

# More Efficient Local Polynomial Estimation in Nonparametric Regression With Autocorrelated Errors

Zhijie XIAO, Oliver B. LINTON, Raymond J. CARROLL, and Enno MAMMEN

---

We propose a modification of local polynomial time series regression estimators that improves efficiency when the innovation process is autocorrelated. The procedure is based on a pre-whitening transformation of the dependent variable that must be estimated from the data. We establish the asymptotic distribution of our estimator under weak dependence conditions. We show that the proposed estimation procedure is more efficient than the conventional local polynomial method. We also provide simulation evidence to suggest that gains can be achieved in moderate-sized samples.

KEY WORDS: Kernel regression; Linear process; Prewhitening; Time series.

---

## 1. INTRODUCTION

Consider the regression model

$$Y_t = m(X_t) + u_t, \quad t = 1, \dots, T, \quad (1)$$

where the stationary residual process  $u_t$  is autocorrelated but satisfies  $E(u_t | X_1, \dots, X_T) = 0$  almost surely. The function  $m(\cdot)$  is assumed to be unknown but smooth and is the object of central interest. There are two leading sampling schemes with regard to the process  $\{X_t\}$ : the “fixed design” case where  $X_t$  is time or some smooth function thereof [i.e.,  $X_t = f(t/T)$  for some smooth  $f$ ], and the “random design” case where  $X_t$  is a stationary stochastic process itself with a nondegenerate marginal distribution. (Opsomer, Wang, and Yang 2001 have discussed the related case where the regressors are multivariate and random but the error covariance is a smooth function of the regressors. This case is more like the fixed design in some respects.) In the fixed design case, both the standard least squares parametric and kernel nonparametric estimator have variances proportional to the long-run variance (i.e., the spectral density at frequency 0) of the process  $\{u_t\}$ . However, adjusting for serial correlation brings no advantage in terms of estimator variance in either the parametric or nonparametric method. Specifically, when the regressors are polynomials in time, ordinary least squares (OLS) = generalized least squares (GLS) (see, e.g., Andersen 1971, p. 581). Much methodologic work in nonparametric statistics has focused on this sampling scheme, especially with regard to bandwidth selection (see Hart 1991 for references).

The focus of this article is the second sampling scheme, where  $X_t$  is a nondegenerate stochastic process. This setting arises in many applications, because time itself is often not the only relevant covariate. Indeed, in the 1970s, the linear regression model with autocorrelated disturbances was one of the

central models of interest, and numerous procedures were created to deal with the estimation and testing issues that ensued, including Cochrane–Orcutt, Hildreth–Lu, Prais–Winsten, and Durbin–Watson. As is well known, when the regression function is parametric, the variance of the parameter estimators is proportional to the long-run variance of the process  $\{X_t u_t\}$ , and least squares standard errors that ignore this fact are inconsistent and need to be modified in a nontrivial way. Also, one can generally improve the efficiency of least squares estimators by using a GLS weighting scheme that reflects the error autocorrelation function. Compare this with the case where  $m(\cdot)$  is nonparametric, which has been analyzed by, for example, Robinson (1983) and Masry (1996a,b). In this case, standard kernel regression smoothers do not take into account the correlation structure in  $X_t$  or  $u_t$  and estimate the regression function in the same way as if these processes were independent. Furthermore, the variance of such estimators is proportional to the short-run variance of  $u_t$ ,  $\sigma_u^2 = \text{var}(u_t)$  and does not depend on the regressor or on error covariance functions  $\gamma_X(j) = \text{cov}(X_t, X_{t-j})$ ,  $\gamma_u(j) = \text{cov}(u_t, u_{t-j})$ ,  $j \neq 0$ . Practitioners accustomed to correcting standard errors for dependence believe that the standard errors in nonparametric regression are therefore suspect. As Conley, Hansen, Luttmer, and Scheinkman (1997, p. 548) stated, “although theoretically correct, the practice of ignoring serial correlation is not likely to work well for the temporal dependence present in our short-term interest rate data.” The purpose of this article is to show that the autocorrelation function of the error process has useful information for improving estimators of the regression function. As a byproduct, one might hope to obtain more accurate standard errors, given that the resulting error process is purged of all correlation.

There is a related literature on estimating nonparametric regression with longitudinal or panel data, including works by Severini and Staniswalis (1994), Zeger and Diggle (1994), Wild and Yee (1996), and Wu, Chiang, and Hoover (1998), among others. The first authors estimated the covariance matrix of the correlated observations and used this in their kernel construction of the nonparametric regression estimate. The other authors effectively ignored the correlation structure entirely and “pretended” that the data were really independent, the so-called “working independence” method. Ruckstuhl, Welsh, and Carroll (2000) and Lin and Carroll (2000) provided theoretical evidence in support of the working independence method. In

---

Zhijie Xiao is Associate Professor, Department of Economics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (E-mail: [zxiao@uiuc.edu](mailto:zxiao@uiuc.edu)). Oliver Linton is Professor of Econometrics, Department of Economics, London School of Economics, London WC2A 2AE, U.K. (E-mail: [linton@lse.ac.uk](mailto:linton@lse.ac.uk)). Ray Carroll is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: [carroll@stat.tamu.edu](mailto:carroll@stat.tamu.edu)). Enno Mammen is Professor, Institute für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, 69120 Heidelberg, Germany. The authors thank the editor, an associate editor, two referees, M. Francisco-Fernandez, Jean Opsomer, and Michael Schimek for very helpful comments and suggestions. The authors also thank the Cowles Foundation, the National Cancer Institute, the National Institute of Environmental Health Sciences, the National Science Foundation, the Economic and Social Science Research Council of Great Britain, and the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 for financial support.

fact, they showed that for many situations and different methods of kernel estimation, the working independence method is most efficient in terms of mean squared error (MSE). That is, for the kernel methods proposed in the literature, it is generally better to ignore the correlation structure entirely. Carroll, Lin, Linton, and Mammen (2004) constructed a kernel-type method that can take advantage of the correlations among the data. The method is a simple modification, and a generalization to an arbitrary covariance matrix, of a method proposed by Ruckstuhl et al. (2000). The resulting estimator is asymptotically more efficient than the working independence estimator.

In this article we propose a new kernel-based procedure for estimating  $m(x)$  in the time series regression model (1) that takes into account the correlation structure of the error terms and is asymptotically more efficient than the usual methods. The basic idea of the proposed estimation is to transform or “prewhiten” the original regression model, so that the filtered regression has a residual term that is uncorrelated. However, because of the nonlinear feature of the regression function  $m(\cdot)$ , the transformation depends on both the function  $m(\cdot)$  and on the parameters of the autoregressive representation of  $u$ . Thus we first estimate these quantities, and then construct a feasible transformation of the dependent variable  $Y_t$ . We shown the resulting estimator to be asymptotically normal and more efficient than the conventional kernel estimator. We allow for an error correlation structure of unknown form; that is, the autoregressive representation of the process need not be of finite order.

The rest of the article is organized as follows. The proposed estimation method is introduced in Section 2. Regularity assumptions and the limiting distribution of the estimator are given in Section 3. An even more efficient estimator is discussed in Section 4. Some numerical results on simulated data are reported in Section 5. Conclusions are given in Section 6. All proofs are given in the Appendix. A technical report (Xiao, Linton, Carroll, and Mammen 2003) contains more detailed proofs and more numerical evidence.

## 2. ESTIMATION METHOD

### 2.1 Motivation and an Infeasible Estimator

Suppose that we have a sample  $\{(X_1, Y_1), \dots, (X_T, Y_T)\}$ , where  $X_t \in \mathbb{R}^d$  and  $Y_t \in \mathbb{R}$ , from the nonparametric regression model (1). We assume that the residual process  $u_t$  is stationary, has mean 0, and has an invertible linear process representation

$$u_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \tag{2}$$

where  $\varepsilon_t$  are independent identically distributed with mean 0 and variance  $\sigma_\varepsilon^2$ . Without loss of generality,  $c_0 = 1$ . It is convenient for this discussion to assume that the process  $\{u_t\}$  is independent of the process  $\{X_t\}$ , but the main results are true under some forms of mutual dependence. The coefficients  $\{c_j\}_{j=0}^{\infty}$  and the regression function  $m(\cdot)$  are unknown except that  $m(\cdot)$  is a smooth function and the coefficients  $c_j$  satisfy certain summability conditions (e.g., the process is short memory), as specified later in our assumptions. Our assumptions permit  $u_t$  to be any finite-order ARMA( $p, q$ ) process, but we allow for the full class of linear processes as is common in much of

the literature on estimating linear regression with correlated errors. The objective is to estimate  $m(x)$  at some interior point  $x$  and to provide confidence intervals for such estimates.

Let  $c(L) = \sum_{j=0}^{\infty} c_j L^j$ , where  $L$  is the usual lag operator. Inverting  $c(L)$ , we obtain an autoregressive representation of  $u_t$  of potentially infinite order. Let

$$c(L)^{-1} = a(L) = a_0 - a_1 L - \dots - a_j L^j - \dots = a_0 - \sum_{j=1}^{\infty} a_j L^j \tag{3}$$

be the inverse, and define that  $a_0 = 1$  without loss of generality, so we have  $a(L)u_t = \varepsilon_t$ . Applying  $a(L)$  to regression (1), we obtain

$$a(L)Y_t = a(L)m(X_t) + \varepsilon_t. \tag{4}$$

The error term in this transformed model is now uncorrelated; however, the immediate usefulness of this is unclear because  $m$  is nonlinear and so does not commute with the operator  $a(L)$ , as would be the case in a linear model.

We rewrite (4) as

$$\underline{Y}_t = m(X_t) + \varepsilon_t, \tag{5}$$

where  $\underline{Y}_t$  is the filtered series

$$\underline{Y}_t = Y_t - \sum_{j=1}^{\infty} a_j (Y_{t-j} - m(X_{t-j})).$$

The transformed model (5) is also a valid regression equation because  $\varepsilon_t$  is independent of  $X_t$ . If  $\underline{Y}_t$  were known, as shown by Theorem 1, then a nonparametric kernel regression of  $\underline{Y}_t$  on  $X_t$  would be more efficient than the conventional kernel estimation.

The approach proposed in this article may be applied to a wide range of nonparametric estimators, including the local polynomial estimator and the Nadaraya–Watson procedure. In this article we give asymptotic analysis based on the local polynomial procedures. (See Fan 1992 and Fan and Gijbels 1996 for discussion on the attractive properties of local polynomials.) For any dataset  $\{Z_t, X_t\}_{t=1}^n$ , the local polynomial regression of  $Z_t$  on  $X_t$  (of order  $p$ , where  $p$  is an integer) can be obtained from the multivariate weighted least squares criterion

$$\sum_{t=1}^n \left[ Z_t - \sum_{0 \leq |k| \leq p} b_k \cdot (X_t - x)^k \right]^2 K \left( \frac{X_t - x}{h} \right), \tag{6}$$

where  $K(u)$  is a nonnegative weight function on  $\mathbb{R}^d$  and  $h$  is a bandwidth parameter. Let  $\widehat{m}(x) = \widehat{b}_0$ , where  $\widehat{b}_0$  is the minimizing intercept in (6) with  $Z_t = Y_t$ , and let  $\overline{m}(x)$  be the corresponding estimator when  $Z_t = \underline{Y}_t$ . For simplicity, we assume that  $K((x - X_t)/h) = \prod_{j=1}^d k((x_j - X_{jt})/h)$ , with  $k$  the univariate kernel function.

Following the notation of Masry (1996a,b), let  $N_\ell = (\ell + d - 1)!/\ell!(d - 1)!$  be the number of distinct  $d$ -tuples  $j$  with  $|j| = \ell$ . Arrange these  $N_\ell$   $d$ -tuples as a sequence in a lexicographic order and let  $\phi_\ell^{-1}$  denote this one-to-one map. For each  $j$  with  $0 \leq |j| \leq 2p$ , let  $\mu_j(K) = \int_{\mathfrak{R}^d} u^j K(u) du$ ,  $\nu_j(K) =$

$\int_{\mathbb{R}^d} u^j K^2(u) du$ , and define the  $N \times N$  dimensional matrices  $\mathbf{M}$  and  $\mathbf{\Gamma}$  and  $N \times 1$  vector  $\mathbf{B}$ , where  $N = \sum_{\ell=0}^p N_\ell \times 1$ , by

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{0,0} & \mathbf{M}_{0,1} & \cdots & \mathbf{M}_{0,p} \\ \mathbf{M}_{1,0} & \mathbf{M}_{1,1} & \cdots & \mathbf{M}_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{M}_{p,0} & \mathbf{M}_{p,1} & \cdots & \mathbf{M}_{p,p} \end{bmatrix},$$

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{0,0} & \mathbf{\Gamma}_{0,1} & \cdots & \mathbf{\Gamma}_{0,p} \\ \mathbf{\Gamma}_{1,0} & \mathbf{\Gamma}_{1,1} & \cdots & \mathbf{\Gamma}_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Gamma}_{p,0} & \mathbf{\Gamma}_{p,1} & \cdots & \mathbf{\Gamma}_{p,p} \end{bmatrix}, \tag{7}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{M}_{0,p+1} \\ \mathbf{M}_{1,p+1} \\ \vdots \\ \mathbf{M}_{p,p+1} \end{bmatrix},$$

where  $\mathbf{M}_{i,j}$  and  $\mathbf{\Gamma}_{i,j}$  are  $N_i \times N_j$  dimensional matrices whose  $(\ell, m)$  elements are  $\mu_{\phi_i(\ell)+\phi_j(m)}$  and  $\nu_{\phi_i(\ell)+\phi_j(m)}$ . Note that the elements of the matrices  $\mathbf{M} = \mathbf{M}(K)$  and  $\mathbf{\Gamma} = \mathbf{\Gamma}(K)$  are simply multivariate moments of the kernel  $K$  and  $K^2$ . In addition, we arrange the  $N_r$  elements of the derivatives

$$\frac{1}{r_1! \cdots r_d!} \frac{\partial^r m(x)}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}} \quad \text{for } r_1 + \cdots + r_d = r$$

as a column vector  $\mathbf{m}^{(r)}(x)$ . Theorem 1 gives the asymptotic distribution of  $\overline{m}(x)$  and shows that it is asymptotically more efficient than  $\widehat{m}(x)$ .

*Theorem 1.* Suppose that the assumptions given in Section 3 hold. Then

$$\sqrt{Th^d}(\overline{m}(x) - m(x) - h^q[\mathbf{M}^{-1}\mathbf{B}\mathbf{m}^{(q)}(x)]_{0,0}) \Rightarrow N\left(0, \frac{\sigma_\varepsilon^2}{f_X(x)}[\mathbf{M}^{-1}\mathbf{\Gamma}\mathbf{M}^{-1}]_{0,0}\right),$$

where  $q = p + 1$  and  $[\mathbf{A}]_{0,0}$  signifies the upper-left element of matrix  $\mathbf{A}$ .

Theorem 1 shows that the bias term  $[\mathbf{M}^{-1}\mathbf{B}\mathbf{m}^{(q)}(x)]_{0,0}$  of the estimator  $\overline{m}(x)$  is the same as that of the conventional  $p$ th order local polynomial estimator  $\widehat{m}(x)$ . In the case with a local linear estimator,  $p = 1$ , and the bias term is simply  $\mu_2(K)m''(x)/2$ . The smoother  $\overline{m}(x)$  has a variance proportional to  $\sigma_\varepsilon^2$  and hence is more efficient than the traditional kernel estimator  $\widehat{m}(x)$ , which has variance proportional to

$$\sigma_u^2 = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} c_j^2 \geq \sigma_\varepsilon^2.$$

Therefore, the relative efficiency (of the proposed estimator relative to the standard estimator) in purely variance terms is  $\sigma_\varepsilon^2/\sigma_u^2 \leq 1$ . For example, when  $u_t = au_{t-1} + \varepsilon_t$ , we have  $\sigma_u^2 = \sigma_\varepsilon^2/(1 - a^2)$ , which strictly exceeds  $\sigma_\varepsilon^2$  except when  $a = 0$ . We now turn to a comparison of the asymptotic MSEs (integrated or pointwise); we make the comparison at the respectively optimal bandwidths. For  $\overline{m}$ , this bandwidth is the same as for  $\widehat{m}$ , but with  $\sigma_\varepsilon^2$  in place of  $\sigma_u^2$ , and hence is larger. The relative efficiency is then  $\{\sigma_\varepsilon^2/\sigma_u^2\}^{2q/(2q+d)}$ , assuming that the bias term

is nonzero. The percentage efficiency gain is less if measured in MSE than in variance, but the two relative efficiencies are monotonically related. In the sequel we will focus on variance comparisons for simplicity.

### 2.2 The Estimator

In practice,  $\underline{Y}_t$  is unknown, and thus the regression (5) and  $\overline{m}(x)$  are infeasible. Here we propose a feasible estimator of regression (5) by replacing the left side of this equation by an approximation of  $\underline{Y}_t$  based on estimates of the coefficients  $a_j$  and a truncation of the infinite sum to a finite but large-order sum. The proposed estimation procedure is as follows:

1. Obtain a preliminary consistent estimate of  $m$  by local  $p$ th order polynomial smoothing  $Y_t$  on  $X_t$  with corresponding kernel  $K_0$  and bandwidth  $h_0$ . Denote the preliminary estimates as  $\widehat{m}(X_t)$  and calculate the estimated residuals

$$\widehat{u}_t = Y_t - \widehat{m}(X_t).$$

2. Let  $\tau = \tau(T)$  be some truncation parameter suitably small relative to the sample size  $T$  but large enough to avoid serious bias (see Assumption 6 in Sec. 3). Conduct a  $\tau$ th order autoregression of  $\widehat{u}_t$ ,

$$\widehat{u}_t = \widehat{a}_1\widehat{u}_{t-1} + \cdots + \widehat{a}_\tau\widehat{u}_{t-\tau} + \text{residual}. \tag{9}$$

Define the estimate  $\widehat{\mathbf{A}}_\tau = (\widehat{a}_1, \dots, \widehat{a}_\tau)'$  of  $\mathbf{A}_\tau = (a_1, \dots, a_\tau)'$ , where

$$\widehat{\mathbf{A}}_\tau = (\widehat{\mathbf{U}}_\tau' \widehat{\mathbf{U}}_\tau)^{-1} \widehat{\mathbf{U}}_\tau' \widehat{\mathbf{u}},$$

where  $\widehat{\mathbf{u}} = (\widehat{u}_\tau, \dots, \widehat{u}_T)'$  and  $\widehat{\mathbf{U}}_\tau$  is the  $(T - \tau) \times \tau$  matrix of regressors with typical element  $\widehat{u}_{t-j}$ .

3. Construct an approximation of  $\underline{Y}_t$  by

$$\widehat{\underline{Y}}_t = Y_t - \sum_{j=1}^{\tau} \widehat{a}_j (Y_{t-j} - \widehat{m}(X_{t-j})).$$

The proposed estimator of  $m(x)$  is then obtained from the local  $p$ th order polynomial smoothing of  $\widehat{\underline{Y}}_t$  on  $X_t$  with corresponding kernel  $K_1$  and bandwidth  $h_1$ , calling the resulting estimator  $\widetilde{m}(x)$ .

The foregoing procedure may be iterated to achieve better finite-sample performance in practice. Also, when estimating the coefficients  $(\widehat{a}_1, \dots, \widehat{a}_\tau)$ , for reasons of parsimony, it may be advantageous to “model” the residual process  $u_t$  by some parametric autoregressive moving average (ARMA) process  $\mathbf{A}(L)u_t = \mathbf{B}(L)\varepsilon_t$ ; estimates of  $a_j$  may be obtained from inverting  $\mathbf{B}(L)$ .

In Section 3 we show that under appropriate assumptions, the proposed estimator  $\widetilde{m}(x)$  is asymptotically equivalent to the infeasible estimator  $\overline{m}(x)$ , which is more efficient than the conventional kernel estimation. In fact, the transformation that we propose is also effective in parametric models, although not as effective as a full GLS transform (see Kristensen and Linton 2001).

Recently, Vilar-Fernandez and Francisco-Fernandez (2002) analyzed an alternative modification of standard local polynomial regression. They included a “GLS weighting” for autocorrelation in the criterion function. The resulting estimator

involves transformation of both  $Y$  and  $X$  processes by a matrix  $\mathbf{P}$ , which is the square root of the inverse covariance matrix of  $(u_1, \dots, u_T)$ . This transformation does not improve the first-order properties of the estimator, although these authors have shown in simulations that it can improve the finite-sample MSE.

### 3. MAIN RESULT

We assume that the error process  $\{u_t\}$  is independent of the process  $\{X_t\}$ , but discuss some extensions at the end of this section. To proceed, we assume that  $\{X_t\}$  is an  $\alpha$ -mixing process. Let  $\mathcal{F}_a^b$  be the  $\sigma$ -algebra of events generated by the random variables  $\{X_t; a \leq j \leq b\}$ . The stationary process  $\{X_t\}$  is called strongly mixing (Rosenblatt 1956) if

$$\sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} |\Pr(A \cap B) - \Pr(A)\Pr(B)| \equiv \alpha(k) \rightarrow 0, \text{ as } k \rightarrow \infty. \quad (10)$$

To facilitate the asymptotic analysis, we make the following assumptions on the residuals and regressors, the kernel function  $k(\cdot)$ , and the bandwidth parameters  $h_0$  and  $h_1$ .

1. The kernels  $K = K_i, i = 0, 1$ , satisfy  $K_i(u) = \prod_{j=1}^d k_i(u_j)$ , where  $k = k_i, i = 0, 1$ , are bounded, have compact support  $[-1, 1]$ , and are symmetric about 0. They also satisfy the property that  $\int k(u) du = 1$ . The functions  $H_{ij}(u) = u^j K_i(u)$  for all  $j$  with  $0 \leq |j| \leq 2p + 1$  are Lipschitz continuous; that is, there exists a positive finite constant  $C$  such that  $|H_{ij}(u) - H_{ij}(v)| \leq C\|u - v\|$ .
2. The process  $\{X_t\}$  is strongly mixing with  $\sum_{i=1}^\infty i^\delta \times \{\alpha(i)\}^{1-2/\nu} < \infty$  for some  $2 < \nu$  and  $\delta > 1 - 2/\nu$ . The density  $f_X$  of  $X_t$  and the joint densities of  $(X_t, X_{t+\ell}), (X_t, X_{t+\ell}, X_{t+j}), (X_t, X_{t+\ell}, X_{t+j}, X_{t+s})$  are uniformly bounded and are bounded away from 0 on their supports.
3. For some  $\theta \geq \nu, E(|u_t|^\theta) < \infty$ .
4. The function  $m(\cdot)$  is  $(p + 1)$  times partially differentiable, and the  $(p + 1)$ st-order partial derivatives are Lipschitz continuous on  $\mathcal{X}$ . The first-order partial derivatives of  $f_X$  exist and are continuous on  $\mathcal{X}$ .
5. The process  $\{u_t\}$  is a stationary invertible linear process representable in the form of (2) and has inverse (3). In addition, there exists some  $\lambda \in (0, 1)$  such that the linear process coefficients  $|a_j|$  are bounded by a constant multiple of  $\lambda^j$ .
6. The truncation parameter  $\tau$  satisfies  $\tau(T) = \kappa \log T$  for some  $\kappa > 0$ .
7. Bandwidths  $h_0$  and  $h_1$  satisfy that  $h_0/h_1 \rightarrow 0, T^{1/2} \times h_1^{d/2} h_0^{2q} (\log T) \rightarrow 0$ , and  $T^{-1/2} h_0^{-d} h_1^{d/2} (\log T) \rightarrow 0$ , where  $q = p + 1$ .

The stationarity condition rules out examples like  $X_t = f(t/T)$  for smooth  $f$ . Assumption 1 is a standard assumption for kernel functions in nonparametric estimation. Under the mixing conditions of assumption 2, the temporal dependence among  $\{X_t\}$  decreases sufficiently fast as the time distance increases and thus is asymptotically ignorable. In particular, the strong law of large numbers and the central limit theorems continue to hold for standardized summations, and uniform

convergence results on the kernel smooth quantities still hold. The differentiability of assumption 4 ensures a Taylor expansion to appropriate order. Whereas assumption 5 is stronger than the summability conditions of, say, Phillips and Solo (1992), the dominance requirement that  $|a_j|$  are bounded by a constant multiple of  $\lambda^j$  is sufficiently general to include leading cases like the widely considered stationary invertible ARMA process. This dominance condition is useful in our technical development and in particular provides a sufficient condition for controlling the order of magnitude of various summations involving  $c_j$ . No doubt this condition could be weakened, but we do not attempt to do so or to find minimal conditions under which our results hold. The expansion rate of the truncation parameter given in assumption 6 is also for convenience, and our results hold for a much wider range of  $\tau$ . In fact, from the proof in the Appendix, we can see that as long as the tail summation  $(\sum_{j=\tau+1}^\infty a_j)$  of the sequence  $a_j$  is controlled under appropriate order, our results still hold. Assumption 7 assumes that we undersmooth in the preliminary estimation stage, so that the bias term coming from preliminary estimation will be smaller than the leading bias term. Consequently, the feasible estimator has the same asymptotic MSE as the infeasible estimator  $\bar{m}$ . Note that if we take  $h_1 = O(T^{-1/(2q+d)})$ , then assumption 7 is satisfied for all  $q, d$  and many sequences  $h_0(T)$ .

*Theorem 2.* Suppose that assumptions 1–7 hold. Then,

$$\sqrt{Th_1^d}(\tilde{m}(x) - m(x) - h_1^q[\mathbf{M}^{-1}\mathbf{Bm}^{(q)}(x)]_{0,0}) \Rightarrow N\left(0, \frac{\sigma_\varepsilon^2}{f_X(x)}[\mathbf{M}^{-1}\mathbf{\Gamma M}^{-1}]_{0,0}\right),$$

where  $\mathbf{M}, \mathbf{\Gamma}, \mathbf{B}$  are defined by (7) with kernel  $K_1$ .

We have a sort of ‘oracle’ property here: The feasible estimator  $\tilde{m}(x)$  is asymptotically equivalent to  $\bar{m}(x)$  and hence is more efficient than  $\hat{m}(x)$ . By undersmoothing the pilot estimator  $\hat{m}(x)$ , we can make the bias of  $\tilde{m}(x)$  the same as the bias of  $\hat{m}(x)$ . Therefore,  $\tilde{m}(x)$  should be preferred to  $\hat{m}(x)$ .

The asymptotic normal distribution given by Theorem 2 can be used to calculate pointwise confidence intervals for estimators described here. To do this, we require an estimate of the asymptotic variance. (For literature on variance estimation under dependency, see, e.g., Müller and Stadtmüller 1988). Let  $\mathcal{M}_T$  and  $\mathbf{G}_T$  be the  $N \times N$ -dimensional matrices

$$\mathcal{M}_T = \begin{bmatrix} \mathcal{M}_{0,0} & \mathcal{M}_{0,1} & \cdots & \mathcal{M}_{0,p} \\ \mathcal{M}_{1,0} & \mathcal{M}_{1,1} & \cdots & \mathcal{M}_{1,p} \\ \vdots & & & \vdots \\ \mathcal{M}_{p,0} & \mathcal{M}_{p,1} & \cdots & \mathcal{M}_{p,p} \end{bmatrix}$$

and

$$\mathbf{G}_T = \begin{bmatrix} \mathbf{G}_{0,0} & \mathbf{G}_{0,1} & \cdots & \mathbf{G}_{0,p} \\ \mathbf{G}_{1,0} & \mathbf{G}_{1,1} & \cdots & \mathbf{G}_{1,p} \\ \vdots & & & \vdots \\ \mathbf{G}_{p,0} & \mathbf{G}_{p,1} & \cdots & \mathbf{G}_{p,p} \end{bmatrix},$$

where  $\mathcal{M}_{i,j}$  and  $\mathbf{G}_{i,j}$  are  $N_i \times N_j$ -dimensional matrices whose  $(l, r)$  elements are

$$\sum_{t=1}^T \left(\frac{x - X_t}{h}\right)^{\phi_{|l|}(l) + \phi_{|r|}(r)} K_1\left(\frac{x - X_t}{h_1}\right)$$

and

$$\sum_{t=1}^T \left( \frac{x - X_t}{h} \right)^{\phi_{|j|}(l) + \phi_{|k|}(r)} K_1 \left( \frac{x - X_t}{h_1} \right)^2 \tilde{\varepsilon}_t^2,$$

where  $\tilde{\varepsilon}_t = \widehat{Y}_t - \tilde{m}(X_t)$ . Let  $\tilde{v}(x) = [\mathcal{M}_T^{-1} \mathbf{G}_T \mathcal{M}_T^{-1}]_{0,0}$  and let  $z_\alpha$  be the standard normal  $1 - \alpha$  critical value. Then

$$\tilde{m}(x) \pm z_{\alpha/2} \sqrt{\tilde{v}(x)} \tag{11}$$

is a valid two-sided level  $\alpha$  pointwise confidence intervals provided that the estimation is undersmoothed, i.e.,  $h_1 = o(T^{-1/(2q+d)})$ . By definition,  $\varepsilon_t$  is supposed to be an uncorrelated sequence, so we might expect these standard errors to be more accurate than those for  $\widehat{m}(x)$ .

One may substitute different smoothers and may use a different estimation scheme to obtain the  $\widehat{a}'_j$ s. One can also expect some improvement by iterating the process. Specifically, define again

$$\tilde{Y}_t = Y_t - \sum_{j=1}^{\tau} \tilde{a}_j (Y_{t-j} - \tilde{m}(X_{t-j})),$$

where  $(\tilde{a}_1, \dots, \tilde{a}_\tau)'$  are obtained from the least squares regression of  $Y_t - \tilde{m}(X_t)$  on  $(Y_{t-1} - \tilde{m}(X_{t-1}), \dots, Y_{t-\tau} - \tilde{m}(X_{t-\tau}))'$  and local polynomial smooth  $\tilde{Y}_t$  against  $X_t$ .

Finally, we discuss the assumption of independence of  $\{X_t\}$  from  $\{u_t\}$ . As we stated after (1), the main assumption is that  $E(u_t | X_1, \dots, X_T) = 0$ , which effectively rules out  $X_t$  containing lagged  $Y_t$ . This is quite natural, because when  $X_t$  contains lagged  $Y_t$  and  $u_t$  is autocorrelated the standard local polynomial estimation method may not even be consistent. For example, consider the case where  $Y_t$  is an ARMA(1, 1) process; even the OLS estimator of  $Y_t$  on  $Y_{t-1}$  is inconsistent. In these cases, we need an alternative estimation strategy. Alternatively, we can include sufficient lagged  $Y$ 's in  $m$  so that the error term is approximately uncorrelated. Either way, our method and approach are not directly relevant. However, there are other types of dependence between the error process and the covariate process that we can allow for, like heteroscedasticity. For example, suppose that  $u_t = \sigma(X_t)v_t$  with  $\sigma(X_t)$  a smooth function bounded away from 0 and  $E(v_t | X_1, \dots, X_T) = 0$  and  $\text{cov}(v_s, v_t | X_1, \dots, X_T) = \gamma_v(|s - t|)$  for some covariance function  $\gamma_v$ . Under some further regularity conditions (see, e.g., Masry 1996a,b) on the dependence of the joint process  $(Y_t, X_t)$ , it can be shown that our main result continues to hold with the same  $\tilde{Y}_t$ , and indeed (11) is still valid as stated.

#### 4. EFFICIENT ESTIMATION

We now discuss how we can improve the efficiency of our estimator even more and approach a sort of GLS bound. Notice that for each  $j$  where  $a_j \neq 0$ , we can rewrite (4) as

$$\underline{Y}_t^j = m(X_{t-j}) + \frac{1}{a_j} \varepsilon_t, \tag{12}$$

where

$$\underline{Y}_t^j = \frac{1}{a_j} \left[ a(L)Y_t - \sum_{k \neq j}^{\infty} a_k m(X_{t-k}) \right].$$

That is, the transformation that we discussed in Section 2 is just a special case of a whole family of transformations, one associated with each lag. Given some estimate of  $\underline{Y}_t^j$ , denoted by  $\widehat{\underline{Y}}_t^j$ , we can now smooth this against  $X_{t-j}$  and call the resulting estimator  $\tilde{m}_j(x)$ . Then we have under the same conditions as before that  $\tilde{m}_j(x)$  has asymptotic variance proportional to  $\sigma_\varepsilon^2/a_j^2$  for any  $j$  where  $a_j \neq 0$ . Furthermore,  $\tilde{m}_j(x)$  and  $\tilde{m}_k(x)$  for  $j \neq k$  will be asymptotically independent, because they are smoothers on different random variables  $X_{t-j}$  and  $X_{t-k}$ ; that is, the sets  $\{x : |x - X_{t-j}| < h\}$  and  $\{x : |x - X_{t-k}| < h\}$  have an intersection that contains fewer observations than either set itself by an order of magnitude. The bias of each  $\tilde{m}_j(x)$  is the same, because the target function is the same. By combining the estimators, we can improve efficiency, specifically the variance. Consider the class containing all estimators of the form  $\sum_{j=0}^J \omega_j \tilde{m}_j(x)$  for some weighting sequence  $\omega_j$  that satisfies  $\sum_{j=0}^J \omega_j = 1$ , where  $J$  is a given integer. It is easy to see that the member of this class of estimators with the smallest asymptotic variance (they all have the same bias) is

$$\tilde{m}_{\text{eff}}(x) = \sum_{j=0}^J \omega_j^{\text{eff}} \tilde{m}_j(x), \quad \text{where } \omega_j^{\text{eff}} = \frac{a_j^2}{\sum_{j=0}^J a_j^2}.$$

Indeed, it follows from the same arguments as in the proof of Theorem 1 that

$$\sqrt{Th_1^d} (\tilde{m}_{\text{eff}}(x) - m(x) - h_1^q [\mathbf{M}^{-1} \mathbf{B} m^{(q)}(x)]_{0,0}) \Rightarrow N \left( 0, \frac{\sigma_\varepsilon^2}{f_X(x) \sum_{j=0}^J a_j^2} [\mathbf{M}^{-1} \mathbf{\Gamma} \mathbf{M}^{-1}]_{0,0} \right) \tag{13}$$

for any fixed  $J$ . Under some conditions, it may be possible to extend this result to the case where  $J \rightarrow \infty$  and where one has estimated weights  $\tilde{a}_j$  instead of  $a_j$ . We would expect to just replace  $J$  by  $\infty$  in (13), and that the estimation of weights would have no effect. (See Chen and Linton 2001 for a result of this type in a different context.) Anticipating this result, we expect that because  $a_0, c_0 = 1$ , we would have [setting  $J = \infty$  in (13)]

$$\begin{aligned} \frac{\text{avar}[\tilde{m}_{\text{eff}}(x)]}{\text{avar}[\widehat{m}(x)]} &= \frac{1}{\sum_{j=0}^{\infty} a_j^2 \sum_{j=0}^{\infty} c_j^2} \\ &\leq \frac{\text{avar}[\tilde{m}(x)]}{\text{avar}[\widehat{m}(x)]} = \frac{1}{\sum_{j=0}^{\infty} c_j^2} \leq 1. \end{aligned}$$

We expect that  $\text{avar}[\tilde{m}_{\text{eff}}(x)]$  provides a lower bound achievable by this sort of method rather like that achieved in the linear regression case by Hannan (1963). In the AR(1) case, the asymptotic variance of  $\tilde{m}(x)$  is  $(\|K\|^2/f_X(x))\sigma_\varepsilon^2/(1 - a^2)$ , whereas that of  $\tilde{m}_{\text{eff}}(x)$  is  $(\|K\|^2/f_X(x))\sigma_\varepsilon^2/(1 + a^2)$ . Compare this with the linear regression model  $Y_t = \beta X_t + u_t$ , where  $X_t$  is an iid process with mean 0. The variance of the OLS estimator of  $\beta x$  is  $(x^2/\sigma_X^2)\sigma_\varepsilon^2/(1 - a^2)$ , and that of the GLS estimator of  $\beta x$  is  $(x^2/\sigma_X^2)\sigma_\varepsilon^2/(1 + a^2)$ . This is suggestive that  $\tilde{m}_{\text{eff}}(x)$  is like GLS.

In practice, the gain of  $\tilde{m}_{\text{eff}}(x)$  over  $\tilde{m}(x)$  may not be so great in comparison with the gain of  $\tilde{m}(x)$  over  $\widehat{m}(x)$ . For example, in the AR(1) case, the improvement of  $\tilde{m}(x)$  over the usual kernel smoother  $\widehat{m}(x)$  can be arbitrarily large, but  $\tilde{m}_{\text{eff}}(x)$  can have at best only half the variance of  $\tilde{m}(x)$ . Therefore, it may be that in

practice, the benefit from computing  $\tilde{m}_{eff}(x)$  may be exceeded by its small sample cost. We investigate this in the simulation experiments reported in the next section.

### 5. NUMERICAL RESULTS

We investigate the proposed estimator  $\tilde{m}(x)$  on simulated data, as well as the estimator  $\tilde{m}_{eff}(x)$  considered in Section 5. We compare these estimators with the conventional local polynomial estimator  $\hat{m}(x)$ . We have not tried to optimize the performance of either the conventional estimator or our own, more efficient modifications. Rather, we have taken what are fairly common choices, in real applications, of bandwidth, and other variables, and demonstrate that even with these implementations there are finite sample gains to be made. The technical report version of this article (Xiao et al. 2003) contains more results and an application to financial data. Here we just report a selection of our results.

The error process  $u_t$  is various special cases of the ARMA(1, 1) process

$$u_t = \alpha_1 u_{t-1} + \varepsilon_t + \gamma_1 \varepsilon_{t-1},$$

where  $\varepsilon_t$  is iid  $N(0, 1)$ . Various values for  $\alpha_1, \gamma_1$  are considered. We take  $m(x) = x$ , where  $X_t$  are chosen to be iid  $U[-1, 1]$ . We report the results for just the  $T = 100$  and 500 cases. The number of replications is 200.

Regarding the implementation of the estimators, we chose exactly the same kernel and bandwidth in all these three estimators, that is,  $h = h_0 = h_1$ , and so on. Specifically, we use the Gaussian kernel and bandwidth  $h = 1.06s_X T^{-1/5}$ . We used a third-order local polynomial in the results presented here. In our estimator we consider AR(2) prewhitening. The AR parameters in the prewhitening process are estimated by least squares. In computing  $\tilde{m}_{eff}$ , we take  $J = \tau = 2$ .

We report the average squared errors (denoted by ISE) (the average over all sample points  $X_1, \dots, X_T$ ), and the relative efficiency calculated based on the ratio of average squared errors (denoted as RE). Tables 1 and 2 correspond to the different sample sizes. In these tables, ISE0, ISE1, and ISE2 give the average squared errors of the conventional local polynomial estimator  $\hat{m}(x)$  and the proposed efficient estimators  $\tilde{m}(x)$  and  $\tilde{m}_{eff}(x)$ , RE1 reports the relative efficiency of the proposed efficient estimator  $\tilde{m}(x)$  over the conventional estimator  $\hat{m}(x)$ , and RE2 reports the relative efficiency of  $\tilde{m}_{eff}(x)$  over  $\hat{m}(x)$ . Qualitatively

Table 1. Comparison With the Conventional Estimator,  $n = 100$

ARMA parameters		Integrated squared errors			Relative efficiency	
$\alpha_1$	$\gamma_1$	ISE0	ISE1	ISE2	RE1	RE2
0	0	.0641	.0658	.0661	1.027	1.031
0	.1	.0770	.0776	.0778	1.007	1.010
0	.2	.0912	.0905	.0901	.992	.988
0	.5	.1413	.1363	.1331	.964	.942
0	.7	.1809	.1733	.1660	.957	.916
0	.9	.2255	.2157	.2044	.956	.906
.1	0	.0784	.0789	.0791	1.006	1.009
.2	0	.0983	.0968	.0970	.984	.986
.5	0	.2393	.2216	.2227	.926	.931
.7	0	.6073	.5409	.5375	.891	.885
.9	0	3.6257	3.1559	3.1297	.870	.863
.2	.2	.1402	.1343	.1341	.957	.956
.2	.5	.1789	.1701	.1675	.951	.936
.5	.2	.5012	.4601	.4616	.918	.921

NOTE: AR(2) prewhitening,  $m(x) = x$ , estimate all points.

Table 2. Comparison With the Conventional Estimator,  $n = 500$

ARMA parameters		Integrated squared errors			Relative efficiency	
$\alpha_1$	$\gamma_1$	ISE0	ISE1	ISE2	RE1	RE2
0	0	.01463	.01498	.01487	1.024	1.016
0	.1	.01802	.01775	.01792	.985	.994
0	.2	.02143	.02076	.02110	.968	.984
0	.5	.03345	.03150	.03140	.942	.938
0	.7	.04295	.04018	.03940	.935	.917
0	.9	.05360	.05006	.04870	.933	.908
.1	0	.01838	.01806	.01827	.982	.994
.2	0	.02323	.02231	.02281	.960	.982
.5	0	.05895	.05318	.05399	.902	.916
.7	0	.16130	.14000	.14021	.868	.869
.9	0	1.32120	1.10900	1.10240	.839	.834
.2	.2	.03340	.03122	.03168	.934	.948
.2	.5	.03812	.03579	.03499	.939	.918
.5	.2	.12181	.10829	.10951	.889	.899

NOTE: AR(2) prewhitening,  $m(x) = x$ , at all points.

similar results are obtained from other cases (see the working paper version of this article for more details).

Some general conclusion can be drawn from our simulation experiments:

1. The results show that the relative efficiency improves with sample size. There is likely a considerable small sample effect that is dominating in this range of parameters, and this requires a very large sample indeed before the asymptotic predictions become reality. Nevertheless, in most cases apart from the iid case (where all parameters are 0's) our estimator improves on the standard kernel procedure.
2. When the underlying process has a nontrivial moving average (MA) part, our method is likely to be quite far from matching the true autocorrelation structure in the errors. Nevertheless, even in those cases there are positive results.
3. In general, the more serial correlation, the larger efficiency gain achieved from our prewhitening procedure. However, consider the autocorrelation 1 [AR(1)] case, for example; note that the relative efficiency first improves as the AR coefficient increases, and then worsens as it approaches 1. This is due in part to the large downward bias in estimating  $\alpha$  in this region. We could perhaps improve the relative efficiency by taking a larger bandwidth in the second step, as would be permitted by our theory.
4. Both  $\tilde{m}(x)$  and  $\tilde{m}_{eff}(x)$  improve the estimation in the presence of serial correlation, especially for large sample sizes, but none of them dominates the other. It seems that  $\tilde{m}(x)$  performs slightly better than  $\tilde{m}_{eff}(x)$  when the true error process is actually an AR process. This is intuitive, because an AR prewhitening was used. But different results were obtained when the error terms are MA processes.

### 6. CONCLUSIONS

We have proposed a new method for improving the efficiency of local polynomial estimators in time series nonparametric regression. The asymptotic improvement in the variance depends only on the autocorrelation function of the error process, and the improvement can be arbitrarily large depending on this. In simulation experiments, we have shown that some improvement

is possible in small samples even with the sort of bandwidth choices and other choices that are widely used in practice but are suboptimal for our method. We expect that the numerical performance of our method can be considerably improved in small samples. First, better bandwidth choice and order of the autoregression should make a big difference to the performance of our method. (Some approaches to this are discussed in the working version of the article.) Second, iterating the procedure may confer benefits through more accurate estimates of the autoregressive coefficients. In summary, we think our method can deliver performance benefits for a wide variety of time series situations in which nonparametric methods are applied.

APPENDIX: PROOFS OF THEOREMS

We use  $\|\bullet\|$  to denote the Euclidean norm of  $\bullet$  and  $C$  to signify a generic positive constant whose exact value may vary from case to case. We denote  $\phi(x, y, z, \dots)$  as a general function whose exact form may change from case to case. For two random variables  $X_T$  and  $Y_T$ , we say that  $X_T \simeq Y_T$  whenever  $X_T = Y_T(1 + o_p(1))$  as  $T \rightarrow \infty$ . Given kernel function  $K$  and bandwidth  $h$ , we define the  $N \times N$ -dimensional matrices  $\mathbf{M}$  and  $\mathbf{\Gamma}$  and  $N \times 1$  vector  $\mathbf{B}$  by (7) and denote  $\mathbf{M}^{-1} = (\mathbf{M}^{i,j})_{i,j=1}^p$ , where the partition of submatrices in  $\mathbf{M}^{-1}$  is conformable to that of  $\mathbf{M}$ . The asymptotic properties of local polynomial estimator have been well developed and documented (see, e.g., Fan and Gijbels 1996; Masry 1996a,b; and the references therein).

Proof of Theorem 1

The proof follows work of Masry (1996a). The  $p$ th-order local polynomial regression of  $\underline{Y}_i$  on  $X_i$  gives  $\bar{m}(x) = \mathbf{e}'_1 \mathbf{M}_T^{-1} \underline{\Psi}_T$ , where  $\mathbf{e}_1 = (1, 0, \dots, 0)'$  and  $\mathbf{M}_T(x)$  and  $\underline{\Psi}_T(x)$  are a symmetric  $N \times N$  ( $N = \sum_{\ell=0}^p N_\ell \times 1$ ) matrix and an  $N \times 1$ -dimensional column vector and are defined as

$$\mathbf{M}_T(x) = \begin{bmatrix} \mathbf{M}_{T,0,0}(x) & \mathbf{M}_{T,0,1}(x) & \dots & \mathbf{M}_{T,0,p}(x) \\ \vdots & \mathbf{M}_{T,1,1}(x) & \dots & \mathbf{M}_{T,1,p}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{M}_{T,p,0}(x) & \dots & \dots & \mathbf{M}_{T,p,p}(x) \end{bmatrix}$$

and

$$\underline{\Psi}_T(x) = \begin{bmatrix} \underline{\Psi}_{T,0}(x) \\ \underline{\Psi}_{T,1}(x) \\ \vdots \\ \underline{\Psi}_{T,p}(x) \end{bmatrix},$$

where  $\mathbf{M}_{T,|j|,|k|}(x)$  is an  $N_{|j|} \times N_{|k|}$ -dimensional submatrix with the  $(l, r)$  element given by

$$[\mathbf{M}_{T,|j|,|k|}]_{l,r} = \frac{1}{Th^d} \sum_{i=1}^T \left(\frac{x - X_i}{h}\right)^{\phi_{|j|}(l) + \phi_{|k|}(r)} K\left(\frac{x - X_i}{h}\right)$$

and  $\underline{\Psi}_{T,|j|}(x)$  is an  $N_{|j|}$ -dimensional subvector whose  $r$ th element is given by

$$[\underline{\Psi}_{T,|j|}]_r = \frac{1}{Th^d} \sum_{i=1}^T \left(\frac{x - X_i}{h}\right)^{\phi_{|j|}(r)} K\left(\frac{x - X_i}{h}\right) \underline{Y}_i.$$

Note that  $\underline{Y}_i = m(X_i) + \varepsilon_i$ ; we obtain  $\bar{m}(x) \equiv m(x) + \bar{B}_x + \bar{V}_x$ , where  $\bar{B}_x$  is the bias term  $\bar{B}_x = \mathbf{e}'_1 \mathbf{M}_T^{-1}(x) \mathbf{B}_T(x)$  and  $\mathbf{B}_T(x)$  is an  $N \times 1$  vector

$$\mathbf{B}_T(x) = \begin{bmatrix} \mathbf{B}_{T,0}(x) \\ \mathbf{B}_{T,1}(x) \\ \vdots \\ \mathbf{B}_{T,d}(x) \end{bmatrix},$$

where  $\mathbf{B}_{T,|j|}(x)$  are an  $N_{|j|}$ -dimensional subvectors whose  $r$ th elements are given by

$$[\mathbf{B}_{T,|j|}]_r = \frac{1}{Th^d} \sum_{i=1}^n \left(\frac{x - X_i}{h}\right)^{\phi_{|j|}(r)} K\left(\frac{x - X_i}{h}\right) \Delta_i(x),$$

where  $\Delta_i(x) = m(X_i) - \frac{1}{k!} \sum_{0 \leq |k| \leq p} (D^k m)(x) (X_i - x)^k$ , and  $\bar{V}_x$  is the variance effect defined by  $\bar{V}_x = \mathbf{e}'_1 \mathbf{M}_T^{-1}(x) \underline{\mathbf{U}}_T(x)$ . The stochastic term  $\underline{\mathbf{U}}_T(x)$  is also an  $N \times 1$  vector

$$\underline{\mathbf{U}}_T(x) = \begin{bmatrix} \underline{\mathbf{U}}_{T,0}(x) \\ \underline{\mathbf{U}}_{T,1}(x) \\ \vdots \\ \underline{\mathbf{U}}_{T,p}(x) \end{bmatrix},$$

where  $\underline{\mathbf{U}}_{T,|j|}(x)$  is an  $N_{|j|}$ -dimensional subvector whose  $r$ th elements are given by

$$[\underline{\mathbf{U}}_{T,|j|}]_r = \frac{1}{Th^d} \sum_{i=1}^n \left(\frac{x - X_i}{h}\right)^{\phi_{|j|}(r)} K\left(\frac{x - X_i}{h}\right) \varepsilon_i.$$

By the results of Masry (1996a),  $\mathbf{M}_T(x)$  converges in mean square to  $f(x)\mathbf{M}$ , and  $h^{-(p+1)}\mathbf{B}_T(x)$  converges in mean square to  $f(x)\mathbf{B} \cdot \mathbf{m}^{(p+1)}(x)$ . In addition,

$$\sqrt{Th^d} \mathbf{M}_T^{-1}(x) \underline{\mathbf{U}}_T(x) \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2}{f_X(x)} \mathbf{M}^{-1} \mathbf{\Gamma} \mathbf{M}^{-1}\right);$$

thus

$$\begin{aligned} &\sqrt{Th^d} (\bar{m}(x) - m(x) - h^{p+1} [\mathbf{M}^{-1} \mathbf{B} \mathbf{m}^{(p+1)}(x)]_{0,0}) \\ &\quad \Rightarrow N\left(0, \frac{\sigma_\varepsilon^2}{f_X(x)} [\mathbf{M}^{-1} \mathbf{\Gamma} \mathbf{M}^{-1}]_{0,0}\right). \end{aligned}$$

Proof of Theorem 2

First, note that  $\tilde{m}(x) = \mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_T$ , where  $\mathbf{M}_T$  and  $\tilde{\Psi}_T$  are defined similarly to  $\mathbf{M}_T$  and  $\underline{\Psi}_T$  in the proof of Theorem 1 but with  $K, h$ , and  $\underline{Y}_i$  replaced by  $K_1, h_1$ , and  $\hat{Y}_i$ . We decompose  $\tilde{m}(x)$  into  $\bar{m}(x)$  plus error terms coming from the preliminary estimation and the truncation, and show that these terms are small-order terms. First, we write

$$\begin{aligned} \hat{Y}_t &= Y_t - \sum_{j=1}^{\tau} \hat{a}_j (Y_{t-j} - \hat{m}(X_{t-j})) \\ &= Y_t - \sum_{j=1}^{\infty} a_j u_{t-j} + \sum_{j=\tau+1}^{\infty} a_j u_{t-j} - \sum_{j=1}^{\tau} (\hat{a}_j - a_j) u_{t-j} \\ &\quad + \sum_{j=1}^{\tau} a_j (\hat{m}(X_{t-j}) - m(X_{t-j})) \\ &\quad + \sum_{j=1}^{\tau} (\hat{a}_j - a_j) (\hat{m}(X_{t-j}) - m(X_{t-j})). \end{aligned}$$

Then  $\tilde{\Psi}_T = \tilde{\Psi}_T(x)$  may be decomposed as

$$\tilde{\Psi}_T(x) = \underline{\Psi}_T(x) + \tilde{\Psi}_{T1}(x) - \tilde{\Psi}_{T2}(x) + \tilde{\Psi}_{T3}(x) + \tilde{\Psi}_{T4}(x),$$

where again  $\tilde{\Psi}_{Tl}(x), l = 1, 2, 3, 4$ , are defined similarly to  $\underline{\Psi}_T(x)$ , with the  $r$ th element of  $\tilde{\Psi}_{Tl,|j|}(x)$  given by

$$\begin{aligned} [\tilde{\Psi}_{T1,|j|}]_r &= \frac{1}{Th_1^d} \sum_{i=1}^T \left(\frac{x - X_i}{h_1}\right)^{\phi_{|j|}(r)} \\ &\quad \times K_1\left(\frac{x - X_i}{h_1}\right) \left(\sum_{j=\tau+1}^{\infty} a_j u_{i-j}\right), \end{aligned}$$

$$[\tilde{\Psi}_{T2,|j|}]_r = \frac{1}{Th_1^d} \sum_{i=1}^T \left(\frac{x-X_i}{h_1}\right)^{\phi_{|j|}(r)} \times K_1\left(\frac{x-X_i}{h_1}\right) \left(\sum_{j=1}^{\tau} (\hat{a}_j - a_j) u_{i-j}\right),$$

$$[\tilde{\Psi}_{T3,|j|}]_r = \frac{1}{Th_1^d} \sum_{i=1}^T \left(\frac{x-X_i}{h_1}\right)^{\phi_{|j|}(r)} K_1\left(\frac{x-X_i}{h_1}\right) \times \left(\sum_{j=1}^{\tau} a_j (\hat{m}(X_{i-j}) - m(X_{i-j}))\right),$$

and

$$[\tilde{\Psi}_{T4,|j|}]_r = \frac{1}{Th_1^d} \sum_{i=1}^T \left(\frac{x-X_i}{h_1}\right)^{\phi_{|j|}(r)} K_1\left(\frac{x-X_i}{h_1}\right) \times \left(\sum_{j=1}^{\tau} (\hat{a}_j - a_j) (\hat{m}(X_{i-j}) - m(X_{i-j}))\right).$$

Substituting the foregoing expression into  $\tilde{m}(x)$ , we have

$$\tilde{m}(x) = \bar{m}(x) + Q_{T1} - Q_{T2} + Q_{T3} + Q_{T4},$$

where  $Q_{Tl} = \mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{Tl}$ ,  $l = 1, 2, 3, 4$ .

We analyze the asymptotic properties of  $Q_{Tl}$ ,  $l = 1, \dots, 4$ , in Lemmas A.1–A.4, which are key results for proof of the theorem.

*Lemma A.1.* Under assumptions 1–7,

$$Q_{T1} = o_p(T^{-1/2}h_1^{-d/2}).$$

*Proof of Lemma A.1.*  $Q_{T1}$  is of smaller order because of the tail properties of the summable sequence  $a_j$ . First, note that

$$Q_{T1} = \mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T1} = \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} \tilde{\Psi}_{T1} (1 + o_p(1)),$$

by the uniform convergence result of  $\mathbf{M}_T(x)$  (Masry 1996b). Because  $f_X(x) > 0$ , we need only verify, for each  $r$ , the order of

$$\frac{1}{Th_1^d} \sum_{i=1}^T \left(\frac{x-X_i}{h_1}\right)^r K_1\left(\frac{x-X_i}{h_1}\right) \left(\sum_{j=\tau+1}^{\infty} a_j u_{i-j}\right).$$

Note that this has mean 0 and

$$\begin{aligned} \text{var} & \left[ \frac{1}{Th_1^d} \sum_{i=1}^T \left(\frac{x-X_i}{h_1}\right)^r K_1\left(\frac{x-X_i}{h_1}\right) \sum_{j=\tau+1}^{\infty} a_j u_{i-j} \right] \\ & = \left(\frac{1}{Th_1^d}\right)^2 \left\{ \sum_{t=s=1}^T \mathbb{E} \left[ \left(\frac{x-X_t}{h_1}\right)^{2r} K_1\left(\frac{x-X_t}{h_1}\right)^2 \right] \right. \\ & \quad \times \left. \sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_i a_j \gamma_u(|i-j|) \right\} \\ & \quad + \left(\frac{1}{Th_1^d}\right)^2 \left\{ \sum_{t=1}^T \sum_{s=1, s \neq t}^T \mathbb{E} \left[ \left(\frac{x-X_t}{h_1}\right)^r \left(\frac{x-X_s}{h_1}\right)^r \right. \right. \\ & \quad \times \left. \left. K_1\left(\frac{x-X_t}{h_1}\right) K_1\left(\frac{x-X_s}{h_1}\right) \right] \right. \\ & \quad \times \left. \sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_j a_i \gamma_u(|t-s+i-j|) \right\}. \end{aligned}$$

The first term is  $o(T^{-1}h_1^{-d})$  because

$$\begin{aligned} & \left(\frac{1}{Th_1^d}\right)^2 \left\{ \sum_{t=1}^T \mathbb{E} \left[ \left(\frac{x-X_t}{h_1}\right)^{2r} K_1\left(\frac{x-X_t}{h_1}\right)^2 \right] \right. \\ & \quad \times \left. \sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_i a_j \gamma_u(|i-j|) \right\} \\ & \leq \left(\frac{1}{Th_1^d}\right)^2 T \cdot \mathbb{E} \left\{ \left(\frac{x-X_1}{h_1}\right)^{2r} K_1\left(\frac{x-X_1}{h_1}\right)^2 \right\} \\ & \quad \times \left\{ \sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_i a_j \right\} \sup_{0 \leq i, j \leq \infty} |\gamma_u(|i-j|)| \end{aligned}$$

and

$$\sup_{0 \leq j, l < \infty} |\gamma_u(|j-l|)| < \infty,$$

by the stationarity/mixing property of  $u$ ;

$$T \cdot \mathbb{E} \left[ \left(\frac{x-X_1}{h_1}\right)^{2r} K_1\left(\frac{x-X_1}{h_1}\right)^2 \right] = O(Th_1^d),$$

by a direct calculation of expectation,

$$\mathbb{E} \left[ \left(\frac{x-X_t}{h_1}\right)^{2r} K_1\left(\frac{x-X_t}{h_1}\right)^2 \right] = h_1^d \int u^{2r} K_1(u)^2 f_X(x-hu) du;$$

and

$$\sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_i a_j = o(1) \quad \text{as } \tau \rightarrow \infty,$$

by summability of  $\{a_j\}_{j=1}^{\infty}$ .

The second term is  $o(T^{-1}h_1^{-d})$  because

$$\begin{aligned} & \left(\frac{1}{Th_1^d}\right)^2 \left\{ \sum_{t=1}^T \sum_{s=1, s \neq t}^T \mathbb{E} \left[ \left(\frac{x-X_t}{h_1}\right)^r \left(\frac{x-X_s}{h_1}\right)^r \right. \right. \\ & \quad \times \left. \left. K_1\left(\frac{x-X_t}{h_1}\right) K_1\left(\frac{x-X_s}{h_1}\right) \right] \right. \\ & \quad \times \left. \sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_j a_i \gamma_u(|t-s+i-j|) \right\} \\ & = \left(\frac{1}{T}\right)^2 \sum_{t=1}^T \sum_{s=1, s \neq t}^T \left[ \int u^r v^r K_1(u) K_1(v) f_{X,|t-s|} \right. \\ & \quad \times \left. (x-uh_1, y-vh_1) dudv \right] \\ & \quad \times \sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_j a_i \gamma_u(|t-s+i-j|) \\ & \leq C \left(\frac{1}{T}\right)^2 \sum_{t=1}^T \sum_{s=1, s \neq t}^T \sum_{i=\tau+1}^{\infty} \sum_{j=\tau+1}^{\infty} a_j a_i \gamma_u(|t-s+i-j|), \end{aligned}$$

where the last inequality follows from the boundedness assumption of the density and joint densities and the fact that

$$\sup_{0 \leq i, j \leq \infty} \left| \sum_{s \neq t, t=1, s=1}^T \sum_{s=1}^T \gamma_u(|t-s+i-j|) \right| = O(T), \quad (\text{A.1})$$

where, again, the result (A.1) comes from the stationarity/mixing property of  $u$ . Thus

$$\text{var} \left[ \frac{1}{Th_1^d} \sum_{i=1}^T \left( \frac{x - X_i}{h_1} \right)^r K_1 \left( \frac{x - X_i}{h_1} \right) \left( \sum_{j=\tau+1}^{\infty} a_j u_{i-j} \right) \right] = o(T^{-1} h_1^{-d}).$$

Therefore, the magnitude of  $Q_{T1}$  is as stated.

*Lemma A.2.* Under assumptions 1–7,

$$Q_{T2} = o_p(T^{-1/2} h_1^{-d/2}).$$

*Proof of Lemma A2.* We decompose  $Q_{T2}$  into  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T2A} + \mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T2B}$ , where

$$[\tilde{\Psi}_{T2A, |j|}]_r = \frac{1}{Th_1^d} \sum_{i=1}^T \left( \frac{x - X_i}{h_1} \right)^{\phi_{|j|(r)}} \times K_1 \left( \frac{x - X_i}{h_1} \right) \left( \sum_{j=1}^{\tau} (\bar{a}_j - a_j) u_{i-j} \right)$$

and

$$[\tilde{\Psi}_{T2B, |j|}]_r = \frac{1}{Th_1^d} \sum_{i=1}^T \left( \frac{x - X_i}{h_1} \right)^{\phi_{|j|(r)}} \times K_1 \left( \frac{x - X_i}{h_1} \right) \left( \sum_{j=1}^{\tau} (\hat{a}_j - \bar{a}_j) u_{i-j} \right),$$

and show that both  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T2A}$  and  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T2B}$  are  $o_p(T^{-1/2} \times h_1^{-d/2})$ . First, we denote  $\bar{\mathbf{A}}_{\tau} = (\mathbf{U}'_{\tau} \mathbf{U}_{\tau})^{-1} \mathbf{U}'_{\tau} \mathbf{u} = (\bar{a}_1, \dots, \bar{a}_{\tau})'$ , where  $\mathbf{u} = (u_{\tau+1}, \dots, u_T)'$  and  $\mathbf{U}_{\tau}$  is like  $\hat{\mathbf{U}}_{\tau}$  with  $\hat{u}_t$  replaced by  $u_t$ , and write  $\hat{a}_j - a_j = (\hat{a}_j - \bar{a}_j) + (\bar{a}_j - a_j)$ ; that is,  $\hat{\mathbf{A}}_{\tau} - \mathbf{A}_{\tau} = (\hat{\mathbf{A}}_{\tau} - \bar{\mathbf{A}}_{\tau}) + (\bar{\mathbf{A}}_{\tau} - \mathbf{A}_{\tau})$ .

For  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T2A}$ , let  $\mathbf{U}_{\tau t} = (u_{t-1}, \dots, u_{t-\tau})'$  and define the  $\tau \times \tau$  matrices

$$\mathbf{G}_{\tau} = \frac{1}{T} \mathbf{U}'_{\tau} \mathbf{U}_{\tau} = \frac{1}{T} \sum_t \mathbf{U}_{\tau t} \mathbf{U}'_{\tau t} = \left( \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} u_{t-l} \right)_{j,l}$$

and

$$\mathbf{\Gamma}_{\tau} = \frac{1}{T} \mathbf{E}(\mathbf{U}'_{\tau} \mathbf{U}_{\tau}) = \frac{1}{T} \sum_t \mathbf{E} \mathbf{U}_{\tau t} \mathbf{U}'_{\tau t} = (\mathbf{E}(u_{t-j} u_{t-l}))_{j,l}.$$

Then there exists a  $c > 0$  such that  $\lambda_{\min}(\mathbf{\Gamma}_{\tau}) \geq c\tau^{-\alpha}$  for some  $\alpha > 0$ . Therefore,  $\|\mathbf{\Gamma}_{\tau}^{-1}\| \leq c^{-1} \tau^{\alpha}$ , and

$$\|\mathbf{G}_{\tau} - \mathbf{\Gamma}_{\tau}\| = O_p(Q_T), \tag{A.2}$$

where  $Q_T = \sqrt{(\log \log T)/T}$ , provided that  $\tau \leq (\log T)^{\kappa}$  for some  $\kappa > 0$  (Hannan and Deistler 1988, sec 5.3). Note that

$$\bar{\mathbf{A}}_{\tau} - \mathbf{A}_{\tau} = \mathbf{G}_{\tau}^{-1} \left[ \frac{1}{T} \sum_t \mathbf{U}_{\tau t} \left( \varepsilon_t + \sum_{j=\tau+1}^{\infty} a_j u_{t-j} \right) \right].$$

We verify the magnitude of  $\frac{1}{T} \sum_t \mathbf{U}_{\tau t} \varepsilon_t$  and  $\frac{1}{T} \sum_t \mathbf{U}_{\tau t} (\sum_{j=\tau+1}^{\infty} a_j u_{t-j})$ . For the first component,

$$\begin{aligned} \mathbf{E} \left\| \frac{1}{T} \sum_t \mathbf{U}_{\tau t} \varepsilon_t \right\|^2 &= \frac{1}{T^2} \sum_{i=1}^{\tau} \mathbf{E} \left[ \sum_t u_{t-i} \varepsilon_t \right]^2 \\ &= \frac{\tau}{T} \gamma_u(0) \sigma_{\varepsilon}^2 = O\left(\frac{\tau}{T}\right); \end{aligned} \tag{A.3}$$

thus  $\frac{1}{T} \sum_t \mathbf{U}_{\tau t} \varepsilon_t$  is of order  $O_p(T^{-1/2} \tau^{1/2})$ . For the second component, note that  $u_t$  is a stationary invertible process whose linear process coefficients satisfy the given summability assumption,

$$\begin{aligned} \mathbf{E} \left\| \frac{1}{T} \sum_t \mathbf{U}_{\tau t} \left( \sum_{j=\tau+1}^{\infty} a_j u_{t-j} \right) \right\|^2 \\ = \frac{1}{T^2} \sum_{i=1}^{\tau} \mathbf{E} \left[ \sum_{j=\tau+1}^{\infty} \sum_{l=\tau+1}^{\infty} a_j a_l \sum_t \sum_s u_{t-i} u_{t-j} u_{s-i} u_{s-l} \right]. \end{aligned}$$

Using the linear process representation of  $u_t$ , we obtain

$$\begin{aligned} \mathbf{E} \left[ \sum_{j=\tau+1}^{\infty} \sum_{l=\tau+1}^{\infty} a_j a_l \sum_t \sum_s u_{t-i} u_{t-j} u_{s-i} u_{s-l} \right] \\ = \mathbf{E} \left[ \sum_{j=\tau+1}^{\infty} \sum_{l=\tau+1}^{\infty} a_j a_l \sum_t \sum_s \left( \sum_{r=0}^{\infty} c_r \varepsilon_{t-i-r} \right) \left( \sum_{p=0}^{\infty} c_p \varepsilon_{t-j-p} \right) \right. \\ \left. \times \left( \sum_{g=0}^{\infty} c_g \varepsilon_{s-i-g} \right) \left( \sum_{h=0}^{\infty} c_h \varepsilon_{s-l-h} \right) \right] \\ = \sum_{j=\tau+1}^{\infty} \sum_{l=\tau+1}^{\infty} a_j a_l \sum_t \sum_s \left( \sum_{r=0}^{\infty} \sum_{p=0}^{\infty} \sum_{g=0}^{\infty} \sum_{h=0}^{\infty} c_r c_p c_g c_h \right. \\ \left. \times \mathbf{E}[\varepsilon_{t-i-r} \varepsilon_{t-j-p} \varepsilon_{s-i-g} \varepsilon_{s-l-h}] \right). \end{aligned} \tag{A.4}$$

Note that the  $\varepsilon_t$ 's are iid with mean 0 and the expectation  $\mathbf{E}[\varepsilon_{t-i-r} \times \varepsilon_{t-j-p} \varepsilon_{s-i-g} \varepsilon_{s-l-h}]$  is nonzero when (a)  $s - i - g = s - l - h$  and  $t - i - r = t - j - p$ , (b)  $s - i - g = t - i - r$  and  $t - j - p = s - l - h$ , (c)  $s - i - g = t - j - p$  and  $t - i - r = s - l - h$ , (d)  $s - i - g = s - l - h = t - i - r = t - j - p$ . By the summability condition of  $\{c_i\}_{i=0}^{\infty}$ , direct calculations show that

$$\mathbf{E} \left\| \frac{1}{T} \sum_t \mathbf{U}_{\tau t} \left( \sum_{j=\tau+1}^{\infty} a_j u_{t-j} \right) \right\|^2 = O\left( \tau \left[ \sum_{j=\tau+1}^{\infty} a_j^2 \right] \right).$$

Under assumption 5, there exists some  $0 < \lambda < 1$  such that  $|a_j|$  is bounded by a constant multiple of  $\lambda^j$ , and we have  $\sum_{j=\tau+1}^{\infty} a_j^2 = O(\lambda^{\tau})$ . Thus, under assumption 6 that  $\tau = \kappa \log T$ , with appropriately chosen  $\kappa$  (say,  $\kappa = -\ln \lambda > 0$ ),  $\sum_{j=\tau+1}^{\infty} a_j^2 = O(T)$ . Thus, combining the result of (A.3),

$$\left\| \frac{1}{T} \sum_t \mathbf{U}_{\tau t} \left( \varepsilon_t + \sum_{j=\tau+1}^{\infty} a_j u_{t-j} \right) \right\| = O_p(T^{-1/2} \tau^{1/2}).$$

Given our choice of  $\tau$ , we have, for any small  $\nu > 0$ ,  $\|\bar{\mathbf{A}}_{\tau} - \mathbf{A}_{\tau}\| = o_p(T^{-1/2+\nu})$ . This concludes the first part that  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T2A} = o_p(T^{-1/2} h_1^{-d/2})$ .

Next, we show that  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T2B} = o_p(T^{-1/2} h_1^{-d/2})$ . We have

$$\begin{aligned} \hat{\mathbf{A}}_{\tau} - \bar{\mathbf{A}}_{\tau} &= \hat{\mathbf{G}}_{\tau}^{-1} \hat{\mathbf{g}}_{\tau} - \mathbf{G}_{\tau}^{-1} \mathbf{g}_{\tau} \\ &= -\mathbf{G}_{\tau}^{-1} [\hat{\mathbf{G}}_{\tau} - \mathbf{G}_{\tau}] \mathbf{G}_{\tau}^{-1} \mathbf{g}_{\tau} + \mathbf{G}_{\tau}^{-1} [\hat{\mathbf{g}}_{\tau} - \mathbf{g}_{\tau}] \\ &\quad + \hat{\mathbf{G}}_{\tau}^{-1} [\hat{\mathbf{G}}_{\tau} - \mathbf{G}_{\tau}] \mathbf{G}_{\tau}^{-1} [\hat{\mathbf{G}}_{\tau} - \mathbf{G}_{\tau}] \mathbf{G}_{\tau}^{-1} \mathbf{g}_{\tau} \\ &\quad - \hat{\mathbf{G}}_{\tau}^{-1} [\hat{\mathbf{G}}_{\tau} - \mathbf{G}_{\tau}] \mathbf{G}_{\tau}^{-1} [\hat{\mathbf{g}}_{\tau} - \mathbf{g}_{\tau}], \end{aligned}$$

where

$$\hat{\mathbf{G}}_{\tau} = \frac{1}{T} \hat{\mathbf{U}}'_{\tau} \hat{\mathbf{U}}_{\tau} = \left( \frac{1}{T} \sum_{t=\tau+1}^T \hat{u}_{t-j} \hat{u}_{t-l} \right)_{j,l},$$

$$\widehat{\mathbf{g}}_\tau = \frac{1}{T} \widehat{\mathbf{U}}'_\tau \widehat{\mathbf{u}} = \left( \frac{1}{T} \sum_{t=\tau+1}^T \widehat{u}_{t-j} \widehat{u}_t \right)_j,$$

and

$$\mathbf{g}_\tau = \frac{1}{T} \mathbf{U}'_\tau \mathbf{u} = \left( \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} u_t \right)_j.$$

Further, define the  $\tau \times 1$  vector  $\boldsymbol{\gamma}_\tau = \frac{1}{T} \mathbf{E}(\mathbf{U}'_\tau \mathbf{u}) = (\mathbf{E}(u_{t-j} u_t))_j$ . Then

$$\|\mathbf{g}_\tau - \boldsymbol{\gamma}_\tau\| = O_p(Q_T). \tag{A.5}$$

Note that

$$(\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau)_{j,l} = \frac{1}{T} \sum_{t=\tau+1}^T (\widehat{u}_{t-j} \widehat{u}_{t-l} - u_{t-j} u_{t-l})$$

and

$$(\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau)_j = \frac{1}{T} \sum_{t=\tau+1}^T (\widehat{u}_{t-j} \widehat{u}_t - u_{t-j} u_t).$$

Now write  $\widehat{u}_t = u_t - \widehat{V}_t - \widehat{B}_t$ , where

$$\widehat{B}_t = \mathbf{e}'_1 \mathbf{M}_T^{*-1}(X_t) \mathbf{B}_T^*(X_t), \quad \widehat{V}_t = \mathbf{e}'_1 \mathbf{M}_T^{*-1}(X_t) \mathbf{U}_T^*(X_t) \tag{A.6}$$

and  $\mathbf{M}_T^*$ ,  $\mathbf{B}_T^*$  and  $\mathbf{U}_T^*$  are defined as  $\mathbf{M}_T$ ,  $\mathbf{B}_T$  and  $\mathbf{U}_T$  but with kernel  $K_0$  and bandwidth  $h_0$ . Then

$$\begin{aligned} & \widehat{u}_{t-j} \widehat{u}_{t-l} - u_{t-j} u_{t-l} \\ &= -u_{t-j} \widehat{V}_{t-l} - u_{t-j} \widehat{B}_{t-l} - u_{t-l} \widehat{V}_{t-j} - u_{t-l} \widehat{B}_{t-j} \\ & \quad + \widehat{V}_{t-l} \widehat{V}_{t-j} + \widehat{B}_{t-j} \widehat{B}_{t-l} + \widehat{V}_{t-l} \widehat{B}_{t-j} + \widehat{B}_{t-j} \widehat{V}_{t-l}. \end{aligned}$$

Clearly,

$$\begin{aligned} & \left| \frac{1}{T} \sum_{t=\tau+1}^T (\widehat{V}_{t-l} \widehat{V}_{t-j} + \widehat{B}_{t-j} \widehat{B}_{t-l} + \widehat{V}_{t-l} \widehat{B}_{t-j} + \widehat{B}_{t-j} \widehat{V}_{t-l}) \right| \\ & \leq \frac{1}{T} \sum_{t=\tau+1}^T \left( \left( \sup_s |\widehat{V}_s| \right)^2 + \left( \sup_s |\widehat{B}_s| \right)^2 + 2 \sup_s |\widehat{V}_s| \sup_s |\widehat{B}_s| \right) \\ & = O_p((\log T) T^{-1} h_0^{-d} + h_0^{2q}), \tag{A.7} \end{aligned}$$

by virtue of the uniform rate of convergence of the terms  $\widehat{V}_s$  and  $\widehat{B}_s$  over  $s$ .

The cross-product terms require more detailed analysis. Note that

$$\begin{aligned} & \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} \widehat{V}_{t-l} \\ & \simeq \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} [\mathbf{e}'_1 [\mathbf{M}(K_0) f_X(X_{t-l})]^{-1} \mathbf{U}_T^*(X_{t-l})], \end{aligned}$$

where  $\mathbf{M}(K_0)$  is defined as  $\mathbf{M}$  but with kernel  $K_0$ . Letting  $[\dots, \omega_0^{0,\kappa}, \dots]$  be the first row of  $\mathbf{M}(K_0)^{-1}$ , we have

$$\begin{aligned} & \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} [\mathbf{e}'_1 [\mathbf{M}(K_0) f_X(X_{t-l})]^{-1} \mathbf{U}_T^*(X_{t-l})] \\ &= \sum_{\kappa} \omega_0^{0,\kappa} \frac{1}{T} \sum_{t=\tau+1}^T \sum_{r=1}^T \frac{1}{T h_0^d} f_X(X_{t-l})^{-1} \left( \frac{X_{t-l} - X_r}{h_0} \right)^\kappa \\ & \quad \times K_0 \left( \frac{X_{t-l} - X_r}{h_0} \right) u_{t-j} u_r, \end{aligned}$$

where the sum over  $\kappa$  is over a finite index set. Note that  $u_r$  has linear process representation  $u_t = \sum_{j=0}^\infty c_j \varepsilon_{t-j}$ . Denoting

$$\frac{1}{T h_0^d} f_X(X_{t-l})^{-1} \left( \frac{X_{t-l} - X_r}{h_0} \right)^\kappa K_0 \left( \frac{X_{t-l} - X_r}{h_0} \right)$$

as  $w_{\kappa,t-l,r}$ , we have

$$\begin{aligned} & \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} [\mathbf{e}'_1 [\mathbf{M}(K_0) f_X(X_{t-l})]^{-1} \mathbf{U}_T^*(X_{t-l})] \\ &= \sum_{\kappa} \omega_0^{0,\kappa} \varphi_{\kappa,T,j,l}, \end{aligned}$$

where

$$\varphi_{\kappa,T,j,l} = \frac{1}{T} \sum_{t=\tau+1}^T \sum_{r=1}^T w_{\kappa,t-l,r} \left( \sum_{s=0}^\infty c_s \varepsilon_{t-j-s} \right) \left( \sum_{b=0}^\infty c_b \varepsilon_{r-b} \right).$$

In addition, notice that  $X$  and  $\varepsilon$  are independent; thus

$$\begin{aligned} & E|\varphi_{\kappa,T,j,l}|^2 \\ &= \frac{1}{T^2} \sum_{a=0}^\infty \sum_{b=0}^\infty \sum_{g=0}^\infty \sum_{s=0}^\infty \sum_{t=\tau+1}^T \sum_{p=\tau+1}^T \sum_{r=1}^T \sum_{h=1}^T c_a c_b c_g c_s \\ & \quad \times E(w_{\kappa,t-l,r} w_{\kappa,p-l,h}) \\ & \quad \times E(\varepsilon_{t-j-s} \varepsilon_{p-j-g} \varepsilon_{r-b} \varepsilon_{h-a}). \end{aligned}$$

Because the  $\varepsilon$ 's are iid, the foregoing expectation is nonzero when (a)  $r - b = h - a$  and  $t - s = p - g$ , (b)  $r - b = t - j - s$  and  $h - a = p - j - g$ , (c)  $r - b = p - j - g$  and  $h - a = t - j - s$ , (d)  $h - a = r - b = t - j - s = p - j - g$ . Simple calculations show that

$$\begin{aligned} \varphi_{\kappa,T,j,l} &= \frac{1}{T} \sum_{t=\tau+1}^T \sum_{r=1}^T w_{\kappa,t-l,r} \left( \sum_{s=0}^\infty c_s \varepsilon_{t-j-s} \right) \left( \sum_{b=0}^\infty c_b \varepsilon_{r-b} \right) \\ &= O_p\left(\frac{1}{T}\right). \tag{A.8} \end{aligned}$$

For the term with bias effects,

$$\begin{aligned} & \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} \widehat{B}_{t-l} \\ & \simeq \sum_{\kappa} \omega_0^{0,\kappa} \frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} \left( \frac{1}{T h_0^d} \sum_{s \neq t-l} f_X(X_{t-l})^{-1} K_0 \right. \\ & \quad \left. \times \left( \frac{X_{t-l} - X_s}{h_0} \right) \left( \frac{X_{t-l} - X_s}{h_0} \right)^{q+\kappa-1} h^q m^{(q)}(X_{t-l}) \right). \end{aligned}$$

By verification of moments, we show that  $\frac{1}{T} \sum_{t=\tau+1}^T u_{t-j} \widehat{B}_{t-l} = O_p(h^q)$ . Therefore, we have

$$\|\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau\| = O_p(T^{-1/2} h_0^q + (\log T) T^{-1} h_0^{-d} + h_0^{2q}) \tau. \tag{A.9}$$

Similarly, we have

$$\|\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau\| = O_p(T^{-1/2} h_0^q + (\log T) T^{-1} h_0^{-d} + h_0^{2q}) \tau. \tag{A.10}$$

Note that

$$\begin{aligned} \widehat{\mathbf{A}}_\tau - \overline{\mathbf{A}}_\tau &= -\mathbf{G}_\tau^{-1} [\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau] \mathbf{G}_\tau^{-1} \mathbf{g}_\tau + \mathbf{G}_\tau^{-1} [\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau] \\ & \quad + \widehat{\mathbf{G}}_\tau^{-1} [\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau] \mathbf{G}_\tau^{-1} [\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau] \mathbf{G}_\tau^{-1} \mathbf{g}_\tau \\ & \quad - \widehat{\mathbf{G}}_\tau^{-1} [\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau] \mathbf{G}_\tau^{-1} [\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau]. \end{aligned}$$

Furthermore, we can substitute  $\Gamma_\tau^{-1}$  and  $\gamma_\tau$  for  $\mathbf{G}_\tau^{-1}$  and  $\mathbf{g}_\tau$ . Using (A.9), (A.10), (A.2) and (A.5), we obtain

$$\begin{aligned} \|\widehat{\mathbf{A}}_\tau - \overline{\mathbf{A}}_\tau + \Gamma_\tau^{-1}[\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau]\Gamma_\tau^{-1}\gamma_\tau - \Gamma_\tau^{-1}[\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau]\| \\ = O_p(\Delta_n^2), \quad (\text{A.11}) \end{aligned}$$

where  $\Delta_n = ((\log T)T^{-1}h_0^{-d} + h_0^{2q})\tau$ .

We can then write each element of  $\widetilde{\Psi}_{T2B}$  as

$$\begin{aligned} \frac{1}{Th_1^d} \sum_{t=1}^T \left(\frac{x-X_t}{h_1}\right)^r K_1\left(\frac{x-X_t}{h_1}\right) \left(\sum_{j=1}^\tau (\widehat{a}_j - \bar{a}_j)u_{t-j}\right) \\ = \frac{1}{Th_1^d} \sum_{t=1}^T \left(\frac{x-X_t}{h_1}\right)^r K_1\left(\frac{x-X_t}{h_1}\right) \mathbf{U}'_{\tau t} \\ \times [\Gamma_\tau^{-1}[\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau]\Gamma_\tau^{-1}\gamma_\tau - \Gamma_\tau^{-1}[\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau]]. \end{aligned}$$

Note that

$$\begin{aligned} \left\| \frac{1}{Th_1^d} \sum_{t=1}^T \left(\frac{x-X_t}{h_1}\right)^r K_1\left(\frac{x-X_t}{h_1}\right) \mathbf{U}'_{\tau t} \right. \\ \left. \times [\Gamma_\tau^{-1}[\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau]\Gamma_\tau^{-1}\gamma_\tau - \Gamma_\tau^{-1}[\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau]] \right\| \\ \leq \frac{1}{T} \sum_{t=1}^T \left\| \frac{1}{h_1^d} \left(\frac{x-X_t}{h_1}\right)^r K_1\left(\frac{x-X_t}{h_1}\right) \right\| \|\mathbf{U}'_{\tau t}\| \\ \times [\|\Gamma_\tau^{-1}\|\|\widehat{\mathbf{G}}_\tau - \mathbf{G}_\tau\|\|\Gamma_\tau^{-1}\gamma_\tau\| + \|\Gamma_\tau^{-1}\|\|\widehat{\mathbf{g}}_\tau - \mathbf{g}_\tau\|]. \end{aligned}$$

Thus  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \widetilde{\Psi}_{T2B}$  is of order  $O_p((\log T)T^{-1}h_0^{-d} + h_0^{2q})\tau^c$ , where  $c$  is a constant. Under assumption 6,  $\mathbf{e}'_1 \mathbf{M}_T^{-1} \widetilde{\Psi}_{T2B}$  is  $o_p(T^{-1/2} \times h_1^{-d/2})$ , which completes the proof.

*Lemma A.3.* Under assumptions 1–7,

$$Q_{T3} = O_p(h_0^q) + o_p(T^{-1/2}h_1^{-d/2}).$$

*Proof of Lemma A.3.* We substitute  $\widehat{m}(X_t) - m(X_t) = \widehat{V}_t + \widehat{B}_t$  into  $\widetilde{\Psi}_{T3}$ , where  $\widehat{B}_t$  and  $\widehat{V}_t$  are defined as in (A.6), and decompose  $\widetilde{\Psi}_{T3}$  into  $\widetilde{\Psi}_{T3V} + \widetilde{\Psi}_{T3B}$ , where

$$\begin{aligned} [\widetilde{\Psi}_{T3V,|j|}]_r \\ = \frac{1}{Th_1^d} \sum_{t=1}^T \left(\frac{x-X_t}{h_1}\right)^{\phi_{|j|}(r)} K_1\left(\frac{x-X_t}{h_1}\right) \left(\sum_{j=1}^\tau a_j \widehat{V}_{t-j}\right), \end{aligned}$$

and

$$\begin{aligned} [\widetilde{\Psi}_{T3B,|j|}]_r \\ = \frac{1}{Th_1^d} \sum_{t=1}^T \left(\frac{x-X_t}{h_1}\right)^{\phi_{|j|}(r)} K_1\left(\frac{x-X_t}{h_1}\right) \left(\sum_{j=1}^\tau a_j \widehat{B}_{t-j}\right). \end{aligned}$$

We have

$$\begin{aligned} Q_{T3} \simeq \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} \widetilde{\Psi}_{T3V} + \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} \widetilde{\Psi}_{T3B} \\ + \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} (\mathbf{M}_T(x) - \mathbf{M}f_X(x)) \widetilde{\Psi}_{T3V} \\ + \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} (\mathbf{M}_T(x) - \mathbf{M}f_X(x)) \widetilde{\Psi}_{T3B}. \end{aligned}$$

We start with the first (“variance”) term, which can be written as

$$\begin{aligned} \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} \widetilde{\Psi}_{T3V} \\ = \sum_\nu \omega^{0,\nu} \sum_\kappa \omega_0^{0,\kappa} \frac{1}{T} \sum_{r=1}^T u_r \sum_{j=1}^\tau a_j w_{\kappa,\nu,T,j,r}, \end{aligned}$$

where

$$\begin{aligned} w_{\kappa,\nu,T,j,r} = \frac{1}{Th_1^d h_0^d} \sum_{t=1}^T \frac{1}{f_X(X_{t-j})f_X(x)} K_1 \\ \times \left(\frac{x-X_t}{h_1}\right) K_0\left(\frac{X_{t-j}-X_r}{h_0}\right) \\ \times \left(\frac{x-X_t}{h_1}\right)^\nu \left(\frac{X_{t-j}-X_r}{h_0}\right)^\kappa, \end{aligned}$$

$\omega^{0,\nu}$  and  $\omega_0^{0,\kappa}$  are elements in the first row of  $\mathbf{M}^{-1}$  and  $\mathbf{M}(K_0)^{-1}$ , and the sum over  $\nu$  and  $\kappa$  are over finite index sets. We need to verify the boundedness of  $E(|w_{\kappa,\nu,T,j,r}|)$ ,

$$\begin{aligned} E|w_{\kappa,\nu,T,j,r}| = \frac{1}{T} \sum_{t=1}^T \int \frac{f_{X,j,t-r}(x-uh_1, z, z-vh_0)}{f_X(x)f_X(z)} \\ \times K_1(u)K_0(v)u^\nu v^\kappa du dz dv. \end{aligned}$$

Again, under assumption 2, that the densities are bounded,  $E|w_{\kappa,\nu,T,j,r}|$  is uniformly bounded over all  $j$  and  $r$ . Because  $w_{\kappa,\nu,T,j,r}$  depends only on  $X_1, \dots, X_T$ , and  $u$  and  $X$  are mutually independent, we have

$$\begin{aligned} \text{var} \left[ \frac{1}{T} \sum_{r=1}^T u_r \sum_{j=1}^\tau a_j w_{\kappa,\nu,T,j,r} \right] \\ \leq \frac{1}{T} \left( \gamma_u(0) + 2 \sum_{j=1}^\infty \gamma_u(j) \right) \left( \sum_{j=1}^\infty |a_j| \right)^2 \left( \sup_{j,r,T} E(|w_{\kappa,\nu,T,j,r}|) \right)^2 \\ = O(T^{-1}) \end{aligned}$$

by the summability of  $a_j$  and  $\gamma_u(j)$  and the boundedness of  $E(|w_{\kappa,\nu,T,j,r}|)$ . Thus the first term  $\mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} \widetilde{\Psi}_{T3V} = O_p(T^{-1/2})$ .

We now turn to the leading bias term. Note that

$$\begin{aligned} \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} \widetilde{\Psi}_{T3B} \\ = \sum_\nu \omega^{0,\nu} \sum_\kappa \omega_0^{0,\kappa} \frac{h^q}{Th_1^d} \frac{1}{f_X(x)} \sum_{t=1}^T \left(\frac{x-X_t}{h_1}\right)^\nu K_1\left(\frac{x-X_t}{h_1}\right) \\ \times \sum_{j=1}^\tau a_j \left[ \frac{1}{Th_0^d} \sum_s K_0\left(\frac{X_{t-j}-X_s}{h_0}\right) \right. \\ \left. \times \left(\frac{X_{t-j}-X_s}{h_0}\right)^{q+\kappa-1} \frac{m^{(q)}(X_{t-j})}{f_X(X_{t-j})} \right], \end{aligned}$$

conditional on  $X_{t-j}$ , for each  $\kappa$  and  $\nu$ ,

$$\begin{aligned} \frac{1}{T} \sum_s \frac{1}{h_0^d} K_0\left(\frac{X_{t-j}-X_s}{h_0}\right) \left(\frac{X_{t-j}-X_s}{h_0}\right)^{q+\kappa-1} \frac{m^{(q)}(X_{t-j})}{f_X(X_{t-j})} \\ \simeq m^{(q)}(X_{t-j}) \frac{f_{X,t-j-s}(X_{t-j})}{f_X(X_{t-j})} \int K_0(u)u^{q+\kappa-1} du, \end{aligned}$$

and

$$\begin{aligned} \frac{1}{Th_1^d} \sum_{t=1}^T \frac{1}{f_X(x)} \left(\frac{x-X_t}{h_1}\right)^\nu K_1\left(\frac{x-X_t}{h_1}\right) \\ \times m^{(q)}(X_{t-j}) \frac{f_{X,t-j-s}(X_{t-j})}{f_X(X_{t-j})} \\ \simeq E \left[ \frac{1}{h_1^d} \frac{1}{f_X(x)} \left(\frac{x-X_t}{h_1}\right)^\nu K_1\left(\frac{x-X_t}{h_1}\right) \right. \\ \left. \times m^{(q)}(X_{t-j}) \frac{f_{X,t-j-s}(X_{t-j})}{f_X(X_{t-j})} \right], \end{aligned}$$

Thus

$$\begin{aligned} & \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} \tilde{\Psi}_{T3B} \\ &= h^q \sum_v \omega^{0,v} \sum_\kappa \omega_0^{0,\kappa} \mu_{q+\kappa-1}(K_0) \sum_{j=1}^\tau a_j \\ & \times \mathbb{E} \left[ \frac{1}{h_1^d} \frac{1}{f_X(x)} \left( \frac{x-X_t}{h_1} \right)^v K_1 \left( \frac{x-X_t}{h_1} \right) \right. \\ & \quad \left. \times m^{(q)}(X_{t-j}) \frac{f_{X,t-j-s}(X_{t-j})}{f_X(X_{t-j})} \right] \\ &= O_p(h_0^q), \end{aligned}$$

because  $\sum_{j=1}^\infty |a_j| < \infty$  and

$$\begin{aligned} & \mathbb{E} \left[ \frac{1}{h_1^d} \frac{1}{f_X(x)} \left( \frac{x-X_t}{h_1} \right)^v K_1 \left( \frac{x-X_t}{h_1} \right) \right. \\ & \quad \left. \times m^{(q)}(X_{t-j}) \frac{f_{X,t-j-s}(X_{t-j})}{f_X(X_{t-j})} \right] = O(1). \end{aligned}$$

Finally, we turn to the remainder terms

$$\mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} (\mathbf{M}_T(x) - \mathbf{M}f_X(x)) \tilde{\Psi}_{T3V}$$

and

$$\mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} (\mathbf{M}_T(x) - \mathbf{M}f_X(x)) \tilde{\Psi}_{T3B}.$$

Note that

$$\sup_{x \in \mathcal{X}} |\mathbf{M}_T(x) - f(x)\mathbf{M}| = O_p(h_1 + T^{-1/2}h_1^{-d/2} \log T), \tag{A.12}$$

$$\sup_t |\widehat{V}_t| = O_p(T^{-1/2}h_0^{-d/2}(\log T)^{1/2}), \tag{A.13}$$

$$\sup_t |\widehat{B}_t| = O_p(h_0^q), \tag{A.14}$$

and

$$\begin{aligned} & \frac{1}{Th_1^d} \sum_{i=1}^T \left| \left( \frac{x-X_i}{h_1} \right)^\kappa K_1 \left( \frac{x-X_i}{h_1} \right) \right| \\ & \rightarrow \mathbb{E} \left| \frac{1}{h_1^d} \left( \frac{x-X_1}{h_1} \right)^\kappa K_1 \left( \frac{x-X_1}{h_1} \right) \right| \\ & = \int |K_1(u)| |u|^\kappa |f_X(x-uh_1)| du. \end{aligned}$$

Under our assumptions,  $f_X(x)$  is bounded away from 0, and we have

$$\begin{aligned} & \left\| \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} (\mathbf{M}_T(x) - \mathbf{M}f_X(x)) \tilde{\Psi}_{T3V} \right\| \\ & \leq \|\mathbf{e}'_1 \mathbf{M}^{-1}\| \sup_{x \in \mathcal{X}} \|(\mathbf{M}_T(x) - \mathbf{M}f_X(x))\| \sup_t |\widehat{V}_t| \left( \sum_{j=1}^\infty |a_j| \right) \\ & \quad \times \sum_\kappa \frac{1}{Th_1^d} \sum_{i=1}^T \frac{1}{f_X(x)} \left| \left( \frac{x-X_i}{h_1} \right)^\kappa K_1 \left( \frac{x-X_i}{h_1} \right) \right| \\ & = O_p(h_1 + T^{-1/2}h_1^{-d/2} \log T) O_p(T^{-1/2}h_0^{-d/2}(\log T)^{1/2}) \end{aligned}$$

and

$$\begin{aligned} & \mathbf{e}'_1 [\mathbf{M}f_X(x)]^{-1} (\mathbf{M}_T(x) - \mathbf{M}f_X(x)) \tilde{\Psi}_{T3B} \\ & \leq \|\mathbf{e}'_1 \mathbf{M}^{-1}\| \sup_{x \in \mathcal{X}} \|(\mathbf{M}_T(x) - \mathbf{M}f_X(x))\| \sup_t |\widehat{B}_t| \left( \sum_{j=1}^\infty |a_j| \right) \end{aligned}$$

$$\begin{aligned} & \times \sum_\kappa \frac{1}{Th_1^d} \sum_{i=1}^T \frac{1}{f_X(x)} \left| \left( \frac{x-X_i}{h_1} \right)^\kappa K_1 \left( \frac{x-X_i}{h_1} \right) \right| \\ & = O_p(h_1 + T^{-1/2}h_1^{-d/2} \log T) O_p(h_0^q). \end{aligned}$$

Lemma A.4. Under assumptions 1–7,

$$Q_{T4} = o_p(T^{-1/2}h_1^{-d/2}).$$

Proof of Lemma A.4. By definition,  $Q_{T4} = \mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T4}$ , where

$$\begin{aligned} [\tilde{\Psi}_{T4,|j|}]_r &= \frac{1}{Th_1^d} \sum_{i=1}^T \left( \frac{x-X_i}{h_1} \right)^{\phi_{|j|}(r)} K_1 \left( \frac{x-X_i}{h_1} \right) \\ & \quad \times \left( \sum_{j=1}^\tau (\widehat{a}_j - a_j) (\widehat{m}(X_{t-j}) - m(X_{t-j})) \right). \end{aligned}$$

We have

$$\begin{aligned} |Q_{T4}| &= \|\mathbf{e}'_1 \mathbf{M}_T^{-1} \tilde{\Psi}_{T4}\| \\ & \leq \frac{1}{f_X(x)} \|\mathbf{e}'_1 \mathbf{M}^{-1}\| \frac{1}{Th_1^d} \sum_{i=1}^T \left| \left( \frac{x-X_i}{h_1} \right)^\kappa K_1 \left( \frac{x-X_i}{h_1} \right) \right| \\ & \quad \times \|\widehat{\mathbf{A}}_\tau - \mathbf{A}_\tau\| \left[ \sum_{j=1}^\tau (\widehat{m}(X_{t-j}) - m(X_{t-j}))^2 \right]^{1/2} \\ & \leq \frac{1}{f_X(x)} \|\mathbf{e}'_1 \mathbf{M}^{-1}\| \frac{1}{Th_1^d} \sum_{i=1}^T \left| \left( \frac{x-X_i}{h_1} \right)^\kappa K_1 \left( \frac{x-X_i}{h_1} \right) \right| \\ & \quad \times \|\widehat{\mathbf{A}}_\tau - \mathbf{A}_\tau\| \cdot \tau \max_s |\widehat{m}(X_s) - m(X_s)|. \end{aligned}$$

Note that  $\|\widehat{\mathbf{A}}_\tau - \mathbf{A}_\tau\| \leq \|\widehat{\mathbf{A}}_\tau - \overline{\mathbf{A}}_\tau\| + \|\overline{\mathbf{A}}_\tau - \mathbf{A}_\tau\|$ , and, from the proof of Lemma A.2, we have  $\|\widehat{\mathbf{A}}_\tau - \overline{\mathbf{A}}_\tau\| = o_p((\log T)T^{-1/2}h_0^{-d/2} + h_0^q)$  and  $\|\overline{\mathbf{A}}_\tau - \mathbf{A}_\tau\| = O_p(T^{-1/2}\tau^{3/2})$ . In addition,

$$\max_s |\widehat{m}(X_s) - m(X_s)| = O_p(h_0^q + T^{-1/2}h_0^{-d/2}(\log T)^{1/2});$$

thus  $|Q_{T4}| = o_p(T^{-1/2}h_1^{-d/2})$ .

[Received June 2002. Revised August 2003.]

## REFERENCES

Adersen, T. W. (1971), *The Statistical Analysis of Time Series*, New York: Wiley.

Carroll, R. J., Lin, X., Linton, O. B., and Mammen, E. (2004), "Accounting for Correlation in Marginal Longitudinal Nonparametric Regression," in *Second Seattle Symposium on Biostatistics*, ed. D. Lin, forthcoming.

Chen, X., and Linton, O. B. (2001), "An Alternative Way of Computing Efficient Semiparametric Instrumental Variables Estimators," working paper available at <http://www.econ.lse.ac.uk/folinton>.

Conley, T. G., Hansen, L. P., Luttmer, E. G. J., and Scheinkman, J. A. (1997), "Short-Term Interest Rates as Subordinated Diffusions," *The Review of Financial Studies*, 10, 525–577.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.

Hannan, E. J. (1963), "Regression for Time Series," in *Time Series Analysis*, ed. M. Rosenblatt, New York: Wiley.

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

- Hart, J. D. (1991), "Kernel Regression Estimation With Time Series Errors," *Journal of the Royal Statistical Society*, 53, 173–187.
- Kristensen, D., and Linton, O. (2001), "An Alternative GLS-Like Transformation in Regression Models With AR(1) Errors," *Econometric Theory*, 17, 853.
- Lin, X., and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor Is Measured Without/With Error," *Journal of the American Statistical Association*, 95, 520–534.
- Masry, E. (1996a), "Multivariate Regression Estimation: Local Polynomial Fitting for Time Series," *Stochastic Processes and Their Applications*, 65, 81–101.
- (1996b), "Multivariate Local Polynomial regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599.
- Müller, H., and Stadtmüller, U. (1988), "Detecting Dependencies in Smooth Regression Models," *Biometrika*, 75, 639–650.
- Opsomer, J., Wang, Y., and Yang, Y. (2001), "Nonparametric Regression With Correlated Errors," *Statistical Science*, 16, 134–153.
- Phillips, P. C. B., and Solo, V. (1992), "Asymptotics for Linear Processes," *The Annals of Statistics*, 20, 971–1001.
- Robinson, P. M. (1983), "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis*, 4, 185–207.
- Rosenblatt, M. (1956), "A Central Limit Theorem and Strong Mixing Conditions," *Proceedings of the National Academy of Science*, 4, 43–47.
- Ruckstuhl, A., Welsh, A. H., and Carroll, R. J. (2000), "Nonparametric Function Estimation of the Relationship Between Two Repeatedly Measured Variables," *Statistica Sinica*, 10, 51–71.
- Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.
- Vilar-Fernandez, J. M., and Francisco-Fernandez, M. (2002), "Local Polynomial Regression Smoothers With AR-Error Structure," *TEST*, 11, 143–165.
- Wild, C. J., and Yee, T. W. (1996), "Additive Extensions to Generalized Estimating Equation Methods," *Journal of the Royal Statistical Society, Ser. B*, 58, 711–725.
- Wu, C. O., Chiang, C. T., and 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.
- Xiao, Z., Linton, O. B., Carroll, R. J., and Mammen, E. (2003), "More Efficient Local Polynomial Estimation in Nonparametric Regression With Autocorrelated Errors," working paper, available at <http://www.econ.lse.ac.uk/~olinton>.
- Zeger, S. L., and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.