JOURNAL OF Econometrics

NONLINEAR AND NONPARAMETRIC METHODS IN ECONOMETRICS

edited by
SONGNIAN CHEN AND QI LI

Nonparametric transfer function models

Jun M. Liu [a,1], Rong Chen [b,c,*,2], Qiwei Yao [b,d,3]

[a] Georgia Southern University, United States

[b] Peking University, China

[c] Rutgers University, United States

[d] London School of Economics, United Kingdom

ARTICLE INFO

ABSTRACT

In this paper a class of nonparametric transfer function models is proposed to model nonlinear relationships between 'input' and 'output' time series. The transfer function is smooth with unknown functional forms, and the noise is assumed to be a stationary autoregressive-moving average (ARMA) process. The nonparametric transfer function is estimated jointly with the ARMA parameters. By modeling the correlation in the noise, the transfer function can be estimated more efficiently. The parsimonious ARMA structure improves the estimation efficiency in finite samples. The asymptotic properties of the estimators are investigated. The finite-sample properties are illustrated through simulations and one empirical example.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Linear transfer function models (Box et al., 1994) have been extensively used to model the relationship between 'output' and 'input' time series. With one input X_t, 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, e_t follows an ARMA process, and $\alpha(B)$, $\beta(B)$ are polynomials of the *backshift operator B* defined as $B^i X_t \equiv X_{t-i}$. Linear transfer function models have been well studied and proved to be 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 $Y_t = f(X_{t-d}, \ldots, X_{t-d-p}; \theta) + N_t$, where $f(\cdot)$ is a parametric function assuming the Volterra series representation, N_t is stationary and modeled as an ARMA process.

A problem common to nonlinear parametric models is that it is difficult to justify the explicit parametric functional forms *a priori*, because of the large number of candidate functions. Following the principle of 'letting the data speak for themselves', nonparametric smoothing methods provide a more flexible alternative to model nonlinear time series (Lewis and Stevens, 1991; Robinson, 1983; Auestad and Tjøstheim, 1990; Masry, 1996a,b; Fan and Gilbels, 1996; Smith et al., 1998, e.g.,). 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 et al., 2000), the *nonlinear additive autoregressive model* (Chen and Tsay, 1993b), the *adaptive functional-coefficient model* (Ichiruma, 1993; Xia and Li, 1999; Fan et al., 2003), the *single index model* (Härdle et al., 1993; Carroll et al., 1997; Newey and Stoker, 1993; Heckman et al., 1998; Xia et al., 2002) and the *partially linear models* (Härdle et al., 2000). The literature about nonlinear and nonparametric time series analysis is vast, some reviews can be found in Tjøstheim (1994), Hädle et al. (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. The processes $\{X_t\}$ and $\{e_t\}$ are assumed to be strictly stationary. The transfer function $f(\cdot)$ is modeled via nonparametric smoothing and the innovation process $\{e_t\}$ is modeled as a stationary and invertible ARMA(p, q) process, i.e., $\phi(B)e_t = \theta(B)\varepsilon_t$, where $\phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i$, $\theta(B) = 1 - \sum_{j=1}^{q}\theta_j B^j$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\tau$ and

* Corresponding address: Department of Statistics, Hill Center, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019, United States. Tel.: +1 732 445 2691; fax: +1 732 445 3428.

E-mail address: rongchen@stat.rutgers.edu (R. Chen).

1 Assistant Professor of Quantitative Methods, Department of Finance & Quantitative Analysis, Georgia Southern University, United States.

2 Professor of Statistics, Department of Business Statistics and Econometrics, Peking University, China and Department of Statistics, Rutgers University, United States.

3 Professor of Statistics, London School of Economics, United Kingdom and Department of Business Statistics and Econometrics, Peking University, China.

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\tau$ are unknown parameters and $\{\varepsilon_t\}$ is a sequence of independent random variables with mean 0 and variance σ^2. Furthermore, $\{X_t\}$ and $\{\varepsilon_t\}$ are assumed to be independent, which implies the independence between $\{X_t\}$ and $\{e_t\}$. A multi-step procedure is used to estimate both the transfer function and the ARMA parameters.

By modeling the transfer function $f(\cdot)$ nonparametrically, the model is flexible therefore may be used to model nonlinear relationships 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. Furthermore, 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. 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 et al., 1998; Ruckstuhl et al., 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 et al. (2003); Su and Ullah (2006) considered problems similar to the one considered in this paper. These studies are closely related, but major difference exists, especially in the treatment of the noise $\{e_t\}$. In Xiao et al. (2003), $\{e_t\}$ 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 noise 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) is a special case of ARMA(p, q), different algorithms are used and different approaches are needed to prove the results. 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 are 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 in one real data example 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. Method for models with AR noise

2.1. The algorithm

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

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

Under normal assumption and with observations $\{(X_t, Y_t)\}_{t=1}^n$, the maximum likelihood estimation for $f(\cdot)$ and $\boldsymbol{\phi}$ boils down to the following optimization problem:

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

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

We first obtain a preliminary estimator for $f(\cdot)$ by local linear regression, ignoring the correlation in $\{e_t\}$. Namely, $\widetilde{f}(x) = \widetilde{a}_0$, and $(\widetilde{a}_0, \widetilde{a}_1)$ minimizes

$$\sum_{t=1}^n \{Y_t - a_0 - a_1(X_t - x)\}^2 K_b(X_t - x), \tag{3}$$

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

$$\widetilde{f}(x) - f(x)$$
$$= \frac{1}{nb} \sum_{t=1}^n W_n\left(\frac{X_t - x}{b}, x\right) \{Y_t - f(x) - \dot{f}(x)(X_t - x)\}, \tag{4}$$

where

$$W_n(t, x) = (1, 0)S_n(x)^{-1}\begin{pmatrix} 1 \\ t \end{pmatrix}K(t). \tag{5}$$

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

$$s_k(x) = \frac{1}{n} \sum_{t=1}^n \left(\frac{X_t - x}{b}\right)^k K_b(X_t - x). \tag{6}$$

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

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

and $\mathbf{W} = \text{diag}\{\prod_{i=0}^p w(X_{t-i})\}$, where $w(\cdot)$ is a weight function controlling the boundary effect in nonparametric estimation. The following estimation procedure is used,

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

$$\widetilde{\boldsymbol{\phi}} = (\mathbf{X}_1^\tau \mathbf{W} \mathbf{X}_1)^{-1} \mathbf{X}_1^\tau \mathbf{W} \mathbf{Y}_1. \tag{7}$$

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

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

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

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

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

Remark 1. In practice, we repeat Steps 2 and 3 above until convergence. 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. Replacing $\widetilde{f}$ with the newly estimated function values may improve the results, though convergence is no longer guaranteed. Furthermore such a replacement is not necessary at least asymptotically; see Theorem 2 below.

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. 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 (3) 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. Otherwise the least squares-based estimators, such as local polynomial estimators, might not be consistent. This assumption essentially forbids the use of lagged Y's as explanatory variables. When lagged Y's 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. See also Xiao et al. (2003).

2.2. Asymptotic results

In the subsequent discussion, we denote the moments of the kernel function by $\mu_i \equiv \int u^i K(u) \mathrm{d}u$. Let

$$
\mathbf{X}_2 = \begin{pmatrix} e_p & e_{p-1} & \cdots & e_1 \\ e_{p+1} & e_p & \cdots & e_2 \\ \cdots & \cdots & \cdots & \\ e_{n-1} & e_{n-2} & \cdots & e_{n-p} \end{pmatrix}, \qquad \mathbf{Y}_2 = \begin{pmatrix} e_{p+1} \\ e_{p+2} \\ \cdots \\ e_n \end{pmatrix}.
$$

Define the "idealized" estimator

$$
\widehat{\boldsymbol{\phi}}_{\mathrm{Ideal}} = (\mathbf{X}_2^\tau \mathbf{W} \mathbf{X}_2)^{-1} \mathbf{X}_2^\tau \mathbf{W} \mathbf{Y}_2,
$$

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 observable. It has been shown (e.g. Brockwell and Davis, 1991) that

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

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 $\widehat{\boldsymbol{\phi}}_{\mathrm{Ideal}}$.

Theorem 1. Let conditions (A1)–(A6) in Appendix A hold, and $\boldsymbol{\phi}$ satisfy the stationarity condition. Then as $n \to \infty$,

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

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

By Theorem 1, $\widetilde{\boldsymbol{\phi}}$ shares the same asymptotic distribution of $\widehat{\boldsymbol{\phi}}_{\mathrm{Ideal}}$, i.e.,

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

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

$$
\widehat{f}(x) - f(x) = \frac{1}{nh} \sum_{t=1}^n W_n^*\left(\frac{X_t - x}{h}, x, X_{t-1}, \ldots, X_{t-p}\right)
$$
$$
\times \left\{\widetilde{Y}_t - f(x) - \dot{f}(x)(X_t - x)\right\}, \qquad (11)
$$

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

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

and $S_n^*(x)$ is defined in the same manner as $S_n(x)$ in (6) with $K_b(X_t - x)$ replaced by $K_h(X_t - x) \prod_{i=1}^p w(X_{t-i})$ (See also (4)). 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.

Theorem 2. Let conditions (A1) to (A6) in Appendix A hold. For any x in the support of X_t, it holds as $n \to \infty$ that

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

where

$$
\sigma(x)^2 = \frac{\sigma^2 \int K(u)^2 \mathrm{d}u}{g_1(x)}
$$
$$
\times \frac{\mathrm{E}\left\{\left[W(X_{t-1})W(X_{t-2})\cdots W(X_{t-p})\right]^2 | X_t = x\right\}}{\left\{\mathrm{E}\left[W(X_{t-1})W(X_{t-2})\cdots W(X_{t-p})|X_t = x\right]\right\}^2}, \qquad (12)
$$

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

This theorem shows that the nonparametric transfer function estimator $\widehat{f}(\cdot)$ is more efficient than the "conventional" local polynomial estimator which ignores the correlation in e_t. If the conventional estimator is used, the asymptotic variance would have the same form as (12), but with σ^2 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. In fact the gain in efficiency of $\widehat{f}(\cdot)$ over the conventional estimator is larger with stronger autocorrelation in e_t.

3. Method for models with ARMA noise

Here we consider the 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 method is different and the proof for the asymptotic results requires different techniques.

3.1. The algorithm

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

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

Since $\{e_t\}$ is assumed to be stationary and invertible, it admits the linear process representations $e_t = -\sum_{i=1}^\infty \pi_i e_{t-i} + \varepsilon_t$ and $e_t = \sum_{i=0}^\infty \psi_i \varepsilon_{t-i}$, where 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 et al., 1994). Let

$\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)^\tau$. $f(\cdot)$ and $\boldsymbol{\beta}$ are estimated by solving the following nonlinear optimization problem

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

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. The three-step estimation procedure is described as follows:

1. Obtain an initial estimate $\widetilde{f}(\cdot)$ by local linear regression, ignoring the serial correlation in $\{e_t\}$ and using b as the bandwidth (see Eq. (3) in the Section 2).
2. Given $\widetilde{f}$, obtain $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\phi}}^\tau, \widehat{\boldsymbol{\theta}}^\tau)^\tau$ by minimizing

$$\sum_{t=1}^{n} \left\{ \frac{\phi(B)}{\theta(B)} \left[Y_t - \widetilde{f}(X_t) \right] \right\}^2 \qquad (14)$$

with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.
3. Given $\widehat{\boldsymbol{\beta}}$, let $\widehat{f_j} \equiv \widehat{f}(X_j) = \widehat{a}_0$, where $(\widehat{a}_0, \widehat{a}_1)$ minimizes

$$\sum_{t=1}^{n} \left\{ Y_t - a_0 - a_1(X_t - X_j) + \left[\frac{\phi(B)}{\theta(B)} - 1 \right] [Y_t - \widetilde{f}(X_t)] \right\}^2 \\ \times K_h(X_t - X_j),$$

where $K_h(\cdot) = 1/h K(\cdot/h)$, h is a bandwidth and h is of larger order than b.

In practice, steps 2 and 3 may be iterated to improve the finite-sample performance, with $\widetilde{f}$ updated by the estimate of the previous iteration. Given $\widetilde{f}$ and (p, q), standard ARMA estimators such as MLE and the least squares estimator (LSE) may be used. In this study, we adopt a nonlinear estimation method which has the same asymptotic distribution as MLE and LSE. A detailed description of this method is given in Appendix A. This estimator is used mainly because it has a closed-form expression, which greatly simplifies the technical proof of the asymptotic properties. We believe that the same asymptotic results hold when MLE or LSE is used, but the technical proofs will be more involved. We also find through simulation that the finite-sample performance of this estimator is similar to that of the MLE (Table 4) when used in nonparametric transfer function models. More details of this nonlinear estimator can be found in textbooks such as Box et al. (1994, Section 7.2) and Brockwell and Davis (1991, Section 8.11).

3.2. Asymptotic results

Similar to the AR(p) case, we construct the "idealized" estimator $\widehat{\boldsymbol{\beta}}_{\text{Ideal}}$ by treating $\{e_t\}$ as observed and using the nonlinear ARMA estimator mentioned above. It has been shown that (e.g. Brockwell and Davis, 1991)

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}}_{\text{Ideal}} - \boldsymbol{\beta} \right) \xrightarrow{D} N \left(0, \sigma^2 \mathbf{V}(\boldsymbol{\beta})^{-1} \right),$$

where

$$\mathbf{V}(\boldsymbol{\beta}) = \mathrm{E} \begin{pmatrix} \mathbf{U}_1 \mathbf{U}_1^\tau & \mathbf{U}_1 \mathbf{V}_1^\tau \\ \mathbf{V}_1 \mathbf{U}_1^\tau & \mathbf{V}_1 \mathbf{V}_1^\tau \end{pmatrix}, \qquad (15)$$

$\mathbf{U}_t = (U_t, U_{t-1}, \ldots, U_{t+1-p})^\tau$, $\mathbf{V}_t = (V_t, V_{t-1}, \ldots, V_{t+1-q})^\tau$. $\{U_t\}$ is an AR(p) process defined by $\phi(B) U_t = a_t$ and $\{V_t\}$ is an AR(q) process defined by $\theta(B) V_t = b_t$. Here both a_t and b_t are white noise processes. Obviously, when the model does not contain the AR component (a pure MA (q) model), we have $\mathbf{V}(\boldsymbol{\beta}) = \mathrm{E}\left(\mathbf{V}_1 \mathbf{V}_1^\tau \right)$.

Theorem 3. *Let conditions* (A1) *to* (A5) *and* (A6*) *in Appendix A hold, $\boldsymbol{\phi}$ satisfy the stationarity condition and $\boldsymbol{\theta}$ satisfy the invertibility condition. Then as $n \to \infty$, it holds that*

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

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

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

where $\boldsymbol{\beta} = (\boldsymbol{\phi}^\tau, \boldsymbol{\theta}^\tau)^\tau$, and $\mathbf{V}(\boldsymbol{\beta})$ is defined in (15).

Theorem 4. *Let conditions* (A1) *to* (A5) *and* (A6*) *in Appendix A hold, $\{e_t\}$ be a stationary and invertible* ARMA(p, q) *process. For any x in the support of X_t, it holds as $n \to \infty$ that*

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

where

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

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) and Su and Ullah (2006) under different assumptions on e_t.

4. Numerical properties

In practice we must determine the ARMA orders p and q first. Since e_t is not observable, we rely on $\widetilde{e}_t \equiv Y_t - \widetilde{f}(X_t)$ instead to identify the model, where $\widetilde{f}$ is the conventional local linear regression estimator. A simulation is conducted to investigate the performance of AIC, BIC and AICC in order determination based on $\widetilde{e}_t$. In the simulation we set

$$f(X_t) = \sin(4X_t) + \cos(2X_t), \qquad (16)$$

$X_t = 0.3 X_{t-1} + a_t$, and $e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$, where a_t are independent $N(0, 0.3^2)$ random variables, and ε_t are independent $N(0, 0.5^2)$ variables. Sample sizes are set at $n = 100, 200$ and 400. Two hundred samples are drawn from each setting. The Gaussian kernel is used with the bandwidth $b = 1.06 s_X n^{-9/40}$. For different ϕ, θ combinations, Table 1 reports the percentage of times that the correct orders are identified by the three criteria (AIC$_1$, BIC$_1$ and AICC$_1$, respectively in the table). As a comparison, the same model selection is done using the real noise e_t in each replication and the percentages are also reported in this table (AIC$_2$, BIC$_2$ and AICC$_2$). As expected, BIC provides the best performance. Overall the performance of those criteria is satisfactory and is in line with the case when the observations from a time series are directly available. Therefore in the simulation study for $\widehat{f}$ below the order for the ARMA model is assumed to be known.

To investigate the finite-sample properties of the proposed estimator $\widehat{f}$, we continue to use the above setting but with $\{e_t\}$ restricted to three cases: 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. Different bandwidths b and h are experimented. Since the results are not very sensitive with respect to the bandwidths, only those with $h = 1.06 s_X n^{-1/5}$ and $b = 1.06 s_X n^{-9/40}$ are reported here. The methods included in the numerical comparison are NPTF—the proposed nonparametric transfer function method, XLCM—the AR approximation approach of Xiao et al. (2003), SU—the nonparametric AR approximation approach of Su and Ullah (2006), and WHITE—the conventional local linear estimator treating e_t as a while noise.

Table 1
ARMA order selection.

ϕ_1	ϕ_2	θ_1	θ_2	n	AIC_1	AIC_2	BIC_1	BIC_2	$AICC_1$	$AICC_2$
0.5				100	51.5	53.0	89.5	90.0	59.0	58.5
				200	58.5	60.0	98.0	98.5	61.5	64.0
				400	62.0	59.5	96.0	96.0	62.5	60.5
0.8				100	48.5	51.0	90.0	91.0	54.5	57.5
				200	53.0	54.0	93.0	95.0	56.5	55.0
				400	63.0	60.5	96.0	94.0	64.5	62.0
0.5	0.25			100	38.0	46.5	48.0	54.0	42.5	49.5
				200	57.5	59.0	73.0	79.0	61.5	61.5
				400	62.5	62.0	97.5	98.0	62.5	62.0
1.1	−0.4			100	47.5	58.5	62.5	88.0	49.5	62.0
				200	52.5	58.0	87.5	95.5	57.5	61.0
				400	46.0	57.5	91.0	97.0	48.5	61.5
		0.5		100	74.5	75.5	95.0	97.0	77.5	78.0
				200	68.0	70.0	97.0	97.5	70.0	71.0
				400	72.0	71.5	98.5	97.5	72.5	72.0
		0.8		100	70.0	58.5	94.5	89.5	73.5	62.0
				200	70.5	71.5	98.0	97.0	71.0	72.5
				400	74.0	76.5	99.0	99.0	75.0	76.5
		0.5	0.25	100	54.0	61.0	47.0	60.0	56.0	64.5
				200	74.0	72.0	82.0	84.0	75.5	74.5
				400	73.5	71.5	98.5	99.5	73.5	71.5
		1.1	−0.4	100	59.0	60.5	73.5	81.0	63.5	62.5
				200	75.0	73.5	96.0	95.0	77.0	75.5
				400	74.0	72.0	99.5	98.5	75.5	72.5
0.5		−0.8		100	23.5	31.5	49.0	64.5	28.5	36.0
				200	25.0	28.0	75.5	85.0	28.5	34.5
				400	25.0	29.0	93.0	95.0	28.5	30.0
0.5		−0.5		100	22.0	27.0	31.5	55.0	24.5	34.0
				200	26.5	28.0	68.5	83.5	30.0	35.5
				400	26.0	27.5	86.5	91.0	28.0	28.5
−0.5	−0.5	0.2	0.7	100	28.5	27.0	17.0	41.5	31.0	33.0
				200	39.5	39.5	64.5	68.0	42.0	44.0
				400	43.5	44.0	94.0	91.0	46.0	44.5

The average mean squares error MSE $= n^{-1} \sum_{1 \le t \le n} \{\widehat{f}(X_t) - f(X_t)\}^2$ over 200 replications are computed for all four methods. Tables 2 and 3 report the simulation results, in which the entries in the AMSE column stand for the average MSE of WHITE, the entries in NPTF, XLCM and SU columns are the ratio of the corresponding average MSE to the average MSE of WHITE. The means and standard deviations of $\widehat{\phi}$ and $\widehat{\theta}$ from NPTF are also included. A histogram of $\widehat{\phi}$ and a plot of a typical simulation are given in Fig. 1.

Some findings from the simulation are summarized as follows: (1) the NPTF estimator $\widehat{f}(\cdot)$ is more efficient than the simple local linear regression estimator WHITE. 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; and (3) some of the ARMA estimates have large bias (e.g., $(\phi, \theta) = (0.5, -0.8)$ and $(0.8, -0.8)$) and larger sample sizes are needed to improve the performance. Fig. 1 indicates that the sampling distributions of $\widehat{\phi}$ are close to the asymptotic normal distribution. In general, NPTF, XLCM and SU are more efficient than WHITE. When $\{e_t\}$ follows an ARMA model with small $|\theta|$ (including pure AR models), NPTF and XLCM have similar efficiency. However when $|\theta|$ is large, NPTF is more efficient. For the seasonal models, NPTF has similar gain in efficiency as in the non-seasonal models, while XLCM and SU often fail to approximate $\{e_t\}$ appropriately, leading to low efficiencies; see Table 3. Numerical results (not reported here) indicate that for the cases with seasonal $\{e_t\}$, the efficiency of XLCM improves as the AR order increases, though it is still not as efficient as NPTF, partially due to the additional errors introduced in estimating more parameters. On the other hand, SU is not as efficient as NPTF and XLCM for $\{e_t\}$ generated from ARMA models, as it is designed for the models with nonlinear AR $\{e_t\}$.

The nonlinear estimator used in the ARMA(p, q) cases essentially minimizes an approximation of the true sum of squares. Although it has the same asymptotic distribution as the MLE and the LSE, it might not be as efficient in finite samples. To better understand the consequence of the approximation, we compare the performances of this nonlinear estimator (NLE in what follows) and the MLE in step 2 of the proposed estimation procedure (Eq. (14)), the results are summarized in Table 4. We can see that although generally the MLE performs better than the NLE, the difference is rather small, especially for larger sample sizes.

5. Example: River flow and rainfall

In this section we use the proposed method 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 (parametric) nonlinear transfer function (NLTF) model. For details of the data, see Ozaki (1985) or Chen and Tsay (1996). The performance of the nonparametric transfer function model is compared with those of the NLTF model and the linear transfer function model (LTF) in this section.

Let Y_t be river flow and X_t be rain fall. The time series plot and the correlogram of Y_t (not shown here to save space) indicate certain non-stationarity (more detailed results, including the omitted plots, can be found in Liu et al. (2005)). To formally test the existence of a unit root, we perform the Augmented Dickey–Fuller (ADF) test (Dickey and Fuller, 1981) using the following model:

Table 2
Simulation results: AR(1) and MA(1) models.

ϕ	θ	n	Mean $(\widehat{\phi})$, $s_{\widehat{\phi}}$	Mean $(\widehat{\theta})$, $s_{\widehat{\theta}}$	AMSE	NPTF	XLCM	SU
		100	−0.795, 0.070		0.030	0.518	0.534	0.683
−0.8		200	−0.801, 0.048		0.020	0.452	0.460	0.594
		400	−0.804, 0.030		0.012	0.468	0.494	0.587
		100	−0.490, 0.090		0.019	0.799	0.824	0.998
−0.5		200	−0.501, 0.065		0.012	0.768	0.781	0.961
		400	−0.502, 0.047		0.007	0.775	0.784	0.945
		100	−0.191, 0.112		0.017	0.974	0.995	1.19
−0.2		200	−0.203, 0.072		0.011	1.08	1.08	1.19
		400	−0.195, 0.052		0.007	0.942	0.952	1.12
		100	0.187, 0.111		0.020	1.00	1.02	1.18
0.2		200	0.191, 0.083		0.012	1.02	1.03	1.15
		400	0.195, 0.050		0.008	1.08	1.08	1.21
		100	0.486, 0.104		0.028	0.941	0.947	1.05
0.5		200	0.488, 0.071		0.018	0.922	0.927	1.02
		400	0.495, 0.045		0.010	0.920	0.922	1.00
		100	0.769, 0.079		0.092	0.858	0.861	0.893
0.8		200	0.787, 0.046		0.057	0.801	0.801	0.830
		400	0.794, 0.036		0.031	0.812	0.813	0.839
		100		−0.747, 0.084	0.032	0.895	0.940	1.04
	−0.8	200		−0.756, 0.056	0.019	0.892	0.877	0.982
		400		−0.774, 0.034	0.010	0.871	0.885	0.972
		100		−0.477, 0.091	0.024	0.965	0.989	1.09
	−0.5	200		−0.490, 0.066	0.014	0.964	0.976	1.05
		400		−0.498, 0.047	0.009	0.945	0.969	1.06
		100		−0.193, 0.128	0.019	1.03	1.13	1.18
	−0.2	200		−0.194, 0.072	0.012	1.04	1.06	1.17
		400		−0.200, 0.049	0.008	1.01	1.02	1.11
		100		0.206, 0.112	0.017	0.969	1.08	1.24
	0.2	200		0.204, 0.071	0.010	0.957	0.970	1.15
		400		0.200, 0.049	0.007	0.959	0.984	1.13
		100		0.494, 0.069	0.018	0.842	0.874	1.09
	0.5	200		0.499, 0.045	0.011	0.797	0.841	1.03
		400		0.494, 0.034	0.010	0.819	0.842	1.00
		100		0.708, 0.094	0.022	0.768	0.761	0.965
	0.8	200		0.722, 0.066	0.014	0.715	0.731	0.865
		400		0.750, 0.043	0.008	0.706	0.727	0.866
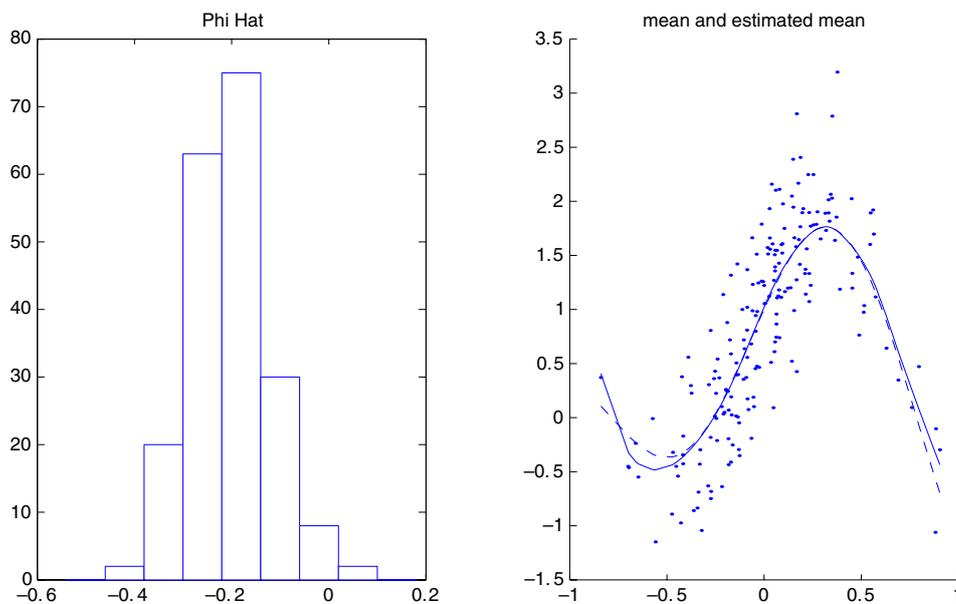
Fig. 1. $\phi = -0.2$, $n = 200$. Left panel: histogram of $\widehat{\phi}$, right panel: true (solid line) and estimated (dashed line) transfer function in a typical simulation.

$Y_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t$, which can be re-written as $\Delta Y_t = \beta_0 + \beta_1 t + \gamma Y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta Y_{t-i} + \varepsilon_t$, where t denotes time, $\Delta Y_{t-i} \equiv Y_{t-i} - Y_{t-i-1}$, $i = 0, \ldots, p-1$, $\gamma = \sum_{i=1}^{p} \phi_i - 1$, and $\gamma_i = -\sum_{k=i+1}^{p} \phi_k$. The AR order $p = 5$ is selected by BIC. A likelihood ratio test is conducted to test the hypothesis $H_0 : \beta_0 = \beta_1 = \gamma = 0$ against the general alternative. The value of the

Table 3
Simulation results: ARMA(1, 1), AR(1)$_4$ and MA(1)$_4$ models.

ϕ	θ	n	Mean $(\widehat{\phi})$, $s_{\widehat{\phi}}$	Mean $(\widehat{\theta})$, $s_{\widehat{\theta}}$	AMSE	NPTF	XLCM	SU
		100	0.172, 0.140	−0.716, 0.115	0.037	0.812	0.872	0.946
0.2	−0.8	200	0.192, 0.094	−0.736, 0.069	0.025	0.846	0.847	0.912
		400	0.197, 0.058	−0.755, 0.045	0.014	0.779	0.820	0.867
		100	0.463, 0.130	−0.668, 0.117	0.079	0.809	0.806	0.855
0.5	−0.8	200	0.487, 0.069	−0.693, 0.077	0.042	0.745	0.725	0.773
		400	0.491, 0.048	−0.722, 0.048	0.022	0.716	0.730	0.789
		100	0.758, 0.083	−0.596, 0.126	0.289	0.767	0.790	0.818
0.8	−0.8	200	0.782, 0.047	−0.633, 0.087	0.169	0.766	0.776	0.805
		400	0.790, 0.029	−0.681, 0.056	0.092	0.714	0.731	0.741
		100	0.172, 0.167	−0.493, 0.163	0.032	0.922	0.957	0.992
0.2	−0.5	200	0.179, 0.111	−0.506, 0.101	0.018	0.894	0.869	0.967
		400	0.197, 0.075	−0.493, 0.070	0.011	0.870	0.900	0.951
		100	0.466, 0.122	−0.476, 0.126	0.052	0.879	0.918	0.910
0.5	−0.5	200	0.488, 0.079	−0.467, 0.092	0.029	0.828	0.834	0.837
		400	0.490, 0.052	−0.484, 0.053	0.016	0.763	0.785	0.845
		100	0.759, 0.074	−0.426, 0.112	0.189	0.787	0.790	0.829
0.8	−0.5	200	0.782, 0.048	−0.452, 0.073	0.104	0.741	0.756	0.782
		400	0.789, 0.031	−0.471, 0.053	0.065	0.730	0.749	0.763

ϕ_4	θ_4	n	Mean $(\widehat{\phi}_4)$, $s_{\widehat{\phi}_4}$	Mean $(\widehat{\theta}_4)$, $s_{\widehat{\theta}_4}$	AMSE	NPTF	XLCM	SU
		100	−0.762, 0.077		0.038	0.503	1.01	1.13
−0.8		200	−0.784, 0.049		0.022	0.465	0.996	1.09
		400	−0.791, 0.032		0.013	0.475	0.992	1.01
		100	−0.481, 0.093		0.021	0.834	1.02	1.20
−0.5		200	−0.491, 0.069		0.012	0.831	0.997	1.16
		400	−0.492, 0.047		0.008	0.819	1.00	1.15
		100		0.493, 0.096	0.019	0.873	1.03	1.23
	0.5	200		0.500, 0.072	0.012	0.905	1.03	1.18
		400		0.492, 0.043	0.007	0.855	1.03	1.18
		100		0.683, 0.090	0.024	0.887	1.08	1.21
	0.8	200		0.715, 0.064	0.015	0.817	1.01	1.14
		400		0.745, 0.038	0.008	0.785	0.994	1.12

Table 4
Performance of the MLE and the NLE in estimating ϕ and θ.

			MLE		NLE	
ϕ	θ	n	Mean $(\widehat{\phi})$, $s_{\widehat{\phi}}$	Mean $(\widehat{\theta})$, $s_{\widehat{\theta}}$	Mean $(\widehat{\phi})$, $s_{\widehat{\phi}}$	Mean $(\widehat{\theta})$, $s_{\widehat{\theta}}$
		100	0.758, 0.078	−0.621, 0.110	0.754, 0.081	−0.603, 0.115
0.8	−0.8	200	0.783, 0.047	−0.657, 0.086	0.783, 0.048	−0.636, 0.088
		400	0.788, 0.031	−0.690, 0.058	0.788, 0.030	−0.681, 0.059
		100	0.194, 0.158	−0.487, 0.154	0.175, 0.163	−0.500, 0.146
0.2	−0.5	200	0.180, 0.119	−0.499, 0.113	0.175, 0.119	−0.500, 0.110
		400	0.200, 0.080	−0.492, 0.073	0.197, 0.080	−0.492, 0.073

test statistic is 4.40 with p-value between 0.05 and 0.10 (Table V in Dickey and Fuller, 1981), which suggests that a unit root may exist in the process. Similar analysis is performed on X_t and the results show that it is stationary (details omitted). Based on the above analysis, we take first-order difference $Z_t = Y_t - Y_{t-1}$ and consider the model

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

Note here a low-dimensional smoothing model is used instead of an univariate smoothing model. Following the proposed estimation procedures, $\widetilde{f}(\cdot)$ is first obtained by local linear regression assuming $\{e_t\}$ i.i.d., then we obtain the partial residual $\widetilde{e}_t = Z_t - \widetilde{f}(X_t)$. Based on $\widetilde{e}_t$, AIC, BIC and AICC are used to select from ARMA models of orders up to 20. BIC identifies an ARMA(1, 1) model. However, after this model is fitted to $\widetilde{e}_t$, the portmanteau test still shows autocorrelation in the residuals. On the other hand, AIC and AICC select an AR model of orders 1, 4, 5, 6, and 14. Considering the ACF of $\widetilde{e}_t$ (Fig. 2), after some refinement, we select an AR model of orders 4, 5, 6, and 14. Although the BIC of this AR model is slightly higher than that of the ARMA(1, 1) model, it does a better job in removing the correlation in the residuals.

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

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

where $\mathbf{S}_h$ is defined by the equation 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 $\text{tr}(\mathbf{S}_h)$ is also calculated. The resulting bandwidth is 6 and the equivalent number of parameters is 31.2. The estimated AR parameters are $\widehat{\phi}_4 = 0.091$, $\widehat{\phi}_5 = 0.137$, $\widehat{\phi}_6 = 0.150$, and $\widehat{\phi}_{14} = 0.076$. The t-ratios are calculated using the limiting distribution in Theorem 1 and are all greater than 2 in absolute value. The Ljung–Box statistic $Q_{LB}(12) = 9.95$ suggests absence of serial correlation in the final residuals. The ACF plot of the final residuals also indicates the residual is roughly "white" (Fig. 3).

To study the forecasting performance of the NPTF model, the following rolling forecasting scheme is employed: for each $t = 180, 181, \ldots, 365$, data available at t are used to build the model and make one-step ahead prediction and the forecasting error
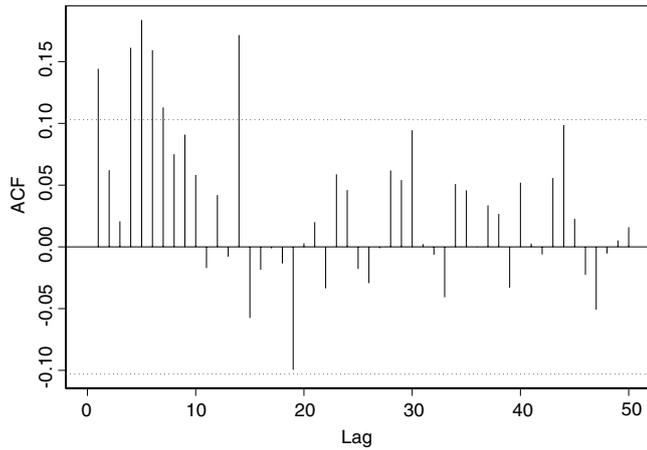
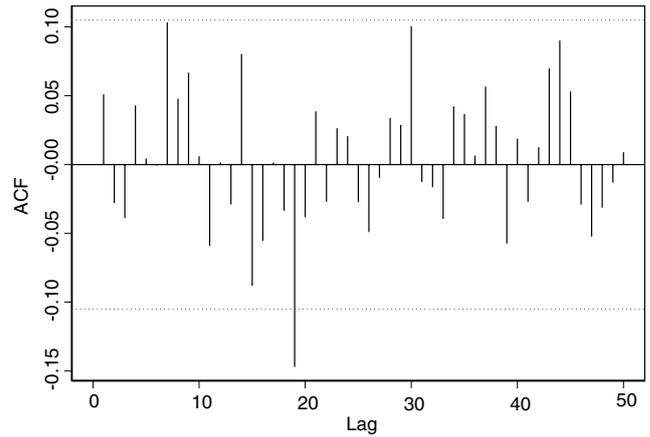**Fig. 2.** Sample ACF plots of the partial residuals $\widetilde{e}_t$.

Fig. 3. Sample ACF plot of the final residuals.

$Y_{t+1} - \widehat{Y}_t(1)$ is calculated. The squared forecasting errors are averaged over t. The square-root of this average is referred to as "post-sample forecasting RMSE". To put the performance of the NPTF model in perspective, the same rolling forecast is computed using the parametric NLTF and the LTF models entertained by Chen and Tsay (1996). The results are summarized in Table 5.

The above results show that the residual variance of the NPTF model is the smallest among the three models. Although it uses more parameters than the other two models, the gain in post-sample forecasting justifies the use of this more sophisticated model.

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

Table 5
Within- and post-sample comparisons.

Models	NPTF	NLTF	LTF
(Equivalent) Number of parameters	31.2	12	10
Residual variance	5.45	6.23	20.81
Forecasting RMSE	8.82	12.56	13.93

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
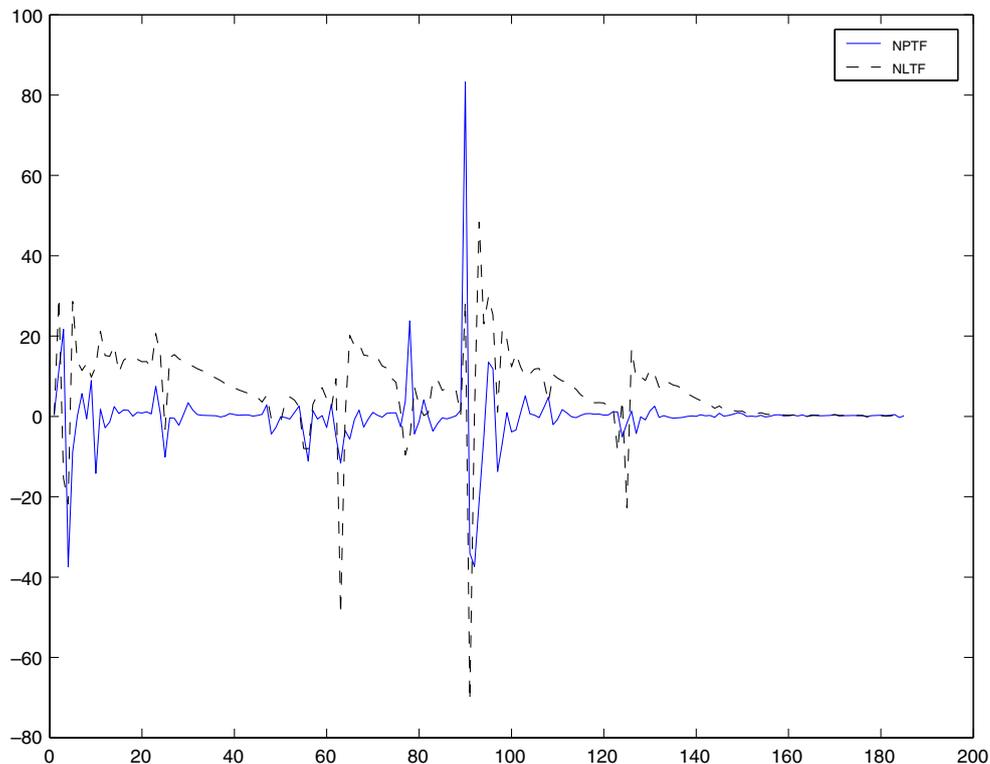

Fig. 4. The one-step ahead forecast errors of the NPTF model and the NLTF model.

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 $Y_t = f(X_{1t}, \ldots, X_{pt}) + e_t$. However, such a direct generalization is often not practical in practice due to the "curse of dimensionality". To solve this problem, more restrictive models, such as the additive model, may be considered. Research addressing this topic is ongoing.

Acknowledgements

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 $g_i(x_{t1}, \ldots, x_{ti})$ be the i-dimensional joint density function of $\{X_{t1}, \ldots, X_{ti}\}$. 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
$$\beta(k) = E\{\sup_{B \in \mathcal{F}_k^\infty} |P(B) - P(B|X_0, X_{-1}, \ldots)|\} \to 0$$
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 $K(\cdot)$ is compactly supported and Lipschitz continuous. Additionally, we require $\mu_0 = \int K(u)du = 1$, and $\mu_1 = \int uK(u)du = 0$, i.e, $K(u)$ is a second-order kernel.

(A3) $f(\cdot)$ has continuous second derivative $\ddot{f}(\cdot)$, $g_1(\cdot)$ is continuous and 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 $g_2(\cdot, \cdot)$, $g_4(\cdot, \cdot, \cdot, \cdot)$ and $g_6(\cdot, \cdot, \cdot, \cdot, \cdot, \cdot)$ are continuous and have bounded first derivative.

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

where $\mu_2 = \int u^2 K(u)du$, and
$$R_n(x) = \frac{1}{nbg_1(x)}\left\{\left|\sum_{t=1}^{n} K\left(\frac{X_t - x}{b}\right) e_t\right|\right.$$
$$\left. + \left|\sum_{t=1}^{n}\left(\frac{X_t - x}{b}\right) K\left(\frac{X_t - x}{b}\right) e_t\right|\right\} + O(b^2).$$

Proof of Lemma 1. It follows from Theorem 5.3 of Fan and Yao (2003) that
$$s_k(x) = g_1(x)\mu_k + O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2} + b^2\right\}$$
uniformly for $x \in A$, where $s_k(x)$ is defined in (6), $\mu_k = \int u^k K(u)du$, and A is any compact set contained in $\{g_1(x) > 0\}$. Hence it holds uniformly for $x \in A$ that
$$S_n(x) = S(x) + O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2} + b^2\right\},$$
where $S(x) = g_1(x)\text{diag}(1, \mu_2)$. Write $Y_t^* = Y_t - f(x) - \dot{f}(x)(X_t - x)$. It is easy to see from (5) that
$$\left|\sum_{t=1}^{n}\left\{W_n\left(\frac{X_t - x}{b}, x\right) - g_1(x)^{-1}K\left(\frac{X_t - x}{b}\right)\right\} Y_t^*\right|$$
$$= \left|(1, 0)\{S_n(x)^{-1} - S(x)^{-1}\} \sum_{t=1}^{n}\left(1, \frac{X_t - x}{b}\right)^\tau K\left(\frac{X_t - x}{b}\right) Y_t^*\right|$$
$$\le \left[(1, 0)\{S_n(x)^{-1} - S(x)^{-1}\}^2(1, 0)^\tau\right]^{1/2}$$
$$\times \left\{\left|\sum_{t=1}^{n} K\left(\frac{X_t - x}{b}\right) Y_t^*\right|^2\right.$$
$$\left. + \left|\sum_{t=1}^{n}\frac{X_t - x}{b}K\left(\frac{X_t - x}{b}\right) Y_t^*\right|^2\right\}^{1/2}$$
$$\le \left[(1, 0)\{S_n(x)^{-1} - S(x)^{-1}\}^2(1, 0)^\tau\right]^{1/2}$$
$$\times \left\{\left|\sum_{t=1}^{n} K\left(\frac{X_t - x}{b}\right) Y_t^*\right| + \left|\sum_{t=1}^{n}\frac{X_t - x}{b}K\left(\frac{X_t - x}{b}\right) Y_t^*\right|\right\}$$
$$\le O_p\left[\left\{\left(\frac{\log n}{nb}\right)^{1/2} + b^2\right\}^{1/2}\right]$$
$$\times \left\{\left|\sum_{t=1}^{n} K\left(\frac{X_t - x}{b}\right) e_t\right|\right.$$
$$\left. + \left|\sum_{t=1}^{n}\frac{X_t - x}{b}K\left(\frac{X_t - x}{b}\right) e_t\right| + O(nb^3)\right\}.$$

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 (4) 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 $\mathbf{A} = \mathbf{X}_1^\tau \mathbf{W} \mathbf{X}_1$ and $\mathbf{B} = \mathbf{X}_1^\tau \mathbf{W} \mathbf{Y}_1$,

where $\mathbf{X}_1$, $\mathbf{Y}_1$ and $\mathbf{W}$ are defined in Section 2.1. From (7) we have $\widetilde{\boldsymbol{\phi}} = \mathbf{A}^{-1}\mathbf{B}$, the (r, s)th element of $\mathbf{A}$ is

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

where

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

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

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

The rth element of $\mathbf{B}$ is

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

where

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

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

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

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 $r, s = 1, 2, \ldots, p$.

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

$$
\begin{aligned}
B_{r11} &= \frac{1}{n^2 b^2} \sum_{i,j,k} K\left(\frac{X_i - X_k}{b}\right) K\left(\frac{X_j - X_{k-r}}{b}\right) \\
&\quad \times \frac{e_i e_j}{g_1(X_k)g_1(X_{k-r})} \prod_{l=0}^{p} w_{k-l} \\
&\equiv \frac{1}{n^2 b^2} \sum_{i,j,k} \zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k),
\end{aligned}
$$

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

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

where $\boldsymbol{\xi}_i = (X_i, X_{i-1}, \ldots, X_{i-p}, e_i)^\tau$. 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

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

Define

$$
\theta(P) = \iiint \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) \mathrm{d}P(\boldsymbol{\xi}_i)\mathrm{d}P(\boldsymbol{\xi}_j)\mathrm{d}P(\boldsymbol{\xi}_k);
$$

$$
\widetilde{\kappa}_1(\boldsymbol{\xi}_i) = \iint \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) \mathrm{d}P(\boldsymbol{\xi}_j)\mathrm{d}P(\boldsymbol{\xi}_k);
$$

$$
\widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \int \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) \mathrm{d}P(\boldsymbol{\xi}_k);
$$

$$
\widetilde{\kappa}_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) = \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k).
$$

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

$$
U_n^{(0)} = \theta(P),
$$

$$
U_n^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\kappa}_1(\boldsymbol{\xi}_i) - \theta(P),
$$

$$
U_n^{(2)} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \frac{2}{n} \sum_{i=1}^{n} \widetilde{\kappa}_1(\boldsymbol{\xi}_i) + \theta(P),
$$

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

We can show the following:

$$
\widetilde{\kappa}_1(\boldsymbol{\xi}_i) = 0,
$$

$$
\begin{aligned}
\widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) &= b^2 \frac{e_i e_j w_i w_j R(X_i, X_j)}{g_1(X_i)g_1(X_j)} \\
&\quad \times \{g_2(X_i, X_j) + g_2(X_j, X_i)\}\{1 + O(b)\},
\end{aligned}
$$

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

$$
U_n^{(1)} = -\theta(P),
$$

$$
U_n^{(2)} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) + \theta(P),
$$

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

$$-\widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_k) - \widetilde{\kappa}_2(\boldsymbol{\xi}_j, \boldsymbol{\xi}_k)] - \theta(P)$$

$$\equiv \frac{6}{n(n-1)(n-2)}$$

$$\times \sum_{1 \le i < j < k \le n} \kappa_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) - \theta(P).$$

Combining the above results, we have

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

$$\times \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)/b^2.$$

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

$$P\left\{ \frac{1}{n^2 b^2} \left| \sum_{1 \le i < j < k \le n} \kappa_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) \right| > \epsilon \sqrt{n} \right\}$$

$$\le \frac{n\epsilon^{-2}}{b^4} \mathrm{E} \left| \frac{1}{n^3} \sum_{1 \le i < j < k \le n} \kappa_3(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k) \right|^2$$

$$= O(n^{-1} b^{-4}) \to 0,$$

and

$$P\left\{ \frac{1}{n} \left| \sum_{1 \le i < j \le n} \widetilde{\kappa}_2(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)/b^2 \right| > \epsilon \sqrt{n} \right\}$$

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

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. $\square$

Proof of Theorem 2. Define

$$\widetilde{Y}_t = Y_t - \sum_{i=1}^p \widetilde{\phi}_i \left[Y_{t-i} - \widetilde{f}(X_{t-i}) \right]$$

$$= Y_t - \sum_{i=1}^p \phi_i \left[Y_{t-i} - \widetilde{f}(X_{t-i}) \right] + \sum_{i=1}^p (\phi_i - \widetilde{\phi}_i) \left[Y_{t-i} - \widetilde{f}(X_{t-i}) \right]$$

$$= f(X_t) + \sum_{i=1}^p \phi_i e_{t-i} + \varepsilon_t - \sum_{i=1}^p \phi_i \left[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i} \right]$$

$$+ \sum_{i=1}^p (\phi_i - \widetilde{\phi}_i) \left[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i} \right].$$

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),

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

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

$$\widehat{f}(x) - f(x) = \frac{1}{nhp_1(x)} \sum_{t=1}^n K\left(\frac{X_t - x}{h}\right) \prod_{l=1}^p w(X_{t-l}) \left\{ \varepsilon_t + f(X_t) \right.$$

$$\left. + \sum_{k=1}^p \phi_k[\widetilde{f}(X_{t-k}) - f(X_{t-k})] - f(x) - \dot{f}(x)(X_t - x) \right\}$$

$$= \frac{1}{nhp_1(x)} \sum_{t=1}^n K\left(\frac{X_t - x}{h}\right) \prod_{l=1}^p w(X_{t-l})$$

$$\times \left\{ \varepsilon_t + f(X_t) - f(x) - \dot{f}(x)(X_t - x) \right\} + \frac{b^2 \mu_2}{2nhp_1(x)}$$

$$\times \sum_{k=1}^p \phi_k \sum_{t=1}^n K\left(\frac{X_t - x}{h}\right) \prod_{l=1}^p w(X_{t-l}) \ddot{f}(X_{t-k})$$

$$+ \frac{1}{n^2 hbp_1(x)} \sum_{k=1}^p \phi_k \sum_{i,j=1}^n K\left(\frac{X_i - x}{h}\right)$$

$$\times \prod_{l=1}^p w(X_{t-l}) K\left(\frac{X_j - X_{i-k}}{b}\right) \frac{e_j}{g_1(X_{i-k})}. \tag{19}$$

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 $k = 1, 2, \ldots, p$. Put

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

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

$$J = \frac{\phi_1}{n^2 bhp_1(x)} \sum_{i,j=1}^n \zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$$

$$= \frac{\phi_1}{n^2 bhp_1(x)} \sum_{1 \le i < j \le n} \left[\zeta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) + \zeta(\boldsymbol{\xi}_j, \boldsymbol{\xi}_i) \right]$$

$$\equiv \frac{\phi_1}{n^2 bhp_1(x)} \sum_{1 \le i < j \le n} \kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j).$$

Then it holds that

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

$$+ \frac{\phi_1(n-1)}{n^2 p_1(x)} \sum_{i=1}^n \kappa_1(\boldsymbol{\xi}_i)/(hb), \tag{20}$$

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 B below,

$$P\left\{ \sqrt{nh}|J_1| > \epsilon \right\} \le \frac{\phi_1^2 \epsilon^{-2} nh}{h^2 b^2 p_1(x)^2}$$

$$\times \mathrm{E} \left| \frac{1}{n^2} \sum_{1 \le i < j \le n} \{\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \kappa_1(\boldsymbol{\xi}_i) - \kappa_1(\boldsymbol{\xi}_j)\} \right|^2$$

$$= O\{(nb^2 h)^{-1}\} \to 0.$$

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

$$\widehat{f}(x) - f(x) = \frac{1}{nh p_1(x)} \sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right) \prod_{l=1}^{p} w(X_{t-l})$$

$$\times \{\varepsilon_t + f(X_t) - f(x) - \dot{f}(x)(X_t - x)\} + o_p\left\{\frac{1}{(nh)^{\frac{1}{2}}}\right\}$$

$$= \frac{1}{nh p_1(x)} \sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right) \prod_{l=1}^{p} w(X_{t-l}) \varepsilon_t + \frac{h^2}{2} \mu_2 \ddot{f}(x)$$

$$+ o_p\left\{\frac{1}{(nh)^{\frac{1}{2}}}\right\}.$$

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

Proof of Theorem 3. In this paper a nonlinear estimation method is used to estimate the ARMA parameters, detailed descriptions of this method can be found in, for example, Box et al. (1994) and Brockwell and Davis (1991). Here we briefly describe how this method is applied in the proposed estimator. Given initial estimate $\boldsymbol{\beta}_0 = (\phi_1^0, \ldots, \phi_p^0, \theta_1^0, \ldots, \theta_q^0)^\tau$, we adopt the following notations

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

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

and we use the approximations

$$\phi_0(B)\theta_0(B)^{-1}e_t = \sum_{i=0}^{t-1} \pi_i^0 e_{t-i},$$

$$\theta_0(B)^{-1}e_t = \sum_{i=0}^{t-1} \xi_i^0 e_{t-i}, \qquad (21)$$

$$\phi_0(B)\theta_0(B)^{-2} = \sum_{i=0}^{t-1} \eta_i^0 e_{t-i}.$$

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

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

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

$$\sum_{t=m}^{n} \left\{ \sum_{i=0}^{t-1} \pi_i^0 e_{t-i} - \sum_{j=1}^{p} \sum_{i=0}^{t-j-1} \xi_i^0 e_{t-j-i}\Delta\phi_i \right.$$

$$\left. + \sum_{j=1}^{q} \sum_{i=0}^{t-j-1} \eta_i^0 e_{t-j-i}\Delta\theta_i \right\}^2$$

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

$$\sum_{j=1}^{n} \sum_{t=m}^{n} \left\{ Y_t - a_0 - a_1(X_t - X_j) + \sum_{l=1}^{t-1} \pi_l^0 \widetilde{e}_{t-l} \right.$$

$$\left. - \sum_{i=1}^{p} \sum_{l=0}^{t-i-1} \xi_l^0 \widetilde{e}_{t-i-l}\Delta\phi_i + \sum_{i=1}^{q} \sum_{l=0}^{t-i-1} \eta_l^0 \widetilde{e}_{t-i-l}\Delta\theta_i \right\}^2 K_h(X_t - X_j)$$

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

$$D_t^\tau = \left(\frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\phi_1}, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\phi_2}, \ldots, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\phi_p}, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\theta_1}, \right.$$

$$\left. \times \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\theta_2}, \ldots, \frac{\partial\varepsilon_t(\boldsymbol{\beta}_0)}{\partial\theta_q} \right),$$

where $\partial\varepsilon_t(\boldsymbol{\beta}_0)/\partial\beta_i$, $i = 1, \ldots, p + q$ 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 $\sum_{t=1}^{n}\{\varepsilon_t(\boldsymbol{\beta}_0) + D_t^\tau\Delta\boldsymbol{\beta}\}^2$. Let $\mathbf{D}$ be as shown in Box I. Let

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

By the same approximations in (21), the 'regressor' matrix becomes that in Box II, and

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

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 et al. (2005). $\square$

Proof of Theorem 4. Define

$$\widetilde{Y}_t = Y_t + \sum_{i=1}^{t-1} \widetilde{\pi}_i[Y_{t-i} - \widetilde{f}(X_{t-i})]$$

$$= f(X_t) - \sum_{i=1}^{\infty} \pi_i e_{t-i} + \varepsilon_t + \sum_{i=1}^{t-1} \pi_i[Y_{t-i} - \widetilde{f}(X_{t-i})]$$

$$+ \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[Y_{t-i} - \widetilde{f}(X_{t-i})]$$

$$= f(X_t) + \varepsilon_t - \sum_{i=1}^{\infty} \pi_i e_{t-i} + \sum_{i=1}^{t-1} \pi_i[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}]$$

$$+ \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}]$$

$$= f(X_t) + \varepsilon_t - \sum_{i=t}^{\infty} \pi_i e_{t-i} + \sum_{i=1}^{t-1} \pi_i[f(X_{t-i}) - \widetilde{f}(X_{t-i})]$$

$$+ \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}].$$

$$\mathbf{D} = - \begin{pmatrix} \dfrac{\partial \varepsilon_m(\boldsymbol{\beta}_0)}{\partial \phi_1} & \dfrac{\partial \varepsilon_m(\boldsymbol{\beta}_0)}{\partial \phi_2} & \cdots & \dfrac{\partial \varepsilon_m(\boldsymbol{\beta}_0)}{\partial \phi_p} & \dfrac{\partial \varepsilon_m(\boldsymbol{\beta}_0)}{\partial \theta_1} & \dfrac{\partial \varepsilon_m(\boldsymbol{\beta}_0)}{\partial \theta_2} & \cdots & \dfrac{\partial \varepsilon_m(\boldsymbol{\beta}_0)}{\partial \theta_q} \\[2mm] \dfrac{\partial \varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial \phi_1} & \dfrac{\partial \varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial \phi_2} & \cdots & \dfrac{\partial \varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial \phi_p} & \dfrac{\partial \varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial \theta_1} & \dfrac{\partial \varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial \theta_2} & \cdots & \dfrac{\partial \varepsilon_{m+1}(\boldsymbol{\beta}_0)}{\partial \theta_q} \\[2mm] \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\[2mm] \dfrac{\partial \varepsilon_n(\boldsymbol{\beta}_0)}{\partial \phi_1} & \dfrac{\partial \varepsilon_n(\boldsymbol{\beta}_0)}{\partial \phi_2} & \cdots & \dfrac{\partial \varepsilon_n(\boldsymbol{\beta}_0)}{\partial \phi_p} & \dfrac{\partial \varepsilon_n(\boldsymbol{\beta}_0)}{\partial \theta_1} & \dfrac{\partial \varepsilon_n(\boldsymbol{\beta}_0)}{\partial \theta_2} & \cdots & \dfrac{\partial \varepsilon_n(\boldsymbol{\beta}_0)}{\partial \theta_q} \end{pmatrix}$$

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

Box I.

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

Box II.

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

$$\widehat{f}(x) - f(x) = \frac{1}{nhg_1(x)} \sum_{t=1}^n K\left(\frac{X_t - x}{h}\right) \Bigg\{ f(X_t) + \varepsilon_t - f(x)$$

$$- \dot{f}(x)(X_t - x) + \sum_{i=1}^{t-1} \pi_i [f(X_{t-i}) - \widetilde{f}(X_{t-i})]$$

$$- \sum_{i=t}^{\infty} \pi_i e_{t-i} + \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i)[f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}] \Bigg\}$$

$$= \frac{1}{nhg_1(x)} \sum_{t=1}^n K\left(\frac{X_t - x}{h}\right) \Big\{ f(X_t) - f(x)$$

$$- \dot{f}(x)(X_t - x) + \varepsilon_t \Big\} + \frac{1}{nhg_1(x)} \sum_{t=2}^n K\left(\frac{X_t - x}{h}\right)$$

$$\times \sum_{i=1}^{t-1} \pi_i [f(X_{t-i}) - \widetilde{f}(X_{t-i})] - \frac{1}{nhg_1(x)}$$

$$\times \sum_{t=2}^n \sum_{i=t}^{\infty} K\left(\frac{X_t - x}{h}\right) \pi_i e_{t-i} + \frac{1}{nhg_1(x)}$$

$$\times \sum_{t=2}^n \sum_{i=1}^{t-1} (\widetilde{\pi}_i - \pi_i) K\left(\frac{X_t - x}{h}\right) [f(X_{t-i}) - \widetilde{f}(X_{t-i}) + e_{t-i}]$$

$$\equiv S_1 + S_2 + S_3 + S_4.$$

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:

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

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 et al. (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) in Yoshihara (1976) hold, and $\sum_{n \geq 1} n\beta(n)^{\delta/(2+\delta)} < \infty$, then we have*

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

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

References

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

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control, 3rd ed.. Prentice-Hall, New Jersey.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, 2nd ed.. Springer-Verlag, New York.

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

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

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

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

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

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

Dickey, D.A., Fuller, W.A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49, 1057–1072.

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

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

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

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

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

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

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

Heckman, J., Ichimura, H., Smith, J., Todd, P., 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.

Lewis, P.A.W., Stevens, J.G., 1991. Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). Journal of the American Statistical Association 86, 864–877.

Liu, J.M., Chen, R., Yao, Q., 2005. Nonparametric transfer function models. Technical Report, Georgia Southern University.

Liu, L.M., Hanssens, D.M., 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., Stoker, T.M., 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 B 51, 29–46.

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

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

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

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

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

Tiao, G.C., Box, G.E.P., 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 and Economic Statistics 3, 228–237.

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

Wu, C.O., Chiang, C.T., Hoover, D.R., 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., Li, W.K., 1999. On single-index coefficient regression models. Journal of the American Statistical Association 94, 1275–1285.

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

Xiao, Z., Linton, O.B., Carroll, R.J., Mammen, E., 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., Diggle, P.J., 1994. Semiparametric models for longitudinal data with application to CD4 cell number in HIV seroconverters. Biometrics 50, 789–699.