Functional-Coefficient Regression Models for Nonlinear Time Series

Zongwu CAI, Jianqing FAN, and Qiwei YAO

The local linear regression technique is applied to estimation of functional-coefficient regression models for time series data. The models include threshold autoregressive models and functional-coefficient autoregressive models as special cases but with the added advantages such as depicting finer structure of the underlying dynamics and better postsample forecasting performance. Also proposed are a new bootstrap test for the goodness of fit of models and a bandwidth selector based on newly defined cross-validatory estimation for the expected forecasting errors. The proposed methodology is data-analytic and of sufficient flexibility to analyze complex and multivariate nonlinear structures without suffering from the "curse of dimensionality." The asymptotic properties of the proposed estimators are investigated under the α-mixing condition. Both simulated and real data examples are used for illustration.

KEY WORDS: α-mixing; Asymptotic normality; Bootstrap; Forecasting; Goodness-of-fit test; Local linear regression; Nonlinear time series; Varying-coefficient models.

1. INTRODUCTION

Until recently, much of time series modeling has been confined to linear autoregressive moving average (ARMA) models (Box and Jenkins 1970). Although the original ARMA framework has been enlarged to include long-range dependence with fractional ARMA (Dahlhaus 1989; Granger and Joyeux 1980), multivariate vector ARMA and vector ARMA models with exogenous variables (Hannan and Deistler 1988), and random walk nonstationarities via cointegration (Engle and Granger 1987), there still exist so-called "nonlinear" features beyond the capacity of linear ARMA modeling. For example, various "nonstandard" phenomena such as nonnormality, asymmetric cycles, bimodality, nonlinear relationship between lagged variables, variation of prediction performance over the state-space, nonreversibility, and sensitivity to initial conditions have been well observed in many real time series data, including some benchmark sets such as the sunspot, lynx, and blowfly data. (See Tjøstheim 1994 and Tong 1990, 1995 for further discussion on this aspect.) Beyond linear domain, there are infinitely many nonlinear forms to be explored. Early development of nonlinear time series analysis focused on various nonlinear (sometimes non-Gaussian) parametric forms (Tjøstheim 1994; Tong 1990; references within). The successful examples include, among others, the autoregressive conditional heteroscedastic (ARCH) modeling of fluctuating structure for financial time series (Bollerslev 1986; Engel 1982), and the threshold modeling for biological and economic data (Tiao and Tsay 1994; Tong 1990). On the other hand, recent development in nonparametric regression techniques provides an alternative to model nonlinear time series (Härdle, Lütkepohl, and Chen 1997; Masry and Fan 1997; Tjøstheim 1994; Yao and Tong 1995). The immediate advantage of this is that little prior information on model structure is assumed. Further, it may provide useful insight for further parametric fitting. But an entirely nonparametric approach is hampered by the requirement of large sample sizes and is often practically useful only for, for example, autoregressive (AR) models with order 1 or 2.

This article adapts the functional-coefficient modeling technique to analyze nonlinear time series data. The approach allows appreciable flexibility on the structure of fitted models without suffering from the "curse of dimensionality." Let $\{\mathbf{U}_i, \mathbf{X}_i, Y_i\}_{i=-\infty}^{\infty}$ be jointly strictly stationary processes with $\mathbf{U}_i$ taking values in $\Re^k$ and $\mathbf{X}_i$ taking values in $\Re^p$. Typically, k is small. Let $E(Y_1^2) < \infty$. We define the multivariate regression function

$$m(\mathbf{u}, \mathbf{x}) = E(Y|\mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}), \tag{1}$$

where $(\mathbf{U}, \mathbf{X}, Y)$ has the same distribution as $(\mathbf{U}_i, \mathbf{X}_i, Y_i)$. In a pure time series context, both $\mathbf{U}_i$ and $\mathbf{X}_i$ consist of some lagged values of Y_i. The functional-coefficient regression model has the form

$$m(\mathbf{u}, \mathbf{x}) = \sum_{j=1}^{p} a_j(\mathbf{u}) x_j, \tag{2}$$

where the functions $\{a_j(\cdot)\}$ are measurable functions from $\Re^k$ to $\Re^1$ and $\mathbf{x} = (x_1, \ldots, x_p)^T$, with T denoting the transpose of a matrix or vector. The idea to model time series in such a form is not new (see, e.g., Nicholls and Quinn 1982). In fact, many useful time series models may be viewed as special cases of model (2) (often with specified parametric forms for the functions $\{a_j(\cdot)\}$; see Sec. 2). But the potential of this modeling technique had not been fully explored until the seminal work of Chen and Tsay (1993), Cleveland, Grosse, and Shyu (1992), and Hastie and Tibshirani (1993), in which nonparametric techniques were developed

Zongwu Cai is Assistant Professor, Department of Mathematics, University of North Carolina, Charlotte, NC 28223 (E-mail: *zcai@uncc.edu*). Jianqing Fan is Professor, Department of Statistics, University of California, Los Angeles, CA 90095 (E-mail: *jfan@stat.unc.edu*). Qiwei Yao is Reader, Department of Statistics, London School of Economics, London WC2A 2AE, U.K. (E-mail: *q.yao@lse.ac.uk*). Fan's research was partially supported by National Science Foundation grant DMS-9803200 and National Science Administration 96-1-0015. Yao's work was supported in part by Engineering and Physical Sciences Research Council grant L16358 and Biotechnology and Biological Sciences Research Council/Engineering and Physical Sciences Research Council grant 96/MMI09785. The authors thank the editor, the associate editor, and two referees for their insightful comments, which led a substantial improvement of the article.

for estimation of the functions $\{a_j(\cdot)\}$. In the context of independent samples, Fan and Zhang (1999) provided an innovative two-step method and insightful asymptotic results for the local polynomial estimation of $\{a_j(\cdot)\}$. They also pointed out that model (2) has strong connections with the functional linear models discussed by Brumback and Rice (1998) and Ramsay and Silverman (1997). Yet few results are available in the time series context.

In this article we adapt local linear regression technique to estimate the coefficient functions $\{a_j(\cdot)\}$. By smoothing $\mathbf{U}$ only, our method is particularly easy to implement. Within the framework of (2), the detailed form of model is determined by data, which will reduce the bias of fitting automatically. Because only k-dimensional functions are estimated, the difficulties associated with the "curse of dimensionality" will be substantially eased. Indeed, our data-analytic approach increases modeling flexibility with little sacrifice of estimability (see Theorem 2 in Sec. 6). The specified form of (2) also facilitates the interpretability of the fitted model when k is small. This is particularly relevant in modeling longitudinal data where it is reasonable to assume that the regression coefficients change over time t. (See Hoover, Rice, Wu, and Yang 1998 for a novel application of functional-coefficient models to longitudinal data.) Model (2) is also important for modeling the population dynamics, where it is reasonable to expect that animals behave differently based on its population size. Thus, using model (2) with $\mathbf{u}$ denoting the population size of a previous year captures such a kind of feature in the population dynamics [see Tong 1990, p. 377, and (8) for further discussion].

An important statistical question in fitting model (2) arises if the coefficient functions are actually varying (namely, if a linear AR model is adequate) or more generally if a parametric model fits the given data. This amounts to testing whether the coefficient functions are constant or in a certain parametric form. A new testing procedure, related to the sieve likelihood ratio statistic of Fan, Zhang, and Zhang (1999), is proposed based on the comparison of the residual sum of squares under the null and alternative models. A bootstrap method is proposed for finding the null distribution of the test statistic. Our simulation shows that the resulting testing procedure is indeed powerful, and the bootstrap method does give the correct null distribution. This is consistent with the Wilks phenomenon observed by Fan et al. (1999).

In Section 2 we mention several familiar nonlinear times series models that are within the framework of (2). Through the famous Canadian lynx data, we illustrate the advantages of the new approach over the existing parametric models in terms of both understanding the underlying dynamics and forecasting the postsample. In Section 3 we present local linear regression estimators for the functional-coefficient functions and offer a simple and fast algorithm for bandwidth selection. In Section 4 we propose a bootstrap method for testing the goodness of fit of a parametric model against model (2). In Section 5 we use both simulated models and real datasets to illustrate the proposed methodology. The applications with real data lend further support to use some well-known parametric models. We study the asymptotic properties of the proposed estimators in Section 6. All technical proofs are relegated to the Appendix.

2. MODELS AND AN ILLUSTRATIVE EXAMPLE

The general setting (2) includes many familiar time series models, a few of which we mention here. We use some of them for numerical illustration in Section 5.

2.1 Functional-Coefficient Autoregressive Model

Chen and Tsay (1993) proposed the functional-coefficient autoregressive (FAR) model

$$x_t = a_1(\mathbf{X}_{t-1}^*)x_{t-1} + \cdots + a_p(\mathbf{X}_{t-1}^*)x_{t-p} + \varepsilon_t, \quad (3)$$

where $\mathbf{X}_{t-1}^* = (x_{t-i_1}, \ldots, x_{t-i_k})^T$, $\{\varepsilon_t\}$ is a sequence of iid random variables and ε_t is independent of $\{x_{t-i}, i > 0\}$. Chen and Tsay studied probabilistic properties of FAR models and proposed an iterative algorithm to estimate the coefficient functions. In fact, their algorithm is in the spirit of local constant fitting, although they did not apply local regression technique directly, but rather constructed estimators based on an iterative recursive formula.

2.2 Threshold Autoregressive Model

One of the simplest nonlinear time series models is the threshold autoregressive (TAR) model,

$$x_t = \phi_1^{(i)}x_{t-1} + \cdots + \phi_p^{(i)}x_{t-p} + \varepsilon_t^{(i)} \quad \text{if } x_{t-d} \in \Omega_i, \quad i = 1, \ldots, k, \quad (4)$$

where $\{\Omega_i\}$ form a (nonoverlapping) partition of the real line. Theoretical properties and practical implementations of TAR modeling have been covered by Tong (1990).

2.3 Exponential Autoregressive Model

The following generalized exponential autoregressive (EXPAR) model was proposed and studied by Haggan and Ozaki (1981) and Ozaki (1982):

$$x_t = \sum_{i=1}^{p} \{\alpha_i + (\beta_i + \gamma_i x_{t-d}) \exp(-\theta_i x_{t-d}^2)\}x_{t-i} + \varepsilon_t, \quad (5)$$

where $\theta_i \geq 0$ for $i = 1, \ldots, p$.

2.4 Regression With Random Coefficients

Consider the model of Granger and Teräsvirta (1993),

$$Y_t = \boldsymbol{\beta}(t)^T\mathbf{X}_t + u_t, \quad (6)$$

where $\{u_t\}$ is a sequence of iid random variables with $E(u_t) = 0$ and $\text{var}(u_t) = \sigma^2$ and is independent of $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\beta}(t)\}$. Further, $E(\boldsymbol{\beta}(t)) = \boldsymbol{\beta}$ and $\text{var}(\boldsymbol{\beta}(t)) = \Phi$, $\text{cov}(\boldsymbol{\beta}(s), \boldsymbol{\beta}(t)) = 0$ for $s \neq t$. The foregoing random coefficient model has received considerable attention in econometrics (see Granger and Teräsvirta 1993). If $\mathbf{X}_t = (Y_{t-1}, \ldots, Y_{t-p})^T$, then (6) is the random coefficient AR model surveyed by Nicholls and Quinn (1982).

To our knowledge, it remains as an open question to derive the general conditions under which a FAR model is stationary. It is well-known that an ergodic Markov process

initiated from its invariant distribution is (strictly) stationary. Note that any AR model can be expressed as a vector-valued Markov model. Thus it is common practice to prove ergodicity to establish the stationarity. Recent results in this direction include those of An and Chen (1997) and An and Huang (1996), which surveyed various sufficient conditions for the ergodicity of nonlinear AR models, including some special cases of FAR models.

All of the foregoing models have proven successful for modeling *some* nonlinear features. For example, the TAR model has received considerable attention due to its easy implementation and often nice interpretation. The application to Canadian lynx data (i.e., the annual fur returns of lynx at auction in 1821–1934) is arguably a showcase of the TAR modeling technique (see Tong 1990). The periodic fluctuation displayed in this time series has profoundly influenced ecological theory. The dataset has been constantly used to examine the concepts as "balance of nature," predator–prey interaction, food web dynamics, and so on (see Stenseth et al. 1998 and references therein). Having incorporated biological evidence, Tong (1990) fitted the following TAR model with two regimes and the delay variable at lag 2 to the lynx data at the logarithmic scale with the base 10:

$$x_t = \begin{cases} .62 + 1.25x_{t-1} - .43x_{t-2} + \varepsilon_t^{(1)}, \\ \qquad\qquad\qquad\qquad x_{t-2} \le 3.25 \\ 2.25 + 1.52x_{t-1} - 1.24x_{t-1} - 1.24x_{t-2} + \varepsilon_t^{(2)}, \\ \qquad\qquad\qquad\qquad x_{t-2} > 3.25 \end{cases} \tag{7}$$

(see Tong 1990, p. 377). This simple model admits nice biological interpretation. Indeed, it can be viewed as derived from basic predator (lynx) and prey (hare) interaction model in ecology (see eq. 2 of Stenseth et al. 1998). The lower regime corresponds roughly to the population increase phase, and the upper regime corresponds to the population decrease phase. Note that the coefficient of x_{t-1} in the model is significantly positive, but less so during the increase phase. The coefficient of x_{t-2} is significantly negative, more so during the decline phase. The signs of those coefficients reflect that lynx and hare relate with each other in a specified prey–predator interactive manner. The difference of the coefficients in the increase and decrease phases reflects the so-called "phase dependence" and "density dependence" in ecology (Stenseth et al. 1998). The phase dependence means that both lynx and hare behave differently (in hunting or escaping) when the lynx population increases or decreases. The density dependence implies that the reproduction rates of animals, as well as their behavior, depend on the abundance of the population. Clearly the foregoing threshold model simplified the varying behavior into two states. With the new technique proposed in this article, we fit the lynx data with the model

$$x_t = a_1(x_{t-2})x_{t-1} + a_2(x_{t-2})x_{t-2} + \varepsilon_t, \tag{8}$$

in which the coefficients $a_1(\cdot)$ and $a_2(\cdot)$ vary with respect to "threshold variable" x_{t-2}. Both $a_1(\cdot)$ and $a_2(\cdot)$ are estimated through a simple one-dimensional kernel regression. The estimators are plotted in Figures 1(a) and 1(b). Except a few points near the low end, $a_1(\cdot)$ is a positive increasing function, which depicts the *smooth* (rather than *radical*) density dependence. The function $a_2(\cdot)$ is negative and largely decreasing. This pictures a gradual change in animal behavior in corresponding to the change in population abundance. By allowing the coefficient to vary with respect to population density, the model presents the lynx–hare interaction in the manner that is one step closer to the reality than the TAR models. The advantages of the new technique on other aspects such as prediction will be reported in Section 5.

In summary, the functional-coefficient model (2) provides a simple alternative to the existing techniques such as TAR and FAR for modeling nonlinear time series in a continuous manner. It allows us to make full use of available information to model local variation at a finer scale. However, simple parametric models such as TAR would be more appealing when the sample size is small or when discontinuities exist genuinely.

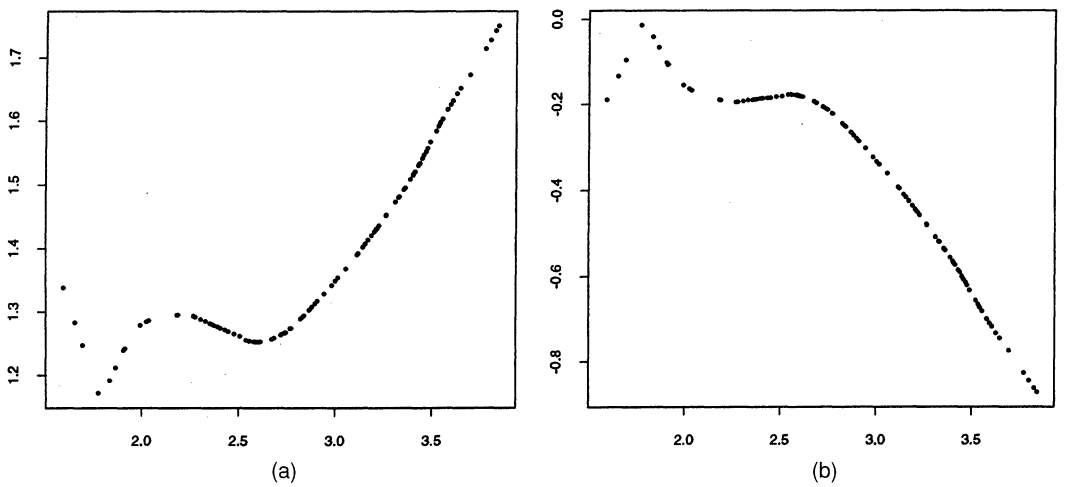

Figure 1. Canadian Lynx Data: Local Linear Estimator of (a) $a_1(x_{t-2})$ and (b) $a_2(x_{t-2})$ in (8).

3. ESTIMATION

For simplicity we consider only the case $k = 1$ in (2). Extension to the case $k > 1$ involves no fundamentally new ideas. Note that models with large k are often not practically useful due to the "curse of dimensionality."

3.1 Local Linear Regression Estimation

Local linear fittings have several nice properties. They possess high statistical efficiency in an asymptotic minimax sense and are design adaptive (Fan 1993). Further, they automatically correct edge effects (Fan and Gijbels 1996; Hastie and Loader 1993; Ruppert and Wand 1994). We estimate the functions $a_j(\cdot)$'s using the local linear regression method from observations $\{U_i, \mathbf{X}_i, Y_i\}_{i=1}^n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$. We assume throughout the article that $a_j(\cdot)$ has a continuous second derivative. Note that we may approximate $a_j(\cdot)$ locally at u_0 by a linear function $a_j(u) \approx a_j + b_j(u - u_0)$. The local linear estimator is defined as $\hat{a}_j(u_0) = \hat{a}_j$, where $\{(\hat{a}_j, \hat{b}_j)\}$ minimize the sum of weighted squares

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{p} \{a_j + b_j(U_i - u_0)\} X_{ij} \right]^2 K_h(U_i - u_0), \quad (9)$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function on $\Re^1$ and $h > 0$ is a bandwidth. It follows from the least squares theory that

$$\hat{a}_j(u_0) = \sum_{k=1}^{n} K_{n,j}(U_k - u_0, \mathbf{X}_k) Y_k, \quad (10)$$

where

$$K_{n,j}(u, \mathbf{x}) = \mathbf{e}_{j,2p}^T (\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}})^{-1} \begin{pmatrix} \mathbf{x} \\ u\mathbf{x} \end{pmatrix} K_h(u). \quad (11)$$

In the foregoing expression, $\mathbf{e}_{j,2p}$ is the $2p \times 1$ unit vector with 1 at the jth position, $\tilde{\mathbf{X}}$ denotes an $n \times 2p$ matrix with $(\mathbf{X}_i^T, \mathbf{X}_i^T(U_i - u_0))$ as its ith row, and $\mathbf{W} = \text{diag}\{K_h(U_1 - u_0), \ldots, K_h(U_n - u_0)\}$.

3.2 Bandwidth Selection

Various existing bandwidth selection techniques for nonparametric regression can be adapted for the above estimation (see, e.g., Fan, Yao, and Cai 2000). But here we propose a simple and quick method for selecting bandwidth h. It can be regarded as a modified multifold cross-validation criterion that is attentive to the structure of *stationary* time series data. Let m and Q be two given positive integers and $n > mQ$. The basic idea is first to use Q subseries of lengths $n - qm$ $(q = 1, \ldots, Q)$ to estimate the unknown coefficient functions and then compute the one-step forecasting errors of the next section of the time series of length m based on the estimated models. More precisely, we choose h that minimizes the average mean squared (AMS) error

$$\text{AMS}(h) = \sum_{q=1}^{Q} \text{AMS}_q(h), \quad (12)$$

where for $q = 1, \ldots, Q$,

$$\text{AMS}_q(h) = \frac{1}{m} \sum_{i=n-qm+1}^{n-qm+m} \left\{ Y_i - \sum_{j=1}^{p} \hat{a}_{j,q}(U_i) X_{i,j} \right\}^2,$$

and $\{\hat{a}_{j,q}(\cdot)\}$ are computed from the sample $\{(U_i, \mathbf{X}_i, Y_i), 1 \le i \le n - qm\}$ with bandwidth equal $h[n/(n - qm)]^{1/5}$. Note that we rescale bandwidth h for different sample sizes according to its optimal rate, i.e. $h \propto n^{-1/5}$. In practical implementations, we may use $m = [0.1n]$ and $Q = 4$. The selected bandwidth does not depend critically on the choice of m and Q, as long as mQ is reasonably large so that the evaluation of prediction errors is stable. A weighted version of AMS(h) can be used, if one wishes to downweight the prediction errors at an earlier time. We take $m = [0.1n]$ rather than $m = 1$ simply because of computation expediency.

3.3 Smoothing Variable Selection

Of importance is to choose an appropriate smoothing variable U in applying functional-coefficient regression models. Knowledge on physical background of the data may be very helpful, as we have witnessed in modeling the lynx data in Section 2. Without any prior information, it is pertinent to choose U in terms of some data-driven methods such as the Akaike information criterion, cross-validation, and other criteria. Ideally, we would choose U as a linear function of given explanatory variables according to some optimal criterion, which we fully explored in earlier work (Fan et al. 2000). Nevertheless, we propose here a simple and practical approach: let U be one of the given explanatory variables such that AMS defined in (12) obtains its minimum value. Obviously, this idea can be also extended to select p as well. Example 4 in Section 5.2 presents the practical implementation of this approach.

4. GOODNESS-OF-FIT TEST

To test whether model (2) holds with a specified parametric form such as the TAR or EXPAR model (see Sec. 2), we propose a goodness-of-fit test based on the comparison of the residual sum of squares (RSS) from both parametric and nonparametric fittings. This method is closely related to the sieve likelihood method proposed by Fan et al. (1999). Those authors demonstrated the optimality of this kind of procedures for independent samples.

Consider the null hypothesis

$$H_0: \; a_j(u) = \alpha_j(u, \boldsymbol{\theta}), \qquad 1 \le j \le p, \quad (13)$$

where $\alpha_j(\cdot, \boldsymbol{\theta})$ is a given family of functions indexed by unknown parameter vector $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$. The RSS under the null hypothesis is

$$\text{RSS}_0 = n^{-1} \sum_{i=1}^{n} \{Y_i - \alpha_1(U_i, \hat{\boldsymbol{\theta}}) X_{i1} - \cdots - \alpha_p(U_i, \hat{\boldsymbol{\theta}}) X_{ip}\}^2.$$

Analogously, the RSS corresponding to model (2) is

$$\text{RSS}_1 = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{a}_1(U_i) X_{i1} - \cdots - \hat{a}_p(U_i) X_{ip}\}^2.$$

The test statistic is defined as

$$\mathbf{T}_n = (\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1 = \mathrm{RSS}_0/\mathrm{RSS}_1 - 1,$$

and we reject the null hypothesis (13) for large value of $\mathbf{T}_n$. We use the following nonparametric bootstrap approach to evaluate the p value of the test:

1. Generate the bootstrap residuals $\{\varepsilon_i^*\}_{i=1}^n$ from the empirical distribution of the centered residuals $\{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}\}_{i=1}^n$, where

$$\hat{\varepsilon}_i = Y_i - \hat{a}_1(U_i)X_{i1} - \cdots - \hat{a}_p(U_i)X_{ip}, \bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i,$$

and define

$$Y_i^* = \alpha_1(U_i, \hat{\boldsymbol{\theta}})X_{i1} + \cdots + \alpha_p(U_i, \hat{\boldsymbol{\theta}})X_{ip} + \varepsilon_i^*.$$

2. Calculate the bootstrap test statistic $\mathbf{T}_n^*$ based on the sample $\{U_i, \mathbf{X}_i, Y_i^*\}_{i=1}^n$.
3. Reject the null hypothesis H_0 when $\mathbf{T}_n$ is greater than the upper-α point of the conditional distribution of $\mathbf{T}_n^*$ given $\{U_i, \mathbf{X}_i, Y_i\}_{i=1}^n$.

The p value of the test is simply the relative frequency of the event $\{\mathbf{T}_n^* \geq \mathbf{T}_n\}$ in the replications of the bootstrap sampling. For the sake of simplicity, we use the same bandwidth in calculating $\mathbf{T}_n^*$ as that in $\mathbf{T}_n$. Note that we bootstrap the centralized residuals from the nonparametric fit instead of the parametric fit, because the nonparametric estimate of residuals is always consistent, no matter whether the null or the alternative hypothesis is correct. The method should provide a consistent estimator of the null distribution even when the null hypothesis does not hold. Kreiss, Neumann, and Yao (1998) considered nonparametric bootstrap tests in a general nonparametric regression setting. They proved that, asymptotically, the conditional distribution of the bootstrap test statistic is indeed the distribution of the test statistic under the null hypothesis. It may be proven that the similar result holds here as long as $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ at the rate $n^{-1/2}$.

5. NUMERICAL PROPERTIES

We illustrate the proposed methods through two simulated and two real data examples. The estimators $\{\hat{a}_j(\cdot)\}$ are assessed via the square root of average squared errors (RASE),

$$\mathrm{RASE}^2 = \sum_{j=1}^p \mathrm{RASE}_j^2, \tag{14}$$

where

$$\mathrm{RASE}_j = \left[n_{\mathrm{grid}}^{-1}\sum_{k=1}^{n_{\mathrm{grid}}} \{\hat{a}_j(u_k) - a_j(u_k)\}^2\right]^{1/2}$$

and $\{u_k, k = 1, \ldots, n_{\mathrm{grid}}\}$ are regular grid points. We also compare the postsample forecasting performance of the new method with existing methods such as the linear AR model, the TAR model, and the FAR model (implemented by Chen and Tsay 1993). We consider three predictors based on functional-coefficient modeling (3): the one-step-ahead predictor

$$\hat{x}_{t+1} = \hat{a}_1(\mathbf{X}_t^*)x_t + \cdots + \hat{a}_p(\mathbf{X}_t^*)x_{t-p+1};$$

the iterative two-step-ahead predictor

$$\hat{x}_{t+2} = \hat{a}_1(\hat{\mathbf{X}}_{t+1}^*)\hat{x}_{t+1} + \hat{a}_2(\hat{\mathbf{X}}_{t+1}^*)x_t + \cdots + \hat{a}_p(\hat{\mathbf{X}}_{t+1}^*)x_{t-p+2}; \tag{15}$$

and the direct two-step-ahead predictor based on the model

$$x_{t+2} = b_1(\mathbf{X}_t^*)x_t + \cdots + b_p(\mathbf{X}_t^*)x_{t-p} + \varepsilon_t'. \tag{16}$$

Note that model (3) does not necessarily imply (16). In this sense the direct two-step-ahead prediction explores the predictive power of the proposed modeling techniques when the model is misspecified. We always use the Epanechnikov kernel $K(u) = .75(1-u^2)_+$. For the two real data examples, we repeat bootstrap sampling 1,000 times in goodness-of-fit tests, and select the bandwidths by the method proposed in Section 3.2.

5.1 Simulated Examples

Example 1. We first consider an EXPAR model. We replicate simulation 400 times and each time draw a time series with length 400 from the model

$$x_t = a_1(x_{t-1})x_{t-1} + a_2(x_{t-1})x_{t-2} + \varepsilon_t, \tag{17}$$

where $a_1(u) = .138 + (.316 + .982u)\mathrm{e}^{-3.89u^2}$, $a_2(u) = -.437 - (.659 + 1.260u)\mathrm{e}^{-3.89u^2}$, and $\{\varepsilon_t\}$ are iid $\mathrm{N}(0, .2^2)$. We choose the optimal bandwidth $h_n = .41$ that minimizes the sum of the integrated squared errors of estimators for $a_1(\cdot)$ and $a_2(\cdot)$. Figures 2(a) and 2(b) present the estimated $a_1(\cdot)$ and $a_2(\cdot)$ from a typical sample. The typical sample is selected in such a way that its RASE value is equal to the median in the 400 replications. The boxplot for 400 RASE values is presented in Figure 2(c). To gauge the performance of our procedure in terms of RASE, we computed the standard deviation of the time series $\{x_t\}$, denoted by σ_X. The mean and standard deviation of the σ_X in the simulation with 400 replications are .5389 and .0480. Overall, the proposed modeling procedure performs fairly well.

To demonstrate the power of the proposed bootstrap test, we consider the null hypothesis

$$H_0\colon\ a_j(u) = \theta_j, \qquad j = 1, 2,$$

namely a linear AR model, versus the alternative

$$H_1\colon\ a_j(u) \neq \theta_j, \quad \text{for at least one } j.$$

The power function is evaluated under a sequence of the alternative models indexed by β,

$$H_1\colon\ a_j(u) = \bar{a}_j^0 + \beta(a_j^0(u) - \bar{a}_j^0), \qquad j = 1, 2 \qquad (0 \leq \beta \leq 1),$$

where $\{a_j^0(u)\}$ are the solid curves given in Figures 2(a) and 2(b) and $\bar{a}_j^0$ is the average height of $a_j^0(u)$. We apply the

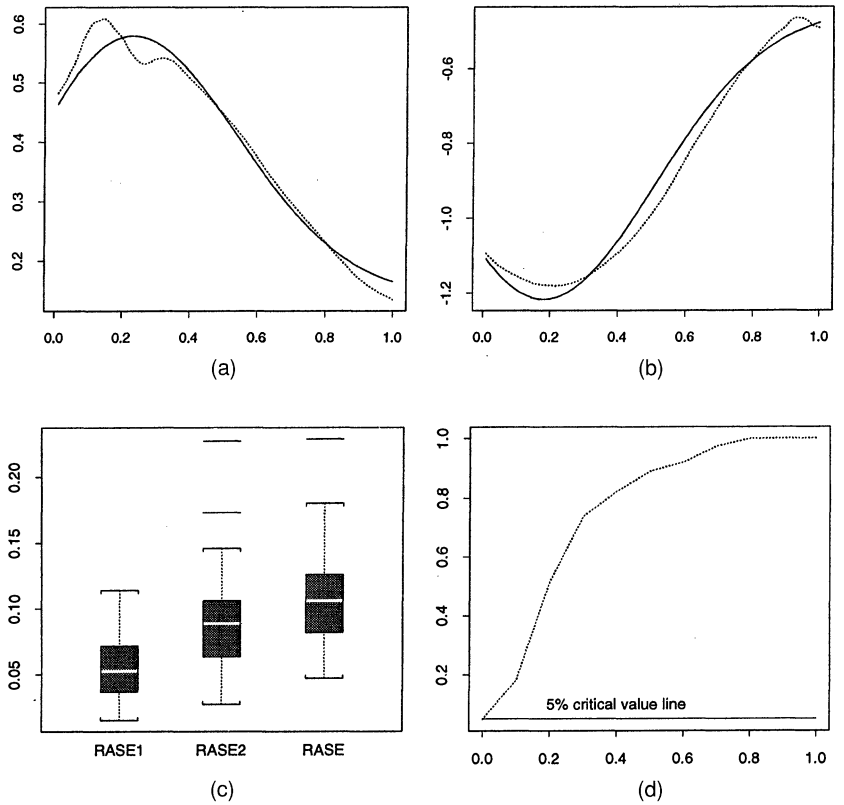

Figure 2. Simulation Results for Example 1. (a) The local linear estimator (dotted line) of the coefficient function $a_1(x_{t-1})$ (solid line); (b) the local linear estimator (dotted line) of $a_2(x_{t-1})$ (solid line); (c) the boxplots of the 400 RASE values in estimation of $a_1(\cdot)$ and $a_2(\cdot)$; (d) the plot of power curve against β for the goodness-of-fit test.

goodness-of-fit test described in Section 4 in a simulation with 400 replications. For each realization, we repeat bootstrap sampling 500 times. Figure 2(d) plots the simulated power function against β. When $\beta = 0$, the specified alternative hypothesis collapses into the null hypothesis. The power is .047, which is close to the significance level of 5%. This demonstrates that bootstrap estimate of the null distribution is approximately correct. The power function shows that our test is indeed powerful. To appreciate why, consider the specific alternative with $\beta = .4$. The functions $\{a_j(u)\}$ under H_1 are shown in Figure 3. The null hypothesis is essentially the constant curves in Figure 3. Even with such a small difference under our noise level, we can correctly detect the alternative over 80% among the 400 simulations. The power increases rapidly to 1 when $\beta = .8$. When $\beta = 1$, we test the constant functions in Figure 3 against the coefficient functions in Figures 2(a) and 2(b).

Example 2. We now consider a TAR model,

$$x_t = a_1(x_{t-2})x_{t-1} + a_2(x_{t-2})x_{t-2} + \varepsilon_t, \qquad (18)$$

where $a_1(u) = .4I\ (u \leq 1) - .8I\ (u > 1), a_2(u) = -.6I$

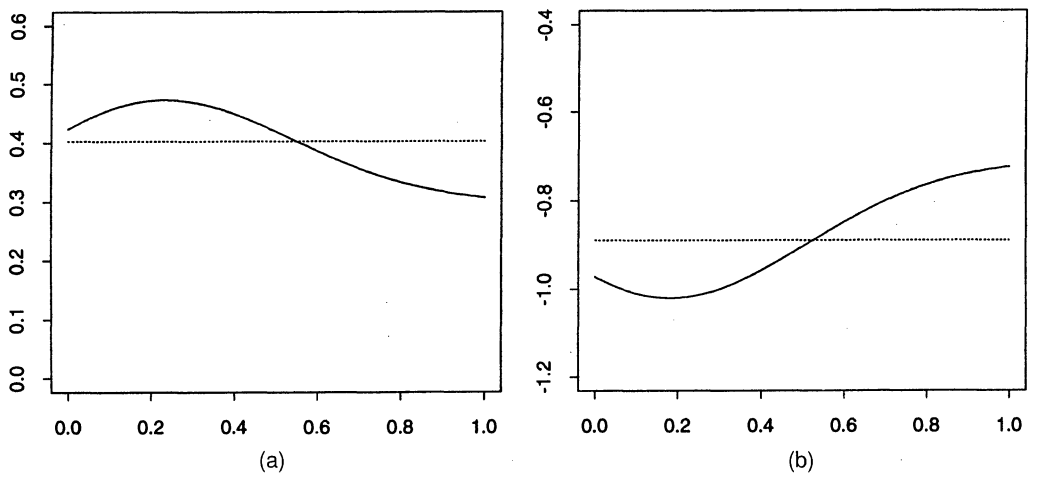

Figure 3. The Coefficient Functions Under (a) the Null Hypothesis and (b) a Specific Alternative Hypothesis With $\beta = .4$. The solid curves are coefficient functions under H_1; the dotted lines are the coefficient functions under H_0.

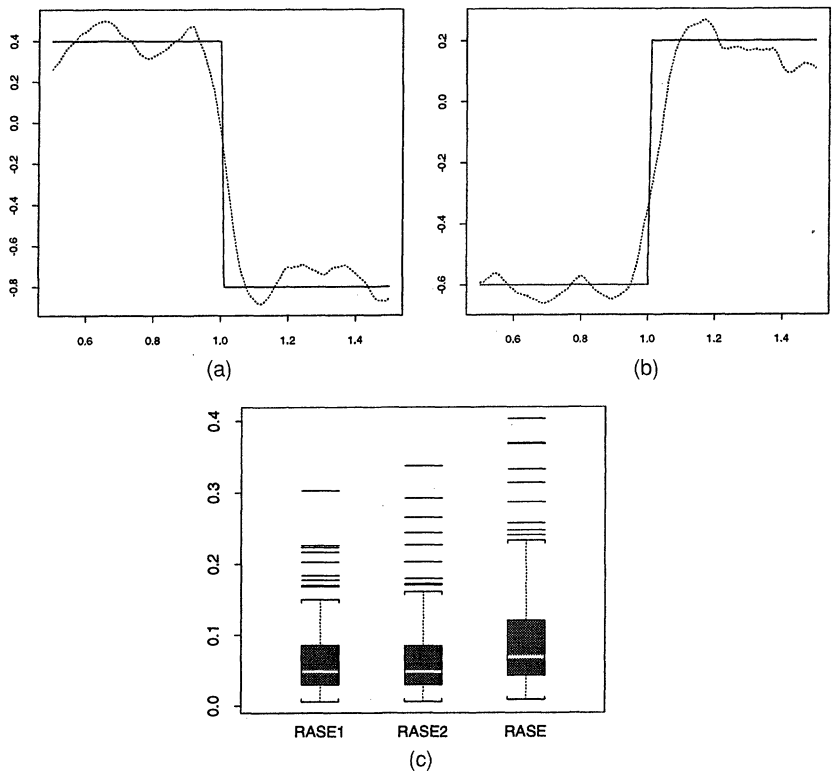

Figure 4. Simulation Results for Example 2. (a) The local linear estimator (dotted line) of the coefficient function $a_1(x_{t-2})$ (solid line); (b) the local linear estimator (dotted line) of $a_2(x_{t-2})$ (solid line); (c) the boxplot of the 400 RASE values in estimation of $a_1(\cdot)$ and $a_2(\cdot)$.

$(u \leq 1) + .2I\ (u > 1)$, and $\{\varepsilon_t\}$ are iid N(0, 1). With sample size $n = 500$, we replicate simulation 400 times. As in Example 1, the optimal bandwidth $h_n = .325$ is used. The boxplot for 400 RASE values is presented in Figure 4(c), and the local linear estimators of $a_1(\cdot)$ and $a_2(\cdot)$ from a typical sample are plotted in Figures 4(a) and 4(b). The typical sample is selected in such a way that its RASE value is equal to the median in the 400 replications.

To compare the prediction performance of the three predictors from functional-coefficient modeling with the best-fitted linear AR(2) model,

$$\hat{x}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t-1} + \hat{\beta}_2 x_{t-2},$$

we predict 10 postsample points in each of the 400 replicated simulations. The mean and standard deviation (SD, in parentheses) of the average absolute predictive errors (AAPE) are recorded in Table 1. Note that $E|\varepsilon_t| = .7979$ and $\mathrm{SD}(|\varepsilon_t|) = .6028$, so that the average of 10 absolute deviation errors has an SD of .1897. These are indeed very close to one-step AAPE and its associated SD using model (18), and imply that the errors in estimating functions $a_1(\cdot)$ and $a_2(\cdot)$ are negligible in the prediction. It is clear that the functional-coefficient AR modeling, although somewhat overparameterized, provides more relevant predictors for the given model (18). Note that the direct predictor based on functional-coefficient model (16) performs reasonably well due to the flexibility of the functional-coefficient models.

Table 1. Mean and Standard Deviation of AAPE Based on 400 Replications

	One-step	*Iterative two-step*	*Direct two-step*
Model (18)	.784 (.203)	.904 (.273)	.918 (.281)
Linear AR(2)	1.131 (.485)	1.117 (.496)	

5.2 Real Data Examples

Example 3. We continue the discussion on Canadian lynx data in Section 2. To fit model (8), we select the bandwidth that minimizes AMS(h) defined in (12). To this end, we let $Q = 4$ and $m = 11$. Figure 5(b) plots the AMS values against h. The selected bandwidth is $h_n = .90$. The fitted values from both functional-coefficient model (8) and TAR model (7) are very close to each other; see Figure 5(a). Our goodness-of-fit test lends further support to using the TAR model. In fact, the RSS_1 for model (8) is .0406, which is slightly smaller than $\mathrm{RSS}_0 = .0414$ for the TAR model (7). The p value of the test is .714. Indeed, the TAR model (7) and the model (8) with the coefficient functions given in Figure 1 are statistically indistinguishable for this dataset. The difference lies in the interpretation of the two models (see Sec. 2). On the other hand, the p value of the goodness-of-fit test to test for the linear AR(2) model against the functional-coefficient model (8) is less than .001, which reinforces the existence of nonlinearity in the lynx data.

To compare the prediction performance of various models, we estimate the functional-coefficient model (8), a TAR model, and a linear AR(2) model using the first 102 data points only. We leave out last 12 points to check the prediction performance. The fitted TAR model is

$$\hat{x}_t = \begin{cases} .424 + 1.255x_{t-1} - .348x_{t-2}, & x_{t-2} \leq 2.981 \\ 1.882 + 1.516x_{t-1} - 1.126x_{t-2}, & x_{t-2} > 2.981. \end{cases} \tag{19}$$

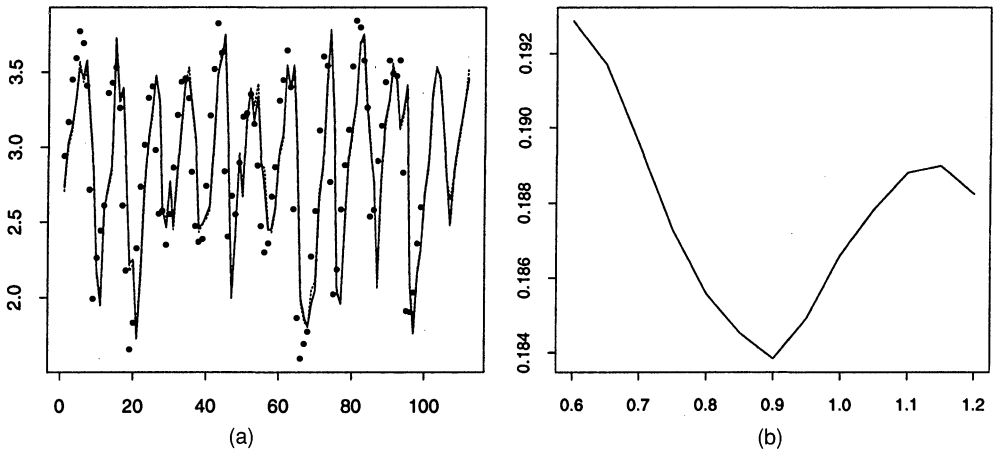

Figure 5. Canadian Lynx Data. (a) Time plots of the fitted values from TAR model (7) (solid line) and the fitted values of functional-coefficient model (8) (dotted line). The true values are indicated by ".". (b) Plot of the AMS against bandwidth.

The fitted linear AR(2) model is $\hat{x}_t = 1.048 + 1.376x_{t-1} - .740x_{t-2}$. Both the TAR and linear models are estimated using the least squares method. The threshold was searched among 60% inner sample points. The absolute prediction errors are reported in Table 2, which shows that the functional-coefficient model has better performance than both the TAR and linear AR(2) models. For example, for one-step-ahead prediction, the AAPE was reduced by 36% when the TAR model was used instead of the linear AR(2) model, and was further reduced by 25% when the functional-coefficient model was used instead of the TAR model.

Models (7) and (8) look different on the surface. However, they provide more or less equally good fits to the data, as evidenced by the goodness-of-fit test conducted earlier. The improvement in prediction by model (8) was due to local smoothing at smaller scales, which is based on the understanding that population dynamics varies in a continuous manner. Although the improvement in terms of relative predictive errors is evident, the difference in the AAPE occurs only at the second decimal place, which is not very substantial with respect to the dynamic range of the data.

Tong (1990) also suggested a more refined model involving seven lagged variables:

$$x_t = \begin{cases} .546 + 1.032x_{t-1} - .173x_{t-2} + .171x_{t-3} \\ \quad - .431x_{t-4} + .332x_{t-5} - .284x_{t-6} \\ \quad + .210x_{t-7} + \varepsilon_t^{(1)}, \\ \qquad\qquad \text{if } x_{t-2} \le 3.116 \\ 2.632 + 1.492x_{t-1} - 1.324x_{t-2} + \varepsilon_t^{(2)}, \\ \qquad\qquad \text{if } x_{t-2} > 3.116 \end{cases} \tag{20}$$

(see Tong 1990, p. 387). We fit the following, more-complex, functional-coefficient model accordingly:

$$x_t = \sum_{j=1}^{7} a_j(x_{t-2})x_{t-j} + \varepsilon_t. \tag{21}$$

The selected bandwidth is $h_n = 1.45$; see Figure 6(c). The estimated functions $a_j(\cdot)$ $(1 \le j \le 7)$ are plotted in Figure 6(a), which shows that the dynamical change is predominantly dictated by $a_1(\cdot)$ and $a_2(\cdot)$. The fitted values of the two models are very close to one another, as shown in Figure 6(b). We apply the goodness-of-fit test to test the TAR model (20) against the functional-coefficient model (21). The p value is .883, which again supports using the TAR model for the lynx data.

Table 2. The Postsample Predictive Errors for the Canadian Lynx Data

		Model (8)			TAR model (19)		Linear AR(2)	
Year	x_t	One-step	Iterative	Direct	One-step	Iterative	One-step	Iterative
1923	3.054	.157	.156	.209	.187	.090	.173	.087
1924	3.386	.012	.227	.383	.035	.269	.061	.299
1925	3.553	.021	.035	.195	.014	.038	.106	.189
1926	3.468	.008	.037	.034	.022	.000	.036	.182
1927	3.187	.085	.101	.295	.059	.092	.003	.046
1928	2.723	.055	.086	.339	.075	.015	.143	.148
1929	2.686	.135	.061	.055	.273	.160	.248	.051
1930	2.821	.016	.150	.318	.026	.316	.093	.434
1931	3.000	.017	.037	.111	.030	.062	.058	.185
1932	3.201	.007	.014	.151	.060	.043	.113	.193
1933	3.424	.089	.098	.209	.076	.067	.191	.347
1934	3.531	.053	.175	.178	.072	.187	.140	.403
AAPE		.055	.095	.206	.073	.112	.114	.214

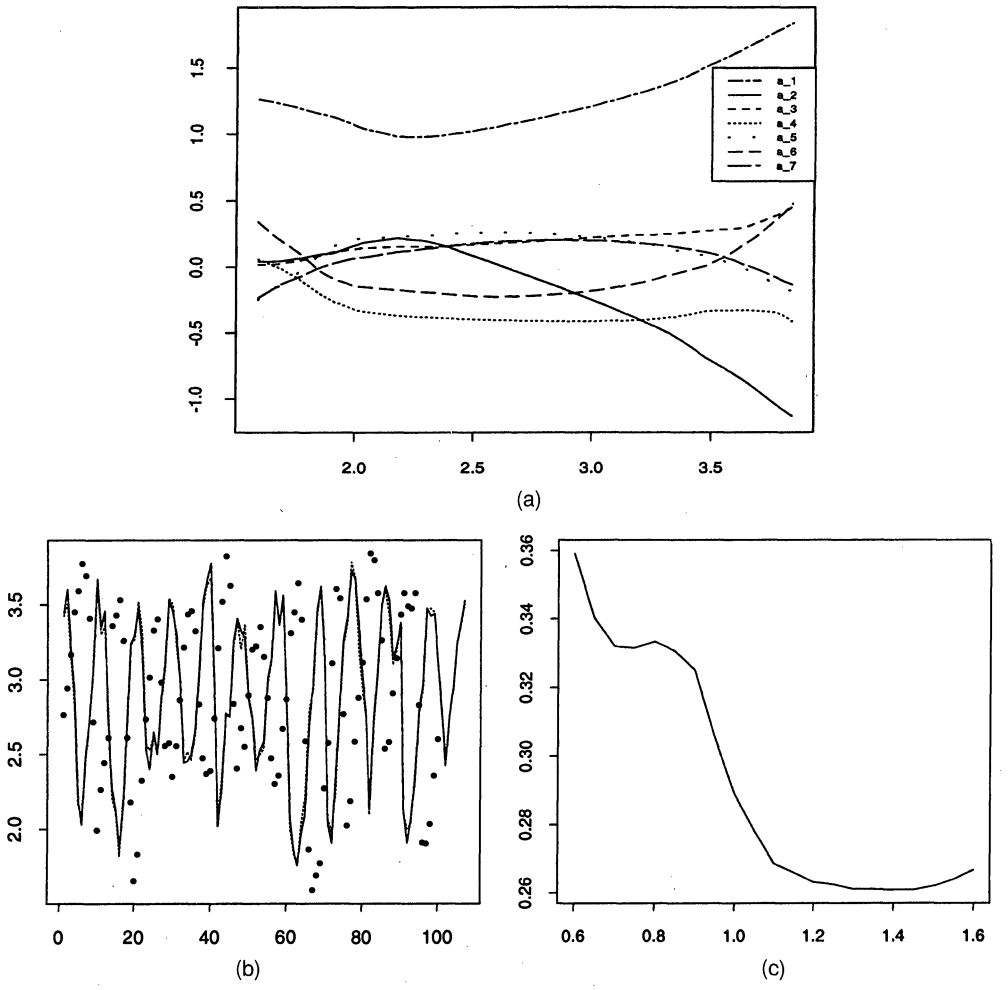

Figure 6. Canadian Lynx Data. (a) The estimated curves for functional coefficients $a_i(x_{t-2})$ ($i = 1, \ldots, 7$) in model (21). — - - a_1; ——— a_2; - - - a_3; · · · · · · · · · a_4; - - a_5; — — a_6; —— - a_7. (b) The time plots of the fitted values from model (20) (solid line) and the fitted values from model (21) (dotted line), with the true values indicated by ".". (c) The plot of the AMS against bandwidth.

Example 4. In many respects, Wolf's annual sunspot numbers are known to be challenging (see, e.g., Tong 1990). Following the convention in the literature, we first apply the transform $x_t = 2(\sqrt{1+y_t} - 1)$ to the 288 annual sunspot numbers in 1700–1987 (see, e.g., Chen and Tsay 1993; Ghaddar and Tong 1981). We apply the technique proposed in Section 3.3 to select the *optimum* functional-coefficient models among the class of models $x_t = \sum_{j=1}^{p} a_j(x_{t-d})x_{t-j} + \varepsilon_t$ with $1 \le d \le p$ and $2 \le p \le 11$. We let $m = 28$ and $Q = 4$ in AMS defined as in (12). Table 3 records the best model with each value of p between 2 and 11. The overall optimum model should be of order $p = 7$ or 8; the smooth variable, at lag $d = 3$.

Note that the FAR model proposed by Chen and Tsay (1993, p. 305) is

$$x_t = \begin{cases} 1.23 + (1.75 - .17|x_{t-3} - 6.6|)x_{t-1} \\ \quad + (-1.28 + .27|x_{t-3} - 6.6|)x_{t-2} \\ \quad + .20x_{t-8} + \varepsilon_t^{(1)}, \\ \qquad\qquad \text{if } x_{t-3} < 10.3 \\ .92 - .24x_{t-3} + .87x_{t-1} + .17x_{t-2} \\ \quad - .06x_{t-6} + .04x_{t-8} + \varepsilon_t^{(2)}, \\ \qquad\qquad \text{if } x_{t-3} \ge 10.3. \end{cases} \tag{22}$$

Combining this with the aforementioned result from the model selection, we fit the data with the functional-coefficient model

$$x_t = a_1(x_{t-3})x_{t-1} + a_2(x_{t-3})x_{t-2} + a_3(x_{t-3})x_{t-3} + a_6(x_{t-3})x_{t-6} + a_8(x_{t-3})x_{t-8} + \varepsilon_t. \tag{23}$$

The estimated coefficient functions are plotted in Figures 7(a)–(e). The selected bandwidth is $h_n = 4.75$ [see Fig. 7(f)], which minimizes the AMS defined as in (12).

Table 3. Selected Functional-Coefficient Models for the Sunspot Data

p	2	3	4	5	6	7	8	9	10	11
d	1	3	3	2	2	3	3	5	3	5
AMS	18.69	13.46	13.90	12.26	13.93	11.68	11.95	14.06	14.26	13.91

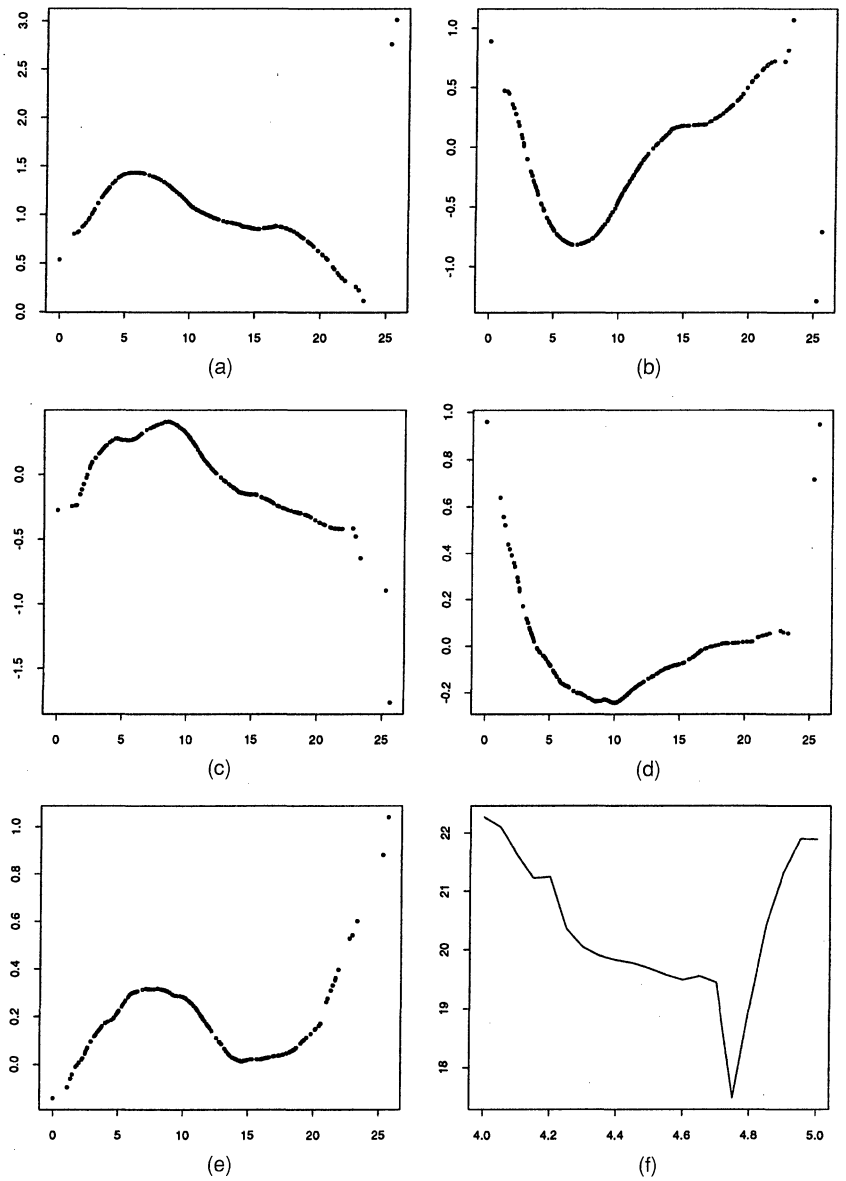

Figure 7. Wolf's Sunspot Numbers. (a)–(e) Estimated functional coefficients in model (23) (a) a_1; (b) a_2; (c) a_3; (d) a_6; (e) a_8. The x-axis is x_{t-3}. (f) The plot of the AMS against bandwidth for estimation of model (23).

To compare the prediction performance, the first 280 data points (in 1700–1979) are used to estimate the coefficient functions in (23). Table 4 reports the absolute errors in predicting the sunspot numbers in 1980–1987 from the newly estimated model (23) as well as those from the FAR model (22) and the following TAR model (Tong 1990, p. 420):

$$x_t = \begin{cases} 1.92 + .84x_{t-1} + .07x_{t-2} - .32x_{t-3} \\ \quad + .15x_{t-4} - .20x_{t-5} - 0x_{t-6} \\ \quad + .19x_{t-7} - .27x_{t-8} + .21x_{t-9} \\ \quad + .01x_{t-10} + .09x_{t-11} + \varepsilon_t^{(1)}, \\ \qquad\qquad\qquad\qquad \text{if } x_{t-8} \le 11.93 \\ 4.27 + 1.44x_{t-1} - .84x_{t-2} + .06x_{t-3} + \varepsilon_t^{(2)}, \\ \qquad\qquad\qquad\qquad \text{if } x_{t-8} > 11.93. \end{cases} \tag{24}$$

Note that both models (22) and (24) also were estimated using the first 280 sample points (Chen and Tsay 1993, p. 304; Tong 1990, p. 420). According to the AAPEs, the functional-coefficient model performs as well as both the TAR and FAR models in one-step-ahead prediction. Furthermore, it performs better in two-step-ahead prediction with both iterative and direct methods.

Finally, we apply the goodness-of-fit technique to test the hypothesis of the FAR model (22) against the nonparametric model (23). The RSS_1 for model (23) is 2.932, in contrast to $\mathrm{RSS}_0 = 3.277$ for the FAR model. The p value for this test is .454, which lends further support to using the FAR model in this context. We also test the hypothesis of the TAR model (24) against the functional-coefficient model

$$x_t = \sum_{j=1}^{11} a_j(x_{t-8})x_{t-j} + \varepsilon_t. \tag{25}$$

The RSS_1 for (25) is 2.077, which is about 43.64% smaller

Table 4. The Postsample Predictive Errors for the Sunspot Data

		Model (23)			FAR model (22)		TAR model (24)	
Year	x_t	One-step	Iterative	Direct	Error	Iterative	Error	Iterative
1980	154.7	1.4	1.4	1.4	13.8	13.8	5.5	5.5
1981	140.5	11.4	10.4	3.7	0	3.8	1.3	0
1982	115.9	15.7	20.7	12.9	10.0	16.4	19.5	22.1
1983	66.6	10.3	.7	11.0	3.3	.8	4.8	6.5
1984	45.9	1.0	1.5	4.3	3.8	5.6	14.8	15.9
1985	17.9	2.6	3.4	7.8	4.6	1.7	.2	2.7
1986	13.4	3.1	.7	1.9	1.3	2.5	5.5	5.4
1987	29.2	12.3	13.1	18.9	21.7	23.6	.7	17.5
AAPE		7.2	6.5	7.7	7.3	8.3	6.6	9.5

than $\mathrm{RSS}_0 = 3.685$ for the TAR model. The p value of the test is .101.

6. ASYMPTOTIC RESULTS

Let $\mathcal{F}_a^b$ be the σ algebra generated by $\{(U_j, \mathbf{X}_j, Y_j);\ a \le j \le b\}$. Let

$$\alpha(k) = \sup\{|P(A\cap B) - P(A)P(B)|:\ A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty\}.$$

The quantity $\alpha(k)$ is called the mixing coefficient of the stationary process $\{U_j, \mathbf{X}_j, Y_j\}_{j=-\infty}^{\infty}$. If $\alpha(k) \to 0$ as $k \to \infty$, then the process $\{U_j, \mathbf{X}_j, Y_j\}_{j=-\infty}^{\infty}$ is called α-mixing.

Among various mixing conditions used in literature, α-mixing is reasonably weak and is known to be fulfilled for many stochastic processes, including many time series models. Gorodetskii (1977) and Withers (1981) derived the conditions under which a linear process is α-mixing. In fact, under very mild assumptions, linear AR and more generally bilinear time series models are strongly mixing, with mixing coefficients decaying exponentially. Auestad and Tjøstheim (1990) provided illuminating discussions on the role of α-mixing (including geometric ergodicity) for model identification in nonlinear time series analysis. Chen and Tsay (1993) showed that the FAR process defined in (3) is geometrically ergodic under certain conditions. Further, Masry and Tjøstheim (1995, 1997) showed that under some mild conditions, both ARCH processes and additive AR processes with exogenous variables (NAARX), which are particularly popular in finance, are stationary and α-mixing.

We first present a result on mean squared convergence that serves as a building block for our main result and is also of independent interest. We now introduce some notation. Let

$$\mathbf{S}_n = \mathbf{S}_n(u_0) = \begin{pmatrix} \mathbf{S}_{n,0} & \mathbf{S}_{n,1} \\ \mathbf{S}_{n,1} & \mathbf{S}_{n,2} \end{pmatrix}$$

and

$$\mathbf{T}_n = \mathbf{T}_n(u_0) = \begin{pmatrix} \mathbf{T}_{n,0}(u_0) \\ \mathbf{T}_{n,1}(u_0) \end{pmatrix}$$

with

$$\mathbf{S}_{n,j} = \mathbf{S}_{n,j}(u_0) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \left(\frac{U_i - u_0}{h}\right)^j K_h(U_i - u_0)$$

and

$$\mathbf{T}_{n,j}(u_0) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \left(\frac{U_i - u_0}{h}\right)^j K_h(U_i - u_0) Y_i. \quad (26)$$

Then, the solution to (9) can be expressed as

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^{-1}\mathbf{S}_n^{-1}\mathbf{T}_n, \quad (27)$$

where $\mathbf{H} = \mathrm{diag}(1, \ldots, 1, h, \ldots, h)$ with p-diagonal elements 1's and p diagonal elements h's. To facilitate the notation, we denote

$$\mu_j = \int_{-\infty}^{\infty} u^j K(u)\, du, \qquad \nu_j = \int_{-\infty}^{\infty} u^j K^2(u)\, du,$$

and

$$\boldsymbol{\Omega} = (\omega_{l,m})_{p\times p} = E(\mathbf{X}\mathbf{X}^T | U = u_0). \quad (28)$$

Also, let $f(u, \mathbf{x})$ denote the joint density of $(U, \mathbf{X})$ and $f_U(u)$ be the marginal density of U. We use the following convention: if $U = X_{j_0}$ for some $1 \le j_0 \le p$, then $f(u, \mathbf{x})$ becomes $f(\mathbf{x})$ the joint density of $\mathbf{X}$.

Theorem 1. Let condition A.1 in the Appendix hold, and let $f(u, \mathbf{x})$ be continuous at the point u_0. Let $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$. Then it holds that

$$E(\mathbf{S}_{n,j}(u_0)) \to f_U(u_0)\boldsymbol{\Omega}(u_0)\mu_j,$$

and

$$nh_n \mathrm{var}(\mathbf{S}_{n,j}(u_0)_{l,m}) \to f_U(u_0)\nu_{2j}\omega_{l,m}$$

for each $0 \le j \le 3$ and $1 \le l, m \le p$.

As a consequence of Theorem 1, we have

$$\mathbf{S}_n \xrightarrow{\mathcal{P}} f_U(u_0)\mathbf{S}$$

and

$$\mathbf{S}_{n,3} \xrightarrow{\mathcal{P}} \mu_3 f_U(u_0)\boldsymbol{\Omega}$$

in the sense that each element converges in probability, where

$$\mathbf{S} = \begin{pmatrix} \boldsymbol{\Omega} & \mu_1\boldsymbol{\Omega} \\ \mu_1\boldsymbol{\Omega} & \mu_2\boldsymbol{\Omega} \end{pmatrix}.$$

Put

$$\sigma^2(u, \mathbf{x}) = \mathrm{var}(Y | U = u, \mathbf{X} = \mathbf{x}) \quad (29)$$

and

$$\mathbf{\Omega}^*(u_0) = E[\mathbf{X}\mathbf{X}^T\sigma^2(U,\mathbf{X})|U=u_0]. \quad (30)$$

Let $c_0 = \mu_2/(\mu_2-\mu_1^2)$ and $c_1 = -\mu_1/(\mu_2-\mu_1^2)$.

Theorem 2. Let $\sigma^2(u,\mathbf{x})$ and $f(u,\mathbf{x})$ be continuous at the point u_0. Then, under conditions A.1 and A.2 in the Appendix,

$$\sqrt{nh_n}\left[\hat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) - \frac{h^2}{2}\frac{\mu_2^2-\mu_1\mu_3}{\mu_2-\mu_1^2}\mathbf{a}''(u_0)\right] \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \mathbf{\Theta}^2(u_0)), \quad (31)$$

provided that $f_U(u_0) \neq 0$, where

$$\mathbf{\Theta}^2(u_0) = \frac{c_0^2\nu_0 + 2c_0c_1\nu_1 + c_1^2\nu_2}{f_U(u_0)} \times \mathbf{\Omega}^{-1}(u_0)\mathbf{\Omega}^*(u_0)\mathbf{\Omega}^{-1}(u_0). \quad (32)$$

Theorem 2 indicates that the asymptotic bias of $\hat{a}_j(u_0)$ is

$$\frac{h^2}{2}\frac{\mu_2^2-\mu_1\mu_3}{\mu_2-\mu_1^2}a_j''(u_0)$$

and the asymptotic variance is $(nh_n)^{-1}\theta_j^2(u_0)$, where

$$\theta_j^2(u_0) = \frac{c_0^2\nu_0 + 2c_0c_1\nu_1 + c_1^2\nu_2}{f_U(u_0)} \times \mathbf{e}_{j,p}^T\mathbf{\Omega}^{-1}(u_0)\mathbf{\Omega}^*(u_0)\mathbf{\Omega}^{-1}(u_0)\mathbf{e}_{j,p}.$$

When $\mu_1 = 0$, the bias and variance expressions can be simplified as $(h^2/2)\mu_2 a_j''(u_0)$ and

$$\theta_j^2(u_0) = \frac{\nu_0}{f_U(u_0)}\mathbf{e}_{j,p}^T\mathbf{\Omega}^{-1}(u_0)\mathbf{\Omega}^*(u_0)\mathbf{\Omega}^{-1}(u_0)\mathbf{e}_{j,p}.$$

The optimal bandwidth for estimating $a_j(\cdot)$ can be defined to be the one that minimizes the squared bias plus variance. The optimal bandwidth is given by

$$h_{j,\mathrm{opt}} = \left[\frac{\mu_2^2\nu_0 - 2\mu_1\mu_2\nu_1 + \mu_1^2\nu_2}{f_U(u_0)(\mu_2^2-\mu_1\mu_3)^2} \times \frac{\mathbf{e}_{j,p}^T\mathbf{\Omega}^{-1}(u_0)\mathbf{\Omega}^*(u_0)\mathbf{\Omega}^{-1}(u_0)\mathbf{e}_{j,p}}{\{a_j''(u_0)\}^2}\right]^{1/5} n^{-1/5}. \quad (33)$$

Recently, Fan and Gijbels (1995) and Ruppert, Sheather, and Wand (1995) developed data-driven bandwidth selection schemes based on asymptotic formulas for the optimal bandwidths, which are less variable and more effective than the conventional data-driven bandwidth selectors such as the cross-validation bandwidth rule. Similar algorithms can be developed for the estimation of functional-coefficient models based on (35); however, this is beyond the scope of this article.

APPENDIX: CONDITIONS AND PROOFS

We first impose some conditions on the regression model. They are not the weakest possible.

Condition A.1

a. The kernel function $K(\cdot)$ is a bounded density with a bounded support $[-1,1]$.

b. $|f(u,v|\mathbf{x}_0,\mathbf{x}_1;l)| \leq M < \infty$, for all $l \geq 1$, where $f(u,v,|\mathbf{x}_0,\mathbf{x}_1;l)$ is the conditional density of $(U_0,U_l))$ given $(\mathbf{X}_0,\mathbf{X}_l)$, and $f(u|\mathbf{x}) \leq M < \infty$, where $f(u|\mathbf{x})$ is the conditional density of U given $\mathbf{X}=\mathbf{x}$.

c. The process $\{U_i,\mathbf{X}_i,Y_i\}$ is α-mixing with $\sum k^c[\alpha(k)]^{1-2/\delta} < \infty$ for some $\delta > 2$ and $c > 1-2/\delta$.

d. $E|\mathbf{X}|^{2\delta} < \infty$, where δ is given in condition A.1c.

Condition A.2

a. Assume that

$$E\{Y_0^2 + Y_l^2|U_0=u,\mathbf{X}_0=\mathbf{x}_0;U_l=v,\mathbf{X}_l=\mathbf{x}_1\} \leq M < \infty, \quad (A.1)$$

for all $l \geq 1, \mathbf{x}_0,\mathbf{x}_1 \in \Re^p, u$, and v in a neighborhood of u_0.

b. Assume that $h_n \to$ and $nh_n \to \infty$. Further, assume that there exists a sequence of positive integers s_n such that $s_n \to \infty, s_n = o((nh_n)^{1/2})$, and $(n/h_n)^{1/2}\alpha(s_n) \to 0$, as $n \to \infty$.

c. There exists $\delta^* > \delta$, where δ is given in condition A.1c, such that

$$E\{|Y|^{\delta^*}|U=u,\mathbf{X}=\mathbf{x}\} \leq M_4 < \infty \quad (A.2)$$

for all $\mathbf{x} \in \Re^p$ and u in a neighborhood of u_0, and

$$\alpha(n) = O(n^{-\theta^*}), \quad (A.3)$$

where $\theta^* \geq \delta\delta^*/\{2(\delta^*-\delta)\}$.

d. $E|\mathbf{X}|^{2\delta^*} < \infty$, and $n^{1/2-\delta/4}h^{\delta/\delta^*-1/2-\delta/4} = O(1)$.

Remark A.1. We provide a sufficient condition for the mixing coefficient $\alpha(n)$ to satisfy conditions 1(c) and 2(b). Suppose that $h_n = An^{-\rho}(0<\rho<1, A>0), s_n = (nh_n/\log n)^{1/2}$ and $\alpha(n) = O(n^{-d})$ for some $d > 0$. Then condition A.1c is satisfied for $d > 2(1-1/\delta)/(1-2/\delta)$ and condition A.2b is satisfied if $d > (1+\rho)/(1-\rho)$. Hence both conditions are satisfied if

$$\alpha(n) = O(n^{-d}), \qquad d > \max\left\{\frac{1+\rho}{1-\rho}, \frac{2(1-1/\delta)}{1-2/\delta}\right\}.$$

Note that this is a trade-off between the order δ of the moment of Y and the rate of decay of the mixing coefficient; the larger the order δ, the weaker the decay rate of $\alpha(n)$.

To study the joint asymptotic normality of $\hat{\mathbf{a}}(u_0)$, we need to center the vector $\mathbf{T}_n(u_0)$ by replacing Y_i with $Y_i - m(U_i,\mathbf{X}_i)$ in the expression (28) of $\mathbf{T}_{n,j}(u_0)$. Let

$$\mathbf{T}_{n,j}^*(u_0) = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\left(\frac{U_i-u_0}{h}\right)^j K_h(U_i-u_0)[Y_i - m(U_i,\mathbf{X}_i)],$$

and

$$\mathbf{T}_n^* = \begin{pmatrix}\mathbf{T}_{n,0}^* \\ \mathbf{T}_{n,1}^*\end{pmatrix}.$$

Because the coefficient functions $a_j(u)$ are conducted in the neighborhood of $|U_i-u_0|<h$, by Taylor's expansion,

$$m(U_i,\mathbf{X}_i) = \mathbf{X}_i^T\mathbf{a}(u_0) + (U_i-u_0)\mathbf{X}_i^T\mathbf{a}'(u_0) + \frac{h^2}{2}\left(\frac{U_i-u_0}{h}\right)^2\mathbf{X}_i^T\mathbf{a}''(u_0) + o_p(h^2),$$

where $\mathbf{a}'(u_0)$ and $\mathbf{a}''(u_0)$ are the vectors consisting of the first and second derivatives of the functions $a_j(\cdot)$. Then,

$$\mathbf{T}_{n,0}-\mathbf{T}^*_{n,0}=\mathbf{S}_{n,0}\mathbf{a}(u_0)+h\mathbf{S}_{n,1}\mathbf{a}'(u_0)+\frac{h^2}{2}\mathbf{S}_{n,2}\mathbf{a}''(u_0)+o_p(h^2)$$

and

$$\mathbf{T}_{n,1}-\mathbf{T}^*_{n,1}=\mathbf{S}_{n,1}\mathbf{a}(u_0)+h\mathbf{S}_{n,2}\mathbf{a}'(u_0)+\frac{h^2}{2}\mathbf{S}_{n,3}\mathbf{a}''(u_0)+o_p(h^2),$$

so that

$$\mathbf{T}_n-\mathbf{T}^*_n=\mathbf{S}_n\mathbf{H}\boldsymbol{\beta}+\frac{h^2}{2}\begin{pmatrix}\mathbf{S}_{n,2}\\ \mathbf{S}_{n,3}\end{pmatrix}\mathbf{a}''(u_0)+o_p(h^2), \quad (A.4)$$

where $\boldsymbol{\beta}=(\mathbf{a}(u_0)^T,\mathbf{a}'(u_0)^T)^T$. Thus it follows from (29), (A.4), and Theorem 1 that

$$\mathbf{H}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=f_U^{-1}(u_0)\mathbf{S}^{-1}\mathbf{T}^*_n+\frac{h^2}{2}\mathbf{S}^{-1}\begin{pmatrix}\mu_2\boldsymbol{\Omega}\\ \mu_3\boldsymbol{\Omega}\end{pmatrix}\mathbf{a}''(u_0)+o_p(h^2), \quad (A.5)$$

from which the bias term of $\hat{\boldsymbol{\beta}}(u_0)$ is evident. Clearly,

$$\hat{\mathbf{a}}(u_0)-\mathbf{a}(u_0)=\frac{\boldsymbol{\Omega}^{-1}}{f_U(u_0)(\mu_2-\mu_1^2)}[\mu_2\mathbf{T}^*_{n,0}-\mu_1\mathbf{T}^*_{n,1}]+\frac{h^2}{2}\frac{\mu_2^2-\mu_1\mu_3}{\mu_2-\mu_1^2}\mathbf{a}''(u_0)+o_p(h^2). \quad (A.6)$$

Thus (A.6) indicates that the asymptotic bias of $\hat{\mathbf{a}}(u_0)$ is

$$\frac{h^2}{2}\frac{\mu_2^2-\mu_1\mu_3}{\mu_2-\mu_1^2}\mathbf{a}''(u_0).$$

Let

$$\mathbf{Q}_n=\frac{1}{n}\sum_{i=1}^n\mathbf{Z}_i, \quad (A.7)$$

where

$$\mathbf{Z}_i=\mathbf{X}_i\left[c_0+c_1\left(\frac{U_i-u_0}{h}\right)\right]K_h(U_i-u_0)[Y_i-m(U_i,\mathbf{X}_i)] \quad (A.8)$$

with $c_0=\mu_2/(\mu_2-\mu_1^2)$ and $c_1=-\mu_1/(\mu_2-\mu_1^2)$. It follows from (A.6) and (A.7) that

$$\sqrt{nh_n}\left[\hat{\mathbf{a}}(u_0)-\mathbf{a}(u_0)-\frac{h^2}{2}\frac{\mu_2^2-\mu_1\mu_3}{\mu_2-\mu_1^2}\mathbf{a}''(u_0)\right]=\frac{\boldsymbol{\Omega}^{-1}}{f_U(u_0)}\sqrt{nh_n}\mathbf{Q}_n+o_p(1). \quad (A.9)$$

We need the following lemma, whose proof is more involved than that for Theorem 1. Therefore, we prove only this lemma. Throughout this Appendix, we let C denote a generic constant, which may take different values at different places.

Lemma A.1 Under conditions A.1 and A.2 and the assumption that $h_n\to 0$ and $nh_n\to\infty$, as $n\to\infty$, if $\sigma^2(u,\mathbf{x})$ and $f(u,\mathbf{x})$ are continuous at the point u_0, then we have

(a) $h_n\ \mathrm{var}(\mathbf{Z}_1)\to f_U(u_0)\boldsymbol{\Omega}^*(u_0)[c_0^2\nu_0+2c_0c_1\nu_1+c_1^2\nu_2]$;

(b) $h_n\sum_{l=1}^{n-1}|\mathrm{cov}(\mathbf{Z}_1,\mathbf{Z}_{l+1})|=o(1)$; and

(c) $nh_n\mathrm{var}(\mathbf{Q}_n)\to f_U(u_0)\boldsymbol{\Omega}^*(u_0)[c_0^2\nu_0+2c_0c_1\nu_1+c_1^2\nu_2]$.

Proof. First, by conditioning on $(U_1,\mathbf{X}_1)$ and using theorem 1 of Sun (1984), we have

$$\begin{aligned}&\mathrm{var}(\mathbf{Z}_1)\\ &=E\left[\mathbf{X}_1\mathbf{X}_1^T\sigma^2(U_1,\mathbf{X}_1)\left\{c_0+c_1\left(\frac{U_1-u_0}{h}\right)\right\}^2K_h^2(U_1-u_0)\right]\\ &=\frac{1}{h}\left[f_U(u_0)\boldsymbol{\Omega}^*(u_0)\{c_0^2\nu_0+2c_0c_1\nu_1+c_1^2\nu_2\}+o(1)\right]. \quad (A.10)\end{aligned}$$

The result (c) follows in an obvious manner from (a) and (b) along with

$$\mathrm{var}(\mathbf{Q}_n)=\frac{1}{n}\mathrm{var}(\mathbf{Z}_1)+\frac{2}{n}\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\mathrm{cov}(\mathbf{Z}_1,\mathbf{Z}_{l+1}). \quad (A.11)$$

It thus remains to prove part (b). To this end, let $d_n\to\infty$ be a sequence of positive integers such that $d_nh_n\to 0$. Define

$$J_1=\sum_{l=1}^{d_n-1}|\mathrm{cov}(\mathbf{Z}_1,\mathbf{Z}_{l+1})|$$

and

$$J_2=\sum_{l=d_n}^{n-1}|\mathrm{cov}(\mathbf{Z}_1,\mathbf{Z}_{l+1})|.$$

It remains to show that $J_1=o(h^{-1})$ and $J_2=o(h^{-1})$.

We remark that because $K(\cdot)$ has a bounded support $[-1,1]$, $a_j(u)$ is bounded in the neighborhood of $u\in[u_0-h,u_0+h]$. Let $B=\max_{1\le j\le p}\sup_{|u-u_0|<h}|a_j(u)|$ and $g(\mathbf{x})=\sum_{j=1}^p|x_j|$. Then $\sup_{|u-u_0|<h}|m(u,\mathbf{x})|\le Bg(\mathbf{x})$. By conditioning on $(U_1,\mathbf{X}_1)$ and $(U_{l+1},\mathbf{X}_{l+1})$, and using (A.1) and condition A.1b, we have, for all $l\ge 1$,

$$\begin{aligned}&|\mathrm{cov}(\mathbf{Z}_1,\mathbf{Z}_{l+1})|\\ &\le CE[|\mathbf{X}_1\mathbf{X}_{l+1}^T|\{|Y_1|+Bg(\mathbf{X}_1)\}\{|Y_{l+1}|+Bg(\mathbf{X}_{l+1})\}\\ &\qquad\times K_h(U_1-u_0)K_h(U_{l+1}-u_0)]\\ &\le CE[|\mathbf{X}_1\mathbf{X}_{l+1}^T|\{M_2+B^2g^2(\mathbf{X}_1)\}^{1/2}\{M_2+B^2g^2(\mathbf{X}_{l+1})\}^{1/2}\\ &\qquad\times K_h(U_1-u_0)K_h(U_{l+1}-u_0)]\\ &\le CE[|\mathbf{X}_1\mathbf{X}_{l+1}^T|\{1+g(\mathbf{X}_1)\}\{1+g(\mathbf{X}_{l+1})\}]\\ &\le C. \quad (A.12)\end{aligned}$$

It follows that

$$J_1\le Cd_n=o(h^{-1})$$

by the choice of d_n. We next consider the upper bound of J_2. To this end, using Davydov's inequality (see Hall and Heyde 1980, cor. A.2), we obtain, for all $1\le j,m\le p$ and $l\ge 1$,

$$|\mathrm{cov}(Z_{1j},Z_{l+1,m})|\le C[\alpha(l)]^{1-2/\delta}[E|Z_j|^\delta]^{1/\delta}[E|Z_m|^\delta]^{1/\delta}. \quad (A.13)$$

By conditioning on $(U,\mathbf{X})$ and using conditions A.1b and A.2c, one has

$$\begin{aligned}E[|Z_j|^\delta]&\le CE[|X_j|^\delta K_h^\delta(U-u_0)\{|Y|^\delta+B^\delta g^\delta(\mathbf{X})\}]\\ &\le CE[|X_j|^\delta K_h^\delta(U-u_0)\{M_3+B^\delta g^\delta(\mathbf{X})\}]\\ &\le Ch^{1-\delta}E[|X_j|^\delta\{M_3+B^\delta g^\delta(\mathbf{X})\}]\\ &\le Ch^{1-\delta}. \quad (A.14)\end{aligned}$$

A combination of (A.13) and (A.14) leads to

$$\begin{aligned}J_2&\le Ch^{2/\delta-2}\sum_{l=d_n}^\infty[\alpha(l)]^{1-2/\delta}\\ &\le Ch^{2/\delta-2}d_n^{-c}\sum_{l=d_n}^\infty l^c[\alpha(l)]^{1-2/\delta}=o(h^{-1}) \quad (A.15)\end{aligned}$$

by choosing d_n such that $h^{1-2/\delta}d_n^c=C$, so the requirement that $d_nh_n\to 0$ is satisfied.

Proof of Theorem 2

We use the small-block and large-block technique—namely, partition $\{1, \ldots, n\}$ into $2q_n + 1$ subsets with large block of size $r = r_n$ and small block of size $s = s_n$. Set

$$q = q_n = \left\lfloor \frac{n}{r_n + s_n} \right\rfloor. \tag{A.16}$$

We now use the Cramér–Wold device to derive the asymptotic normality of $\mathbf{Q}_n$. For any unit vector $\mathbf{d} \in \Re^p$, let $Z_{n,i} = \sqrt{h}\mathbf{d}^T\mathbf{Z}_{i+1}, i = 0, \ldots, n-1$. Then

$$\sqrt{nh}\mathbf{d}^T\mathbf{Q}_n = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} Z_{n,i},$$

and, by Lemma A.1,

$$\begin{aligned}\operatorname{var}(Z_{n,0}) &= f_U(u_0)\mathbf{d}^T\mathbf{\Omega}^*(u_0)\mathbf{d}[c_0^2\nu_0 + 2c_0c_1\nu_1 + c_1^2\nu_2](1 + o(1)) \\ &\equiv \theta^2(u_0)(1 + o(1)) \end{aligned} \tag{A.17}$$

and

$$\sum_{l=0}^{n-1} |\operatorname{cov}(Z_{n,0}, Z_{n,l})| = o(1). \tag{A.18}$$

Define the random variables, for $0 \le j \le q-1$,

$$\eta_j = \sum_{i=j(r+s)}^{j(r+s)+r-1} Z_{n,i},$$

$$\xi_j = \sum_{i=j(r+s)+r}^{(j+1)(r+s)} Z_{n,i},$$

and

$$\zeta_q = \sum_{i=q(r+s)}^{n-1} Z_{n,i}.$$

Then,

$$\begin{aligned}\sqrt{nh}\mathbf{d}^T\mathbf{Q}_n &= \frac{1}{\sqrt{n}} \left\{ \sum_{j=0}^{q-1} \eta_j + \sum_{j=0}^{q-1} \xi_j + \zeta_q \right\} \\ &\equiv \frac{1}{\sqrt{n}} \{Q_{n,1} + Q_{n,2} + Q_{n,3}\}. \end{aligned} \tag{A.19}$$

We show that as $n \to \infty$,

$$\frac{1}{n}E[Q_{n,2}]^2 \to 0, \qquad \frac{1}{n}E[Q_{n,3}]^2 \to 0, \tag{A.20}$$

$$\left| E[\exp(itQ_{n,1})] - \prod_{j=0}^{q-1} E[\exp(it\eta_j)] \right| \to 0, \tag{A.21}$$

$$\frac{1}{n} \sum_{j=0}^{q-1} E(\eta_j^2) \to \theta^2(u_0), \tag{A.22}$$

and

$$\frac{1}{n} \sum_{j=0}^{q-1} E[\eta_j^2 I\{|\eta_j| \ge \varepsilon\theta(u_0)\sqrt{n}\}] \to 0 \tag{A.23}$$

for every $\varepsilon > 0$. (A.20) implies that $Q_{n,2}$ and $Q_{n,3}$ are asymptotically negligible in probability, (A.21) shows that the summands η_j in $Q_{n,1}$ are asymptotically independent, and (A.22) and (A.23) are the standard Lindeberg–Feller conditions for asymptotic normality of $Q_{n,1}$ for the independent setup.

We first establish (A.20). For this purpose, we choose the large-block size. Condition A.2b implies that there is a sequence of positive constants $\gamma_n \to \infty$ such that

$$\gamma_n s_n = o(\sqrt{nh_n})$$

and

$$\gamma_n(n/h_n)^{1/2}\alpha(s_n) \to 0. \tag{A.24}$$

Define the large-block size r_n by $r_n = \lfloor (nh_n)^{1/2}/\gamma_n \rfloor$ and the small-block size s_n. Then it can easily be shown from (A.24) that as $n \to \infty$,

$$s_n/r_n \to 0, \qquad r_n/n \to 0, \qquad r_n(nh_n)^{-1/2} \to 0, \tag{A.25}$$

and

$$(n/r_n)\alpha(s_n) \to 0. \tag{A.26}$$

Observe that

$$\begin{aligned}E[Q_{n,2}]^2 &= \sum_{j=0}^{q-1} \operatorname{var}(\xi_j) + 2 \sum_{0 \le i < j \le q-1} \operatorname{cov}(\xi_i, \xi_j) \\ &\equiv I_1 + I_2. \end{aligned} \tag{A.27}$$

It follows from stationarity and Lemma A.1 that

$$\begin{aligned}I_1 &= q_n \operatorname{var}(\xi_1) = q_n \operatorname{var}\left(\sum_{j=1}^{s_n} Z_{n,j} \right) \\ &= q_n s_n [\theta^2(u_0) + o(1)]. \end{aligned} \tag{A.28}$$

Next consider the second term I_2 in the right side of (A.27). Let $r_j^* = j(r_n + s_n)$, then $r_j^* - r_i^* \ge r_n$ for all $j > i$, we thus have

$$\begin{aligned}|I_2| &\le 2 \sum_{0 \le i < j \le q-1} \sum_{j_1=1}^{s_n} \sum_{j_2=1}^{s_n} |\operatorname{cov}(Z_{n, r_i^* + r_n + j_1}, Z_{n, r_j^* + r_n + j_2})| \\ &\le 2 \sum_{j_1=1}^{n-r_n} \sum_{j_2=j_1+r_n}^{n} |\operatorname{cov}(Z_{n,j_1}, Z_{n,j_2})|. \end{aligned}$$

By stationarity and Lemma A.1, one obtains

$$|I_2| \le 2n \sum_{j=r_n+1}^{n} |\operatorname{cov}(Z_{n,1}, Z_{n,j})| = o(n). \tag{A.29}$$

Hence, by (A.25)–(A.29), we have

$$\frac{1}{n}E[Q_{n,2}]^2 = O(q_n s_n n^{-1}) + o(1) = o(1). \tag{A.30}$$

It follows from stationarity, (A.25), and Lemma A.1 that

$$\begin{aligned}\operatorname{var}[Q_{n,3}] &= \operatorname{var}\left(\sum_{j=1}^{n - q_n(r_n + s_n)} Z_{n,j} \right) \\ &= O(n - q_n(r_n + s_n)) = o(n). \end{aligned} \tag{A.31}$$

Combining (A.25), (A.30), and (A.31), we establish (A.20). As for (A.22), by stationarity, (A.25), (A.26), and Lemma A.1, it is easily seen that

$$\frac{1}{n} \sum_{j=0}^{q_n-1} E(\eta_j^2) = \frac{q_n}{n}E(\eta_1^2) = \frac{q_n r_n}{n} \cdot \frac{1}{r_n} \operatorname{var}\left(\sum_{j=1}^{r_n} Z_{n,j} \right) \to \theta^2(u_0).$$

To establish (A.21), we use Lemma 1.1 of Volkonskii and Rozanov (1959) (see also Ibragimov and Linnik 1971, p. 338) to obtain

$$\left| E[\exp(itQ_{n,1})] - \prod_{j=0}^{q_n-1} E[\exp(it\eta_j)] \right| \le 16(n/r_n)\alpha(s_n)$$

tending to 0 by (A.26).

It remains to establish (A.23). For this purpose, we use theorem 4.1 of Shao and Yu (1996) and condition 2 to obtain

$$E[\eta_1^2 I\{|\eta_1| \geq \varepsilon\theta(u_0)\sqrt{n}\}]$$

$$\leq Cn^{1-\delta/2}E(|\eta_1|^{\delta}) \leq Cn^{1-\delta/2}r_n^{\delta/2}\{E(|Z_{n,0}|^{\delta^*})\}^{\delta/\delta^*}. \quad \text{(A.32)}$$

As in (A.14),

$$E(|Z_{n,0}|^{\delta^*}) \leq Ch^{1-\delta^*/2}. \quad \text{(A.33)}$$

Therefore, by (A.32) and (A.33),

$$E[\eta_1^2 I\{|\eta_1| \geq \varepsilon\theta(u_0)\sqrt{n}\}] \leq Cn^{1-\delta/2}r_n^{\delta/2}h^{(2-\delta^*)\delta/(2\delta^*)}.$$

Thus, by (A.16) and the definition of r_n, and using conditions A.2c and A.2d, we obtain

$$\frac{1}{n}\sum_{j=0}^{q-1} E[\eta_j^2 I\{|\eta_j| \geq \varepsilon\theta(u_0)\sqrt{n}\}]$$

$$\leq C\gamma_n^{1-\delta/2}n^{1/2-\delta/4}h_n^{\delta/\delta^*-1/2-\delta/4} \to 0$$

because $\gamma_n \to \infty$. This completes the proof of the theorem.

[Received April 1998. Revised November 1999.]

REFERENCES

An, H. Z., and Chen, S. G. (1997), "A Note on the Ergodicity of Nonlinear Autoregressive Models," *Statistics and Probability Letters*, 34, 365–372.

An, H. Z., and Huang, F. C. (1996), "The Geometrical Ergodicity of Nonlinear Autoregressive Models," *Statistica Sinica*, 6, 943–956.

Auestad, B., and Tjøstheim, D. (1990), "Identification of Nonlinear Time Series: First-Order Characterization and Order Determination," *Biometrika*, 77, 669–687.

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroscedasticity," *Journal of Econometrics*, 31, 307–327.

Box, G. E. P., and Jenkins, G. M. (1970), *Time Series Analysis, Forecasting, and Control*, San Francisco: Holden Day.

Brumback, B., and Rice, J. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves" (with discussion), *Journal of the American Statistical Association*, 93, 961–976.

Chen, R., and Tsay, R. S. (1993), "Functional-Coefficient Autoregressive Models," *Journal of the American Statistical Association*, 88, 298–308.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), "Local Regression Models," in *Statistical Models in S*, eds. J. M. Chambers and T. J. Hastie, Pacific Grove, CA: Wadsworth & Brooks/Cole, pp. 309–376.

Dahlhaus, R. (1989), "Efficient Parameter Estimation for Self-Similar Processes," *The Annals of Statistics*, 17, 1749–1766.

Engle, R. F. (1982), "Autoregressive Conditional Heteroscedasticity With Estimates of the Variance of U.K. Inflation," *Econometrica*, 50, 987–1008.

Engle, R. F., and Granger, C. W. J. (1987), "Cointegration and Error Correction: Representation, Estimation and Testing," *Econometrica*, 55, 251–276.

Fan, J. (1993), "Local Linear Regression Smoothers and Their Minimax," *The Annals of Statistics*, 21, 196–216.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth Spatial Adaptation," *Journal of the Royal Statistical Society*, Ser. B, 57, 371–394.

——— (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman and Hall.

Fan, J., Yao, Q., and Cai, Z. (2000), "Adaptive Varying-Coefficient Linear Models," unpublished manuscript.

Fan, J., Zhang, C., and Zhang, J. (1999), "Sieve Likelihood Ratio Statistics and Wilks Phenomenon," technical report, Department of Statistics, University of California at Los Angeles.

Fan, J., and Zhang, W. (1999), "Statistical Estimation in Varying-Coefficient Models," *The Annals of Statistics*, 27, 1491–1518.

Ghaddar, D. K., and Tong, H. (1981), "Data Transformation and Self-Exciting Threshold Autoregression," *Journal of the Royal Statistical Society*, Ser. C, 30, 238–248.

Gorodetskii, V. V. (1977), "On the Strong Mixing Property for Linear Sequences," *Theory of Probability and Its Applications*, 22, 411–413.

Granger, C. W. J., and Joyeux, R. (1980), "An Introduction to Long-Memory Time Series Models and Fractional Differencing," *Journal of Time Series Analysis*, 1, 15–29.

Granger, C. W. J., and Teräsvirta, T. (1993), *Modeling Nonlinear Economic Relationships*, Oxford, U.K.: Oxford University Press.

Haggan, V., and Ozaki, T. (1981), "Modeling Nonlinear Vibrations Using an Amplitude-Dependent Autoregressive Time Series Model," *Biometrika*, 68, 189–196.

Hall, P., and Heyde, C. C. (1980), *Martingale Limit Theory and Its Applications*, New York: Academic Press.

Hannan, E. J., and Deistler, M. (1988), *The Statistical Theory of Linear Systems*, New York: Wiley.

Härdle, W., Lütkepohl, H., and Chen, R. (1997), "A Review of Nonparametric Time Series Analysis," *International Statistical Review*, 65, 49–72.

Hastie, T. J., and Loader, C. (1993), "Local Regression: Automatic Kernel Carpentry" (with discussion), *Statistical Science*, 8, 120–143.

Hastie, T. J., and Tibshirani, R. J. (1993), "Varying-Coefficient Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 757–796.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809–822.

Ibragimov, I. A., and Linnik, Yu. V. (1971), *Independent and Stationary Sequences of Random Variables*, Groningen, the Netherlands: Walters-Noordhoff.

Kreiss, J. P., Neumann, M., and Yao, Q. (1998), "Bootstrap Tests for Simple Structures in Nonparametric Time Series Regression," unpublished manuscript.

Masry, E., and Fan, J. (1997), "Local Polynomial Estimation of Regression Functions for Mixing Processes," *Scandinavian Journal of Statistics*, 24, 165–179.

Masry, E., and Tjøstheim, D. (1995), "Nonparametric Estimation and Identification of Nonlinear ARCH Time Series: Strong Convergence and Asymptotic Normality," *Econometric Theory*, 11, 258–289.

——— (1997), "Additive Nonlinear ARX Time Series and Projection Estimates," *Econometric Theory*, 13, 214–252.

Nicholls, D. F., and Quinn, B. G. (1982), *Random Coefficient Autoregressive Models: An Introduction* (Lecture Notes in Statistics 11), New York: Springer-Verlag.

Ozaki, T. (1982), "The Statistical Analysis of Perturbed Limit Cycle Processes Using Nonlinear Time Series Models," *Journal of Time Series Analysis*, 3, 29–41.

Ramsay, J. O., and Silverman, B. W. (1997), *Functional Data Analysis*, Berlin: Springer-Verlag.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selection for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Ruppert, D., and Wand, M. P. (1994), "Multivariate Weighted Least Squares Estimation," *The Annals of Statistics*, 22, 1346–1370.

Shao, Q., and Yu, H. (1996), "Weak Convergence for Weighted Empirical Processes of Dependent Sequences," *The Annals of Probability*, 24, 2098–2127.

Stenseth, N. C., Falck, W., Chan, K. S., Bjørnstad, O. N., O'Donoghue, M., Tong, H., Boonstra, R., Boutin, S., Krebs, C. J., and Yoccoz, N. G. (1998), "From Ecological Patterns to Ecological Processes: Phase- and Density-Dependencies in Canadian Lynx Cycle," *Proceedings of National Academy of Science, Washington*, 95, 15430–15435.

Sun, Z. (1984), "Asymptotic Unbiased and Strong Consistency for Density Function Estimator," *Acta Mathematica Sinica*, 27, 769–782.

Tiao, G. C., and Tsay, R. S. (1994), "Some Advances in Nonlinear and Adaptive Modeling in Time Series," *Journal of Forecasting*, 13, 109–131.

Tjøstheim, D. (1994), "Non-Linear Time Series: A Selective Review," *Scandinavian Journal of Statistics*, 21, 97–130.

Tong, H. (1990), *Nonlinear Time Series: A Dynamical System Approach*, Oxford, U.K.: Oxford University Press.

——— (1995), "A Personal Overview of Non-Linear Time Series Analysis From a Chaos Perspective" (with discussion), *Scandinavian Journal of Statistics*, 22, 399–445.

Volkonskii, V. A., and Rozanov, Yu. A. (1959), "Some Limit Theorems for Random Functions. I," *Theory of Probability and Its Applications*, 4, 178–197.

Withers, C. S. (1981), "Conditions for Linear Processes to be Strong Mixing," *Zeitschrift fur Wahrscheinlichkeitstheorie Verwandte Gebiete*, 57, 477–480.

Yao, Q., and Tong, H. (1995), "On Initial-Condition Sensitivity and Prediction in Nonlinear Stochastic Systems," *Bulletin of the International Statistical Institute*, IP 10.3, 395–412.