Nonparametric Transfer Function Models

Jun M. Liu[1], Rong Chen[2] and Qiwei Yao[3]
[1]Georgia Southern University,
[2,3]Peking University
[2]University of Illinois at Chicago,
[3]London School of Economics [1]

Abstract

In this paper a class of nonparametric transfer function models is proposed to model nonlinear relationships between 'input' and 'output' time series. In this approach, the functional form of the transfer function is assumed to be unknown but smooth, and the noise is assumed to be stationary with a parametric autoregressive-moving average (ARMA) form. A new method is developed to jointly estimate the transfer function nonparametrically and the ARMA parameters parametrically. By modeling the transfer function nonparametrically, the model is flexible and can be used to model nonlinear relationship of unknown functional forms; by modeling the noise explicitly as a parsimonious ARMA model, the correlation in the data is removed so the transfer function can be estimated more efficiently. Additionally, the estimated ARMA parameters can be used to improve the forecasting performance. Estimation procedures are introduced and the asymptotic properties of the estimators are investigated. The finite-sample properties of the estimators are studied through simulations and one real example.

JEL Classification: C14, C22
Keywords: Nonparametric smoothing, Time series, Transfer Function

[1]Jun M. Liu is Assistant Professor of Quantitative Methods, Department of Finance & Quantitative Analysis, Georgia Southern University. Rong Chen is Professor of Statistics, Department of Business Statistics and Econometrics, Peking University and Department of Information & Decision Sciences, University of Illinois at Chicago. Qiwei Yao is Professor of Statistics, London School of Economics and Department of Business Statistics and Econometrics, Peking University. Corresponding author: Rong Chen, 601 South Morgan Street (M/C 294), Chicago, IL 60607, USA. Tel: (312)996-2323, Fax: (312)413-0385, Email: rongchen@uic.edu.

1 Introduction

Linear transfer function models (Box and Jenkins, 1976) have been extensively used to model the relationship between one 'output' time series and several 'input' time series. With one input series, it assumes the form $Y_t = \alpha(B)\beta(B)^{-1}X_t + e_t$, where Y_t is the observed output series of interest, X_t is an observed input time series, e_t follows an ARMA process, and $\alpha(B)$ and $\beta(B)$ are polynomials of the *backshift operator* B defined as $B^iX_t \equiv X_{t-i}$. Linear transfer function models have been well studied and proven successful in many fields (e.g., Newbold, 1973; Tiao and Box, 1981; Tsay, 1985; Poskitt, 1989; Liu and Hanssens, 1982). However, its linear nature limits its applicability because many nonlinear features encountered in practice cannot be well approximated by linear models. To model nonlinear relationships between time series, Chen and Tsay (1996) proposed the *nonlinear transfer function model* of the form $Y_t = f(X_{t-d}, \cdots, X_{t-d-p}; \boldsymbol{\theta}) + \varepsilon_t$, where $f(\cdot)$ is a parametric function assuming the Volterra series representation, ε_t is stationary and modeled by an ARMA model.

There are infinitely many candidate nonlinear functions beyond the linear domain. Therefore, it is usually difficult to justify the explicit parametric functional forms a priori for nonlinear models. Following the "letting the data speak for themselves" principle, nonparametric smoothing methods provide a more flexible alternative to model nonlinear time series (e.g., Robinson, 1983; Auestad and Tjøstheim, 1990; Lewis and Stevens, 1991; Masry, 1996a,b; Fan and Gilbels, 1996; Smith, Wong, and Kohn, 1998). To overcome the 'curse of dimensionality', various specially structured nonparametric models have been proposed, including the *functional-coefficient autoregressive (FAR) model* (Chen and Tsay, 1993a; Cai, Fan and Yao, 2000), the *nonlinear additive autoregressive model* (Chen and Tsay, 1993b), the *adaptive functional-coefficient model* (Ichimura, 1993; Xia and Li, 1999; Fan, Yao and Cai, 2003), the *single index model* (e.g., Härdle, Hall, and Ichimura, 1993; Carroll, Fan, Gijbels, and Wand, 1997; Newey and Stoker, 1993; Heckman, Ichimura, Smith, and Todd, 1998; Xia, Tong, Li, and Zhu, 2002) and the *partially linear models* (Härdle, Liang and Gao, 2000). There is vast literature about nonlinear and nonparametric time series analysis. Some reviews can be found in Tjøstheim (1994), Härdle, Lütkepohl and Chen (1997) and Fan and Yao (2003).

In this paper a class of nonparametric transfer function models is proposed. Consider the model

$$Y_t = f(X_t) + e_t, \tag{1}$$

where $f(\cdot)$ is an unknown and smooth function, $\{X_t\}$ and $\{e_t\}$ are strictly stationary processes. The transfer function $f(\cdot)$ is modeled via nonparametric smoothing and the innovation process $\{e_t\}$ is assumed to follow a stationary and invertible ARMA(p, q) process, i.e., $\phi(B)e_t = \theta(B)\varepsilon_t$, where $\phi(B) = 1 - \sum_{i=1}^{p}\phi_iB^i$, $\theta(B) = 1 - \sum_{j=1}^{q}\theta_jB^j$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_p)^\tau$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_q)^\tau$ are unknown parameters and $\{\varepsilon_t\}$ is a sequence of independent $(0, \sigma^2)$ random variables. An iterative

procedure is used to estimate both the transfer function and the ARMA parameters. Because of its close connections to the Box-Jenkins transfer function model and nonparametric smoothing, the proposed method is named *nonparametric transfer function model*. $\{X_t\}$ and $\{\varepsilon_t\}$ are assumed to be independent, which implies the independence between $\{Y_t\}$ and $\{e_t\}$.

By modeling the transfer function $f(\cdot)$ nonparametrically, the model is flexible therefore can be used to model nonlinear relationship of unknown functional forms. By modeling $\{e_t\}$ as an ARMA(p, q) process, the autocorrelation in the data is removed so $f(\cdot)$ can be estimated more efficiently. Additionally, the explicit correlation structure can be used to improve the forecasting performance.

The problem of estimating $f(\cdot)$ in (1) can be viewed as a regression with correlated noise problem. Under certain mixing conditions, the *windowing-and-whitening* effect (Hart, 1996) makes the local smoothing method valid even when the correlation is ignored (Zeger and Diggle, 1994; Wild and Yee, 1996; Wu, Chiang and Hoover, 1998; Ruchstuhl, Welsh and Caroll, 2000). To take advantage of the correlation in the data, Severini and Staniswalis (1994) proposed to estimate the covariance matrix and incorporate the estimated covariance structure in the kernel weights.

Recently Xiao, Linton, Carroll and Mammen (2003) and Su and Ullah (2006) considered a problem similar to the one considered in this paper. These studies are closely related, but major difference exists, especially in the handling of the noise $\{e_t\}$. In Xiao et al. (2003) the noise series $\{e_t\}$ is assumed to be a general linear process and is approximated by a truncated AR process; in Su and Ullah (2006) $\{e_t\}$ is modeled as a finite-order nonparametric AR process. In this paper $\{e_t\}$ is modeled explicitly as an ARMA(p, q) process. This parsimonious representation allows us to improve the efficiency of estimation in finite samples. It has special advantages over Xiao et al. (2003) when the innovation process cannot be approximated with small-order AR models (e.g., seasonal ARMA models or ARMA models with roots close to one in the MA part). Comparing to the approach of Su and Ullah (2006), an explicit parametric form of the noise process allows faster convergence in the estimation of the innovation structure, hence the ability of generating more accurate predictions using the model.

This paper is organized as follows. In section 2, the estimation procedure and the asymptotic properties of the proposed estimator when e_t follows an AR(p) process are presented. In section 3 the results for the AR(p) case are extended to the general case when e_t follows an ARMA(p, q) process. Although AR(p) case is a special case of ARMA(p, q), different algorithms are used and different approaches are needed to prove the theorems. The pure AR structure provides a better algorithm and simpler proof of the asymptotic results. The performance of the proposed estimators are studied through simulation and compared with those of Xiao et al. (2003) and Su and Ullah (2006), the results are presented in section 4. The proposed procedures are applied on one real-life application and the results are presented in section 5. Section 6 contains summary and discussion.

The technical proofs are given in Appendix A. In the proof one important result of Yoshihara (1976) is used and an account of this result is given in Appendix B.

2 Estimation procedure in the pure AR case

2.1 The algorithm

When $\{e_t\}$ is a stationary AR(p) process, model (1) can be written as

$$Y_t = f(X_t) + e_t, \;\; \phi(B)e_t = \varepsilon_t.$$

With observations $\{(X_t, Y_t)\}_{t=1}^n$, first a preliminary estimator for $f(\cdot)$ is obtained by local linear regression, ignoring the correlation in $\{e_t\}$. Namely, $\widetilde{f}(x) = \widetilde{a}_0$, where $(\widetilde{a}_0, \widetilde{a}_1)$ minimizes

$$\sum_{t=1}^{n} \{Y_t - a_0 - a_1(X_t - x)\}^2 K_b(X_t - x), \tag{2}$$

where $K_b(\cdot) = b^{-1}K(\cdot/b)$, $K(\cdot)$ is a kernel function in $\mathcal{R}$, and $b > 0$ is a bandwidth. By simple algebra,

$$\widetilde{f}(x) - f(x) = \frac{1}{nb}\sum_{t=1}^{n} W_n\Big(\frac{X_t - x}{b}, x\Big)\{Y_t - f(x) - \dot{f}(x)(X_t - x)\}, \tag{3}$$

where

$$W_n(t, x) = (1, 0)S_n(x)^{-1}\begin{pmatrix} 1 \\ t \end{pmatrix} K(t). \tag{4}$$

In the above expression, $S_n(x)$ is a 2×2 matrix with $s_{i+j-2}(x)$ as its (i, j)-th element, and

$$s_k(x) = \frac{1}{n}\sum_{t=1}^{n} \Big(\frac{X_t - x}{b}\Big)^k K_b(X_t - x). \tag{5}$$

Under normal assumption, the maximum likelihood estimation for $f(\cdot)$ and $\boldsymbol{\phi}$ boils down to the following optimization problem:

$$\inf_{f,\phi} \sum_{t=1}^{n} \{Y_t - f(X_t) - \sum_{i=1}^{p} \phi_i(Y_{t-i} - f(X_{t-i}))\}^2, \tag{6}$$

where the infimum is taken over all smooth function f and $\boldsymbol{\phi} \in \mathcal{R}^p$ satisfies the stationary condition.

Let $\widetilde{e}_t = Y_t - \widetilde{f}(X_t)$ be the initial estimate of the innovation series e_t. Define

$$\mathbf{X_1} = \begin{pmatrix} \widetilde{e}_p & \widetilde{e}_{p-1} & \cdots & \widetilde{e}_1 \\ \widetilde{e}_{p+1} & \widetilde{e}_p & \cdots & \widetilde{e}_2 \\ \cdots & \cdots & \cdots & \\ \widetilde{e}_{n-1} & \widetilde{e}_{n-2} & \cdots & \widetilde{e}_{n-p} \end{pmatrix}, \qquad \mathbf{Y_1} = \begin{pmatrix} \widetilde{e}_{p+1} \\ \widetilde{e}_{p+2} \\ \cdots \\ \widetilde{e}_n \end{pmatrix},$$

and $\mathbf{W}$=diag$\Big\{\prod_{i=0}^{p} w(X_{t-i})\Big\}$, where $w(\cdot)$ is a weight function controlling the boundary effect in nonparametric estimation. An iterative estimation procedure is defined as follows:

1. Specify an initial value $\boldsymbol{\phi} = \widetilde{\boldsymbol{\phi}}$ defined as

$$\widetilde{\boldsymbol{\phi}} = (\mathbf{X_1^\tau W X_1})^{-1}\mathbf{X_1^\tau W Y_1}. \tag{7}$$

2. For given $\boldsymbol{\phi}$, let $\check{f}_j \equiv \check{f}(X_j) = \widehat{a}_0$, where $(\widehat{a}_0, \widehat{a}_1)$ minimizes

$$\sum_{t=1}^{n}\Big\{Y_t - a_0 - a_1(X_t - X_j) - \sum_{i=1}^{p}\phi_i\Big[Y_{t-i} - \widetilde{f}(X_{t-i})\Big]\Big\}^2 K_h(X_t - X_j)\prod_{i=1}^{p} w(X_{t-i}), \tag{8}$$

   where $K_h(\cdot) = h^{-1}K(\cdot/h)$, and $h > 0$ is a bandwidth. Obviously $\widehat{a}_1$ is an estimator for $\dot{f}_j \equiv \check{f}(X_j)$.

3. Obtain $\check{\boldsymbol{\phi}}$ by minimizing

$$\sum_{j=1}^{n}\sum_{t=1}^{n}\Big\{Y_t - \check{f}_j - \check{\dot{f}}_j(X_t - X_j) - \sum_{i=1}^{p}\phi_i\Big[Y_{t-i} - \widetilde{f}(X_{t-i})\Big]\Big\}^2 K_h(X_t - X_j)w(X_j)\prod_{i=1}^{p} w(X_{t-i}). \tag{9}$$

4. Repeat Steps 2 and 3 above until convergence. The terminal values are defined as estimators $\widehat{f}(X_j) = \check{f}_j$ and $\widehat{\boldsymbol{\phi}} = \check{\boldsymbol{\phi}}$.

Remark 1: Note that in (8) and (9), the values of $\widetilde{f}(X_{t-i})$ are fixed at the initial estimate throughout the iterations. This setting guarantees that the sum of squares is non-increasing in every iteration, hence guarantees the convergence. In practice, replacing $\widetilde{f}$ with the newly estimated function values may improve the results, though convergence is no longer guaranteed, and asymptotically it is not necessary.

Remark 2: In practice, only those $\widehat{f}(X_j)$ with $w(X_j) > 0$ will be calculated in order to eliminate the boundary bias in nonparametric estimation. One may let $w(\cdot)$ be an indicator function on, for example, the 80% inner sample range of X_t.

Remark 3: There are two bandwidths b and h in the estimation procedure. The asymptotic results below show that the bandwidth h in the iteration step should be of the standard order of $n^{-1/5}$. However, the bandwidth at the preliminary step (2) should be of smaller order $b = o(h)$ but $nb^4 \to \infty$ (Condition A4 in Appendix A). This requirement controls the bias in the preliminary step of the estimation. In practice, standard bandwidth selection in the iteration steps can be utilized. Experiments show that the final results are usually not very sensitive to the choice of bandwidth b. A fraction of the usual optimal bandwidth often works well.

Remark 4: In this paper $\{e_t\}$ and $\{X_t\}$ are assumed to be independent. For otherwise, the least squares-based estimators, such as local polynomial estimators, may not be consistent. Unfortunately this assumption essentially forbids the use of lagged Ys as explanatory variables. When lagged Ys are needed on the right-hand side of the model, alternative approaches are needed. For example, one may consider including enough lags of Y on the RHS of the model so that the innovation process becomes nearly uncorrelated and standard smoothing methods can be applied. Xiao et al. (2003) made a similar observation, here we share their view.

2.2 Asymptotic results

Let

$$\mathbf{X_2} = \begin{pmatrix} e_p & e_{p-1} & \cdots & e_1 \\ e_{p+1} & e_p & \cdots & e_2 \\ \cdots & \cdots & \cdots & \\ e_{n-1} & e_{n-2} & \cdots & e_{n-p} \end{pmatrix}, \quad \mathbf{Y_2} = \begin{pmatrix} e_{p+1} \\ e_{p+2} \\ \cdots \\ e_n \end{pmatrix}.$$

Define the "idealized" estimator

$$\widehat{\boldsymbol{\phi}}_{\text{Ideal}} = (\mathbf{X_2^\tau W X_2})^{-1}\mathbf{X_2^\tau W Y_2},$$

where $\mathbf{W}$ is the boundary weight matrix defined in section 2.1. This would be the 'idealized' least square estimator of the AR coefficients if $\{e_t\}$ is actually observable. It has been shown (e.g., Brockwell and Davis, 1987) that

$$\sqrt{n}(\widehat{\boldsymbol{\phi}}_{\text{Ideal}} - \boldsymbol{\phi}) \xrightarrow{D} N\Big(0,\ \frac{\mathrm{E}(\Pi_{i=0}^p w(X_{t-i}))^2}{[\mathrm{E}(\Pi_{i=0}^p w(X_{t-i}))]^2}\sigma^2\mathbf{V}(\boldsymbol{\phi})^{-1}\Big),$$

where $\mathbf{V}(\boldsymbol{\phi})$ is a $p \times p$ matrix and its (i,j)-th element is $\mathrm{Cov}(e_i, e_j)$. The following theorem links our estimator to $\widehat{\boldsymbol{\phi}}_{\text{Ideal}}$.

Theorem 1 *Under the conditions (A1)-(A6) in Appendix A, and that $\boldsymbol{\phi}$ satisfies the stationarity condition, then as $n \to \infty$,*

$$\sqrt{n}(\widetilde{\boldsymbol{\phi}} - \widehat{\boldsymbol{\phi}}_{Ideal}) = o_p(1),$$

where $\widetilde{\boldsymbol{\phi}}$ is the preliminary estimator defined in (7).

As a result of Theorem 1, $\widetilde{\boldsymbol{\phi}}$ shares the same asymptotic distribution of $\widehat{\boldsymbol{\phi}}_{\text{Ideal}}$, i.e.,

$$\sqrt{n}\Big(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}\Big) \xrightarrow{D} N\Big(0,\ \frac{\mathrm{E}\Big(\Pi_{i=0}^p w(X_{t-i})\Big)^2}{\Big[\mathrm{E}\Big(\Pi_{i=0}^p w(X_{t-i})\Big)\Big]^2}\sigma^2\mathbf{V}(\boldsymbol{\phi})^{-1}\Big). \tag{10}$$

As for the nonparametric function f, note that the local linear estimator defined by (8) may be expressed, for a generic x, as follows:

$$\widehat{f}(x) - f(x) = \frac{1}{nh}\sum_{t=1}^n W_n^*\Big(\frac{X_t - x}{h}, x, X_{t-1}, \cdots, X_{t-p}\Big)\Big\{\widetilde{Y}_t - f(x) - \dot{f}(x)(X_t - x)\Big\}, \tag{11}$$

where $\widetilde{Y}_t = Y_t - \sum_{i=1}^p \widetilde{\phi}_i\{Y_{t-i} - \widetilde{f}(X_{t-i})\}$, and

$$W_n^*(t, x, y_1, y_2, \cdots, y_p) = (1,0)S_n^*(x)^{-1}(1,t)^\tau K(t)\Pi_{i=1}^p w(y_i),$$

and $S_n^*(x)$ is defined in the same manner as $S_n(x)$ in (5) with $K_b(X_t - x)$ replaced by $K_h(X_t - x)\prod_{i=1}^p w(X_{t-i})$ (See also (3)). Theorem 2 below indicates that the above estimator is *asymptotically efficient* in the sense that the estimator admits the same (the first order) asymptotic distribution as if $\{Y_t\}$ would be defined by a simpler model with i.i.d. noise, namely $Y_t = f(X_t) + \varepsilon_t$.

Theorem 2 *Under the conditions (A1) to (A6) in Appendix A, for any point x in the support of X_t, as $n \to \infty$,*

$$\sqrt{nh}\Big\{\widehat{f}(x) - f(x) - \frac{h^2\mu_2}{2}\ddot{f}(x)\Big\} \xrightarrow{D} N\Big(0,\ \sigma(x)^2\Big),$$

where

$$\sigma(x)^2 = \frac{\sigma^2 \int K(u)^2 du}{g_1(x)} \frac{E\Big\{\Big[W(X_{t-1})W(X_{t-2})\cdots W(X_{t-p})\Big]^2 | X_t = x\Big\}}{\Big\{E\Big[W(X_{t-1})W(X_{t-2})\cdots W(X_{t-p})|X_t = x\Big]\Big\}^2}, \tag{12}$$

and $g_1(x)$ is the marginal density of X_t.

This theorem shows that the nonparametric transfer function estimator $\widehat{f}(\cdot)$ is indeed more efficient than the conventional local polynomial estimator $\widetilde{f}(\cdot)$. If $\widetilde{f}(\cdot)$ is used, the resulting asymptotic variance would have the same form as (12), but the white noise variance σ^2 in (12) would be replaced by the variance of e_t, which is strictly greater than σ^2 for a nontrivial AR(p) model. On the other hand, the asymptotic bias is not affected by the correlation structure. As a result, $\widehat{f}$ is more efficient than the conventional estimator $\widetilde{f}$ in the sense of mean square error. It can also be seen that the gain in efficiency of $\widehat{f}(\cdot)$ over $\widetilde{f}(\cdot)$ will be greater if the correlation is stronger.

3 Estimation procedure in the ARMA(p, q) case

Here we consider the general case when $\{e_t\}$ follows an ARMA(p, q) process. The estimation shares the similar "pre-whitening" idea with the AR(p) case and the asymptotic results are also similar. However the estimation procedures are more complicated in details and different techniques are required to establish the asymptotic results.

3.1 The algorithm

Modeling $\{e_t\}$ as a stationary, invertible ARMA(p, q) process, model (1) becomes

$$Y_t = f(X_t) + e_t, \quad e_t = \phi^{-1}(B)\theta(B)\varepsilon_t.$$

$\{e_t\}$ is assumed to be stationary and invertible, so $\{e_t\}$ admits the linear process representations $e_t = -\sum_{i=1}^{\infty} \pi_i e_{t-i} + \varepsilon_t$ and $e_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$, π_i and ψ_i are absolutely summable, i.e., $\sum_{i=0}^{\infty} |\pi_i| < \infty$ and $\sum_{i=0}^{\infty} |\psi_i| < \infty$ (Box and Jenkins, 1976). Denote $\boldsymbol{\beta} = (\phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q)^\tau$. $f(\cdot)$ and $\boldsymbol{\beta}$ are estimated by solving the following nonlinear optimization problem

$$\inf_{f, \boldsymbol{\beta}} \sum_{t=1}^{n} \Big\{Y_t - f(X_t) + \Big[\frac{\phi(B)}{\theta(B)} - 1\Big]\Big[Y_t - f(X_t)\Big]\Big\}^2, \tag{13}$$

where the infimum is taken over all smooth function f and all $\boldsymbol{\beta} \in \mathcal{R}^{p+q}$ satisfying the stationary and invertible conditions. To initiate the iteration, an initial estimate $\widetilde{f}(\cdot)$ is obtained by local

linear regression, ignoring the serial correlation in $\{e_t\}$ (see also (2)). The iterative procedure is described as follows:

1. Obtain an initial estimate $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}})$ by minimizing
$$\sum_{t=1}^{n} \Big\{ \frac{\phi(B)}{\theta(B)} \big[Y_t - \widetilde{f}(X_t) \big] \Big\}^2 \tag{14}$$
with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

2. Given $\boldsymbol{\beta}$, let $\check{f}_j \equiv \check{f}(X_j) = \hat{a}_0$, where $(\hat{a}_0, \hat{a}_1)$ minimizes
$$\sum_{t=1}^{n} \Big\{ Y_t - a_0 - a_1(X_t - X_j) + \Big[\frac{\phi(B)}{\theta(B)} - 1 \Big] \big[Y_t - \widetilde{f}(X_t) \big] \Big\}^2 K_h(X_t - X_j),$$
where $K_h(\cdot) = 1/hK(\cdot/h)$, h is a bandwidth and h is of larger order than b.

3. Define $\check{\boldsymbol{\beta}}$ to minimize
$$\sum_{j=1}^{n} \sum_{t=1}^{n} \Big\{ Y_t - \check{f}_j - \dot{\check{f}}_j (X_t - X_j) + \Big[\frac{\phi(B)}{\theta(B)} - 1 \Big] \big[Y_t - \widetilde{f}(X_t) \big] \Big\}^2 K_h(X_t - X_j). \tag{15}$$

4. Repeat steps 2 and 3 until $\{\check{f}_j\}$ and $\check{\boldsymbol{\beta}}$ change only by a small amount in two successive iterations. The terminal values of $\widehat{f}(X_j) = \check{f}_j$ and $\widehat{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}}$ are the estimators of $f(\cdot)$ and $\boldsymbol{\beta}$, respectively.

Several algorithms can be used to solve the nonlinear optimization problems presented in equations (13) to (15). In this study, a nonlinear estimation method based on the Gauss-Newton algorithm is used. In this method, steps 1 and 3 can be iterated to improve the finite sample performance. The details of this method can be found in Appendix A.

3.2 Asymptotic results

Similar to the AR(p) case, the "idealized" estimator of $\boldsymbol{\beta}$ is defined as the solution of $\widehat{\boldsymbol{\beta}}_{\text{Ideal}} = \inf_{\boldsymbol{\beta}} \Big\{ \phi(B)\theta(B)^{-1} e_t \Big\}^2$, assuming $\{e_t\}$ observable. As a standard estimator of an ARMA model, it has been shown that (e.g., Brockwell and Davis, 1987)
$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{Ideal}} - \boldsymbol{\beta}) \xrightarrow{D} N\Big(0,\ \sigma^2 \mathbf{V}(\boldsymbol{\beta})^{-1}\Big),$$
where
$$\mathbf{V}(\boldsymbol{\beta}) = \mathrm{E} \begin{pmatrix} \mathbf{U}_1 \mathbf{U}_1^\tau & \mathbf{U}_1 \mathbf{V}_1^\tau \\ \mathbf{V}_1 \mathbf{U}_1^\tau & \mathbf{V}_1 \mathbf{V}_1^\tau \end{pmatrix}, \tag{16}$$
$\mathbf{U}_t = (U_t, U_{t-1}, \cdots, U_{t+1-p})^\tau$, $\mathbf{V}_t = (V_t, V_{t-1}, \cdots, V_{t+1-q})^\tau$. $\{U_t\}$ is an AR(p) process defined by $\phi(B)U_t = a_t$ and $\{V_t\}$ is an AR(q) process defined by $\theta(B)V_t = b_t$, a_t and b_t are white noise processes. Obviously, when the model does not contain the AR component (pure MA(q) model), $\mathbf{V}(\boldsymbol{\beta}) = \mathrm{E}\left(\mathbf{V}_1 \mathbf{V}_1^\tau\right)$. Using this result, the following asymptotic results for the ARMA(p, q) case can be established.

Theorem 3 *Under the conditions (A1) to (A5) and (A6*) in Appendix A, and that $\boldsymbol{\phi}$ satisfies the stationarity condition and $\boldsymbol{\theta}$ satisfies the invertibility condition, then as $n \to \infty$,*

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{Ideal}) = o_p(1).$$

As a result of Theorem 3, $\widetilde{\boldsymbol{\beta}}$ shares the same asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\text{Ideal}}$, i.e.,

$$\sqrt{n}\Big(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\Big) \xrightarrow{D} N\Big(0,\ \sigma^2 \mathbf{V}(\boldsymbol{\beta})^{-1}\Big),$$

where $\mathbf{V}(\boldsymbol{\beta})$ is defined in (16).

Theorem 4 *Under the conditions (A1) to (A5) and (A6*) in Appendix A and that $\{e_t\}$ is a stationary, invertible ARMA(p,q) process, then for any point x in the support of X_t, as $n \to \infty$,*

$$\sqrt{nh}\Big\{\widehat{f}(x) - f(x) - \frac{h^2\mu_2}{2}\ddot{f}(x)\Big\} \xrightarrow{D} N\Big(0,\ \sigma(x)^2\Big),$$

where

$$\sigma(x)^2 = \frac{\sigma^2 \int K(u)^2 du}{g_1(x)},$$

and $g_1(x)$ is the marginal density function of X_t.

Theorems 3 and 4 show that similar results as those in the AR(p) case continue to hold in the ARMA(p,q) case, despite the more complicated correlation structure. Results similar to Theorems 2 and 4 are established by Xiao et al. (2003, Theorem 2) and Su and Ullah (2006, Theorem 3.1) under different assumptions on e_t.

4 Numerical properties

To study the finite-sample properties of the proposed estimator, simulation studies are conducted using model (1), where

$$f(X_t) = \sin(4X_t) + \cos(2X_t),$$

and X_t is generated from an AR(1) model $X_t = 0.3X_{t-1} + a_t,\ a_t \sim$ i.i.d. N$(0, 0.3^2)$. For $\{e_t\}$, an ARMA(1,1) model ($e_t = \phi e_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}$) and two simple seasonal models ($e_t = \phi_4 e_{t-4} + \varepsilon_t$ and $e_t = \varepsilon_t - \theta_4\varepsilon_{t-4}$, denoted as AR$(1)_4$ and MA$(1)_4$, respectively) are considered. In these models, $\varepsilon_t \sim N(0, 0.5^2)$.

Three sample sizes (100, 200 and 400) are considered and 200 replications are used in each case. The standard normal density function is used as the kernel function. Different bandwidths b and h are experimented. Due to the fact that the results are not very sensitive to the bandwidths, only the case of $h = 1.06 s_X n^{-1/5}$ and $b = 1.06 s_X n^{-1/4}$ is reported here.

For comparison, under the same setting specified above, simulations are run using the proposed nonparametric transfer function approach, the AR approximation approach of Xiao et al. (2003), the nonparametric AR approximation approach of Su and Ullah (2006), and the "conventional" local linear estimator, in which $\{e_t\}$ is assumed to be white noise. In the sequel, the approaches will be abbreviated as NPTF, XLCM, SU and WHITE, respectively.

The mean squared errors (MSE$\equiv \frac{1}{n}\sum_{t=1}^{n}\{\widehat{f}(X_t) - f(X_t)\}^2$) of all four estimators are averaged over the replications. As a measure of relative efficiency, the relative MSEs of NPTF, XLCM, and SU are calculated by dividing their average MSEs by that of WHITE. The relative MSEs are reported in Tables 1 and 2 under the corresponding procedure names. The means and standard deviations of $\widehat{\phi}$ and $\widehat{\theta}$ from NPTF are also reported, as well as the average mean squared error of WHITE (AMSE), which is the common denominator of the relative MSEs. A histogram of $\widehat{\phi}$ and a plot of a typical simulation are given in Figure 1.

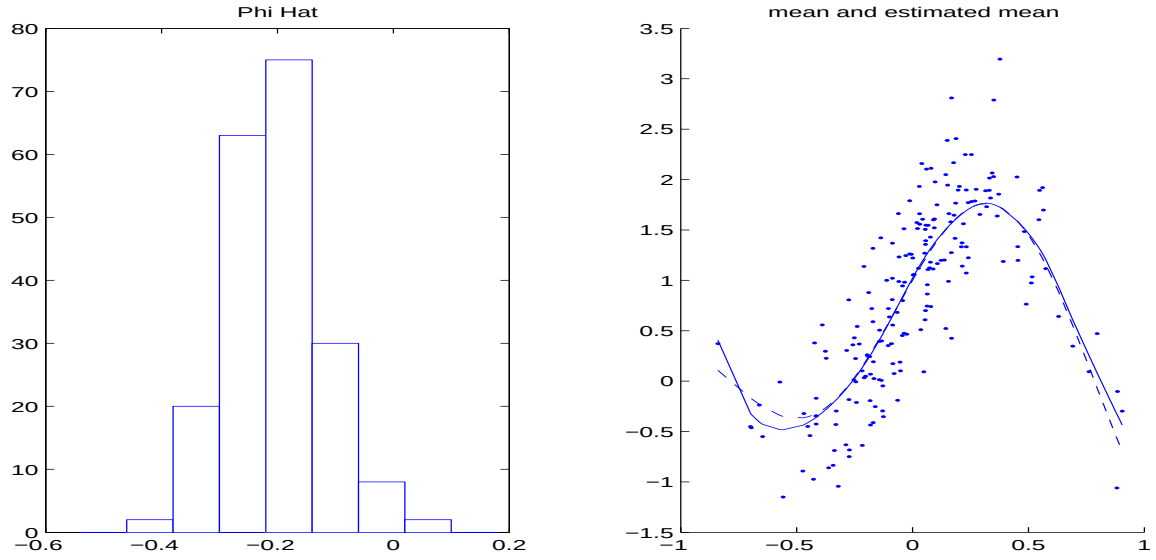

Figure 1: $\phi = -0.2$, n=200. Left panel: histogram of $\hat{\phi}$, right panel: true (solid line) and estimated (dashed line) transfer function in a typical simulation.

The following phenomena are also observed in Xiao et al. (2003) and Su and Ullah (2006) so they are only briefly mentioned here. (1) The NPTF estimator $\widehat{f}(\cdot)$ is more efficient than the conventional local linear regression estimator, the stronger the autocorrelation, the larger the gain in efficiency of $\widehat{f}(\cdot)$. (2) the performance of the estimators improves with the increase of sample size. (3) The MA estimates may have large bias and larger sample sizes are needed to improve the performance. In this study we model e_t explicitly as an ARMA(p, q) process. As illustrated in Figure 1, the sampling distributions of $\hat{\phi}$ and $\hat{\theta}$ are close to their asymptotic normal distributions. For a comparison between NPTF, XLCM and SU, the simulation shows that generally they are all more efficient than the conventional estimator. When $\{e_t\}$ follows an ARMA model with small

Table 1: Simulation results: AR(1) and MA(1) models

ϕ	θ	n	mean($\hat{\phi}$), $s_{\hat{\phi}}$	mean($\hat{\theta}$), $s_{\hat{\theta}}$	AMSE	NPTF	XLCM	SU
		100	-.786, .070		.033	.466	.484	.553
-.8		200	-.802, .049		.023	.405	.414	.464
		400	-.799, .034		.015	.395	.400	.482
		100	-.507, .100		.019	.792	.810	.895
-.5		200	-.508, .065		.011	.801	.809	.900
		400	-.501, .047		.006	.756	.767	.836
		100	-.216, .105		.018	.992	1.01	1.12
-.2		200	-.210, .082		.010	.966	.970	1.04
		400	-.200, .054		.006	.981	.982	1.05
		100	.196, .107		.020	1.01	1.07	1.10
.2		200	.198, .078		.012	1.05	1.06	1.12
		400	.198, .054		.007	1.01	1.01	1.06
		100	.483, .096		.031	.912	.926	.943
.5		200	.493, .066		.019	.904	.910	.944
		400	.494, .048		.010	.898	.902	.921
		100	.774, .076		.092	.835	.837	.845
.8		200	.792, .049		.053	.758	.761	.776
		400	.799, .032		.030	.738	.740	.745
		100		-.712, .091	.120	.818	.883	.916
	-.8	200		-.742, .057	.069	.753	.816	.859
		400		-.765, .035	.038	.746	.797	.847
		100		-.492, .099	.092	.884	.921	.961
	-.5	200		-.497, .069	.052	.849	.885	.933
		400		-.496, .048	.029	.833	.872	.927
		100		-.184, .115	.064	.990	1.02	1.05
	-.2	200		-.198, .075	.039	.953	.954	1.03
		400		-.200, .052	.023	.950	.949	1.03
		100		.219, .123	.058	.955	.965	1.06
	.2	200		.209, .078	.034	.936	.950	1.03
		400		.204, .054	.021	.919	.926	1.02
		100		.516, .099	.059	.791	.824	.987
	.5	200		.497, .073	.037	.757	.773	.866
		400		.501, .048	.023	.742	.759	.858
		100		.725, .094	.053	.691	.737	.843
	.8	200		.745, .061	.047	.665	.696	.773
		400		.757, .040	.029	.643	.682	.740

Table 2: Simulation results: ARMA(1,1), $AR(1)_4$ and $MA(1)_4$ models

ϕ	θ	n	mean($\hat{\phi}$), $s_{\hat{\phi}}$	mean($\hat{\theta}$), $s_{\hat{\theta}}$	AMSE	NPTF	XLCM	SU
		100	.217, .151	-.645, .127	.039	.836	.869	.944
.2	-.8	200	.211, .093	-.685, .076	.026	.802	.873	.985
		400	.209, .065	-.739, .055	.013	.737	.815	.909
		100	.512, .123	-.537, .147	.076	.737	.780	.839
.5	-.8	200	.518, .083	-.592, .104	.044	.703	.748	.784
		400	.522, .056	-.639, .075	.025	.669	.728	.771
		100	.819, .079	-.286, .163	.259	.761	.819	.857
.8	-.8	200	.816, .053	-.397, .142	.161	.692	.748	.761
		400	.812, .039	-.482, .083	.083	.666	.710	.726
		100	.207, .183	-.457, .168	.029	.882	.930	1.04
.2	-.5	200	.210, .127	-.469, .111	.016	.865	.907	1.03
		400	.207, .086	-.485, .080	.011	.859	.886	.975
		100	.516, .141	-.381, .147	.059	.771	.823	.885
.5	-.5	200	.507, .086	-.422, .088	.034	.759	.781	.841
		400	.503, .056	-.447, .069	.019	.758	.783	.802
		100	.806, .094	-.235, .148	.210	.784	.813	.830
.8	-.5	200	.811, .051	-.305, .103	.113	.714	.753	.783
		400	.812, .038	-.359, .079	.060	.663	.703	.730
ϕ_4	θ_4	n	mean($\hat{\phi}_4$), $s_{\hat{\phi}_4}$	mean($\hat{\theta}_4$), $s_{\hat{\theta}_4}$	AMSE	NPTF	XLCM	SU
		100	-.764, .072		.039	.471	.927	1.03
-.8		200	-.783, .048		.025	.434	.895	.978
		400	-.791, .034		.015	.421	.896	.962
		100	-.484, .094		.020	.874	1.01	1.13
-.5		200	-.488, .065		.013	.836	.996	1.10
		400	-.495, .048		.008	.815	.989	1.07
		100		.495, .113	.018	.959	1.05	1.23
	.5	200		.493, .064	.011	.875	1.02	1.19
		400		.495, .050	.007	.866	1.02	1.18
		100		.698, .098	.023	.872	1.03	1.17
	.8	200		.721, .061	.013	.842	1.01	1.20
		400		.741, .045	.008	.809	1.00	1.16

$|\theta|$ (including pure AR models), NPTF and XLCM have similar efficiency, however when $|\theta|$ is large, the NPTF estimator is more efficient. For the seasonal models, NPTF has similar gain in efficiency as in the non-seasonal models, while in many cases XLCM and SU fail to approximate e_t appropriately and the estimate is no longer efficient (Table 2). In the simulation higher-order AR approximations are also used in XLCM, but the performance does not always improve, partially due to the additional error introduced in estimating more parameters. Since the finding is similar, the detailed results are omitted. In the simulation, SU is not as efficient as NPTF and XLCM, mainly because here e_t is generated from ARMA models of finite order. In a separate study, e_t is generated from nonlinear finite order AR processes and SU is found to be more efficient.

5 Example: river flow and rainfall

In this section the proposed nonparametric transfer function approach is used to analyze the effect of daily rain fall on river flow of Kanna river (Japan) in year 1956. The effect of rainfall on river flow is usually highly nonlinear, mainly because the soil moisture varies from rainy period to dry period. This dataset was analyzed by Ozaki (1985) and later used by Chen and Tsay (1996) as an example of the nonlinear transfer function (NLTF) model. For details of the data, see Chen and Tsay (1996).

The proposed nonparametric transfer function model is used to analyze this dataset and the performance is compared with those of the NLTF model and the linear transfer function model (LTF). The sample autocorrelation function (ACF) of Y_t indicates non-stationarity. After taking first order difference of Y_t, the resulting series appears to be stationary. Let $Z_t = Y_t - Y_{t-1}$ and consider the following model

$$Z_t = f(X_t, X_{t-1}, X_{t-2}) + e_t. \tag{17}$$

Note here a low-dimensional smoothing model is used instead of an univariate smoothing model. Following the proposed estimation procedures, $f(\cdot)$ is first estimated assuming $\{e_t\}$ i.i.d., then the resulting preliminary estimate $\widetilde{f}(\cdot)$ is removed from Z_t and a model is identified for $\{e_t\}$ based on the sample autocorrelation function of the partial residuals (Figure 2). The resulting model is an AR model with lagged variables at lags 4, 5, 6 and 14.

The bandwidth is selected via the generalized cross validation (GCV) criteria (Craven and Wahba, 1979).

$$h = \arg\min_h \frac{(\mathbf{Y} - \widehat{\mathbf{f}})^\tau (\mathbf{Y} - \widehat{\mathbf{f}})}{n[1 - \mathrm{tr}(\mathbf{S}_h)/n]^2},$$

where $\mathbf{S}_h$ is the smoother matrix associated with h such that $\widehat{f} = \mathbf{S}_h \mathbf{Y}$, and $\mathbf{Y}$ is the vector of observations. In order to compare with the parametric models, the equivalent number of parameters defined as $\mathrm{tr}(\mathbf{S}_h)$ is also calculated. The resulting bandwidth is 5 and the equivalent number of

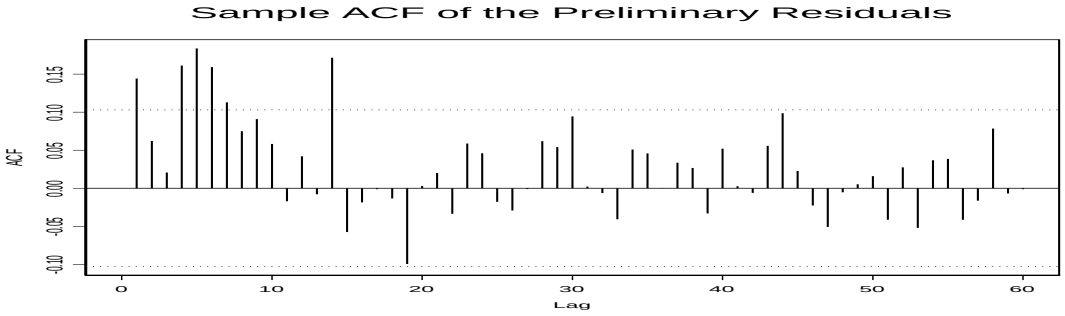

Figure 2: Sample ACF plot of the partial residuals after removing $\widetilde{f}(\cdot)$

parameters is 33.46. The estimated AR parameters are $\widehat{\phi}_4 = .0912$, $\widehat{\phi}_5 = .1264$, $\widehat{\phi}_6 = .1593$ and $\widehat{\phi}_{14} = .0704$. Figure 3 is the ACF plot of the final residuals.

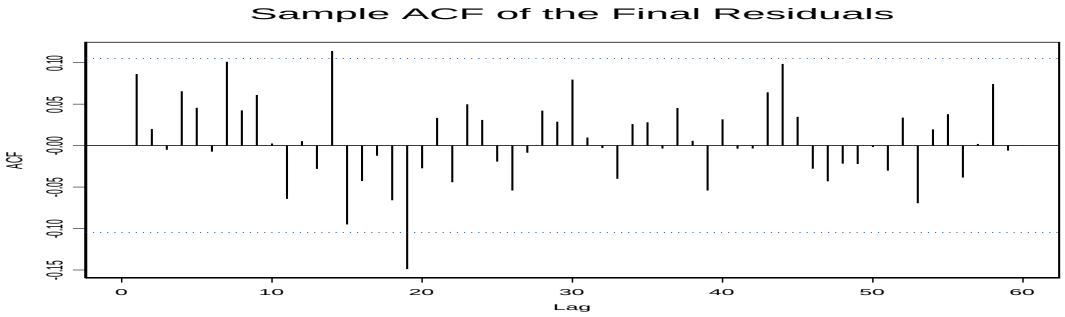

Figure 3: Sample ACF plot of the final residuals

To study the forecasting performance of the NPTF model, the following rolling forecasting scheme is employed: for each $t = 180, 181, \cdots, 365$, data available at t are used to build the model and make one-step ahead prediction. For convenience, actual values of X_{t+1} are used in the prediction. For each t, the forecasting error $Y_{t+1} - \widehat{Y}_t(1)$ is calculate. Finally, the squared forecasting errors are averaged over t. The square-root of this average is referred to as "post-sample forecasting RMSE".

Table 3 shows a comparison between the NPTF model with a parametric nonlinear transfer function model (NLTF) and a linear transfer function model (LTF) fitted by Chen and Tsay (1996). Residual variances and RMSEs from rolling forecasts are obtained using the model settings detailed in Chen and Tsay (1996).

The above results show that the NPTF has smaller residual variance, but large equivalent number of parameters. This may indicate overfitting. However, the better forecasting performance of the NPTF model justifies its use of more parameters.

The one-step ahead forecast errors of the NPTF model and the NLTF model are plotted against the forecasting origins in Figure 4. The performance of the LTF model is not as good as the NLTF and NPTF models, so its errors are not plotted in this figure for clearer presentation. From this

Table 3: Within- and Post-Sample Comparisons

	NPTF	NLTF	LTF
(Equivalent) Number of Parameters	33.46	12	10
Residual variance	4.58	6.23	20.81
Forecasting RMSE	8.80	12.56	13.93

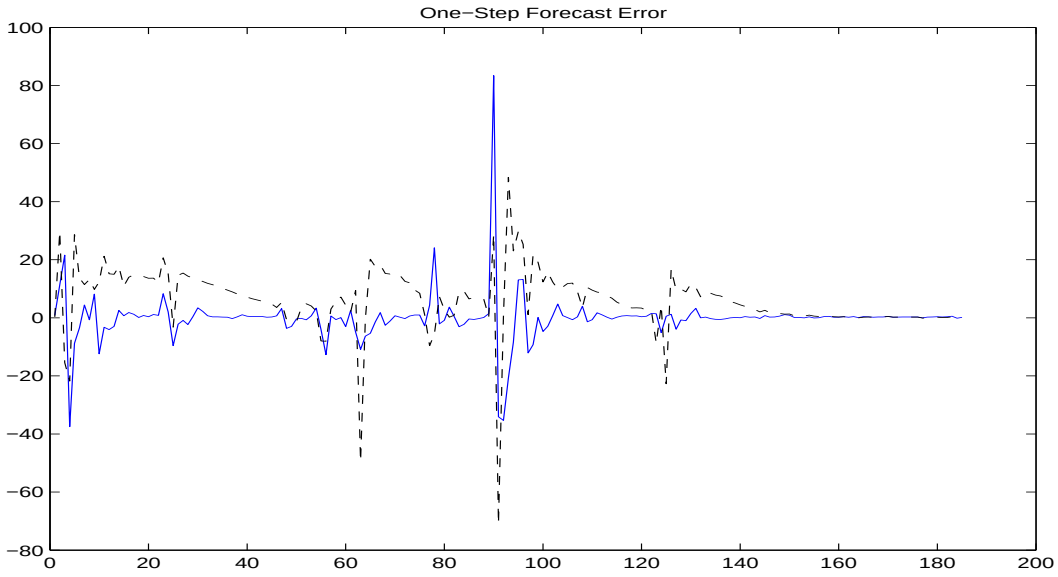

Figure 4: The one-step ahead forecast errors of the NPTF model (solid line) and the NLTF model (dashed line)

figure it is clear that the NPTF model outperforms the NLTF model most of the time. On average, the NPTF model performs better than the NLTF and LTF models in that it produces not only smaller within-sample RMSE but also smaller post-sample RMSE. This example shows the potential of the nonparametric transfer function model in modeling nonlinear time series.

6 Summaries and discussions

In this paper a new method is proposed to model nonlinear relationships between an input and an output time series. The transfer function $f(\cdot)$ is modeled by nonparametric smoothing and the innovation process $\{e_t\}$ is modeled as a stationary ARMA(p, q) process. The nonparametric feature of this model allows us to model highly nonlinear relationships of unknown functional forms, while modeling $\{e_t\}$ as an ARMA model improves not only the efficiency in estimating $f(\cdot)$ but also the forecasting performance. The simulations and empirical study show good potential of this model

in analyzing nonlinear time series.

There are some issues in the nonparametric transfer function model that deserve further study. For example, in this study the transfer function is univariate. It is easy, though tedious, to generalize the results to multi-dimensional cases, under the general model $Y_t = f(X_{1t}, \cdots, X_{pt}) + e_t$. However, such a direct generalization is often not practical in practice due to the aforementioned "curse of dimensionality". To solve this problem, more restrictive models, such as the additive model, must be considered. Research addressing this topic is ongoing.

Acknowledgement

Rong Chen's research is partially supported by NSF grant DMS-0244541 and NIH grant R01 GM068958. Qiwei Yao's research is partially supported by EPSRC grants GR/R97436 and EP/C549058. We would like to thank the editor and two anonymous referees for their valuable comments and suggestions which led to a substantial improvement of the paper.

Appendix A – Technical Proofs

In the proofs that follow, $C > 0$ denotes a generic constant that may vary from line to line. Let $g_1(\cdot)$ be the density function of X_t and $g_i(x_{t1}, \cdots, x_{ti})$ be the i-dimensional joint density function of $\{X_{t1}, \cdots, X_{ti}\}$. The following assumptions are needed, of which (A1) to (A5) are needed for both the pure AR(p) and the ARMA(p, q) cases, (A6) is needed for the pure AR(p) case and (A6*) is needed for the ARMA(p, q) case.

(A1) $\{X_t\}$ is β-mixing in the sense that

$$\beta(k) = \mathrm{E}\{ \sup_{B \in \mathcal{F}_k^\infty} |P(B) - P(B|X_0, X_{-1}, \cdots)| \} \to 0$$

as $k \to \infty$, where $\mathcal{F}_i^j$ is the σ-algebra generated by $\{X_i, \cdots, X_j\}$ for $i \le j$. In addition, $\sum_{k \ge 1} k\beta(k)^{\delta/(2+\delta)} < \infty$ for some $\delta \in (0, 8)$.

(A2) The kernel function is symmetric, compactly supported and Lipschitz continuous.

(A3) $f(\cdot)$ has continuous second derivative $\ddot{f}(\cdot)$ and $g_1(\cdot)$ is bounded away from zero.

(A4) As $n \to \infty$, $h = O(n^{-1/5})$, $b = o(n^{-1/5})$, and $nb^4 \to \infty$.

(A5) $\{X_t\}$ and $\{\varepsilon_t\}$ are two independent processes.

(A6) The weight function $w(\cdot)$ is continuous on its compact support contained in $\{g_1(x) > 0\}$.

(A6*) X_t has bounded support $[a,b]$. The density functions $g_1(\cdot)$, $g_2(\cdot,\cdot)$, $g_4(\cdot,\cdot,\cdot,\cdot)$ and $g_6(\cdot,\cdot,\cdot,\cdot,\cdot,\cdot)$ are continuous and have continuous first two derivatives.

The following lemma is needed to prove the theorems:

Lemma 1 *As $n \to \infty$, it holds uniformly for x in any compact subset of $\{g_1(x) > 0\}$ that*

$$\widetilde{f}(x) - f(x) = \frac{1}{nbg_1(x)} \sum_{t=1}^{n} K\Big(\frac{X_t - x}{b}\Big) e_t + \frac{b^2}{2}\mu_2 \ddot{f}(x) + O_p\Big[R_n(x)\Big\{(\frac{\log n}{nb})^{1/4} + b\Big\}\Big],$$

where $\mu_2 = \int u^2 K(u)du$, and

$$R_n(x) = \frac{1}{nbg_1(x)}\Big\{\Big|\sum_{t=1}^{n} K\Big(\frac{X_t - x}{b}\Big)e_t\Big| + \Big|\sum_{t=1}^{n}\Big(\frac{X_t - x}{b}\Big)K\Big(\frac{X_t - x}{b}\Big)e_t\Big|\Big\} + O(b^2).$$

Proof of Lemma 1

It follows from Theorem 5.3 of Fan and Yao (2003) that

$$s_k(x) = g_1(x)\mu_k + O_p\Big\{\Big(\frac{\log n}{nb}\Big)^{1/2} + b^2\Big\}$$

uniformly for $x \in A$, where $s_k(x)$ is defined in (5), $\mu_k = \int u^k K(u)du$, and A is any compact set contained in $\{g_1(x) > 0\}$. Hence it holds uniformly for $x \in A$ that

$$S_n(x) = S(x) + O_p\Big\{\Big(\frac{\log n}{nb}\Big)^{1/2} + b^2\Big\},$$

where $S(x) = g_1(x)\mathrm{diag}(1, \mu_2)$. Write $Y_t^* = Y_t - f(x) - \dot{f}(x)(X_t - x)$. It is easy to see from (4) that

$$\begin{aligned}
&\Big|\sum_{t=1}^{n}\Big\{W_n\Big(\frac{X_t - x}{b}, x\Big) - g_1(x)^{-1}K\Big(\frac{X_t - x}{b}\Big)\Big\}Y_t^*\Big| \\
= \;& \Big|(1,0)\{S_n(x)^{-1} - S(x)^{-1}\}\sum_{t=1}^{n}\Big(1, \frac{X_t - x}{b}\Big)^{\tau} K\Big(\frac{X_t - x}{b}\Big)Y_t^*\Big| \\
\le \;& [(1,0)\{S_n(x)^{-1} - S(x)^{-1}\}^2(1,0)^{\tau}]^{1/2}\Big\{\Big|\sum_{t=1}^{n} K\Big(\frac{X_t - x}{b}\Big)Y_t^*\Big|^2 + \Big|\sum_{t=1}^{n}\frac{X_t - x}{b}K\Big(\frac{X_t - x}{b}\Big)Y_t^*\Big|^2\Big\}^{1/2} \\
\le \;& [(1,0)\{S_n(x)^{-1} - S(x)^{-1}\}^2(1,0)^{\tau}]^{1/2}\Big\{\Big|\sum_{t=1}^{n} K\Big(\frac{X_t - x}{b}\Big)Y_t^*\Big| + \Big|\sum_{t=1}^{n}\frac{X_t - x}{b}K\Big(\frac{X_t - x}{b}\Big)Y_t^*\Big|\Big\} \\
\le \;& O_p\Big[\Big\{\Big(\frac{\log n}{nb}\Big)^{1/2} + b^2\Big\}^{1/2}\Big]\Big\{\Big|\sum_{t=1}^{n} K\Big(\frac{X_t - x}{b}\Big)e_t\Big| + \Big|\sum_{t=1}^{n}\frac{X_t - x}{b}K\Big(\frac{X_t - x}{b}\Big)e_t\Big| + O(nb^3)\Big\}.
\end{aligned}$$

The last inequality follows from the fact that $Y_t = f(X_t) + e_t$, $K(\cdot)$ has a compact support. Now the lemma follows from (3) and a simple Taylor expansion. The proof is completed.

Proof of Theorem 1

Since $\{e_t\}$ is a stationary Gaussian AR(p) process, it is also β-mixing with exponentially decaying mixing coefficients. Put $w_t = w(X_t)$, let $\mathbf{A} = \mathbf{X_1^\tau W X_1}$ and $\mathbf{B} = \mathbf{X_1^\tau W Y_1}$, where $\mathbf{X_1}$, $\mathbf{Y_1}$ and $\mathbf{W}$ are defined in section 2.1. From (7) we have $\widetilde{\boldsymbol{\phi}} = \mathbf{A^{-1}B}$, the (r,s)-th element of $\mathbf{A}$ is

$$\begin{aligned} A_{rs} &= \sum_{t=1}^{n} \Big[Y_{t-r} - \widetilde{f}(X_{t-r})\Big]\Big[Y_{t-s} - \widetilde{f}(X_{t-s})\Big] \prod_{k=0}^{p} w_{t-k} \\ &= \sum_{t=1}^{n} \Big[e_{t-r} + f(X_{t-r}) - \widetilde{f}(X_{t-r})\Big]\Big[e_{t-s} + f(X_{t-s}) - \widetilde{f}(X_{t-s})\Big] \prod_{k=0}^{p} w_{t-k} \\ &= \sum_{t=1}^{n} e_{t-r}e_{t-s} \prod_{k=0}^{p} w_{t-k} + A_{rs1} + A_{rs2} + A_{rs3}, \end{aligned}$$

where

$$A_{rs1} = \sum_{t=1}^{n} \{f(X_{t-r}) - \widetilde{f}(X_{t-r})\}\{f(X_{t-s}) - \widetilde{f}(X_{t-s})\} \prod_{k=0}^{p} w_{t-k},$$

$$A_{rs2} = \sum_{t=1}^{n} e_{t-r}\{f(X_{t-s}) - \widetilde{f}(X_{t-s})\} \prod_{k=0}^{p} w_{t-k}, \qquad A_{rs3} = \sum_{t=1}^{n} e_{t-s}\{f(X_{t-r}) - \widetilde{f}(X_{t-r})\} \prod_{k=0}^{p} w_{t-k}.$$

The r-th element of $\mathbf{B}$ is

$$\begin{aligned} B_{r} &= \sum_{t=1}^{n} \Big[Y_{t} - \widetilde{f}(X_{t})\Big]\Big[Y_{t-r} - \widetilde{f}(X_{t-r})\Big] \prod_{k=0}^{p} w_{t-k} \\ &= \sum_{t=1}^{n} \Big[e_{t} + f(X_{t}) - \widetilde{f}(X_{t})\Big]\Big[e_{t-r} + f(X_{t-r}) - \widetilde{f}(X_{t-r})\Big] \prod_{k=0}^{p} w_{t-k} \\ &= \sum_{t=1}^{n} e_{t}e_{t-r} \prod_{k=0}^{p} w_{t-k} + B_{r1} + B_{r2} + B_{r3}, \end{aligned}$$

where

$$B_{r1} = \sum_{t=1}^{n} \{f(X_{t}) - \widetilde{f}(X_{t})\}\{f(X_{t-r}) - \widetilde{f}(X_{t-r})\} \prod_{k=0}^{p} w_{t-k},$$

$$B_{r2} = \sum_{t=1}^{n} e_{t}\{f(X_{t-r}) - \widetilde{f}(X_{t-r})\} \prod_{k=0}^{p} w_{t-k}, \qquad B_{r3} = \sum_{t=1}^{n} e_{t-r}\{f(X_{t}) - \widetilde{f}(X_{t})\} \prod_{k=0}^{p} w_{t-k}.$$

The Theorem follows immediately from the two statements below:

(i) $B_{r1} + B_{r2} + B_{r3} = o_p(\sqrt{n})$, and
(ii) $A_{rs1} + A_{rs2} + A_{rs3} = o_p(\sqrt{n})$.

for all $r, s = 1, 2, \cdots, p$.

Here only (i) is established. The proof for (ii) is similar and simpler. By Lemma 1, we may write

$$B_{r1} = \{B_{r11} + B_{r12} + B_{r13} + O_p(nb^4)\}\{1 + o_p(1)\}, \tag{18}$$

where

$$B_{r11} = \frac{1}{n^2b^2}\sum_{i,j,k} K\Big(\frac{X_i - X_k}{b}\Big)K\Big(\frac{X_j - X_{k-r}}{b}\Big)\frac{e_ie_j}{g_1(X_k)g_1(X_{k-r})}\prod_{l=0}^{p} w_{k-l} \equiv \frac{1}{n^2b^2}\sum_{i,j,k}\zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k),$$

$$B_{r12} = \frac{b\mu_2}{2n}\sum_{i,k}\frac{e_i\ddot{f}(X_{k-r})}{g_1(X_k)}K\Big(\frac{X_i - X_k}{b}\Big)\prod_{l=0}^{p} w_{k-l}, \quad B_{r13} = \frac{b\mu_2}{2n}\sum_{i,k}\frac{e_i\ddot{f}(X_k)}{g_1(X_{k-r})}K\Big(\frac{X_i - X_{k-r}}{b}\Big)\prod_{l=0}^{p} w_{k-l},$$

where $\boldsymbol{\xi}_i = (X_i, X_{i-1}, \cdots, X_{i-p}, e_i)^\tau$. B_{r11} is split into two sums B_{r111} and B_{r112} consisting of, respectively, the terms with different i, j, k and the terms with at least two of i, j, k the same. To perform the Hoeffding decomposition on the U-statistic B_{r111}, put

$$\begin{aligned}\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) &= \zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) + \zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_k, \boldsymbol{\xi}_j) + \zeta(\boldsymbol{\xi}_j, \boldsymbol{\xi}_i, \boldsymbol{\xi}_k)\\ &+ \zeta(\boldsymbol{\xi}_j, \boldsymbol{\xi}_k, \boldsymbol{\xi}_i) + \zeta(\boldsymbol{\xi}_k, \boldsymbol{\xi}_i, \boldsymbol{\xi}_j) + \zeta(\boldsymbol{\xi}_k, \boldsymbol{\xi}_j, \boldsymbol{\xi}_i).\end{aligned}$$

Define

$$\begin{aligned}\theta(P) &= \int\int\int \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k)\, dP(\boldsymbol{\xi}_i)\, dP(\boldsymbol{\xi}_j)\, dP(\boldsymbol{\xi}_k);\\ \widetilde{\kappa}_1(\boldsymbol{\xi}_i) &= \int\int \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k)\, dP(\boldsymbol{\xi}_j)\, dP(\boldsymbol{\xi}_k);\\ \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) &= \int \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k)\, dP(\boldsymbol{\xi}_k);\\ \widetilde{\kappa}_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) &= \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k),\end{aligned}$$

Then $\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k)$ satisfies the following:

$$\binom{n}{3}^{-1}\sum_{1\le i<j<k\le n}\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) = \sum_{c=0}^{3}\binom{3}{c}U_n^{(c)},$$

where

$$\begin{aligned}U_n^{(0)} &= \theta(P),\\ U_n^{(1)} &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{\kappa}_1(\boldsymbol{\xi}_i) - \theta(P),\\ U_n^{(2)} &= \frac{2}{n(n-1)}\sum_{1\le i<j\le n}\widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \frac{2}{n}\sum_{i=1}^{n}\widetilde{\kappa}_1(\boldsymbol{\xi}_i) + \theta(P),\\ U_n^{(3)} &= \frac{6}{n(n-1)(n-2)}\sum_{1\le i<j<k\le n}\widetilde{\kappa}_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) - \frac{6}{n(n-1)}\sum_{1\le i<j\le n}\widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) + \frac{3}{n}\sum_{i=1}^{n}\widetilde{\kappa}_1(\boldsymbol{\xi}_i) - \theta(P).\end{aligned}$$

We can show the following:

$$\begin{aligned}\widetilde{\kappa}_1(\boldsymbol{\xi}_i) &= 0,\\ \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) &= b^2\frac{e_ie_jw_iw_jR(X_i, X_j)}{g_1(X_i)g_1(X_j)}\{g_2(X_i, X_j) + g_2(X_j, X_i)\}\{1 + O(b)\},\end{aligned}$$

where $R(x_i, x_j) = \mathrm{E}(w(X_{k-1}) \cdots w(X_{k-i+1}) w(X_{k-i-1}) \cdots w(X_{k-p}) | X_k = x_i, X_{k-i} = x_j)$. Thus

$$\begin{aligned}
U_n^{(1)} &= -\theta(P), \\
U_n^{(2)} &= \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) + \theta(P), \\
U_n^{(3)} &= \frac{6}{n(n-1)(n-2)} \sum_{1 \le i < j < k \le n} \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) - \frac{6}{n(n-1)} \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \theta(P) \\
&= \frac{6}{n(n-1)(n-2)} \sum_{1 \le i < j < k \le n} [\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) - \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_k) - \widetilde{\kappa}_2(\boldsymbol{\xi}_j, \boldsymbol{\xi}_k)] - \theta(P) \\
&\equiv \frac{6}{n(n-1)(n-2)} \sum_{1 \le i < j < k \le n} \kappa_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) - \theta(P).
\end{aligned}$$

Combining the above results, we have

$$B_{r111} = \frac{1}{n^2 b^2} \sum_{1 \le i < j < k \le n} \kappa_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) + \frac{n-2}{n^2} \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)/b^2.$$

It follows from Lemma 2 of Yoshihara (1976) (Appendix B) that for any $\epsilon > 0$,

$$\begin{aligned}
P\Big\{ \frac{1}{n^2 b^2} \Big| \sum_{1 \le i < j < k \le n} \kappa_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) \Big| > \epsilon \sqrt{n} \Big\} &\le \frac{n \epsilon^{-2}}{b^4} \mathrm{E} \Big| \frac{1}{n^3} \sum_{1 \le i < j < k \le n} \kappa_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) \Big|^2 \\
&= O(n^{-1} b^{-4}) \to 0,
\end{aligned}$$

and

$$P\Big\{ \frac{1}{n} \Big| \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)/b^2 \Big| > \epsilon \sqrt{n} \Big\} \le n \epsilon^{-2} \mathrm{E} \Big| \frac{1}{n^2} \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)/b^2 \Big|^2 = O(n^{-1}).$$

Thus $B_{r111} = o_p(\sqrt{n})$. Similar (but simpler) arguments may show that $B_{r112} = o_p(\sqrt{n})$ (therefore $B_{r11} = o_p(\sqrt{n})$), $B_{r12} = o_p(\sqrt{n})$ and $B_{r13} = o_p(\sqrt{n})$. Note that Assumption A4 implies $\sqrt{n} b^4 \to 0$. Now argument (i) holds due to (18). The proof is completed.

Proof of Theorem 2

Define

$$\begin{aligned}
\widetilde{Y}_t &= Y_t - \sum_{i=1}^{p} \widetilde{\phi}_i \Big[Y_{t-i} - \widetilde{f}(X_{t-i}) \Big] \\
&= Y_t - \sum_{i=1}^{p} \phi_i \Big[Y_{t-i} - \widetilde{f}(X_{t-i}) \Big] + \sum_{i=1}^{p} (\phi_i - \widetilde{\phi}_i) \Big[Y_{t-i} - \widetilde{f}(X_{t-i}) \Big] \\
&= f(X_t) + \sum_{i=1}^{p} \phi_i e_{t-i} + \varepsilon_t - \sum_{i=1}^{p} \phi_i \Big[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i} \Big] \\
&\quad + \sum_{i=1}^{p} (\phi_i - \widetilde{\phi}_i) \Big[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i} \Big].
\end{aligned}$$

By Theorem 1, $\widetilde{\boldsymbol{\phi}} = \boldsymbol{\phi} + O_p(n^{-1/2})$, the convergence rate is faster than that for the nonparametric estimator $\widehat{f}(x)$. Therefore we may treat $\widetilde{\boldsymbol{\phi}} = \boldsymbol{\phi}$ in the proof, so $\widetilde{Y}_t = \varepsilon_t + f(X_t) + \sum_{i=1}^p \phi_i\{\widetilde{f}(X_{t-i}) - f(X_{t-i})\}$. By Theorem 5.3 of Fan and Yao (2003),

$$s_k^*(x) = p_1(x)\mu_k + O_p\Big\{(\frac{\log n}{nh})^{1/2} + h)\Big\},$$

where $p_1(x) = g_1(x)\mathrm{E}\{w(X_{t-1})w(X_{t-2})\cdots w(X_{t-p})|X_t = x\}$. From Lemma 1 and (11), it holds that

$$\begin{aligned}
\widehat{f}(x) - f(x) &= \frac{1}{nhp_1(x)}\sum_{t=1}^n K\Big(\frac{X_t - x}{h}\Big)\prod_{l=1}^p w(X_{t-l})\Big\{\varepsilon_t + f(X_t) \\
&\quad + \sum_{k=1}^p \phi_k[\widetilde{f}(X_{t-k}) - f(X_{t-k})] - f(x) - \dot{f}(x)(X_t - x)\Big\} \\
&= \frac{1}{nhp_1(x)}\sum_{t=1}^n K\Big(\frac{X_t - x}{h}\Big)\prod_{l=1}^p w(X_{t-l})\Big\{\varepsilon_t + f(X_t) - f(x) - \dot{f}(x)(X_t - x)\Big\} \\
&\quad + \frac{b^2\mu_2}{2nhp_1(x)}\sum_{k=1}^p \phi_k \sum_{t=1}^n K\Big(\frac{X_t - x}{h}\Big)\prod_{l=1}^p w(X_{t-l})\ddot{f}(X_{t-k}) \\
&\quad + \frac{1}{n^2hbp_1(x)}\sum_{k=1}^p \phi_k \sum_{i,j=1}^n K\Big(\frac{X_i - x}{h}\Big)\prod_{l=1}^p w(X_{t-l})K\Big(\frac{X_j - X_{i-k}}{b}\Big)\frac{e_j}{g_1(X_{i-k})}. \quad (19)
\end{aligned}$$

By an ergodic theorem, the second term on the RHS of the above expression is of the order $O_p(b^2) = o_p(h^2)$. To show that the third term on the RHS is of the desired order, we prove it for some particular k, say $k = 1$, the same argument holds for all $k = 1, 2, \cdots, p$. Put

$$\zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = K\Big(\frac{X_i - x}{h}\Big)\prod_{l=1}^p w(X_{i-l})K\Big(\frac{X_j - X_{i-1}}{b}\Big)\frac{e_j}{g_1(X_{i-1})},$$

where $\boldsymbol{\xi}_i = (X_i, X_{i-1}, \cdots, X_{i-p}, e_i)$. Denote the third term on the RHS of (19) as J.

$$\begin{aligned}
J &= \frac{\phi_1}{n^2bhp_1(x)}\sum_{i,j=1}^n \zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \frac{\phi_1}{n^2bhp_1(x)}\sum_{1\le i<j\le n}\Big[\zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) + \zeta(\boldsymbol{\xi}_j, \boldsymbol{\xi}_i)\Big] \\
&\equiv \frac{\phi_1}{n^2bhp_1(x)}\sum_{1\le i<j\le n}\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j).
\end{aligned}$$

Then it holds that

$$J = \frac{\phi_1}{n^2hbp_1(x)}\sum_{1\le i<j\le n}\{\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \kappa_1(\boldsymbol{\xi}_i) - \kappa_1(\boldsymbol{\xi}_j)\} + \frac{\phi_1(n-1)}{n^2p_1(x)}\sum_{i=1}^n \kappa_1(\boldsymbol{\xi}_i)/(hb), \quad (20)$$

where

$$\kappa_1(\boldsymbol{\xi}_i) \equiv \int \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)dP(\boldsymbol{\xi}_j) = hb\, e_i w(X_i)p_2(x, X_i)/g_1(X_i)\{1 + O(h)\},$$

where $p_2(x, X_i) = \mathrm{E}\{w(X_{j-2})\cdots w(X_{j-p})|X_j = x, X_{j-1} = X_i\}g_2(x, X_i)$. Denote the two terms on the RHS of (20) by J_1 and J_2, respectively. By a CLT for mixing processes (e.g., Theorem 2.21(i)

of Fan and Yao 2003), $J_2 = O_p(n^{-1/2}) = o_p\{(nh)^{-1/2}\}$. By Lemma 2 in Appendix 2 below,

$$\begin{aligned} P\{\sqrt{nh}|J_1| > \epsilon\} &\leq \frac{\phi_1^2\epsilon^{-2}nh}{h^2b^2p_1(x)^2}\mathrm{E}\Big|\frac{1}{n^2}\sum_{1\leq i<j\leq n}\{\kappa(\boldsymbol{\xi}_i,\boldsymbol{\xi}_j)-\kappa_1(\boldsymbol{\xi}_i)-\kappa_1(\boldsymbol{\xi}_j)\}\Big|^2 \\ &= O\{(nb^2h)^{-1}\} \to 0. \end{aligned}$$

Hence $J_1 = o_p\{(nh)^{-1/2}\}$. Note $h^2 = O\{(nh)^{-1/2}\}$ under Assumption A4. Now it follows from (19) that

$$\begin{aligned} \widehat{f}(x) - f(x) &= \frac{1}{nhp_1(x)}\sum_{t=1}^{n}K\Big(\frac{X_t-x}{h}\Big)\prod_{l=1}^{p}w(X_{t-l})\{\varepsilon_t + f(X_t) - f(x) - \dot{f}(x)(X_t-x)\} + o_p\Big\{\frac{1}{(nh)^{\frac{1}{2}}}\Big\} \\ &= \frac{1}{nhp_1(x)}\sum_{t=1}^{n}K\Big(\frac{X_t-x}{h}\Big)\prod_{l=1}^{p}w(X_{t-l})\varepsilon_t + \frac{h^2}{2}\mu_2\ddot{f}(x) + o_p\Big\{\frac{1}{(nh)^{\frac{1}{2}}}\Big\}. \end{aligned}$$

Now the theorem follows from, for example, Theorem 2.21(i) of Fan and Yao (2003). The proof is completed.

Proof of Theorem 3

Several algorithms are available to solve the nonlinear optimization problem needed for estimating the ARMA case. Here a nonlinear estimator based on the Gauss-Newton method is adopted. Specifically, given initial estimate $\boldsymbol{\beta}_0 = (\phi_1^0, \cdots, \phi_p^0, \theta_1^0, \cdots, \theta_q^0)^\tau$, we adopt the following notations

$$\phi_0(B)\theta_0(B)^{-1} = \sum_{i=0}^{\infty}\pi_i^0 B^i, \quad \theta_0(B)^{-1} = \sum_{i=0}^{\infty}\xi_i^0 B^i, \quad \phi_0(B)\theta_0(B)^{-2} = \sum_{i=0}^{\infty}\eta_i^0 B^i,$$

and we use the approximations

$$\phi_0(B)\theta_0(B)^{-1}e_t = \sum_{i=0}^{t-1}\pi_i^0 e_{t-i}, \quad \theta_0(B)^{-1}e_t = \sum_{i=0}^{t-1}\xi_i^0 e_{t-i}, \quad \phi_0(B)\theta_0(B)^{-2} = \sum_{i=0}^{t-1}\eta_i^0 e_{t-i}. \tag{21}$$

By a linear Taylor expansion at $\boldsymbol{\beta}_0$, we have

$$\varepsilon_t \approx \frac{\phi_0(B)}{\theta_0(B)}e_t - \sum_{i=1}^{p}\frac{1}{\theta_0(B)}e_{t-i}\Delta\phi_i + \sum_{j=1}^{q}\frac{\phi_0(B)}{\theta_0^2(B)}e_{t-j}\Delta\theta_j,$$

where $\Delta\phi_i = \phi_i - \phi_i^0$ and $\Delta\theta_j = \theta_j - \theta_j^0$. By the approximations in (21), we have the following regression equation

$$\sum_{i=0}^{t-1}\pi_i^0 e_{t-i} = \sum_{j=1}^{p}\sum_{i=0}^{t-j-1}\xi_i^0 e_{t-j-i}\Delta\phi_i - \sum_{j=1}^{q}\sum_{i=0}^{t-j-1}\eta_i^0 e_{t-j-i}\Delta\theta_i + \varepsilon_t.$$

Let $m = \max(p,q)+1$, $\Delta\boldsymbol{\beta}$ can be estimated by minimizing

$$\sum_{t=m}^{n}\Big\{\sum_{i=0}^{t-1}\pi_i^0 e_{t-i} - \sum_{j=1}^{p}\sum_{i=0}^{t-j-1}\xi_i^0 e_{t-j-i}\Delta\phi_i + \sum_{j=1}^{q}\sum_{i=0}^{t-j-1}\eta_i^0 e_{t-j-i}\Delta\theta_i\Big\}^2$$

with respect to $\Delta\boldsymbol{\phi}$ and $\Delta\boldsymbol{\theta}$, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \widehat{\Delta\boldsymbol{\beta}}$ serves as the estimate of $\boldsymbol{\beta}$. Therefore we minimize

$$\sum_{j=1}^{n}\sum_{t=m}^{n}\Big\{Y_t - a_0 - a_1(X_t - X_j) + \sum_{l=1}^{t-1}\pi_l^0\widetilde{e}_{t-l} - \sum_{i=1}^{p}\sum_{l=0}^{t-i-1}\xi_l^0\widetilde{e}_{t-i-l}\Delta\phi_i + \sum_{i=1}^{q}\sum_{l=0}^{t-i-1}\eta_l^0\widetilde{e}_{t-i-l}\Delta\theta_i\Big\}^2 K_h(X_t - X_j)$$

to estimate $f(\cdot)$ and $\boldsymbol{\beta}$. Re-express the above in matrix notation, for initial estimate $\boldsymbol{\beta}_0$, let

$$D_t^\tau = \Big(\frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\phi_1}, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\phi_2}, \cdots, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\phi_p}, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\theta_1}, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\theta_2}, \cdots, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\theta_q}\Big),$$

where $\partial\varepsilon_t(\boldsymbol{\beta}_0)/\partial\beta_i,\ i = 1, \cdots, p+q$ means $\partial\varepsilon_t/\partial\beta_i$ evaluated at $\boldsymbol{\beta}_0$. By a Taylor expansion,

$$\varepsilon_t \approx \varepsilon_t(\boldsymbol{\beta}_0) + D_t^\tau(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \varepsilon_t(\boldsymbol{\beta}_0) + D_t^\tau\Delta\boldsymbol{\beta},$$

where $\varepsilon_t(\boldsymbol{\beta}_0) = \theta_0(B)^{-1}\phi_0(B)e_t$. Re-arranging terms, we have $\varepsilon_t(\boldsymbol{\beta}_0) = -D_t^\tau\Delta\boldsymbol{\beta} + \varepsilon_t$. An estimate of $\Delta\boldsymbol{\beta}$ can be obtained by minimizing the sum of squares $\sum_{t=1}^{n}\{\varepsilon_t(\boldsymbol{\beta}_0) + D_t^\tau\Delta\boldsymbol{\beta}\}^2$. Define

$$\mathbf{D} = -\begin{pmatrix} \frac{\partial\varepsilon_m(\boldsymbol{\beta}_0)}{\partial\phi_1} & \frac{\partial\varepsilon_m(\boldsymbol{\beta}_0)}{\partial\phi_2} & \cdots & \frac{\partial\varepsilon_m(\boldsymbol{\beta}_0)}{\partial\phi_p} & \frac{\partial\varepsilon_m(\boldsymbol{\beta}_0)}{\partial\theta_1} & \frac{\partial\varepsilon_m(\boldsymbol{\beta}_0)}{\partial\theta_2} & \cdots & \frac{\partial\varepsilon_m(\boldsymbol{\beta}_0)}{\partial\theta_q} \\ \frac{\partial\varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial\phi_1} & \frac{\partial\varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial\phi_2} & \cdots & \frac{\partial\varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial\phi_p} & \frac{\partial\varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial\theta_1} & \frac{\partial\varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial\theta_2} & \cdots & \frac{\partial\varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial\theta_q} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial\varepsilon_n(\boldsymbol{\beta}_0)}{\partial\phi_1} & \frac{\partial\varepsilon_n(\boldsymbol{\beta}_0)}{\partial\phi_2} & \cdots & \frac{\partial\varepsilon_n(\boldsymbol{\beta}_0)}{\partial\phi_p} & \frac{\partial\varepsilon_n(\boldsymbol{\beta}_0)}{\partial\theta_1} & \frac{\partial\varepsilon_n(\boldsymbol{\beta}_0)}{\partial\theta_2} & \cdots & \frac{\partial\varepsilon_n(\boldsymbol{\beta}_0)}{\partial\theta_q} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{e_{m-1}}{\theta_0(B)} & \frac{e_{m-2}}{\theta_0(B)} & \cdots & \frac{e_{m-p}}{\theta_0(B)} & -\frac{\phi_0(B)e_{m-1}}{\theta_0^2(B)} & -\frac{\phi_0(B)e_{m-2}}{\theta_0^2(B)} & \cdots & -\frac{\phi_0(B)e_{m-q}}{\theta_0^2(B)} \\ \frac{e_m}{\theta_0(B)} & \frac{e_{m-1}}{\theta_0(B)} & \cdots & \frac{e_{m-p+1}}{\theta_0(B)} & -\frac{\phi_0(B)e_m}{\theta_0^2(B)} & -\frac{\phi_0(B)e_{m-1}}{\theta_0^2(B)} & \cdots & -\frac{\phi_0(B)e_{m-q+1}}{\theta_0^2(B)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{e_{n-1}}{\theta_0(B)} & \frac{e_{n-2}}{\theta_0(B)} & \cdots & \frac{e_{n-p}}{\theta_0(B)} & -\frac{\phi_0(B)e_{n-1}}{\theta_0^2(B)} & -\frac{\phi_0(B)e_{n-2}}{\theta_0^2(B)} & \cdots & -\frac{\phi_0(B)e_{n-q}}{\theta_0^2(B)} \end{pmatrix}.$$

Let

$$\mathbf{u} = \Big(\frac{\phi_0(B)}{\theta_0(B)}e_m, \frac{\phi_0(B)}{\theta_0(B)}e_{m+1}, \cdots, \frac{\phi_0(B)}{\theta_0(B)}e_n\Big)^\tau.$$

By the same approximations in (21), we have the "regressor" matrix

$$\underline{\mathbf{D}} = \begin{pmatrix} \sum_{i=0}^{m-2}\xi_i^0 e_{m-1-i} & \cdots & \sum_{i=0}^{m-p-1}\xi_i^0 e_{m-p-i} & -\sum_{i=0}^{m-2}\eta_i^0 e_{m-2-i} & \cdots & -\sum_{i=0}^{m-q-1}\eta_i^0 e_{m-q-i} \\ \sum_{i=0}^{m-1}\xi_i^0 e_{m-i} & \cdots & \sum_{i=0}^{m-p}\xi_i^0 e_{m-p+1-i} & -\sum_{i=0}^{m-1}\eta_i^0 e_{m-1-i} & \cdots & -\sum_{i=0}^{m-q}\eta_i^0 e_{m-q+1-i} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=0}^{n-2}\xi_i^0 e_{n-1-i} & \cdots & \sum_{i=0}^{n-p-1}\xi_i^0 e_{n-p-i} & -\sum_{i=0}^{n-2}\eta_i^0 e_{n-2-i} & \cdots & -\sum_{i=0}^{n-q-1}\eta_i^0 e_{n-q-i} \end{pmatrix},$$

and

$$\underline{\mathbf{u}} = \Big(\sum_{i=0}^{m-1}\pi_i^0 e_{m-i}, \sum_{i=0}^{m}\pi_i^0 e_{m+1-i}, \cdots, \sum_{i=0}^{n-1}\pi_i^0 e_{n-i}\Big)^\tau.$$

The estimate of $\boldsymbol{\beta}$ can be obtained by $\boldsymbol{\beta}_0 + \widehat{\Delta\boldsymbol{\beta}}_{\text{Ideal}}$, where $\widehat{\Delta\boldsymbol{\beta}}_{\text{Ideal}}$ is the "idealized" estimator of $\Delta\boldsymbol{\beta}$ obtained from "observations" $\{e_t\}$:

$$\widehat{\Delta\boldsymbol{\beta}}_{\text{Ideal}} = (\underline{\mathbf{D}}^\tau\underline{\mathbf{D}})^{-1}\underline{\mathbf{D}}^\tau\underline{\mathbf{u}}.$$

The estimate of $\boldsymbol{\beta}$ based on the initial estimate of the innovation process $\widetilde{e}_t = Y_t - \widetilde{f}(X_t)$, denoted by $\widetilde{\boldsymbol{\beta}}$, is obtained similarly as $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \widetilde{\Delta\boldsymbol{\beta}}$, where $\widetilde{\Delta\boldsymbol{\beta}} = (\mathbf{D}_1^\tau \mathbf{D}_1)^{-1}\mathbf{D}_1^\tau \mathbf{u}_1$, $\mathbf{D}_1$ and $\mathbf{u}_1$ are defined similarly as $\underline{\mathbf{D}}$ and $\underline{\mathbf{u}}$, with e_t replaced by $\widetilde{e}_t$.

The proof of the theorem is complete by showing

(i) $\mathbf{D}_1^\tau \mathbf{D}_1 = \underline{\mathbf{D}}^\tau \underline{\mathbf{D}} + o_p(\sqrt{n})$, and
(ii) $\mathbf{D}_1^\tau \mathbf{u}_1 = \underline{\mathbf{D}}^\tau \underline{\mathbf{u}} + o_p(\sqrt{n})$.

However, to save the space we have to omit the quite lengthy proof here. For detailed proof, please see a technical report by Liu, Chen and Yao (2005).

Proof of Theorem 4

Define

$$\begin{aligned}
\widetilde{Y}_t &= Y_t + \sum_{i=1}^{t-1} \widetilde{\pi}_i [Y_{t-i} - \widetilde{f}(X_{t-i})] \\
&= f(X_t) - \sum_{i=1}^{\infty} \pi_i e_{t-i} + \varepsilon_t + \sum_{i=1}^{t-1} \pi_i [Y_{t-i} - \widetilde{f}(X_{t-i})] + \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[Y_{t-i} - \widetilde{f}(X_{t-i})] \\
&= f(X_t) + \varepsilon_t - \sum_{i=1}^{\infty} \pi_i e_{t-i} + \sum_{i=1}^{t-1} \pi_i [f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}] \\
&\quad + \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}] \\
&= f(X_t) + \varepsilon_t - \sum_{i=t}^{\infty} \pi_i e_{t-i} + \sum_{i=1}^{t-1} \pi_i [f(X_{t-i}) - \widetilde{f}(X_{t-i})] + \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}]
\end{aligned}$$

By Theorem 5.3 of Fan and Yao (2003), we have

$$\begin{aligned}
&\widehat{f}(x) - f(x) \\
&= \frac{1}{nhg_1(x)} \sum_{t=1}^{n} K(\frac{X_t - x}{h}) \Big\{ f(X_t) + \varepsilon_t - f(x) - \dot{f}(x)(X_t - x) + \sum_{i=1}^{t-1} \pi_i [f(X_{t-i}) - \widetilde{f}(X_{t-i})] \\
&\quad - \sum_{i=t}^{\infty} \pi_i e_{t-i} + \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}] \Big\} \\
&= \frac{1}{nhg_1(x)} \sum_{t=1}^{n} K(\frac{X_t - x}{h}) \Big\{ f(X_t) - f(x) - \dot{f}(x)(X_t - x) + \varepsilon_t \Big\} \\
&\quad + \frac{1}{nhg_1(x)} \sum_{t=2}^{n} K(\frac{X_t - x}{h}) \sum_{i=1}^{t-1} \pi_i [f(X_{t-i}) - \widetilde{f}(X_{t-i})] - \frac{1}{nhg_1(x)} \sum_{t=2}^{n} \sum_{i=t}^{\infty} K(\frac{X_t - x}{h}) \pi_i e_{t-i} \\
&\quad + \frac{1}{nhg_1(x)} \sum_{t=2}^{n} \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i) K(\frac{X_t - x}{h}) [f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}] \\
&\equiv S_1 + S_2 + S_3 + S_4
\end{aligned}$$

By a Taylor expansion and Lemma 1, we can show that the remainder term in S_1 related to $R_n(\cdot)$ is ignorable and we only need to consider the leading term of S_1:

$$\frac{1}{nhg_1(x)}\sum_{t=1}^{n} K(\frac{X_t - x}{h})\varepsilon_t + \frac{h^2}{2}\mu_2 \ddot{f}(x).$$

By Theorem 2.21 of Fan and Yao (2003), the proof is complete by showing $S_2 + S_3 + S_4$ is of order $o_p\{(nh)^{-1/2}\}$. Again, the proof of this theorem is quite lengthy, hence omitted here. For detailed proof, please refer to Liu, Chen and Yao (2005).

Appendix B – A note on Lemma 2 of Yoshihara (1976)

Yoshihara (1976) is influential as it establishes asymptotic properties of U-statistics for strictly stationary and β-mixing processes. Its lemma 2, which estimates the orders for the second moments of residual terms in the Hoeffding decomposition, appears to have an error in presentation, since γ in (2.12) of Yoshihara (1976) may be arbitrarily large by choosing $\delta' > 0$ arbitrarily small. (Note that we may let $\delta' > 0$ arbitrarily small for, for example, independent processes.) We state below a rectified version of the lemma, which can be derived in the same manner as the proof in the original paper. All the notation and citation below are referred to Yoshihara (1976).

Lemma 2 *(Yoshihara 1976) . If there is a positive number δ such that for $r = 2 + \delta$ (2.3) and (2.4) hold, and $\sum_{n\geq 1} n\beta(n)^{\delta/(2+\delta)} < \infty$, then we have*

$$E(U_n^{(c)})^2 = O(n^{-2}), \qquad 2 \leq c \leq m.$$

Note that we impose a stronger condition on the mixing coefficients $\beta(n)$, and the rate $O(n^{-2})$ is optimal.

References

Auestad, B. and D. Tjöstheim, 1990, Identification of nonlinear time series: First order characterization and order estimation. *Biometrika* **77**, 669-687.

Box, G.E.P. and G.M. Jenkins, 1976, *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1st ed.

Brockwell, P.J. and R.A. Davis, 1987, *Time Series: Theory and Methods*. Springer-Verlag, New York.

Cai, Z., J. Fan and Q. Yao, 2000, Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**, 941–956.

Carroll, R.J., J. Fan, I. Gijbels and M.P. Wand, 1997, Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489.

Chen, R. and R.S. Tsay, 1996, Nonlinear transfer functions. *Journal of Nonparametric Statistics* **66**, 193-204.

Chen, R. and R.S. Tsay, 1993a, Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**, 298-308.

Chen, R. and R.S. Tsay, 1993b, Nonlinear additive ARX models, *Journal of the American Statistical Association* **88**, 955-967.

Craven, P. and G.Wahba, 1979, Smoothing noisy data with spline functions. *Numerical Mathematics* **31**, 377-403.

Fan, J. and I. Gilbels, 1996, *Local Polynomial Modeling and Its Applications.* Chapman and Hall, Suffolk.

Fan, J. and Q. Yao, 2003, *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

Fan, J., Q. Yao and Z. Cai, 2003, Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B* **65**, 57–80.

Hädle, W., H. Lütkepohl, and R. Chen, 1997, A review of nonparametric time series analysis. *International Statistical Review* **65**, 49-72.

Härdle, W., P. Hall, and H. Ichimura, 1993, Optimal smoothing in single-index models. *The Annals of Statistics* **21**, 157–178.

Härdle, W., H. Liang, and J. Gao, 2000, *Partially Linear Models.* Physica-Verlag, Heidelberg.

Haggan V. and T. Ozaki, 1981, Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68**, 189196.

Hart, J.D., 1996, Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics* **6** 115-142, 1996.

Heckman, J., H. Ichimura, J. Smith, and P. Todd, 1998, Characterizing selection bias using experimental data. *Econometrica* **66**, 1017–1098.

Ichiruma, H., 1993, Semiparametric least-squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Liu, J.M., R. Chen and Q. Yao, 2005, Nonparametric Transfer Function Models. *Technical report, Georgia Southern University.*

Liu, L.M. and D.M. Hanssens, 1982, Identification of multiple-input transfer function models. *Communications in Statistics* **A11**, 297-314.

Masry, E., 1996a, Multivariate local polynomial regression for time series: Uniform consistency and rates. *Journal of Time Series Analysis* **17**, 571-599.

Masry, E., 1996b, Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and Their Applications* **65**, 81-101.

Newbold, P., 1973, Bayesain estimation of Box-Jenkins transfer function-noise models. *Journal of the Royal Statistical Society* **35**, 323-336.

Newey, W.K., and T.M. Stoker, 1993, Efficiency of weighted average derivative estimators and index models. *Econometrica* **61**, 1199-1223.

Ozaki, T., 1985, Statistical identification of storage models with application to stochastic hydrology. *Water Resources Bulletin* **21**, 663-675.

Poskitt, D.S., 1989, A method for the estimation and identification of transfer function models. *Journal of the Royal Statistical Society* **B51**, 29-46.

Robinson, P.M., 1983, Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185-207.

Ruckstuhl, A., A.H. Welsh, and R.J. Carroll, 2000, Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica* **10**, 51-71.

Severini, T.A. and J.G. Staniswalis, 1994, Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* **89**, 501-511.

Smith, M., C.M. Wong, and R. Kohn, 1998, Additive nonparametric regression with autocorrelated errors. *Journal of the Royal Statistical Society* **60**, 311-331.

Su, L. and A. Ullah, 2006, More efficient estimation in nonparametric regression with nonparametric autocorrelated errors. *Econometric Theory* **22**, 98-126.

Tiao, G.C. and G.E.P. Box, 1981, Modeling multiple time series with applications. *Journal of the American Statistical Association* **76**, 802-816.

Tjøstheim, D., 1994, Nonlinear time series: a selective review. *Scandinavian Journal of Statistics* **21**, 97-130.

Tsay, R.S., 1985, Model identification in dynamic regression (distributed lag) models. *Journal of Business Economic Statistics* **3**, 228-237.

Wild, C.J. and T.W. Yee, 1996, Additive extensions to generalized estimation equation methods. *Journal of the Royal Statistical Society: Series B* **58**, 711-725.

Wu, C.O., C.T. Chiang, and D.R. Hoover, 1998, Asymptotic confidence regions for kernel smoothing of a varying coefficient model with longitudinal data. *Journal of the American Statistical Association* **93**, 1388-1402.

Xia, Y. and W.K. Li, 1999, On single-index coefficient regression models. *Journal of the American Statistical Association* **94**, 1275-1285.

Xia, Y., H. Tong, W.K. Li and L. Zhu, 2002, An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 363-410.

Xiao, Z., O.B. Linton, R.J. Carroll, and E. Mammen, 2003, Model efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association* **98**, 980-992.

Yoshihara, K., 1976, Limiting behavior of U-statistics for a stationary absolutely regular process. *Zeitschrift fur Wahrscheinlichkeitstheorie verw.* Gebiete, **35**, 237-252.

Zeger, S.L. and P.J. Diggle, 1994, Semiparametric models for longitudinal data with application to CD4 cell number in HIV seroconverters. *Biometrics* **50**, 789-699.