the regression function.*

$\hat{\lambda} \simeq 0.10$. $\hat{m}_{\hat{\lambda}}$ is plotted in the given graph along with the scatter diagram (Figure 1). In this example, we have not pursued any additional steps to avoid the boundary effects as discussed in Remark 2.3.

We have also considered generalized cross-validation (for the homoscedastic case) and cross-validation estimates of λ. For both these methods, the optimum values of λ turned out to be about 0.11. We do not present here the graph of $\hat{m}_\lambda$ for $\lambda = 0.11$, since it is very close to the one for $\lambda = 0.10$.

6. Proofs of the main results (i.e., results of Section 4).

PROOF OF LEMMA 4.1. Let us note that because of (A5)(i), $V_n(k) \geq c(\lambda_k/n) + \|\tilde{m}_k - m\|^2$ for some $c > 0$, with probability tending to 1. Here and elsewhere we will denote the generic constants by $c, c_1, c_2, \dots$. Let

$$\varepsilon_i = Y_i - m(X_i), \qquad a_{1k}(x_1, x_2) = \int \overline{W}_k(x, x_1)\overline{W}_k(x, x_2)\, dF(x)$$

and $b_k(y) = \int \overline{W}_k(x, y)(\tilde{m}_k(x) - m(x))\, dF(x)$.

Then by (A6)(i) and (A7),

$$\int a_{1k}^2(x_1, x_2)\, dF_n(x_1)\, dF_n(x_2) = \int a_k^2(x_1, x_2)\, dF(x_1)\, dF(x_2) = O_p(\lambda_k), \tag{6.1}$$

$$\begin{aligned} \int b_k^2\, dF_n &\le \left\{ \sup_y \int |\overline{W}_k(x, y)|\, dF(x) \right\} \left\{ \sup_x \int |\overline{W}_k(x, y)|\, dF_n(y) \right\} \\ &\quad \times \|\tilde{m}_k - m\|^2 \\ &= O_p(1) \|\tilde{m}_k - m\|^2. \end{aligned} \tag{6.2}$$

$$\begin{aligned} L_n(k) - V_n(k) &= \left\{ \frac{1}{n^2} \sum \varepsilon_i \varepsilon_j a_{1k}(X_i, X_j) - \frac{1}{n^2} \sum \sigma^2(X_j) a_{1k}(X_j, X_j) \right\} \\ &\quad + 2\left\{ \frac{1}{n} \sum \varepsilon_j b_k(X_j) \right\} \\ &= S_1(k) + 2S_2(k), \quad \text{say}. \end{aligned}$$

$$\begin{aligned} &P\left[\max_{k \le K_n} |[S_1(k)]/[V_n(k)]| > \varepsilon \middle| \underline{X}_n \right] \\ &\quad \le \varepsilon^{-2s} \sum_{k \le K_n} E(nS_1(k))^{2s} / (nV_n(k))^{2s} \\ &\quad \le O(1) \sum_{k \le K_n} n^{-2s} \left(\sum a_{1k}^2(X_i, X_j) \right)^s / (nV_n(k))^{2s} \\ &\qquad\qquad\qquad [\text{by Theorem 4.7 (Whittle)}] \\ &\quad \le O(1) \sum_{k \le K_n} \left(\int a_{1k}^2(x_1, x_2)\, dF_n(x_1)\, dF_n(x_2) \right)^s / (nV_n(k))^{2s} \\ &\quad \le O_p(1) \sum_{k \le K_n} (nV_n(k))^{-s} = o_p(1) \quad [\text{by (6.1) and Lemma 4.5}]. \end{aligned} \tag{6.3}$$

Hence $S_1(k) = o_p(V_n(k))$. Now

$$\begin{aligned} &P\left[\max_{k \le K_n} |S_2(k)| / V_n(k) > \varepsilon \middle| \underline{X}_n \right] \\ &\quad \le \varepsilon^{-2s} \sum_{k \le K_n} E(nS_2(k))^{2s} / (nV_n(k))^{2s} \\ &\quad \le O(1) \sum_{k \le K_n} \left(\sum b_k^2(X_j) \right)^s / (nV_n(k))^{2s} \quad (\text{by Theorem 4.7}) \\ &\quad \le O(1) \sum_{k \le K_n} (n\|\tilde{m}_k - m\|^2)^s / (nV_n(k))^{2s} \\ &\quad \le O_p(1) \sum_{k \le K_n} (nV_n(k))^{-s} = o_p(1) \quad (\text{by Lemma 4.5}). \end{aligned} \tag{6.4}$$

Hence $S_2(k) = o_p(V_n(k))$ and this proves Lemma 4.1. □

Proof of Lemma 4.3. Let us note that

$$V_n(k) = \frac{1}{n}\int \sigma^2(x')\overline{W}_k^2(x,x')\,dF_n(x')\,dF(x) + \|\tilde{m}_k - m\|^2.$$

Also $\hat{L}_n(k)$ can be written as

$$\begin{aligned}&\|m\|_n^2 - n^{-1}\sum Y_j^2 + n^{-1}\sum \varepsilon_j^2 + \|\hat{m}_k - \tilde{m}_k\|_n^2 + \|\tilde{m}_k - m\|_n^2\\ &\quad + 2\langle \hat{m}_k - \tilde{m}_k, \tilde{m}_k - m\rangle\\ &\quad - \frac{2}{n}\sum \varepsilon_j\big(\hat{m}_k(X_j) - m(X_j)\big) - 2n^{-1}\sum \varepsilon_j\big(\tilde{m}_k(X_j) - m(X_j)\big)\\ &\quad + 2n^{-2}\sum\big(Y_j - \hat{m}_k(X_j)\big)^2\overline{W}_k(X_j, X_j).\end{aligned}$$

Hence

$$\begin{aligned}&\hat{L}_n(k) - V_n(k)\\ &\quad = \left\{\|\hat{m}_k - \tilde{m}_k\|_n^2 - \frac{1}{n}\int \sigma^2(x')\overline{W}_k^2(x,x')\,dF_n(x)\,dF_n(x')\right\}\\ &\qquad + \left\{\|\tilde{m}_k - m_k\|_n^2 - \|\tilde{m}_k - m\|^2\right\} + 2\{\langle \hat{m}_k - \tilde{m}_k, \tilde{m}_k - m\rangle_n\}\\ &\qquad - 2\left\{\frac{1}{n}\sum \varepsilon_i(\hat{m}_k(X_i) - \tilde{m}_k(X_i)) - \frac{1}{n}\int \sigma^2(x)\overline{W}_k(x,x)\,dF_n(x)\right\}\\ &\qquad - 2\left\{\frac{1}{n}\sum \varepsilon_i(\tilde{m}_k(X_i) - m(X_i))\right\} - 2\left\{\frac{1}{n}\sum \varepsilon_i m(X_i)\right\}\\ &\qquad + 2\left\{n^{-2}\sum\big(Y_j - \hat{m}_k(X_j)\big)^2\overline{W}_k(X_j,X_j) - \frac{1}{n}\int \sigma^2(x)\overline{W}_k(x,x)\,dF_n(x)\right\}\\ &\qquad + \left\{\frac{1}{n}\iint \sigma^2(x')\overline{W}_k^2(x,x')\,dF_n(x')d(F_n - F)(x)\right\}\\ &\quad = T_1(k) + T_2(k) + 2T_3(k) - 2T_4(k) - 2T_5(k) - 2T_6(k)\\ &\qquad + 2T_7(k) + T_8(k).\end{aligned}$$

We note that T_6 does not depend on k and the difference $\{\hat{L}_n(k) - V_n(k)\} - \{\hat{L}_n(k^*) - V_n(k^*)\}$ does not contain T_6. Because of (A9) and (A5)(ii), $T_2(k) = o_p(V_n(k))$ and $T_8(k) = o_p(V_n(k))$. Let us note that

$$T_1(k) = n^{-2}\sum \varepsilon_i\varepsilon_j\tilde{a}_{1k}(X_i, X_j) - n^{-2}\sum \sigma^2(X_j)\tilde{a}_{1k}(X_j, X_j),$$

where $\tilde{a}_{1k}(x_1, x_2) = \int \overline{W}_k(x, x_1)\overline{W}_k(x, x_2)\,dF_n(x)$. Now arguing as in (6.3) and by (A6)(ii) we can show $T_1(k) = o_p(V_n(k))$. Now

$$T_4(k) = \frac{1}{n^2}\sum \varepsilon_i\varepsilon_j\overline{W}_k(X_i, X_j) - \frac{1}{n^2}\sum \sigma^2(X_j)\overline{W}_k(X_j, X_j).$$

(A5)(i), (A5)(ii) and another use of Whittle's theorem will tell us

$$T_4(k) = o_p(V_n(k)).$$

Now

$$T_3(k) = \frac{1}{n}\sum \varepsilon_i \tilde{b}_k(X_i) \quad \text{where} \quad \tilde{b}_k(y) = \int \overline{W}_k(x, y)(\tilde{m}_k(x) - m(x))\, dF_n(x).$$

Arguing as in (6.2) and by (A9), we get

$$\int \tilde{b}_k^2\, dF_n = \|\tilde{m}_k - m\|^2 + o_p(V_n(k)).$$

Now, an argument similar to that in (6.4) will show $T_3(k) = o_p(V_n(k))$. By Whittle's theorem and (A9) we can show that $T_5(k) = o_p(V_n(k))$. Now

$$\begin{aligned}
T_7(k) &= n^{-2}\sum\Big\{\big(\varepsilon_j^2 - \sigma^2(X_j)\big) + \big(\hat{m}_k(X_j) - \tilde{m}_k(X_j)\big)^2 + \big(\tilde{m}_k(X_j) - m(X_j)\big)^2 \\
&\qquad + 2\big(\hat{m}_k(X_j) - \tilde{m}_k(X_j)\big)\big(\tilde{m}_k(X_j) - m(X_j)\big) - 2\varepsilon_j\big(\hat{m}_k(X_j) - \tilde{m}_k(X_j)\big) \\
&\qquad - 2\varepsilon_j\big(\tilde{m}_k(X_j) - m(X_j)\big)\Big\}\overline{W}_k(X_j, X_j) \\
&= T_{71}(k) + T_{72}(k) + T_{73}(k) + 2T_{74} - 2T_{75} - 2T_{76}, \quad \text{say.}
\end{aligned}$$

Using Whittle's theorem and (A5)(iii) we can show that $T_{71}(k) = o_p(V_n(k))$. Because of (A5)(iii), $T_{72}(k) = O_p(\lambda_k n^{-1})\|\hat{m}_k - \tilde{m}_k\|_n^2$. We know that $T_1(k) = \|\hat{m}_k - \tilde{m}_k\|_n^2 - E\{\|\hat{m}_k - \tilde{m}_k\|_n^2 | \underline{X}_n\} = o_p(V_n(k))$. Since by (A5)(ii) $E\{\|\hat{m}_k - \tilde{m}_k\|_n^2 | \underline{X}_n\} = O_p(\lambda_k n^{-1})$, we conclude that

$$T_{72}(k) = O_p(\lambda_k n^{-1})\big[O_p(\lambda_k n^{-1}) + o_p(V_n(k))\big] = o_p(V_n(k))$$

$$\Big(\text{since } \sup_{k \le K_n} \lambda_k n^{-1} \to 0\Big).$$

By (A5)(iii) and (A9),

$$T_{73}(k) = O_p(\lambda_k n^{-1})\|\tilde{m}_k - m\|_n^2 = o_p(V_n(k)).$$

Arguing similarly,

$$\begin{aligned}
|T_{74}(k)| &\le O_p(\lambda_k n^{-1})\|\hat{m}_k - \tilde{m}_k\|_n^2 \|\tilde{m}_k - m\| = o_p(V_n(k)), \\
|T_{75}(k)| &\le O_p(\lambda_k n^{-1})\big(n^{-1}\textstyle\sum \varepsilon_j^2\big)^{1/2}\|\hat{m}_k - \tilde{m}_k\|_n = o_p(V_n(k)) \\
&\qquad\qquad \big[\text{since } n^{-1}\textstyle\sum \varepsilon_j^2 = O_p(1)\big], \\
|T_{76}(k)| &\le O_p(\lambda_k n^{-1})\big(n^{-1}\textstyle\sum \varepsilon_j^2\big)^{1/2}\|\tilde{m}_k - m\|_n = o_p(V_n(k))
\end{aligned}$$

and this completes the proof of Lemma 4.3. □

PROOF OF LEMMA 4.5. (i) Because of (A8) it is enough to prove only part (a). If $\inf_k \lambda_k \to \infty$ as $n \to \infty$, the result is trivial. Otherwise, by (A3) and (A4) we can find a sequence $k_n \to \infty$ (i.e., $k_{in} \to \infty$ for $i = 1, \dots, p$) such that

$\inf_{k \le k_n} n\|m_k - m\|^2 \to \infty$; hence,

$$\inf_{k \le K_n} n\bar{V}_n(k) \ge \inf\left\{ \inf_{k \le k_n} n\|m_k - m\|^2, \inf_{k \nleq k_n} \int \sigma^2(x')\overline{W}_k^2(x, x')\, dF(x)\, dF_n(x') \right\}$$

($k \le k_n$ means $k_i \le k_{in}$, $i = 1, \dots, p$ and $k \nleq k_n$ means $k_i > k_{in}$ for some $1 \le i \le p$).

The above expression converges to infinity in probability because of (A5)(i).

Because of (A8) it is enough to prove part (b). We note that $\lambda_{k^*} \to_P \infty$ as a consequence of part (i):

$$\sum_{k \le K_n} (nV_n(k))^{-s} \le \sum_{k \le k^*} (nV_n(k^*))^{-s} + \sum_{k \nleq k^*} (nV_n(k))^{-s}.$$

The first expression goes to zero because of (A5)(i) and (A4)(i), whereas the second expression goes to zero because of (A4)(i). □

PROOF OF LEMMA 4.8. Let us first notice that it is enough to prove the result for elementary functions of the form $\phi(x_1, \dots, x_t) = \Sigma\alpha(i_1, \dots, i_t)I(x_t \in A_{i_t})$, where $\{A_j\}$ forms a partition of I and $n^2F(A_j) < 1$ and for all j. Hence

$$\int \phi(x_1, \dots, x_t) \prod_{i=1}^{t} d(F_n - F)(x_i) = n^{-t} \sum \alpha(i_1, \dots, i_t) d_{i_1} \cdots d_{i_t},$$

where $d_j = n\{F_n(A_j) - F(A_j)\}$. This integral can be written as

$$n^{-t} \sum_{l=1}^{t} \left\{ \sum \alpha(i_1, \dots, i_t) d_{i_1} \cdots d_{i_t} \right\},$$

where the inside summation is over all $(i_1, \dots, i_t)$ such that $(i_1, \dots, i_t)$ form only l-distinct indices. Hence for each l and $(i_1, \dots, i_t)$, the term $d_{i_1} \cdots d_{i_t}$ looks like $d_1^{k_1} \cdots d_l^{k_l}$, k_i's positive integers and $\Sigma_{i=1}^{l} k_i = t$. For the moment let us assume (Lemma 6.1) that for $l \le t$, there exists a constant c depending only on t such that

$$|Ed_1^{k_1} \cdots d_l^{k_l}| \le c \prod_{i=1}^{l} F(A_i) n^{l - [(a+1)/2]},$$

where a is the number of $k_i = 1$. This tells us that the expected value of the above integral is bounded above in absolute value by

$$cn^{-t} \sum_{l=1}^{t} \sum_{\xi \in Z(t, l)} n^{l - [(a(\xi)+1)/2]} \int |\phi(x_\xi)|\, dF(x_1) \cdots dF(x_l). \qquad \square$$

LEMMA 6.1. *Let $(n_1, n_2, \dots)$ be multinomial $(n, p_1, p_2, \dots)$ with $n^2 p_u < 1$ for all u and $d_u = n_u - np_u$. Then for $k_u \ge 1$, k_u's integers,*

$$|Ed_1^{k_1} \cdots d_i^{k_i}| \le c(k_1, \dots, k_i) n^{i - [(a_i+1)/2]} (p_1 \cdots p_i)$$

for all i, where a_i is the total number of $k_u = 1$ for $1 \le u \le i$ and $c(k_1, \dots, k_i)$ depends only on $k_1, \dots, k_i$.

PROOF. It is easy to see that it is true for $i = 1$, since if $k_1 = 1$ it is obvious, whereas if $k_1 \geq 2$,

$$Ed_1^{k_1} = \sum_{u=1}^{[k_1/2]} \sum_{v=0}^{[k_1/2]-u} (np_1)^u p_1^v g_{k_1, u, v},$$

where $g_{k_1, u, v}$ are constants independent of n and p_1. Our result holds for $i = 1$ since $n^2 p_1 < 1$. Now let us assume that it is true for i; we will show that it is true for $i + 1$.

Let us note that the conditional distribution of n_{i+1} given $n_1, \dots, n_i$ is binomial (m_i, q_i) where $m_i = n - n_1 - \cdots - n_i$ and $q_i = p_{i+1}/(1 - p_1 - \cdots - p_i)$. Let $\delta_i = d_1 + \cdots + d_i$. Then $d_i + 1 = (n_{i+1} - m_i q_i) - q_i \delta_i$. So if $k_{i+1} = 1$, $E(d_{i+1} | n_1, \dots, n_i) = -q_i \delta_i$ and

$$E\left(d_1^{k_1} \cdots d_i^{k_i} d_{i+1}\right) = q_i E\left(d_1^{k_1} \cdots d_i^{k_i}(d_1 + \cdots + d_i)\right).$$

Now for $1 \leq u \leq i$,

$$|Ed_1^{k_1} \cdots d_{u-1}^{k_{u-1}} d_u^{k_u+1} d_{u+1}^{k_{u+1}} \cdots d_i^{k_i}| \leq \begin{cases} c_1 n^{i-[(a_i+1)/2]}(p_1 \cdots p_i) & \text{if } k_u \geq 2, \\ c_1 n^{i-[a_i/2]}(p_1 \cdots p_i) & \text{if } k_u = 1, \end{cases}$$

where c_1 depends only on $k_1, \dots, k_i$. This tells us

$$\begin{aligned} |E\left(d_1^{k_1} \cdots d_i^{k_i} d_{i+1}\right)| &\leq q_i c_2 n^{i-[a_i/2]}(p_1 \cdots p_i) \\ &\leq c_2' n^{i+1-[(a_i+2)/2]}(p_1 \cdots p_{i+1}), \end{aligned}$$

where c_2 and c_2' depend only on on $k_1, \dots, k_i$. Since $a_i + 1 = a_{i+1}$ here, our result holds.

Now let us assume that $k_{i+1} \geq 2$. Then $a_{i+1} = a_i$ and

$$\begin{aligned} &E\left(d_{i+1}^{k_{i+1}} | n_1, \dots, n_i\right) \\ &\quad = (-q_i\delta_i)^{k_{i+1}} + \sum_{j=2}^{k_{i+1}} \binom{k_{i+1}}{j} E\left\{(n_{i+1} - m_i q_i)^j | n_1, \dots, n_i\right\}(-q_i\delta_i)^{k_{i+1}-j} \\ &\quad = (-q_i\delta_i)^{k_{i+1}} + \sum_{j=2}^{k_{i+1}} \binom{k_{i+1}}{j} \sum_{u=1}^{[j/2]} \sum_{v=0}^{[j/2]-u} (m_i q_i)^u q_i^v g_{j,u,v}(-q_i\delta_i)^{k_{i+1}-j}. \end{aligned}$$

Since $m_i q_i = np_{i+1} - q_i \delta_i$, a little calculation will show that

$$E\left(d_{i+1}^{k_{i+1}} | n_1, \dots, n_i\right) = \sum_{j=0}^{k_{i+1}} r_j (q_i \delta_i)^j,$$

where r_j's are constants and can be bounded above in absolute value by constants independent of n and q_i. Moreover, $|r_0| \leq c_3(np_{i+1})$ for some constant c_3 independent of n and q_i.

$$\begin{aligned} E\left(d_1^{k_1} \cdots d_i^{k_i} d_{i+1}^{k_{i+1}} | n_1, \dots, n_i\right) &= \sum_{j=0}^{k_{i+1}} r_j (q_i\delta_i)^j d_1^{k_1} \cdots d_i^{k_i} \\ &= \sum_{j=0}^{k_{i+1}} r_j q_i^j E(d_1 + \cdots + d_i)^j d_1^{k_1} \cdots d_i^{k_i}. \end{aligned}$$

For $j = 0$, there exist constants depending only on $k_1, \dots, k_i$ so that

$$|r_0 E d_1^{k_1} \cdots d_i^{k_i}|| \le |r_0| c_4 n^{i-[(a_i+1)/2]}(p_1 \cdots p_{i+1}) \le c_4' n^{i+1-[(a_i+2)/2]}(p_1 \cdots p_{i+1}).$$

For $j \ge 1$, $E\{(d_1 + \cdots + d_i)^j d_1^{k_1} \cdots d_i^{k_i}\}$ is a linear combination of terms of the form $E(d_1^{k_1+l_1} \cdots d_i^{k_i+l_i})$, where l_u, $1 \le u \le i$, are nonnegative integers and $l_1 + \cdots + l_i = j$. For $1 \le j \le a_i$ apart from constants, the dominant terms are of the form $n^{i-[(a_i-j+1)/2]}(p_1 \cdots p_i)$ (since these terms have $a_i - j$ exponents equal to 1). Hence we get

$$q_i^j |E(d_1 + \cdots + d_i)^j d_1^{k_1} \cdots d_i^{k_i}| \le q_i^j c_5 n^{i-[(a_i-j+1)/2]} \le c_5' n^{i+1-[(a_i+1)/2]}(p_1 \cdots p_{i+1}),$$

since $q_i^j \le n^{-j+1} p_{i+1}(1 + i/n^2)$ and $[(a_i - j + 1)/2] + j \ge [(a_j + 1)/2]$ (c_5 and c_5' depend only on $k_1, \dots, k_i$). For $j \ge a_i$, apart from constants, the dominant terms are of the form $n^i(p_1 \cdots p_i)$ (since these terms have all exponents greater than or equal to 2). Hence for $j \ge a_i$,

$$q_i^j |E(d_1 + \cdots + d_i)^j d_1^{k_1} \cdots d_i^{k_i}| \le q_i^j c_6 n^i (p_1 \cdots p_i) \le c_6' n^{i+1-[(a_i+1)/2]}(p_1 \cdots p_{i+1}),$$

since $q_i^j \le n^{-a_i+1} p_{i+1}(1 + i/n^2)$ and $a_i \ge [(a_i + 1)/2]$ (c_6 and c_6' depend only on $k_1, \dots, k_i$). Our result is true for $i + 1$ by collecting all the terms. □

7. Proofs of the results in Section 3. We will first prove two important lemmas, then Theorems 3.2, 3.3 and 3.4 and finally we will prove Theorem 3.1. Throughout this section, we will denote $\gamma'\gamma$ by $\|\gamma\|^2$ for any vector of real numbers γ.

LEMMA 7.1. (i) *Let* $\{\xi_{nk}\}$ *be a class of functions on* $I \times I$ *such that* $\sup_x \int |\xi_{nk}(x, y)|\, dF(y) = O(1)$ *and* $\sup_y \int |\xi_{nk}(x, y)|\, dF(x) = O(1)$. *Then*

$$\iint [\xi_{nk}(x, y)(m_k(y) - m(y))\, dF(y)]\, d(F_n - F)(x) = o_p(V_n(k)).$$

(ii) $\int (m_k - m)^2\, d(F_n - F) = o_p(\overline{V}_n(k))$.

PROOF. Let us note that because of (A5)(i), $\overline{V}_n(k) \ge c\lambda_k/n + \|m_k - m\|^2$ for some $c > 0$, with probability approaching 1. The proofs for both parts of this lemma follow easily from the following result in Burman (1985). If $\{g_{nk}\}$ is a class of functions uniformly bounded by a constant, say 1, then

$$E\left[\int g_{nk}\, d(F_n - F)\right]^{2s} \le c n^{-2s} \sum_{l=1}^{s} n^l \|g_{nk}\|^{2l},$$

where c depends only on s. □

LEMMA 7.2. *For all the regression methods (except for piecewise polynomial regression with random partitioning) described in Section* 3,

$$\int(\tilde{m}_k - m)^2 d(F_n - F) = o_p(\overline{V}_n(k)).$$

PROOF. First let us note that

$$\begin{aligned}\int(\tilde{m}_k - m)^2 d(F_n - F) &= \int(\tilde{m}_k - m_k)^2 d(F_n - F) + \int(m_k - m)^2 d(F_n - F)\\ &\quad + 2\int(\tilde{m}_k - m_k)(m_k - m)d(F_n - F)\\ &= I_1 + I_2 + 2I_3, \quad \text{say.}\end{aligned}$$

By part (ii) of Lemma (7.1), $I_2 = o_p(\overline{V}_n(k)]$. Let us denote $\tilde{m}_k - m_k$ and $m_k - m$ by $\tilde{g}_k$ and g_k, respectively. Let $M = n^{1-\delta_1}$ for some small $\delta_1 > 0$. Divide $[0,1]^p$ into M cubes of equal sizes and denote them by A_j, $j = 1, \dots, M$. Let δ_{nk} be as in Theorem 3.1. For all the regression estimates in Section 3, the following are true (by choosing δ_1 properly):

$$a_k = \sup_j |\tilde{g}_k(x_j)| = o_p(\lambda_k \delta_{nk}) \quad \text{where } x_j \text{ is the middle point of } A_j,$$

$$b_k = \sup_j \sup\{|\tilde{g}_k(x_1) - g_k(x_2)|: x_1, x_2 \in A_j\} = o_p(\lambda_k \delta_{nk} n^{-\delta}).$$

Let $\int\tilde{g}_k^2\, dF_n = \sum_j \tilde{g}_k^2(x_{1j})F_n(A_j)$ and $\int\tilde{g}_k^2\, dF = \sum_j \tilde{g}_k^2(x_{2j})F(A_j)$. Then

$$\begin{aligned}I_1 &= \sum\{\tilde{g}_k^2(x_{1j}) - \tilde{g}_k^2(x_{2j})\}F_n(A_j) + \sum\tilde{g}_k^2(x_{2j})(F_n(A_j) - F(A_j))\\ &= I_{11} + I_{12}, \quad \text{say.}\end{aligned}$$

Obviously $I_{11} = o_p(\overline{V}_n(k))$. Lemma 4.6 tells us that for all j, $|F_n(A_j) - F(A_j)| \le c(\log n)^{-2}F(A_j)$ for some constant $c > 0$. Hence,

$$|I_{12}| \le c(\log n)^{-2}\int\tilde{g}_k^2\, dF = o_p(\overline{V}_n(k)).$$

Let I_j be the indicator function of A_j. Then

$$|I_3| \le b_k\int|g_k|d(F_n + F) + \sum|\tilde{g}_k(x_{2j})|\left|\int I_j g_k d(F_n - F)\right| = I_{31} + I_{32}, \quad \text{say.}$$

Obviously $I_{31} = o_p(\overline{V}_n(k))$.

$$I_{32} \le c_1 M^{1/2}\left(\int\tilde{g}_k^2\, dF\right)^{1/2}\left(\sum\left(\int I_j g_k d(F_n - F)\right)^2\right)^{1/2} \quad \text{for some } c_1 > 0.$$

Note that $\int\tilde{g}_k^2\, dF = O_p(\overline{V}_n(k))$ and a use of Lemma 4.8 gives us

$$E\left[\sum\left(\int I_j g_k d(F_n - F)\right)^2\right]^u \le c_2 n^{-2u}\sum_{t=1}^{u} n^t\|g_k\|^{2t} \quad \text{for some } c_2 > 0.$$

This shows that $I_{32} = o_p(\overline{V}_n(k))$ and completes the proof of Lemma 7.2. □

PROOF OF THEOREM 3.2. We need to verify conditions (i) and (ii) of Theorem 3.1. We will prove this result for fixed partitioning. The case for random partitioning is similar with some changes. Note that ψ_k is the vector of $I_{kt}\phi_{ktu}$'s and it is enough to prove (i) and (ii) for A_{kt} and $\hat{A}_{kt}$.

(i) Recalling that f is bounded below and above by b and B and denoting the Lebesgue measure by $\bar{\mu}$ we get, for any $\gamma \in R^{\lambda_k}$ with $\|\gamma\| = 1$,

$$b\int\left(\sum \gamma_u \phi_{ktu}\right)^2 I_{kt}\, d\bar{\mu} \le \gamma' A_{kt}\gamma \le B\int\left(\sum \gamma_u \phi_{ktu}\right)^2 I_{kt}\, d\bar{\mu}.$$

An easy calculation will show that

$$\int\left(\sum \gamma_u \phi_{ktu}\right)^2 I_{kt}\, d\bar{\mu} = 2^{-3p}\lambda_k^{-1}\int_{[-1,1]}\left(\sum \gamma_u \prod_{i=1}^{p} z_i^{u_i}\right) dz_1 \cdots dz_p,$$

$$u = (u_1, \ldots, u_p).$$

The result follows once we notice that $(\Sigma\gamma_u \Pi_{i=1}^p z_i^{u_i})^2 \ge c(\Sigma|\gamma_u|\Pi_{i=1}^p z_i^{u_i})^2$ for some $c > 0$ which depends only on v, the degree of the piecewise polynomials (the proof uses a compactness argument).

(ii) Since $\int I_{kt}\phi_{ktu_2}^2\phi_{ktu_1}^2\, dF = O(\lambda_k^{-1})$, Lemma 4.6 tells us

$$\int I_{kt}\phi_{ktu_1}\phi_{ktu_2} d(F_n - F) = o_p(\varepsilon_{nk})$$

and so the result follows. □

PROOF OF THEOREM 3.3. We will show the proof for equispaced knots. The proofs for random knots are similar. All we need to do is to verify conditions (i) and (ii) of Theorem 3.1.

(i) Using property (viii) of de Boor [(1978), page 155] and Lemma 5.1 of Burman (1985), we can conclude that for any $\gamma \in R^{\lambda_k}$ there exist constants $0 < c_1 < c_2$ such that

$$c_1\lambda_k^{-1}\|\gamma\|^2 \le \int\left(\gamma'\psi_k(x)\right)^2 dx \le c_2\lambda_k^{-1}\|\gamma\|^2.$$

Since the marginal density f of X is bounded below and above by constants b and B, we conclude the proof by noting that

$$c_3 b\lambda_k^{-1}\|\gamma\|^2 \le \int\left(\gamma'\psi_k(x)\right)^2 dF(x) \le c_4 b\lambda_k^{-1}\|\gamma\|^2.$$

(ii) $\|\hat{A}_k - A_k\| \le \sup_{t_1}\Sigma_{t_2}|\int B_{kt_1}B_{kt_2} d(F_n - F)|$. Since $B_{kt_1}B_{kt_2} \equiv 0$ for $|t_1 - t_2| > M$ for some $M > 0$, the result follows from Lemma 4.6 and the fact that $\int B_{kt_1}^2 B_{kt_2}^2\, dF = O_p(\lambda_k^{-1})$ for $|t_1 - t_2| \le M$. □

PROOF OF THEOREM 3.4. We will prove the result only for trigonometric polynomials. So we will verify conditions (i) and (ii) of Theorem 3.1. The proof for (i) is quite obvious. So let us prove (ii).

(ii) Let $R_k(x_1, x_2) = \sum_{j=0}^{\lambda_k} \phi_j(x_1)\phi_j(x_2)$. Then $\|\hat{A}_k - A_k\|^{2u}$ is bounded above by

$$\operatorname{tr}\Big(\big(\hat{A}_k - A_k\big)^{2u}\Big) = \lambda_k^{-2u} \int R_k(x_1, x_{u+1}) R_k(x_u, x_{2u}) \prod_{t=1}^{2u-1} R_k(x_t, x_{t+1}) \times \prod_{t=1}^{2u} d(F_n - F)(x_t)$$
$$= \lambda_k^{-2u} \int \overline{R}_k(\underline{x}) \prod_{t=1}^{2u} d(F_n - F)(x_t), \quad \text{say,}$$

where $\overline{R}_k(\underline{x}) = R_k(x_1, x_{u+1}) R_k(x_u, x_{2u}) \prod_{t=1}^{2u-1} R_k(x_t, x_{t+1})$. Using Lemma 4.8,

$$E \int \overline{R}_k(\underline{x}) \prod_{t=1}^{2u} d(F_n - F)(x_t)$$
$$\leq c \sum_{l=1}^{2u} n^{-2u+l} \sum_{\nu \in Z(2u, l)} n^{-[(a(\nu)+1)/2]} \int |\overline{R}_k(x_\nu)| \, dF(x_1) \cdots dF(x_l).$$

Now, $\int |\overline{R}_k(x_\nu)| \, dF(x_1) \cdots dF(x_l) \leq$ const. λ_k^{2+2u-l} for $\nu \in Z(2u, l)$, by noting that

$$R_k(x_1, x_2) = \sin[(n + 1/2)(x_1 - x_2)] / \{2\pi \sin[(x_1 - x_2)/2]\}.$$

Since $a(\nu) > 2l - 2u$ for $l > u$, we get

$$E \int \overline{R}_k(\underline{x}) \prod_{t=1}^{2u} d(F_n - F)(x_t) \leq O(1) \left\{ \lambda_k^2 \sum_{l=1}^{u} (\lambda_k/n)^{2u-l} + \lambda_k^2 (\lambda_k/n)^u \right\}$$
$$= O(1) \lambda_k^2 (\lambda_k/n)^u \quad \Big(\text{since } \sup_{k \leq K_n} \lambda_k/n \to 0\Big).$$

Our result follows by taking $u > 3\delta^{-1}$ and by noting that

$$P\Big[\sup_{k \leq K_n} \delta_{nk}^{-1} \|\hat{A}_k - A_k\| > \varepsilon\Big] \leq \sum_{k \leq K_n} \varepsilon^{-2u} \delta_{nk}^{-2u} E \|\hat{A}_k - A_k\|^{2u}$$
$$\leq O(1) n^{-u\delta} \sum_{k \leq K_n} \lambda_k^2 = o(1). \qquad \square$$

Proof of Theorem 3.1. It is clear that in order to prove the theorem all we need to do is to check (A5)–(A9).

Let us note that $\|\hat{A}_k\| = O_p(\lambda_k^{-1})$, $\|\hat{A}_k^{-1}\| = O_p(\lambda_k)$ and $\|\hat{A}_k - A_k\| \, \|A_k\| = o_p(1)$. Let $\bar{d}_k = \int m \psi_k \, dF_n$. Then by Lemma 4.6 and the fact that $\sup_{k,x} \|\psi_k(x)\| < \infty$,

$$\|\bar{d}_k - d_k\|^2 = \sum \Big[\int m \psi_{kt} d(F_n - F)\Big]^2 = \sum \Big[o_p\big(\alpha_n^4\big) + o_p\big(\alpha_n^2\big) \int m^2 \psi_{kt}^2 \, dF\Big]$$
$$= o_p\big(\alpha_n^2\big).$$

Also, $0 \le \int (m - d'A_k^{-1}\psi)^2\, dF = \|m\|^2 - d'A_k^{-1}d$. By assumption (i) of this theorem $\|d_k\|^2 = O(\lambda_k^{-1})$. Similarly, $\|d_k\|^2 = O_p(\lambda_k^{-1})$.

$$\text{(A5)(i)}\quad \lambda_k^{-1} \int \overline{W}_k^2(x, x')\, dF_n(x')\, dF(x) = \lambda_k^{-1} \operatorname{tr}\left(A_k \hat{A}_k^{-1}\right)$$
$$= 1 + \lambda_k^{-1} \operatorname{tr}\left(\left(A_k - \hat{A}_k\right)\hat{A}^{-1}\right)$$
$$= 1 + o_p(1).$$

(ii) It can be shown that $\|\int \sigma^2(x)\psi_k(x)\psi_k'(x)\, dF_n(x)\| = o_p(\lambda_k^{-1})$. So

$$|\lambda_k^{-1} \int \sigma^2(x')\overline{W}_k^2(x, x')\, dF_n(x')d(F_n - F)(x)|$$
$$\le \|\hat{A}_k^{-1}\left(\hat{A}_k - A_k\right)\hat{A}_k^{-1}\| \left\| \int \sigma^2 \psi_k \psi_k'\, dF_n \right\|$$
$$\le O_p(1)\|\hat{A}_k^{-1}\|\, \|\hat{A}_k - A_k\| = o_p(1).$$

$$\text{(iii)}\quad \sup_x \lambda_k^{-1}|\psi_k'(x)\hat{A}_k^{-1}\psi_k(x)| \le \sup_x \lambda_k^{-1}\|\psi_k(x)\|^2\|\hat{A}_k^{-1}\| = O_p(1).$$

(A6) Let us note that $a_k(x_1, x_2) = \psi_k'(x_1)\hat{A}_k^{-1}\psi_k(x_2)$.

$$\text{(i)}\quad \int a_k^2(x_1, x_2)\, dF(x_1)\, dF(x_2) = \operatorname{tr}\left(\hat{A}_k^{-1}A_k\hat{A}_k^{-1}A_k\right)$$
$$= O(\lambda_k)\|\hat{A}_k^{-1}A_k\hat{A}_k^{-1}A_k\| = O_p(\lambda_k).$$

$$\text{(ii)}\quad \int a_k^2(x_1, x_2)\, dF_n(x_1)\, dF_n(x_2) = \lambda_k.$$

(A7) The proofs of all the parts are trivial and so we will skip them.

$$\text{(A8)}\quad V_n(k) - \overline{V}_n(k) = \int (\hat{m}_k - m_k)(m_k - m)\, dF$$
$$= \left(\hat{A}_k^{-1}d_k A_k^{-1}d_k\right)' \int \psi_k(m_k - m)\, dF \equiv 0,$$

since

$$\int \psi_k(m_k - m)\, dF = \int \psi_k\psi_k' A_k^{-1}d_k\, dF - \int \psi_k m\, dF = d_k - d_k = 0.$$

$$\text{(A9)}\quad \int (\tilde{m}_k - m)^2 d(F_n - F) = \int (\tilde{m}_k - m_k)^2 d(F_n - F)$$
$$+ \int (m_k - m)^2 d(F_n - F)$$
$$+ 2\int (\tilde{m}_k - m_k)(m_k - m)d(F_n - F)$$
$$= I_1 + I_2 + 2I_3, \quad \text{say}.$$

By part (b) of Lemma 7.1, $I_2 = o_p(V_n(k))$.

Let $\tilde{Q}_k = \hat{A}_k^{-1}\bar{d}_k$ and $Q_k = A_k^{-1}\bar{d}_k$. Then $\tilde{m}_k(x) = \tilde{Q}_k'\psi_k(x)$ and $m_k(x) = Q_k'\psi_k(x)$. Since $\|\bar{d}_k - d_k\| = o_p(\alpha_n)$ and $\|\hat{A}_k - A_k\| = o_p(\delta_{nk})$, an easy calculation will give $\|\tilde{Q}_k - Q_k\| = o_p(\lambda_k^{3/2}\delta_{nk})$.

$$|I_1| \le \|\tilde{Q}_k - Q_k\|^2\|\hat{A}_k - A_k\| = o_p\big(\lambda_k^3\delta_{nk}^3\big) = o_p(V_n(k)),$$

$$|I_3|^2 \le \|\tilde{Q}_k - Q_k\|^2\left[\sum_t\left\{\int\psi_{kt}(m_k - m)\,d(F_n - F)\right\}^2\right]$$

$$= o_p\big(\lambda_k^3\delta_{nk}^2\big)\left[\sum_t\left\{o_p\big(\alpha_n^4\big) + o_p\big(\alpha_n^2\big)\int\psi_{kt}^2(m_k - m)^2\,dF\right\}\right] \quad \text{(by Lemma 4.6)}$$

$$= o_p\big(\lambda_k^3\delta_{nk}^2\big)\big[o_p\big(\lambda_k\alpha_n^4\big) + o_p\big(\alpha_n^2\big)\|m_k - m\|^2\big].$$

This clearly proves that $I_3 = o_p(V_n(k))$. □

PROOF OF LEMMA 3.6. Without loss of generality let us assume that

$$w(x) \le I\big(|x_1| \le 1,\dots,|x_p| \le 1\big) = U(x,\mathbf{1}), \quad \text{say},$$

where $\mathbf{1} = (1,\dots,1) \in R^p$.

Let $q_n \to \infty$ such that $h_{K_n}/q_n \to 0$ and $(h_{K_n}/q_n)^\beta = 1/\sqrt{n}$. For $t = (t_1,\dots,t_p)$, let $S_t = ((t_1 - 1)q_n^{-1}, t_1q_n^{-1}] \times \cdots \times ((t_p - 1)q_n^{-1}, t_pq_n^{-1}]$, $1 \le t_i \le q_n$, $i = 1,\dots,p$. Let us note that

$$\sup_{x\in S_t}|f_k(x) - f_k(x_t)| \le \lambda_k\|h_k\|^\beta O\big(q_n^{-\beta}\big)\int U\big(x_t - x', h_k^{-1} + q_n^{-1}\mathbf{1}\big)\,dF(x')$$

$$= O(n^{-1/2}),$$

where x_t is the middle point of S_t and $h_k^{-1} = (h_{k_1}^{-1},\dots,h_{k_p}^{-1})$. Similarly

$$\sup_{x\in S_t}|\hat{f}_k(x) - \hat{f}_k(x_t)| \le \lambda_k\|h_k\|^\beta O\big(q_n^{-\beta}\big)\int U\big(x_t - x', h_k^{-1} + q_n^{-1}\mathbf{1}\big)\,dF_n(x').$$

By Lemma 4.6,

$$\left|\int U\big(x_t - x', h_k^{-1} + q_n^{-1}\mathbf{1}\big)d(F_n - F)(x')\right| = o_p(\varepsilon_{nk}).$$

So we get

$$\sup_{x\in S_t}|\hat{f}_k(x) - \hat{f}_k(x_t)| = O_p(n^{-1/2})$$

and

$$\sup_t\sup_{x\in S_t}|\hat{f}_k(x) - f_k(x_t)| \le \sup_t|\hat{f}_k(x_t) - f_k(x_t)| + O_p(n^{-1/2})$$

$$= o_p(\lambda_k\varepsilon_{nk}) + o_p(n^{-1/2}) \quad \text{(by Lemma 4.6)}.$$

This proves that $\sup_x|\hat{f}_k(x) - f_k(x)| = o_p(\lambda_k\varepsilon_{nk})$. Similarly we can show $\sup_x|\bar{m}_k(x) - m(x)| = o_p(\lambda_k\varepsilon_{nk})$. □

PROOF OF THEOREM 3.5. To prove this theorem all we need to do is to check the conditions (A1)–(A9) for the discrete bandwidths (in view of Lemma 3.7). Note that (A9) is proved in Lemma 7.2.

$$\text{(A5)(i)}\quad \lambda_k^{-1}\int \overline{W}_k^2(x,x')\,dF_n(x')\,dF(x)$$

$$= \lambda_k^{-1}\int\left[\int \hat{f}_k^{-2}(x)w^2(h_k(x-x'))\,dF(x)\right]dF_n(x').$$

The above expression is bounded above and below by positive constants.

(ii) By Lemma 3.6,

$$\sup_x |\hat{f}_k^{-2}(x) - f_k^{-2}(x)| = o_p(\lambda_k\varepsilon_{nk}).$$

Now

$$\left|\lambda_k^{-1}\int\sigma^2(x')\overline{W}_k^2(x,x')\,dF_n(x')d(F_n-F)(x)\right|$$

$$\leq \left|\lambda_k^{-1}\int f_k^{-2}(x)\sigma^2(x')w^2(h_k(x-x'))\,dF_n(x')d(F_n-F)(x)\right|$$

$$+\,o_p(\lambda_k\varepsilon_{nk})\lambda_k\left[\int\sigma^2(x')w^2(h_k(x-x'))\,dF_n(x')d(F_n+F)(x)\right].$$

Both terms on the above expression are $o_p(1)$ since a repetition of the arguments of Lemma 3.6 will give us

$$\sup_{x'}\left|\lambda_k\int f_k^{-2}(x)\sigma^2(x')w^2(h_k(x-x'))\,d(F_n-F)(x)\right| = o_p(\lambda_k\varepsilon_{nk})$$

and

$$\sup_x\left|\lambda_k\int\sigma^2(x')w^2(h_k(x-x'))d(F_n-F)(x')\right| = o_p(\lambda_k\varepsilon_{nk}).$$

$$\text{(iii)}\qquad \sup_x|\lambda_k^{-1}\overline{W}_k(x,x)| = \sup_x|\hat{f}_k^{-1}(x)|w(0) = O_p(1).$$

(A6) As in the proof of Lemma 3.6 let us assume that $w(z) \leq U(z,\mathbf{1})$. It is not hard to see that

$$\int a_k^2(x_1,x_2)\,dF(x_1)\,dF(x_2)$$

$$\leq O_p(1)\int U(x_1-x_2, 2h_k^{-1}\mathbf{1})$$

$$\times\left[\lambda_k^2\int U\left(x-\frac{x_1+x_2}{2}, h_k^{-1}\mathbf{1}\right)dF(x)\right]^2 dF(x_1)\,dF(x_2).$$

Since $\sup_z \lambda_k \int U(x - z, h_k^{-1}\mathbf{1})\, dF(x) = O_p(1)$, the last term is $O_p(\lambda_k)$. Similarly, $\int a_k^2(x_1, x_2)\, dF_n(x_1)\, dF_n(x_2) = O_p(\lambda_k)$.

(A7) The proofs are simple and so we will skip them.

(A8) $V_n(k) - \overline{V}_n(k) = 2\langle \tilde{m}_k - \overline{m}_k, m_k - m\rangle + 2\langle \overline{m}_k - m_k, m_k - m\rangle$. Using part (i) of Lemma 7.1 we get $\langle \overline{m}_k - m_k, m_k - m\rangle = o_p(\overline{V}_n(k))$ by taking $\xi_k(x, y) = \lambda_k f_k^{-2}(x) w(h_k(x - y))$.

Since $\tilde{m}_k = \hat{f}_k^{-1} f_k \overline{m}_k$, we can write

$$\begin{aligned}\int(\tilde{m}_k - \overline{m}_k)(m_k - m)\, dF = &\int\left(\hat{f}_k^{-1} f_k^{-1}\right)(\overline{m}_k - m_k)(m_k - m)\, dF\\ &- \int f_k^{-1}\left(\hat{f}_k - f_k\right)\overline{m}_k(m_k - m)\, dF\\ &- \int\left(\hat{f}_k^{-1} - \hat{f}_k^{-1}\right)\left(\hat{f}_k - f_k\right)(m_k - m)\, dF\\ = &\; I_1 - I_2 - I_3, \quad \text{say.}\end{aligned}$$

By Lemma 3.6,

$$|I_1| = o_p\left(\lambda_k^2 \varepsilon_{nk}^2\right)\|m_k - m\| = o_p\left(\overline{V}_n(k)\right),$$

$$|I_3| = o_p\left(\lambda_k^2 \varepsilon_{nk}^2\right)\|m_k - m\| = o_p\left(\overline{V}_n(k)\right).$$

Another application of part (i) of Lemma 7.1 will show $I_2 = o_p(\overline{V}_n(k))$. □

Proof of Lemma 3.7. The proofs for (i), (ii) and (iii) are the same in the sense that the same arguments have to be repeated. So we will only prove (ii).

We will need the following two results in order to prove (ii):

$$\sup_k \sup_{h \in C_k} \|\hat{f}_h - \hat{f}_k\|_\infty = O_p(n^{-1})$$

and

$$\sup_{h \in C_k} \|\hat{f}_h - f_h\|_\infty = o_p\left(\sqrt{\lambda}\,(\log n)^{1/2+\delta}/\sqrt{n}\right)$$

$$\left(\lambda = h_1 \cdots h_p \text{ and } \|\cdot\|_\infty \text{ is the sup-norm}\right).$$

The second result follows from the first result, Lemma 3.6 and the facts that if $h \in C_k$, then $|\lambda - \lambda_k| = \lambda_k O(n^{-1})$ and $\sup_k \sup_{h \in C_k} \|f_h - f_k\|_\infty = O(n^{-1})$.

The proof of the first result is very similar to that of Lemma 3.6 and so we will skip its proof. Now let us prove (ii).

$$|L_n(h) - L_n(k)| \le \|\hat{m}_h - \hat{m}_k\|^2 + 2\|\hat{m}_h - \hat{m}_k\|\,\|\hat{m}_k - m\|.$$

Since $\sup_k |(L_n(k) - V_n(k))/V_n(k)| = o_p(1)$, it is enough to show that

$$\sup_{h \in C_k} \|\hat{m}_h - \hat{m}_k\|^2 = o_p\left(V_n(k)\right).$$

Now

$$
\begin{aligned}
&\|\hat{m}_h - \hat{m}_k\|^2 \\
&\quad = \int \Big[n^{-1} \sum Y_j \big\{ \lambda \hat{f}_h^{-1}(X_j) w\big(h(x - X_j)\big) \\
&\qquad\qquad -\lambda_k \hat{f}_k^{-1}(X_j) w\big(h(x - X_j)\big)\big\} \Big]^2 \, dF(x) \\
&\quad \le 8 \int \Big[n^{-1} \sum Y_j \lambda \hat{f}_h^{-1}(X_j) \big\{ w\big(h(x - X_j)\big) - w\big(h_k(x - X_j)\big) \big\} \Big]^2 \, dF(x) \\
&\qquad + 8 \int \Big[n^{-1} \sum Y_j \lambda \big\{ \hat{f}_h^{-1}(X_j) - \hat{f}_k^{-1}(X_j) \big\} w\big(h_k(x - X_j)\big) \Big]^2 \, dF(x) \\
&\qquad + 8 \int \left[\frac{1}{n} \sum Y_j (\lambda - \lambda_k) \hat{f}_k^{-1}(X_j) w\big(h_k(x - X_j)\big) \right]^2 \, dF(x) \\
&\quad = 8(I_1 + I_2 + I_3), \quad \text{say.}
\end{aligned}
$$

$$
\begin{aligned}
|I_1| \le \int \Big[n^{-1} \sum |Y_j| \lambda_k \| h(x - X_j) - h_k(x - X_j) \|^{\beta} \\
\times U\big(x - X_j, h_k^{-1}\mathbf{1}\big) \Big]^2 \, dF(x) (\lambda/\lambda_k) \sup_x \hat{f}_h^{-2}(x).
\end{aligned}
$$

For $h \in C_k$, $(\lambda/\lambda_k)^2 \sup_x \hat{f}_h^{-2}(x) = O_p(1)$. Hence

$$
|I_1| \le O_p(n^{-2-\beta}) \int \Big[n^{-1} \sum |Y_j| \lambda_k U\big(x - X_j, h_k^{-1}\mathbf{1}\big) \Big]^2 \, dF(x).
$$

If $\mu(x) = E(|Y|\,|X = x)$, then by analogy to $\hat{m}_k$ we will denote

$$
n^{-1} \sum |Y_j| \lambda_k U\big(x - X_j, h_k^{-1}\underline{1}\big) / \lambda_k \int U\big(x - x', h_k^{-1}\mathbf{1}\big) \, dF_n(x')
$$

by $\hat{\mu}_k(x)$. It is easy to show that $\int \hat{\mu}_k^2(x)\, dF(x) = O_p(1)$. Hence we have $|I_1| \le O_p(n^{-2-\beta}) = o_p(V_n(k))$. Arguing similarly, we get $I_2 = o_p(V_n(k))$ [since $\sup_k \sup_{h \in C_k} \| \hat{f}_k - \hat{f}_h \|_\infty = O_p(n^{-1})$] and $I_3 = o_p(V_n(k))$ [since $|\lambda - \lambda_k| = \lambda_k O_p(n^{-1})$ for $h \in C_k$]. □

Proof of Theorem 3.8. In order to prove this theorem we will need to check conditions (A1)–(A9). What we should keep in mind is that unlike other cases we have only one smoothing parameter here. It is easy to see that conditions (A1)–(A4) hold. Before we check the rest, let us write down the following useful lemma which is a corollary of Lemma 3.9.

Let

$$
A_1 = \big\{ (x, x_1) \colon 0 \le \lambda_k G(x, \|x - x_1\|) \le 1 - \varepsilon_{nk} \big\}
$$

and

$$
A_2 = \big\{ (x, x_1) \colon 1 - \varepsilon_{nk} \le \lambda_k G(x, \|x - x_1\|) \le 1 + \varepsilon_{nk} \big\}.
$$

LEMMA 7.3. *With probability approaching* 1:

(i) *On the set* A_1, $\lambda_k G_n(x, \|x - x_1\|) \le 1$.
(ii) *On the set* $(A_1 \cup A_2)^c$, $\lambda_k G_n(x, \|x - x_1\|) > 1$.

Note that (A9) follows from Lemma 7.2. Now let us proceed to prove (A5)–(A8).

(A5) It is quite easy to see that (i), (ii) and (iii) hold and hence we omit the proofs.

$$\text{(A6)} \qquad a_k(x_1, x_2) = \nu_k^{-2}\lambda_k^2 \int w(\lambda_k G_n(x, \|x - x_1\|)) \times w(\lambda_k G_n(x, \|x - x_2\|))\, dF_n(x).$$

It is quite easy to see that $\inf_k |\nu_k| > 0$. Let us note that $|a_k(x_1, x_2)| \le c\lambda_k$ for some $c > 0$ for all x_1 and x_2. Also because of Lemmas 3.9 and 7.3, there exists $M > 0$ such that $a_k = 0$ whenever $\|x_1 - x_2\|^p > M\lambda_k^{-1}$. By denoting I to be the indicator function of the set $\{(x_1, x_2): \|x_1 - x_2\|^p > M\lambda_k^{-1}\}$, we get

$$\lambda_k^{-1} \int a_k^2(x_1, x_2)\, dF_n(x_1)\, dF_n(x_2) \le c^2 \lambda_k \int I(x_1, x_2)\, dF_n(x_1)\, dF_n(x_2) = O_p(1).$$

The other part of (A6) can be similarly proved.

(A7) The proofs of all the parts are trivial and so we will skip them.

$$\begin{aligned}\text{(A8)} \quad V_n(k) - \bar{V}_n(k) &= 2\int (\tilde{m}_k - \bar{m}_k)(m_k - m)\, dF \\ &\quad + 2\int (\bar{m}_k - m_k)(m_k - m)\, dF \\ &= 2T_1 + 2T_2, \quad \text{say}, \\ T_2 &= \nu_k^{-1}\lambda_k \int w(\lambda_k G(x, \|x - x'\|)) m(x') \\ &\quad \times (m_k(x) - m(x))\, d(F_n - F)(x')\, dF(x).\end{aligned}$$

By part (i) of Lemma 7.1, $T_2 = o_p(\bar{V}_n(k))$.

For notational convenience we will write $g_k = m_k - m$.

$$T_1 = \nu_k^{-1}\lambda_k \int m(x_1)[w(\lambda_k G_n(x, \|x - x_1\|)) - w(\lambda_k G(x, \|x - x_1\|))] g_k(x)\, dF_n(x_1)\, dF(x).$$

Let us now refer to Lemma 7.3. T_1 can be written as the sum of integrals over the regions A_1 and A_2. Obviously the integral over the region A_2 is $o_p(\bar{V}_n(k))$. If w is uniform, then the proof ends here. If w is not uniform, then let w' be Hölder continuous. Let I_1 be the indicator function of the region A_1. By denoting

$$l_k(x, x_1) = \nu_k^{-1}\lambda_k^2 m(x_1) w'(\lambda_k G(x, \|x - x_1\|)) I_1(x, x_1)$$

and

$$(G_n - G)(x, r) = G_n(x, r) - G(x, r),$$

for any $r > 0$, we can write the integral over the region A_1 as

$$\int l_k(x, x_1)(G_n - G)(x, \|x - x_1\|)g_k(x)\, dF_n(x_1)\, dF(x) + o_p(\bar{V}_n(k))$$

$$= \int l_k(x, x_1)(G_n - G)(x, \|x - x_1\|)g_k(x)\, dF(x_1)\, dF(x)$$

$$+ \int l_k(x, x_1)(G_n - G)(x, \|x - x_1\|)g_k(x)\, d(F_n - F)(x_1)\, dF(x)$$

$$+ o_p(\bar{V}_n(k))$$

$$= T_{11} + T_{12} + o_p(\bar{V}_n(k)), \quad \text{say.}$$

By part (i) of Lemma 7.1, $T_{11} = o_p(\bar{V}_n(k))$. Now

$$T_{12} = \int \bar{l}_k(x_1, x_2)\, d(F_n - F)(x_1)\, d(F_n - F)(x_2),$$

where

$$\bar{l}_k(x_1, x_2) = \int l_k(x, x_1) I(x_2: \|x - x_2\| \le \|x - x_1\|) g_k(x)\, dF(x).$$

First note that $|\bar{l}_k(x_1, x_2)| \le c\lambda_k^{3/2}\|g_k\|$ for some $c > 0$. But $\bar{l}_k = 0$ whenever $\|x_1 - x_2\|^p > M_1\lambda_k^{-1}$ for some $M_1 > 0$. An application of Lemma 4.8 will give us

$$ET_{12}^{2s} \le c_1 n^{-4s}\Big\{ \sum_{l \le 2s} n^l \lambda_k^{3s - l + 1} + \sum_{l > 2s} n^{-l+2s}\lambda_k^{3s-l+1}\Big\}\|g_k\|^{2s}$$

$$\le c_2 n^{-2s}\lambda_k^{s+1}\|g_k\|^{2s} \quad (\text{since } \lambda_k < n).$$

This obviously shows $T_{12} = o_p(\bar{V}_n(k))$ and this completes the proof of Theorem 3.8. □

Acknowledgments. The authors would like to thank Professor Charles J. Stone for his encouragement during the course of this work and the referees for making valuable suggestions for improving the presentation of this paper.

REFERENCES

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, Calif.

BURMAN, P. (1985). A data dependent approach to density estimation. *Z. Wahrsch. verw. Gebiete* **69** 609–628.

CHEN, K.-W. (1983). Asymptotically optimal selection of a piecewise polynomial estimator of a regression function. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

DE BOOR, C. (1978). *A Practical Guide to Splines.* Springer, New York.

HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1485.

JACKSON, D. (1930). *Theory of Approximations. Amer. Math. Soc. Colloq. Publ.* **11.** Amer. Math. Soc., New York.

LI, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** 958–975.

MARDIA, K., KENT, J. and BIBBY, J. (1979). *Multivariate Analysis.* Academic, New York.

MARRON, J. S. and HÄRDLE, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multivariate Anal.* **20** 91–113.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

DIVISION OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NORTH CAROLINA
CHARLOTTE, NORTH CAROLINA 28213