Journal of Multivariate Analysis **70**, 177–201 (1999)
Article ID jmva.1999.1813, available online at http://www.idealibrary.com on IDEAL®

Nonparametric Regression with Singular Design

Zhan-Qian Lu*

The Hong Kong University of Science and Technology, Kowloon, Hong Kong

Received March 6, 1998

Theories of nonparametric regression are usually based on the assumption that the design density exists. However, in some applications such as those involving high-dimensional or chaotic time series data, the design measure may be singular and may be likely to have a fractal (nonintegral) dimension. In this paper, the popular Nadaraya–Watson estimator is studied under the general setup that the continuity of the design measure is governed by the local or pointwise dimension. It will be shown in the iid setup that the nonparametric regression estimator achieves a convergence rate which is dependent only on the pointwise dimension. The case of time series data is also studied. For the latter case, a new mixing condition is introduced, and an assumption of marginal or joint density is completely avoided. Three examples, a fractal regression and two applications for predicting chaotic time series, are used to illustrate the implications of the obtained results.

AMS 1991 subject classifications: 62G07; 62M10; 28A80; 58F13.

Key words and phrases: rate of convergence; high-dimensional data; pointwise dimension; fractal design; chaotic systems; nonlinear prediction; strong mixing.

1. INTRODUCTION

We consider the following setup: we have given random vectors $\{(X_i, Y_i), i=1, ..., n\}$, where $X_i \in \Re^p$ consists of explanatory or design variables and $Y_i \in \Re^1$ is the response variable. Different dependence conditions on the vector sequence $\{(X_i, Y_i)\}$ can be imposed. For the moment we assume that $\{(X_i, Y_i)\}$ has identical marginal distribution and $\{X_i\}$ has marginal probability measure ρ. Our interest is estimation of the regression function $m(\mathbf{x}) = \mathrm{E}(Y \mid X = \mathbf{x})$ for $\mathbf{x}$ in the domain of interest in $\Re^p$. Alternatively, we write

$$Y_i = m(X_i) + v^{1/2}(X_i)\,\varepsilon_i, \qquad i=1, 2, ..., \tag{1.1}$$

where $v(\mathbf{x}) = \mathrm{Var}(Y \mid X = \mathbf{x})$ is the variance function and $\{\varepsilon_i\}$ is a noise sequence with zero mean and unit variance.

Various nonparametric estimators of m have been proposed in the literature, among which the simplest one may be the *Nadaraya–Watson*

* Current address: MathSoft, Inc., Data Analysis Products Division, 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109-3044, USA. E-mail: jlu@statsci.com.

0047-259X/99 $30.00

(N-W) regression estimator (Nadaraya, 1964; Watson, 1964). This kernel regression estimator is defined by, for a given bandwidth h and a kernel function K at any given point $\mathbf{x} \in \Re^p$,

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - \mathbf{x}}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - \mathbf{x}}{h}\right)}. \tag{1.2}$$

Nonparametric regression estimators have been widely studied. Most studies make the assumption that ρ has a density f with respect to the Lebesgue measure; see, e.g., Stone (1980). This has strong implications about the structure of data clouds. For example, under this assumption and the assumption that f is continuous, at *any* location in the support of ρ the data points contained in a neighborhood of size r when r is small is proportional to nr^p apart from a constant. This imposes certain homogeneity in the distribution of design points. Consequently, the *curse of dimensionality* arises in the sense that there are not enough data points in any given local neighborhood, a severe problem in the analysis of multivariate data (Friedman and Stuetzle, 1981; Huber, 1985). On the other hand, the stylized fact of the "cluster" tendency in multivariate data clouds may suggest a different model—the singular probability model based on a *fractal* or *self-similar* measure (Mandelbrot, 1982). In the latter framework, the issue of the curse of dimensionality is alleviated, as will be demonstrated in this study. (The fractal model arises naturally in the chaotic time series context which will be treated later.) Another approach to dimension reduction is the detection of the nonlinear relationships among high-dimensional variables, a situation termed *nonlinear confounding* by Li (1997).

Interesting problems are being raised in the area of chaotic systems, as reflected, for example, in the special *Journal of the Royal Statistical Society, Series B*, issue on chaos edited by Tong and Smith (1992). As a simple illustration, consider the logistic system $x_i = 4x_{i-1}(1 - x_{i-1})$. The time series from this simple system is known to appear random and is hard to distinguish from a stochastic sequence. In general, the study of time series $\{x_t, t = 1, 2, ..., n\}$ from observing a multi-dimensional dynamical system is usually through the technique of *state space reconstruction*. By appealing to Takens' theorem, there exist suitable choices of *embedding dimension* q and *time delay* τ such that valuable information about the original system can be extracted from the reconstructed state vectors $\{X_i = (x_{i-\tau}, ..., x_{i-q\tau})', (q\tau + 1) \leqslant i \leqslant (n + \tau)\}$. In particular, it is of interest to learn the dynamics in the reconstructed space, which reduces to estimating the function m in the system $x_i = m(x_{i-\tau}, ..., x_{i-q\tau})$. In practice, this theory is of limited use

because there may be some measurement error associated with x_t, and q and τ are in general hard to decide. The noisy model (1.1) provides a more general framework with $Y_i = x_i$, $X_i = (x_{i-\tau}, ..., x_{i-q\tau})$. The error term in this context may represent the consequence of embedding or measurement errors. Precise dependence conditions in the time series context will be discussed in Section 4. For the moment we also note that model (1.1) is also closely related to nonlinear prediction. For example, if one is interested in an L-step-ahead forecasting procedure based on values of past p lags, one may apply a nonparametric regression method to the historical data $\{Y_i = x_{i+L},\ X_i = (x_i, ..., x_{i-p+1}),\ p \leqslant i \leqslant n-L\}$ to obtain a best pointwise (conditional mean) predictor in the sense of the mean squared error.

Farmer and Sidorowich (1988) gave a nice survey on various nonlinear prediction procedures which can be regarded as variants of nonparametric regression methods. However, there is a lack of a theoretical basis for this type of application, which stimulated our interest in a new statistical theory. For example, it is well known from chaotic dynamical system theory that the state vectors X_i from this kind of time series usually do not possess a density. Coming back to the logistic example, it is obvious that X_i lies in a one-dimensional manifold for any $q \geqslant 2$ and $\tau \geqslant 1$. More generally, the invariant measure corresponding to state vectors $\{X_i\}$ based on observations of a chaotic system typically exhibits some self-similar property and has a fractal (nonintegral) dimension (Ruelle, 1989). A similar problem also arises in noisy chaotic systems (cf. Subsection 4.1 or Smith, 1992).

In summary, concerning the above situations the nonparametric estimation theory based on a density assumption does not apply. It is thus desirable to have a general theory which does not depend on this condition and which hopefully includes the singular design measure as a special case. This goal is achieved in this paper by introducing a general setup for design measurement, which is based on the theory of fractal geometry (e.g., Falconer, 1990; Cutler, 1993). In this setup (cf. Subsection 2.2), the design measure ρ is assumed to have a *pointwise* dimension at the given point of interest. This covers the usual regular case as well as the cases mentioned above.

Intuitively, the convergence rate of nonparametric estimators should depend only on the actual dimension of the design measure, and not on the number of explanatory variables, since it is the former that controls the number of data points contained in a local neighborhood. Indeed, a conjecture of Farmer and Sidorowich (1988) says that the convergence rate of estimating a chaotic map depends only on the fractal dimension of its invariant measure. In Section 3 the N-W estimator will be extended to the general case by showing that, if the design measure ρ has a local or

pointwise dimension $d = d(\mathbf{x}) \leqslant p$, that is, roughly the number of data points contained in the neighborhood of size r around $\mathbf{x}$ scales according to nr^d, the N-W estimator attains the pointwise convergence rate $O(h^s) + O_p((nh^d)^{-1/2})$ which is faster than that when ρ has a density, as given e.g. in Stone (1980). An example of fractal regression is used to illustrate this result (cf. Subsection 5.1).

The theory is also extended to time series data in Section 4 under general conditions. Thus, Farmer and Sidorowich's enlightening hypothesis is partially confirmed in the N-W estimator. The result in the time series context has interesting implications for predicting chaotic time series. These are illustrated through two examples in Subsections 5.2 and 5.3. Further discussion on other estimators is given in Section 6.

2. A GENERAL SETUP

2.1. *Smoothness Conditions*

For a given distance norm $\|\cdot\|$, say the Euclidean norm in $\Re^p$, let $U = U_{\mathbf{x}}$ denote an open set containing $\mathbf{x}$. The class of continuous functions on U is denoted by $C^0(U)$. The following smoothness conditions are routinely imposed.

(A) There exists a number $0 < s \leqslant 1$ such that m is Lipschitz continuous with exponent s. That is, there exists a constant $\gamma_m > 0$, such that $|m(\mathbf{x}_1) - m(\mathbf{x}_2)| \leqslant \gamma_m \|\mathbf{x}_1 - \mathbf{x}_2\|^s$ for any $\mathbf{x}_1, \mathbf{x}_2 \in U$.

(B) The variance function v is continuous in U, i.e., $v \in C^0(U)$.

For simplicity, we also impose the following condition on the kernel function.

(C) K is a spherically symmetric density function with bounded support. In other words, there exists a univariate function k with finite support, $k(0) > 0$, $k(x) = 0$ for $x > 1$, such that $K(\mathbf{x}) = k(\|\mathbf{x}\|)$. It is also imposed that k satisfies the Lipschitz condition; there exists γ_k, $0 < \alpha \leqslant 1$, such that $|k(x) - k(y)| \leqslant \gamma_k |x - y|^\alpha$ for all $x, y \in [0, 1]$.

Remark 1. Assumption (A) can be relaxed to some extent. As one referee points out, our theory still holds if the restriction $s \leqslant 1$ is removed. We choose to use the given setup since it is more natural (see Remark 2). It is less stringent and is often assumed in the dynamical system literature.

Remark 2. If stronger smoothness on m is available such as m being differentiable, one may prefer to apply the local linear and other local polynomial-type estimators, which presumably further reduce the effect of bias.

However, one should be cautious in applying this type of higher-order polynomial estimator in small-to-moderate sample cases when the variance effect may dominate. See Fan and Gijbels (1998) for more details on local polynomial modeling.

Stone (1980) showed that the N-W estimator attains the optimal rate of convergence in some minimax sense under some additional technical assumptions, including that ρ is absolutely continuous (a.c.). To remove the density assumption, we will discuss in the next subsection a general setup on the design measure which is used throughout this paper.

2.2. *A New Setup for Design Measure*

A general model for design measure can be given which will include the usual case when ρ has a joint density, the nonlinear confounding case, and the fractal case when ρ has a fractal dimension. The main tool is based on the notion of a pointwise dimension from fractal geometry (e.g. Cutler, 1993).

The pointwise dimension defines the local *continuity* behavior of a measure. Given a point $\mathbf{x}$ in $\Re^p$, let $B_r(\mathbf{x}) = \{\mathbf{u}: \|\mathbf{u}-\mathbf{x}\| \leqslant r\}$ denote the ball of radius r centered at $\mathbf{x}$. The following assumption on the design measure is imposed.

(D) The small-ball probability has an exact power-law behavior. That is,

$$\rho(B_{\mathbf{x}}(r)) = cr^d, \qquad \text{as} \quad r \to 0, \tag{2.1}$$

where $c = c(\mathbf{x}) > 0$ is some constant. We will write simply $\rho(B_{\mathbf{x}}(r)) \approx cr^d$. The number $d = d(\mathbf{x})$ is the (local) pointwise dimension of ρ at $\mathbf{x}$. Obviously $d \leqslant p$.

The following remarks are in order.

Remark 3. If ρ has a continuous density f, (2.1) holds at any point in $\{\mathbf{x}: f(\mathbf{x}) > 0\}$ with $d = p$ and $c(\mathbf{x}) = f(\mathbf{x})\, v_p$ as $r \to 0$ (where $v_p = \pi^{p/2}/\Gamma((p+2)/2)$).

Remark 4. If ρ has a continuous density f on a d_t-dimensional manifold, (2.1) holds at any point in the support of f with $d = d_t$. This case contains the situation of nonlinear confounding discussed in Li (1997).

Remark 5. A general definition of pointwise dimension is given by

$$\limsup_{r \to 0} \frac{\rho(B_{\mathbf{x}}(r))}{r^d} < \infty, \qquad \liminf_{r \to 0} \frac{\rho(B_{\mathbf{x}}(r))}{r^d} > 0, \tag{2.2}$$

where $d = d(\mathbf{x})$ is the pointwise dimension. (An apparently more general definition is also given in Cutler, 1993, Sect. 3.2, Definition 3.2.1.) We will write this as $\rho(B_{\mathbf{x}}(r)) \sim r^d$.

In particular, this covers the *lacunar* phenomenon

$$\rho(B_{\mathbf{x}}(r) = r^d H(r), \qquad \text{for} \quad r \leqslant \delta, \tag{2.3}$$

where $H(r)$ is bounded but does not tend to any limit as $r \to 0$. See example 1.

Note that for a singular self-similar measure, (2.1) or (2.2) holds only for ρ-almost all $\mathbf{x}$. On the other hand, the support of ρ may be contained in a set of zero Lebesgue measure.

There are occasions when the fractal pointwise dimension $d = d(\mathbf{x})$ may depend on the position of $\mathbf{x}$ in the state space, in which case the fractal set is said to be *inhomogeneous* and the measure is said to have the so-called *multifractal* property (cf. Ruelle, 1989). However, for many chaotic systems it can be shown that the pointwise dimension is a constant with probability one. Naturally, work on dimension estimation has focused on estimating this common value based on limited data, as emphasized in Smith (1992) and Cutler (1993). Certainly, there are other definitions of dimension which may assign different values in some situations, and they can address other aspects of a measure. For theoretical discussions on local regression estimation, we think that the local pointwise dimension is the most natural one to consider and so will be the only focus in this paper.

2.3. *A Lemma*

Condition (D) has important implications for a general kernel function as given by the following lemma.

LEMMA 2.1. *Assumptions* (C) *and* (D) *imply that*

$$EK\left(\frac{(X-\mathbf{x})}{h}\right) = h^d c(\mathbf{x})\, d \left(\int_0^\infty k(y)\, y^{d-1}\, dy\right)(1 + o(1)), \qquad \text{as} \quad h \to 0, \tag{2.4}$$

where we assume $k(y)\, y^{d-1} \in L^1(0, \infty)$.

The proof is given in Subsection 1 of the Appendix. The lemma develops the scaling property of a kernel function based only on the specification of probabilities on spherically symmetric sets.

Remark 6. An analogous generalization of the Lemma 1-based Remark 2 is also available. Actually, by the assumption (C), there exist positive constants L_1, L_2, a satisfying

$$L_1 1_{\{x < a\}} \leqslant k(x) \leqslant L_2 1_{\{x \leqslant 1\}}.$$

It follows that

$$\limsup_{r\to 0} E\frac{k\left(\frac{\|X-\mathbf{x}\|}{h}\right)}{r^d} \leqslant \limsup_{r\to 0} \frac{L_2\rho(B_{\mathbf{x}}(r))}{r^d} < \infty,$$

and

$$\liminf_{r\to 0} \frac{Ek\left(\frac{\|X-\mathbf{x}\|}{h}\right)}{r^d} \geqslant \liminf_{ar\to 0} \frac{a^d L_1 \rho(B_{\mathbf{x}}(ar))}{(ar)^d} > 0.$$

Thus, $EK((X_{\mathbf{x}})/h) \sim r^d$.

The following corollary can be established based on Lemma 1.

COROLLARY 1. *Under the setup of Lemma* 1, *the generalization*

$$\mathrm{Var}\left\{g(X_1)\,K\left(\frac{X_1-\mathbf{x}}{h}\right)\right\} = dcg^2(\mathbf{x})\,h^d\left(\int_0^\infty k^2(y)\,y^{d-1}\,dy\right)(1+o(1)), \tag{2.5}$$

as $h\to 0$, *holds for any* $g\in C^0(U)$, *where we assume* $k^2(y)\,y^{d-1}\in L^1(0,\infty)$.

From now on, we will modify the condition (C) to include the moment condition $k(y)\,y^{d-1}\in L^1(0,\infty)$, $k^2(y)y^{d-1}\in L^1(0,\infty)$.

3. INDEPENDENT OBSERVATIONS

In this section, the case of independent observations is considered. That is, we assume the following:

(E) (X_i, Y_i) are iid random vectors. Furthermore, there exists some constant $\delta>0$ such that $E|\varepsilon_1|^{2+\delta}<\infty$.

The following theorem on the N-W estimator is established. We use $\xrightarrow{\text{dist}}$ to denote convergence in distribution.

THEOREM 3.1. *Under assumptions* (A)–(E), *it follows that there exists a sequence of constants* $b_n = O(h^s)$ *such that, as* $h\to 0$, $nh^d\to\infty$,

$$\sqrt{nh^d}\,\{\hat{m}(\mathbf{x}) - m(\mathbf{x}) - b_n\} \xrightarrow{\text{dist}} N(0,\delta^2),$$

where $\delta^2 = v(\mathbf{x}) \int_0^\infty k^2(y)\, y^{d-1}\, dy / \{dc(\int_0^\infty k(y)\, y^{d-1}\, dy)^2\}$. *By choosing* $h = O(n^{-1/(d+2s)})$, *the achieved pointwise convergence rate is* $O_p(n^{-s/(d+2s)})$.

The proof is given in Subsection 2 of the Appendix. Theorem 1 complements Stone (1980) by proving a convergence rate for the N-W regression estimator which does not depend on the assumption of design density. The conjecture of Farmer and Sidorowich given in the Introduction is partially answered under the regression setup. An example is given in Subsection 5.1 to illustrate the implications of this theorem.

4. TIME SERIES DATA

Section 3 has discussed the regression problem with fractal design. This section will consider the same general setup except for time series data for which dependence in observations needs to be taken into account. General conditions including a new mixing condition are introduced in Subsection 4.2. This setup may also include the situation of when the time series data actually come from a noisy chaotic system as discussed in the next section.

4.1. *Fractality in Systems with Little Noise*

Simple systems with a *few* degrees of freedom rarely exist in isolation, so there is increasing interest in investigating the behavior of chaotic systems with a little noise; see e.g. Kapitaniak (1990). Since time series data are often subject to dynamical noise, nonlinear systems which are stable under perturbations of dynamic noises are of particular interest; cf. Tong and Smith (1992) and Chan and Tong (1994).

The model which has often been entertained for time series data,

$$x_i = m(x_{i-1}, ..., x_{i-p}) + v^{1/2}(x_{i-1}, ..., x_{i-p})\, \varepsilon_i, \tag{4.1}$$

has been considered by a number of authors, including Chan and Tong (1994) and An and Huang (1996). It is a special case of (1.1) with $Y_i = x_i$, $X_i = (x_{i-1}, ..., x_{i-p})'$.

In order to discuss the theory of noisy chaos, let us assume $v = \sigma^2$ for the moment. (This is not necessary for our latter theory.) In the asymptotic approximations which follow there is an implicit assumption that r or h is of about the same order or of larger magnitude than the noise amplitude σ (and both are small). Heuristically, we would expect that for $r \gg \sigma$ the invariant measure would behave as if noise were not present. For $r \ll \sigma$, the noise component is dominant and one would expect the invariant measure to scale according to the embedding dimension q. In the intermediate range, depending on the underlying dynamics as well as noise,

there may exist a critical value r_0 such that for $r \leqslant r_0$ the invariant measure behaves according to d while for $r > r_0$ the invariant measure scales with the embedding dimension q. Smith (1992) gave further discussions including analyses of a number of simulated and real data sets. Note that the embedding dimension q may be different from a true model dimension p, which may be a consequence of some estimation on selection process.

In kernel regression problems, it is necessary that the bandwidth must be chosen large enough to contain enough data points locally, and for most common multivariate data sets it is *fairly large*. This suggests that the chosen bandwidth is often in the range where the fractal behavior of the invariant measure takes effect. Thus the fractal setup for design measurement in Subsection 2.2 may be pertinent for most practical purposes. This seems to be the case for Examples 2 and 3 considered in Section 5.

4.2. *Dependence Condition*

We need to impose some conditions on the dependence structure. The following assumption on $\{\varepsilon_i\}$ is used for simplicity.

(F) There exists a nondecreasing sequence of σ-fields $\{\mathscr{F}_i\}$ such that $X_i \in \mathscr{F}_{i-1}$, $\varepsilon_i \in \mathscr{F}_i$ for all $i \geqslant 1$, and the sequence $\{\varepsilon_i, \mathscr{F}_i\}$ is a martingale difference satisfying

$$E\{\varepsilon_i \mid \mathscr{F}_{i-1}\} = 0, \qquad E\{\varepsilon_i^2 \mid \mathscr{F}_{i-1}\} = 1. \tag{4.2}$$

Further, there exists some $\delta > 0$ such that

$$\sup_{i \geqslant 1} E\{|\varepsilon_i|^{2+\delta} \mid \mathscr{F}_{i-1}\} < \infty.$$

Assumption (4.2) is natural in the autoregressive model (4.1), since it is equivalent to the assumption that $E\{x_i \mid x_{i-1}, x_{i-2}, \ldots\}$ and $E\{x_i^2 \mid x_{i-1}, x_{i-2}, \ldots\}$ are functions of variables $x_{i-1}, \ldots, x_{i-p}$ only for some integer p or that the underlying model is of finite order.

Since the design density is not assumed, and thus the joint densities between any two vectors involving predictors or responses at different times do not exist, the assumptions that are usually imposed on joint densities in the literature cannot be used. Consequently, we introduce a more general mixing condition next.

Let P denote the joint probability measure of the sequence $\{X_i\}$ so that marginally $P(X_1 \in A) = \rho(A)$. Also let μ denote the d_{H}-dimensional Hausdorff measure where $d_{\mathrm{H}} \geqslant d(\mathbf{x})$. For example, d_{H} can be the Hausdorff dimension for the support of ρ. (See Falconer, 1990, Chap. 2, for a definition.) If there exists a constant d_{p} such that $\rho(d(\mathbf{x}) = d_{\mathrm{p}}) = 1$ (d_{p} is the

(global) pointwise dimension), naturally we require $d_H \geqslant d_p$. More generally, for estimation at more than one location we may require $d_H \geqslant \sup_{\mathbf{x}} d(\mathbf{x})$, where $\mathbf{x}$ goes over the point of interest in the support of ρ. We define a generalized β_n-mixing condition in the following way.

(G) The sequence $\{X_i\}$ is strictly stationary and there exists a sequence of constants β_n such that

$$\sum_{j=1}^{n} |P(X_{j+1} \in A, X_1 \in B) - P(A)\, P(B)| \leqslant \beta_n \mu(A)\, \mu(B), \tag{4.3}$$

for all Borel sets A, B in $\mathfrak{R}^p$, and $\beta_n \leqslant M$ for some constant M.

If the joint density of X_1, X_{j+1} exists, denoted by $f_j(\cdot, \cdot)$, and the marginal density is denoted by $f(\cdot)$, one can set

$$\beta_n = \sup_{\mathbf{u},\, \mathbf{v} \in \mathfrak{R}^p} \sum_{j=1}^{n} |f_j(\mathbf{u}, \mathbf{v}) - f(\mathbf{u})\, f(\mathbf{v})|.$$

Assumption (G) is similar to the requirement that the α-mixing rate (e.g., Bradley, 1986) be summable, except that β_n-mixing imposes an additional constraint on the sets of small probability. In the context of estimating a deterministic system (i.e., $v = 0$), Bosq and Guégan (1995) used a condition similar to (G) but with μ taken to be the Lebesgue measure. Their condition may be too restrictive considering that P may have support on a set of Lebesgue measure zero. When $\{\varepsilon_i\}$ is an independent sequence, the process $\{X_t\}$ is a Markov chain. An and Huang (1996) gave a number of conditions which ensure the geometric ergodicity of a Markov chain, which in turn implies α-mixing with a geometric rate.

4.3. *Theorem*

Asymptotic normality of the N-W estimator in the general setup is given in the next theorem. The proof is given in Subsection 3 of the Appendix.

THEOREM 2. *Under assumptions* (A)–(D), (F), and (G), *there exists a sequence of constants* $b_n = O(h^s)$ *such that, as* $h \to 0$, $nh^d \to \infty$,

$$\sqrt{nh^d}\, \{\hat{m}(\mathbf{x}) - m(\mathbf{x}) - b_n\} \xrightarrow{\text{dist}} N(0, \delta^2),$$

where $\delta^2 = v(\mathbf{x}) \int_0^\infty k^2(y)\, y^{p-1}\, dy / \{pc(\mathbf{x})(\int_0^\infty k(y)\, y^{p-1}\, dy)^2\}$. *Further, the estimators at different points are jointly normal and asymptotically independent.*

In the more realistic noisy situations, Theorem 2 answers the conjecture of Farmer and Sidorowich affirmatively for the N-W-type estimator which includes, in particular, the *nearest-neighbor* estimator with uniform weight.

This result implies that the accuracy of estimating a strongly nonlinear map using a nonparametric regression procedure does not necessarily depend on the number of predictor variables, and if the actual dimension of the measure of predictor variables is *fixed* and not large, the precision of a nonlinear nonparametric prediction procedure can be very good no matter how many predictor variables are included in the model. The theoretical revelation is backed by numerical studies given in Section 5 (Examples 2 and 3).

Our results raise some interesting questions about bandwidth selection in the singular design case. Intuitively, a smaller bandwidth should be used in the singular design case, and in particular when the pointwise dimension is small. Thus, it is useful to apply a dimension estimation method such as that of Smith (1992) to get an idea of the fractalness of the design space. A more formal data-based selection method for h may also be worth investigating, as mentioned by a referee. For the latter, one approach may be based on some assumption of smoothness of regression and plug-in or substitution of estimated quantities so as to achieve a tradeoff between asymptotic bias and variance. See, for example, Chapter 4 of Fan and Gijbels (1995). A particular issue is the estimation of the variance function v. Fan and Yao (1998) presented an estimation method which is efficient even when m is estimated based on the same data set and they applied it to some bandwidth selection methods. Their approach is likely applicable to our general context as well. Another often-used approach of data-based bandwidth selection is the cross-validation method, which does not require any prior model assumption, and in particular does not require knowledge of the pointwise dimension of the design measure.

This paper does not go into any details about bandwidth selection issues, a likely topic for future investigation. For the examples which follow, we simply try a range of different bandwidths and choose the one that works best. For example, in Example 1 we choose h which appears to give a good fit or to minimize the prediction errors at the selected design points.

5. EXAMPLES

The following three examples illustrate the performance of the N-W kernel predictor in singular design models. The first one considers the case of a fractal regression, in which the regressors are self-similar random variables. The next two examples deal with predicting noisy chaotic time series, where we adopt the ℓ-NN method instead of a global bandwidth. Locally, the performance of the ℓ-NN method is exactly the same as the fixed bandwidth method discussed earlier. Globally, the ℓ-NN method allows different amounts of smoothing in different locations. This may be

desirable when the design space is fairly inhomogeneous. For simplicity, we use the uniform kernel in all examples.

5.1. EXAMPLE 1 (Fractal Regression). The Cantor-type distribution is defined by the random variable

$$\eta(\alpha) = (\alpha - 1) \sum_{i=1}^{\infty} \alpha^{-i} \xi_i,$$

where $\alpha > 1$ and ξ_i's are Bernoulli trials with probability of success 0.5. The uniform measure corresponds to $\alpha = 2$. The famous middle-third Cantor measure corresponds to $\alpha = 3$. It has pointwise dimension $d = \log 2/\log 3$.

Feller (1971, Example I.11(d)) discussed the case $\alpha = 4$. To calculate the local dimension, we note that, at any point x on the Cantor-type set,

$$\rho(|\eta(4) - x| \leqslant 4^{-n}) = 2^{-n}$$

and so for any $0 < r < 1$,

$$\rho(B_x(r)) = 2^{-n}, \text{ if } 4^{-n-1} < r \leqslant 4^{-n},$$

and thus, by defining $n = [-\log r/\log 4]$, the integer part, and $H(r) = (4^n r)^{-1/2}$, (2.3) is satisfied. Defining $G(\log(r)) = H(r)$, it follows that G is a periodic function with period $\log 4$. The pointwise dimension is obviously $\log 2/\log 4 = 0.5$. In general, for any $\alpha > 2$, the corresponding measure has pointwise dimension $d = \log 2/\log \alpha$.

We define four independent random vectors of length 500: C_1, C_2, C_3, and C_4, each consisting of iid random samples from $\eta(3)$, $\eta(4)$, $\eta(5)$, and $\eta(7)$, respectively.

Define four regression models by

$$Y_i = \cos(2\pi C_i) + \sin(2\pi C_{i+1}) + 0.1N(\mathbf{0}, I),$$

for $i = 1, 2, 3, 4$, corresponding respectively to Cases 1–4. The design space (C_i, C_{i+1}) for each case is plotted in Fig. 1. Due to self-similarity in the distribution of design points, details are hidden in most plots. To better appreciate the difference in details of the four designs, we also present the enlarged details inside the small squares, and these are shown in Fig. 2.

We compute fits using the kernel regression method for each data set at selected data points from 100, 101, ..., 500. The respective bandwidth in each case is:

Case 1. $h = 0.03$

Case 2. $h = 0.01$

Case 3. $h = 0.005$

Case 4. $h = 0.001$.

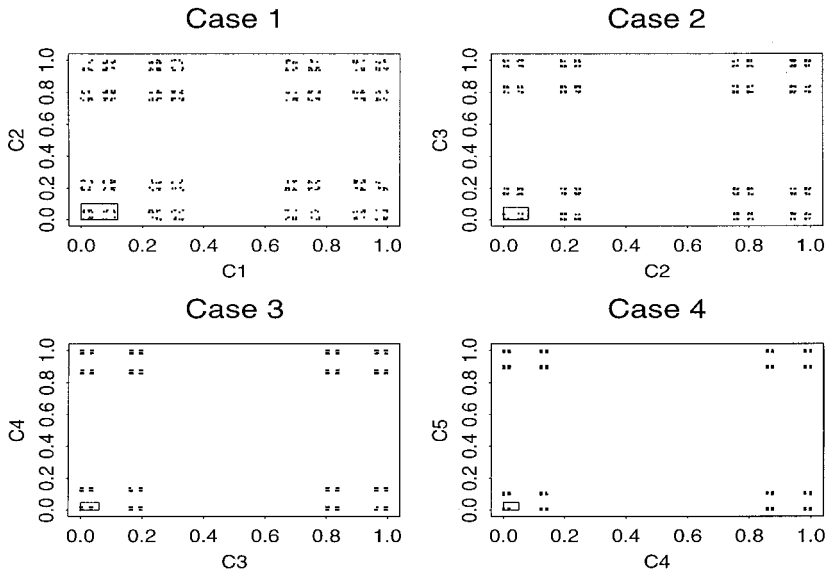

FIG. 1. Plots of design space for each case in Example 1.

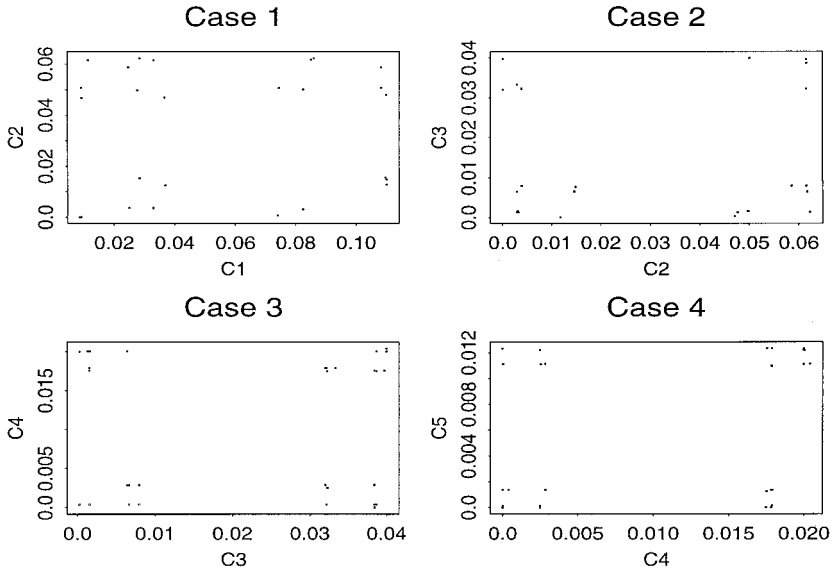

Fig. 2. Enlarged details inside the squares in plots of Fig. 1.

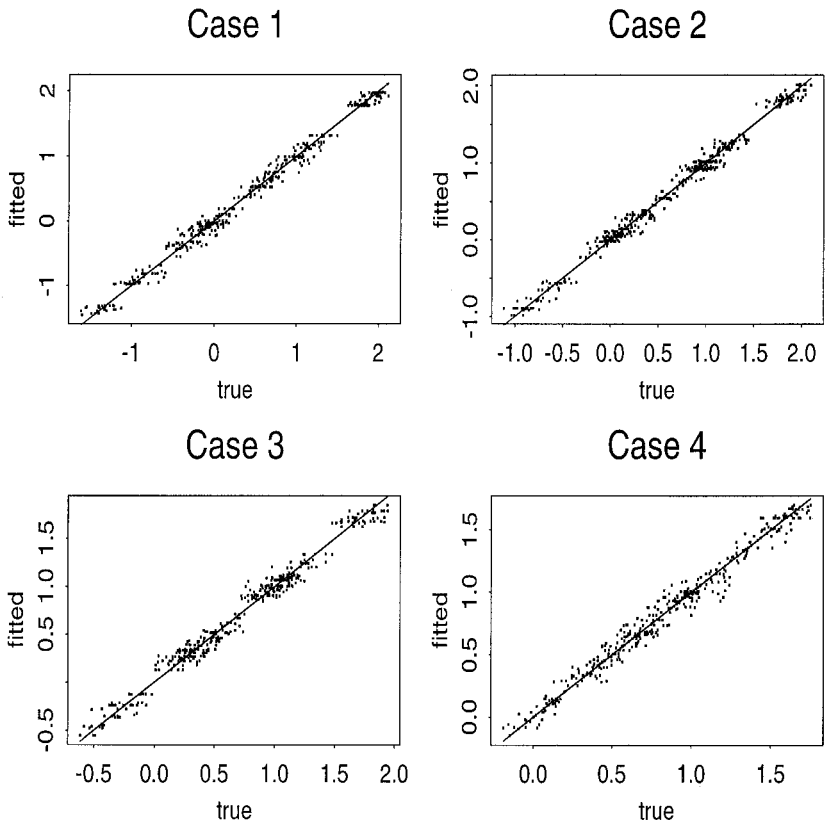

FIG. 3. Scatter plots of fitted values versus true values for each case in Example 1. The solid lines denote $\hat{m} = y$.

The scatter plots of fitted values versus true values are given in Fig. 3. The bandwidth reflects some tuning to achieve a visually good fit (fitted values and true values fall around the straight diagonal line, solid lines in the figures) in each case.

The bandwidth choice is consistent with the implications of our theory. Since the design points are more scattered in the first two cases than in the last two cases, the bandwidth is smaller for the last two. In other words, as the fractal dimension of the design space for each case, denoted using the formula

$$d(C_i, C_{i+1}) = d(C_i) + d(C_{i+1}),$$

gets smaller as i increases, a smaller and smaller bandwidth can be employed (hence there is a smaller bias effect). This example partly conveys

the message that *fractal design reduces the effect of dimensionality curse* and may be "good" for multivariate regression.

5.2. EXAMPLE 2 (Simulations from a Noisy Hénon System). Consider the noisy Hénon system $x_i = 1 - ax_{i-1}^2 + bx_{i-2} + \sigma e_i$ with parameters $a = 1.4$, $b = 0.3$, and $\sigma = 0.03$. The dynamical noises e_i are assumed to be iid with uniform distribution on $[-0, 5, 0.5]$. A time series of length 3000 is generated (after discarding the first 500 transient steps). The reconstructed state vectors $\{X_i = (x_i, ..., x_{i-p+1})'\}$ from this time series are known to have a fixed dimension around 1.25 for any p larger than 2 (e.g., Smith, 1992).

A prediction experiment is carried out as follows: the first 2000 values are used as training data and $\ell = 15$ is chosen. For various embedding dimensions $p = 1, 2, ..., 8$, one-step out-sample predictions are computed at 17 selected time steps between steps 2000 and 2100. Table I gives the results of the prediction for the selected time steps. Note that the standard deviation of the time series is about 0.73, so except for the two points 4 and

TABLE I

Prediction Results for the Hénon Series

Time index	Time step	True value	Predicted value							
			$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$
1	2003	0.16	0.02	0.18	0.20	0.15	0.09	0.09	0.16	0.16
2	2004	0.70	1.14	0.69	0.69	0.68	0.68	0.65	0.58	0.59
3	2016	−1.28	−1.28	−1.23	−1.27	−1.26	−1.25	−1.22	−1.22	−1.22
4	2017	−0.90	−0.92	−0.92	−0.90	−0.90	−0.89	−0.86	−0.76	−0.70
5	2018	−0.50	0.07	−0.52	−0.52	−0.56	−0.56	−0.52	−0.51	−0.28
6	2019	0.37	0.71	0.33	0.33	0.33	0.33	0.33	0.33	0.33
7	2044	1.15	0.88	1.14	1.14	1.13	1.10	1.10	1.10	1.06
8	2048	0.32	0.33	0.31	0.31	0.31	0.32	0.32	0.33	0.33
9	2059	1.08	0.98	1.07	1.08	1.08	1.05	1.05	1.01	1.01
10	2063	1.12	0.77	1.13	1.13	1.13	1.13	1.13	1.07	1.00
11	2064	−0.86	−0.71	−0.88	−0.91	−0.89	−0.89	−0.89	−0.88	−0.87
12	2076	0.58	0.74	0.55	0.55	0.55	0.57	0.56	0.55	0.55
13	2077	0.69	0.57	0.72	0.71	0.71	0.71	0.67	0.69	0.64
14	2078	0.51	0.33	0.53	0.44	0.44	0.48	0.48	0.50	0.45
15	2086	1.10	1.07	1.09	1.08	1.08	1.08	1.08	1.08	1.01
16	2087	−0.62	−0.67	−0.59	−0.60	−0.57	−0.57	−0.57	−0.55	−0.54
17	2089	−0.09	0.13	−0.09	−0.11	−0.10	−0.12	−0.11	−0.11	−0.11

Note: The first column is the time index, the second column is the actual time steps at which predictions are made, the third column is the true values. The fourth through the eleventh columns are the one-step predicted values corresponding to the embedding dimensions $p = 1, 2, ..., 8$, respectively.

5 which are at the boundary of the phase space the predicted values are very close to the true values as p increases from 2 to 8. This example confirms Theorem 2 in that the accuracy of kernel prediction is independent of the embedding dimension p when p is greater than the actual pointwise dimension.

5.3. EXAMPLE 3 (Laboratory Data). In this example, high-precision measurements of a temperature time series from a fluid mechanics experiment are used (cf. Read *et al.*, 1992, and Smith, 1992). The first 2000 observations are used for model fitting, while one-step out-sample predictions

TABLE II

Prediction Results for the Temperature Series

Starting index	Starting time step	Prediction error							
		$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$
1	2000	0.149	0.136	0.151	0.167	0.190	0.200	0.210	0.218
3	2040	0.059	0.055	0.062	0.071	0.074	0.075	0.075	0.077
5	2080	0.136	0.128	0.130	0.130	0.144	0.155	0.167	0.162
7	2120	0.039	0.038	0.041	0.040	0.047	0.050	0.051	0.049
9	2160	0.096	0.084	0.081	0.095	0.103	0.099	0.109	0.118
11	2200	0.043	0.039	0.034	0.036	0.038	0.039	0.038	0.044
13	2240	0.251	0.261	0.279	0.286	0.286	0.285	0.294	0.305
15	2280	0.067	0.054	0.067	0.068	0.071	0.076	0.077	0.082
17	2320	0.118	0.116	0.099	0.105	0.120	0.130	0.135	0.133
19	2360	0.044	0.033	0.032	0.031	0.032	0.033	0.038	0.042
21	2400	0.174	0.193	0.208	0.217	0.227	0.239	0.234	0.239
23	2440	0.064	0.056	0.061	0.070	0.075	0.077	0.077	0.087
25	2480	0.096	0.104	0.105	0.110	0.119	0.133	0.133	0.136
27	2520	0.078	0.073	0.072	0.073	0.072	0.075	0.077	0.072
29	2560	0.047	0.051	0.045	0.053	0.057	0.057	0.057	0.064
31	2600	0.050	0.040	0.034	0.038	0.043	0.046	0.051	0.054
33	2640	0.139	0.137	0.129	0.122	0.121	0.142	0.143	0.142
35	2680	0.081	0.075	0.073	0.070	0.080	0.087	0.089	0.093
37	2720	0.104	0.090	0.096	0.111	0.106	0.103	0.096	0.102
39	2760	0.054	0.050	0.055	0.067	0.079	0.081	0.088	0.084
41	2800	0.139	0.121	0.126	0.109	0.120	0.143	0.138	0.123
43	2840	0.059	0.058	0.059	0.062	0.071	0.072	0.075	0.070
45	2880	0.107	0.085	0.105	0.104	0.115	0.119	0.126	0.127
47	2920	0.088	0.073	0.082	0.085	0.089	0.093	0.104	0.110
49	2960	0.113	0.094	0.091	0.075	0.103	0.114	0.112	0.098
51	3000	0.088	0.081	0.088	0.097	0.100	0.101	0.093	0.092

Note: The first column is the starting time index, and the second column is the actual starting time step from which one-step predictions are computed at the next 20 time steps. The third through the ninth columns are the root mean squared prediction errors of 20 predicted values corresponding to embedding dimensions $p = 1, 2, ..., 8$, respectively.

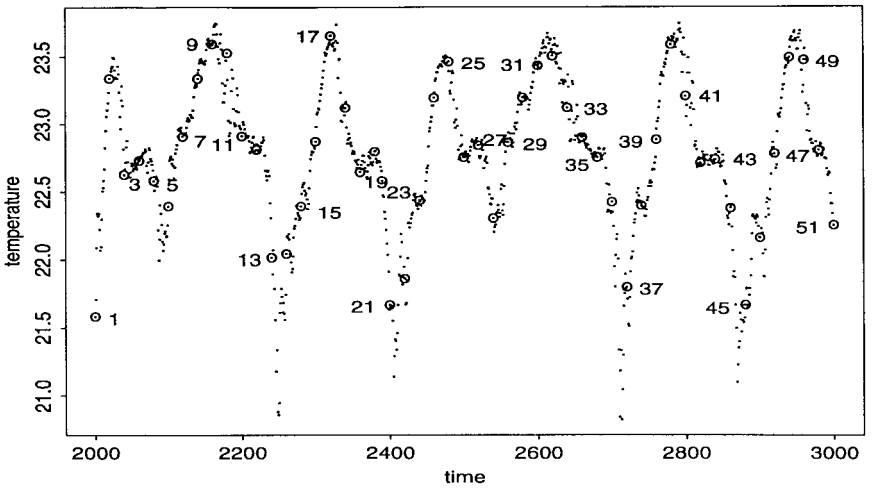

FIG. 4. Time series plot of the temperature series at steps 2000–3000. The circled points denote the starting points at which predictions are computed.

for the time steps from 2000 through 3000 are calculated. We use the ℓ-nearest neighbor method with $\ell = 20$. To assess the prediction error, the square roots of the means of 20 squared prediction errors (root mse) starting at every 20th step are computed. Table II shows the prediction errors at 26 selected time steps. Note that the standard deviation of this series is about 0.52. The prediction error is minimized for $p = 2$ or $p = 3$ most of the time

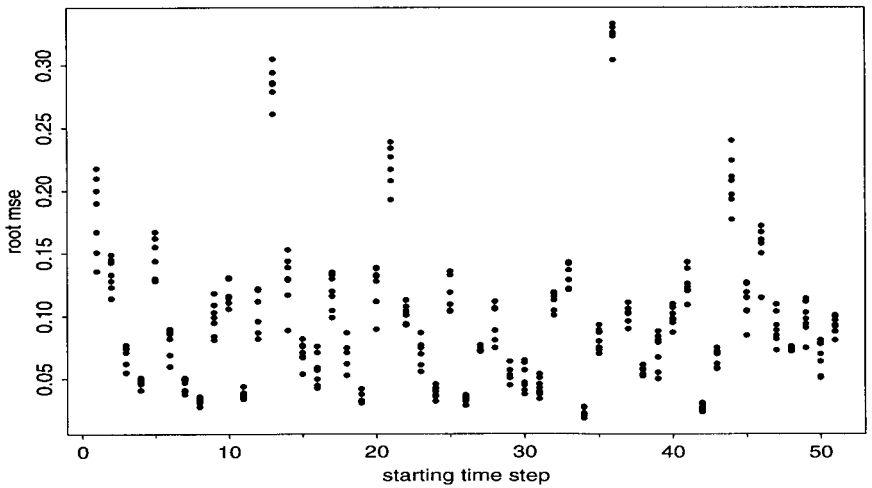

FIG. 5. Root mean squared prediction errors for $p = 2, 3, ..., 8$ at selected time points. The positions of these points are shown in Fig. 4.

and it does not deteriorate much as p is increased from 2 to 8. This is consistent with previous studies' determination that this data set may have a low fractal dimension (cf. Smith, 1992, or Read *et al.*, 1992) and is anticipated by Theorem 2. Interestingly, the prediction errors appear to depend upon which part of the time series is predicted: for example, the smallest prediction errors occur during the rise to the secondary peak within each cycle (e.g., 4, 11, 19, 26, 34, 42) while the largest errors occur during the fall to the longer tail (e.g., 13, 21, 36, 44). See Figs. 4 and 5. This may indicate that different predictability regimes exist in this time series. This is in contrast to the Hénon example for which the prediction errors appear to be relatively homogeneous.

6. CONCLUDING REMARKS

This paper develops a general theory for the popular Nadaraya–Watson estimator which is applicable to the usual case as well as to the nonregular situation in which the design measure has a fractal dimension. Our results demonstrate the potential for applying the nonparametric regression method to a high-dimensional model as long as the actual dimensionality of the design measure is not large.

In principle, the results of this paper can be expected to be valid for other nonparametric estimators such as higher-order polynomial fitting as well. However, in order to obtain similar results for such estimators, an extension of Lemma 1 to more general functions, and in particular to moments of a spherically symmetric smooth function, is needed. Our preliminary work indicates that additional conditions on the design measurement beyond the specification of the probability of the locally spherically symmetric sets may be necessary. This is an interesting open problem involving the calculus of fractal measures (cf. Hutchinson, 1981).

In conclusion, we think that the issue of the curse of dimensionality may be alleviated to some extent when the joint density assumption is not attainable and the actual dimension of the probability measure is substantially smaller than the number of variables. Our study reveals that singular or fractal design is "good" for multivariate data. We also point out that there is substantial scope for further investigation in high-dimensional and fractal modelings.

APPENDIX: PROOFS

A.1. *Proof of Lemma* 1. Given any partition on $[0, 1]$,

$$0 = a_0 < a_1 < a_2 < \cdots < a_{n-1} < a_n = 1,$$

and let

$$\Delta = \max_i (a_{i+1} - a_i).$$

Write

$$k(y) = \sum_{i=0}^{n-1} k(a_i)\, 1_{\{a_i < y < a_{i+1}\}} + \sum_{i=0}^{n-1} (k(y) - k(a_i))\, 1_{\{a_i < y < a_{i+1}\}} \triangleq I(y) + II(y),$$

and by Assumption (C), the LHS of (2.4) is equal to

$$Ek\left(\frac{\|X - \mathbf{x}\|}{h}\right) = EI\left(\frac{\|X - \mathbf{x}\|}{h}\right) + EII\left(\frac{\|X - \mathbf{x}\|}{h}\right). \tag{A.1}$$

The first term is equal to

$$\sum_{i=0}^{n-1} k(a_i)\{\rho(B_{\mathbf{x}}(ha_{i+1})) - \rho(B_{\mathbf{x}}(ha_i))\}$$

$$= ch^d \left\{\sum_{i=0}^{n-1} k(a_i)(a_{i+1}^d - a_i^d)\right\}$$

$$= ch^d\, d \left\{\int_0^\infty k(y)\, y^{d-1}\, dy + o_\Delta(1)\right\},$$

where $o_\Delta(1) \to 0$ as $\Delta \to 0$. Assumption (D) is used in the second equation.

On the other hand, the second term in (A.1) is bounded by

$$\gamma_k\, \Delta^\alpha \sum_{i=0}^{n-1} \rho\{ha_i < \|X - \mathbf{x}\| < ha_{i+1}\} = \gamma_k\, \Delta^\alpha \rho\{\|X - \mathbf{x}\| < h\}$$

$$= \gamma_k\, \Delta^\alpha c(\mathbf{x})\, h^d.$$

Thus, (A.1) becomes

$$ch^d \left\{d \int_0^\infty k(y)\, y^{d-1}\, dy + o_\Delta(1)\right\}(1 + o(1)) + O(\Delta^\alpha h^\alpha)(1 + o(1))$$

$$\to ch^d \left\{d \int_0^\infty k(y)\, y^{d-1}\, dy\right\}$$

by taking $\Delta \to 0$. The proof of the lemma is complete. ∎

A.2. *Proof of Theorem* 1. Write

$$\hat{m}(\mathbf{x}) - m(\mathbf{x}) = \frac{\sum_{i=1}^{n} (m(X_i) - m(\mathbf{x}))\, K\left(\frac{X_i - \mathbf{x}}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - \mathbf{x}}{h}\right)} + \frac{\sum_{i=1}^{n} v^{1/2}(X_i)\, K\left(\frac{X_i - \mathbf{x}}{h}\right) \varepsilon_i}{\sum_{i=1}^{n} K\left(\frac{X_i - \mathbf{x}}{h}\right)}$$

$$\triangleq B_n + R_n. \tag{A.2}$$

First, consider R_n which we further write as $(nh^d)^{-1}\, T_n/S_n$ where

$$S_n = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - \mathbf{x}}{h}\right) \tag{A.3}$$

$$T_n = \sum_{i=1}^{n} v^{1/2}(X_i)\, K\left(\frac{X_i - \mathbf{x}}{h}\right) \varepsilon_i. \tag{A.4}$$

It follows that $ET_n = 0$,

$$\mathrm{Var}\{T_n\} = nh^d v(\mathbf{x})\, dc(\mathbf{x}) \left(\int_0^\infty k^2(y)\, y^{d-1} dy \right) (1 + o(1)).$$

Defining

$$\Lambda_n \triangleq \sum_{i=1}^{n} E \left| v^{1/2}(\mathbf{x})\, K\left(\frac{X_i - \mathbf{x}}{h}\right) \varepsilon_i \right|^{2+\delta}$$

$$= nh^d v(\mathbf{x})^{(2+\delta)/2}\, dc(\mathbf{x})\, E\,|\varepsilon_1|^{2+\delta} \left(\int_0^\infty k(y)^{2+\delta}\, y^{d-1}\, dy \right) (1 + o(1)),$$

we have

$$\frac{\Lambda_n}{(\mathrm{Var}\, T_n)^{(2+\delta)/2}} = O((nh^d)^{-\delta/2}) \to 0 \tag{A.5}$$

as $nh^d \to \infty$. This verifies the conditions of the central limit theorem for sums of triangular arrays (Serfling, 1981, p. 32, Corollary 1.9.3).

Thus, we prove

$$\frac{1}{\sqrt{nh^d}}\, T_n \xrightarrow{d} N(0, \delta_1^2), \qquad \text{where} \quad \delta_1^2 = v(\mathbf{x})\, dc(\mathbf{x}) \int_0^\infty k^2(y)\, y^{d-1}\, dy. \tag{A.6}$$

Since

$$S_n = h^{-d} EK\left(\frac{X_1 - \mathbf{x}}{h}\right) + O_p((nh^d)^{-1/2}),$$

we have

$$\sqrt{nh^d}\ R_n = \frac{(\sqrt{nh^d})^{-1}\ T_n}{S_n} \xrightarrow{\text{dist}} N(0, \delta^2),$$

where $\delta^2 = v(\mathbf{x}) \int k^2(y)\, y^{d-1}\, dy / \{dc(\mathbf{x})(\int k(y)\, y^{d-1}\, dy)^2\}$. That is, we obtain

$$\sqrt{nh^d}\ \{\hat{m}(\mathbf{x}) - m(\mathbf{x}) - B_n\} \xrightarrow{\text{dist}} N(0, \delta^2). \tag{A.7}$$

Next we show how to replace B_n with a constant. Note that

$$\begin{aligned}
&\left| E\{m(X_1) - m(\mathbf{x})\}\ K\left(\frac{X_1 - \mathbf{x}}{h}\right)\right| \\
&\quad \leqslant E\,|m(X_1) - m(\mathbf{x})|\ K\left(\frac{X_1 - \mathbf{x}}{h}\right) \\
&\quad \leqslant \gamma_m E\,\|X_1 - \mathbf{x}\|^s\ K\left(\frac{X_1 - \mathbf{x}}{h}\right), \\
&\quad = \gamma_m\, dc(\mathbf{x})\, h^{d+s} \left(\int_0^\infty y^s k(y)\, y^{d-1}\, dy\right)(1 + o(1)).
\end{aligned}$$

Similarly,

$$\text{Var}\left\{(m(X_1) - m(\mathbf{x}))\ K\left(\frac{X_1 - \mathbf{x}}{h}\right)\right\} = O(h^{d+2s}).$$

We obtain that

$$\begin{aligned}
&\sum_{i=1}^{n} (m(X_i) - m(\mathbf{x}))\ K\left(\frac{X_i - \mathbf{x}}{h}\right) \\
&\quad = nE\{m(X_1) - m(\mathbf{x})\}\ K\left(\frac{X_1 - \mathbf{x}}{h}\right) + O_p(nh^{2s+d})
\end{aligned}$$

is of order $nh^d(O(h^s) + O_p(h^s(nh^d)^{-1/2}))$.

Defining

$$b_n = \frac{E\{m(X_1) - m(\mathbf{x})\}\ K\left(\frac{X_1 - \mathbf{x}}{h}\right)}{EK\left(\frac{X_1 - \mathbf{x}}{h}\right)}, \tag{A.8}$$

which is seen to be of order $O(h^s)$, it can be shown that

$$B_n = b_n + h^s O_p((nh^d)^{-1/2}), \qquad \text{as} \quad h \to 0, \quad nh^d \to \infty.$$

Thus, replacing B_n with the constant b_n in (A.7) will not affect the asymptotic normality. The theorem is thus proved. ▌

A.3. *Proof of Theorem* 2. The following lemmas will be used.

LEMMA 2. *We have that, for a* d_{H}*-dimensional Hausdorff measure* μ,

$$\int K\left(\frac{\mathbf{w} - \mathbf{x}}{h}\right) \mu(d\mathbf{w}) \sim h^{d_{\mathrm{H}}}, \qquad \text{as} \quad h \to 0.$$

The proof follows from Falconer (1990, Property 2.1, p. 27) and the fact that $L_1 1_{\{\|\mathbf{u}\| \leqslant a\}} \leqslant K(\mathbf{u}) \leqslant L_2 1_{\{\|\mathbf{u}\| \leqslant 1\}}$ for some a, L_1, L_2 from assumptions on K.

LEMMA 3. *Assumption* (G) *implies that*

$$\sum_{j=1}^{n} |\mathrm{Cov}\{g_n(X_{j+1}), h_n(X_1)\}| \leqslant \beta_n \int g_n(\mathbf{u})\, \mu(d\mathbf{u}) \int h_n(\mathbf{w})\, \mu(d\mathbf{w}), \tag{A.9}$$

for any functions $g_n, h_n \in L^\infty(\Re^p)$ *and* $g_n \geqslant 0$, $h_n \geqslant 0$.

Proof. Note that (4.3) implies that (A.9) is satisfied by $g_n = 1_A$, $h_n = 1_B$ for any Borel sets A, B. It follows that (A.9) holds for any positive simple functions. The general case follows from approximating g_n, h_n, respectively, by a monotone sequence of positive simple functions. ▌

The following lemma is immediate.

LEMMA 4. *For* $g_n \in L^\infty(\Re^p)$ *and* $g_n \geqslant 0$, *Condition* (G) *implies that*

$$\sum_{i=1}^{n} g_n(X_i) = nEg_n(X_1) + O_p\left(n\beta_n \left[\int g_n(\mathbf{w})\, \mu(d\mathbf{w})\right]^2\right). \tag{A.10}$$

Proof. Note that

$$\sum_{i=1}^{n} g_n(X_i) = nEg_n(X_1) + O_p\left(n \sum_{j=1}^{n-1} |\mathrm{Cov}\{g_n(X_{j+1}), g_n(X_1)\}|\right),$$

where the remainder term is of order $O_p(n\beta_n[\int g_n(\mathbf{w})\,\mu(d\mathbf{w})]^2)$ by virtue of Lemma 3. ∎

Proof of Theorem 2. We use the same notations as those used in the proof of Theorem 1. By assumption (F), $(nh^d)^{-1/2}\,T_n$ is the array sum of the martingale difference

$$\xi_{ni} = \frac{1}{\sqrt{nh^d}}\, v^{1/2}(X_i)\, K\left(\frac{X_i - \mathbf{x}}{h}\right)\varepsilon_i.$$

Now we check the Lindberg condition in Theorem 3 of Shiryayev (1984, p. 511):

$$\begin{aligned}
\sum_{i=1}^{n} E\{\xi_{ni}^2 1_{|\xi_{ni}|>\varepsilon} \mid \mathscr{F}_{i-1}\} &\leqslant \sum_{i=1}^{n} E\left\{\frac{|\xi_{ni}|^{2+\delta}}{\varepsilon^{\delta}} \,\middle|\, \mathscr{F}_{i-1}\right\} \\
&= \frac{1}{(nh^d)^{1+\delta/2}\,\varepsilon^{\delta}} \sum_{i=1}^{n} |v^{1/2}(X_i)|^{2+\delta} \\
&\quad \times K\left(\frac{X_i - \mathbf{x}}{h}\right) E\{|\varepsilon_i|^{2+\delta} \mid \mathscr{F}_{i-1}\} \\
&\leqslant \frac{1}{(nh^d)^{1+\delta/2}\,\varepsilon^{\delta}} \sup_{i \geqslant 1} E\{|\varepsilon_i|^{2+\delta} \mid \mathscr{F}_{i-1}\} \\
&\quad \times \sum_{i=1}^{n} |v^{1/2}(X_i)|^{2+\delta} K\left(\frac{X_i - \mathbf{x}}{h}\right). \qquad \text{(A.11)}
\end{aligned}$$

Applying Lemma 4 with $g_n(X_i) = |v^{1/2}(X_i)|^{2+\delta} K((X_i - \mathbf{x})/h)$ and also using Lemma 1, we see that

$$\sum_{i=1}^{n} |v^{1/2}(X_i)|^{2+\delta} K\left(\frac{X_i - \mathbf{x}}{h}\right) = O(nh^d) + O_p(n\beta_n h^{2d_H}).$$

So under Assumption (F), the RHS of (A.11) tends to zero in probability as $nh^d \to \infty$. We have thus verified the Lindberg condition.

In addition,

$$\sum_{i=1}^{n} E\{\xi_{ni}^2 \mid \mathscr{F}_{i-1}\} = \frac{1}{nh^d} \sum_{i=1}^{n} v(X_i)\, K^2\left(\frac{X_i - \mathbf{x}}{h}\right) = \delta_1^2(1 + o_p(1))$$

where $\delta_1 = v(\mathbf{x})\, pc(\mathbf{x}) \int_0^\infty k^2(y)\, y^{p-1}\, dy$. Thus, by Theorem 4 of Shiryayev (1984, p. 511),

$$\frac{1}{\sqrt{nh^d}}\, T_n \xrightarrow{d} N(0, \delta_1^2).$$

Using Lemma 4, the rest of the proof follows the same steps as those in the proof of Theorem 1. Employing the Cramer–Wold device, joint asymptotic normality can be proved similarly. Joint asymptotic independence follows by using Lemma 3 and the fact that K has finite support. We thus complete the proof. ▮

ACKNOWLEDGMENTS

The author gratefully thanks Professor Richard L. Smith for raising the question discussed in this paper and for his valuable comments. He also thanks Professor L. Mark Berliner, two anonymous referees, and the editor in charge of this paper for their useful comments and suggestions which improved the presentation.

REFERENCES

1. H. Z. An and F. C. Huang, The geometric ergodicity of nonlinear autoregressive models, *Statist. Sinica* **6** (1996), 943–956.
2. D. Bosq and D. Guégan, Nonparametric estimation of the chaotic function and the invariant measure of a dynamical system, *Statist. Probab. Lett.* **25**, No. 3 (1995), 201–212.
3. R. C. Bradley, Basic properties of strong mixing conditions, *in* "Dependence in Probability and Statistics, A Survey of Recent Results" (E. Eberlein and M. S. Taqqu, Eds.), pp. 165–192, Birkhäuser, Boston, 1986.
4. K. S. Chan and H. Tong, A note on noisy chaos, *J. R. Statist. Soc. B* **56**, No. 2 (1994), 301–311.
5. C. D. Cutler, A review of the theory and estimation of fractal dimension, *in* "Dimension Estimates and Models" (H. Tong, Ed.), pp. 1–107, World Scientific, Singapore, 1993.
6. K. Falconer, "Fractal Geometry, Mathematical Foundation and Applications," Wiley, Chichester, 1990.
7. J. Fan and Z. Gijbels, "Local Polynomial Modelling and Its Applications," Chapman & Hall, London, 1995.
8. J. Fan and Y. Yao, Efficient estimation of conditional variance functions in stochastic regression, *Biometrika* **85** (1998), 645–660.
9. J. D. Farmer and J. J. Sidorowich, Exploiting chaos to predict the future and reduce noise, *in* "Evolution, Learning, and Cognition" (Y. C. Lee, Ed.), pp. 277–330, World Scientific, Singapore, 1988.
10. J. H. Friedman and W. Stuetzle, Projection pursuit regression, *J. Am. Statist. Assoc.* **76**, No. 376 (1981), 817–823.
11. J. E. Hitchinson, Fractals and self-similarity, *Indiana Univ. Math. J.* **30** (1981), 713–747.
12. P. J. Huber, Projection pursuit (with discussions), *Ann. Statist.* **13** (1985), 435–525.
13. T. Kapitaniak, "Chaos in Systems with Noise," 2nd ed., World Scientific, Singapore, 1990.
14. K. C. Li, Nonlinear confounding in high-dimensional regression, *Ann. Statist.* **25**, No. 2 (1997), 577–612.

15. B. Mandelbrot, "The fractal Geometry of Nature," Freedman, San Francisco, 1982.
16. E. A. Nadaraya, On estimating regression, *Theory Probab. Appl.* **9** (1964), 141–142.
17. P. L. Read, M. J. Bell, D. W. Johnson, and R. M. Small, Quasi-periodic and chaotic flow regimes in a thermally driven, rotating fluid annulus, *J. Fluid Mech.* **238** (1992), 599–632.
18. D. Ruelle, "Chaotic Evolution and Strange Attractors," Cambridge Univ. Press, Cambridge, UK, 1989.
19. R. J. Serfling, "Approximation Theorems of Mathematical Statistics," Wiley, New York, 1980.
20. R. L. Smith, Estimating dimension in noisy chaotic time series, *J. R. Statist. Soc. B* **54**, No. 2 (1992), 329–351.
21. C. J. Stone, Optimal rate of convergence for nonparametric estimators, *Ann. of Statist.* **8**, No. 6 (1980), 1348–1360.
22. H. Tong and R. L. Smith, Royal Statistical Society meeting on chaos, *J. R. Statist. Soc. B* **54**, No. 2 (1992), 301–474.
23. G. S. Watson, Smooth regression analysis, *Sankhyā Ser. A* **26** (1964), 359–372.