Semi-Parametric Time Series Regression

By

Young K. Truong[1] and Charles J. Stone[2]

University of North Carolina, Chapel Hill and University of California, Berkeley

Technical Report No. 336
February 1992

[1]This research was supported in part by NSF Grant DMS-89-14344

[2]This research was supported in part by NSF Grant DMS-89-02016

Department of Statistics
University of California
Berkeley, California 94720

SEMI-PARAMETRIC TIME SERIES REGRESSION

by YOUNG K. TRUONG[1] and CHARLES J. STONE[2]

University of North Carolina, Chapel Hill, and University of California, Berkeley

February 12, 1992

Abstract. Let $(\mathbf{X}_i, Y_i)$, $i = 0, \pm 1, \ldots$ denote a bivariate stationary time series with $\mathbf{X}_i$ being $\mathbb{R}^d$-valued and Y_i being real-valued. We consider the regression model

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

where $\theta(\cdot)$ is an unknown function and Z_i is an $AR(\nu)$ process. Given a realization of length n, we examine the problem of estimating the nonparametric function $\theta(\cdot)$ and the parametric component Z_i. Under appropriate regularity conditions, it is shown that both components can be optimally estimated.

Keywords. Nonparametric regression, local polynomial estimator, optimal rate of convergence, time series, autoregressive process.

[1]This research was supported in part by National Science Foundation Grant DMS-89-14344.

[2]This research was supported in part by National Science Foundation Grant DMS-8902016.

1 Introduction

Nonlinear time series modelling provides a useful alternative to traditional data analysis and it has been an active area of research lately. Comprehensive accounts of this important direction are described in Priestley (1988) and Tong (1990), where specific properties of models, testing and estimation are also obtained. There are, however, fewer results on model identification. This important issue has led a number of authors to consider nonparametric problems of estimating conditional mean and median functions. The usefulness of this approach in model identification is discussed, for example, in Robinson (1983, 1984), Auestad and Tjøstheim (1990), Truong and Stone (1992) and the references cited therein.

The present paper considers a semiparametric approach to regression problems involving bivariate stationary time series. To motivate this approach, let (X_i, Y_i) be a bivariate stationary time series in which

$$\begin{aligned} X_i &= \gamma X_{i-1} + \varepsilon_i, \quad |\gamma| < 1, \quad \varepsilon_i \sim_{\text{iid}} (0, \sigma^2), \\ Y_i &= \theta(X_i) + Z_i, \\ Z_i &= \eta Z_{i-1} + \xi_i, \quad |\eta| < 1, \quad \xi_i \sim_{\text{iid}} (0, \tau^2). \end{aligned} \tag{1.1}$$

Suppose $\{\varepsilon_i\}$ and $\{\xi_i\}$ are independent sequences of random variables, so that $\theta(X_i) = E(Y_i|X_i)$. Consider now the problem of estimating the parameter η as well as the regression function $\theta(\cdot)$. The solution to such problems would lead to a better understanding of the underlying structures of the response series Y_i and the bivariate series (X_i, Y_i).

In classical time series analysis, the estimation is carried out by putting parametric (linear) assumptions on $\theta(\cdot)$. This approach, known as parametric regression analysis, is described in detail by Brillinger (1981), Brockwell and Davis (1987), Fuller (1976), Hannan (1970), Priestley (1981), Shumway (1988) and the references cited therein; see also Cochran and Orcutt (1949). The approach can be made more flexible or general by adopting nonlinear models as considered by Priestley (1988) and Tong (1983, 1990). In practical applications, one would have to address the issue of model identification.

To develop a general approach to the above problem, we extend model (1.1) as follows. Let $(\mathbf{X}_i, Y_i)$, $i = 0, \pm 1, \ldots$ denote a bivariate stationary time series with $\mathbf{X}_i$ being $\mathbb{R}^d$-valued and Y_i being real-valued. Also, let $\theta(\cdot)$ be a smooth real-valued function on $\mathbb{R}^d$ and consider the following regression model:

$$Y_i = \theta(\mathbf{X}_i) + Z_i, \tag{1.2}$$

where $Z_i = \sum_{u=-\infty}^{\infty} a_{i-u}\varepsilon_u$. Here $\{a_u : u = 0, \pm 1, \pm 2, \ldots\}$ is a sequence of unknown parameters and $\{\varepsilon_u : u = 0, \pm 1, \ldots\}$ is a sequence of iid random variables with mean zero and variance σ^2. Furthermore, suppose the sequences $\{\mathbf{X}_i, i = 0, \pm 1, \pm 2, \ldots\}$ and $\{\varepsilon_i, i = 0, \pm 1, \pm 2, \ldots\}$ are independent, so that $\theta(\mathbf{X}_i) = E(Y_i|\mathbf{X}_i)$.

To enhance the flexibility in describing the relationship between the response and predictive series, as well as the ability to identify various nonlinear models, it is desirable to eliminate the parametric assumptions on the mean function $\theta(\cdot)$. With this semiparametric approach, the present paper generalizes the results presented in Truong and Stone (1992) in several directions. In particular, inference for the autocorrelated structures in the response series [such as the estimation of η in model (1.1)] can be given via the usual parametric approach. Moreover, results on rates of convergence in the estimation of the regression function $\theta(\cdot)$ can be improved. More specifically, under appropriate conditions to be given in the following sections, it is shown that local polynomial estimators of the regression function $\theta(\cdot)$ can be chosen to achieve both local and global optimal rates of convergence as given in Stone (1980, 1982). Furthermore, if the errors Z_i form part of a parametric time series model with finitely many unknown parameters, then these parameters can be estimated with the usual root-n rate of convergence.

The rest of this paper is organized as follows. Results on rates of convergence are given in Section 2, where the local polynomial estimators and conditions required for the achievability of the convergence rates are also described. The fitting of the error process is treated in Section 3, and proofs are given in Section 4.

2 Estimation of the Non-parametric Regression $\theta(\cdot)$

The estimator of the function $\theta(\cdot)$ will now be described. For each given $n \geq 1$, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ denote observations obtained from (1.2) and let δ_n be positive numbers that tend to zero as $n \to \infty$. For a given $\mathbf{x} \in \mathbb{R}^d$, set

$$I_n(\mathbf{x}) = \{i : 1 \leq i \leq n \text{ and } \|\mathbf{X}_i - \mathbf{x}\| \leq \delta_n\}$$

and let $N_n(\mathbf{x}) = \#I_n(\mathbf{x})$ denote the number of points in $I_n(\mathbf{x})$, where $\|\mathbf{x}\| = (x_1^2 + \cdots + x_d^2)^{1/2}$ for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Given $k \geq 1$, let $\hat{P}_{n,\mathbf{x}}(\mathbf{u})$ denote the polynomial in $(u_i - x_i)$ [here $\mathbf{u} = (u_1, \ldots, u_d)$] of degree $(k-1)$ such that it minimizes

$$\sum_{i \in I_n(\mathbf{x})} [Y_i - P_{\mathbf{x}}(\mathbf{X}_i)]^2$$

over the class of $(k-1)$-th degree polynomials $P_{\mathbf{x}}(\mathbf{u})$. (The existence of $\hat{P}_{n,\mathbf{x}}(\cdot)$ follows from Lemma 1 in Section 4.) Define an estimator of $\theta(\mathbf{x})$ based on $\hat{P}_{n,\mathbf{x}}(\cdot)$ by

$$\hat{\theta}_n(\mathbf{x}) = \hat{P}_{n,\mathbf{x}}(\mathbf{x}).$$

For example, suppose $d = 1$ and $k = 2$. Then $\hat{P}_n(X) = \hat{\alpha} + \hat{\beta}(X - x)$ where $\hat{\alpha}$ and $\hat{\beta}$ are obtained by minimizing

$$\sum_{i \in I_n(x)} [Y_i - \alpha - \beta(X_i - x)]^2.$$

Thus

$$\hat{\theta}_n(x) = \hat{\alpha} = \sum_{i \in I_n(x)} w_{n,i} Y_i \Big/ \sum_{i \in I_n(x)} w_{n,i},$$

where

$$w_{n,i} = s_{n,2} - (X_i - x) s_{n,1}, \qquad i \in I_n(x)$$

and

$$s_{n,j} = \sum_{i \in I_n(x)} (X_i - x)^j, \qquad j = 1, 2.$$

The estimator defined by $\hat{\theta}_n(x) = \hat{\alpha}$ is called the local linear estimator.

The conditions required for the rates of convergence of the nonparametric estimators $\hat{\theta}_n(\cdot)$ treated here depend on the following conditions.

Let U be a nonempty open subset of the origin of $\mathbb{R}^d$. The following smoothness condition is imposed on the function $\theta(\cdot)$ so that the bias of $\hat{\theta}_n(\cdot)$ can be dominated.

Condition 1 *$\theta(\cdot)$ has bounded partial derivatives of order k on U.*

The following two conditions are required to show the existence of $\hat{P}_{n,\mathbf{x}}(\cdot)$. Consequently, it also guarantees the existence of the nonparametric estimator $\hat{\theta}_n(\cdot)$. These conditions are similar to those used in ordinary least square theory for random effect models.

Condition 2 *The distribution of $\mathbf{X}_1$ is absolutely continuous and its density $f(\cdot)$ is bounded away from zero and infinity on U; that is, there is a positive constant M_1 such that $M_1^{-1} \leq f(\mathbf{x}) \leq M_1$ for $\mathbf{x} \in U$.*

Condition 3 *For $j \geq 1$, the conditional distribution of $\mathbf{X}_j$ given $\mathbf{X}_0 = \mathbf{x}$ has a density $f_j(\cdot|\mathbf{x})$; there is a positive constant M_2 such that*

$$M_2^{-1} \leq f_j(\mathbf{x}'|\mathbf{x}) \leq M_2 \quad \textit{for } \mathbf{x}, \mathbf{x}' \in U \textit{ and } j \geq 1.$$

Let $\mathcal{F}_j$ and $\mathcal{F}^j$ denote the σ-fields generated respectively by $(\mathbf{X}_i, Y_i)$, $-\infty < i \leq j$, and $(\mathbf{X}_i, Y_i)$, $j \leq i < \infty$. Given a positive integer u set

$$\alpha_u = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_j \text{ and } B \in \mathcal{F}^{j+u}\}.$$

The stationary sequence is said to be α-mixing or strongly mixing if $\alpha_u \to 0$ as $u \to \infty$.

Condition 4 *$\{\mathbf{X}_i, i = 0, \pm 1, \pm 2, \ldots\}$ is α-mixing with*

$$\alpha_u = O(\rho^u), \qquad 0 < \rho < 1, \quad u = 0, 1, 2, \ldots.$$

Conditions 3 and 4 are not required if $\{\mathbf{X}_i, i = 0, \pm 1, \ldots\}$ is a sequence of independent random variables. In connection with model (1.1) described in Section 1, it is useful to note that Gorodetskii (1977) and Withers (1981) have given sufficient conditions for linear processes to be α-mixing. See also Auestad and Tjøstheim (1990) for an illuminating discussion on the role of α-mixing (or geometric ergodicity) for model identification in nonlinear time series analysis.

The following condition is necessary for working with the variance term:

Condition 5 $\sum_u |a_u| < \infty$.

Finally, it is necessary to impose a condition similar to Condition 2 of Stone (1982) under which the L_∞ rate is achievable. We recall that if ξ is a random variable with $E|\xi|^q < \infty$, then the qth order cumulant of ξ is defined by

$$\mathrm{cum}_q(\xi) = \sum (-1)^{j-1}(j-1)! E\left(\xi^{\nu_1}\right) \cdots E\left(\xi^{\nu_j}\right)$$

where the sum extends over all partitions $\nu_1, \ldots, \nu_j, j = 1, 2, \ldots, q$, of $\{1, \ldots, q\}$. See Brillinger (1981) or Shiryayev (1984). Let $\phi(\cdot)$ denote the characteristic function of ε_0.

Condition 6 *The function* $\log \phi(\cdot)$ *has a Taylor expansion:*

$$\sum_q C_q z^q / q! < \infty, \qquad z \textit{ in some neighborhood of the origin,}$$

where $C_q = |\mathrm{cum}_q(\varepsilon_0)|$.

In particular, the above condition is satisfied when ε_0 is a Gaussian random variable.

Given positive numbers b_n and c_n, $n \geq 1$, $b_n \sim c_n$ means that b_n/c_n is bounded away from zero and infinity. Given random variables V_n, $n \geq 1$, $V_n = O_p(b_n)$ means that the random variables $b_n^{-1} V_n$, $n \geq 1$, are bounded in probability; that is, that

$$\lim_{c \to \infty} \limsup_n P(|V_n| > c b_n) = 0.$$

Also, $V_n = o_p(b_n)$ means that the random variables $b_n^{-1} V_n$, $n \geq 1$ converges to zero in probability:

$$\lim_{n \to \infty} P(|V_n| > c b_n) = 0, \qquad c > 0.$$

Set $r = k/(2k+d)$. The pointwise rate of convergence for the estimator $\hat{\theta}_n(\cdot)$ will now be given.

Theorem 1 *Suppose Conditions* 1–5 *hold and that* $\delta_n \sim n^{-1/(2k+d)}$. *Then*

$$|\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| = O_p(n^{-r}), \qquad \mathbf{x} \in U.$$

Let C be a fixed compact subset of U having a nonempty interior. Given a real-valued function $g(\cdot)$ on C, set

$$\|g\|_2 = \left(\int_C |g(\mathbf{x})|^2 \, d\mathbf{x}\right)^{1/2} \quad \text{and} \quad \|g\|_\infty = \sup_{\mathbf{x} \in C} |g(\mathbf{x})|.$$

The L_2 and L_∞ rates of convergence are given in the following results.

Theorem 2 *Suppose Conditions* 1–5 *hold and that* $\delta_n \sim n^{-1/(2k+d)}$. *Then*

$$\|\hat{\theta}_n(\cdot) - \theta(\cdot)\|_2 = O_p(n^{-r}).$$

Theorem 3 *Suppose Conditions* 1–6 *hold and that* $\delta_n \sim (n^{-1}\log n)^{1/(2k+d)}$. *Then there is a positive constant c such that*

$$\lim_n P\left(\|\hat{\theta}_n(\cdot) - \theta(\cdot)\|_\infty \geq c(n^{-1}\log n)^r\right) = 0.$$

Proofs of Theorems 1–3 will be given in Section 4. According to Stone (1980, 1982), the rates presented in Theorems 1–3 are optimal. In the fixed design setup (non-random $\mathbf{X}_i$), similar results are obtained by Müller and Stadtmüller (1987) [based on a finite-order moving average process] and Truong (1991). Based on the α-mixing of the bivariate series (instead of $\{\mathbf{X}_i\}$ alone), Theorem 2 was established by Truong and Stone (1992); Theorem 3 was also established in this paper under the additional assumption that the series $\{Y_i\}$ be bounded.

The above results can also be extended to the estimation of the derivatives of the regression function $\theta(\cdot)$. Given nonnegative integers $\alpha_1, \ldots, \alpha_d$, set $\alpha = (\alpha_1, \ldots, \alpha_d)$ and

$[\alpha] = \alpha_1 + \cdots + \alpha_d$. Let the regression function $\theta(\cdot)$ be k-times ($k \geq 1$) differentiable on $\mathbb{R}^d$. Set

$$T\theta(\cdot) = \sum_{[\alpha]<k} q_\alpha D^\alpha \theta(\cdot),$$

where q_α are constants and D^α denotes the differential operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Let m denote the order of T; that is, set $m = \max\{[\alpha] : 0 \leq [\alpha] < k \text{ and } q_\alpha \neq 0\}$. Note that the function $T\theta(\cdot) = \theta(\cdot)$ corresponds to $m = 0$, while $T\theta(\cdot) = \partial\theta(\cdot)/\partial x_1$ corresponds to $m = 1$. Under the conditions given in Theorems 1–3, a sequence of local polynomial estimators of $T(\theta)$ can be chosen to achieve optimal rates $n^{-(k-m)/(2k+d)}$ both pointwise and in L_2 norm. Furthermore, it can also be chosen to achieve the optimal L_∞ rate $(n^{-1}\log n)^{(k-m)/(2k+d)}$.

3 Estimation of the AR Parameters

Suppose that $\{Z_i\}$ is a stationary time series in which

$$Z_i = \beta_1 Z_{i-1} + \cdots + \beta_\nu Z_{i-\nu} + \varepsilon_i, \tag{3.1}$$

where ν is a positive integer, the ν roots of the polynomial $1 - \beta_1 z - \cdots - \beta_\nu z^\nu$ are all greater than 1 in the absolute value, and ε_i, $i = 0, \pm 1, \pm 2, \ldots$ are independent and identically distributed with mean zero and finite variance σ^2. Then $\{Z_i\}$ is an autoregressive process of order ν. By Theorem 3.1.1 of Brockwell and Davis (1987), there is a sequence a_u, $u = 0, \pm 1, \pm 2, \ldots$ of real numbers such that $\sum |a_u| < \infty$ and $Z_i = \sum_u a_{i-u}\varepsilon_u$ for $i = 0, \pm 1, \pm 2, \ldots$. Thus Condition 5 is automatically satisfied.

Let C be a compact subset of $\mathbb{R}^d$ having a nonempty interior (for example, a d-dimensional rectangle), let $\mathbf{X}_i$, $i = 0, \pm 1, \pm 2, \ldots$ be a C-valued stationary time series that is independent of ε_i, $i = 0, \pm 1, \pm 2, \ldots$, and set

$$Y_i = \theta(\mathbf{X}_i) + Z_i, \quad i = 0, \pm 1, \pm 2, \ldots. \tag{3.2}$$

As an example, X_i and Y_i could be bimonthly compliance and cholesterol measurements, respectively; here compliance is the percent of medication consumed by the patient between consecutive visits to the physician's office.

In the present context, we set $U = C$ in Conditions 1–3 and Theorem 1. Then Theorems 1–3 are valid in this context, as can be seen by examining their proofs.

Consider the ν-dimensional column vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_\nu)^t$. Given $b_1, \ldots, b_\nu \in \mathbb{R}$, set $\mathbf{b} = (b_1, \ldots, b_\nu)^t$. Then $\boldsymbol{\beta}$ is the unique value of $\mathbf{b}$ that minimizes $E[(Z_i - b_1 Z_{i-1} - \cdots - b_\nu Z_{i-\nu})^2]$. Consequently, $\beta_1, \ldots, \beta_\nu$ satisfy the system of normal equations

$$E[Z_{i-h}(\beta_1 Z_{i-1} + \cdots + \beta_\nu Z_{i-\nu})] = E(Z_{i-h} Z_i), \qquad h = 1, \ldots, \nu.$$

Set $\gamma_h = E(Z_i Z_{i+h})$. Then the system of normal equations can be written as

$$\sum_{l=1}^{\nu} \beta_l \gamma_{|h-l|} = \gamma_h, \qquad h = 1, \ldots, \nu.$$

Observe that the $\nu \times \nu$ matrix $\boldsymbol{\Gamma} = (\gamma_{|h-l|})$ is the variance-covariance matrix of $Z_1, \ldots, Z_\nu$. Set $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_\nu)^t$. Then the system of normal equations for $\boldsymbol{\beta}$ can be written in matrix form as $\boldsymbol{\Gamma}\boldsymbol{\beta} = \boldsymbol{\gamma}$, whose unique solution is given by $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}$.

Set $\hat{Z}_i = Y_i - \hat{\theta}_n(\mathbf{X}_i) = Z_i - (\hat{\theta}_n(\mathbf{X}_i) - \theta(\mathbf{X}_i))$ for all i and

$$\hat{\gamma}_{nh} = \frac{1}{n-h} \sum_{i=1}^{n-h} \hat{Z}_i \hat{Z}_{i+h}, \qquad h = 0, 1, \ldots, \nu.$$

Consider the estimates $\hat{\boldsymbol{\Gamma}}_n = (\hat{\gamma}_{n,|h-l|})$ and $\hat{\boldsymbol{\gamma}}_n = (\hat{\gamma}_{n1}, \ldots, \hat{\gamma}_{n\nu})^t$ of $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$, respectively, and consider the estimate $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\Gamma}}_n^{-1} \hat{\boldsymbol{\gamma}}_n$ of $\boldsymbol{\beta}$. (Given sequences $\{b_n\}$ and $\{c_n\}$, let $b_n \ll c_n$ mean that $b_n / c_n \to 0$ as $n \to \infty$.)

Theorem 4 *Suppose $k > d/2$, Conditions* 1–5 *and* (3.1) *and* (3.2) *hold, and that $n^{-1/2d} \ll \delta_n \ll n^{-1/4k}$. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2 \boldsymbol{\Gamma}^{-1})$.*

The above result on parametric inference for the error process provides an interesting justification for using local polynomials of degree $k - 1 > d/2 - 1$ to estimate $\theta(\cdot)$. Note that the condition on δ_n in this result is quite broad; in particular, it is satisfied if $\delta_n \sim n^{-1/(2k+d)}$

or $\delta_n \sim (n^{-1}\log n)^{1/(2k+d)}$. The proof of the result will be given in Section 4.4. A special case of the result was presented in Truong (1991), where $d = 1$ and local averages are used to estimate $\theta(\cdot)$. When $d = 2$ or $d = 3$, local-linear estimates of $\theta(\cdot)$ can be used, but when $d \geq 4$, it is necessary to use local polynomials of higher degree. Moreover, the achievability of the $n^{-1/2}$ rate of convergence for estimation of $\boldsymbol{\beta}$ presumably requires Condition 1 or some other suitable smoothness condition on $\theta(\cdot)$. Theorem 4 raises the interesting issue of determining the order ν from the observed data. Simulation results have shown that the parametric estimates are insensitive to δ_n, which is compatible with Theorem 4. In these simulations, AIC has been quite consistent in selecting the order ν and this selection also is insensitive to the value of δ_n.

4 Proofs

Given nonnegative integers $\alpha_1, \ldots, \alpha_d$, set $\alpha = (\alpha_1, \ldots, \alpha_d)$, $[\alpha] = \alpha_1 + \cdots + \alpha_d$, $\alpha! = (\alpha_1)! \cdots (\alpha_d)!$ and $\mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Also, set $\mathcal{A} = \{\alpha : [\alpha] < k\}$. Given $\mathbf{x} \in C$, let $\mathbf{Y}_n(\mathbf{x})$, $\mathbf{X}_n(\mathbf{x})$ and $\mathbf{A}_n(\mathbf{x})$ be defined as follows: $\mathbf{Y}_n(\mathbf{x}) = (Y_{ni}(\mathbf{x}))$ is the n-dimensional column vector given by $Y_{ni}(\mathbf{x}) = Y_i$ if $i \in I_n(\mathbf{x})$ and $Y_{ni}(\mathbf{x}) = 0$ otherwise; $\mathbf{X}_n(\mathbf{x}) = (X_{ni\alpha}(\mathbf{x}))$ is the $n \times \#(\mathcal{A})$ matrix given by $X_{ni\alpha}(\mathbf{x}) = (\mathbf{X}_i - \mathbf{x})^\alpha / \delta_n^{[\alpha]}$ if $i \in I_n(\mathbf{x})$ and $\alpha \in \mathcal{A}$ and $X_{ni\alpha}(\mathbf{x}) = 0$ otherwise; $\mathbf{A}_n(\mathbf{x}) = (A_{n\alpha\beta}(\mathbf{x})) = \mathbf{X}_n^t(\mathbf{x})\mathbf{X}_n(\mathbf{x})$, which is a $\#(\mathcal{A}) \times \#(\mathcal{A})$ matrix. Let $\hat{P}_{n,\mathbf{x}}(\mathbf{u})$ denote the polynomial of degree $k-1$, in terms of $(\mathbf{u} - \mathbf{x})^\alpha / \delta_n^{[\alpha]}$, minimizing

$$\sum_{I_n(\mathbf{x})} [Y_i - \hat{P}_{n,\mathbf{x}}(\mathbf{X}_i)]^2.$$

We have that

$$\hat{P}_{n,\mathbf{x}}(\mathbf{u}) = \sum_{[\alpha]<k} \hat{b}_{n\alpha}(\mathbf{x}) \frac{(\mathbf{u} - \mathbf{x})^\alpha}{\delta_n^{[\alpha]}},$$

where

$$\hat{b}_{n\alpha}(\mathbf{x}) = (\mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x})\mathbf{Y}_n(\mathbf{x}))_\alpha.$$

Let $\mathbf{Q} = (1, 0, \ldots, 0)^t$ denote the $\#(\mathcal{A})$-dimensional vector so that

$$\hat{\theta}_n(\mathbf{x}) = \hat{P}_{n,\mathbf{x}}(\mathbf{x}) = \mathbf{Q}^t \mathbf{A}_n^{-1}(\mathbf{x}) \mathbf{X}_n^t(\mathbf{x}) \mathbf{Y}_n(\mathbf{x}).$$

Let $\theta_k(\cdot; \mathbf{x})$ denote the $(k-1)$th-degree Taylor polynomial of $\theta(\cdot)$ about $\mathbf{x}$. That is,

$$\theta_k(\cdot; \mathbf{x}) = \sum_{[\alpha]<k} \frac{D^\alpha \theta(\mathbf{x})}{\alpha!} (\cdot - \mathbf{x})^\alpha, \quad \mathbf{x} \in C;$$

where

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Define the n-dimensional column vectors $\mathbf{T}_n(\mathbf{x}) = (T_{ni}(\mathbf{x}))$ and $\mathbf{T}_{kn}(\mathbf{x}) = (T_{kni}(\mathbf{x}))$ by

$$\begin{aligned} T_{ni}(\mathbf{x}) &= \theta(\mathbf{X}_i), \qquad i \in I_n(\mathbf{x}), \\ T_{kni}(\mathbf{x}) &= \theta_k(\mathbf{X}_i; \mathbf{x}), \qquad i \in I_n(\mathbf{x}), \end{aligned}$$

and $T_{ni}(\mathbf{x}) = T_{kni}(\mathbf{x}) = 0$, otherwise. Then

$$T_{kni}(\mathbf{x}) = \sum_{[\alpha]<k} \frac{(\mathbf{X}_i - \mathbf{x})^\alpha}{\delta_n^{[\alpha]}} \frac{\delta_n^{[\alpha]} D^\alpha \theta(\mathbf{x})}{\alpha!}, \qquad i \in I_n(\mathbf{x}),$$

so

$$(\mathbf{A}_n^{-1}(\mathbf{x}) \mathbf{X}_n^t(\mathbf{x}) \mathbf{T}_{kn}(\mathbf{x}))_\alpha = \frac{\delta_n^{[\alpha]} D^\alpha \theta(\mathbf{x})}{\alpha!}$$

and

$$\theta(\mathbf{x}) = \mathbf{Q}^t \mathbf{A}_n^{-1}(\mathbf{x}) \mathbf{X}_n^t(\mathbf{x}) \mathbf{T}_{kn}(\mathbf{x}). \tag{4.1}$$

Consequently,

$$\begin{aligned} \hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x}) &= \mathbf{Q}^t \mathbf{A}_n^{-1}(\mathbf{x}) \mathbf{X}_n^t(\mathbf{x}) [\mathbf{Y}_n(\mathbf{x}) - \mathbf{T}_n(\mathbf{x})] \\ &\quad + \mathbf{Q}^t \mathbf{A}_n^{-1}(\mathbf{x}) \mathbf{X}_n^t(\mathbf{x}) [\mathbf{T}_n(\mathbf{x}) - \mathbf{T}_{kn}(\mathbf{x})], \quad \mathbf{x} \in C. \end{aligned} \tag{4.2}$$

To show that the solution to the least squares problem:

$$\sum_{I_n(\mathbf{x})} [Y_i - \hat{P}_{n,\mathbf{x}}(\mathbf{X}_i)]^2$$

does exist, we recall that $N_n(\mathbf{x}) = \#\{i : 1 \le i \le n \text{ and } \|\mathbf{X}_i - \mathbf{x}\| \le \delta_n\}$, $\mathbf{x} \in C$. Suppose that $1/n\delta_n^d = o(n^{-\epsilon})$ for some $\epsilon > 0$. Then, according to Conditions 2–4 and the argument of Lemma 7 in Truong and Stone (1992), there are positive constants c_1 and c_2 such that

$$\lim_n P(\Omega_n) = 1, \tag{4.3}$$

where $\Omega_n = \Big\{c_1 n\delta_n^d \le N_n(\mathbf{x}) \le c_2 n\delta_n^d \text{ for } \mathbf{x} \in C\Big\}$. The existence of $\hat{P}_{n,\mathbf{x}}(\mathbf{u})$ is given in the following result.

Lemma 1 *Under Conditions 2–4, there is a positive constant c_3 such that*

$$\lim_n P\Big(N_n(\mathbf{x})(\mathbf{A}_n^{-1}(\mathbf{x}))_{\alpha\beta} \le c_3 \quad \textit{for } \mathbf{x} \in C \textit{ and } \alpha, \beta \in \mathcal{A}\Big) = 1.$$

Proof We may assume that $C = [-1/2, 1/2]^d$. Let λ be a positive integer so that $L_n = n^\lambda$. Let W_n be the collection of $(2L_n + 1)^d$ points in $[-1/2, 1/2]^d$ each of whose coordinates is of the form $j/(2L_n)$ for some integer j such that $|j| \le L_n$. Then $[-1/2, 1/2]^d$ can be written as the union of $(2L_n)^d$ subcubes, each having length $1/2L_n$ and all of its vertices in W_n. For each $\mathbf{x} \in [-1/2, 1/2]^d$ there is a subcube $Q_{\mathbf{w}}$ with center $\mathbf{w}$ such that $\mathbf{x} \in Q_{\mathbf{w}}$. Let C_n denote the collection of centers of these subcubes.

Define the $\#(\mathcal{A}) \times \#(\mathcal{A})$ matrix $\mathbf{A} = (A_{\alpha\beta})$ by

$$A_{\alpha\beta} = \frac{\int_{|\mathbf{u}|\le 1} \mathbf{u}^\alpha \mathbf{u}^\beta \, d\mathbf{u}}{\int_{|\mathbf{u}|\le 1} d\mathbf{u}}.$$

Then $\det(\mathbf{A}) > 0$ (see pp. 1354–1355 of Stone, 1980). Given $\epsilon > 0$, it follows from the argument of Lemma 9 in Truong and Stone (1992) that

$$\lim_n P\Big(|N_n^{-1}(\mathbf{w})A_{n\alpha\beta}(\mathbf{w}) - A_{\alpha\beta}| \le \epsilon \quad \text{for all } \mathbf{w} \in C_n \text{ and } \alpha, \beta \in \mathcal{A}\Big) = 1.$$

According to (2.13) of Truong and Stone (1992) and by expanding $(\mathbf{X}_i - \mathbf{x})^\alpha$ in terms of $\mathbf{X}_i - \mathbf{w}$ and $\mathbf{w} - \mathbf{x}$, where $\mathbf{x} \in Q_{\mathbf{w}}$,

$$\lim_n P\Big(|N_n^{-1}(\mathbf{x})A_{n\alpha\beta}(\mathbf{x}) - A_{\alpha\beta}| \le \epsilon \quad \text{for all } \mathbf{x} \in C \text{ and } \alpha, \beta \in \mathcal{A}\Big) = 1, \quad \epsilon > 0.$$

Consequently,

$$\lim_n P\Big(\det(N_n^{-1}(\mathbf{x})\mathbf{A}_n(\mathbf{x})) > 0 \quad \text{for all } \mathbf{x} \in C\Big) = 1$$

and hence

$$\lim_n P\Big(|N_n(\mathbf{x})(\mathbf{A}_n^{-1}(\mathbf{x}))_{\alpha\beta} - (\mathbf{A}^{-1})_{\alpha\beta}| \le \epsilon \quad \text{for } \mathbf{x} \in C \text{ and } \alpha, \beta \in \mathcal{A}\Big) = 1, \quad \epsilon > 0. \quad \square$$

Set

$$\mathbf{M}_n(\mathbf{x}) = (M_i(\mathbf{x})) = \mathbf{Q}^t \mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x}), \qquad \mathbf{x} \in C.$$

Then

$$\hat{\theta}_n(\mathbf{x}) = \sum_i M_i(\mathbf{x}) Y_i. \tag{4.4}$$

Note that $M_i(\mathbf{x}) = 0$ for $i \notin I_n(\mathbf{x})$ and that

$$|\mathbf{M}_n(\mathbf{x})|^2 = O(N_n^{-1}(\mathbf{x})\mathbf{Q}^t(\mathbf{x})[N_n(\mathbf{x})\mathbf{A}_n^{-1}(\mathbf{x})]\mathbf{Q}_n(\mathbf{x})).$$

By Lemma 1 and (4.3), there is a positive constant c_4 such that

$$\lim_n P(\Psi_n) = 1, \tag{4.5}$$

where $\Psi_n = \{|\mathbf{M}_n(\mathbf{x})|^2 \le c_4 n^{-1}\delta_n^{-d} \quad \text{and} \quad \max_i |M_i(\mathbf{x})| \le c_4 n^{-1}\delta_n^{-d}, \quad \mathbf{x} \in C\}$.

In a number of results below, we need to condition on $\mathbf{X}_1, \ldots, \mathbf{X}_n$. In order to simplify the notation, we set $P_n(\cdot) = P(\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $E_n(\cdot) = E(\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Also, for positive numbers b_n and random variables V_n, $n \ge 1$, we write

$$E_n(V_n) \le b_n \quad \text{on } \Psi_n$$

to mean that

$$P\Big(\{E_n(V_n) > b_n\} \cap \Psi_n\Big) = 0.$$

Similarly, for events A_n,

$$P_n(A_n) \le b_n \quad \text{on } \Psi_n$$

is equivalent to

$$P\Big(\{P_n(A_n) > b_n\} \cap \Psi_n\Big) = 0.$$

Remark. The following argument will be used repeatedly in the proofs. Suppose $E_n(|V_n|^2) \le b_n^2$ on Ψ_n. By Markov's inequality

$$
\begin{aligned}
P\Big(\{|V_n| \ge cb_n\} \cap \Psi_n\Big) &= \int_{\Psi_n \cap \{E_n(|V_n|^2) \le b_n^2\}} P_n(|V_n| \ge cb_n)dP \\
&\le \frac{1}{c^2} P\Big(\Psi_n \cap \{E_n(|V_n|^2) \le b_n^2\}\Big).
\end{aligned}
$$

Lemma 2 *Suppose Conditions* 2–5 *hold. Then there is a positive constant* c_5 *such that, for* $\mathbf{x} \in C$,

$$
E_n\Big[\Big(\textstyle\sum_i M_i(\mathbf{x})[Y_i - \theta(\mathbf{X}_i)]\Big)^2\Big] \le c_5 n^{-1}\delta_n^{-d} \quad on\ \Omega_n \cap \Psi_n.
$$

Proof Set $U_i = Y_i - \theta(\mathbf{X}_i) = \sum_u a_{i-u}\varepsilon_u$ and $||\mathbf{a}||^2 = \sum a_u^2$. By (4.3), (4.5), and the assumptions on the ε_u's,

$$
\begin{aligned}
E_n\Big[\Big(\textstyle\sum_i M_i(\mathbf{x})[Y_i - \theta(\mathbf{X}_i)]\Big)^2\Big] &= \textstyle\sum_i E_n[M_i(\mathbf{x})U_i]^2 + 2\sum_i\sum_j E_n(M_i(\mathbf{x})U_i M_{i+j}(\mathbf{x})U_{i+j}) \\
&= c_4 n^{-1}\delta_n^{-d}\sigma^2||\mathbf{a}||^2 + 2\sigma^2 \textstyle\sum_i\sum_j M_i(\mathbf{x})M_{i+j}(\mathbf{x})\sum_u a_{i-u}a_{i+j-u} \\
&\le c_4 n^{-1}\delta_n^{-d}\sigma^2||\mathbf{a}||^2 + 2\sigma^2||\mathbf{a}||^2 \max_j |M_j(\mathbf{x})| \textstyle\sum_i |M_i(\mathbf{x})| \\
&= O(n^{-1}\delta_n^{-d}) \quad \text{on } \Omega_n \cap \Psi_n. \quad \square
\end{aligned}
$$

4.1 Proof of Theorem 1

According to Condition 1, there is a positive constant c_6 such that

$$
|T_{ni}(\mathbf{x}) - T_{kni}(\mathbf{x})| = |\theta(\mathbf{X}_i) - \theta_k(\mathbf{X}_i;\mathbf{x})| \le c_6\delta_n^k, \quad i \in I_n(\mathbf{x}), \quad \mathbf{x} \in C. \tag{4.6}
$$

By Lemma 1, there is a positive constant c_7 such that

$$
\mathbf{Q}^t\mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x})[\mathbf{T}_n(\mathbf{x}) - \mathbf{T}_{kn}(\mathbf{x})] \le c_7\delta_n^k \quad \text{on } \Omega_n, \quad \mathbf{x} \in C. \tag{4.7}
$$

By Markov's inequality and Lemma 2,

$$
\Big|\mathbf{Q}^t\mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x})[\mathbf{Y}_n(\mathbf{x}) - \mathbf{T}_n(\mathbf{x})]\Big| = O_p(\delta_n^k). \tag{4.8}
$$

The desired result follows from (4.7) and (4.8). $\square$

4.2 Proof of Theorem 2

By Lemma 2, there is a positive constant c_8 such that

$$E_n[\, Z_n^2(\mathbf{x})\,] \le c_8 n^{-1}\delta_n^{-d} \quad \text{on } \Omega_n \cap \Psi_n, \quad \mathbf{x} \in C,$$

where $Z_n(\mathbf{x}) = \mathbf{Q}^t \mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x})[\mathbf{Y}_n(\mathbf{x}) - \mathbf{T}_n(\mathbf{x})]$. Hence

$$E_n\left(\int_C |Z_n(\mathbf{x})|^2\, d\mathbf{x}\right) = \int_C E_n[\,|Z_n(\mathbf{x})|^2\,]\, d\mathbf{x} = O(n^{-1}\delta_n^{-d}) \quad \text{on } \Omega_n \cap \Psi_n.$$

Consequently, by (4.3), (4.5) and $\delta_n \sim n^{-1/(2k+d)}$

$$\left(\int_C |Z_n(\mathbf{x})|^2\, d\mathbf{x}\right)^{\frac{1}{2}} = O_p(\delta_n^k). \tag{4.9}$$

The conclusion of Theorem 2 now follows from (4.7) and (4.9). □

Lemma 3 *Suppose Conditions* 2–6 *hold and that* $1/n\delta_n^d = o(n^{-\epsilon})$ *for some* $\epsilon > 0$. *Then there is a positive constant* c_9 *such that, for* $b > 0$, *if* n *is sufficiently large, then*

$$P_n\left(|\textstyle\sum_i M_i(\mathbf{x})[Y_i - \theta(\mathbf{X}_i)]| \ge b\sqrt{(\log n)/n\delta_n^d}\right) \le 2n^{-c_9 b^2} \quad \text{on } \Omega_n \cap \Psi_n.$$

Proof Set $C_q = |\mathrm{cum}_q(\varepsilon_0)|$ and $A = \sum |a_u|$. Also set $U_i = Y_i - \theta(\mathbf{X}_i)$, $M_i^*(\mathbf{x}) = M_i(\mathbf{x})/\max_j |M_j(\mathbf{x})|$, $\mathrm{cum}_{n,q}(\cdot) = \mathrm{cum}_q(\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $\mathrm{var}_n(\cdot) = \mathrm{var}(\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Then

$$\mathrm{cum}_{n,1}\left(\textstyle\sum_i M_i^*(\mathbf{x})U_i\right) = E_n\left(\textstyle\sum_i M_i^*(\mathbf{x})U_i\right) = 0$$

and

$$\mathrm{cum}_{n,2}\left(\textstyle\sum_i M_i^*(\mathbf{x})U_i\right) = \mathrm{var}_n\left(\textstyle\sum_i M_i^*(\mathbf{x})U_i\right).$$

Let $q \ge 3$. By Theorem 2.3.1(iv) of Brillinger (1981), the absolute value of the qth cumulant of $\sum_i M_i^*(\mathbf{x})U_i$ has the form

$$\begin{aligned} &\left|\sum_{i_1,\ldots,i_q} \mathrm{cum}_{n,q}\left(M_{i_1}^*(\mathbf{x})U_{i_1}, \ldots, M_{i_q}^*(\mathbf{x})U_{i_q}\right)\right| \\ &= \left|\sum_{i_1,\ldots,i_q} \sum_{u_1} M_{i_1}^*(\mathbf{x})a_{i_1-u_1} \cdots \sum_{u_q} M_{i_q}^*(\mathbf{x})a_{i_q-u_q}\mathrm{cum}(\varepsilon_{u_1}, \ldots, \varepsilon_{u_q})\right| \end{aligned}$$

$$
\begin{aligned}
&= O\Big(C_q \sum_{i_1 \in I_n(\mathbf{x})} \cdots \sum_{i_q \in I_n(\mathbf{x})} \sum_u |a_{i_1-u}| \cdots |a_{i_q-u}|\Big) \\
&= O\Big(C_q \sum_{i_1 \in I_n(\mathbf{x})} \sum_u |a_{i_1-u}| \sum_{i_2 \in I_n(\mathbf{x})} |a_{i_2-u}| \cdots \sum_{i_q \in I_n(\mathbf{x})} |a_{i_q-u}|\Big) \\
&= O\Big(C_q N_n(\mathbf{x})(\sum_j |a_j|)^q\Big) = C_q A^q O(n\delta_n^d) \quad \text{on } \Omega_n.
\end{aligned}
$$

Thus, by Taylor's expansion, for ζ positive and sufficiently small,

$$
\left|\log E_n[\exp(\zeta \textstyle\sum_i M_i^*(\mathbf{x})U_i)] - \dfrac{\zeta^2}{2}\mathrm{var}_n(\textstyle\sum_i M_i^*(\mathbf{x})U_i)\right| \le c_2 n\delta_n^d \sum_{q\ge 3} A^q C_q \zeta^q/q! \quad \text{on } \Omega_n.
$$

Let b_n, $n \ge 1$, be a sequence of positive numbers. According to Markov's inequality,

$$
\begin{aligned}
&P_n\Big(\textstyle\sum_i M_i(\mathbf{x})U_i \ge b_n\Big) \\
&\quad \le \exp\Big(-sb_n\Big) E_n\Big(\exp(s\textstyle\sum_i M_i(\mathbf{x})U_i)\Big) \\
&\quad \le \exp\Big(-sb_n\Big) E_n\Big(\exp(s \max_j M_j(\mathbf{x}) \textstyle\sum_i M_i^*(\mathbf{x})U_i)\Big) \\
&\quad \le \exp\Big(-sb_n\Big) \exp\left(\frac{s^2}{2}\mathrm{var}_n(\textstyle\sum_i M_i(\mathbf{x})U_i)\right) \exp\left(c_2 n\delta_n^d \displaystyle\sum_{q\ge 3} A^q C_q \zeta^q/q!\right) \quad \text{on } \Omega_n,
\end{aligned}
$$

provided that $\zeta = s\max_j |M_j(\mathbf{x})|$ is positive and sufficiently small. By Lemma 2,

$$
\mathrm{var}_n\Big(\textstyle\sum_i M_i(\mathbf{x})U_i\Big) \le c_5 n^{-1}\delta_n^{-d} \quad \text{on } \Omega_n \cap \Psi_n.
$$

By Condition 6, for ζ positive and sufficiently small,

$$
c_2 n\delta_n^d \sum_{q\ge 3} A^q C_q \zeta^q/q! = c_2 n\delta_n^d \zeta^2 \sum_{q\ge 3} A^q C_q \zeta^{q-2}/q! = O\Big(n\delta_n^d \zeta^2\Big).
$$

Thus there is a positive constant a such that, for $sn^{-1}\delta_n^{-d}$ positive and sufficiently small,

$$
P_n\Big(\textstyle\sum_i M_i(\mathbf{x})U_i \ge b_n\Big) \le \exp\Big(-sb_n + as^2 n^{-1}\delta_n^{-d}\Big) \quad \text{on } \Omega_n \cap \Psi_n.
$$

In particular, by choosing $s = (b_n/2a)n\delta_n^d$, we see that if b_n is positive and sufficiently small, then

$$
P_n\Big(\textstyle\sum_i M_i(\mathbf{x})U_i \ge b_n\Big) \le \exp\Big(-(b_n^2/4a)n\delta_n^d\Big) \quad \text{on } \Omega_n \cap \Psi_n.
$$

Similarly,

$$
P_n\Big(\textstyle\sum_i M_i(\mathbf{x})U_i \le -b_n\Big) \le \exp\Big(-(b_n^2/4a)n\delta_n^d\Big) \quad \text{on } \Omega_n \cap \Psi_n,
$$

so

$$P_n\Big(|\sum_i M_i(\mathbf{x})U_i| \geq b_n\Big) \leq \exp\Big(-(b_n^2/4a)n\delta_n^d\Big) \quad \text{on } \Omega_n \cap \Psi_n.$$

By choosing $b_n = b\sqrt{(\log n)/n\delta_n^d}$ for n sufficiently large and $c_9 = 1/4a$, we get the desired result. □

4.3 Proof of Theorem 3

Set $T_n(\mathbf{x}) = \mathbf{Q}^t\mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x})\mathbf{T}_n(\mathbf{x})$. Then

$$T_n(\mathbf{x}) - \theta(\mathbf{x}) = \mathbf{Q}^t\mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x})[\mathbf{T}_n(\mathbf{x}) - \mathbf{T}_{nk}(\mathbf{x})].$$

Thus, by (4.3) and (4.7),

$$\lim_n P\Big(\sup_{\mathbf{x}\in C} |T_n(\mathbf{x}) - \theta(\mathbf{x})| \geq c_7\delta_n^k\Big) = 0. \tag{4.10}$$

Set $\mathbf{U}_n(\mathbf{x}) = (U_{ni}(\mathbf{x}))$, where $U_{ni}(\mathbf{x}) = U_i = Y_i - \theta(\mathbf{X}_i)$ for $i \in I_n(\mathbf{x})$ and $U_{ni}(\mathbf{x}) = 0$ otherwise. Then $\mathbf{U}_n(\mathbf{x}) = \mathbf{Y}_n(\mathbf{x}) - \mathbf{T}_n(\mathbf{x})$ and

$$\hat{T}_n(\mathbf{x}) - T_n(\mathbf{x}) = \mathbf{Q}^t\mathbf{A}_n^{-1}(\mathbf{x})\mathbf{X}_n^t(\mathbf{x})\mathbf{U}_n(\mathbf{x}) = \mathbf{M}_n(\mathbf{x})\mathbf{U}_n(\mathbf{x}). \tag{4.11}$$

Let C_0 be a compact subset of U containing C in its interior. We can assume that if $i \in I_n(\mathbf{x})$ for some $\mathbf{x} \in C$, then $\mathbf{X}_i \in C_0$. Choose $\gamma > 0$. It follows from Conditions 5 and 6 by applying Markov's inequality to a suitable analog of Lemma 3 that

$$\max\{|Y_i - \theta(\mathbf{X}_i)| : \mathbf{X}_i \in C_0\} = O_p(n^\gamma)$$

and hence that

$$\sup_{\mathbf{x}\in C}\max_i |U_{ni}(\mathbf{x})| = O_p(n^\gamma). \tag{4.12}$$

Consider the symmetric difference

$$I_n(\mathbf{x})\triangle I_n(\mathbf{w}) = (I_n(\mathbf{x}) \setminus I_n(\mathbf{w})) \cup (I_n(\mathbf{w}) \setminus I_n(\mathbf{x})), \quad \mathbf{w} \in C_n \text{ and } \mathbf{x} \in Q_{\mathbf{w}},$$

where C_n is as in the proof of Lemma 1. It follows from Condition 2 and 4 (see the proof of (2.13) of Truong and Stone (1992)) that, for λ sufficiently large,

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}} \#(I_n(\mathbf{x})\triangle I_n(\mathbf{w})) = O_p(n^\gamma). \tag{4.13}$$

We conclude from (4.5), (4.12) and (4.13) that, for λ sufficiently large,

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}|\mathbf{M}_n(\mathbf{x})[\mathbf{U}_n(\mathbf{x})-\mathbf{U}_n(\mathbf{w})]|=O_p\Big(n^{2\gamma-1}\delta_n^{-d}\Big).$$

Consequently,

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}|\mathbf{M}_n(\mathbf{x})[\mathbf{U}_n(\mathbf{x})-\mathbf{U}_n(\mathbf{w})]|=o_p\Big(\sqrt{1/n\delta_n^d}\Big). \tag{4.14}$$

Observe next that

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}\max_{i\in I_n(\mathbf{x})\triangle I_n(\mathbf{w})}\max_{\alpha\in\mathcal{A}}|X_{ni\alpha}(\mathbf{x})-X_{ni\alpha}(\mathbf{w})|=O\Big(n^{-\lambda}\delta_n^{-1}\Big). \tag{4.15}$$

We conclude from Lemma 1, (4.3), (4.12), (4.13) and (4.15) that, for λ sufficiently large,

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}|\mathbf{Q}^t\mathbf{A}_n^{-1}(\mathbf{x})[\mathbf{X}_n^t(\mathbf{x})-\mathbf{X}_n^t(\mathbf{w})]\mathbf{U}_n(\mathbf{w})|=o_p\Big(\sqrt{1/n\delta_n^d}\Big). \tag{4.16}$$

Write $\mathbf{A}_n(\mathbf{x})=(A_{n\alpha\beta}(\mathbf{x}))$. It follows from (4.13) and (4.15) that, for λ sufficiently large,

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}\max_{\alpha,\beta\in\mathcal{A}}|A_{n\alpha\beta}(\mathbf{x})-A_{n\alpha\beta}(\mathbf{w})|=o_p\Big(n^{\gamma}\Big). \tag{4.17}$$

We conclude from Lemma 1, (4.3) and (4.17) that, for λ sufficiently large,

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}\max_{\alpha,\beta\in\mathcal{A}}|(A_n^{-1}(\mathbf{x}))_{\alpha\beta}-(A_n^{-1}(\mathbf{w}))_{\alpha\beta}|=o_p\Big(n^{\gamma}(n\delta_n^d)^{-2}\Big)$$

and hence that

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}|\mathbf{Q}^t[\mathbf{A}_n^{-1}(\mathbf{x})-\mathbf{A}_n^{-1}(\mathbf{w})]\mathbf{X}_n^t(\mathbf{w})\mathbf{U}_n(\mathbf{w})|=o_p\Big(\sqrt{1/n\delta_n^d}\Big). \tag{4.18}$$

It follows from (4.11), (4.14), (4.16) and (4.18) that

$$\max_{\mathbf{w}\in C_n}\sup_{\mathbf{x}\in Q_{\mathbf{w}}}|\hat{T}_n(\mathbf{x})-T_n(\mathbf{x})-\hat{T}_n(\mathbf{w})+T_n(\mathbf{w})|=o_p\Big(\sqrt{1/n\delta_n^d}\Big). \tag{4.19}$$

For each fixed $\lambda>0$, it follows from Lemma 3 that there is a positive constant c_{10} such that

$$\lim_n P\Big(\max_{\mathbf{w}\in C_n}|\hat{T}_n(\mathbf{w})-T_n(\mathbf{w})|\ge c_{10}\sqrt{(\log n)/n\delta_n^d}\Big)=0. \tag{4.20}$$

We conclude from (4.19) and (4.20) that there is a positive constant c_{11} such that

$$\lim_n P\Big(\sup_{\mathbf{x}\in C}|\hat{T}_n(\mathbf{x})-T_n(\mathbf{x})|\ge c_{11}\sqrt{(\log n)/n\delta_n^d}\Big)=0. \tag{4.21}$$

The conclusion of Theorem 3 follows from (4.10) and (4.21). □

4.4 Proof of Theorem 4

The main idea behind the proof of Theorem 4 is to approximate the sample covariance structure constructed based on $\{\hat{Z}_i\}$ by the corresponding one using Z_i, and show that this approximation has an error rate no more than $n^{-1/2}$.

Observe that, by (3.2), (4.1) and (4.4),

$$\begin{aligned}\hat{Z}_i &= Y_i - \hat{\theta}_n(\mathbf{X}_i)\\ &= \theta(\mathbf{X}_i) + Z_i - \sum_j M_j(\mathbf{X}_i)\theta(\mathbf{X}_j) - \sum_j M_j(\mathbf{X}_i)Z_j\\ &= Z_i - \sum_j M_j(\mathbf{X}_i)Z_j - \sum_j M_j(\mathbf{X}_i)[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_i)],\end{aligned}$$

where $\theta_k(\cdot;\mathbf{x})$ is the Taylor polynomial of $\theta(\cdot)$ about $\mathbf{x}$. Thus

$$n^{-1}\sum_{i=1}^{n-h}\hat{Z}_i\hat{Z}_{i+h} - n^{-1}\sum_{i=1}^{n-h}Z_iZ_{i+h} = B_{1,h} + B_{2,h} + \cdots + B_{5,h}, \quad h = 0,1,\ldots,\nu, \tag{4.22}$$

where

$$\begin{aligned}B_{1,h} &= n^{-1}\sum_{i=1}^{n-h}\Big\{\sum_j M_j(\mathbf{X}_i)[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_i)]\Big\}\Big\{\sum_j M_j(\mathbf{X}_{i+h})[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_{i+h})]\Big\},\\ B_{2,h} &= n^{-1}\sum_{i=1}^{n-h}\Big\{\sum_j M_j(\mathbf{X}_i)[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_i)]\Big\}\Big\{\sum_j M_j(\mathbf{X}_{i+h})Z_j\Big\}\\ &\quad + n^{-1}\sum_{i=1}^{n-h}\Big\{\sum_j M_j(\mathbf{X}_{i+h})[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_{i+h})]\Big\}\Big\{\sum_j M_j(\mathbf{X}_i)Z_j\Big\}\\ B_{3,h} &= n^{-1}\sum_{i=1}^{n-h}\Big\{\sum_j M_j(\mathbf{X}_{i+h})Z_j\Big\}\Big\{\sum_j M_j(\mathbf{X}_i)Z_j\Big\}\\ B_{4,h} &= -n^{-1}\sum_{i=1}^{n-h}\Big(Z_i\sum_j M_j(\mathbf{X}_{i+h})Z_j + Z_{i+h}\sum_j M_j(\mathbf{X}_i)Z_j\Big)\\ B_{5,h} &= n^{-1}\sum_{i=1}^{n-h}Z_i\sum_j M_j(\mathbf{X}_{i+h})[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_{i+h})]\\ &\quad + n^{-1}\sum_{i=1}^{n-h}Z_{i+h}\sum_j M_j(\mathbf{X}_i)[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_i)].\end{aligned}$$

By (4.3), (4.5) and (4.6),

$$\max_{i,h}\sum_j M_j(\mathbf{X}_{i+h})(\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_{i+h})) = O_p(\delta_n^k) \qquad \text{on } \Psi_n. \tag{4.23}$$

Consequently,

$$B_{1,h} = O_p(\delta_n^{2k}), \quad h = 0, 1, \ldots, \nu. \tag{4.24}$$

By (4.23) and Lemma 2,

$$B_{2,h} = O_p\Big(\delta_n^k/\sqrt{n\delta_n^d}\Big), \quad h = 0, 1, \ldots, \nu. \tag{4.25}$$

By Lemma 2,

$$B_{3,h} = O_p\Big((n\delta_n^d)^{-1}\Big), \quad h = 0, 1, \ldots, \nu. \tag{4.26}$$

The last two terms $B_{4,h}$ and $B_{5,h}$ will be treated in the following lemmas respectively.

Lemma 4 *Suppose Conditions* 2–5 *and* (3.1) *and* (3.2) *hold, and that* $1/n\delta_n^d = o(n^{-\epsilon})$ *for some* $\epsilon > 0$. *Then*

$$B_{4,h} = O_p\Big(1/n\delta_n^d\Big), \quad h = 0, 1, \ldots, \nu.$$

Proof Since $Z_i = \sum_i a_{i-u}\varepsilon_u$ and $\{\varepsilon_u\}$ is a sequence of iid random variables with mean zero,

$$E\Big(Z_i Z_l\Big) = \sigma^2 \textstyle\sum_u a_{i-u} a_{l-u}$$

and

$$\begin{aligned} E\Big(Z_i Z_l Z_j Z_m\Big) &= \textstyle\sum_u \sum_v \sum_s \sum_t a_{i-u} a_{l-v} a_{j-s} a_{m-t} E\Big(\varepsilon_u \varepsilon_v \varepsilon_s \varepsilon_t\Big) \\ &= (\gamma_{i-l})(\gamma_{j-m}) + (\gamma_{i-j})(\gamma_{l-m}) + (\gamma_{i-m})(\gamma_{j-l}) \\ &\qquad + \Big(E(\varepsilon_0^4) - 3\sigma^4\Big) \textstyle\sum_u a_{i-u} a_{l-u} a_{j-u} a_{m-u}. \end{aligned}$$

Thus,

$$\begin{aligned} &E_n\left[\Big(n^{-1} \textstyle\sum_i Z_i \sum_j M_j(\mathbf{X}_{i+h}) Z_j\Big)^2\right] \\ &= n^{-2} \textstyle\sum_i \sum_j \sum_l \sum_m M_l(\mathbf{X}_{i+h}) M_m(\mathbf{X}_{j+h}) E(Z_i Z_l Z_j Z_m) \\ &= O_p\Big(n^{-4}\delta_n^{-2d} \textstyle\sum_i \sum_j \sum_l \sum_m (|\gamma_{i-l}||\gamma_{j-m}| + |\gamma_{i-j}||\gamma_{l-m}| + |\gamma_{i-m}||\gamma_{l-m}|) \\ &\qquad + n^{-4}\delta_n^{-2d} \textstyle\sum_i (\sum_u |a_u|)^4\Big) \quad \text{on } \Psi_n. \end{aligned}$$

Hence

$$E_n\left[\left(n^{-1}\sum_i Z_i \sum_j M_j(\mathbf{X}_{i+h})Z_j\right)^2\right] = O_p\left(n^{-2}\delta_n^{-2d}\right) \quad \text{on } \Psi_n.$$

Similarly,

$$E_n\left[\left(n^{-1}\sum_i Z_{i+h} \sum_j M_j(\mathbf{X}_i)Z_j\right)^2\right] = O_p\left(n^{-2}\delta_n^{-2d}\right) \quad \text{on } \Psi_n.$$

The desired result follows from (4.5). □

Lemma 5 *Suppose Conditions* 1–5 *and* (3.1) *and* (3.2) *hold, and that* $1/n\delta_n^d = o(n^{-\epsilon})$ *for some* $\epsilon > 0$. *Then*

$$B_{5,h} = O_p\left(\delta_n^k/\sqrt{n}\right), \quad h = 0, 1, \ldots, \nu.$$

Proof Write

$$n^{-1}\sum_i Z_i \sum_j M_j(\mathbf{X}_{i+h})[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_{i+h})] = n^{-1}\sum_i W_{nih} Z_i,$$

where

$$W_{nih} = \sum_j M_j(\mathbf{X}_{i+h})[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_{i+h})], \quad i = 0, \pm 1, \ldots.$$

By (4.23),

$$\max_{i,h} |W_{nih}| = O(\delta_n^k) \quad \text{on } \Psi_n.$$

Since

$$\sum_i \sum_j |E(Z_i Z_j)| \le \sigma^2 \sum_i \sum_j \left|\sum_u a_{i-u} a_{j-u}\right| = O\left(n(\sum |a_j|)^2\right) = O(n),$$

we conclude that

$$E_n\left[\left(n^{-1}\sum_i W_{nih} Z_i\right)^2\right] = n^{-2}\sum_i \sum_j W_{nih} W_{njh} E(Z_i Z_j) = O_p\left(n^{-1}\delta_n^{2k}\right) \quad \text{on } \Psi_n.$$

By (4.5) and Markov's inequality,

$$n^{-1}\sum_{i=1}^{n-h} Z_i \sum_j M_j(\mathbf{X}_{i+h})[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_{i+h})] = O_p\left(\delta_n^k/\sqrt{n}\right), \quad h = 0, 1, \ldots, \nu.$$

Similarly,

$$n^{-1}\sum_{i=1}^{n-h} Z_{i+h} \sum_j M_j(\mathbf{X}_i)[\theta(\mathbf{X}_j) - \theta_k(\mathbf{X}_j;\mathbf{X}_i)] = O_p\left(\delta_n^k/\sqrt{n}\right), \quad h = 0, 1, \ldots, \nu.$$

This completes the proof of Lemma 5. $\square$

According to (4.22)–(4.26), Lemmas 4 and 5,

$$n^{-1}\sum_{i=1}^{n-h}\hat{Z}_i\hat{Z}_{i+h}-n^{-1}\sum_{i=1}^{n-h}Z_iZ_{i+h}=O_p\Big(\delta_n^{2k}+\delta_n^k(n\delta_n^d)^{-1/2}+(n\delta_n^d)^{-1}+\delta_n^k/\sqrt{n}\Big),\quad h=0,1,\ldots,\nu. \tag{4.27}$$

We are now ready to prove Theorem 4. Set

$$\tilde{\gamma}_{nh}=\frac{1}{n-h}\sum_{i=1}^{n-h}Z_iZ_{i+h},\quad h=0,1,\ldots,\nu.$$

Consider the consistent estimates $\tilde{\mathbf{\Gamma}}_n=(\tilde{\gamma}_{n,|h-l|})$ and $\tilde{\boldsymbol{\gamma}}_n=(\tilde{\gamma}_{n1},\ldots,\tilde{\gamma}_{n\nu})^t$ of $\mathbf{\Gamma}$ and $\boldsymbol{\gamma}$, respectively, and consider the Yule-Walker estimate $\tilde{\boldsymbol{\beta}}_n=\tilde{\mathbf{\Gamma}}_n^{-1}\tilde{\boldsymbol{\gamma}}_n$ of $\boldsymbol{\beta}$. By Proposition 8.10.1 of Brockwell and Davis (1987),

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\xrightarrow{\mathcal{D}}N(0,\sigma^2\mathbf{\Gamma}^{-1})\quad\text{as } n\to\infty. \tag{4.28}$$

Since $\hat{\gamma}_{nh}=(n-h)^{-1}\sum_{i=1}^{n-h}\hat{Z}_i\hat{Z}_{i+h}$, it follows from (4.27) that

$$\hat{\gamma}_{nh}-\tilde{\gamma}_{nh}=O_p\Big(\delta_n^{2k}+\delta_n^k(n\delta_n^d)^{-1/2}+(n\delta_n^d)^{-1}+\delta_n^k/\sqrt{n}\Big). \tag{4.29}$$

Under the supposition that $n^{-1/2d}\ll\delta_n\ll n^{-1/4k}$, we conclude from (4.29) that $\hat{\gamma}_{nh}-\tilde{\gamma}_{nh}=o_p\big(1/\sqrt{n}\big)$ and hence that $\hat{\boldsymbol{\beta}}_n-\tilde{\boldsymbol{\beta}}_n=o_p\big(1/\sqrt{n}\big)$. The desired result now follows from (4.28). (The quantity $\delta_n^{2k}+\delta_n^k(n\delta_n^d)^{-1/2}+(n\delta_n^d)^{-1}+\delta_n^k/\sqrt{n}$ in the right side of (4.29) has the fastest possible rate of convergence when $\delta_n\sim n^{-1/(2k+d)}$.) $\square$

REFERENCES

1. Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika* **77**, 669–687.

2. Brillinger, D. R. (1981). *Time series: Data analysis and theory.* Holden-Day, San Francisco.

3. Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods.* Springer, New York.

4. Cochran, D. and G. H. Orcutt (1949). Applications of least squares regression to relationships containing autocorrelated errors. *J. Amer. Stat. Assoc.* **44**, 32–61.

5. Fuller, W. A. (1976). *Introduction to statistical time series analysis.* Wiley, New York.

6. Gorodetskii, V. V. (1977). On the strong mixing property for linear sequences. *Th. Prob. Appl.* **22**, 411–413.

7. Hannan, E. T. (1970). *Multiple time series.* Wiley, New York.

8. Müller, H. G. and U. Stadtmüller (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–625.

9. Priestley, M. B. (1981). *Spectral analysis and time series.* Academic Press, New York.

10. Priestley, M. B. (1988). *Non-linear and Non-stationary Time series.* Academic Press, New York.

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

12. Robinson, P. M. (1984). Robust nonparametric regression. *Robust and nonlinear time series analysis.* Franke, J., Härdle, W. and Martin, D., (eds.) Springer Lecture Notes, New York.

13. Shiryayev, A. N. (1984). *Probability.* Springer, New York.

14. Shumway, R. H. (1988). *Applied statistical time series analysis.* Prentice Hall, New Jersey.

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

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

17. Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis.* Lecture Notes in Statistics **21**. Springer, New York.

18. Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach.* Oxford University Press.

19. Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *J. Statist. Planning and Inference* **28**, 167–183.

20. Truong, Y. K. and Stone, C. J. (1992). Nonparametric function estimation involving time series. *Ann. Statist.* **20.** To appear.

21. Withers, C. S. (1981). Conditions for Linear Processes to be Strong Mixing. *Z. Wahrsch. verw. Gebiete* **57**, 477–480.