Statistica Sinica **6**(1996), 603-615

A NONPARAMETRIC MULTI-STEP PREDICTION ESTIMATOR IN MARKOVIAN STRUCTURES

Rong Chen

Texas A & M University

Abstract: In this paper, a multistage kernel smoother is proposed to estimate the conditional mean $E(Z \mid X)$ in a Markovian structure where the observations (X_i, Y_i, Z_i) are i.i.d. samples from a distribution that possesses the Markov property $E(Z \mid Y, X) = E(Z \mid Y)$. We prove that the asymptotic mean squared error of the proposed estimator is smaller than that using the Nadaraya-Watson estimator directly on the pairs (X_i, Z_i). A simulation study is also given.

Key words and phrases: Kernel smoothing, mean squared errors, multi-stage smoother.

1. Introduction

In this paper we study the nonparametric multi-step ahead prediction problem in a Markovian structure. Specifically, assume the joint distribution of (X, Y, Z) possesses the Markov property $E(Z \mid Y, X) = E(Z \mid Y)$. We are interested in predicting a future observation Z given $X = x$, without the knowledge of Y, while a complete data set (X_i, Y_i, Z_i), $i = 1, \ldots, n$, is available.

This problem is of interest in many situations. For example, it arises from the measurement error models where Z is the response variable, X is the predictor measured with error and Y is the corresponding true predictor. The underlying assumption is that the response Z is related to X only through the true predictor Y. Most of the time the true predictor Y is difficult and expensive to observe. Hence it is of interest to predict Z using only X, the predictor measured with error. The prediction function $m(x)$ is to be estimated nonparametrically using the complete data set where both Y and X are available.

A similar problem can be found in nonlinear time series analysis. With a first order nonlinear autoregressive model $X_t = f(X_{t-1}) + \varepsilon_t$, two-step ahead least squares prediction requires one to estimate $E(X_{N+2} \mid X_N = x)$ using the triples (X_{t+2}, X_{t+1}, X_t) for $t = 1, \ldots, N-2$. However, the correlation between the triples complicates the problem. In this paper, we concentrate on the regression problem where available data are i.i.d. triples.

The best predictor in the least squares sense is, of course, the conditional mean $m(x) = E(Z \mid X = x)$. There is a very large body of literature on the

nonparametric estimation of the conditional mean function in a regression setting. For example, see Eubank (1988), Härdle (1990), Hall (1984), Stone(1980), Cleveland and Devlin (1988), Fan (1993) and the references therein. However, researchers have been focused mainly on estimating $m(x)$ directly from the pairs (X_i, Z_i), which ignores the information on the variable Y. One commonly used estimator is the Nadaraya-Watson (N-W) kernel estimator (Nadaraya (1964), Watson (1964)).

$$\tilde{m}_h(x) = \frac{\sum_{i=1}^n K((x - X_i)/h)Z_i}{\sum_{i=1}^n K((x - X_i)/h)}, \tag{1}$$

where $K(\cdot)$ is a Kernel function and h is the smoothing parameter. Note that the variable Y contains substantial information, particularly when (X, Y, Z) possesses the Markov property. In this paper, we propose a multistage kernel smoother which utilizes the information on Y to estimate the conditional mean function $E(Z \mid X)$; and we prove that the proposed smoother has a smaller mean squared error than the direct estimator (1).

Let $f(y) = E(Z \mid Y = y)$. Due to the Markov property, we have

$$\begin{aligned} m(x) \equiv E(Z \mid X = x) &= E[E(Z \mid Y, X) \mid X = x] = E[E(Z \mid Y) \mid X = x] \\ &= E(f(Y) \mid X = x). \end{aligned}$$

Ideally, if the function $f(\cdot)$ were known, we would use the pairs $\{X_i, f(Y_i)\}$, $i = 1, \ldots, n$ to estimate $E[f(Y) \mid X] = E(Z \mid X)$, since $f(Y)$ has all the available information of $m(\cdot)$, with less variation, due to the fact that $\mathrm{Var}\,[f(Y) \mid X = x] \le \mathrm{Var}\,(Z \mid X = x)$. When we use the pairs (X_i, Z_i) to estimate $E[f(Y) \mid X]$ directly as in (1), we are actually using Z_i in place of $f(Y_i)$, which has an error rate of $O(1)$. On the other hand, if we estimate the function $f(\cdot)$ with a suitable estimator $\hat{f}(\cdot)$ that has a smaller error rate than Z and use the pairs $\{X_i, \hat{f}(Y_i)\}$ to estimate $E[f(Y) \mid X]$, we obtain smaller errors. We shall prove that this is indeed the case. In fact, with a proper estimator of $f(\cdot)$, the mean squared error of the proposed estimator is "almost" the same as if the true function $f(\cdot)$ were known. The above observation motivates the following estimator for $E(Z \mid X)$, which will be referred to as the "multistage smoother". It is defined as

$$\hat{m}_{h_1,h_2}(x) = \frac{\sum_{k=1}^n K((x - X_k)/h_2)\hat{f}_{h_1}(Y_k)}{\sum_{k=1}^n K((x - X_k)/h_2)}, \tag{2}$$

where

$$\hat{f}_{h_1}(y) = \frac{\sum_{j=1}^n K((y - Y_j)/h_1)Z_j}{\sum_{j=1}^n K((y - Y_j)/h_1)}.$$

Note that $\hat{f}$ is the regular N-W kernel estimator of $f(y) = E(Z \mid Y = y)$. Here, different smoothing parameters are used for each stage of smoothing.

The above estimator can be derived in the following way, analogous to that of the N-W estimator (Eubank (1988), p169-170). Let $p_X(x)$, $p_Y(y)$ be the marginal densities of X and Y respectively and $p_{X,Y}(x,y)$ and $p_{Y,Z}(x,z)$ be the joint densities of (X,Y) and (Y,Z) respectively. If we are to use the following one and two dimensional multiplicative kernel density estimators

$$\hat{p}_X(x) = \frac{1}{nh_2}\sum_{k=1}^{n} K((x-X_k)/h_2); \qquad \hat{p}_Y(y) = \frac{1}{nh_1}\sum_{k=1}^{n} K((y-Y_k)/h_1);$$

$$\hat{p}_{X,Y}(x,y) = \frac{1}{nh_1h_2}\sum_{k=1}^{n} K((x-X_k)/h_2)K((y-Y_k)/h_1);$$

and

$$\hat{p}_{Y,Z}(y,z) = \frac{1}{nh_1h_3}\sum_{k=1}^{n} K((y-Y_k)/h_1)K((z-Z_k)/h_3),$$

then, under the Markovian structure $p_{Z|X,Y}(z \mid x,y) = p_{Z|Y}(z \mid y)$, we have

$$\begin{aligned}
m(x) &= \frac{1}{p_X(x)}\int\int z p_{Y,Z}(y,z)dz\frac{p_{X,Y}(x,y)}{p_Y(y)}dy \\
&\sim \frac{1}{\hat{p}_X(x)}\int\int z\frac{1}{nh_1h_3}\sum_{j=1}^{n} K((y-Y_j)/h_1)K((z-Z_j)/h_3)dz\frac{p_{X,Y}(x,y)}{p_Y(y)}dy \\
&\sim \frac{1}{\hat{p}_X(x)}\int\frac{1}{nh_1}\sum_{j=1}^{n} K((y-Y_j)/h_1)Z_j\frac{p_{X,Y}(x,y)}{p_Y(y)}dy \\
&\sim \frac{1}{\hat{p}_X(x)}\int\frac{1}{nh_1}\sum_{j=1}^{n} K((y-Y_j)/h_1) \\
&\qquad\qquad \times Z_j\frac{nh_1}{nh_1h_2}\frac{\sum_{k=1}^{n}K((x-X_k)/h_2)K((y-Y_k)/h_1)}{\sum_{l=1}^{n}K((y-Y_l)/h_1)}dy \\
&\sim \frac{1}{\hat{p}_X(x)}\frac{1}{nh_2}\sum_{j=1}^{n} Z_j\sum_{k=1}^{n}\Big[\frac{K((x-X_k)/h_2)K((Y_k-Y_j)/h_1)}{\sum_{l=1}^{n}K((Y_k-Y_l)/h_1)}\Big] \qquad (3) \\
&= \frac{1}{\hat{p}_X(x)}\frac{1}{nh_2}\sum_{k=1}^{n} K((x-X_k)/h_2)\Big[\frac{\sum_{j=1}^{n}K((Y_k-Y_j)/h_1)Z_j}{\sum_{l=1}^{n}K((Y_k-Y_l)/h_1)}\Big],
\end{aligned}$$

which is that in (2). Also note that by (3), $\hat{m}(x)$ is a linear smoother.

The rest of the paper is organized as follows. In Section 2, the pointwise and the integrated mean squared error of the multistage estimator (2) is calculated and compared to that of the N-W estimator (1). Section 3 carries out a simulation study for empirical comparisons of the two estimators. Section 4 provides a brief summary. The detailed conditions and an outline of the proof of the main theorem are given in the appendix.

2. Asymptotic Properties

Let $k_1 = \int K^2(z)dz$, $k_2 = \int z^2 K(z)dz$. The following conditional means and variances are used.

$$m(x) = E(Z \mid X = x); \quad f(y) = E(Z \mid Y = y); \quad v(y) = \mathrm{Var}\,(Z \mid Y = y);$$

$$u(x) = \mathrm{Var}\,(f(Y) \mid X = x); \; w(x) = E(v(Y) \mid X = x); \; \sigma^2(x) = \mathrm{Var}\,(Z \mid X = x).$$

It is important to note that $\sigma^2(x) = u(x) + w(x)$, due to the Markov property. We have the following theorems.

Theorem 1. *Under Conditions* (C1)-(C5) *given in the Appendix, if* $nh_1 \to \infty$, $h_1 = o(h_2)$, *and* $h_2 = \{D_1(x)/4D_2(x)\}^{1/5} n^{-1/5}$, *where* $D_1(x) = k_1 u(x)/p_X(x)$ *and* $D_2(x) = k_2^2\{m'(x)p'_X(x) + \frac{1}{2}m''(x)p_X(x)\}^2/p_X^2(x)$, *then for a given* x, *the asymptotic mean squared error of estimator* (2) *is*

$$E(\hat{m}(x) - m(x))^2 = D_1(x)^{4/5} D_2(x)^{1/5}(4^{1/5} + 4^{-4/5})n^{-4/5} + o(n^{-4/5}). \qquad (4)$$

An outline of the proof of the theorem is given in the Appendix.

Corollary 1. *The ratio of the minimum asymptotic mean squared errors of estimators* (1) *and* (2) *is*

$$r(x) = \Big\{1 + \frac{w(x)}{u(x)}\Big\}^{4/5}.$$

Proof. Let $D_3(x) = k_1\sigma^2(x)/p_X(x) = k_1\{u(x) + w(x)\}/p_X(x)$. It is well known (e.g. Collomb (1977), Härdle (1990)) that, for $h = (D_3/4D_2)^{1/5} n^{-1/5}$, the mean squared error of the N-W estimator (1) is

$$E(\tilde{m}(x) - m(x))^2 = D_3(x)^{4/5} D_2(x)^{1/5}(4^{1/5} + 4^{-4/5})n^{-4/5} + o(n^{-4/5}).$$

The result follows immediately by comparing the above expression with (4).

Note that, if the true function $f(\cdot)$ is known, then the asymptotic mean squared error of estimating $E[f(Y) \mid X = x]$ using the N-W estimator on the pairs $\{X_i, f(Y_i)\}$ is exactly that in (4). Hence, asymptotically, the proposed multistage smoother provides results as good as if we knew the function $f(\cdot)$. The corollary also shows that although the mean squared error of the two estimators are of the same order $O(n^{-4/5})$, $\hat{m}(x)$ is smaller by a factor which depends on the ratio of the conditional variances $w(x)$ and $u(x)$. Note that the asymptotic result is quite insensitive to the extra bandwidth h_1. As long as $nh_1 \to \infty$ and $h_1 = o(h_2)$, the result holds.

Theorem 2. *Let* $D_1 = \int_A D_1(x)dx$ *and* $D_2 = \int_A D_2(x)dx$, *where* A *is the interval of interest. Under Conditions* (C1)-(C5), *if* $h_2 = (D_1/4D_2)^{1/5}n^{-1/5}$ *and* $nh_1 \to \infty$, $h_1 = o(h_2)$, *then the integrated mean squared error of estimator* (2) *is*

$$\int_A E(\hat{m}(x) - m(x))^2 dx = D_1^{4/5} D_2^{1/5}(4^{1/5} + 4^{-4/5})n^{-4/5} + o(n^{-4/5}).$$

Corollary 2. *The ratio of the minimum asymptotic integrated mean squared errors of estimators* (1) *and* (2) *is*

$$r = \Big\{1 + \frac{\int_A w(x)/p_X(x)dx}{\int_A u(x)/p_X(x)dx}\Big\}^{4/5}. \tag{5}$$

3. Simulation Study

In this section, we compare the estimators (1) and (2) through a simulation study. Consider the following model:

$$X \sim \text{Uniform}(-3, 3); \quad Y = 3\sin(X) + \varepsilon; \quad Z = c_1 \sin(Y) + c_2 e; \tag{6}$$

where ε and e follow the standard normal distribution. For each combination of (c_1, c_2) where $c_1 = 3, 6$ and $c_2 = 1, 3, 6$, two hundred sets of samples, each with three hundred observations, are generated. For each set, estimators (1) and (2) are evaluated at 250 equally spaced points on the interval $(-2.5, 2.5)$ using a quartic Kernel $K(u) = 0.9375(1-u^2)^2 I(|u| \le 1|)$. We did not use the full interval $[-3, 3]$ to avoid the complication of the edge effect. For the multistage estimator, the bandwidth h_1 for estimating $f(\cdot)$ is chosen to be h^*/b, $b = 1, 5, 10$, where h^* is the optimal generalized cross validation (GCV) bandwidth (Craven and Wahba (1979), Li (1985)) for estimating $E(Z|Y)$ using the pairs (Y_i, Z_i). The GCV is computed using the WARPing approximation (Härdle (1990)). Note that by Theorem 1, the order of h_1 should be smaller than that of the usual optimal bandwidth. The second bandwidth h_2 is chosen to be the optimal GCV bandwidth for the pairs $\{X_i, \hat{f}(Y_i)\}$. For the N-W estimator (1), the optimal GCV bandwidth is used. For each of the 250 points evaluated, the squared errors are computed to the true conditional mean $m(x) = c_1 e^{-1/2}\sin\{3\sin(x)\}$. Then the average of those squared errors are computed, denoted by mse_1 and mse_2, for the ordinary N-W estimator (1) and the multistage estimator (2), respectively. Then the ratio $r = mse_1/mse_2$ is obtained. Table 1 shows the first quartile, the median and the third quartile of the ratios r from the two hundred sets of samples under each coefficient and bandwidth combination. The column under "true" is the theoretical asymptotic ratio

$$r_{(c_1,c_2)}(x) = \Big(1 + \frac{2c_2^2}{c_1^2(1-e^{-1})[1 + e^{-1}\int \cos\{6\sin(x)\}dx]}\Big)^{4/5},$$

using (5) for the above model on the interval $(-2.5, 2.5)$.

Table 1. The quartiles (first, median, third) of the ratios r using model (6) with sample size 300

(c_1, c_2)	true	$h_1 = h^*$			$h_1 = h^*/5$			$h_1 = h^*/10$		
(3,1)	1.25	1.03	1.19	1.39	1.01	1.15	1.32	1.00	1.11	1.24
(3,3)	2.98	1.60	2.06	2.81	1.39	1.69	2.04	1.19	1.41	1.72
(3,6)	7.62	2.08	2.89	4.38	1.61	2.05	2.79	1.38	1.68	2.14
(6,1)	1.06	0.96	1.04	1.12	0.99	1.01	1.09	0.97	1.02	1.08
(6,3)	1.55	1.18	1.44	1.85	1.13	1.32	1.53	1.07	1.24	1.44
(6,6)	2.98	1.58	2.06	2.73	1.22	1.55	1.88	1.12	1.28	1.54

From the table we can see that the performance of multistage smoother depends on the ratio of c_1/c_2. The improvement can be quite significant in some cases. With finite samples, the bandwidth h_1 does have certain effect on the improvement rate. It is surprised to see that the $h_1 = h_1^*$ provides the best improvement, though theoretically the bandwidth should be of smaller order than the optimal bandwidth for the regular smoother. Note that, the cross-validation selection of bandwidth was designed for one-stage smoothing. It may not be a good choice in our situation. A more sophisticated bandwidth selection procedure may be needed. We tried a leave-one-out cross-validation procedure to select h_1 and h_2 simultaneously. The computation is much more time consuming and there is no apparent improvement over the above procedure which selects the bandwidth separately.

We observe that the sample size has little effect on the improvement rate. Table 2 show the improvement rate using sample size 100, 300 and 500 in the above example. The optimal GCV bandwidth is used for the first stage smoothing in this example.

Table 2. The quartiles (first, median, third) of the ratios r using model (6) with different sample sizes

(c_1, c_2)	true	100			300			500		
(3,1)	1.25	1.02	1.21	1.39	1.03	1.19	1.39	1.04	1.20	1.38
(3,3)	2.98	1.58	1.98	2.82	1.60	2.06	2.81	1.70	2.23	2.83
(3,6)	7.62	1.91	2.81	4.06	2.08	2.89	4.38	2.33	3.19	4.92
(6,1)	1.06	1.00	1.08	1.17	0.96	1.04	1.12	0.99	1.06	1.14
(6,3)	1.55	1.18	1.40	1.77	1.18	1.44	1.85	1.16	1.36	1.65
(6,6)	2.98	1.70	2.17	2.89	1.58	2.06	2.73	1.60	2.03	2.68

To see the effect of the function $E(Y \mid X)$ and $\mathrm{Var}\,(Y \mid X)$, we investigated the model

$$X \sim Uniform(-3, 3); \quad Y = c_3 \sin(X) + c_4 \varepsilon; \quad Z = 3 \sin(Y) + 3e; \qquad (7)$$

with different combinations of (c_3, c_4). The ε and e are generated from the standard normal distribution. Table 3 is constructed the same way as Table 1, using sample size 300. We observe similar results.

Table 3. The quartiles (first, median, third) of the ratios r using model (7) with sample size 300

(c_3, c_4)	true	$h_1 = h^*$			$h_1 = h^*/5$			$h_1 = h^*/10$		
(2,0.5)	8.76	1.87	2.99	4.64	1.62	2.27	3.25	1.40	1.91	2.58
(2,1.0)	3.51	1.68	2.46	3.79	1.52	2.08	2.89	1.32	1.80	2.47
(2,2.0)	2.44	1.59	2.23	2.96	1.32	1.79	2.32	1.21	1.54	1.96
(3,0.5)	5.95	1.61	2.23	3.03	1.38	1.81	2.47	1.22	1.54	2.05
(3,1.0)	3.05	1.58	2.05	2.82	1.35	1.68	2.12	1.16	1.42	1.78
(3,2.0)	2.43	1.52	1.98	2.45	1.30	1.54	1.89	1.17	1.35	1.61

4. Summary

In this paper, we propose a multistage kernel smoother to estimate the conditional mean function in a Markovian structure. It is shown, both theoretically and empirically, that the proposed smoother has smaller mean squared error than the regular smoother. We point out that the proposed smoother is not restricted to two-stage smoothing. If the Markov chain is longer, say, $(X_1, \ldots, X_p)$, then a $p-1$ stage smoother can be used to estimate the $p-1$ step ahead prediction $E(X_p \mid X_1)$. It is also simple to extend the estimator to the cases where X or Y are multi-dimensional.

Acknowledgement

This work is partially supported by the NSF grant DMS-9301193. The author wishes to thank the associate editor and the referee for helpful comments.

Appendix. Conditions and Proofs

Slightly different from the notation we used before, let $p(x, y)$ be the joint density of (X, Y) and $p_1(y), p_2(x)$ the marginal densities of Y and X respectively. The derivatives of the functions are denoted by $^{(k)}$. For example, $m^{(1)}(x)$ is the first derivative of $m(\cdot)$ evaluated at x.

The following functions are needed.

$$
\begin{aligned}
g_1(x) &= m^{(1)}(x)p_2^{(1)}(x) + \frac{1}{2}m^{(2)}(x)p_2(x);\\
g_2(x) &= \int \frac{p(x,y)}{p_1(y)}dy; \qquad g_3(x) = \int \frac{p^2(x,y)}{p_1^2(y)}dy;\\
g_4(x) &= E\Big[\frac{f^{(1)}(Y)p_1^{(1)}(Y) + \frac{1}{2}f^{(2)}(Y)p_1(Y)}{p_1(Y)} \mid X = x\Big].
\end{aligned}
$$

The set of interest is a finite interval $A = [a_1, a_2]$, i.e. we are only interested in estimating $E(Z \mid X = x)$ for $x \in A$. Let $A_\varepsilon = [a_1 - \varepsilon, a_2 + \varepsilon]$. The following conditions are assumed.

(C1): $K(\cdot)$ is a bounded symmetric density function with compact support.

(C2): The marginal density $p_1(x)$ of X has finite second derivative and is bounded away from zero in A_ε. The joint density of $p(x, y)$ is Hölder continuous both in the x and y variables, i.e. $|p(x_1, y) - p(x_2, y)| \le c_1|x_1 - x_2|^{\gamma_1}$ uniformly for y and $|p(x, y_1) - p(x, y_2)| \le c_2|y_1 - y_2|^{\gamma_2}$ uniformly for x where $\gamma_1 > 0$ and $\gamma_2 > 0$.

(C3): The functions $m(x) = E(Z \mid X = x)$ and $f(y) = E(Z \mid Y = y)$ are twice differentiable and the second derivative is Hölder continuous such that $|f^{(2)}(y_1) - f^{(2)}(y_2)| \le c_3|y_1 - y_2|^{\gamma_3}$, $|m^{(2)}(x_1) - m^{(2)}(x_2)| \le c_4|x_1 - x_2|^{\gamma_4}$ where $\gamma_3 > 0$ and $\gamma_4 > 0$. In addition, $f(y)$ is Hölder continuous such that $|f(y_1) - f(y_2)| \le c_5|y_1 - y_2|^{\gamma_5}$.

(C4): The functions $\sigma^2(x)$, $u(x)$, and $w(x)$ are all well defined and bounded in A_ε. The function $v(y)$ is bounded.

(C5): The functions $g_i(x)$, $i = 1, \ldots, 4$ are all well defined and bounded in A_ε.

Conditions (C1) to (C4) are standard in nonparametric inference. In Condition (C5), the requirement for $g_1(x)$ is usual. The conditions for $g_2(x)$ to $g_4(x)$ are less obvious. Note $p(x, y)/p_1(y) = p_{X|Y}(x \mid y)$. Hence these conditions concern the conditional density of X given $Y = y$ as a function of y for fixed x. Note that, given Condition (C2) and (C3), when the marginal density $p_1(y)$ is bounded below from zero on a finite support, or when $p(x, y)$ follows a joint normal distribution, then g_2, g_3 and g_4 are bounded in A_ε.

Let $\hat{p}_2(x) = (nh)^{-1} \sum_{k=1}^n K((x - X_k)/h)$ and $\hat{p}_1(y) = (nh)^{-1} \sum_{k=1}^n K((y - Y_k)/h)$. To prove the main theorem, we need the following lemmas.

Lemma 1. *Under Conditions* C1-C5, *and* $h \to 0$ *and* $nh \to \infty$,

$$|\hat{p}_2(x) - p_2(x)| \to 0 \quad \textit{and} \quad |\hat{p}_1(y) - p_1(y)| \to 0 \ \textit{in probability.}$$

Conditional on Y_1, *we have* $|\hat{p}_1(Y_1) - p_1(Y_1)| \to 0$ *a.s.*

These are well known results. For example, see Parzen (1962), Nadaraya (1964), Silverman(1986). The last result is due to the fact that $\hat{p}_1(\cdot)$ is independent of Y_1.

In what follows, $A_n \sim B_n$ means $A_n = B_n + o_p(B_n)$, i.e. A_n equals B_n plus a term that goes to zero in a faster order than B_n as n goes to ∞. Note that if $A_n \sim B_n$ and B_n has finite support, then $E(A_n) = E(B_n) + o(E(B_n))$. In this case, we write $E(A_n) \sim E(B_n)$ as well. Also note that if $A_n \sim B_n$, then $A_n^2 \sim B_n^2$.

Lemma 2. *Under Conditions* C1-C5 *and* $n \to \infty$, $h \to 0$, $nh \to \infty$, *for any identically distributed random variables* W_i, *we have*

$$\frac{\sum_{i=1}^n K((x-X_i)/h)W_i}{\sum_{i=1}^n K((x-X_i)/h)} \sim \frac{1}{nh}\frac{\sum_{i=1}^n K((x-X_i)/h)W_i}{p_2(x)}.$$

Proof. Following Härdle and Marron(1985),

$$\begin{aligned}&\frac{\sum_{i=1}^n K((x-X_i)/h)W_i}{\sum_{i=1}^n K((x-X_i)/h)}\\ =\ &\frac{1}{nh}\frac{\sum_{i=1}^n K((x-X_i)/h)W_i}{p_2(x)} + \frac{1}{nh}\frac{\sum_{i=1}^n K((x-X_i)/h)W_i}{p_2(x)}\Big(\frac{p_2(x)}{\hat{p}_2(x)} - 1\Big).\end{aligned}$$

By Lemma 1, it is easy to see that the second term is negligible compared to the first.

Lemma 3. *Under Conditions* C1-C5 *and* $n \to \infty$, $h \to 0$, $nh \to \infty$, *for any identically distributed random variables* Z_j, *we have*

$$\frac{\sum_{j=1}^n K((Y_1-Y_j)/h)Z_j}{\sum_{j=1}^n K((Y_1-Y_j)/h)} \sim \frac{1}{nh}\frac{\sum_{j=1}^n K((Y_1-Y_j)/h)Z_j}{p_1(Y_1)}.$$

Lemma 4. *Under Conditions* (C1) *and* (C3), *if* $h_1 \to 0$, *then*

$$K(z)\{f(Y+h_1z) - f(Y)\} = o(1)K(z).$$

This is due to the fact that $K(\cdot)$ has finite support.

Proof of Theorem 1. Let $\varepsilon_j = Z_j - f(Y_j)$. We have ε_j i.i.d. with $E(\varepsilon_j) = 0$ and $\mathrm{Var}\,(\varepsilon_j \mid Y_j) = v(Y_j) = O(1)$, by Condition C4. In the following derivation, we repeatedly use lemmas 2 and 3 and the fact that (X_i, Y_i, Z_i) are i.i.d. and $\int zK(z)dz = 0$. The notations k_1 and k_2 are defined in Section 2. We write $K_h(u) = K(u/h)/h$.

For fixed $x \in A$, the mean squared error is

$$\begin{aligned}E[\hat{m}(x) - m(x)]^2 \ &= \ E\Big[\Big\{\hat{m}(x) - \frac{\sum_{k=1}^n K_{h_2}(x-X_k)f(Y_k)}{\sum_{k=1}^n K_{h_2}(x-X_k)}\Big\}\\ &\quad + \Big\{\frac{\sum_{k=1}^n K_{h_2}(x-X_k)f(Y_k)}{\sum_{k=1}^n K_{h_2}(x-X_k)} - m(x)\Big\}\Big]^2\\ &= \ E\Big[\frac{\sum_{k=1}^n K_{h_2}(x-X_k)\{\hat{f}(Y_k) - f(Y_k)\}}{\sum_{k=1}^n K_{h_2}(x-X_k)}\Big]^2\\ &\quad + E\Big[\frac{\sum_{k=1}^n K_{h_2}(x-X_k)\{f(Y_k) - m(x)\}}{\sum_{k=1}^n K_{h_2}(x-X_k)}\Big]^2\end{aligned}$$

$$+2E\Big[\Big(\frac{\sum_{k=1}^n K_{h_2}(x-X_k)\{\hat f(Y_k)-f(Y_k)\}}{\sum_{k=1}^n K_{h_2}(x-X_k)}\Big) \times\Big(\frac{\sum_{k=1}^n K_{h_2}(x-X_k)\{f(Y_k)-m(x)\}}{\sum_{k=1}^n K_{h_2}(x-X_k)}\Big)\Big]$$
$$\equiv A+B+2C. \tag{8}$$

Here we outline the calculation of term A in detail. The approximation of term B is standard and well known (e.g. Härdle (1989)). The calculation of term C is similar to term A and hence omitted. By Lemma 2, we have

$$\begin{aligned} A &\sim \frac{1}{p_2^2(x)}E\Big[\frac{1}{n}\sum_{k=1}^n K_{h_2}(x-X_k)\{\hat f(Y_k)-f(Y_k)\}\Big]^2 \\ &= \frac{1}{n^2p_2^2(x)}\Big\{nE[K_{h_2}(x-X_1)\{\hat f(Y_1)-f(Y_1)\}]^2 + \\ &\quad +n(n-1)(E[K_{h_2}(x-X_1)\{\hat f(Y_1)-f(Y_1)\}K_{h_2}(x-X_2)\{\hat f(Y_2)-f(Y_2)\}])\Big\} \\ &\equiv \frac{1}{n^2p_2^2(x)}\{nA_1+n(n-1)A_2\}, \end{aligned}$$

where, by Lemma 3,

$$\begin{aligned} A_1 &= E\Big[K_{h_2}(x-X_1)\frac{\sum_{j=1}^n K_{h_1}(Y_1-Y_j)\{Z_j-f(Y_1)\}}{\sum_{j=1}^n K_{h_1}(Y_1-Y_j)}\Big]^2 \\ &\sim \frac{1}{n^2}E\Big[\frac{K_{h_2}(x-X_1)}{p_1(Y_1)}[\sum_{j=1}^n K_{h_1}(Y_1-Y_j)\varepsilon_j \\ &\qquad + \sum_{j=2}^n K_{h_1}(Y_1-Y_j)\{f(Y_j)-f(Y_1)\}]\Big]^2 \\ &\equiv \frac{1}{n^2}E\Big[\frac{K_{h_2}(x-X_1)}{p_1(Y_1)}(A_{11}+A_{12})\Big]^2 \\ &= \frac{1}{n^2}E\Big[\frac{K_{h_2}^2(x-X_1)}{p_1^2(Y_1)}(A_{11}^2+A_{12}^2+2A_{11}A_{12})\Big]. \end{aligned}$$

It is obvious that $E(A_{11}A_{12})=0$, due to the independence of the ε_i's. Taking the expectation on ε_k and Y_k, $k=2,\dots,n$, conditioning on Y_1, we have

$$\begin{aligned} E[A_{11}^2 \mid Y_1] &= \frac{1}{h_1^2}K^2(0)E[\varepsilon_1^2 \mid Y_1] + (n-1)E_{Y_2}[K_{h_1}^2(Y_1-Y_2)v(Y_2) \mid Y_1] \\ &= O(1/h_1^2)+O(n/h_1) \end{aligned}$$

and

$$E[A_{12}^2 \mid Y_1] = E\Big[\Big(\sum_{j=2}^n K_{h_1}(Y_1-Y_j)\{f(Y_j)-f(Y_1)\}\Big)^2\Big|Y_1\Big]$$

$$= (n-1)E_{Y_2}[K_{h_1}^2(Y_1-Y_2)\{f(Y_2)-f(Y_1)\}^2 \mid Y_1] + (n-1)(n-2)(E_{Y_2}[K_{h_1}(Y_1-Y_2)\{f(Y_2)-f(Y_1)\} \mid Y_1])^2,$$

where, by Condition C3, the first term can be easily shown to be $o(n/h_1)$ and the second term $o(n^2)$. From now on we assume that $nh_1 \to \infty$. It will be shown later that we, indeed, need this condition. Under this assumption, we have

$$\begin{aligned} A_1 &\sim O(\frac{1}{nh_1})E_{X_1,Y_1}\Big[\frac{K_{h_2}^2(x-X_1)}{p_1(Y_1)}\Big] + o(1)E_{X_1,Y_1}\Big[\frac{K_{h_2}^2(x-X_1)}{p_1^2(Y_1)}p_1^2(Y_1)\Big] \\ &= O(\frac{1}{nh_1h_2})\int K^2(z)\frac{p(x+h_2z,y)}{p_1(y)}dzdy + o(1)\int K_{h_2}^2(x-X_1)dF(X_1) \\ &= O(1/nh_1h_2)(g_2(x)+o(1)) + o(1/h_2) \\ &= O(1/nh_1h_2) + o(1/h_2). \end{aligned}$$

For A_2, we have

$$\begin{aligned} A_2 &\sim \frac{1}{n^2}E\Big[\frac{K_{h_2}(x-X_1)K_{h_2}(x-X_2)}{p_1(Y_1)p_1(Y_2)} \times \\ &\qquad \Big(\sum_{j=1}^n K_{h_1}(Y_1-Y_j)\{Z_j-f(Y_1)\}\Big)\Big(\sum_{j=1}^n K_{h_1}(Y_2-Y_j)\{Z_j-f(Y_2)\}\Big)\Big] \\ &= \frac{1}{n^2}E\Big[\frac{K_{h_2}(x-X_1)K_{h_2}(x-X_2)}{p_1(Y_1)p_1(Y_2)} \times [2K_{h_1}(0)\varepsilon_1 K_{h_1}(Y_1-Y_2)\{Z_1-f(Y_2)\} \\ &\quad +(n-2)K_{h_1}(Y_1-Y_3)\{Z_3-f(Y_1)\}K_{h_1}(Y_2-Y_3)\{Z_3-f(Y_2)\} \\ &\quad +K_{h_1}^2(0)\varepsilon_1\varepsilon_2 + K_{h_1}(Y_2-Y_1)\{Z_2-f(Y_1)\}K_{h_1}(Y_1-Y_2)\{Z_1-f(Y_2)\} \\ &\quad +2(n-2)K_{h_1}(0)\varepsilon_1 K_{h_1}(Y_2-Y_3)\{Z_3-f(Y_2)\} \\ &\quad +2(n-2)K_{h_1}(Y_2-Y_1)\{Z_2-f(Y_1)\}K_{h_1}(Y_3-Y_2)\{Z_3-f(Y_2)\} \\ &\quad +(n-2)(n-3)K_{h_1}(Y_3-Y_1)\{Z_3-f(Y_1)\}K_{h_1}(Y_4-Y_2)\{Z_4-f(Y_2)\}]\Big] \\ &\equiv \frac{1}{n^2}(2A_{21}+(n-2)A_{22}+A_{23}+A_{24}+2(n-2)A_{25}+2(n-2)A_{26} \\ &\qquad +(n-2)(n-3)A_{27}). \end{aligned}$$

It is obvious that $A_{23}=0$ and $A_{25}=0$. Since $E_{Z_1}[\varepsilon_1\{Z_1-f(Y_2)\} \mid Y_1,Y_2] = E_{Z_1}[\varepsilon_1^2 \mid Y_1,Y_2] = v(Y_1) = O(1)$ and let $X_1 = x+h_2z_1$, $X_2 = x+h_2z_2$ and $Y_2 = Y_1+h_1z_3$, we can show

$$\begin{aligned} A_{21} &= K_{h_1}(0)E\Big[\frac{K_{h_2}(x-X_1)}{p_1(Y_1)}\frac{K_{h_2}(x-X_2)}{p_1(Y_2)}v(Y_1)K_{h_1}(Y_2-Y_1)\Big] \\ &= O(1)\frac{1}{h_1}\int K(z_1)K(z_2)K(z_3)\frac{p(x+h_2z_1,y_1)p(x+h_2z_2,y_1+h_1z_3)}{p_1(y_1)p_1(y_1+h_1z_3)}dz_1dz_2dz_3dy_1 \end{aligned}$$

$$= O(1/h_1)\int \Big\{\frac{p^2(x,y_1)}{p_1^2(y_1)} + o(1)\Big\}dy_1$$
$$= O(1/h_1)(g_3(x) + o(1)) = O(1/h_1).$$

Similarly, let $X_1 = x + h_2 z_1$, $X_2 = x + h_2 z_2$, $Y_1 = Y_3 + h_1 z_3$, $Y_2 = Y_3 + h_1 z_4$; then we have

$$A_{221} = E\Big[\frac{K_{h_2}(x - X_1)}{p_1(Y_1)}\frac{K_{h_2}(x - X_2)}{p_1(Y_2)} v(Y_3) K_{h_1}(Y_3 - Y_1) K_{h_1}(Y_3 - Y_2)\Big] = O(1).$$

Using Lemma 4, we can show similarly that

$$A_{222} = E\Big[\frac{K_{h_2}(x-X_1)}{p_1(Y_1)}\frac{K_{h_2}(x-X_2)}{p_1(Y_2)} K_{h_1}(Y_3-Y_1)\{f(Y_3) - f(Y_1)\}K_{h_1}(Y_3-Y_2)\{f(Y_3)-f(Y_2)\}\Big] = o(1).$$

Hence $A_{22} = A_{221} + A_{222} = O(1)$. Similarly, we can prove $A_{24} = O(1)$ and $A_{26} = O(1)$. Lastly,

$$\begin{aligned} A_{27} &= \Big\{E\Big[\frac{K_{h_2}(x - X_1)}{p_1(Y_1)} K_{h_1}(Y_1 - Y_3)\{f(Y_3) - f(Y_1)\}\Big]\Big\}^2 \\ &= \Big\{E_{X_1,Y_1}\Big[\frac{K_{h_2}(x - X_1)}{p_1(Y_1)}\int K(z)\{f(Y_1 - h_1 z) - f(Y_1)\}p_1(Y_1 - h_1 z)dz\Big]\Big\}^2 \\ &= \Big\{E_{X_1,Y_1}\Big[\frac{K_{h_2}(x-X_1)}{p_1(Y_1)} k_2 h_1^2\{f^{(1)}(Y_1)p_1^{(1)}(Y_1) + \frac{1}{2}f^{(2)}(Y_1)p_1(Y_1) + o(1)\}\Big]\Big\}^2 \\ &= k_2^2 h_1^4\{E_{X_1}[K_{h_2}(x - X_1)\{g_4(X_1) + o(1)\}]\}^2 = O(h_1^4). \end{aligned}$$

Hence, $A_2 = O(1/n^2 h_1) + O(1/n) + O(h_1^4)$. By ignoring smaller order terms, under the assumption that $nh_1 \to \infty$, we have

$$A = O\Big(\frac{1}{n^2 h_1 h_2}\Big) + o\Big(\frac{1}{nh_2}\Big) + O(h_1^4).$$

For B in (8), it is well known (e.g. Collomb (1977), Härdle (1989)) that

$$B = \frac{k_1}{nh_2}\frac{u(x)}{p_2(x)} + \frac{h_2^4 k_2^2}{p_2^2(x)} g_1^2(x) + o\Big(\frac{1}{nh_2}\Big) + o(h_2^4).$$

Similar computation yields $C = o(1/nh_2) + O(h_1^2 h_2^2)$. Putting everything together, we have

$$E[\hat{m}(x) - m(x)]^2 = \frac{1}{nh_2} D_1(x) + h_2^4 D_2(x) + O\Big(\frac{1}{n^2 h_1 h_2} + h_1^4 + h_1^2 h_2^2\Big) + o\Big(\frac{1}{nh_2} + h_2^4\Big), \tag{9}$$

where $D_1(x) = k_1 u(x)/p_2(x)$ and $D_2(x) = k_2^2 g_1^2(x)/p_2^2(x)$. To minimize the asymptotic mean squared error, the optimal bandwidths are

$$h_{2opt} = \{D_1(x)/4D_2(x)\}^{1/5} n^{-1/5} \quad \text{and} \quad h_1 = o(h_2),\ nh_1 \to \infty.$$

The corresponding minimum mean squared error is

$$\min_{h_1,h_2} E(\hat{m}(x) - m(x))^2 = D_1(x)^{1/5} D_2(x)^{4/5} (4^{1/5} + 4^{-4/5}) n^{-4/5} + o(n^{-4/5}).$$

The $O(1/n^2 h_1 h_2)$ term in (9) shows that the condition $nh_1 \to \infty$ is indeed necessary.

References

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83**, 596-610.

Collomb, G. (1977). Quelques propriétés de la méthode du noyau pour l'estimation nonparamétrique de la régression en un point fixé. *Comptes Rendus Acad. Sci. Paris* **285**, 289-292.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377-403.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, Inc.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Härdle, W and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465-1481.

Hall, P. (1984). Integrated square error properties of kernel estimators of regression functions. *Ann. Statist.* **12**, 241-260.

Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.* **13**, 1352-1377.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141-142.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065-1076.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348-1360.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser.A* **26**, 359-372.

Department of Statistics, Texas A & M University, College Station, TX 77843, U.S.A.

(Received June 1994; accepted July 1995)