

Kansas State University Libraries

New Prairie Press

---

Conference on Applied Statistics in Agriculture

1996 - 8th Annual Conference Proceedings

---

## WAVELET NONPARAMETRIC REGRESSION WITH DEPENDENT DATA

Chengjie Xiong

George A. Milliken

Follow this and additional works at: <https://newprairiepress.org/agstatconference>



Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

---

### Recommended Citation

Xiong, Chengjie and Milliken, George A. (1996). "WAVELET NONPARAMETRIC REGRESSION WITH DEPENDENT DATA," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1326>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact [cads@k-state.edu](mailto:cads@k-state.edu).

# WAVELET NONPARAMETRIC REGRESSION WITH DEPENDENT DATA

Chengjie Xiong and George A. Milliken  
Department of Statistics  
Kansas State University  
Manhattan, KS 66506

## Abstract

Estimation of the regression function has many applications in agriculture and industry. Usually, the regression function is assumed a known functional form which depends on unknown parameters. Nonparametric regression theory makes no such assumption and often uses some kernel functions to form the so-called Watson-Nadaraya type estimators. Such estimators were extensively studied by Watson (1964), Nadaraya (1964, 1989) and Collomb (1981, 1985). When the data are independent, these estimators have nice asymptotic convergence properties. When the data are dependent, Györfi et al (1989) gave some large sample properties for the Watson-Nadaraya estimators. In this paper, the recently developed theory of wavelet will be used to estimate the regression function when the data are dependent. Large sample properties for the wavelet estimator will be proved, and the wavelet smoothing will be compared with the other well known nonparametric smoothing methods.

**Key Words:** Wavelet, Multiresolution Analysis, Mixing Conditions, Complete Convergence in Probability.

## 1 Introduction

One of the primary goals in statistics is estimation of the regression function. More specifically, Given a random vector  $\mathbf{W} = (\mathbf{X}, y) \in \mathbb{R}^{d+1}$  such that the conditional expectation  $E(y|\mathbf{X} = \mathbf{x}) = r(\mathbf{x})$  exists, how should we estimate the function  $r(\mathbf{x})$  based on the data?

There are many different situations to be considered. If we are willing to assume that the data are jointly normally distributed, then  $r(\mathbf{x})$  is a linear function of  $\mathbf{x}$ , and the question becomes how to estimate unknown intercept and partial slope parameters. Several packages can be used to estimate the unknown parameters by the method of least squares when the data are independent. If we believe that a linear function should be used to estimate  $r(\mathbf{x})$ , but that the data are not independent (e.g. time series data), classical time series analysis (Box, Jenkins, 1976) and ARMA models (Doob, 1953, Akaike, 1974) can be used to estimate the unknown parameters. If a plot of the data suggests very clear curvature, none of the above techniques directly apply. If the data  $\{\mathbf{X}_i, y_i\}_{i=1}^n$  are independent, Watson (1964), Nadaraya (1964, 1989) and Collomb (1981, 1985) studied the following so-called Watson-Nadaraya estimator and obtained asymptotic normality and pointwise consistency:

$$\hat{r}_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)},$$

where  $K(\mathbf{x})$  is some nice kernel function on  $R^d$  and  $h_n$  is the bandwidth. If the data  $\{\mathbf{X}_i, y_i\}_{i=1}^n$  are dependent, Gyorfi et al (1989) studied the above estimator and obtained some consistency results, and more importantly, they proved the uniform consistency of the estimator on any compact set  $G$  in  $R^d$  in the sense of complete convergence in probability.

Wavelet analysis is a technique which approximates a signal in a step by step fashion. More specifically, a nested sequence of closed subspaces of  $L^2(R^d)$  can be constructed by spanning the family of functions formed by the scale changes and the translations of a father wavelet. The projections on these subspaces of the regression function are used as the estimator. With independent data, Antoniadis et al. (1994) studied the estimator of  $r(\mathbf{x})$  given by

$$\hat{r}_n^m(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i E_m(\mathbf{x}, \mathbf{X}_i)}{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i)},$$

where  $E_m$  is the  $m$ -th resolution projection from a multiresolution analysis on  $L^2(R^d)$ . They obtained consistency and asymptotical normality.

We consider the problem of nonparametric estimation of regression function by using the theory of wavelets with dependent data. This problem has the potential application in nonlinear autoregressive time series. We state the uniform consistency of the above estimator in the sense of complete convergence in probability in Section 3. The proofs of these results will be given in Section 4. Some simulation study results about the performance of the wavelet regression estimator and a comparison between some well known regression estimators and wavelet estimators are given in Section 5.

## 2 Multiresolution Analysis and Wavelets

The subject of wavelets has evolved very rapidly in the last ten years. In one sense wavelets are no different than other orthogonal systems, such as Legendre polynomials. But there are notable differences: wavelet series have very good pointwise convergence properties, wavelet series are more localized and pick up edge effect better, wavelets use fewer coefficients to represent certain signals and images. Because of such characteristics, wavelet theory has potentially important applications in statistics, especially to nonparametric inference. In this section, the theory of wavelets will be briefly discussed, and then a nonparametric regression estimator based on a scaling function will be discussed. In the sequel,  $L^2(R^d)$  denotes all measurable functions on  $R^d$  which are square integrable with respect to Lebesgue measure.

Definition 2.1:

A multiresolution analysis of  $L^2(R)$  consists of an increasing sequence of closed subspaces  $V_j, j \in Z$ , of  $L^2(R)$  such that

- (a)  $\cap V_j = \{0\}$ ;
- (b)  $\overline{\cup V_j} = L^2(R)$ ;
- (c) There exists a scaling function  $\varphi \in V_0$  such that  $\{\varphi(\cdot - k), k \in Z\}$  is an orthonormal basis of  $V_0$  ;
- (d) for  $h \in L^2(R)$  ,  $h(x) \in V_0 \leftrightarrow h(x - k) \in V_0$  for all  $k \in Z$  , and
- (e)  $h(x) \in V_j \leftrightarrow h(2x) \in V_{j+1}$ .

The intuitive meaning of (e) is that in passing from  $V_j$  to  $V_{j+1}$ , the resolution of the approximation is doubled, and the approximation to the signal becomes more accurate. Mallat (1989) and Meyer (1990) have shown that given any multiresolution analysis, it is possible to derive a function  $\psi$  such that the family  $\{\psi_{j,k}, k \in Z\}$  is an orthonormal basis of the orthogonal complement  $W_j$  of  $V_j$  in  $V_{j+1}$ , and so  $\{\psi_{j,k}, j, k \in Z\}$  is an orthonormal basis of  $L^2(R)$ , where

$$\psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x - k).$$

These  $\psi_{j,k}$  are called wavelets. The function  $\psi$  is called the mother wavelet and the function  $\varphi(x)$  is called the father wavelet. For any function  $f(x) \in L^2(R)$ , there exists a unique sequence of constants  $\{c_{j,k}\}$  such that

$$f(x) = \sum_{j,k \in Z} c_{j,k} \psi_{j,k}.$$

This expansion is called the wavelet expansion of  $f(x)$ .

Since the wavelet series approximation to a function  $f$  is equivalent to the approximation of  $f$  through the subspaces  $V_j$ 's from the multiresolution analysis. We will primarily use the scaling function  $\varphi$  in this paper.

Definition 2.2

A scaling function  $\varphi$  is called  $r$ -regular for some positive integer  $r$  if it is  $r$  times differentiable and its derivatives satisfy for any  $p \in \mathbb{N}$ ,

$$\left| \frac{d^k}{dx^k} \varphi(t) \right| \leq C_{pk} (1 + |t|)^{-p},$$

where  $C_{pk}$  is some constant,  $k \leq r$  and  $t \in \mathbb{R}$ .

Given a multiresolution analysis on  $\mathbb{R}$ , one can construct a sequence of projection operators  $E_j : L^2(\mathbb{R}) \rightarrow V_j$  which are associated with some integral kernel  $E_j(x, y) :$

$$h \rightarrow E_j(h) = \int_{\mathbb{R}} E_j(\cdot, y) h(y) dy.$$

The kernel  $E_j(x, y)$  can be explicitly expressed in terms of the scaling function  $\varphi$  as:

$$E_j(x, y) = 2^j \sum_{k \in \mathbb{Z}} \varphi(2^j x - k) \overline{\varphi}(2^j y - k).$$

It is easy to see that  $E_j(x, y) = 2^j E_0(2^j x, 2^j y)$  and that  $E_0(x, y) = E_0(x + k, y + k)$  for any  $k \in \mathbb{Z}$  and

$$\sum_{k \in \mathbb{Z}} \varphi(x - k) = 1.$$

If  $\varphi$  is  $r$ -regular for some  $r > 0$ , it can be shown that, for any polynomial  $p$  of degree  $\leq r$ ,

$$E_j(p) = p.$$

The multiresolution analysis on  $L^2(\mathbb{R}^d)$  can be similarly defined. If  $\varphi$  is the scaling function that defines a multiresolution on  $L^2(\mathbb{R})$ , then  $\phi(t_1, t_2, \dots, t_d) = \varphi(t_1)\varphi(t_2) \dots \varphi(t_d)$  can be used as a scaling function that gives a multiresolution analysis on  $L^2(\mathbb{R}^d)$ . In general, given a multiresolution analysis on  $L^2(\mathbb{R}^d)$  with scaling function  $\phi(\mathbf{t})$ , the projection operators  $E_m$  and the corresponding kernels  $E_m(\mathbf{x}, \mathbf{y})$  can be expressed as:

$$E_m(f)(\mathbf{x}) = \int_{\mathbb{R}^d} E_m(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y},$$

and

$$E_m(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} \phi(\mathbf{x} - \mathbf{k}) \overline{\phi}(\mathbf{y} - \mathbf{k}). \tag{1}$$

The following inequality is needed later.

Theorem 2.1

If the scaling function  $\phi(\mathbf{t})$  from a multiresolution analysis on  $L^2(\mathbb{R}^d)$  is  $r$ -regular for some  $r > 0$ , then there exist constants  $C_n$  such that

$$|E_0(\mathbf{x}, \mathbf{y})| \leq \frac{C_n}{(1 + |\mathbf{x} - \mathbf{y}|)^n}$$

for any  $n \in N$ .

The proof of this theorem can be found in Walter (1994).

Compactly supported scaling functions  $\varphi$  and wavelet functions  $\psi$  are of particular interest mainly because of the good localization properties and the computational convenience they possess. Notice that if  $\varphi$  is compactly supported, the kernel  $E_m(x, y)$  is essentially a finite sum for any  $x$  and  $y$ . Daubechies (1988) has constructed a family of such  $\varphi_N$  and  $\psi_N$  with the following property: there exists  $\nu > 0$  such that  $\varphi_N, \psi_N \in C^{\nu N}$ , where  $\varphi \in C^{n+\gamma}$  means  $\varphi \in C^n$  and  $\varphi^{(n)}$  is Hölder continuous with exponent  $\gamma$  ( $0 \leq \gamma < 1$ ). For example, Daubechies and Lagarias (1988) obtained

$$\varphi_2 \in C^{.5500} \quad \varphi_3 \in C^{1.0878\dots} \quad \varphi_4 \in C^{1.6179\dots}$$

The graphs of  $\varphi_4(x)$  and  $\varphi_7(x)$  are shown on Figure 1 and 2.

An algorithm given by Daubechies and Lagarias (1988), the cascade algorithm, allows us to compute the compactly supported scaling function as a limit of step functions that are finer and finer scale approximations of  $\varphi_N$ . Given a finite sequence of filter coefficients,  $c_0, c_1, \dots, c_N$ , define the linear operator  $A$  by

$$(Aa)_n = \sum_{k \in Z} c_{n-2k} a_k, \quad a = (a_k)_{k \in Z},$$

where it is understood that  $c_k = 0$  if  $k < 0$  or  $k > N$ . Define  $a^j = A^j a^0$ , where  $(a^0)_0 = 1$  and  $(a^0)_k = 0$  for  $k \neq 0$ . Set

$$\varphi^j(x) = 2^{\frac{j}{2}} \sum_{k \in Z} a_k^j \chi(2^j x - k),$$

where  $\chi$  is the indicator function of the interval  $[\frac{-1}{2}, \frac{1}{2})$ . Then, pointwise we have:

$$\varphi_N(x) = \lim_{j \rightarrow \infty} \varphi^j(x).$$

The coefficients  $c_0, \dots, c_N$  should satisfy some analytical conditions so that  $\varphi_N$  is a scaling function. These sequences can be found in Daubechies (1988). Table 1 and 2 give these filter coefficients for  $N = 4$  and  $N = 7$ .

### 3 Mixing Conditions and Wavelet Nonparametric Regression

Let  $(Z_i)_{i=-\infty}^{\infty}$  be a sequence of random variables taking values either in  $R$  or  $R^d$ . Let  $F_n^m$  ( $n, m \in Z, n \leq m$ ) denote the  $\sigma$ -algebra generated by  $\{Z_i, n \leq i \leq m\}$ .

Definition 3.1

$(Z_i)_{i=1}^{\infty}$  is said to be uniformly mixing if the mixing coefficients

$$\varphi_k = \sup_n \sup_{A \in F_{-\infty}^n, P(A) > 0, B \in F_{n+k}^{\infty}} |P(B|A) - P(B)|,$$

satisfy

$$\lim_{k \rightarrow \infty} \varphi_k = 0.$$

A very important example of uniformly mixing process is the so-called  $m$ -dependence process, which means that there exists some integer  $m > 0$  such that

$$\varphi_k = 0,$$

for all  $k > m$ .

A fundamental inequality that will be used in the proof of the consistency of the wavelet nonparametric regression estimator is due to Collomb as given below in Theorem 3.1. (Collomb, 1984)

Theorem 3.1

If  $(Z_i)_{i=1}^{\infty}$  is uniformly mixing and there exist constants  $d, D$  and  $\delta$  such that

$$EZ_i = 0, \quad |Z_i| \leq d, \quad EZ_i^2 \leq D,$$

and

$$E|Z_i| \leq \delta,$$

then for any  $\epsilon > 0$ ,

$$P\left(\left|\sum_{i=1}^n Z_i\right| > \epsilon\right) \leq e^{3\sqrt{\epsilon n} \frac{\varphi_m}{m} - \alpha\epsilon + 6\alpha^2 n(D + 4\delta d \sum_{i=1}^m \varphi_i)},$$

where  $\alpha$  is a real and  $m$  is an integer satisfying

$$1 \leq m \leq n \quad \alpha m d \leq \frac{1}{4}.$$

Now, given a sequence of random variables  $(Z_i)_{i=1}^n$ , we will assume that the sequence is uniformly mixing, and

$$Z_i = (\mathbf{X}_i, Y_i),$$

where  $\mathbf{X}_i \in R^d$  and  $Y_i \in R$ . Moreover, we will also assume that the conditional expectation of  $Y$  given  $\mathbf{X}$

$$E(Y_i | \mathbf{X}_i = \mathbf{x}) = r(\mathbf{x}).$$

exists and is finite.

Many authors have investigated nonparametric techniques to estimate the function  $r$  when the random variables  $Z_i$  are independent and identically distributed. The most extensively investigated nonparametric regression estimator (Watson, 1964 and Nadaraya 1964) is defined from a convolution kernel  $K$  of  $R^d$  and a sequence of smoothing parameters  $\{h_n\}$  by

$$\hat{r}_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)}.$$

The parameter  $h_n$ , which controls the degree of smoothness of the estimator  $\hat{r}_n$ , is called the bandwidth. Under several mixing conditions, Hardle, Luckaus, Collomb and Gyorfi et. al. also studied the property of the above convolution kernel estimate  $\hat{r}_n$  and obtained the uniform complete convergence in probability on compact sets in  $R^d$  (See Hardle and Luckaus 1984, Collomb and Hardle 1986, Gyorfi et al 1989). With i.i.d. data, Antoniadis et al. (1994) studied the estimator of  $r(\mathbf{x})$  given by

$$\hat{r}_n^m(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i E_m(\mathbf{x}, \mathbf{X}_i)}{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i)},$$

where  $E_m$  is the m-th resolution projection kernel from a multiresolution analysis on  $L^2(R^d)$ . They also obtained consistency and asymptotic normality results. In next section we will study the uniform complete convergence in probability on compact sets of the same estimator of  $r(\mathbf{x})$  based on a multiresolution analysis under some mixing condition of the data.

## 4 Consistency of Wavelet Regression Estimators

Section 3 presented a wavelet estimator of the regression function  $r(\mathbf{x})$ . The first segment of this section presents some regularity conditions and some necessary lemmas. Next, the consistency results are stated, and then the proofs are provided. The same notations as in section 3 are used here. A sequence of random vectors is denoted by  $\{\mathbf{X}_i, y_i\}_{i=1}^\infty$ , where  $\mathbf{X}_i \in R^d$  and  $y_i \in R$ . It is assumed that the conditional expectation

$$E(y_i | \mathbf{X}_i = \mathbf{x}) = r(\mathbf{x})$$

exists and is finite.

Let  $G$  denote a compact set in  $R^d$  and  $G_{\epsilon_0}$  denote the  $\epsilon_0$ -neighborhood of  $G$  for some  $\epsilon_0 > 0$ . Recall that, given a  $r$ -regular multiresolution analysis on  $L^2(R^d)$  with  $r > 0$  and the data  $\{\mathbf{X}_i, y_i\}_{i=1}^n$ , the wavelet estimator for  $r(\mathbf{x})$  at m-th resolution is defined as

$$\hat{r}_n^m(\mathbf{x}) = \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i}{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i)}.$$

Definition 4.1:

A family of functions  $\{f_i(\mathbf{x})\}_{i=1}^\infty$  is said to be equally uniformly continuous on  $G$  if for any  $\epsilon > 0$ , there exists a common  $\delta(\epsilon) > 0$  such that for any  $\mathbf{x}_1$  and  $\mathbf{x}_2$  in  $G$ ,  $|f_i(\mathbf{x}_1) - f_i(\mathbf{x}_2)| \leq \epsilon$  for all  $i \in N$ , whenever  $|\mathbf{x}_1 - \mathbf{x}_2| \leq \delta(\epsilon)$ .

Definition 4.2:

A sequence of random vectors  $\{\mathbf{Z}_i\}_{i=1}^\infty$  is said to converge completely in probability to a random vector  $\mathbf{Z}$  if for any  $\epsilon > 0$

$$\sum_{i=1}^{\infty} P(|\mathbf{Z}_i - \mathbf{Z}| > \epsilon) < \infty.$$



It is denoted by  $\mathbf{Z}_i \xrightarrow{\text{P.co.}} \mathbf{Z}$ .

Three simple lemmas are stated next without proof.

Lemma 4.1:

If a sequence of random vectors  $\{\mathbf{Z}_i\}_{i=1}^{\infty}$  is uniformly mixing, then there exists a constant  $A > 0$  and a nondecreasing sequence of positive integers  $\{k_n\}_{n=1}^{\infty}$  such that  $1 \leq k_n \leq n$  for any  $n \in N$  and

$$\frac{n\varphi_{k_n}}{k_n} \leq A,$$

where  $\varphi_{k_n}$  are the mixing coefficients.

Lemma 4.2:

If a sequence of random vectors  $\{\mathbf{Z}_i\}_{i=1}^{\infty}$  converges completely in probability to  $\mathbf{Z}$ , then it converges to  $\mathbf{Z}$  in probability and almost surely.

Lemma 4.3

If two sequences of random vectors  $(\mathbf{S}_i)_{i=1}^{\infty}$  and  $(\mathbf{T}_i)_{i=1}^{\infty}$  converge completely to 0 in probability, respectively, then

$$\mathbf{S}_i + \mathbf{T}_i \xrightarrow{\text{P.co.}} 0,$$

and

$$\mathbf{S}_i \mathbf{T}_i \xrightarrow{\text{P.co.}} 0.$$

The following regularity conditions are needed in the proof of the consistency results:

(1)  $(\mathbf{X}_i, y_i)_{i=1}^{\infty}$  is uniformly mixing. i.e.

$$\lim_{n \rightarrow \infty} \varphi_n = 0,$$

where  $\varphi_n$  are defined on Section 3.

(2) the regression function  $r(\mathbf{x})$  is uniformly continuous and bounded on  $R^d$ .

(3) There exists a constant  $\Gamma > 0$  such that for any  $i \in N$  and any  $B \in B(R^d)$

$$P(\mathbf{X}_i \in B) \leq \Gamma \mu(B),$$

and there exists another constant  $\gamma > 0$  such that for any  $i \in N$  and any  $B \in B(G_{\epsilon_0})$ ,

$$P(\mathbf{X}_i \in B) \geq \gamma \mu(B),$$

where  $B(\mathbb{R}^d)$  and  $B(G_{\epsilon_0})$  are the  $\sigma$ -field of the Borel sets on  $\mathbb{R}^d$  and  $G_{\epsilon_0}$ , respectively, and  $\mu$  is the Lebesgue measure.

(4) The probability density functions  $\{h_i(\mathbf{x})\}_{i=1}^{\infty}$  of  $\{\mathbf{X}_i\}_{i=1}^{\infty}$  are equally uniformly continuous on  $G_{\epsilon_0}$ .

(5) There exists a constant  $\beta > 2$  and a constant  $C_1 < \infty$  such that for any  $i \in N$

$$E|y_i|^\beta \leq C_1,$$

and there exists another constant  $C_2 < \infty$  such that for any  $i \in N$  and any  $\mathbf{x} \in G_{\epsilon_0}$ ,

$$E[(y_i - r(\mathbf{x}))^2 | \mathbf{X}_i = \mathbf{x}] \leq C_2.$$

(6)  $m \rightarrow \infty$  as  $n \rightarrow \infty$  and there exists a constant  $\xi \in (\frac{4}{\beta+2}, 1)$  such that

$$\lim_{n \rightarrow \infty} \frac{n2^{-md}}{n^\xi k_n \ln n} = \infty,$$

where  $k_n$  is from Lemma 4.1 and  $\beta$  is from condition (5).

The above conditions are in some sense natural. Condition (1) refers primarily to time series data. It implies that the correlation between two observations becomes smaller and smaller as the time gap between the observations gets bigger and bigger. Condition (3) guarantees that the data will be available to estimate the regression function at every point on the compact set  $G$ . Condition (4) requires in some sense the homogeneity of distributions for all observations, although not necessarily identical distributions. Condition (5) is a moment condition. Condition (6) tells us how big the subspace  $V_j$  has to be used for the regression estimators, which basically controls the smoothness of the data smoothing. The following gives an example of processes which satisfies the above regularity conditions.

Example:

Let  $(x_i, y_i) = (x_i, r(x_i))$  for  $i = 1, 2, \dots$ , where  $r(x)$  is a continuously bounded function on  $R$ . Let the process  $\{x_i\}_{i=1}^{\infty}$  be  $m$ -dependence for some positive integer  $m$ . Suppose that  $x_i$  has a normal distribution with mean  $\mu_i$  and variance  $\sigma^2$ , so  $x_i$  has the density function:

$$h_i(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}},$$

for  $i = 1, 2, \dots$ . Further, assume

$$\lim_{i \rightarrow \infty} \mu_i = \mu_0,$$

for some finite  $\mu_0$ . Then all the conditions (1), (2), (3), (4) and (5) are satisfied on any compact set  $G \subset R$ .

Theorem 4.1:

Under the conditions (1), (2), (3), (4), (5) and (6), the wavelet regression estimator  $\hat{r}_n^m(\mathbf{x})$  converges completely in probability to  $r(\mathbf{x})$  uniformly on  $G$ . i.e.

$$\sup_{\mathbf{x} \in G} |\hat{r}_n^m(\mathbf{x}) - r(\mathbf{x})| \xrightarrow{P.co.} 0,$$

as  $n \rightarrow \infty$ .

We now prove Theorem 4.1 through a series of lemmas. Rewrite

$$\hat{r}_n^m(\mathbf{x}) = \frac{f_n(\mathbf{x})}{g_n(\mathbf{x})},$$

where

$$f_n(\mathbf{x}) = \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i}{n}$$

and

$$g_n(\mathbf{x}) = \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i)}{n}.$$

For any  $\xi > 0$ , write:

$$y_i = y_i^{(1)} + y_i^{(2)},$$

and

$$f_n^{(1)}(\mathbf{x}) = \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i^{(1)}}{n},$$

$$f_n(\mathbf{x}) = f_n^{(1)}(\mathbf{x}) + f_n^{(2)}(\mathbf{x}),$$

where  $y_i^{(1)} = y_i \chi_{[|y_i| \geq n^\xi]}$ ,  $y_i^{(2)} = y_i - y_i^{(1)}$  and  $\chi_A$  is the indicator function of the set  $A$ . In the sequel,  $C$  will be a universal constant which may differ from step to step.

Lemma 4.4

Under conditions (1), (2), (3) and (5), there exist constants  $a$  and  $b$  not dependent on  $n$  such that for any  $\epsilon > 0$ ,

$$\sup_{\mathbf{x} \in G} P(|f_n^{(2)}(\mathbf{x}) - E f_n^{(2)}(\mathbf{x})| > \epsilon) \leq a e^{-b \frac{n^{1-\xi} 2^{-md}}{k_n}},$$

where  $k_n$  is given in Lemma 4.1.

Proof: Write

$$\begin{aligned}
 f_n^{(2)}(\mathbf{x}) - \text{Ef}_n^{(2)}(\mathbf{x}) &= \sum_{i=1}^n \frac{E_m(\mathbf{x}, \mathbf{X}_i)y_i^{(2)} - E(E_m(\mathbf{x}, \mathbf{X}_i)y_i^{(2)})}{n} \\
 &\equiv \sum_{i=1}^n \Delta_i.
 \end{aligned} \tag{2}$$

Now by Schwartz's inequality and condition (2) (3) and (5) we have that for any  $\mathbf{x} \in G$ ,

$$E\Delta_i = 0 \tag{3}$$

$$\begin{aligned}
 |\Delta_i| &\leq C \frac{n^{\xi} 2^{md}}{n} \\
 &= d_1,
 \end{aligned} \tag{4}$$

$$\begin{aligned}
 E|\Delta_i| &\leq C \frac{E|E_m(\mathbf{x}, \mathbf{X}_i)y_i|}{n} \\
 &= C \frac{E(E(|E_m(\mathbf{x}, \mathbf{X}_i)y_i||\mathbf{X}_i))}{n} \\
 &\leq Cn^{-1} E\{(Ey_i^2|\mathbf{X}_i)^{\frac{1}{2}}|E_m(\mathbf{x}, \mathbf{X}_i)|\} \\
 &\leq Cn^{-1} E\{(r^2(\mathbf{X}_i) + C_2)^{\frac{1}{2}}|E_m(\mathbf{x}, \mathbf{X}_i)|\} \\
 &\leq Cn^{-1} \\
 &\equiv \delta,
 \end{aligned}$$

$$\begin{aligned}
 E|\Delta_i|^2 &\leq CE \left( \frac{|E_m(\mathbf{x}, \mathbf{X}_i)y_i|}{n} \right)^2 \\
 &= Cn^{-2} EE((E_m(\mathbf{x}, \mathbf{X}_i)y_i)^2|\mathbf{X}_i) \\
 &\leq Cn^{-2} 2^{md} \int_{R^d} (r^2(\mathbf{u}) + C_2) E_m(\mathbf{x}, \mathbf{u}) h_i(\mathbf{u}) d\mathbf{u} \\
 &\leq Cn^{-2} 2^{md} \\
 &\equiv D.
 \end{aligned}$$

Since  $\{\mathbf{X}_i, y_i\}$  is uniformly mixing,  $\{\Delta_i\}$  is uniformly mixing. By Collomb's inequality in Theorem 3.1, for  $1 \leq m^* \leq n$  and  $\alpha m^* d_1 \leq \frac{1}{4}$ ,

$$\begin{aligned}
 &\sup_{\mathbf{x} \in G} P(|f_n^{(2)}(\mathbf{x}) - \text{Ef}_n^{(2)}(\mathbf{x})| > \epsilon) \\
 &\leq e^{3\sqrt{\epsilon n} \frac{\varphi_{m^*}}{m^*} - \alpha \epsilon + 6n\alpha^2 (D + 4\delta d_1 \sum_{i=1}^{m^*} \varphi_i)} \\
 &= e^{3\sqrt{\epsilon n} \frac{\varphi_{m^*}}{m^*} - \alpha \epsilon + 6n\alpha^2 (Cn^{-2} 2^{md} + 4Cn^{-1} Cn^{\xi-1} 2^{md} \sum_{i=1}^{m^*} \varphi_i)}.
 \end{aligned}$$

Now, take  $\alpha$  to be  $\frac{1}{Cm^*n^{\xi-1}2^{md}}$ . Then for some large  $C$ ,

$$\sup_{\mathbf{x} \in G} P(|f_n^{(2)}(\mathbf{x}) - \text{Ef}_n^{(2)}(\mathbf{x})| > \epsilon)$$

$$\begin{aligned} &\leq e^{3\sqrt{\epsilon n} \frac{\varphi_{m^*}}{m^*} - C m^{*-1} n^{1-\xi} 2^{-md} \epsilon + m^{*-1} n^{1-\xi} 2^{-md} [C n^{-\xi} m^{*-1} + C m^{*-1} \sum_{i=1}^{m^*} \varphi_i]} \\ &\leq e^{3\sqrt{\epsilon n} \frac{\varphi_{m^*}}{m^*}} e^{\frac{-2^{-md} n^{1-\xi}}{m^*} [C \epsilon - (C n^{-\xi} + C \frac{\sum_{i=1}^{m^*} \varphi_i)}]}. \end{aligned} \tag{5}$$

Now, appeal to Lemma 4.1 and take  $m^*$  to be the  $k_n$  from the lemma 4.1. There are two possibilities:

Possibility 1:

$$\lim_{n \rightarrow \infty} k_n = \infty$$

In this case, condition (1) gives

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^{k_n} \varphi_i}{k_n} = 0.$$

Thus, when  $n$  is large enough, there exist constants  $a$  and  $b$  such that

$$\sup_{\mathbf{x} \in G} P(|f_n^{(2)}(\mathbf{x}) - \text{Ef}_n^{(2)}(\mathbf{x})| > \epsilon) \leq a e^{-b \frac{n^{1-\xi} 2^{-md}}{k_n}}.$$

Possibility 2:

$$k_n = k_0,$$

for all  $n \geq n_0$ , for some  $n_0$ . From (4.5), if  $m^* > k_0$

$$\frac{n \varphi_{m^*}}{m^*} \leq \frac{n \varphi_{k_0}}{k_0} \leq A,$$

for all  $n \geq n_0$ . Thus there are some other constants  $a'$  and  $b'$  such that when  $n$  is large enough,

$$\sup_{\mathbf{x} \in G} P(|f_n^{(2)}(\mathbf{x}) - \text{Ef}_n^{(2)}(\mathbf{x})| > \epsilon) \leq a' e^{-b' \frac{n^{1-\xi} 2^{-md}}{k_n}}.$$

The proof is now complete.

Lemma 4.5

Under conditions (1), (2), (3), (5) and (6),

$$\sup_{\mathbf{x} \in G} |f_n^{(2)}(\mathbf{x}) - \text{Ef}_n^{(2)}(\mathbf{x})| \xrightarrow{P.co} 0.$$

Proof: By the compactness of  $G$ ,  $G$  can be covered by a set of finite  $d$ -dimensional balls  $\{B_k\}_{k=1}^{l_n}$  of radius  $r$ . Suppose  $B_k$  centers at  $\xi_k$ . Next take  $r$  to be  $\frac{1}{n^\xi 2^{m(d+2)}}$ . Since

$$C l_n r^d \leq \mu(G),$$

then

$$l_n \leq C n^{\xi d} 2^{md(d+2)}.$$

Define for any  $\mathbf{x} \in G$ :

$$S_n(\mathbf{x}) = f_n^{(2)}(\mathbf{x}) - E f_n^{(2)}(\mathbf{x}),$$

$$\tilde{S}_n(\mathbf{x}) = S_n(\mathbf{x}) - S_n(\xi_k),$$

where  $\mathbf{x} \in B_k$  for some  $k = 1, 2, \dots, l_n$ .

Now by the fact that  $E_0(\mathbf{x}, \mathbf{y})$  is Lipschitz,

$$\begin{aligned} & |\tilde{S}_n(\mathbf{x})| \\ = & \left| \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i^{(2)} - E_m(\xi_k, \mathbf{X}_i) y_i^{(2)}}{n} \right. \\ & \left. - E \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i^{(2)} - E_m(\xi_k, \mathbf{X}_i) y_i^{(2)}}{n} \right| \\ \leq & C \frac{\sum_{i=1}^n n^\xi 2^{md} |2^m \mathbf{x} - 2^m \xi_k|}{n} \\ \leq & C n^\xi 2^{md} 2^m r \\ \rightarrow & 0. \end{aligned} \tag{6}$$

as  $n \rightarrow \infty$ . Thus

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x} \in G} |\tilde{S}_n(\mathbf{x})| = 0,$$

which implies that

$$\sup_{\mathbf{x} \in G} |\tilde{S}_n(\mathbf{x})| \xrightarrow{P.co.} 0.$$

Now by Lemma 4.4, if  $n$  is large enough,

$$\begin{aligned} & P\left(\max_{k=1,2,\dots,l_n} |S_n(\xi_k)| > \epsilon\right) \\ \leq & \sum_{k=1}^{l_n} P(|S_n(\xi_k)| > \epsilon) \\ \leq & l_n \sup_{\mathbf{x} \in G} P(|S_n(\mathbf{x})| > \epsilon) \\ \leq & l_n a e^{-b \frac{n^{1-\xi_2-md}}{k_n}} \\ \leq & C (n^\xi 2^{m(d+2)})^d e^{-b \frac{n^{1-\xi_2-md}}{k_n}}. \end{aligned}$$

Let  $\xi$  be from condition (6). If  $M$  is large enough,

$$\begin{aligned} & \sum_{n > M} P\left(\max_{k=1,2,\dots,l_n} |S_n(\xi_k)| > \epsilon\right) \\ \leq & C \sum_{n > M} (n^\xi 2^{m(d+2)})^d e^{-b \frac{n^{1-\xi_2-md}}{k_n}} \\ \leq & C \sum_{n > M} n^{-s} \\ < & \infty, \end{aligned}$$

for some  $s > 1$ . Now the lemma is proved by the following inequality:

$$|S_n(\mathbf{x})| \leq |S_n(\xi_k)| + |\tilde{S}_n(\mathbf{x})|$$

and Lemma 4.3.

Lemma 4.6

Under conditions (1), (2), (3), (5) and (6),

$$\sup_{\mathbf{x} \in G} |f_n(\mathbf{x}) - \text{Ef}_n(\mathbf{x})| \xrightarrow{\text{P.co.}} 0.$$

Proof: By Lemma 4.3, Lemma 4.6 follows if the next statement is true,

$$\sup_{\mathbf{x} \in G} |f_n^{(1)}(\mathbf{x}) - \text{Ef}_n^{(1)}(\mathbf{x})| \xrightarrow{\text{P.co.}} 0. \tag{7}$$

To verify (7), write

$$\begin{aligned} & |f_n^{(1)}(\mathbf{x}) - \text{Ef}_n^{(1)}(\mathbf{x})| \\ = & \left| \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i \chi_{[|y_i| \geq n^\xi]} }{n} - E \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i \chi_{[|y_i| \geq n^\xi]} }{n} \right|. \end{aligned}$$

Since

$$\begin{aligned} & \sup_{\mathbf{x} \in G} \left| \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i \chi_{[|y_i| \geq n^\xi]} }{n} \right| \\ & \leq C \frac{\sum_{i=1}^n 2^{md} |y_i| \chi_{[|y_i| \geq n^\xi]} }{n}, \end{aligned}$$

from Schwartz's inequality, Chebyshev's inequality and condition (5):

$$\begin{aligned} & E \sup_{\mathbf{x} \in G} \left| \frac{\sum_{i=1}^n E_m(\mathbf{x}, \mathbf{X}_i) y_i \chi_{[|y_i| \geq n^\xi]} }{n} \right| \\ & \leq C 2^{md} \frac{\sum_{i=1}^n (E y_i^2)^{\frac{1}{2}} (P(|y_i| \geq n^\xi))^{\frac{1}{2}}}{n} \\ & \leq C 2^{md} \frac{\sum_{i=1}^n n^{-\frac{\xi\beta}{2}}}{n} \\ & \leq C 2^{md} n^{-\frac{\xi\beta}{2}}. \end{aligned}$$

Now, by the inequality

$$|E f_n^{(1)}(\mathbf{x})| \leq E \sup_{\mathbf{x} \in G} |f_n^{(1)}(\mathbf{x})|,$$

for any  $\epsilon > 0$ , we have that

$$\begin{aligned} & P(\sup_{\mathbf{x} \in G} |f_n^{(1)}(\mathbf{x}) - \text{Ef}_n^{(1)}(\mathbf{x})| \geq \epsilon) \\ & \leq \frac{E(\sup_{\mathbf{x} \in G} |f_n^{(1)}(\mathbf{x}) - \text{Ef}_n^{(1)}(\mathbf{x})|)}{\epsilon} \\ & \leq C \cdot 2^{md} n^{-\frac{\xi\beta}{2}}. \end{aligned}$$

By condition (6), there exists a constant  $C > 0$  such that

$$2^{md} n^{-\frac{\xi\beta}{2}} \leq C n^{1 - \frac{\xi(\beta+2)}{2}},$$

so that

$$\begin{aligned} & \sum_{n=1}^{\infty} P(\sup_{\mathbf{x} \in G} |f_n^{(1)}(\mathbf{x}) - \text{Ef}_n^{(1)}(\mathbf{x})| \geq \epsilon) \\ & \leq C \sum_{n=1}^{\infty} n^{1 - \frac{\xi(\beta+2)}{2}} \\ & < \infty. \end{aligned}$$

This finishes the proof.

Remark: A similar proof works for  $g_n(\mathbf{x})$  since if  $y_i = 1$ , then  $f_n(\mathbf{x}) = g_n(\mathbf{x})$ . i.e.

$$\sup_{\mathbf{x} \in G} |g_n(\mathbf{x}) - \text{E}g_n(\mathbf{x})| \xrightarrow{\text{P.co.}} 0,$$

as  $n \rightarrow \infty$ .

Lemma 4.7

Under conditions (2) (3) and (4),

$$\sup_{\mathbf{x} \in G, n \in \mathbb{N}} \left| \text{E}g_n(\mathbf{x}) - \frac{\sum_{i=1}^n h_i(\mathbf{x})}{n} \right| \rightarrow 0,$$

as  $m \rightarrow \infty$ .

Proof: For any  $D > 0$ ,

$$\begin{aligned} & \left| \text{E}g_n(\mathbf{x}) - \frac{\sum_{i=1}^n h_i(\mathbf{x})}{n} \right| \\ & = \left| \frac{\sum_{i=1}^n \int_{R^d} E_m(\mathbf{x}, \mathbf{u})(h_i(\mathbf{u}) - h_i(\mathbf{x})) d\mathbf{u}}{n} \right| \\ & = \left| \frac{\sum_{i=1}^n \int_{R^d} 2^{md} E_0(2^m \mathbf{x}, 2^m \mathbf{u})(h_i(\mathbf{u}) - h_i(\mathbf{x})) d\mathbf{u}}{n} \right| \\ & = \left| \frac{\sum_{i=1}^n \int_{R^d} E_0(2^m \mathbf{x}, 2^m \mathbf{x} - \mathbf{v})(h_i(\mathbf{x} - 2^{-m} \mathbf{v}) - h_i(\mathbf{x})) d\mathbf{v}}{n} \right| \end{aligned}$$



$$\begin{aligned} &\leq \left| \frac{\sum_{i=1}^n \int_{|\mathbf{v}| \leq D} E_0(2^m \mathbf{x}, 2^m \mathbf{x} - \mathbf{v})(h_i(\mathbf{x} - 2^{-m} \mathbf{v}) - h_i(\mathbf{x})) d\mathbf{v}}{n} \right| \\ &\quad + \left| \frac{\sum_{i=1}^n \int_{|\mathbf{v}| > D} E_0(2^m \mathbf{x}, 2^m \mathbf{x} - \mathbf{v})(h_i(\mathbf{x} - 2^{-m} \mathbf{v}) - h_i(\mathbf{x})) d\mathbf{v}}{n} \right| \\ &\equiv I + II, \end{aligned}$$

say.

Now, for any  $\epsilon > 0$ , choose  $D$  large enough so that for any  $\mathbf{x} \in G$ :

$$\begin{aligned} II &= \left| \frac{\sum_{i=1}^n \int_{|\mathbf{v}| > D} E_0(2^m \mathbf{x}, 2^m \mathbf{x} - \mathbf{v})(h_i(\mathbf{x} - 2^{-m} \mathbf{v}) - h_i(\mathbf{x})) d\mathbf{v}}{n} \right| \\ &\leq C \int_{|\mathbf{v}| > D} \frac{1}{(1 + |\mathbf{v}|)^{2d}} d\mathbf{v} \\ &\leq \epsilon. \end{aligned}$$

By condition (4), for the above  $\epsilon > 0$ , there exists a  $\delta > 0$  such that for all  $\mathbf{x}_1, \mathbf{x}_2 \in G_{\epsilon_0}$

$$|h_i(\mathbf{x}_1) - h_i(\mathbf{x}_2)| \leq \epsilon,$$

whenever  $|\mathbf{x}_1 - \mathbf{x}_2| \leq \delta$ . Thus there exists an  $N_0 \in N$  such that  $2^{-m} D < \delta$ , and so for all  $\mathbf{x} \in G$  and  $n \in N$  when  $m > N_0$ ,

$$\begin{aligned} I &= \left| \frac{\sum_{i=1}^n \int_{|\mathbf{v}| \leq D} E_0(2^m \mathbf{x}, 2^m \mathbf{x} - \mathbf{v})(h_i(\mathbf{x} - 2^{-m} \mathbf{v}) - h_i(\mathbf{x})) d\mathbf{v}}{n} \right| \\ &\leq \epsilon \int_{|\mathbf{v}| \leq D} E_0(2^m \mathbf{x}, 2^m \mathbf{x} - \mathbf{v}) d\mathbf{v} \\ &\leq C\epsilon \int_{|\mathbf{v}| \leq D} \frac{1}{(1 + |\mathbf{v}|)^{2d}} d\mathbf{v} \\ &\leq C\epsilon. \end{aligned}$$

The proof is now complete.

Remark: By Lemma 4.7,  $Eg_n(\mathbf{x})$  is bounded when  $n \in N, \mathbf{x} \in G$ , so by using the inequality

$$|g_n(\mathbf{x})| \leq |Eg_n(\mathbf{x})| + |g_n(\mathbf{x}) - Eg_n(\mathbf{x})|$$

and Lemma 4.3, it can be shown that:

$$\sup_{\mathbf{x} \in G} |g_n(\mathbf{x})(g_n(\mathbf{x}) - Eg_n(\mathbf{x}))| \xrightarrow{P.co.} 0. \tag{8}$$

Lemma 4.8

Under conditions (2) (3) and (4), there exists some  $\delta_0 > 0$  and some constant integer  $N_0$  such that for  $m > N_0$

$$\sum_{n=1}^{\infty} P(\inf_{\mathbf{x} \in G} g_n(\mathbf{x}) \leq \delta_0) < \infty.$$

Proof: By lemma 4.7 and condition (3), there exists some constant  $\gamma_0$  such that for all  $\mathbf{x} \in G$  and  $n \in N$  when  $m > N_0$

$$|Eg_n(\mathbf{x})| \geq \gamma_0.$$

Now, take  $\delta_0 < \gamma_0$ , from Lemma 4.6 and

$$\inf_{\mathbf{x} \in G} g_n(\mathbf{x}) \geq \inf_{\mathbf{x} \in G} Eg_n(\mathbf{x}) - \sup_{\mathbf{x} \in G} (Eg_n(\mathbf{x}) - g_n(\mathbf{x})),$$

then

$$\begin{aligned} & \sum_{n=1}^{\infty} P(\inf_{\mathbf{x} \in G} g_n(\mathbf{x}) \leq \delta_0) \\ & \leq \sum_{n=1}^{\infty} P(\sup_{\mathbf{x} \in G} (Eg_n(\mathbf{x}) - g_n(\mathbf{x})) \geq \gamma_0 - \delta_0) \\ & < \infty. \end{aligned}$$

Thus the proof is complete.

Similar to the proof to Lemma 4.8, one can also prove when  $m$  is large,

$$\sum_{n=1}^{\infty} P(\inf_{\mathbf{x} \in G} |g_n(\mathbf{x})| \leq \delta_0) < \infty.$$

#### Lemma 4.9

Under condition (2) and (3),

$$\sup_{\mathbf{x} \in G} |Ef_n(\mathbf{x}) - r(\mathbf{x})Eg_n(\mathbf{x})| \rightarrow 0,$$

as  $n \rightarrow \infty$  and  $m \rightarrow \infty$ .

Proof: By condition (2), for any  $\epsilon > 0$ , there exists some  $\delta(\epsilon) > 0$  such that for  $\mathbf{x}_1, \mathbf{x}_2 \in G_{\epsilon_0}$

$$|r(\mathbf{x}_1) - r(\mathbf{x}_2)| < \epsilon,$$

whenever  $|\mathbf{x}_1 - \mathbf{x}_2| \leq \delta(\epsilon)$ .

Now, for any  $\mathbf{x} \in G$ ,

$$\begin{aligned} & |Ef_n(\mathbf{x}) - r(\mathbf{x})Eg_n(\mathbf{x})| \\ & = \left| \frac{\sum_{i=1}^n (E(E_m(\mathbf{x}, \mathbf{X}_i)y_i) - r(\mathbf{x})E(E_m(\mathbf{x}, \mathbf{X}_i)))}{n} \right| \\ & = \left| \frac{\sum_{i=1}^n (E(E(E_m(\mathbf{x}, \mathbf{X}_i)y_i|\mathbf{X}_i)) - r(\mathbf{x})E(E_m(\mathbf{x}, \mathbf{X}_i)))}{n} \right| \\ & \leq \left| \frac{\sum_{i=1}^n (E(r(\mathbf{X}_i)E_m(\mathbf{x}, \mathbf{X}_i)) - r(\mathbf{x})E_m(\mathbf{x}, \mathbf{X}_i))}{n} \right| \end{aligned}$$

$$\begin{aligned}
 &\leq \frac{|\sum_{i=1}^n \int_{R^d} (r(\mathbf{u}) - r(\mathbf{x})) E_m(\mathbf{x}, \mathbf{u}) h_i(\mathbf{u}) d\mathbf{u}|}{n} \\
 &= \frac{|\sum_{i=1}^n \int_{|\mathbf{u}-\mathbf{x}| \leq \delta(\epsilon)} (r(\mathbf{u}) - r(\mathbf{x})) E_m(\mathbf{x}, \mathbf{u}) h_i(\mathbf{u}) d\mathbf{u}|}{n} \\
 &\quad + \frac{|\sum_{i=1}^n \int_{|\mathbf{u}-\mathbf{x}| > \delta(\epsilon)} (r(\mathbf{u}) - r(\mathbf{x})) E_m(\mathbf{x}, \mathbf{u}) h_i(\mathbf{u}) d\mathbf{u}|}{n} \\
 &\equiv I + II,
 \end{aligned} \tag{9}$$

say.

since

$$\begin{aligned}
 |I| &\leq \frac{\epsilon \sum_{i=1}^n \int_{|\mathbf{u}-\mathbf{x}| \leq \delta(\epsilon)} 2^{md} |E_0(2^m \mathbf{x}, 2^m \mathbf{u})| h_i(\mathbf{u}) d\mathbf{u}}{n} \\
 &\leq C\epsilon,
 \end{aligned}$$

and for the given  $\epsilon > 0$ , there exists some  $N(\epsilon) > 0$  such that if  $m > N(\epsilon)$  then

$$\begin{aligned}
 |II| &\leq \frac{\sum_{i=1}^n \int_{|\mathbf{u}-\mathbf{x}| > \delta(\epsilon)} |(r(\mathbf{u}) - r(\mathbf{x})) E_m(\mathbf{x}, \mathbf{u})| h_i(\mathbf{u}) d\mathbf{u}}{n} \\
 &\leq C \frac{\sum_{i=1}^n \int_{|\mathbf{u}-\mathbf{x}| > \delta(\epsilon)} \frac{C 2^{md}}{(1+2^m|\mathbf{u}-\mathbf{x}|)^{3d}} h_i(\mathbf{u}) d\mathbf{u}}{n} \\
 &\leq C\epsilon,
 \end{aligned}$$

Lemma 4.9 is proved.

Proof of Theorem 4.1

By the following equality and condition (2), for any  $\mathbf{x} \in G$ ,

$$\begin{aligned}
 \hat{r}_n^m(\mathbf{x}) - r(\mathbf{x}) &= \\
 &= \frac{[(f_n(\mathbf{x}) - E f_n(\mathbf{x})) - r(\mathbf{x})(g_n(\mathbf{x}) - E g_n(\mathbf{x})) + (E f_n(\mathbf{x}) - r(\mathbf{x}) E g_n(\mathbf{x}))]}{g_n(\mathbf{x})}.
 \end{aligned}$$

We therefore have that

$$\begin{aligned}
 &\sup_{\mathbf{x} \in G} |\hat{r}_n^m(\mathbf{x}) - r(\mathbf{x})| \\
 &\leq \frac{\sup_{\mathbf{x} \in G} |f_n(\mathbf{x}) - E f_n(\mathbf{x})|}{\inf_{\mathbf{x} \in G} |g_n(\mathbf{x})|} \\
 &\quad + C \frac{\sup_{\mathbf{x} \in G} |g_n(\mathbf{x}) - E g_n(\mathbf{x})|}{\inf_{\mathbf{x} \in G} |g_n(\mathbf{x})|} \\
 &\quad + \frac{\sup_{\mathbf{x} \in G} |E f_n(\mathbf{x}) - r(\mathbf{x}) E g_n(\mathbf{x})|}{\inf_{\mathbf{x} \in G} |g_n(\mathbf{x})|} \\
 &\equiv I + III + III,
 \end{aligned}$$

say.

For any  $\epsilon > 0$ ,

$$\begin{aligned} P(I > \epsilon) &= P(I > \epsilon, \inf_{\mathbf{x} \in G} |g_n(\mathbf{x})| > \delta_0) + P(I > \epsilon, \inf_{\mathbf{x} \in G} |g_n(\mathbf{x})| \leq \delta_0) \\ &\equiv I' + II', \end{aligned}$$

say, where  $\delta_0$  is from Lemma 4.8. By Lemma 4.6, Lemma 4.8 and the remarks after them,

$$\begin{aligned} \sum_{n=1}^{\infty} I' &\leq \sum_{n=1}^{\infty} P(|\sup_{\mathbf{x} \in G} |f_n(\mathbf{x}) - \text{E}f_n(\mathbf{x})| > \delta_0 \epsilon) \\ &< \infty, \end{aligned}$$

and

$$\begin{aligned} \sum_{n=1}^{\infty} II' &\leq \sum_{n=1}^{\infty} P(\inf_{\mathbf{x} \in G} |g_n(\mathbf{x})| \leq \delta_0) \\ &< \infty, \end{aligned}$$

so that:

$$\sum_{n=1}^{\infty} P(I > \epsilon) < \infty.$$

Similarly,

$$\sum_{n=1}^{\infty} P(II > \epsilon) < \infty.$$

By Lemma 4.8, Lemma 4.9 and the remarks after them,

$$\sum_{n=1}^{\infty} P(III > \epsilon) < \infty.$$

Thus, Theorem 4.1 is proved.

## 5 Simulation

Some simulation studies about the wavelet nonparametric regression estimator are conducted. The compactly supported scaling functions given by Daubechies (1988) are used to construct the nonparametric regression estimate. More specifically, in the family of scaling functions given by Daubechies,  $\varphi_7(x)$  is used. In addition, in the cascade algorithm to compute  $\varphi_7(x)$  (Section 2), eight iterations are used.

The following autoregressive time series model is used to generate 100 data pairs  $(x_t, y_t) = (x_t, x_{t+1}), t = 1, 2, \dots, 100$ :

$$x_t = \frac{x_{t-1} + \sin(x_{t-1})}{1 + x_{t-1}^2} + e_t,$$

where  $\{e_i\}_{i=1}^n$  are i.i.d. uniform random variables from -0.5 to 0.5. The regression function is estimated by the proposed wavelet method. To choose the parameter  $m$  optimally, the classical cross-validation method is used, i.e. the best  $m$  should minimize the following cross-validation function:

$$CV(m) = n^{-1} \sum_{i=1}^n (y_i - \hat{r}_{n,i}^m(x_i))^2,$$

where  $\hat{r}_{n,i}^m(x)$  is the leave-one-out estimator obtained by estimating  $r(x)$  with the  $i$ -th data point removed. We found that it is only necessary to consider  $m=1,2,3,4,5$ . Table 3 gives the cross validation mean square error for different  $m$ 's.

Based on the Table 3, it is clear that the wavelet regression estimator with  $m=2$  is the best. To compare our method with some well-known curve estimation methods, we compute the following mean square error for different estimators:

$$MSE = n^{-1} \sum_{i=1}^n (y_i - \hat{r}(x_i))^2.$$

The smoothness parameter is bandwidth (ban) for the kernel estimators, degrees of freedom (df) for the cubic spline,  $m$  for the wavelet estimator and the fraction of data (f) used in the estimation at each point for the lowess estimator. Table 4 gives the MSE for several kernel estimators and the lowess estimator. Figure 3, 4, 5, 6, 7 and 8 show the best estimated curve by the wavelet method and the corresponding comparisons. (solid line is the wavelet estimated curve, dashed lines are the estimated curves from other methods.)

Our simulation results show that the wavelet regression estimator is very competitive compared with the other well known estimators. In addition, as expected, the wavelet regression estimator shows a better performance in picking up the edge effect, especially on the left boundary of the scatter plot. We also see that wavelet estimator can pick up some very small fluctuations on the right portion of the scatter plot while keeping the smoothness of the estimated curve, this is in contrast with the fact that most other smoothers tend to oversmooth in that portion of the data.

## 6 Conclusions

It is well known that most currently available data smoothers do not perform very well on the boundary. Since all smoothers use the weighted averages to estimate

the regression function, it is expected the estimates at the boundary points will be somehow skewed since only data from one side are available. Although the wavelet regression estimator still uses the weighted averages to estimate the regression function, it is superior in the sense that the wavelet kernel is compactly supported, very smooth and highly concentrated around the peak point. Thus, even though it still uses the one side data to estimate the regression function on the boundary, it can more efficiently use the data from the one side to perform the estimation by concentrating on the data very close to the points being estimated. This is also why the wavelet estimator is very sensitive to the fluctuations of the data even in the interior part of the data. On the other hand, because of the smoothness of the wavelet kernel, the sensitivity to the fluctuations of the data does not tend to undersmooth the data overall. In this sense, the wavelet estimator nicely balances the smoothness and sharpness of the data smoothing. Another advantage of the wavelet estimator is in the cross validation. Since the smoothing parameter  $m$  is discrete, it is feasible to choose  $m$  by using cross validations.

## 7 Acknowledgements

We wish to thank Dr. Paul I. Nelson and Dr. S. S. Yang for a careful reading of the manuscript.

## 8 References

- Akaike, H. Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Ann. Inst. Stat. Math.*, 23, pp363 – 387.
- Antoniadis, A., Gregoire, G., and Mckeague, I.W. (1994), Wavelet Methods for Curve Estimation, *Journal of the American Statistical Association*. 89, p1340 – 1353.
- Billingsley, P. (1968) *Convergence of Probability Measures*, John Wiley.
- Box, G. E. P., Jenkins, G. M. (1976) *Time series analysis: Forecasting and control*. Holden-Day, San Francisco.
- Chui, K. (1992) *Wavelet: A tutorial in Theory and Applications*, Boston: Academic Press.
- Collomb, G. (1981) Estimation non parametrique de la regression: revue bibliographique. *Inter. Stat. Review*, 49, pp75 – 93.
- Collomb, G. (1984) Proprietes de convergence presque complete du predicteur a noyau. *Z. wahrscheinlichkeitstheorie und verw. Gebiete*, 66, pp.441 – 460.
- Collomb, G. (1985) Nonparametric regression: an up to date bibliography. *Statistics*, 2, pp309 – 324.
- Collomb, G. and Hardle, W. (1986) Strong Uniform Convergence Rates in Robust Nonparametric Time Series Analysis, *Stochastic Processes and Their*

- Applications*, 23, pp77 – 89.
- Daubechies, I. (1988). Orthonormal Bases of Compactly Supported Wavelets. *Communications on Pure and Applied Mathematics*.
- Daubechies, I. and Lagarias, J.C. (1988a), Two-scale Difference Equations: I. Global Regularity of Solutions. *Technical Report, Bell Laboratories*.
- Daubechies, I. and Lagarias, J.C. (1988b), Two-scale Difference Equations: II. Local Regularity, Infinity Products of Matrices and Fractals. *Technical Report, Bell Laboratories*.
- Doob, J. L. (1953) *Stochastic Process*. Wiley, New York.
- Doukhan, P. (1991) Consistency of Delta-Sequence Estimates of a Density or of a Regression Function for a Weakly Dependent Stationary Sequence, *Seminaire de Statistique'orsay, Universite Paris Sud*.
- Gasser, T., and Muller, H. (1979), *Curve Estimation*, Heidelberg: Springer-Verlag.
- Gasser, T., and Muller, H. (1984), Estimating Regression Functions and Their Derivatives by the Kernel Method, *Scandinavian Journal of Statistics*, 11, 171 – 185.
- Gyorfi, L., Hardle, W., Sarda, P., and Vieu, P., (1989) Nonparametric Curve Estimation From Time Series, *Lecture Notes in Statistics*, 60, Springer-Verlag.
- Hall, H., and Heyde, C. (1980) *Martingale Limit Theory and its Applications*, Academic Press.
- Hardle, W., and Luckhaus, S. (1984) Uniform Consistency of a Class of Regression Estimators, *Annals of Statistics*, 12, pp612 – 623.
- Mallat, S. (1989), Multiresolution Approximations and Wavelet Orthonormal Bases of  $L^2(\mathbb{R})$ , *Transactions of the American mathematical Society*, 315, 69 – 87.
- Meyer, Y. (1990) *Ondelettes et Operateurs I : Ondelettes*, Paris: Hermann.
- Nadaraya, E.A. (1964) On estimating regression. *Theory of Probability and its Application*, 9, pp141 – 142.
- Nadaraya, E.A. (1989) *Nonparametric Estimation of Probability Densities and Regression Curves*, Kluwer Academic Publishers.
- Rao, P. (1983) *Nonparametric Functional Estimation*, Academic Press.
- Tong, H. (1990) *Non – linear Time Series, A Dynamical System Approach*, Clarendon Press Oxford.
- Walter, G. (1994) *Wavelets and Other Orthogonal Systems With Applications*, CRC Press.
- Watson, G.S. (1964) Smooth regression analysis. *Sankhya A*, 26, pp359 – 372.

Table 1: Filter Coefficients for  $\varphi_4(x)$ 

$c_0$	0.230377813309
$c_1$	0.714846570553
$c_2$	0.630880767930
$c_3$	-0.027983769417
$c_4$	-0.187034811719
$c_5$	0.030841381836
$c_6$	0.032883011667
$c_7$	-0.010597401785

Table 2: Filter Coefficients for  $\varphi_7(x)$ 

$c_0$	0.077852054085
$c_1$	0.396539319482
$c_2$	0.729132090846
$c_3$	0.469782287405
$c_4$	-0.143906003929
$c_5$	-0.224036184994
$c_6$	0.071309219267
$c_7$	0.080612609151
$c_8$	-0.038029936935
$c_9$	-0.016574541631
$c_{10}$	0.012550998556
$c_{11}$	0.000429577973
$c_{12}$	-0.001801640704
$c_{13}$	0.000353713800



Table 3: Cross-validation mean square error

m	CV
1	0.11353
2	0.10355
3	0.12127
4	0.33212
5	1.21126

Table 4: Mean square error for different estimators

Method	Smoothness	MSE
wavelet	m=2	0.0835154
normal kernel	ban=0.25	0.0822207
box kernel	ban=0.25	0.08865488
triangle kernel	ban=0.25	0.08144265
cubic spline	df=10	0.08095377
lowess	f=1/2	0.0889217

Figure 1: Daubechies Scaling Function  $\varphi_4(x)$

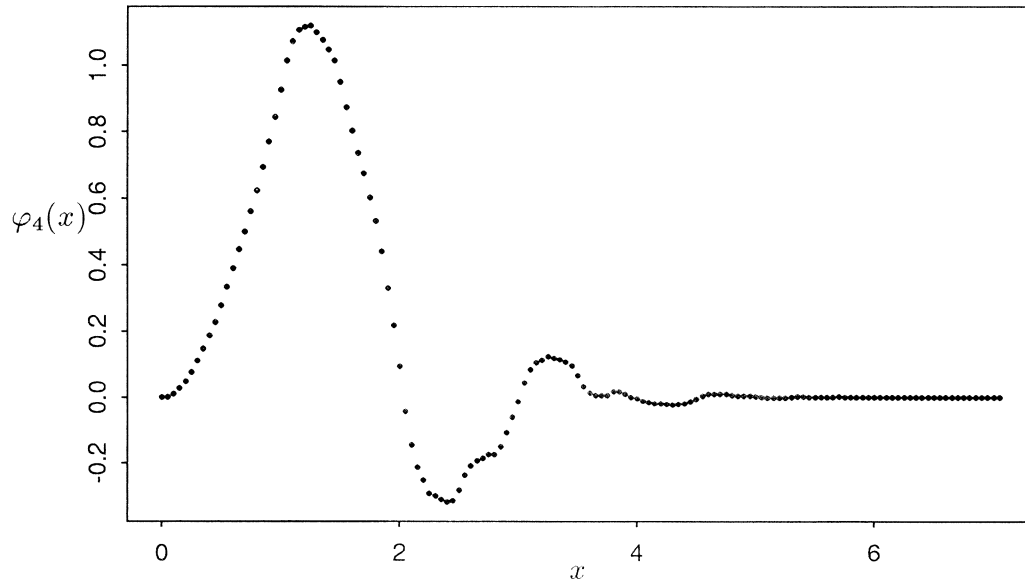


Figure 2: Daubechies Scaling Function  $\varphi_7(x)$

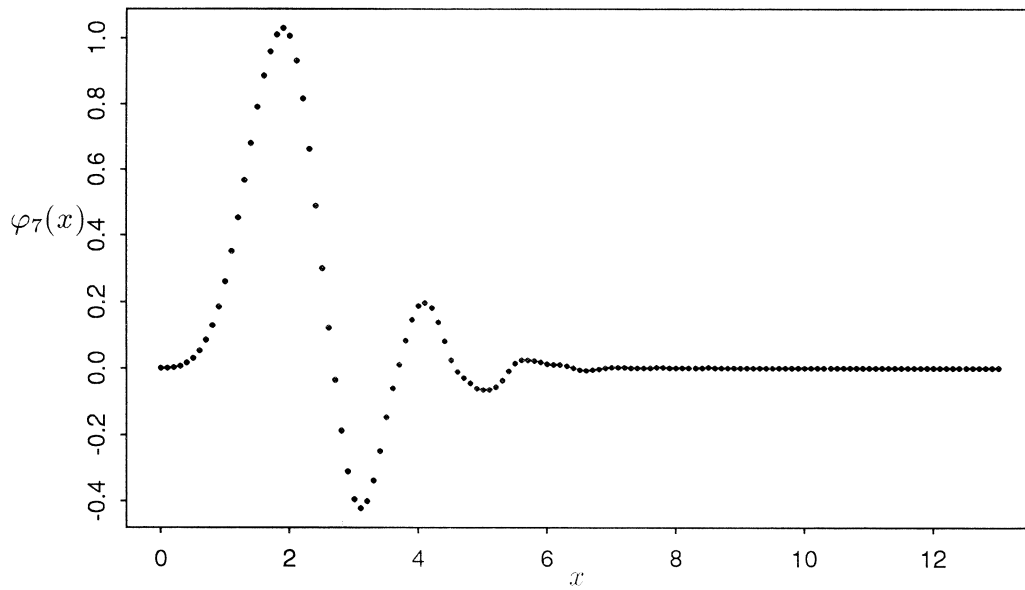


Figure 3: Regression Curve With Optimal  $m=2$

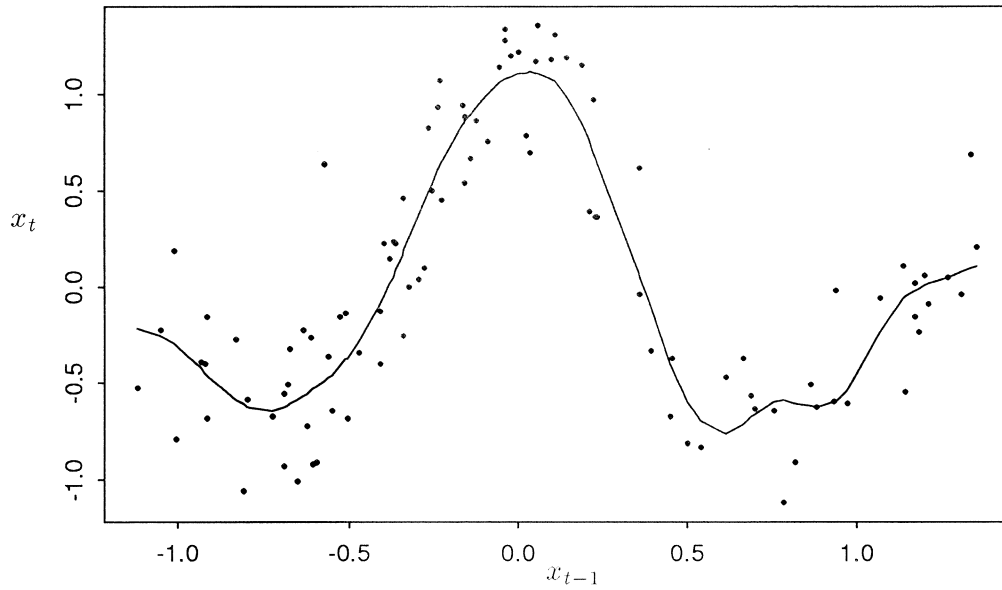


Figure 4: Comparison of the wavelet method and the normal kernel method

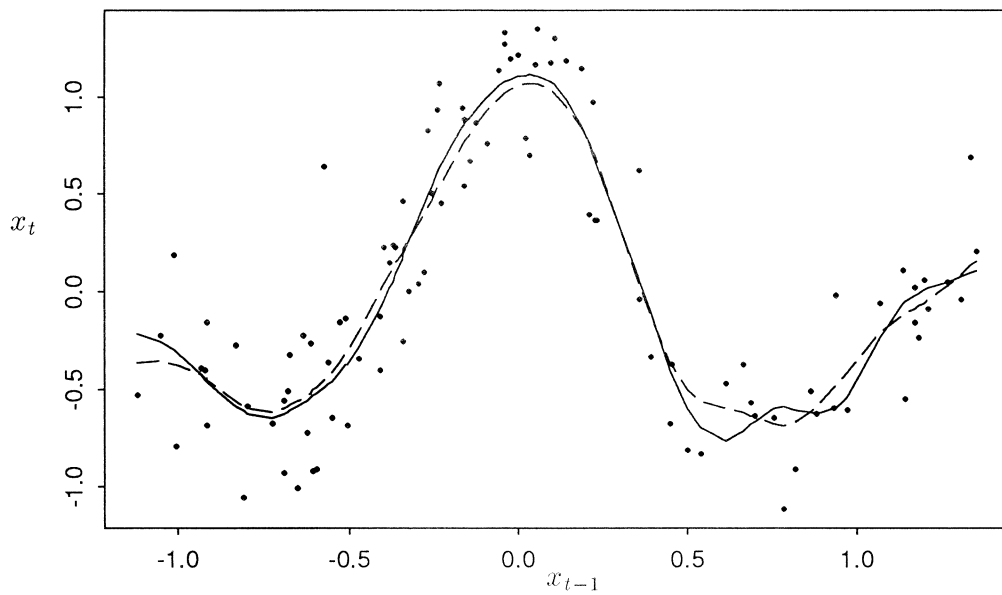


Figure 5: Comparison of the wavelet method and the box kernel method

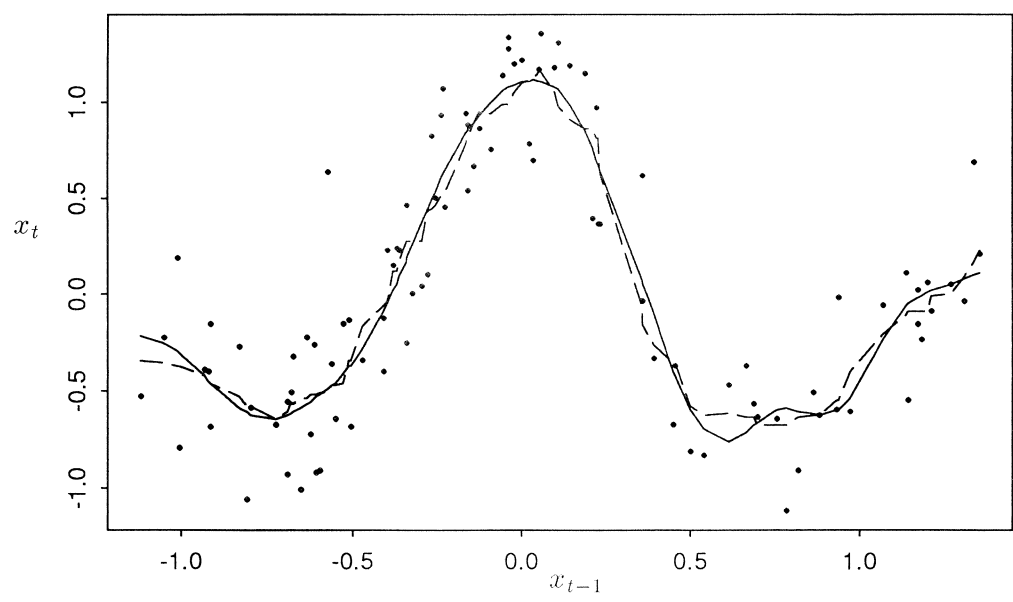


Figure 6: Comparison of the wavelet method and the triangle kernel method

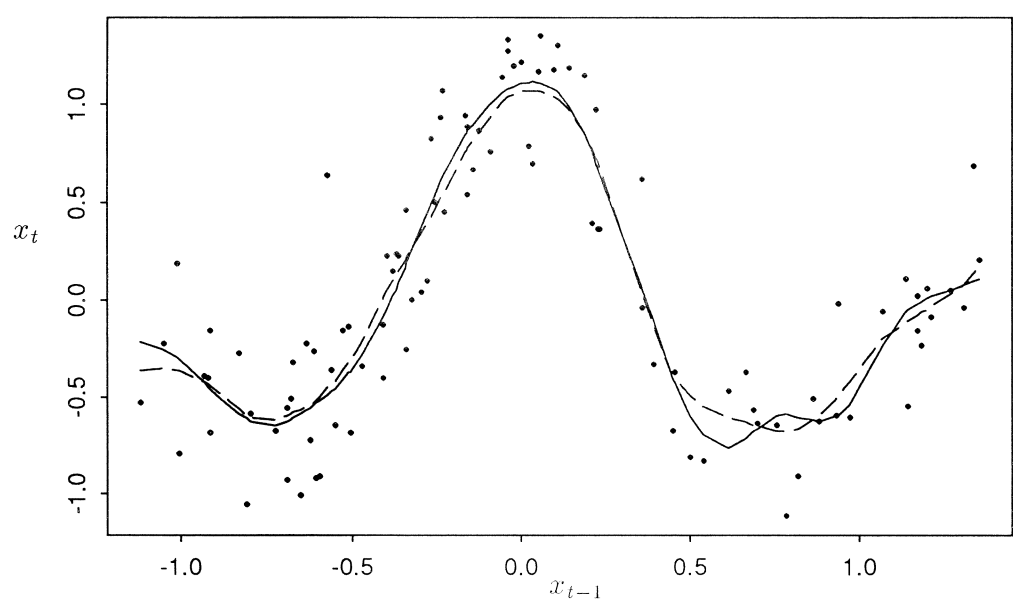


Figure 7: Comparison of the wavelet method and the lowess method

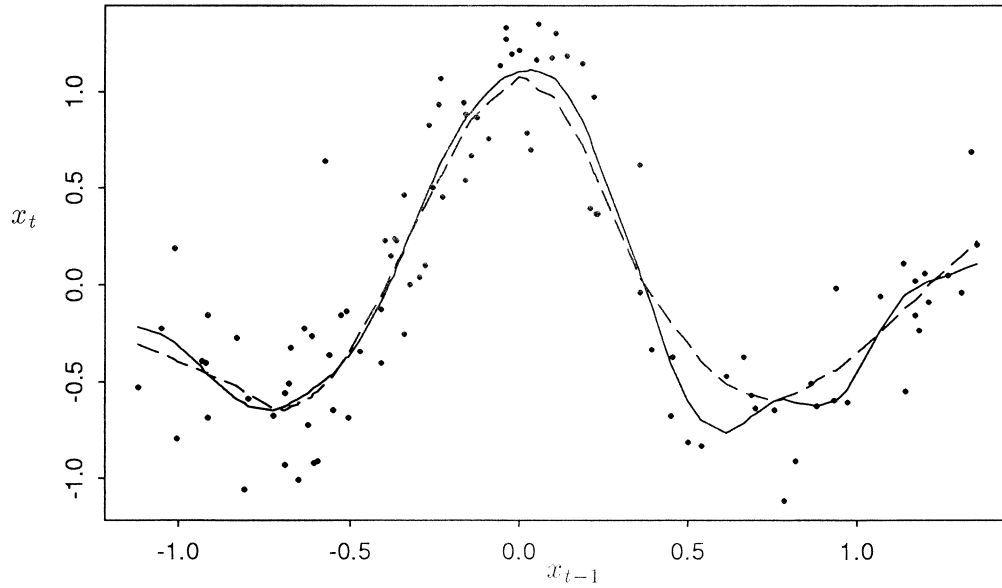


Figure 8: Comparison of the wavelet method and the cubic spline method

