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 we consider a class of nonparametric transfer function models with correlated noise. Transfer function models have been used to study the relationship between several time series, typically one as output and the others as inputs. To avoid the subjectivity in specifying a parametric form for the (nonlinear) transfer function, we assume the functional form of the transfer function is unknown but smooth. 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 applied on highly nonlinear relationship of unknown functional forms; by modeling the noise explicitly as an 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. The estimation procedures are introduced and the asymptotic properties of the estimators are discussed. The finite-sample properties of the estimators are studied through simulations and one real example.

[1]Jun M. Liu is Assistant Professor, 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. Rong Chen's research is partially supported by NSF grant DMS-0244541 and NIH grant R01 GM068958.

1 Introduction

Linear transfer function models (Box and and Jenkins, 1976) have been extensively used to model the relationship between one 'output' time series and several other 'input' time series. It has 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 $B^iX_t \equiv X_{t-i}$. There has been extensive research on linear transfer function models (e.g., Newbold, 1973; Tiao and Box, 1981; Tsay, 1985; Poskitt, 1989; Liu and Hanssens, 1982). Linear transfer function models have been successful in many fields, such as economics, business, engineering and biology. However, its linear nature limits its applicability, because many nonlinear features 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* (NLTF) model of the form $Y_t = C + f(X_{t-d}, \cdots, X_{t-d-p}; \boldsymbol{\theta}) + N_t$, where $f(\cdot)$ is a parametric function assuming the Volterra series representation, N_t is stationary and modeled by an ARMA model.

One problem with nonlinear parametric models is, beyond the linear domain there are infinitely many candidate nonlinear functions so it is usually difficult to justify the explicit functional forms a priori. Following the *let-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 Tjöstheim, 1990; Lewis and Stevens, 1991; Masry, 1996a,b; Fan and Gilbels, 1996). However for nonparametric methods, severe difficulty arises when dealing with high-dimensional model $Y_t = f(X_{1t}, \cdots, X_{pt}) + \varepsilon_t$. To have variance-stable estimation when p is large, the sample size has to increases exponentially. This problem was termed by Bellman (1961) as the *curse of dimensionality*. To solve this problem, Chen and Tsay (1993a,b) proposed two restrictive nonparametric models, the *functional-coefficient autoregressive* (FAR) model and the *nonlinear additive autoregressive* model. Cai, Fan and Yao (2000) applied the functional-coefficient regression model to nonlinear time series data. The FAR model is generalized to the adaptive functional-coefficient model (Ichimura, 1993; Xia and Li, 1999; and Fan, Yao and Cai, 2002). The *single index model* has also been used to model nonlinear time series data (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) extended the single index model to the multiple index model. *Partially linear models* take advantage of the known linearity of the input variables so they are usually more parsimonious, as a result, they are widely used in nonlinear time series modeling (Härdle, Liang and Gao, 2000). Smith et. al. (1998) presented a Bayesian approach to semiparametrically estimate a regression model with additive autocorrelated noise. There is a 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 we propose a class of nonparametric transfer function model (NPTF). Consider

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. We model the transfer function $f(\cdot)$ 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 $\phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i$, $\theta(B) = 1 - \sum_{j=1}^{q}\theta_j B^j$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_p)^\tau$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_q)^\tau$ are unknown parameters, $\{\varepsilon_t\}$ is a sequence of independent $(0, \sigma^2)$ random variables. We propose an iterative procedure to estimate both the transfer function and the ARMA parameters. Because its connection to the Box-Jenkins transfer function model and nonparametric smoothing, we name the proposed method *nonparametric transfer function model*. We also assume that $\{X_t\}$ and $\{\varepsilon_t\}$ are independent, this implies the independence between $\{X_t\}$ and $\{e_t\}$. The main reason for this is, when $\{e_t\}$ is autocorrelated and is correlated with $\{X_t\}$, local polynomial estimator may not be consistent. This assumption essentially forbids the use of lagged Ys as explanatory variable. When lagged Y is needed on the right-hand-side of the model, alternative approach will be 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.

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

The problem of estimating f in (1) can be viewed as a regression with correlated noise. For local nonparametric smoothers such as kernel and local polynomial smoothers, the dependence of the data within a local neighborhood is often weaker than the dependence in the original series. When certain mixing conditions are satisfied, the local data can be treated as independent asymptotically, therefore results such as the Law of Large Numbers and the Central Limit Theorem for independent data hold asymptotically. This is the *windowing-and-whitening* principle by Hart (1996). Partly because of this, some researchers use nonparametric smoother, ignoring the correlation structure in $\{X_t\}$ and $\{e_t\}$. For example, in longitudinal data analysis, the idea of "working independence" is widely adopted (e.g., Zeger and Diggle, 1994; Wild and Yee 1996; Wu, Chiang and Hoover, 1998). There are theoretical justifications for this concept in longitudinal data analysis (e.g., Ruchstuhl, Welsh and Caroll, 2000), but ignoring correlation in other applications may be problematic. 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.

Xiao, Linton, Carroll and Mammen (2003) considered a similar problem. Their approach is similar to ours in spirit but in their study $\{e_t\}$ is approximated by an AR process. By allowing

the order of the AR approximation to go to infinity, this includes the ARMA case. Indeed, Xiao et. al. (2003) established that asymptotically f can be estimated as well as if e_t is iid, which is essentially the same result as Theorems 3 and 4 below. Recently Su and Ullah (2006) modeled e_t as a finite-order nonparametric AR process, they showed similar asymptotic results when $\{e_t\}$ follows a finite order nonlinear process. In this paper we model $\{e_t\}$ explicitly as an ARMA(p,q) process, the parsimonious representation allows us to improve the efficiency of estimation in finite sample. Another advantages of having explicit and parsimonious innovation structure is in the ability of generating accurate predictions using the model.

This paper is organized as follows. In section 2 we present the estimation procedure and the asymptotic properties of the proposed estimator when e_t follows an AR(p) process. In section 3 we extend the results to the case when e_t follows an ARMA(p,q) process. Although AR(p) case is a special case of ARMA(p,q), taking advantage of the pure AR structure provides a better algorithm and simpler proof of the asymptotic results. Section 4 contains some simulation results. We apply the proposed procedures to one real example and the results are presented in Section 5. Section 6 contains summary and discussions. The technical proofs are given in Appendix A. In our proof we use one important result of Yoshihara (1976), 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 $\{(X_t, Y_t)\}_{t=1}^n$, we first ignore the correlation in $\{e_t\}$ and construct a preliminary estimator for $f(\cdot)$ by local linear regression, 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. The least squares property implies that

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

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$(\prod_{i=0}^{p} w(X_{t-i}))$, where $w(\cdot)$ is a weight function, controlling the boundary effect in nonparametric estimation. We define an iterative procedure 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. Define $\check{\boldsymbol{\phi}}$ minimizes

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

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 reduces 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, we only compute those $\widehat{f}(X_j)$ with $w(X_j) > 0$, to eliminate the boundary bias in nonparametric estimation. We may let $w(\cdot)$ be an indicator function on, for example, the 80% inner sample range of X_t.

Remark 3: There are two bandwidth b and h in the estimation procedure. The asymptotic results below shows 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). Such a requirement controls the bias in the preliminary step of the estimation. In practice, standard optimal 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.

2.2 Asymptotic Results

Let

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

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 estimator if we can actually observe $\{e_t\}$. Obviously $\widehat{\boldsymbol{\phi}}_{\text{Ideal}}$ is the least squares estimator of an AR(p) model. It has been shown (e.g., Brockwell and Davis, 1987) that

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

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

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

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

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

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

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

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

and $S_n^*(x)$ is defined in the same manner as $S_n(x)$ in (5) with $K(\frac{X_t - x}{b})$ replaced by $K(\frac{X_t - x}{h})\prod_{i=1}^{p} w(X_{t-i})$ (See also (3)). Theorem 2 below indicates that the above estimator is *asymptotically efficient* in the sense 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

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

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

This theorem shows that the nonparametric transfer function estimator $\widehat{f}(\cdot)$ is indeed more efficient than the conventional local polynomial estimator $\widetilde{f}(\cdot)$. If we ignore the serial correlation in $\{e_t\}$ and use local linear regression to estimate $f(\cdot)$ then the resulting asymptotic variance has the same form as (12), but the white noise variance σ^2 in (12) will be replaced by the variance of e_t, which is strictly greater than σ^2 if e_t follows 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 variance or mean square error. We can also see that the gain in efficiency of $\widehat{f}(\cdot)$ over $\widetilde{f}(\cdot)$ is only affected by the correlation structure of the data, the stronger the correlation, the larger the gain.

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

Here we consider the general case when $\{e_t\}$ follows an ARMA(p, q) process. We will see that in this case, the estimation shares the similar "prewhitening" idea with the AR(p) case, and later in this paper we will see that asymptotic results similar to those of the AR(p) case continue to hold.

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

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

We assume that $\{e_t\}$ is 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 $\sum_{i=0}^{\infty} |\psi_i| < \infty$ (see, e.g., Box and Jenkins, 1976). Denote $\boldsymbol{\beta} = (\phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q)^\tau$. We estimate $f(\cdot)$ and $\boldsymbol{\beta}$ 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 convertible conditions. To start the iteration, we ignore the serial correlation in $\{e_t\}$ and obtain the initial estimate $\widetilde{f}(\cdot)$ by local linear regression as defined in (2). The iterative procedure is described as follows

1. Obtain an initial estimate $\widetilde{\boldsymbol{\beta}}$ 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), (14) and (15). In this paper we use one nonlinear regression method based on the Gauss-Newton algorithm. 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 properties

Similar to the AR(p) case, we define the "idealized" estimator of $\boldsymbol{\beta}$ as the solution of the following optimization problem, assuming $\{e_t\}$ is observable,

$$\widehat{\boldsymbol{\beta}}_{\text{Ideal}} = \inf_{\boldsymbol{\beta}} \Big\{ \frac{\phi(B)}{\theta(B)}\, e_t \Big\}^2.$$

$\widehat{\boldsymbol{\beta}}_{\text{Ideal}}$ is an usual estimator of an ARMA model and 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 $\phi(B)U_t = \varepsilon_t$ and $\{V_t\}$ is an AR(q) process defined by $\theta(B)V_t = \varepsilon_t$. 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, we have the following asymptotic results for the ARMA(p, q) case.

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 $n \to \infty$, we have*

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

where

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

and $g_1(x)$ is the stationary 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.

4 Numerical Properties

To study the finite-sample properties of the proposed estimator, we conduct simulations using model (1), where

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

and X_t is generated from an AR(1) model $X_t = 0.3X_{t-1} + a_t,\ a_t \sim$ iid $\mathrm{N}(0, 0.3^2)$. $\{e_t\}$ is generated from an ARMA(1,1) model $e_t = \phi e_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}$ with different parameters, $\varepsilon_t \sim N(0, 0.5^2)$. In the simulations three sample sizes (100, 200 and 400) are considered, 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, we find that the results are not very sensitive to the bandwidths, so to save the space we only report the case of $h = 1.06 s_X n^{-1/5}$ and $b = 1.06 s_X n^{-1/4}$. Xiao et al. (2003) (subsequently XLCM) proposed to estimate $f(\cdot)$ using local polynomial regression and approximate e_t by an AR process, whose order τ is allowed to go to infinity. Su and Ullah (2006) (subsequently SU) proposed to model e_t as a nonparametric AR model of finite order d, $e_t = g(e_{t-1}, \cdots, e_{t-d}) + \varepsilon_t$, both $g(\cdot)$ and $f(\cdot)$ are estimated via local polynomial regression. For comparison, under the same setting specified above we run simulation using the proposed estimator (subsequently NPTF), XLCM, SU and the "conventional" local linear estimator, in which $\{e_t\}$ is assumed to be i.i.d. We then average the mean squared errors (MSE=$\frac{1}{n}\sum_{t=1}^{n}\{\widehat{f}(X_t) - f(X_t)\}^2$) of all four estimators over the replications. The average mean squared errors of the efficient estimators NPTF, XLCM, and SU are divided by that of the conventional estimator. The resulting ratios are the relative average mean squared errors and are reported under the estimator names in Tables 1 and 2. In the simulation we set $\tau = 2$ and $d = 2$. We also report the means and standard deviations of $\widehat{\phi}$ and $\widehat{\theta}$ from NPTF, as well as the average mean squared error of the conventional estimator (AMSE), which is the common denominator of the relative MSEs. A histogram of $\widehat{\phi}$ and a typical simulation is given in Figure 1.

We summarize the observations from the simulation studies below:

1. The NPTF estimator $\widehat{f}(\cdot)$ is more efficient than the conventional local linear regression estimator in that the relative MSE is less than one in most of the cases.

2. The stronger the autocorrelation, the larger the gain in efficiency of $\widehat{f}(\cdot)$. This can be seen most directly from the AR(1) and the MA(1) cases, where generally, the relative MSE decreases when $|\phi|$ or $|\theta|$ increases.

3. The sampling distributions of $\hat{\phi}$ and $\hat{\theta}$ appear to be normal, with the means close to ϕ or θ, as illustrated in the histogram in Figure 1.

4. The performance of the estimation improves with the increase of sample size. The mean squared errors of $\widehat{f}(\cdot)$ decreases when the sample size increases; the bias and the standard

Table 1: Simulation results I: 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) 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
		100	.166, .301	-.213, .306	.025	.985	1.05	1.17
.2	-.2	200	.185, .206	-.209, .202	.016	.987	1.01	1.15
		400	.193, .145	-.203, .151	.009	.982	1.02	1.14
		100	.469, .168	-.180, .181	.040	.861	.882	.955
.5	-.2	200	.488, .131	-.191, .136	.021	.853	.894	.983
		400	.496, .085	-.192, .091	.014	.829	.865	.948
		100	.807, .099	-.102, .157	.133	.795	.818	.858
.8	-.2	200	.803, .065	-.123, .098	.075	.778	.787	.819
		400	.806, .041	-.149, .063	.039	.715	.736	.763

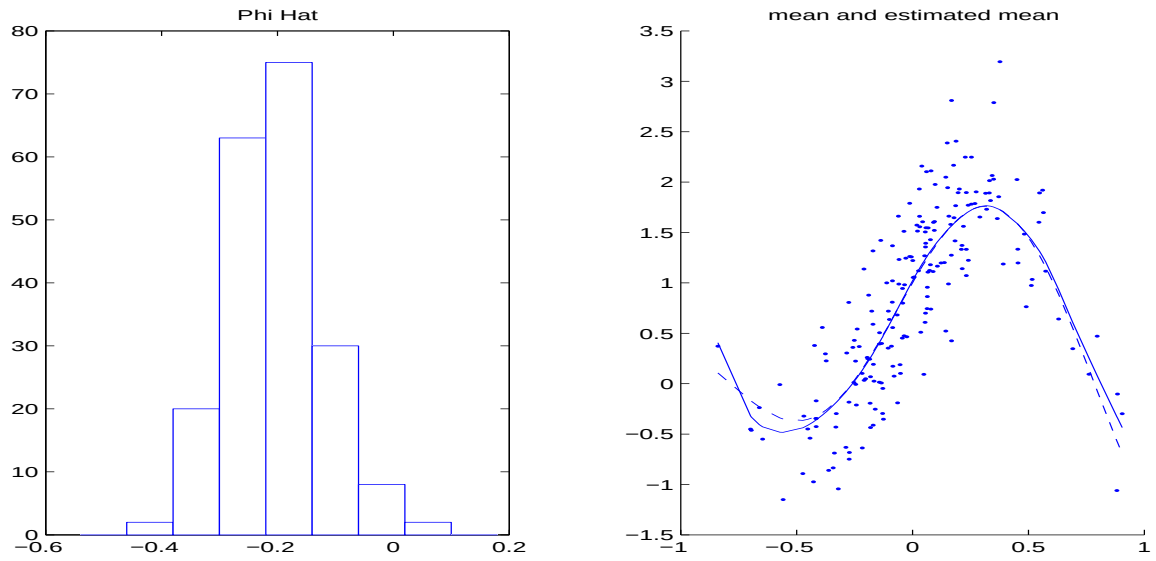

Figure 1: $\phi = -0.2$, n=200. Left panel: histogram of $\hat{\phi}$, right panel: true (solid line) and estimated (dashed line) transfer function in a typical simulation.

deviation of ϕ and θ also decrease when the sample size increases.

5. In most of the cases NPTF, XCLM and SU are more efficient than the conventional estimator. When $\{e_t\}$ follows a an ARMA model with small $|\theta|$ (including pure AR models), NPTF and XCLM have similar efficiency. When $|\theta|$ is large, the NPTF estimator is more efficient. During the simulation we also used larger τ, but the performance of XCLM does not always improve, because of the additional error introduced in estimating more parameters. Since the observations are very similar, we omitted the detailed results. In the above simulation SU is not as efficient as NPTF and XCLM, mainly because here e_t is generated from ARMA models of finite order. In a separate study we found that when e_t is generated from nonlinear finite order AR process SU is more efficient than NPTF and XCLM.

6. When there are nontrivial MA components in $\{e_t\}$, the MA estimates can have large bias. Larger sample sizes are needed to achieve better performance.

These observations confirm our asymptotic theory.

5 Example: River Flow and Rainfall

In this example we study the effect of daily rain fall on river flow of Kanna river in 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, please refer to Chen and Tsay (1996).

5.1 Competing models

Chen and Tsay (1996) considered nonlinear transfer function model of the form

$$Y_t = C + f(X_{t-d}, X_{t-d-1}, \ldots) + N_t,$$

where f is a measurable function in $\mathcal{R}$, $\{N_t\}$ is a stationary time series and independent of $\{X_t\}$, C is a constant, $d \geq 0$ denote the delay. $f(\cdot)$ is approximated by a *Volterra* series

$$f(X_t, X_{t-1}, \ldots) = \sum_{i=0}^{\infty} f_i X_{t-i} + \sum_{i,j=0}^{\infty} f_{ij} X_{t-i} X_{t-j} + \sum_{i,j,k=0}^{\infty} f_{ijk} X_{t-i} X_{t-j} X_{t-k} + \ldots.$$

N_t is modeled as an stationary ARMA process. In their study Chen and Tsay used the following nonlinear transfer function model

$$\begin{aligned} Y_t &= 1.38 + \frac{.41}{1-.87B} X_t - .39X_t^2 + (.015B + .007B^2) X_t X_{t-1} + (.0008 + .0001B) X_t^3 \\ &\quad -(.0005 + .0003B) X_t X_{t-1} X_{t-2} + \frac{1}{1 - .70B - .14B^6}\, \varepsilon_t\ , \end{aligned} \tag{17}$$

where Y_t is river flow and X_t is the daily rain fall. The estimated residual variance is 6.23. For comparison purpose, they also built the following linear transfer function model

$$Y_t = .79 + \frac{.18 + .65B - .30B^2 - .10B^3 - .07B^4}{1 - .90B} X_t + \frac{1}{1 - .63B + .12B^2 - .22B^6}\, \varepsilon_t, \tag{18}$$

and found that the estimated residual variance is 20.81. Since these two models use about the same number of parameters, the nonlinear transfer function model outperforms the classical transfer function model substantially. This is mainly due to the highly nonlinear nature of the data, as indicated by the nonlinearity test proposed by Chen and Tsay (1996).

5.2 Nonparametric transfer function model building

Here we use the proposed nonparametric transfer function model to analyze this dataset and compare the performances with the NLTF model. 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

$$Z_t = f(X_t, X_{t-1}, X_{t-2}) + e_t. \tag{19}$$

Note here we use a low dimensional smoothing model instead of an univariate smoothing model. We first estimate $f(\cdot)$ assuming $\{e_t\}$ iid, then remove this preliminary estimate $\widetilde{f}(\cdot)$ from Z_t and

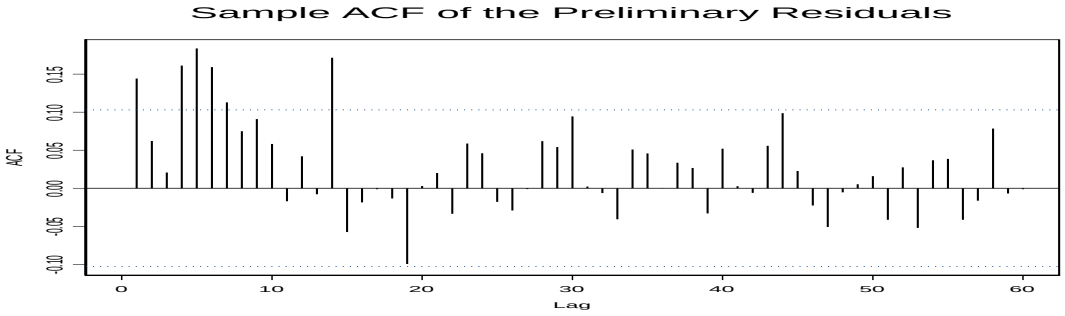

Figure 2: Sample ACF plot of the partial residuals after removing $\widetilde{f}(\cdot)$

identify a model for $\{e_t\}$ based on the sample autocorrelation function (ACF) of the partial residuals (Figure 2). The resulting model is an AR model with lagged terms 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 $\widehat{f} = \mathbf{S}_h \mathbf{Y}$, where $\mathbf{Y}$ is the vector of observations. In order to compare with the parametric models, we also calculate the *equivalent number of parameters* defined as $\mathrm{tr}(\mathbf{S}_h)$. The resulting bandwidth is 5 and the equivalent number of parameters is 33.46. The resulting residual variance is 4.58, which is smaller than that of the nonlinear transfer function model. 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, it indicates that the residual is roughly a white noise process.

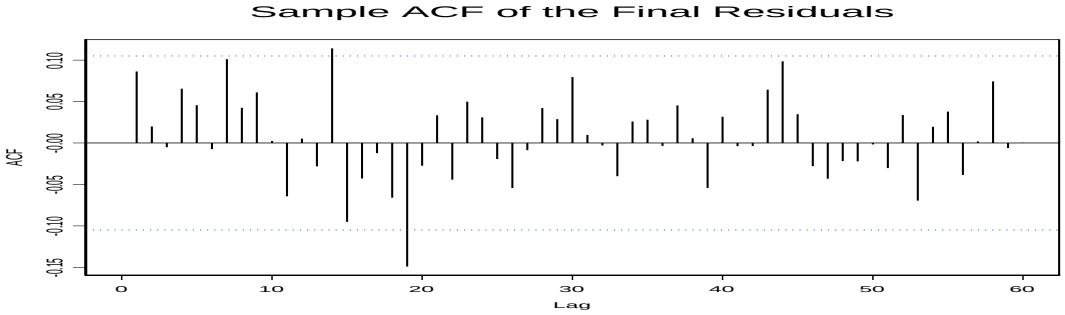

Figure 3: Sample ACF plot of the final residuals

The fitted values are plotted in Figure 4. From this plot, we see that the identified NPTF model fits the data very well.

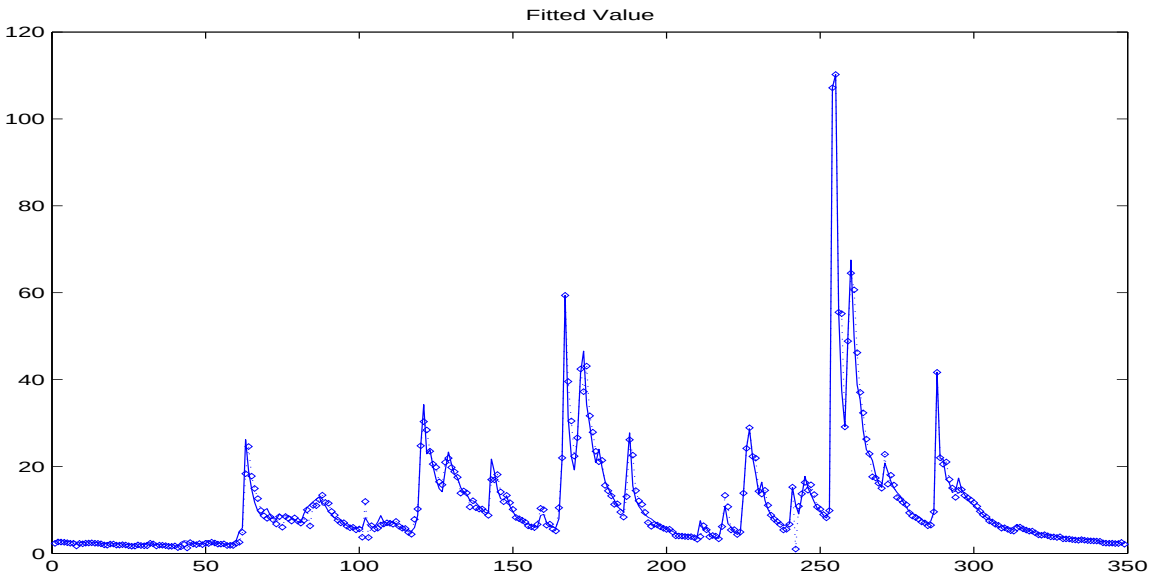

Figure 4: The fitted values of the nonparametric transfer function model
Solid line: the actual observations, Dashed line: the fitted values

5.3 Forecasting performance

We have seen an improvement in the residual variance by using the nonparametric transfer function model. But we also notice that the equivalent number of parameters in this model is far larger than the number of parameters in the NLTF model (12 in total). It is possible that this improvement is the result of overfitting. To see if this is the case, we need to consider the forecasting performance. To this end, we compare the one-step ahead forecasting performance of the NLTF model and the NPTF model. Specifically, we use data available at time t to build the model, then X_{t+1} and its lags are plugged in the estimated model to predict Z_{t+1}. Forecasts are made at $t = 180, 181, \cdots, 365$. For each t, we calculate the forecasting error $Y_{t+1} - \widehat{Y}_t(1)$ for both models, where $\widehat{Y}_t(1)$ is the one-step ahead forecast of Y_{t+1} at time t. Finally the forecasting *root mean squared errors* (RMSEs) of the NPTF model and the NLTF model are calculated and compared. The forecasting RMSE is calculated as

$$\text{RMSE} \equiv \frac{1}{185} \sum_{t=180}^{365} \left\{ Y_{t+1} - \widehat{Y}_t(1) \right\}^2,$$

The post-sample RMSEs for the NPTF model and the NLTF model are 8.80 and 12.56, respectively. The one-step ahead forecast errors are plotted against the forecasting origins in Figure 5.

From Figure 5 we can see that the NPTF model outperforms the NLTF model most of the time. On average, the NPTF model performs better than the NLTF model in that it produce 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.

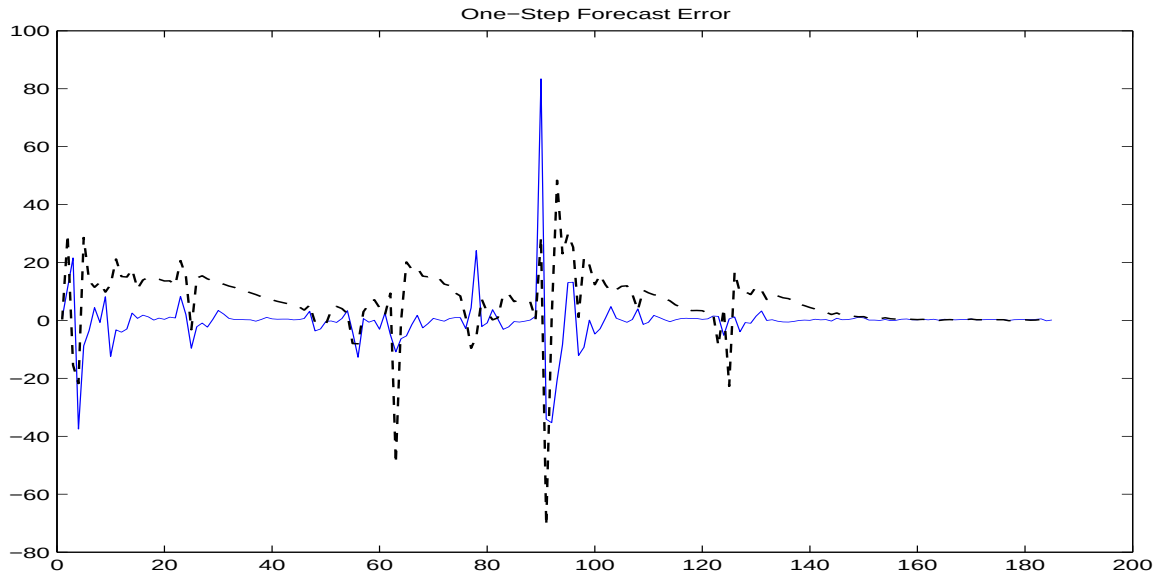

Figure 5: One-step ahead forecast errors
Solid line: the NPTF model, Dashed line: the NLTF model

6 Summaries and Discussions

In this paper we propose a new method to model nonlinear relationship 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. Our 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 our study we have only considered the case where 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, we must consider more restrictive models, such as the additive model. Research addressing this topic is ongoing.

Appendix A – Technical Proofs

In the proofs that follow, we use $C > 0$ to denote 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 $\{X_{t1}, \cdots, X_{ti}\}$. We need the following assumptions, 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 $i \le j$.
Also $\sum_{k \ge 1} k\beta(k)^{\delta/(2+\delta)} < \infty$ for some $\delta \in (0, 8)$.

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

(A3) $f(\cdot)$ has continuous second derivative $\ddot{f}(\cdot)$, $g_1(\cdot)$ 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 $r, s = 1, 2, \cdots, p$.

We only establish (i). The proof for (ii) is similar and simpler, therefore is omitted. By Lemma 1, we may write

$$B_{r1} = \{B_{r11} + B_{r12} + B_{r13} + O_p(nb^4)\}\{1 + o_p(1)\}, \tag{20}$$

where

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

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

where $\boldsymbol{\xi}_i = (X_i, X_{i-1}, \cdots, X_{i-p}, e_i)^\tau$. We split B_{r11} 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

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

$\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\xi}_k)$ 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:

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

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

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

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

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

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 (20). 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 we have $\widetilde{Y}_t = \varepsilon_t + f(X_t) + \sum_{i=1}^{p} \phi_i \{ \widetilde{f}(X_{t-i}) - f(X_{t-i}) \}$. 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

$$\begin{aligned}
\widehat{f}(x) - f(x) &= \frac{1}{nhp_1(x)} \sum_{t=1}^{n} K\Big(\frac{X_t - x}{h} \Big) \prod_{l=1}^{p} w(X_{t-l}) \Big\{ \varepsilon_t + f(X_t) \\
&\quad + \sum_{k=1}^{p} \phi_k [\widetilde{f}(X_{t-k}) - f(X_{t-k})] - f(x) - \dot{f}(x)(X_t - x) \Big\} \\
&= \frac{1}{nhp_1(x)} \sum_{t=1}^{n} K\Big(\frac{X_t - x}{h} \Big) \prod_{l=1}^{p} w(X_{t-l}) \Big\{ \varepsilon_t + f(X_t) - f(x) - \dot{f}(x)(X_t - x) \Big\} \\
&\quad + \frac{b^2 \mu_2}{2nhp_1(x)} \sum_{k=1}^{p} \phi_k \sum_{t=1}^{n} K\Big(\frac{X_t - x}{h} \Big) \prod_{l=1}^{p} w(X_{t-l}) \ddot{f}(X_{t-k}) \\
&\quad + \frac{1}{n^2 hbp_1(x)} \sum_{k=1}^{p} \phi_k \sum_{i,j=1}^{n} K\Big(\frac{X_i - x}{h} \Big) \prod_{l=1}^{p} w(X_{t-l}) K\Big(\frac{X_j - X_{i-k}}{b} \Big) \frac{e_j}{g_1(X_{i-k})}. \quad (21)
\end{aligned}$$

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 (21) as J.

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

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

Then it holds that

$$J = \frac{\phi_1}{n^2 hbp_1(x)} \sum_{1\le i<j\le n} \{\kappa(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) - \kappa_1(\boldsymbol{\xi}_i) - \kappa_1(\boldsymbol{\xi}_j)\} + \frac{\phi_1(n-1)}{n^2 p_1(x)} \sum_{i=1}^{n} \kappa_1(\boldsymbol{\xi}_i)/(hb), \tag{22}$$

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 (22) 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,

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

Hence $J_1 = o_p\{(nh)^{-1/2}\}$. Note $h^2 = O\{(nh)^{-1/2}\}$ under Assumption A4. Now it follows from (21) 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 we prove the theorem by adopting the Gauss-Newton nonlinear regression method. 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

$$\phi_0(B)\theta_0(B)^{-1} e_t = \sum_{i=0}^{t-1} \pi_i^0 e_{t-i}, \quad \theta_0(B)^{-1} e_t = \sum_{i=0}^{t-1} \xi_i^0 e_{t-i}, \quad \phi_0(B)\theta_0(B)^{-2} = \sum_{i=0}^{t-1} \eta_i^0 e_{t-i}. \tag{23}$$

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 (23), 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 $\sum_{t=1}^{n}\{\varepsilon_t(\boldsymbol{\beta}_0) + D_t^\tau \Delta\boldsymbol{\beta}\}^2$. Let $m = \max(p, q) + 1$, 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 (23), 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

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

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

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

R. Bellman 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

P. Newbold 1973. Bayesain estimation of box-jenkins transfer function-noise models. *Journal of the Royal Statistical Society* **35**:323-336.

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

T. Ozaki 1985. Statistical identification of storage models with application to stochastic hydrology. *Water Resources Bulletin* **21**:663675.

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

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

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

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

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

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

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

D. Tjøstheim 1994. Nonlinear time series: A selective review. *Scandanavian Journal of Statistics* **21**:97-130.

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

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

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

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

Y. Xia, 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.

Z. Xiao, 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.

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

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