2-2006

More efficient estimation in nonparametric regression with nonparametric autocorrelated errors

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

Aman ULLAH
University of California, Riverside

Citation

SU, Liangjun and ULLAH, Aman. More efficient estimation in nonparametric regression with nonparametric autocorrelated errors. (2006). *Econometric Theory*. 22, (1), 98-126.
Available at: https://ink.library.smu.edu.sg/soe_research/1998

Published in Econometric Theory, Volume 22, Issue 1, February 2006, Pages 227-231
https://doi.org/10.1017/S026646660606004X

MORE EFFICIENT ESTIMATION IN NONPARAMETRIC REGRESSION WITH NONPARAMETRIC AUTOCORRELATED ERRORS

Liangjun Su
Guanghua School of Management, Peking University

Aman Ullah
University of California, Riverside

We define a three-step procedure for more efficient estimation of the nonparametric regression mean with nonparametric autocorrelated errors. The procedure is based upon a nonparametric prewhitening transformation of the dependent variable that has to be estimated from the data by a local polynomial technique. We establish the asymptotic distribution of our estimator under weak dependence conditions and show that it is more efficient than the conventional local polynomial estimator. Furthermore, we consider criterion functions based on the linear exponential family, which include the local polynomial least squares criterion as a special case. Simulation evidence suggests that significant gains can be achieved in finite samples with our approach.

1. INTRODUCTION

Consider the following regression model:

$$Y_t = m_1(X_t) + U_t, \qquad t = 1,\dots,n, \tag{1.1}$$

where the d_1-dimension process $\{X_t\}$ is a stationary process, the stationary scalar error process $\{U_t\}$ is autocorrelated but satisfies $E(U_t|X_t, X_{t-1},\dots) = 0$ almost surely, and $U_t = m_2(U_{t-1},\dots,U_{t-d_2}) + \varepsilon_t$. For the moment, one can assume that $\{\varepsilon_t\}$ is independent and identically distributed (i.i.d.) with mean zero and variance σ_ε^2 and $E(\varepsilon_t|U_{t-1}, U_{t-2},\dots, X_t, X_{t-1},\dots) = 0$. The functions $m_1(\cdot)$ and $m_2(\cdot)$ are assumed to be unknown but smooth. We are interested in estimating

The authors thank Oliver Linton for his many constructive and helpful suggestions. The very insightful comments from the referees are also gratefully acknowledged. The second author gratefully acknowledges financial support from the Academic Senate, UCR. Address correspondence to Liangjun Su, Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing 100871, P.R. China; e-mail: lsu@gsm.pku.edu.cn. Aman Ullah, Department of Economics, UCR, Riverside, CA 92521-0427, USA; e-mail: aman.ullah@ucr.edu.

the infinite-dimensional parameter $m_1(\cdot)$ more efficiently than some conventional approaches.

When the regression function $m_1(\cdot)$ is parametrically specified, it is standard to use generalized least squares (GLS) that reflects the autocorrelation structure in the error process to improve the efficiency of the least squares estimators. When $m_1(\cdot)$ is nonparametrically specified, Xiao, Linton, Carroll, and Mammen (2003) showed that the autocorrelation function of the error process can help improve estimators of the regression function. They assumed that the error process U_t is stationary and mean zero and has an invertible linear process representation: $U_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}$, where $\{\varepsilon_t\}$ are i.i.d. $(0, \sigma_\varepsilon^2)$. This assumption is fairly general because by the Wold decomposition theorem any stationary linear or nonlinear process has a linear (infinite-order) representation. For example, it permits U_t to be any finite-order $\text{ARMA}(p,q)$ process and allows for the full class of linear processes, as is common in much literature on estimating linear regression with linearly autocorrelated errors. In this paper we argue that it may be better, say, in finite-sample applications, to model the error process nonparametrically than to model it as an infinite-order linear process when the error process has a finite-order nonlinear structure: $U_t = m_2(U_{t-1}, \ldots, U_{t-d_2}) + \varepsilon_t$.

One intuitive explanation for this approach is that the factors omitted from the time series regression, like those included, are correlated across periods in an unknown way, which will bring errors with autocorrelation of an unknown form (also see Hong, 1996). As Wooldridge (2003) argued, if interest rates are unexpectedly high for this period, then they are likely to be above average (for the given levels of inflation, deficits, etc.) for the next period. But there is no theory that suggests any parametric form for the correlation structure. The empirical literature is overwhelmingly dominated by the linear $\text{AR}(1)$ model, which in our opinion is mostly a matter of convenience. It is sometimes argued that models can be improved by more elaborate error processes, but the results are often sensitive to the specification chosen (see Greene, 1997, p. 584). To avoid this issue, we can leave the form of m_2 unspecified.

Our second motivation for modeling $m_2(\cdot)$ nonparametrically is due to Hidalgo (1992). As Hidalgo (1992, p. 48) remarked, linear time series models are appropriate in a Gaussian environment, and in the absence of Gaussianity it is most probable that the autocorrelation function is nonlinear. We share the opinion that the Gaussianity assumption in econometrics or statistics is most often for mathematical convenience. Once Gaussianity is abandoned, nonlinear time series are frequently required; and once linearity is abandoned, the form of $m_2(\cdot)$ is hard to identify in view of the many possibilities that exist. Therefore, we leave $m_2(\cdot)$ to be unspecified and shall estimate it nonparametrically.

Third, when $m_1(\cdot)$ is parametric (e.g., $m_1(x) = x'\beta$, where β is a $d_1 \times 1$ parameter vector) but $m_2(\cdot)$ is of our form with $d_2 = 1$, Hidalgo (1992) showed that the finite-dimensional parameter in the regression function can be adaptively estimated. As he argued, by allowing the error process to be of nonlinear feature, the regression model of interest may capture some interesting characteristics such as cycles and jumping behavior that are observed in many time series.

Also, it is worth mentioning that our efficient result does not rely heavily upon the correct specification of $m_2(\cdot)$ (or the correct choice of d_2). We comment in Remark 2 in Section 3 that for any invertible moving average (MA) process or strictly stationary autoregressive (AR) process, our proposed estimator has an efficiency gain over the conventional polynomial estimator.

Nevertheless, due to the "curse of dimensionality," we expect the dimensions d_1 and d_2 to be low unless one imposes further assumptions, say, an additive structure in $m_i(\cdot)$, for $i = 1,2$. In this sense, we say that our result complements that of Xiao et al. (2003). In practice, one can use a nonlinearity test to determine whether the error process is linear or not and then determine which procedure to pursue.

Now we come to the methodology adopted in the paper. Throughout our presentation (with an exception in Sect. 5), we assume that the order d_2 is known. We propose a local-polynomial-based procedure for estimating $m_1(\cdot)$ in the time series regression model (1.1) that takes account of the correlation structure of the error terms. The resulting estimator is asymptotically more efficient than the conventional polynomial estimator. The basic idea is to write (1.1) as

$$Y_t = m_1(X_t) + m_2(U_{t-1},\dots,U_{t-d_2}) + \varepsilon_t, \qquad t = d_2 + 1,\dots,n, \tag{1.2}$$

and then to explore the additive structure in the preceding model. If $(U_{t-1},\dots,U_{t-d_2})$ is observable counterfactually, the model is simply the additive model widely studied in the literature. Linton (1997, 2000) defined a two-step procedure for estimating generalized additive nonparametric regression models that is more efficient than the marginal integration-based method (e.g., Linton and Härdle, 1996). However, because $(U_{t-1},\dots,U_{t-d_2})$ are not observed, we replace them by the residual series $(\hat{U}_{t-1},\dots,\hat{U}_{t-d_2})$ obtained by regressing Y_t on X_t nonparametrically. We show that such a replacement does not affect the first-order asymptotic efficiency of the resulting estimator.

The rest of the paper is organized as follows. We introduce an infeasible local polynomial estimator in Section 2 and a feasible one in Section 3, both of which are more efficient than the conventional one-step local polynomial estimator. In Section 4 we define criterion functions based on the linear exponential family and discuss a class of efficient estimators. We then address some practical issues related to the implementation of our estimators and report some Monte Carlo simulation results in Section 5. All proofs are given in the Appendix.

2. AN INFEASIBLE EFFICIENT ESTIMATOR

Suppose that we have a sample $\{(X_1,Y_1),\dots,(X_n,Y_n)\}$, where $X_t \in \mathbb{R}^{d_1}$ and $Y_t \in \mathbb{R}$, from the nonparametric regression model (1.1). The objective is to estimate $m_1(x)$ at some interior point x more efficiently than some conventional approach and to provide tight confidence intervals for such estimates. Noticing that the error terms in (1.1) are correlated, the conventional nonparametric estimator for $m_1(x)$ does not take into account the correlation structure and thus is supposed to be dominated by some more efficient estimators. Nevertheless, the error terms in the transformed model (1.2) are uncorrelated. Equation (1.2) is also a valid regression equation by assumption. Let $\underline{U}_{t-1} \equiv (U_{t-1},\dots,U_{t-d_2})$. If $m_2(\underline{U}_{t-1})$ were known, a nonparametric regression of $Y_t - m_2(\underline{U}_{t-1})$ on X_t would be more efficient in the sense of mean squared errors than the nonparametric regression of Y_t on X_t. In this paper, we give asymptotic analysis based on the local polynomial procedure. See Fan (1992) and Fan and Gijbels (1996) for discussion on the attractive properties of local polynomials.

For any data set $\{\tilde{Y}_t, X_t\}_{t=1}^n$, the local polynomial regression of $\tilde{Y}_t$ on X_t of order q can be obtained from the multivariate weighted least squares criterion:

$$nh_0^{-d_1} \sum_{t=1}^{n} K_0((x - X_t)/h_0) \left[\tilde{Y}_t - \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}} (X_t - x)^{\mathbf{j}} \right]^2, \tag{2.1}$$

where K_0 is a nonnegative kernel function on $\mathbb{R}^{d_1}$ and $h_0 = h_0(n)$ is a scalar bandwidth sequence. Here, we use the notation of Masry (1996a, 1996b):[1] $\mathbf{j} = (j_1,\dots,j_{d_1})$, $|\mathbf{j}| = \sum_{i=1}^{d_1} j_i$, $x^{\mathbf{j}} = \mathbf{\Pi}_{i=1}^{d_1} x_i^{j_i}$, and $\sum_{0\le|\mathbf{j}|\le q} = \sum_{k=0}^{q} \sum_{\substack{j_1=0 \\ j_1+\dots+j_{d_1}=k}}^{k} \dots \sum_{j_{d_1}=0}^{k}$. The term $\theta_{\mathbf{j}} = \theta_{\mathbf{j}}(x) = (1/\mathbf{j}!)(\partial^{|\mathbf{j}|} m_1(x)/\partial^{j_1} x_1 \dots, \partial^{j_{d_1}} x_{d_1})$, where $\mathbf{j}! \equiv \mathbf{\Pi}_{i=1}^{d_1} j_i!$. Let $\tilde{m}_1(x) = \tilde{\theta}_0$, where $\tilde{\theta}_0$ is the minimizing intercept in (2.1) with $\tilde{Y}_t = Y_t$, and let $\bar{m}_1(x)$ be the corresponding estimator when $\tilde{Y}_t = Y_t - m_2(\underline{U}_{t-1})$. For simplicity, we suppose that $K_0(u) = \Pi_{j=1}^{d_1} k_0(u_j)$, with k_0 being the univariate kernel function. For later use, we denote

$$Q_{n,loc}(\theta) \equiv nh_0^{-d_1} \sum_{t=1}^{n} K_0((x - X_t)/h_0) \left[Y_t - m_2(\underline{U}_{t-1}) - \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}} (X_t - x)^{\mathbf{j}} \right]^2, \tag{2.2}$$

where $\theta = \theta(x)$ is a collection of all the parameters $\theta_{\mathbf{j}}$, $0 \le |\mathbf{j}| \le q$, in a lexicographical order introduced subsequently. In particular, the first element in θ is denoted as $\theta_0 = \theta_0(x)$ throughout our presentation.

Following the notation of Masry (1996a, 1996b), let $N_l = (l + d_1 - 1)!/(l!(d_1 - 1)!)$ be the number of distinct d_1-tuples $\mathbf{j}$ with $|\mathbf{j}| = l$. Arrange the N_l d_1-tuples as a sequence in a lexicographical order (with highest priority to last position so that $(0,0,\dots,l)$ is the first element in the sequence and $(l,0,\dots,0)$ is the last element) and let ϕ_l^{-1} denote this one-to-one map. For each $\mathbf{j}$ with $0 \le |\mathbf{j}| \le 2q$, let $\mu_{\mathbf{j}}(K_0) = \int_{\mathbb{R}^{d_1}} u^{\mathbf{j}} K_0(u)\, du$, $\nu_{\mathbf{j}}(K_0) = \int_{\mathbb{R}^{d_1}} u^{\mathbf{j}} K_0^2(u)\, du$, and define the $N \times N$-dimensional matrices M and Γ and $N \times N_{q+1}$ matrix B, where $N = \sum_{l=1}^{p} N_l$, by

$$M = \begin{bmatrix} M_{0,0} & M_{0,1} & \dots & M_{0,q} \\ M_{1,0} & M_{1,1} & \dots & M_{1,q} \\ \vdots & \vdots & & \vdots \\ M_{q,0} & M_{q,1} & \dots & M_{q,q} \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \dots & \Gamma_{0,q} \\ \Gamma_{1,0} & \Gamma_{1,1} & \dots & \Gamma_{1,q} \\ \vdots & \vdots & & \vdots \\ \Gamma_{q,0} & \Gamma_{q,1} & \dots & \Gamma_{q,q} \end{bmatrix},$$

$$B = \begin{bmatrix} M_{0,q+1} \\ M_{1,q+1} \\ \vdots \\ M_{q,q+1} \end{bmatrix}, \tag{2.3}$$

where $M_{i,j}$ and $\Gamma_{i,j}$ are $N_i \times N_j$-dimensional matrices whose (l,r) elements are, respectively, $\mu_{\phi_i(l)+\phi_j(r)}$ and $\upsilon_{\phi_i(l)+\phi_j(r)}$. Note that the elements of the matrices $M = M(K_0,q)$ and $\Gamma = \Gamma(K_0,q)$ are simply multivariate moments of the kernel K_0 and K_0^2, respectively, and the matrix $B = B(K_0,q)$ depends on the kernel and the order of the local polynomial we use. In addition, we arrange the N_r elements of the derivatives

$$(D^{\mathbf{r}}m_1)(x) \equiv \frac{\partial^{|\mathbf{r}|}m_1(x)}{\partial^{r_1}x_1 \dots, \partial^{r_{d_1}}x_{d_1}}$$

as an $N_r \times 1$ column vector $m_1^{(r)}(x)$ in the lexicographical order.

The following theorem gives the asymptotic distribution of $\overline{m}_1(x)$ and shows that it is asymptotically more efficient than $\widetilde{m}_1(x)$.

THEOREM 2.1. *Suppose that (1.2) holds. Then, under Assumption A in the next section, we have*

$$\sqrt{nh_0^{d_1}}(\overline{m}_1(x) - m_1(x) - h_0^{q+1}[M^{-1}Bm_1^{(q+1)}(x)]_{0,0}) \\ \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2}{f_X(x)}[M^{-1}\Gamma M^{-1}]_{0,0}\right), \tag{2.4}$$

where $[A]_{0,0}$ signifies the upper-left element of matrix A, $M = M(K_0,q)$, $\Gamma = \Gamma(K_0,q)$, and $B = B(K_0,q)$.

Theorem 2.1 can be proved under much weaker assumptions, and it shows that the bias of the estimator $\overline{m}_1(x)$ is the same as that of the conventional qth-order local polynomial estimator $\widetilde{m}_1(x)$ and the variance of $\overline{m}_1(x)$ is smaller than that of $\widetilde{m}_1(x)$.[2] In the case with a local linear estimator, $q = 1$ and the bias term is $\frac{1}{2}h_0^2\int_{\mathbb{R}} u^2k_0(u)\,du \sum_{i=1}^{d_1}(\partial^2 m_1(x)/\partial x_i^2)$. Let $\|K_0\|_2^2 \equiv \int_{\mathbb{R}^{d_1}} K_0^2(u)\,du$. Then $\overline{m}_1(x)$ has a variance $\sigma_\varepsilon^2\|K_0\|_2^2/f_X(x)$, and hence it is more efficient than the traditional local linear estimator $\widetilde{m}_1(x)$, which has variance $\sigma_U^2\|K_0\|_2^2/f_X(x)$, where $\sigma_U^2 = \mathrm{var}(m_2(\underline{U}_{t-1})) + \sigma_\varepsilon^2$. 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 $U_t = \alpha U_{t-1} + \varepsilon_t$, we have $\sigma_\varepsilon^2/\sigma_U^2 = (1 - \alpha^2)$, which is strictly less than one except when $\alpha = 0$. The percentage efficiency gain is less if measured in mean squared errors than in variances, but the two relative efficiencies are monotonically related.[3]

We call $\overline{m}_1(x)$ an "oracle" estimator because its definition uses knowledge that only an oracle could have. A variety of smoothing paradigms could have been chosen here, and each will result in an "oracle" estimate. See Section 4 for more discussions.

3. A FEASIBLE EFFICIENT ESTIMATOR

In practice, $m_2(\cdot)$ is unknown and $\underline{U}_{t-1} = (U_{t-1}, \ldots, U_{t-d_2})'$ is not observed. In this section we propose a feasible estimator by substituting a suitable pilot estimator of $m_2(\cdot)$ in (2.2) and replacing $\underline{U}_{t-1}$ by the residuals obtained from regressing Y_t on X_t only. The proposed three-step estimation procedure is as follows.

1. Obtain a preliminary consistent estimator of m_1 by qth-order local polynomial smoothing Y_t on X_t with corresponding kernel K_1 and bandwidth sequence $h_1 = h_1(n)$. Denote the preliminary estimates as $\hat{m}_1(X_t)$ and calculate the estimated residuals $\hat{U}_t = Y_t - \hat{m}_1(X_t)$ for $t = 1, \ldots, n$.[4]
2. Obtain a consistent estimator of m_2 by pth-order local polynomial smoothing $\hat{U}_t$ on $\hat{\underline{U}}_{t-1} \equiv (\hat{U}_{t-1}, \ldots, \hat{U}_{t-d_2})$ with corresponding kernel K_2 and bandwidth sequence $h_2 = h_2(n)$. Denote the estimates as $\hat{m}_2(\hat{\underline{U}}_{t-1})$ for $t = d_2 + 1, \ldots, n$.
3. Replace $m_2(\underline{U}_{t-1})$ in (2.2) by $\hat{m}_2(\hat{\underline{U}}_{t-1})$ to obtain

$$\hat{Q}_{n,loc}(\theta) \equiv nh_0^{-d_1} \sum_{t=1}^{n} K_0((x - X_t)/h_0) \times \hat{I}_t \left[Y_t - \hat{m}_2(\hat{\underline{U}}_{t-1}) - \sum_{0 \leq |\mathbf{j}| \leq q} \theta_{\mathbf{j}} (X_t - x)^{\mathbf{j}} \right]^2, \tag{3.1}$$

where $\hat{I}_t = 1\{\hat{f}_U(\hat{\underline{U}}_{t-1}) \geq b\}$ for some constant $b = b(n) > 0$ and $\hat{f}_U$ is the nonparametric kernel estimator for the density f_U of $\underline{U}_{t-1}$ with bandwidth h_2 and kernel K_2 based on the residual series $\{\hat{U}_t\}$. Let $\hat{\theta}^* = \hat{\theta}^*(x)$ minimize $\hat{Q}_{n,loc}(\theta)$ and let $\hat{m}_1^*(x) = \hat{\theta}_0^*(x)$ be our feasible estimate of $m_1(x)$.

Like Robinson (1988) and Hidalgo (1992), we use $\hat{I}_t$ to trim out small values of $\hat{f}_U$ to get a desirable result for $\hat{m}_1^*$. When $\underline{U}_{t-1}$ has a compact support, m_2 can be estimated consistently over its full support by the local polynomial regression. So there is no need for trimming in this case.

The preceding procedure may be iterated to achieve better finite-sample performance in practice. We shall show subsequently that, under appropriate assumptions, the proposed estimator $\hat{m}_1^*(x)$ is asymptotically equivalent to the infeasible estimator $\bar{m}_1(x)$. To facilitate the asymptotic analysis, let $W_t = (X_t', \underline{U}_{t-1}')'$. We denote the densities of X_t, $\underline{U}_{t-1}$, and W_t by f_X, $f_{\underline{U}}$, and f_W, respectively. We make the following assumptions on the error terms, regressors, kernel functions, and bandwidth sequences.

Assumption A.

1. The kernels $K_i, i = 0,1,2$, satisfy $K_i(u) = \Pi_{j=1}^{d_i} k_i(u_j)$, where $d_0 \equiv d_1$ and k_i, $i = 0,1,2$, are bounded, symmetric about zero, and of compact support $[-c_i, c_i]$ and satisfy the property that $\int_{\mathbb{R}} k_i(u)\,du = 1$. The functions $H_{i\mathbf{j}}(u) = u^{\mathbf{j}} K_i(u)$ for all $\mathbf{j}$ with $0 \leq |\mathbf{j}| \leq 2p + 1$ for $i = 2$, and $0 \leq |\mathbf{j}| \leq 2q + 1$ for $i = 0$ and 1, are Lipschitz continuous. The first and second partial derivatives $K_2^{(i)}(u)$, $i = 1,2$, of $K_2(u)$ satisfy $\int \| u^{\mathbf{j}} K_2^{(i)}(u) \| du < \infty$ for $0 \leq |\mathbf{j}| \leq 2p$. The functions $\tilde{H}_{\mathbf{j}}(u) \equiv u^{\mathbf{j}} K_2^{(2)}(u)$ for all $\mathbf{j}$ with $0 \leq |\mathbf{j}| \leq 2p$ are Lipschitz continuous.
2. The strictly stationary process $\{(X_t', U_t)'\}$ is strong mixing with mixing coefficients $\alpha(j)$ that satisfy $\sum_{j=1}^{\infty} j^2 \alpha(j)^{\delta/(1+\delta)} < \infty$ for some $\delta > 0$. The term f_X is Lipschitz continuous and bounded away from zero on its compact support. The first- and second-order partial derivatives of $f_{\underline{U}}$ are continuous and bounded on its support. The density f_W of W_t and the joint densities $f_{t_1,\ldots,t_l}(\cdot,\ldots,\cdot)$ of $(V_0, V_{t_1}, \ldots, V_{t_l})$ $(1 \leq l \leq 5)$ are continuous and bounded where $V_t \equiv (X_t', X_{t-1}', \ldots, X_{t-d_2}', \underline{U}_{t-1}', \varepsilon_t)'$.
3. The function $m_1(\cdot)$ is $(q + 1)$ times partially continuously differentiable, and the function $m_2(\cdot)$ is $(p + 1)$ times partially continuously differentiable. The $(q + 1)$th-order partial derivatives of m_1 and the $(p + 1)$th-order partial derivatives of m_2 are Lipschitz continuous on their supports. The first- and second-order partial derivatives of m_2 are bounded on its support.
4. The process $\{U_t\}$ satisfies $U_t = m_2(U_{t-1}, \ldots, U_{t-d_2}) + \varepsilon_t$, where $\{\varepsilon_t\}$ is an i.i.d. process with mean zero and variance σ_ε^2 and $E[\varepsilon_t | U_{t-1}, U_{t-2}, \ldots, X_t, X_{t-1}, \ldots] = 0$ for all t. Here $E|U_t|^{4(1+\delta)} > 0$.
5. The bandwidth sequences $h_{i,}$ $i = 0,1,2$, go to zero as $n \to \infty$ and satisfy (i) $nh_0^{d_1} h_1^{2(q+1)} \to 0$, $nh_0^{d_1} h_2^{2(p+1)} b^{-2} \to 0$; (ii) $nh_0^{-d_1} h_1^{2d_1}/(\ln n)^2 \to \infty$, $nh_0^{-d_1} h_2^{2d_2} b^4/(\ln n)^2 \to \infty$; (iii) $nh_1^{d_1} h_2^2/\ln n \to \infty$, $nh_2^{d_2+2} b^4/\ln n \to \infty$; (iv) $nh_0^{-d_1} h_1^{d_1} h_2^{d_2+2} b^4/(\ln n)^2 \to \infty$, $h_0^{d_1} h_1^{-d_1} h_2^{2p} b^{-2} \ln n \to 0$; (v) $nh_0^{d_1+2(q+1)} \to C \in [0,\infty)$, $h_0^{d_1} h_2^{-d_2+2} b^{-2} \ln n \to 0$.

Assumptions A1–A3 are comparable with Conditions 1–3 in Masry (1996a) except that our assumptions are a bit stronger than his. Unlike Li and Wooldridge (2002), who assume a β-mixing condition, which is handy for applying a result of Yoshihara (1976), we assume a strong mixing condition, which suffices to use the Davydov inequality (Bosq, 1996, p. 19) and one lemma due to Gao and King (2002) for U-statistics. The differentiability of A3 ensures a Taylor expansion to appropriate orders. Although A4 is a high-level assumption, it includes the stationary linear or nonlinear AR processes of finite order. If $\{U_t\}$ is geometrically ergodic with its initial measure equal to its invariant measure, then $\{U_t\}$ is strictly stationary. Primitive conditions on the geometric ergodicity of nonlinear autoregressive processes can be found from Appendix A1.4 of Tong (1990).

Assumption A5 looks complicated and deserves some comment. Let $v_{1n} = n^{-1/2}h_1^{-d_1/2}\sqrt{\ln n}$, $v_{2n} = n^{-1/2}h_2^{-d_2/2}\sqrt{\ln n}$, and $I_t = 1\{f_{\underline{U}}(\underline{U}_{t-1}) > b\}$. Then by Masry (1996b) and Hansen (2004), $\max_{1\le t\le n}|\hat{m}_1(X_t) - m_1(X_t)| = O_p(v_{1n} + h_1^{q+1})$ and $\max_{d_2+1\le t\le n} I_t|\hat{m}_2(\underline{U}_{t-1}) - m_2(\underline{U}_{t-1})| = O_p(b^{-1}v_{2n} + b^{-1}h_2^{p+1})$ if $\{U_1,\dots,U_n\}$ were used in forming $\hat{m}_2(\cdot)$. Assumptions A5(i) and (ii) require that the bias and variance terms in the first and second estimation stage should be $o(n^{-1/2}h_0^{-d_1/2})$. Because $\{U_1,\dots,U_n\}$ is not observed and $\{\hat{U}_1,\dots,\hat{U}_n\}$ is used instead in forming $\hat{m}_2(\cdot)$, there are compound estimation errors associated with the first two-stage estimation. We restrict them to be small in Assumptions A5(iii) and (iv): $h_2^{-1}v_{1n} = o(1)$, $h_2^{-1}b^{-2}v_{2n} = o(1)$, $h_2^{-1}b^{-2}v_{1n}v_{2n} = o(n^{-1/2}h_0^{-d_1/2})$, $b^{-1}h_2^{p}v_{1n} = o(n^{-1/2}h_0^{-d_1/2})$, where the appearance of h_2^{-1} is due to the use of Taylor expansion in our proof. The last part of Assumption A5 will facilitate the proof. Like Hansen (2004), one can set $b \propto (\ln n)^{-1/2}$ and consider the case when $p = q = 1$. If $d_1 = d_2 = 1$, one can set $h_0 \propto n^{-1/5}$, $h_1 \propto n^{-1/4}$, and $h_2 \propto n^{-1/4}$; if $d_1 = 1$ and $d_2 = 2$, one can set $h_0 \propto n^{-1/5}$, $h_1 \propto n^{-1/4}$, and $h_2 \propto n^{-1/5}$; if $d_1 = 2$ and $d_2 = 1$ or 2, one can set $h_0 \propto n^{-1/6}$, $h_1 \propto n^{-1/5}$, and $h_2 \propto n^{-1/4}$ or $h_2 \propto n^{-1/5}$, respectively, etc.; then Assumption A5 is satisfied. When $d_1 = d_2$, undersmoothing is needed to achieve bias reductions in the first- and second-stage estimations, which is familiar from the multistep nonparametric estimation literature. In contrast, when $d_2 < d_1$, undersmoothing may not be required in the second-stage estimation. When $f_{\underline{U}}$ has a compact support and is bounded away from zero on its support, the trimming technique is not needed. See Remark 7 in Section 4.

THEOREM 3.1. *Under Assumption A,*

$$\sqrt{nh_0^{d_1}}(\hat{m}_1^*(x) - m_1(x) - h_0^{q+1}[M^{-1}Bm_1^{(q+1)}(x)]_{0,0}) \xrightarrow{d} N\left(0, \frac{\sigma_\varepsilon^2}{f_X(x)}[M^{-1}\Gamma M^{-1}]_{0,0}\right). \qquad \textbf{(3.2)}$$

Remark 1. We have a sort of “oracle” property here: the feasible estimator $\hat{m}_1^*(x)$ is asymptotically equivalent to $\overline{m}_1(x)$ and hence is more efficient than $\widetilde{m}_1(x)$. Therefore $\hat{m}_1^*(x)$ should be preferred to $\widetilde{m}_1(x)$. The asymptotic normal distribution given by Theorem 3.1 can be used to calculate pointwise confidence intervals for estimators described here. To do this we require an estimation of the asymptotic variance. The procedure is standard, and we omit it for brevity.

Remark 2. It is worth mentioning that our three-stage approach has a close analogy with the well-known prewhitening method in the time series literature. See Press and Tukey (1956), Andrews and Monahan (1992), and more recently

Xiao et al. (2003). Also, our efficiency gain may not require the correct specification of m_2 (or the correct choice of d_2) in the second stage. Interestingly, for any invertible MA process or strictly stationary AR process, the proposed estimator has an efficiency gain over the conventional one-step local polynomial estimator. To see this more clearly, suppose $U_t = (1 - 0.7L)\varepsilon_t$, where L is the lag operator. The process is invertible, so that we can write it as the AR(∞) process: $U_t = -\sum_{j=1}^{\infty} 0.7^j U_{t-j} + \varepsilon_t$. Now if we fit a misspecified nonlinear AR(1) model to the AR(∞) process $\{U_t\}: U_t = m_2(U_{t-1}) + e_t$, where $e_t \equiv U_t - E(U_t|U_{t-1})$, we can tell that $\text{var}(e_t) < \text{var}(U_t)$. So even for this misspecified model, the variance of our third-stage estimator is proportional to $\text{var}(e_t)$, which is strictly smaller than the variance of the preliminary estimator (proportional to $\text{var}(U_t)$). Though not presented here the efficiency gain in such misspecified models was verified in a separate study through Monte Carlo simulations.

Remark 3. Horowitz, Klemela, and Mammen (2002) obtained asymptotic minimax results for estimating additive components in nonparametric regression models under the assumption that the error term is Gaussian. Fan, Gasser, Gijbels, Brockmann, and Engel (1997) showed that, with an appropriate choice of the bandwidth matrix and the kernel function, the univariate local polynomial regression estimator and the multivariate local linear regression estimator achieve the asymptotic linear minimax risk. To apply the latter result to our context, we assume for clarity that $q = 1$ and that the regression function m_1 is in the class

$$\mathcal{C} \equiv \{m_1 : |m_1(z) - m_1'(x)(z-x)^T| \leq 0.5(z-x)C(z-x)^T\},$$

where C is a positive definite $(d_1 \times d_1)$ matrix. Heuristically speaking, this class includes regression functions that have a Hessian matrix bounded by C. If we choose the last-stage bandwidth and kernel function according to equation (3.1) and Theorem 3.1 of Fan et al. (1997), respectively, the resulting estimator $(\hat{m}_1^*)$ will be minimax efficient. Although minimax efficiency is an important property in theory, it is hard to work with and even harder to justify (cf. Linton, 2000). For this reason, we do not stress the minimax efficiency of our estimator.

4. MORE EFFICIENT ESTIMATION BASED ON GENERAL CRITERION FUNCTIONS

As Linton (2000) remarked, the notion of efficiency in nonparametric models is not as clear and well understood as it is in parametric models. One example he gave is that pointwise mean squared error comparisons do not provide a simple ranking between estimators such as kernels, splines, and nearest neighbors. Our purpose is to measure our procedure against the given infeasible ("oracle") procedure for estimating $m_1(x)$ based on the knowledge of $m_2(\cdot)$ and show that the estimator of $m_1(x)$ based on our procedure is more efficient than the estimator obtained by ignoring $m_2(\cdot)$. This result does not pertain to the least squares criterion function only.

Like Linton (2000), we can extend the results in preceding sections and work with criterion functions motivated by the likelihood function of a complete specification of the conditional distribution of Y_t given $W_t \equiv (X_t', \underline{U}_t')'$ along with the additive restriction (1.2). In the following discussion, we shall write $w \equiv (x', \underline{u}')'$. We assume that the conditional distribution of Y_t given $W_t = w$ comes from the one-parameter linear exponential family (e.g., Gourieroux, Monfort, and Trognon, 1984), admitting a density with respect to some fixed measure μ:

$$P(y,m) = \exp\{A(m) + B(y) + C(m)y\}, \tag{4.1}$$

where $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$ are known functions on the real line and $m \in \mathcal{M}$, a suitable parameter space, is the mean of the distribution whose density is $P(y,m)$. Equation (4.1) suggests the following class of criterion functions:

$$Q_n(\theta) = nh_0^{-d_1} \sum_{t=1}^{n} K_0((x - X_t)/h_0)\{Y_t C_t(\theta) + A_t(\theta)\}, \tag{4.2}$$

where $C_t(\theta) = C(m_2(\underline{U}_{t-1}) + \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}}(X_t - x)^{\mathbf{j}})$, $A_t(\theta) = A(m_2(\underline{U}_{t-1}) + \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}}(X_t - x)^{\mathbf{j}})$, and $B(Y_t)$ is absent from (4.2) because it does not depend on θ. Let $\vec{\theta} = \vec{\theta}(x)$ maximize $Q_n(\theta)$ and let $\vec{m}_1(x) = \vec{\theta}_0(x)$ be our infeasible estimate of $m_1(x)$. We have the following result.

THEOREM 4.1. *Suppose that (1.2) holds and the functions $A(\cdot)$ and $C(\cdot)$ have bounded continuous derivatives up to order* $\max(q+1,3)$ *over any compact interval. Then, under Assumption A, we have*

$$\sqrt{nh_0^{d_1}}\left(\vec{m}_1(x) - m_1(x) - h_0^{q+1}[M^{-1}Bm_1^{(q+1)}(x)]_{0,0}\right)$$
$$\xrightarrow{d} N(0, \sigma_\varepsilon^2 a(x)[M^{-1}\Gamma M^{-1}]_{0,0}),$$

where $a(x) = a_2(x)/[a_1(x)]^2$, $a_1(x) = \int C'(m_1(x) + m_2(\underline{u}))f_W(x,\underline{u})\, d\underline{u}$, and $a_2(x) = \int C'(m_1(x) + m_2(\underline{u}))^2 f_W(x,\underline{u})\, d\underline{u}$.

Remark 4. The preceding theorem indicates that ignoring $m_2(\cdot)$ will result in a loss of efficiency in estimating $m_1(\cdot)$ for a general class of criterion functions. Corresponding to each density P (see (4.1)), we have an estimator, $m_1^+(x)$, for $m_1(x)$ that is obtained from maximizing (4.2) when $m_2(\underline{U}_{t-1})$ is absent in the definitions of $A_t(\theta)$ and $C_t(\theta)$. The term $m_1^+(x)$ will have the same bias as $\vec{m}_1(x)$ but the variance that is proportional to σ_U^2. Thus $m_1^+(x)$ is dominated by $\vec{m}_1(x)$ in terms of both variance and mean squared error. For example, if we choose the density P to be normal with mean m and variance 1, then $\vec{m}_1(x) = \bar{m}_1(x)$ and $m_1^+(x) = \tilde{m}_1(x)$. From the preceding section, we know that $\bar{m}_1(x)$ is more efficient than the conventional one-step local polynomial estimator $\tilde{m}_1(x)$.

Remark 5. From the proof of the preceding theorem ((A.5)–(A.9) in particular), we can see that a stronger conclusion can be drawn: all derivatives of $m_1(x)$ up to order q can be estimated more efficiently in the procedure of obtaining $\vec{m}_1(x)$ than in the procedure of obtaining $m_1^+(x)$. Because our primary interest is the efficient estimation of $m_1(x)$, we state the theorem as it is.

Like $\bar{m}_1(x)$ introduced before, $\vec{m}_1(x)$ is an "oracle" estimate. In practice, we can follow the procedure in Section 3 to obtain a feasible estimator for $m_1(x)$. Specifically, we replace $m_2(\underline{U}_{t-1})$ in (4.2) by $\hat{m}_2(\hat{\underline{U}}_{t-1})$ to obtain

$$\hat{Q}_n(\theta) = nh^{-d_1} \sum_{t=1}^{n} K_0((x - X_t)/h_0) I_t\{Y_t \hat{C}_t(\theta) + \hat{A}_t(\theta)\}, \qquad \textbf{(4.3)}$$

where $\hat{C}_t(\theta) = C(\hat{m}_2(\hat{\underline{U}}_{t-1}) + \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}} (X_t - x)^{\mathbf{j}})$ and $\hat{A}_t(\theta) = A(\hat{m}_2(\hat{\underline{U}}_{t-1}) + \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}} (X_t - x)^{\mathbf{j}})$. Let $\hat{\theta}^{**} = \hat{\theta}^{**}(x)$ maximize $\hat{Q}_n(\theta)$ in (4.3) and let $\hat{m}_1^{**}(x) = \hat{\theta}_0^{**}(x)$ be our feasible estimate of $m_1(x)$. A result similar to Theorem 3.1 can be stated as follows.

THEOREM 4.2. *Suppose that the functions* $A(\cdot)$ *and* $C(\cdot)$ *have bounded continuous derivatives up to order* $\max(q+1, 3)$ *over any compact interval. Then, under Assumption A, we have*

$$\sqrt{nh_0^{d_1}}\left(\hat{m}_1^{**}(x) - m_1(x) - h_0^{q+1}[M^{-1}Bm_1^{(q+1)}(x)]_{0,0}\right)$$

$$\xrightarrow{d} N(0, \sigma_\varepsilon^2 a(x)[M^{-1}\Gamma M^{-1}]_{0,0}),$$

where $a(x)$ *is defined in Theorem 4.1.*

Remark 6. This says that $\hat{m}_1^{**}(x)$ is asymptotically equivalent to $\vec{m}_1(x)$ and thus more efficient than $m_1^+(x)$. The expression $\hat{m}_1^{**}(x)$ is constructed by the three-step procedure, implying that the estimation of the infinite-dimensional nuisance parameter $m_2(\cdot)$ has no impact on the limiting distribution of $\hat{m}_1^{**}(x)$. This is not generally the case in parametric estimation problems, unless there is some orthogonality between the estimating equations. One important reason for our procedure to work is that the bias from the first two-step estimation can be made asymptotically negligible by undersmoothing in the first one or two stages. As before, if we choose the density P to be normal with mean m and variance 1, then $\hat{m}_1^{**}(x) = \hat{m}_1^{*}(x)$. From the preceding section, we know that $\hat{m}_1^{*}(x)$ is more efficient than the conventional one-step local polynomial estimator $\tilde{m}_1(x)$.

Remark 7. Recently, Hansen (2004) provides a uniform convergence result for a sample average functional, which can easily be used for density and regression estimation with infinite support. Applying Theorem 3 of Hansen (2004), one can easily show

$$\sup_{\{\underline{u}: f_{\underline{U}}(\underline{u}) \geq b\}} |\hat{m}_2(u) - m_2(u)| = O_p(v_{1,n} + b^{-1} v_{2,n}). \tag{4.4}$$

The preceding result plays a key role in our proof of Theorem 4.2. It is easy to see that the local polynomial estimation of m_2 can be replaced by the standard nonparametric kernel regression. We stick to the local polynomial estimation because there is no need for trimming when $f_{\underline{U}}$ is compactly supported.

5. MONTE CARLO SIMULATIONS

In this section we first address the issue of determining the order d_2 in the second-stage estimation. Then we investigate the proposed estimator $\hat{m}_1^*(x)$ on simulated data and compare it with the conventional one-step local polynomial estimator $\tilde{m}_1(x)$ and the conventional kernel estimator.

5.1. Choice of d_2

In the preceding sections we assume that the order of serial correlation d_2 is known. In practice, however, d_2 is generally unknown. Recently Gao and Tong (2004) have developed a novel cross-validation-based model selection procedure for semiparametric nonlinear time series when all regressors are observables. One can in theory extend their results by allowing for nonparametrically generated regressors. However, the theoretical investigation of this problem can be quite tedious. So we use simulations to examine whether the Gao and Tong method is useful in our context.[5]

We consider a small class of data generating processes (DGPs) for the error process $\{U_t\}$, including both linear and nonlinear AR(1) processes and two linear AR(2) processes:

DGP 1: $U_t = -0.8 U_{t-1} + \varepsilon_t$;
DGP 2: $U_t = 0.95 U_{t-1} \exp(-U_{t-1}^2) + \varepsilon_t$;
DGP 3: $U_t = -0.5 U_{t-1} 1\{|U_{t-1}| > 0.1\} + 0.95 U_{t-1} 1\{|U_{t-1}| \leq 0.1\} + \varepsilon_t$;
DGP 4: $U_t = 2\varphi(U_{t-1}) U_{t-1} + \varepsilon_t$;
DGP 5: $U_t = 0.6 U_{t-1} - 0.3 U_{t-2} + \varepsilon_t$;
DGP 6: $U_t = 0.7 U_{t-1} + 0.2 U_{t-2} + \varepsilon_t$;

where $\varphi(\cdot)$ is the standard normal density function, $\{\varepsilon_t\}$ is i.i.d., and ε_t is equal to the sum of 100 independent random variables each uniformly distributed on $[-0.1, 0.1]$. According to the central limit theorem (CLT), we can treat ε_t as being nearly a normal random variable with truncated support on $[-10, 10]$ (see Yao and Tong, 1994, p. 59). DGPs 1, 5, and 6 allow us to examine the effect of linear dependence on the proposed estimator. DGPs 2–4 are nonlinear processes. The stationary and mixing conditions for the linear processes can be justified easily, and those for the nonlinear processes can be verified by using existing results from Masry and Tjøstheim (1995, 1997). In all cases, the dependent variables are generated according to $Y_t = m_1(X_t) + U_t$, where $m_1(X_t) = 5\exp(X_t)/(1 + \exp(X_t))$ and $\{X_t\}$ is i.i.d. uniform on $[-0.5, 0.5]$ and is independent of the process $\{\varepsilon_t\}$. Simple algebra shows that $\text{var}(m_1(X_t)) = 0.127$, which is comparable to $\text{var}(\varepsilon_t) \simeq 0.333$.

We first obtain a local linear estimator of m_1 by choosing $q = 1$, the Gaussian kernel $K_1(u) = (2\pi)^{-1/2}\exp(-u^2/2)$, and the rule of thumb bandwidth for Gaussian kernel $h_1 = 1.06\hat{s}_X n^{-1/5}$, where $\hat{s}_X$ is the sample standard error for $\{X_t\}_{t=1}^n$. We denote the local linear estimates as $\hat{m}_1(X_t)$ and calculate the residuals $\hat{U}_t = Y_t - \hat{m}_1(X_t)$ for $t = 1,\ldots,n$. Then we apply the procedure of Gao and Tong (2004) to choose the order d_2 based on $\{\hat{U}_t\}_{t=1}^n$.

To be specific, we consider the case where U_{t-1}, U_{t-2}, and U_{t-3} are selected as candidate regressors in generating U_t. Like Gao and Tong (2004), we split the observed data set $\{(\hat{U}_t, R_t) \equiv (\hat{U}_t, \hat{U}_{t-1}, \hat{U}_{t-2}, \hat{U}_{t-3})\}_{t=4}^n$ into two parts: $\{(\hat{U}_t, R_t),\ t \in S\}$ and $\{(\hat{U}_t, R_t),\ t \in S^c\}$, where S is a subset of $\{4,5,\ldots,n\}$ containing n_v integers and S^c is its complement containing n_c integers: $n_c + n_v = n - 3$. The basic idea of Gao and Tong is to use the "construction" data $\{(\hat{U}_t, R_t),\ t \in S^c\}$ to run the regression of $\hat{U}_t$ on potential candidate regressors chosen from R_t and then to assess the prediction error by using the data $\{(\hat{U}_t, R_t),\ t \in S\}$, treated as if they were future values.

Let $\mathcal{D}$ denote all nonempty subsets of $\{1,2,3\}$. For any $D \in \mathcal{D}$, denote $R_{t,D}$ as a column vector consisting of $\hat{U}_{t-i}$ such that $i \in D$. Let K_D be a multivariate kernel function defined on $\mathbb{R}^{|D|}$ and h be a bandwidth parameter. Denote $W_D(t,s) \equiv K_D((R_{t,D} - R_{s,D})/h)/\sum_{r=4}^n K_D((R_{t,D} - R_{r,D})/h)$, $\hat{\phi}_t^c(D) \equiv \sum_{s\in S^c} W_D(t,s)\hat{U}_s$, and $Z_{t,c}(D) \equiv \hat{U}_t - \hat{\phi}_t^c(D)$. The average squared prediction error is

$$CV(D; h, n_v) \equiv \frac{1}{n_v}\sum_{t\in S}\{Z_{t,c}(D)\}^2 w(R_t), \tag{5.1}$$

analogous to equation (2.9) in Gao and Tong (2004), where $w(\cdot)$ is a weighting function. In the special case where $n_v = 1$, we get the conventional leave-one-out cross-validation function (CV1). Let us randomly draw a collection $\mathcal{R}$ of m subsets of $\{4,5,\ldots,n\}$ that have size n_v. Let $MCCV(D; h, n_v) \equiv m^{-1}n_v^{-1}\sum_{S\in\mathcal{R}}\sum_{t\in S}\{Z_{t,c}(D)\}^2 w(R_t)$ and denote $(\hat{D}, \hat{h}) \equiv \arg\min_{\{D\in\mathcal{D},\, h\in H_D^c\}} MCCV(D; h, n_v)$, where H_D^c is a set of bandwidth parameters. Under suitable conditions, we conjecture that with probability converging to one, $\hat{D}$ will pick up the right regressors in R_t.

We choose K_D, H_D^c, m, and n_c according to Section 3.1 in Gao and Tong (2004): K_D is a product of $|D|$ standard normal kernels, $H_D^c \equiv [0.1n_c^{-2/9}, 3n_c^{-1/9}]$, $m = n - 3$, and $n_c = \lceil (n-3)^{0.9} \rceil$, the largest integer part of $(n-3)^{0.9}$. We use $w(R_t) = \Pi_{i=1}^3 1(|\hat{U}_{t-i}| \le 2\hat{s}_U)$, where $\hat{s}_U$ is the sample standard error for $\{\hat{U}_t\}_{t=1}^n$. To save time, we specify $\mathcal{D} = \{\{1\},\{1,2\},\{1,2,3\}\}$, and we choose $h \in H_D^c$ by discretizing H_D^c with 50 equally spaced points. Table 1 reports the

Table 1. Relative frequencies based on the semiparametric MCCV order selection of Gao and Tong (2004)

	DGP1 $(d_2 = 1)$	DGP2 $(d_2 = 1)$	DGP3 $(d_2 = 1)$	DGP4 $(d_2 = 1)$	DGP5 $(d_2 = 2)$	DGP6 $(d_2 = 2)$
$n = 100$						
$\{U_{t-1}\}$	0.842	0.860	0.776	0.842	0.315	0.508
$\{U_{t-1}, U_{t-2}\}$	0.134	0.114	0.164	0.120	0.486	0.394
$\{U_{t-1}, U_{t-2}, U_{t-3}\}$	0.024	0.026	0.060	0.038	0.199	0.098
$n = 200$						
$\{U_{t-1}\}$	0.932	0.948	0.872	0.908	0.144	0.462
$\{U_{t-1}, U_{t-2}\}$	0.060	0.052	0.124	0.084	0.748	0.516
$\{U_{t-1}, U_{t-2}, U_{t-3}\}$	0.008	0	0.004	0.008	0.108	0.024
$n = 500$						
$\{U_{t-1}\}$	1	1	0.960	0.970	0.010	0.300
$\{U_{t-1}, U_{t-2}\}$	0	0	0.040	0.030	0.960	0.700
$\{U_{t-1}, U_{t-2}, U_{t-3}\}$	0	0	0	0	0.030	0

Note: The results are based on 500, 250, and 100 repetitions for n = 100, 200, and 500, respectively.

relative frequencies that each element of $\{\{\hat{U}_{t-1}\}, \{\hat{U}_{t-1}, \hat{U}_{t-2}\}, \{\hat{U}_{t-1}, \hat{U}_{t-2}, \hat{U}_{t-3}\}\}$ is selected by using the Monte Carlo cross validation (MCCV) criterion function.

Table 1 shows that when the true order (d_2) of the time series $\{U_t\}$ is 1, the MCCV procedure of Gao and Tong (2004) can pick up the right order even when the sample size is fairly small ($n = 200$) given the fact that three-dimensional densities have to be estimated in the selection procedure. When $d_2 = 2$, the success of the procedure depends on the persistence in the data: for low persistent data (DGP 5), the Gao and Tong procedure still works well for n as small as 200, but for highly persistent data (DGP 6), a large sample ($n \geq 500$) is required.

5.2. Efficiency Comparison

Now we suppose that the right number of lags, d_2, has been chosen and investigate the proposed estimator $\hat{m}_1^*(x)$ on simulated data and compare it with the conventional one-step local polynomial estimator $\tilde{m}_1(x)$ and the conventional kernel estimator, $\hat{m}_{1,\text{ker}}(x) \equiv \sum_{s=1}^n K_0((x - X_s)/h_0)/\sum_{t=1}^n K_0((x - X_t)/h_0)Y_t$. We do not try to optimize the performance of either the conventional estimators or our own. Rather, we take what are fairly common choices, in real applications, of bandwidth sequences and kernels and demonstrate that even with these implements there are significant finite-sample gains to be made. We choose the same Epanechnikov kernel in estimating $\hat{m}_1^*(x)$, $\tilde{m}_1(x)$, and $\hat{m}_{1,\text{ker}}(x)$:

$K_i(u) = 0.75(1 - u^2)1(|u| \le 1)$, $i = 0,1,2$. Note that the error term has a compact support; we do not use the trimming device in our simulations.

For bandwidth sequences, noticing that only h_0 shows up in the limiting distribution of our more efficient estimator, we recommend choosing h_0 based upon the nonparametric leave-one-out cross-validation method. To be specific, let $h_0 = c_0 \hat{s}_X n^{-1/(4+d_1)}$; we choose h_0 to be

$$\hat{h}_0 \equiv \arg \min_{h_0 \in H^{d_1}} CV(h_0) \equiv \frac{1}{n} \sum_{t=1}^{n} \{Y_t - \hat{m}_{1,\text{ker}}^{(-t)}(X_t)\}^2 w(X_t), \tag{5.2}$$

where $\hat{m}_{1,\text{ker}}^{(-t)}(x)$ is obtained as $\hat{m}_{1,\text{ker}}(x)$ by deleting the tth observation, $w(X_t)$ is defined analogously as in Section 5.1, and $\hat{h}_0$ is obtained by grid search over the interval $H^{d_1} \equiv [0.1\hat{s}_X n^{-1/(4+d_1)}, 5\hat{s}_X n^{-1/(4+d_1)}]$ in 50 steps. We shall use the same bandwidth $\hat{h}_0$ in obtaining $\widetilde{m}_1(x)$ and $\hat{m}_{1,\text{ker}}(x)$ as in the third step of obtaining $\hat{m}_1^*(x)$.

To obtain $\hat{m}_1^*(x)$, we also need to choose h_1 and h_2 in the first two-stage estimation. As we shall see, our efficient results are not sensitive to the choice of these two bandwidth parameters, and so we recommend some rules of thumb based on the optimal bandwidth for nonparametric kernel density estimation. It is well known that the optimal bandwidth for estimating a one-dimension density with the Epanechnikov kernel is given by $h_{opt} = 2.34\hat{s}_X n^{-1/5}$. So when $d_1 = 1$ and $q = p = 1$, by rule of thumb we can set $h_1 = 2.34\hat{s}_X n^{-1/4}$, $h_2 = 2.34\hat{s}_U n^{-1/(3+d_2)}$, where we have imposed undersmoothing conditions. This choice of h_1 and h_2, together with the choice of $h_0 = \hat{h}_0$, will meet Assumption A5 for $d_2 = 1$, 2, or 3, where we conjecture that this data-driven bandwidth selection does not alter the validity of our theorems. If $d_1 \ge 2$, we can modify the rule correspondingly, e.g., set $h_1 = 2.34\hat{s}_X n^{-1/(3+d_1)}$ and $h_2 = 2.34\hat{s}_U n^{-1/(3+d_2)}$, where one should understand that h_1 is a diagonal matrix and that $\hat{s}_X$ is a diagonal matrix with the standard deviation of, say, $\{X_{i,t}\}$ in the (i,i)th place. In the following experiment ($d_1 = 1$, $d_2 = 1$ or 2), we set $h_1 = c\hat{s}_X n^{-1/4}$, $h_2 = c\hat{s}_U n^{-1/(3+d_2)}$, where $c \in \{1, 2, 4\}$, and consider three sample sizes: $n = 100$, 200, and 500.

First we consider the estimation of $m_1(x)$ at the interior point $x = 0$ and report in Table 2 the relative efficiency: the ratio of average mean squared errors over 500 replications. In the table, row RE1 reports the relative efficiency of the proposed efficient estimator $\hat{m}_1^*(x)$ over the conventional kernel estimator $\hat{m}_{1,\text{ker}}(x)$, and row RE2 reports the relative efficiency of the proposed efficient estimator $\hat{m}_1^*(x)$ over the conventional local polynomial estimator $\widetilde{m}_1(x)$. For comparison purposes, we also report in Table 2 the infeasible asymptotic relative efficiency (RE0) calculated based upon the asymptotic variance of $\hat{m}_1^*(x)$ and $\widetilde{m}_1(x)$. It is worth mentioning that the second derivative of $m_1(x)$ at $x = 0$ is 0. So RE0 is also the asymptotic relative efficiency based on the mean squared error of $\hat{m}_1^*(x)$ and $\widetilde{m}_1(x)$ at $x = 0$.

Table 2. Relative efficiency of the efficient nonparametric estimator

		DGP1	DGP2	DGP3	DGP4	DGP5	DGP6
	RE0	(0.360)	(0.182)	(0.524)	(0.382)	(0.716)	(0.225)
$n = 100$							
$c = 1$	RE1	0.525	0.918	0.898	0.851	0.805	0.978
	RE2	0.495	0.912	0.865	0.850	0.792	0.963
$c = 2$	RE1	0.403	0.920	0.832	0.861	0.937	0.923
	RE2	0.396	0.920	0.816	0.827	0.886	0.933
$c = 4$	RE1	0.341	0.977	0.931	0.843	0.799	0.915
	RE2	0.349	0.969	0.914	0.850	0.755	0.913
$n = 200$							
$c = 1$	RE1	0.432	0.957	0.776	0.837	0.913	0.922
	RE2	0.390	0.961	0.778	0.823	0.861	0.926
$c = 2$	RE1	0.374	0.892	0.750	0.909	0.777	0.960
	RE2	0.364	0.892	0.722	0.914	0.765	0.947
$c = 4$	RE1	0.314	0.925	0.819	0.813	0.804	0.923
	RE2	0.287	0.912	0.827	0.803	0.800	0.922
$n = 500$							
$c = 1$	RE1	0.305	0.901	0.697	0.828	0.926	0.919
	RE2	0.297	0.897	0.703	0.827	0.927	0.918
$c = 2$	RE1	0.322	0.878	0.714	0.802	0.825	0.907
	RE2	0.321	0.877	0.726	0.797	0.836	0.910
$c = 4$	RE1	0.383	0.905	0.766	0.805	0.779	0.931
	RE2	0.377	0.908	0.784	0.800	0.771	0.930

Note: RE0 is the infeasible asymptotic relative efficiency, and RE1 and RE2 are the relative efficiency of the efficient estimator $\hat{m}_1^*(x)$ over the conventional kernel estimator $\hat{m}_{1,\mathrm{ker}}(x)$ and conventional local polynomial estimator $\widetilde{m}_1(x)$, respectively.

We summarize some general findings from our simulation experiments. (1) The proposed efficient estimator performs very well across various DGPs under our investigation and across different c's that control the degree of smoothing in the first two-stage estimations. (2) In general, the more the serial dependence in the error process, the larger is the efficiency gain achieved. In particular, for the linear AR(1) processes, like Xiao et al. (2003), we find in a separate study that the relative efficiency first improves as the AR coefficient in absolute value increases and then degenerates as the coefficient approaches one. This suggests that a different asymptotic theory may apply when we have a near unit error process. (3) There are more substantial gains that can be achieved for negative serial dependence cases (DGP1) than for positive serial dependence cases (DGPs 2, 4–6). The RE1 and RE2 can reach RE0 in the case where the negative serial dependence is large (DGP1). (4) The relative efficiency generally improves when the sample sizes n increase from 100 to 200.

We also consider the estimation of $m_1(x)$ over all sample points $X_1, \ldots, X_n$ and calculate the integrated relative efficiency (IRE) as the ratios of mean squared errors that are averaged over a smaller number of replications and over all data points. The results are qualitatively similar to the preceding analysis.

NOTES

1. We do not use boldface letters to denote random variables such as X_t and $\underline{U}_{t-1}$ and their values, x and $\underline{u}$.

2. It is straightforward to show that $\sqrt{nh_0^{d_1}}(\widetilde{m}_1(x) - m_1(x) - h_0^{q+1}[M^{-1}Bm_1^{(q+1)}(x)]_{0,0}) \xrightarrow{d} N(0, (\sigma_U^2/f_X(x)) \times [M^{-1}\Gamma M^{-1}]_{0,0})$.

3. We can make the comparison at the respectively optimal bandwidths. By Masry (1996a), the optimal bandwidth for estimating $m_1(x)$ by $\widetilde{m}_1(x)$ is given by $h_{0,opt} = \{(\sigma_U^2 d_1 [M^{-1}\Gamma M^{-1}]_{0,0})/(2(q+1)\{[M^{-1}Bm_1^{(q+1)}(x)]_{0,0}\}^2)\}n^{-1/(2q+2+d_1)}$. For $\overline{m}_1$ the optimal bandwidth is the same as that for $\widetilde{m}_1$ but with σ_ε^2 in place of σ_U^2, and hence it is smaller. With this, one can obtain the formula for the mean squared errors when the optimal bandwidth is used.

4. Noting that we need to obtain $\hat{m}_1(X_t)$ and $\hat{U}_t = Y_t - \hat{m}_1(X_t)$ for all $t = 1, \ldots, n$, we can tell that the Nadaraya–Watson kernel estimator is less desirable than the local polynomial estimator: we need to correct the boundary bias in the case where $\{X_t\}$ is compactly supported or apply some trimming techniques otherwise. See Fan and Gijbels (1996) or Pagan and Ullah (1999) for more comparisons between the two types of estimators.

5. The authors are thankful to a referee for suggesting the use of the Gao and Tong (2004) simulation procedure.

REFERENCES

Andrews, D.W.K. & J.C. Monahan (1992) An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* 60, 953–966.

Bosq, D. (1996) *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction.* Lecture Notes in Statistics 110. Springer-Verlag.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87, 998–1004.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, & J. Engel (1997) Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics* 49, 79–99.

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

Gao, J. & M. King (2002) Estimation and Model Specification Testing in Nonparametric and Semiparametric Regression Models. Technical Report, Department of Mathematics and Statistics, University of Western Australia.

Gao, J. & H. Tong (2004) Semiparametric non-linear time series model selection. *Journal of the Royal Statistical Society, Series B* 66, 321–336.

Gourieroux, C., A. Monfort, & A. Trognon (1984) Pseudo maximum likelihood methods: Theory. *Econometrica* 52, 681–700.

Gozalo, P. & O. Linton (2000) Local non-linear least squares: Using parametric information in nonparametric regression. *Journal of Econometrics* 99, 63–106.

Gozalo, P. & O. Linton (2001) Testing additivity in generalized nonparametric regression models with estimated parameters. *Journal of Econometrics* 104, 1–48.

Greene, W.H. (1997) *Econometric Analysis*. 3rd ed. Prentice-Hall.

Hansen, B.E. (2004) Uniform Convergence Rates for Kernel Regression. Working paper, Department of Economics, University of Wisconsin, Madison.

Hidalgo, F.J. (1992) Adaptive semiparametric estimation in the presence of autocorrelation of unknown form. *Journal of Time Series Analysis* 13, 47–78.

Hong, Y. (1996) Consistent testing for serial correlation of unknown form. *Econometrica* 64, 837–864.

Horowitz, J., J. Klemela, & E. Mammen (2002) Optimal estimation in additive regression models. Working paper, Northwestern University.

Li, Q. & J.M. Wooldridge (2002) Semiparametric estimation of partially linear models for dependent data with generated regressors. *Econometric Theory* 18, 625–645.

Linton, O. (1997) Efficient estimation of additive nonparametric regression models. *Biometrika* 84, 469–473.

Linton, O. (2000) Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* 16, 502–523.

Linton, O. & W. Härdle (1996) Estimating additive regression models with known links. *Biometrika* 83, 529–540.

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

Masry, E. (1996b) Multivariate local polynomial regression for time series: Uniform strong consistency rates. *Journal of Time Series Analysis* 17, 571–599.

Masry, E. & D. Tjøstheim (1995) Nonparametric estimation and identification of non-linear ARCH time series. *Econometric Theory* 11, 258–289.

Masry, E. & D. Tjøstheim (1997) Additive non-linear ARX time series and projection estimates. *Econometric Theory* 13, 214–252.

Pagan, A. & A. Ullah (1999) *Nonparametric Econometrics*. Cambridge University Press.

Press, H. & J.W. Tukey (1956) Power spectral methods of analysis and their application to problems in airline dynamics. *Flight Test Manual*, NATO, Advisory Group for Aeronautical Research and Development, vol. IV-C, pp. 1–41.

Robinson, P.M. (1988) Root-*n*-consistent semiparametric regression. *Econometrica* 56, 931–954.

Tong, H. (1990) *Non-linear Time Series: A Dynamic Systems Approach.* Oxford University Press.

Wooldridge, J.M. (2003) *Introductory Econometrics: A Modern Approach*. 2nd ed. South-Western.

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

Yao, Q. & H. Tong (1994) On subset selection in non-parametric stochastic regression. *Statistica Sinica* 4, 51–70.

Yoshihara, K. (1976) Limiting behavior of U-statistics for stationary absolutely regular processes. *Zeitschrift für Wahrscheinlichskeitstheorie und Verwandte Gebiete* 35, 237–252.

APPENDIX

We use $\|\cdot\|$ to denote the euclidean norm of $\cdot$, c to signify a generic constant whose exact value may vary from case to case, and we use a^T to denote the transpose of a. Let $W_t = (X_t^T, \underline{U}_{t-1}^T)^T$, $w = (x^T, \underline{u}^T)^T$, $m(w) = m_1(x) + m_2(\underline{u})$, $G(w;z) = m(w)C(z) + A(z)$, $Z_t = m(W_t)$, $\bar{Z}_t(\theta) = m_2(\underline{U}_{t-1}) + \sum_{0\le|\mathbf{j}|\le q}\theta_{\mathbf{j}}(X_t - x)^{\mathbf{j}}$, and $\hat{Z}_t(\theta) = \hat{m}_2(\hat{\underline{U}}_{t-1}) + \sum_{0\le|\mathbf{j}|\le q}\theta_{\mathbf{j}}(X_t - x)^{\mathbf{j}}$. Further, let $G^{(j)}(w;z)$, $j = 1,2,\dots,\max(q + 1,3)$, denote partial derivatives of G with respect to z.

Proof of Theorem 2.1. The proof follows from the argument of Masry (1996a) or the proof of Theorem 1 in Xiao et al. (2003) with $\tilde{Y}_t = Y_t - m_2(\underline{U}_{t-1})$ in place of $\underline{Y}_t$ in their paper. Alternatively, one can apply Theorem 4.1. To see this, write $Q_{n,loc}(\theta) = nh_0^{-d_1}\sum_{t=d_2+1}^{n} K_0((x - X_t)/h_0)Y_t^2 - 2nh_0^{-d_1}\sum_{t=1}^{n} K_0((x - X_t)/h_0)\{Y_t C_t(\theta) + A_t(\theta)\} \equiv$

$nh_0^{-d_1}\sum_{t=d_2+1}^{n} K_0((x - X_t)/h_0)Y_t^2 - 2Q_{n,1}(\theta)$, where $C_t(\theta) \equiv C(\bar{Z}_t(\theta)) = \bar{Z}_t(\theta)$, $A_t(\theta) \equiv A(\bar{Z}_t(\theta)) = -\bar{Z}_t(\theta)^2/2$ with $\bar{Z}_t(\theta) \equiv m_2(\underline{U}_{t-1}) + \sum_{0\le|\mathbf{j}|\le q}\theta_{\mathbf{j}}\ (X_t - x)^{\mathbf{j}}$. It is immediate to see that $Q_{n,1}(\theta)$ plays the role of $Q_n(\theta)$ in Theorem 4.1 and $a(x) = 1/f_X(x)$ as desired. ■

Proof of Theorem 3.1. Write $\hat{Q}_{n,loc}(\theta) = nh_0^{-d_1}\sum_{t=d_2+1}^{n} K_0((x - X_t)/h_0)Y_t^2 - 2nh_0^{-d_1}\sum_{t=1}^{n} K_0((x - X_t)/h_0)\{Y_t\hat{C}_t(\theta) + \hat{A}_t(\theta)\} \equiv nh_0^{-d_1}\sum_{t=d_2+1}^{n} K_0((x - X_t)/h_0)Y_t^2 - 2\hat{Q}_{n,1}(\theta)$, where $\hat{C}_t(\theta) \equiv C(\hat{Z}_t(\theta)) = \hat{Z}_t(\theta)$ and $\hat{A}_t(\theta) \equiv A(\hat{Z}_t(\theta)) = -\hat{Z}_t(\theta)^2/2$ with $\hat{Z}_t(\theta) \equiv \hat{m}_2(\hat{\underline{U}}_{t-1}) - \sum_{0\le|\mathbf{j}|\le q}\theta_{\mathbf{j}}(X_t - x)^{\mathbf{j}}$. Then $\hat{Q}_{n,1}(\theta)$ plays the role of $\hat{Q}_n(\theta)$ in Theorem 4.2, and the result follows immediately. ■

Proof of Theorem 4.1. Let $\theta^0(x)$ be the collection of $\theta_{\mathbf{j}}^0(x) = (1/\mathbf{j}!)(D^{\mathbf{j}}m_1)(x)$, $0 \le |\mathbf{j}| \le q$, the true local parameters, using the lexicographical order introduced in the text. For example, $|\mathbf{j}| = 0$ corresponds to the first element in the collection, to be denoted as $\theta_0^0(x)$. We first show that $\bar{\theta}_0(x) (= \bar{m}_1(x))$ consistently estimates $\theta_0^0(x)$ and that $\bar{\theta}_{\mathbf{j}}(x)$ consistently estimates $\theta_{\mathbf{j}}^0(x)$ for $1 \le |\mathbf{j}| \le q$. By Assumption A and the continuity of $A(\cdot)$ and $C(\cdot)$, we can apply the uniform law of large numbers (e.g., Gozalo and Linton, 2000, p. 101) to get

$$\sup_{\theta\in\Theta}|Q_n(\theta) - \bar{Q}_n(\theta)| \to_p 0, \tag{A.1}$$

where $\bar{Q}_n(\theta) = E\{Q_n(\theta)\}$ and $Q_n(\theta)$ is as in (4.2). Furthermore, noting that $E\{Y_tC_t(\theta) + A_t(\theta)|W_t\} = m(W_t)C_t(\theta) + A_t(\theta) = G(W_t;\bar{Z}_t(\theta))$ by the definitions of G and $\bar{Z}_t(\theta)$, we have

$$\begin{aligned}\bar{Q}_n(\theta) &= \int G\left(W_t; m_2(\underline{U}_{t-1}) + \theta_0 + \sum_{1\le|\mathbf{i}|\le q}\theta_{\mathbf{i}}(X_t - x)^{\mathbf{i}}\right)h_0^{-d_1}K_0\left(\frac{X_t - x}{h_0}\right)f_W(W_t)\,dW_t\\ &= \int G\left(x - vh_0, \underline{u}; m_2(\underline{u}) + \theta_0 + \sum_{1\le|\mathbf{i}|\le q}\theta_i v^{\mathbf{i}}h_0^{|\mathbf{i}|}\right)K_0(v)f_W(x - vh_0, \underline{u})\,dvd\underline{u}\\ &\to \int G(w; m_2(\underline{u}) + \theta_0)f_W(w)\,d\underline{u} \equiv Q_0(\theta_0)\end{aligned} \tag{A.2}$$

uniformly in $\theta \in \Theta$. By Property 4 of Gourieroux et al. (1984), $Q_0(\theta_0) \le Q_0(\theta_0^0)$ with equality if and only if $\theta_0 = \theta_0^0$. This establishes consistency of $\bar{\theta}_0(x)$ for $\theta_0^0(x)$.

The derivative parameters $\theta_{\mathbf{i}}(x)$, $1 \le |\mathbf{i}| \le q$, are determined by subsequent order terms (in h_0) through a Taylor expansion of (A.2). For example, let $\mathbf{j} = (1,0,\ldots,0) \in \mathbb{R}^{d_1}$ and $\theta_{-\mathbf{j}}$ denote all the elements in θ except $\theta_{\mathbf{j}}$. We consider the consistency of $\bar{\theta}_{\mathbf{j}}(x)$ for $\theta_{\mathbf{j}}^0(x)$. When evaluated at $(\theta_{\mathbf{j}}, \theta_{-\mathbf{j}}^0)$, $\theta_{\mathbf{j}} = \theta_{\mathbf{j}}(x)$ is determined by the order term that is, apart from terms that do not depend on θ_1 or are of smaller orders, a constant times

$$Q_1(\theta_{\mathbf{j}}) \equiv h_0^2\int\left\{\alpha(w)\theta_{\mathbf{j}} + \frac{1}{2}\beta(w)\theta_{\mathbf{j}}^2\right\}f_W(w)\,d\underline{u}, \tag{A.3}$$

where $\alpha(w) = (\partial m/\partial x_1)(w)C'(m(w))$ and $\beta(w) = G''(w; m(w))$. By Property 3 of Gourieroux et al. (1984), $C'(m) > 0$. By Properties 1 and 2 of Gourieroux et al. (1984), $G''(w; m(w)) = -C'(m(w))$. We can see that the unique maximum of $Q_1(\theta_{\mathbf{j}})$ is $\theta_{\mathbf{j}}(x) =$

$(\partial m_1/\partial x_1)(x)$, where $\mathbf{j} = (1,0,\ldots,0)$. This establishes the consistency of $\bar{\theta}_{\mathbf{j}}(x)$. Similarly, one can establish the consistency of other elements in $\bar{\theta}(x)$.

We now turn to the asymptotic normality. To facilitate the proof, we denote $\mathcal{M}_t = \mathcal{M}_t(x)$ and $\mathcal{K}_{0,t} = \mathcal{K}_{0,t}(x)$ as a symmetric $N \times N$ matrix and an $N \times 1$ vector, respectively:

$$\mathcal{M}_t(x) = \begin{bmatrix} \mathcal{M}_{t,0,0}(x) & \mathcal{M}_{t,0,1}(x) & \ldots & \mathcal{M}_{t,0,q}(x) \\ \mathcal{M}_{t,1,0}(x) & \mathcal{M}_{t,1,1}(x) & \ldots & \mathcal{M}_{t,1,q}(x) \\ \vdots & \vdots & & \vdots \\ \mathcal{M}_{t,q,0}(x) & \mathcal{M}_{t,q,1}(x) & \ldots & \mathcal{M}_{t,q,q}(x) \end{bmatrix}, \qquad \mathcal{K}_{0,t}(x) = \begin{bmatrix} \mathcal{K}_{0,t,0}(x) \\ \mathcal{K}_{0,t,1}(x) \\ \vdots \\ \mathcal{K}_{0,t,q}(x) \end{bmatrix}, \tag{A.4}$$

where $\mathcal{M}_{t,j,k}(x)$ is an $N_j \times N_k$-dimensional submatrix with the (l,r) element given by

$$[\mathcal{M}_{t,j,k}]_{l,r} = \left(\frac{X_t - x}{h_0}\right)^{\phi_j(l)+\phi_k(r)} K_0\left(\frac{X_t - x}{h_0}\right)$$

and $\mathcal{K}_{0,t,j}(x)$ is an N_j-dimensional subvector whose rth element is given by

$$[\mathcal{K}_{0,t,j}(x)]_r = \left(\frac{X_t - x}{h_0}\right)^{\phi_j(r)} K_0\left(\frac{X_t - x}{h_0}\right).$$

By asymptotic expansion of $Q_n(\theta)$ at θ^0, we have

$$H_n[\bar{\theta}(x) - \theta^0(x)] = -\left[H_n^{-1}\frac{\partial^2 Q_n(\theta^*(x))}{\partial\theta\partial\theta^T}H_n^{-1}\right]^{-1} H_n^{-1}\frac{\partial Q_n(\theta^0(x))}{\partial\theta}, \tag{A.5}$$

where $\theta^*(x)$ is a vector intermediate between $\bar{\theta}(x)$ and $\theta^0(x)$ and $H_n = \operatorname{diag}(1, h_0, \ldots, h_0, \ldots, h_0^q, \ldots, h_0^q)$ with N_l terms of h_0^l, $0 \le l \le q$ (so H_n is an $N \times N$ matrix). The presentation of (A.5) assumes that the matrix in the square brackets is invertible with probability tending to one, which we will show subsequently. The score function is

$$\frac{\partial Q_n(\theta^0(x))}{\partial\theta} = \frac{1}{nh_0^{d_1}} H_n \sum_{t=1}^{n} \mathcal{K}_{0,t}(x)\{Y_t C'(\bar{Z}_t) + A'(\bar{Z}_t)\},$$

whereas the Hessian matrix is

$$\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta^T} = \frac{1}{nh_0^{d_1}} H_n \sum_{t=1}^{n} \mathcal{M}_t(x)\{Y_t C''(\bar{Z}_t(\theta)) + A''(\bar{Z}_t(\theta))\} H_n,$$

where $\bar{Z}_t(\theta) = m_2(\underline{U}_{t-1}) + \sum_{0 \le |\mathbf{j}| \le q} \theta_{\mathbf{j}}(X_t - x)^{\mathbf{j}}$, $\bar{Z}_t = \bar{Z}_t(\theta^0(x))$.

We next show that the score vector satisfies a CLT. Noting that $Y_t C'(\bar{Z}_t) + A'(\bar{Z}_t) = \varepsilon_t C'(\bar{Z}_t) + G'(W_t;\bar{Z}_t)$, $H_n^{-1}[\partial Q_n(\theta^0(x))/\partial\theta] = n^{-1}h_0^{-d_1}\sum_{t=1}^{n}\mathcal{K}_{0,t}(x)C'(\bar{Z}_t)\varepsilon_t + n^{-1}h_0^{-d_1}\sum_{t=1}^{n}\mathcal{K}_{0,t}(x)G'(W_t;\bar{Z}_t) \equiv T_{n1} + T_{n2}$. By the law of iterated expectation and the dominated convergence theorem, we have $E[\mathcal{K}_{0,t}(x)C'(\bar{Z}_t)\varepsilon_t] = 0$ and $\operatorname{var}[\mathcal{K}_{0,t}(x)C'(\bar{Z}_t)\varepsilon_t] = \sigma_\varepsilon^2 E\{\mathcal{K}_{0,t}(x)[\mathcal{K}_{0,t}(x)]^T C'(\bar{Z}_t)^2\} = h_0^{d_1}\sigma_\varepsilon^2 a_2(x)\Gamma\{1 + o(1)\}$, where $a_2(x) = \int C'(m(w))^2 f_w(x, \underline{u})\, d\underline{u}$, and Γ is defined in (2.3). Under Assumptions A1, A2, and A5, we can follow the argument of Masry (1996a, Thm. 3) to show that the covariance terms are asymptotically negligible and that a CLT applies to T_{n1}:

$$\sqrt{nh_0^{d_1}}T_{n1} \xrightarrow{d} N(0, \sigma_\varepsilon^2 a_2(x)\Gamma). \tag{A.6}$$

The second vector in the score function contributes to the bias. Noting that $G'(W_t;Z_t) = 0$ (by Property 1 of Gourieroux et al., 1984), $G''(W_t;Z_t) = -C'(m(W_t))$, and $\theta_{\mathbf{k}}^0 = (1/\mathbf{k}!)(D^{\mathbf{k}}m_1)(x)$, using Taylor expansion twice gives us

$$\begin{aligned} G'(W_t;\bar{Z}_t) &= -C'(m(W_t))\left[-m_1(X_t) + \sum_{0\le|\mathbf{k}|\le q} \frac{1}{\mathbf{k}!}(D^{\mathbf{k}}m_1)(x)(X_t - x)^{\mathbf{k}}\right] \\ &\quad + G'''(W_t;Z_t^*)\left[-m_1(X_t) + \sum_{0\le|\mathbf{k}|\le q} \frac{1}{\mathbf{k}!}(D^{\mathbf{k}}m_1)(x)(X_t - x)^{\mathbf{k}}\right]^2 \\ &= C'(m(W_t))\left[\sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k}!}(D^{\mathbf{k}}m_1)(x)(X_t - x)^{\mathbf{k}} + o_p(h_0^{q+1})\right] \\ &\quad + G'''(W_t;Z_t^*)\left[\sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k}!}(D^{\mathbf{k}}m_1)(x)(X_t - x)^{\mathbf{k}} + o_p(h_0^{q+1})\right]^2, \end{aligned}$$

where Z_t^* is intermediate between $\bar{Z}_t$ and Z_t, the second equality follows by expanding $m_1(X_t)$ around x for $\|X_t - x\| \le c_0 h_0$ because the kernel K_0 has compact support on $[-c_0, c_0]$ by Assumption A.1 and $m_1(x)$ has continuous derivatives of total order $q+1$ by Assumption A.3. So

$$\begin{aligned} T_{n2} &= \frac{1}{nh_0^{d_1}}\sum_{t=1}^{n} \mathcal{K}_{0,t}\, C'(m(W_t))\left\{\sum_{|\mathbf{k}|=q+1} \frac{1}{\mathbf{k}!}(D^{\mathbf{k}}m_1)(x)(X_t - x)^{\mathbf{k}}\right\} + o_p(h_0^{q+1}) \\ &= h_0^{q+1} a_1(x) B m_1^{(q+1)}(x) + o_p(h_0^{q+1}), \end{aligned} \tag{A.7}$$

where $a_1(x) = \int C'(m(w)) f_w(x, \underline{u})\, d\underline{u}$ and B is defined in (2.3), and the last equality follows by a standard law of large numbers, change of variables, and dominated convergence arguments.

Now by the uniform law of large numbers and dominated convergence arguments, we have

$$\sup_{\theta\in\Theta_n(x)} \left| H_n^{-1}\left(\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 Q_n(\theta^0(x))}{\partial\theta\partial\theta^T}\right) H_n^{-1}\right| \to_p 0, \tag{A.8}$$

where $\Theta_n(x)$ is a shrinking neighborhood of $\theta^0(x)$. Furthermore, by the law of large numbers and dominated convergence theorem,

$$
\begin{aligned}
H_n^{-1}\frac{\partial^2 Q_n(\theta^0(x))}{\partial\theta\partial\theta^T}H_n^{-1} &= \frac{1}{nh_0^{d_1}}\sum_{t=1}^{n}\mathcal{M}_t(x)\{G''(\bar{Z}_t)+\varepsilon_t C''(\bar{Z}_t)\}\\
&= \frac{1}{nh_0^{d_1}}\sum_{t=1}^{n}E[\mathcal{M}_t(x)G''(\bar{Z}_t)]+o_p(1)\\
&= -\frac{1}{nh_0^{d_1}}\sum_{t=1}^{n}E[\mathcal{M}_t(x)C'(m(W_t))]+o_p(1)\\
&= -a_1(x)M+o_p(1), \qquad \textbf{(A.9)}
\end{aligned}
$$

where $a_1(x) = \int C'(m(w))f_w(x,\underline{u})\,d\underline{u}$ and M is defined in (2.3). The theorem then follows from equations (A.5)–(A.9). ■

For any $\theta = \theta(x) \in \Theta$, let $\hat{Z}_t(\theta) = \hat{m}_2(\hat{\underline{U}}_{t-1}) + \sum_{0\le|\mathbf{j}|\le q}\theta_{\mathbf{j}}(X_t - x)^{\mathbf{j}}$, $\hat{Z}_t = \hat{Z}_t(\theta^0)$, $\bar{Z}_t = \bar{Z}_t(\theta^0)$, $Z_t = m(W_t) = m_1(X_t) + m_2(\underline{U}_{t-1})$ as before. Clearly, $\eta_{nt}(\theta) \equiv \hat{Z}_t(\theta) - \bar{Z}_t(\theta) = \hat{Z}_t - \bar{Z}_t = \hat{m}_2(\hat{\underline{U}}_{t-1}) - m_2(\underline{U}_{t-1}) \equiv \eta_{nt}$. Decompose $\eta_{nt} = [\hat{m}_2(\hat{\underline{U}}_{t-1}) - m_2(\hat{\underline{U}}_{t-1})] + [m_2(\hat{\underline{U}}_{t-1}) - m_2(\underline{U}_{t-1})] \equiv \eta_{nt,1} + \eta_{nt,2}$. Denote $\nu_{1n} = n^{-1/2}h_1^{-d_1/2}\sqrt{\ln n}$ and $\nu_{2n} = n^{-1/2}h_2^{-d_2/2}\sqrt{\ln n}$. Let $\mathcal{A}_n = \{\underline{u} : f_{\underline{U}}(\underline{u}) > b\}$ and $I_t = 1\{f_{\underline{U}}(\underline{U}_{t-1}) > b\}$. Our proof of Theorem 4.2 relies on the following propositions and lemmas.

PROPOSITION A.1. *Under Assumption A,* $\eta_{nt,1}I_t = \tilde{e}_1'[f_{\underline{U}}(\underline{U}_{t-1})M(K_2,p)]^{-1}\mathcal{U}_{n,1}(\underline{U}_{t-1})I_t + o_p(n^{-1/2}h_0^{-d_1/2})$ *uniformly in* t *and* $\max_{d_2+1\le t\le n}\eta_{nt,1}I_t = O_p(b^{-1}\nu_{2n})$, *where* $M(K_2,p)$ *is defined as* M *in (2.3) but with kernel* K_2 *and local polynomial order* p *and* $\mathcal{U}_{n,1}$ *is defined subsequently in Lemma A.2.*

Proof. To facilitate the proof, let $\mathcal{K}_{2,t}(\underline{u})$ be an $\tilde{N} \times 1$ vector, $\mathcal{K}_{2,t}'(\underline{u})$ be an $\tilde{N} \times d_2$ matrix, and $M_n(\underline{u})$ be a symmetric $\tilde{N} \times \tilde{N}$ matrix:

$$
\mathcal{K}_{2,t}(\underline{u}) = \begin{bmatrix} \mathcal{K}_{2,t,0}(\underline{u}) \\ \mathcal{K}_{2,t,1}(\underline{u}) \\ \vdots \\ \mathcal{K}_{2,t,p}(\underline{u}) \end{bmatrix}, \qquad \mathcal{K}_{2,t}'(\underline{u}) = \begin{bmatrix} \mathcal{K}_{2,t,0}'(\underline{u}) \\ \mathcal{K}_{2,t,1}'(\underline{u}) \\ \vdots \\ \mathcal{K}_{2,t,p}'(\underline{u}) \end{bmatrix},
$$

$$
M_n(\underline{u}) = \begin{bmatrix} M_{n,0,0}(\underline{u}) & M_{n,0,1}(\underline{u}) & \dots & M_{n,0,p}(\underline{u}) \\ M_{n,1,0}(\underline{u}) & M_{n,1,1}(\underline{u}) & \dots & M_{n,1,p}(\underline{u}) \\ \vdots & \vdots & & \vdots \\ M_{n,p,0}(\underline{u}) & M_{n,p,1}(\underline{u}) & \dots & M_{n,p,p}(\underline{u}) \end{bmatrix},
$$

where $\tilde{N} \equiv \sum_{j=0}^{p}N_j$, $\mathcal{K}_{2,t,j}(\underline{u})$ is a N_j-dimensional subvector whose rth element is given by

$$
[\mathcal{K}_{2,t,j}(\underline{u})]_r = \left(\frac{\underline{U}_{t-1} - \underline{u}}{h_2}\right)^{\phi_j(r)} K_2\left(\frac{\underline{U}_{t-1} - \underline{u}}{h_2}\right),
$$

$\mathcal{K}'_{2,t,j}(\underline{u})$ is an $N_j \times d_2$ matrix with the (r,l) element being the partial derivative of $[\mathcal{K}_{2,t,j}(\underline{u})]_r$ with respect to the lth element in $\underline{u}$, and $M_{n,j,k}(\underline{u})$ is an $N_j \times N_k$-dimensional submatrix with the (l,r) element given by

$$[M_{n,j,k}(\underline{u})]_{l,r} = \frac{1}{(n-d_2)h_2^{d_2}} \sum_{t=d_2+1}^{n} \left(\frac{\underline{U}_{t-1} - \underline{u}}{h_2} \right)^{\phi_j(l)+\phi_k(r)} K_2\left(\frac{\underline{U}_{t-1} - \underline{u}}{h_2} \right).$$

The terms $\hat{\mathcal{K}}_{2,t}(\underline{u})$ and $\hat{M}_n(\underline{u})$ are defined analogously as $\mathcal{K}_{2,t}(\underline{u})$ and $M_n(\underline{u})$, respectively, but with the residual series $\{\hat{U}_1,\ldots,\hat{U}_n\}$ in place of the latent variables $\{U_1,\ldots,U_n\}$.

Like Xiao et al. (2003, Lem. A.3), we write for $\underline{u} \in \mathcal{A}_n$,

$$\begin{aligned}
\hat{m}_2(\underline{u}) - m_2(\underline{u}) &= \tilde{e}_1' \hat{M}_n^{-1}(\underline{u}) \hat{V}_n(\underline{u}) + \tilde{e}_1' \hat{M}_n^{-1}(\underline{u}) \hat{B}_n(\underline{u}) \\
&\simeq \tilde{e}_1' [f_{\underline{U}}(\underline{u}) M(K_2,p)]^{-1} \hat{V}_n(\underline{u}) + \tilde{e}_1' [f_{\underline{U}}(\underline{u}) M(K_2,p)]^{-1} \hat{B}_n(\underline{u}) \\
&\quad + \tilde{e}_1' [f_{\underline{U}}(\underline{u}) M(K_2,p)]^{-1} [\hat{M}_n(\underline{u}) - f_{\underline{U}}(\underline{u}) M(K_2,p)] \hat{V}_n(\underline{u}) \\
&\quad + \tilde{e}_1' [f_{\underline{U}}(\underline{u}) M(K_2,p)]^{-1} [\hat{M}_n(\underline{u}) - f_{\underline{U}}(\underline{u}) M(K_2,p)] \hat{B}_n(\underline{u}) \\
&\equiv T_{n,1}(\underline{u}) + T_{n,2}(\underline{u}) + T_{n,3}(\underline{u}) + T_{n,4}(\underline{u}),
\end{aligned}$$

where the "variance" term $\hat{V}_n(\underline{u})$ and the "bias" term $\hat{B}_n(\underline{u})$ are $\tilde{N} \times 1$ vectors defined by $\hat{V}_n(\underline{u}) = (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \hat{\mathcal{K}}_{2,t}(\underline{u}) \varepsilon_t$, and $\hat{B}_n(\underline{u}) = (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \hat{\mathcal{K}}_{2,t}(\underline{u}) \hat{\Delta}_t(\underline{u})$, where $\hat{\Delta}_t(\underline{u}) = m_2(\hat{\underline{U}}_{t-1}) - \sum_{0 \le |\mathbf{k}| \le p} (1/\mathbf{k}!)(D^{\mathbf{k}} m_2)(\underline{u})(\hat{\underline{U}}_{t-1} - \underline{u})^{\mathbf{k}}$. We analyze the properties of $T_{n,i}(\underline{u})$, $i = 1,\ldots,4$, in Lemmas A.2–A.5, which will complete the proof of the proposition. ■

LEMMA A.2. *Under Assumption A,* $T_{n,1}(\hat{\underline{U}}_{t-1}) I_t = \tilde{e}_1' [f_{\underline{U}}(\underline{U}_{t-1}) M(K_2,p)]^{-1} \mathcal{U}_{n,1}(\underline{U}_{t-1}) I_t + o_p(n^{-1/2} h_0^{-d_1/2})$ *uniformly in t, and* $\sup_{\underline{u} \in \mathcal{A}_n} T_{n,1}(\underline{u}) = O_p(b^{-1} \nu_{2n})$*, where* $\mathcal{U}_{n,1}(\underline{u}) = (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \mathcal{K}_{2,t}(\underline{u}) \varepsilon_t$.

Proof. Let $V_n(\underline{u})$ be defined as $\hat{V}_n(\underline{u})$ but with $\mathcal{K}_{2,t}(\underline{u})$ in place of $\hat{\mathcal{K}}_{2,t}(\underline{u})$. By the Taylor expansion and Assumptions A.1 and A.3, we have

$$\begin{aligned}
\hat{V}_n(\underline{u}) - V_n(\underline{u}) &= \frac{1}{(n-d_2)h_2^{d_2}} \sum_{t=d_2+1}^{n} [\hat{\mathcal{K}}_{2,t}(\underline{u}) - \mathcal{K}_{2,t}(\underline{u})] \varepsilon_t \\
&= \frac{1}{(n-d_2)h_2^{d_2}} \sum_{t=d_2+1}^{n} [\mathcal{K}'_{2,t}(\underline{u})] \{\hat{\underline{U}}_{t-1} - \underline{U}_{t-1}\} \varepsilon_t \{1 + o_p(1)\} \\
&= \frac{1}{(n-d_2)h_2^{d_2}} \sum_{t=d_2+1}^{n} \sum_{i=1}^{d_2} \frac{\partial \mathcal{K}_{2,t}(\underline{u})}{\partial \underline{u}_i} \varepsilon_t \{m_1(X_{t-i}) - \hat{m}_1(X_{t-i})\} \{1 + o_p(1)\},
\end{aligned} \tag{A.10}$$

where $o_p(1)$ is uniformly in $\underline{u}$ and the second equality follows from the property of K_2 and the fact that $\sup_{d_2+1 \le t \le n} \|\hat{\underline{U}}_{t-1} - \underline{U}_{t-1}\| \le d_2 \max_{i=1,\ldots,d_2} \sup_{d_2+1 \le t \le n} \|\hat{m}_1(X_{t-i}) - m_1(X_{t-i})\| = O_p(\nu_{1n} + h_1^{q+1}) = o_p(1)$.

Now, let $e_1 = (1,0,\ldots,0)' \in \mathbb{R}^N$. By Masry (1996b), uniformly in x, $\hat{m}_1(x) - m_1(x) = n^{-1} h_1^{-d_1} e_1' [M f_X(x)]^{-1} \sum_{t=1}^{n} \mathcal{K}_{1,t}(x) \{U_t + \sum_{|\mathbf{k}|=q+1} (1/\mathbf{k}!)(D^{\mathbf{k}} m_1)(x)(X_t - x)^{\mathbf{k}}\}$

$\{1 + o_p(1)\}$, where $\mathcal{K}_{1,t}(x)$ is defined analogously to $\mathcal{K}_{2,t}(\underline{u})$ with a typical element: $[\mathcal{K}_{1,t,j}(x)]_r = ((X_t - x)/h_1)^{\phi_j(r)} K_1((X_t - x)/h_1)$. So $\hat{V}_n(\underline{u}) - V_n(\underline{u}) = V_{n,1}(\underline{u}) + V_{n,2}(\underline{u})$, in which $V_{n,1}(\underline{u}) \equiv n^{-1}(n - d_2)^{-1} h_1^{-d_1} h_2^{-d_2-1} \sum_{t=d_2+1}^{n} \sum_{s=1}^{n} \alpha_n(V_t, V_s; \underline{u})$, $V_{n,2}(\underline{u}) \equiv (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \sum_{i=1}^{d_2} (\partial \mathcal{K}_{2,t}(\underline{u})/\partial \underline{u}_i) \varepsilon_t\ \beta_n(X_{t-i})$, where for $V_t \equiv (X_t', X_{t-1}, \ldots, X_{t-d_2}', \underline{U}_{t-1}', \varepsilon_t)'$, $\alpha_n(V_t, V_s; \underline{u}) \equiv \sum_{i=1}^{d_2} h_2 \varepsilon_t (\partial \mathcal{K}_{2,t}(\underline{u})/\partial \underline{u}_i) e_1' [M f_X(X_{t-i})]^{-1} \mathcal{K}_{1,s}(X_{t-i}) U_s$, and $\beta_n(x) \equiv n^{-1} h_1^{-d_1} e_1' [M f_X(x)]^{-1} \sum_{s=1}^{n} \mathcal{K}_{1,s}(x) \{\sum_{|\mathbf{k}|=q+1} (1/\mathbf{k}!)(D^{\mathbf{k}} m_1)(x)(X_s - x)\}$.

For fixed $\underline{u}$, $V_{n,1}(\underline{u})$ is a second-order U statistic, and it is easy to show that $V_{n,1}(\underline{u}) = O_p(n^{-1} h_1^{-d_1/2} h_2^{-d_2/2-1})$. For a uniform bound on $V_{n,1}(\underline{u})$, one can modify the proof of (A.10) in Gozalo and Linton (2001) to show that $\sup_{\underline{u}} |V_{n,1}(\underline{u})| = O_p(n^{-1} h_1^{-d_1/2} h_2^{-d_2/2-1} \ln n) = o_p(n^{-1/2} h_0^{-d_1/2})$. It is straightforward to extend the proof of Theorem 2 in Masry (1996b) to show that $\sup_x |h_1^{-(q+1)} \beta_n(x) - \beta(x)| = O_p(\nu_{1n})$, where $\beta(x) = e_1' M^{-1} B m_1^{(q+1)}(x)$. So $V_{n,2}(\underline{u}) = h_1^{q+1} (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \sum_{i=1}^{d_2} (\partial \mathcal{K}_{2,t}(\underline{u})/\partial \underline{u}_i) \varepsilon_t\ \beta(X_{t-i}) + o_p(n^{-1/2} h_0^{-d_1/2})$. By a similar argument as used in the proof of Theorem 3 in Hansen (2004), the first term is $O_p(n^{-1/2} h_2^{-d_2/2-1} \sqrt{\ln n}\, h_1^{q+1}) = o_p(n^{-1/2} h_0^{-d_1/2})$ uniformly in $\underline{u}$ by Assumption A.5(i) and (ii), and thus $\sup_{\underline{u}} |V_{n,2}(\underline{u})| = o_p(n^{-1/2} h_0^{-d_1/2})$.

Also, by an application of Theorem 3 in Hansen (2004), $\sup_{\underline{u}} \|V_n(\underline{u})\| = O_p(\nu_{2n})$. So by the triangle inequality, $\sup_{\underline{u}} \|\hat{V}_n(\underline{u})\| = O_p(\nu_{2n}) + o_p(n^{-1/2} h_0^{-d_1/2}) = O_p(\nu_{2n})$, and the second part of the lemma follows. The first part of the lemma follows because, by the Taylor expansion, Assumptions A5(i), (iii), and (iv), $T_{n,1}(\hat{\underline{U}}_{t-1}) I_t = T_{n,1}(\underline{U}_{t-1}) I_t + O_p(b^{-2} \nu_{2n}) O_p(h_2^{-1} \nu_{1n} + h_2^{-1} h_1^{q+1}) = T_{n,1}(\underline{U}_{t-1}) I_t + o_p(n^{-1/2} h_0^{-d_1/2})$. This concludes the proof. ■

LEMMA A.3. *Under Assumption A,* $\sup_{\underline{u} \in \mathcal{A}_n} |T_{n,2}(\underline{u})| = o_p(n^{-1/2} h_0^{-d_1/2})$.

Proof. Let $\Delta_t(\underline{u}) = m_2(\underline{U}_{t-1}) - \sum_{0 \le |\mathbf{k}| \le p} (1/\mathbf{k}!)(D^{\mathbf{k}} m_2)(\underline{u})(\underline{U}_{t-1} - \underline{u})^{\mathbf{k}}$. Then by Assumption A3, $\Delta_t(\underline{u}) = \sum_{|\mathbf{k}|=p+1} (1/\mathbf{k}!)(D^{\mathbf{k}} m_2)(\underline{u}_t^*)(\underline{U}_{t-1} - \underline{u})^{\mathbf{k}}$ for some $\underline{u}_t^*$ that lies between $\underline{U}_{t-1}$ and $\underline{u}$, and $\hat{\Delta}_t(\underline{u}) = \sum_{|\mathbf{k}|=p+1} (1/\mathbf{k}!)(D^{\mathbf{k}} m_2)(\hat{\underline{u}}_t^*)(\hat{\underline{U}}_{t-1} - \underline{u})^{\mathbf{k}}$ for some $\hat{\underline{u}}_t^*$ that lies between $\hat{\underline{U}}_{t-1}$ and $\underline{u}$. Clearly $\|\hat{\underline{u}}_t^* - \underline{u}_t^*\| = O_p(\nu_{1n} + h_1^{q+1})$ uniformly in t and $\underline{u}$. So by Assumption A3,

$$
\begin{aligned}
\hat{\Delta}_t(\underline{u}) - \Delta_t(\underline{u}) &= \left\{ \sum_{|\mathbf{k}|=p+1} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}} m_2)(\hat{\underline{u}}_t^*)[(\hat{\underline{U}}_{t-1} - \underline{u})^{\mathbf{k}} - (\underline{U}_{t-1} - \underline{u})^{\mathbf{k}}] \right\} \\
&\quad + \left\{ \sum_{|\mathbf{k}|=p+1} \frac{1}{\mathbf{k}!} [(D^{\mathbf{k}} m_2)(\hat{\underline{u}}_t^*) - (D^{\mathbf{k}} m_2)(\underline{u}_t^*)](\underline{U}_{t-1} - \underline{u})^{\mathbf{k}} \right\} \\
&= h_2^p O_p(\nu_{1n} + h_1^{q+1}) \quad \text{uniformly in } \underline{u} \text{ and } t \text{ for } \|\underline{U}_{t-1} - \underline{u}\| \le c h_2.
\end{aligned}
$$

Because $|\Delta_t(\underline{u})| = O_p(h_2^{p+1})$ for $\|\underline{U}_{t-1} - \underline{u}\| \le c h_2$, it follows that $|\hat{\Delta}_t(\underline{u})| \le O_p(h_2^p \nu_{1n} + h_2^p h_1^{q+1} + h_2^{p+1})$ for $\|\underline{U}_{t-1} - \underline{u}\| \le c h_2$. Now, write $\hat{B}_n(\underline{u}) = (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \{\hat{\mathcal{K}}_{2,t}(\underline{u}) - \mathcal{K}_{2,t}(\underline{u})\} \hat{\Delta}_t(\underline{u}) + (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \mathcal{K}_{2,t}(\underline{u}) \{\hat{\Delta}_t(\underline{u}) - \Delta_t(\underline{u})\} + (n - d_2)^{-1} h_2^{-d_2} \sum_{t=d_2+1}^{n} \mathcal{K}_{2,t}(\underline{u}) \Delta_t(\underline{u}) \equiv B_{n,1}(\underline{u}) + B_{n,2}(\underline{u}) + B_{n,3}(\underline{u})$. It is easy to show that $\sup_{\underline{u}} \|B_{n,1}(\underline{u})\| = O_p(n^{-1/2} h_1^{-d_1/2} h_2^{-1} \sqrt{\ln n}) O_p(h_2^p \nu_{1n} + h_2^{p+1}) = o_p(b n^{-1/2} h_0^{-d_1/2})$ by Assumptions A1 and A.5(i), (ii), and (iv), $\sup_{\underline{u}} \|B_{n,2}(\underline{u})\| = O_p(h_2^p \nu_{1n}) = o_p(b n^{-1/2} h_0^{-d_1/2})$ by Assumptions A1 and A5(iv), and $\sup_{\underline{u}} \|B_{n,3}(\underline{u})\| = O_p(h_2^{p+1}) = o_p(b n^{-1/2} h_0^{-d_1/2})$ by Assumptions A1 and A.5(i). The lemma follows. ■

LEMMA A.4. *Under Assumption A,* $\sup_{\underline{u}\in\mathcal{A}_n}|T_{n,3}(\underline{u})| = o_p(n^{-1/2}h_0^{-d_1/2})$.

Proof. For a typical element of $\hat{M}_n(\underline{u}) - M_n(\underline{u})$, we have

$$[\hat{M}_{n,j,k}(\underline{u})]_{l,r} - [M_{n,j,k}(\underline{u})]_{l,r}$$

$$= \frac{1}{(n-d_2)h_2^{d_2}} \sum_{t=d_2+1}^{n} \left[\left(\frac{\hat{\underline{U}}_{t-1} - \underline{u}}{h_2} \right)^{\phi_j(l)+\phi_k(r)} K_2\left(\frac{\hat{\underline{U}}_{t-1} - \underline{u}}{h_2} \right) \right.$$
$$\left. - \left(\frac{\underline{U}_{t-1} - \underline{u}}{h_2} \right)^{\phi_j(l)+\phi_k(r)} K_2\left(\frac{\underline{U}_{t-1} - \underline{u}}{h_2} \right) \right].$$

By Assumption A1, we can show that $\sup_{\underline{u}}|[\hat{M}_{n,j,k}(\underline{u})]_{l,r} - [M_{n,j,k}(\underline{u})]_{l,r}| = O_p(h_2^{-1}v_{1n} + h_2^{-1}h_1^{q+1})$. By an application of Theorem 3 of Hansen (2004), $\sup_{\underline{u}}\|M_n(\underline{u}) - f_{\underline{U}}(\underline{u})M(K_2,p)\| = O_p(h_2 + v_{2n})$. So by the triangle inequality, $\sup_{\underline{u}}|\hat{M}_n(\underline{u}) - f_{\underline{U}}(\underline{u})M(K_2,p)| = O_p(h_2 + v_{2n} + h_2^{-1}v_{1,n} + h_2^{-1}h_1^{q+1})$, analogous to Masry (1996b). From the proof of Lemma A.2, $\sup_{\underline{u}}\|\hat{V}_n(\underline{u})\| = O_p(v_{2n})$. Consequently, $\sup_{\underline{u}}|T_{n,3}(\underline{u})| = O_p(h_2 + v_{2n} + h_2^{-1}v_{1n} + h_2^{-1}h_1^{q+1})O_p(b^{-1}v_{2n}) = o_p(n^{-1/2}h_0^{-d_1/2})$, where the expression in the last sentence follows from Assumptions A5(i)–(v). ■

LEMMA A.5. *Under Assumption A,* $\sup_{\underline{u}\in\mathcal{A}_n}|T_{n,4}(\underline{u})| = o_p(n^{-1/2}h_0^{-d_1/2})$.

Proof. The lemma follows from the proofs of Lemmas A.3 and A.4: $\sup_{\underline{u}}\|\hat{B}_n(\underline{u})\| = o_p(bn^{-1/2}h_0^{-d_1/2})$ and $\sup_{\underline{u}}|\hat{M}_n(\underline{u}) - f_{\underline{U}}(\underline{u})M(K_2,p)| = O_p(h_2 + v_{2n} + h_2^{-1}v_{1,n} + h_2^{-1}h_1^{q+1}) = o_p(1)$. ■

PROPOSITION A.6. *Under Assumption A,* $\eta_{nt,2} = \mathcal{U}_{n,2}(\underline{U}_{t-1}) + o_p(n^{-1/2}h_0^{-d_1/2})$ *uniformly in t, and* $\max_{d_2+1\le t\le n}|\eta_{nt,2}| = O_p(\nu_{1n})$*, where* $\mathcal{U}_{n,2}(\underline{U}_{t-1}) = -n^{-1}h_1^{-d_1} \sum_{s=1}^{n}\sum_{i=1}^{d_2}(\partial m_2(\underline{U}_{t-1})/\partial U_{t-i})e_1'[Mf_X(X_{t-i})]^{-1}\mathcal{K}_{1,s}(X_{t-i})U_s$.

Proof. The second part follows from the fact that $\max_{d_2+1\le t\le n}\|\hat{\underline{U}}_{t-1} - \underline{U}_{t-1}\| = O_p(\nu_{1n} + h_1^{q+1}) = O_p(\nu_{1n})$ by Assumption A5. The first part follows because $\eta_{nt,2} = m_2'(\underline{U}_{t-1})(\hat{\underline{U}}_{t-1} - \underline{U}_{t-1}) + O_p((\nu_{1n})^2) = \sum_{i=1}^{d_2}(\partial m_2(\underline{U}_{t-1})/\partial U_{t-i})\{m_1(X_{t-i}) - \hat{m}_1(X_{t-i})\} + o_p(n^{-1/2}h_0^{-d_1/2}) = \mathcal{U}_{n,2}(\underline{U}_{t-1}) + o_p(n^{-1/2}h_0^{-d_1/2})$. ■

Remark. Propositions A.1 and A.6 imply that

$$\eta_{nt}I_t = \tilde{e}_1'[f_{\underline{U}}(\underline{U}_{t-1})M(K_2,p)]^{-1}\mathcal{U}_{n,1}(\underline{U}_{t-1})I_t + \mathcal{U}_{n,2}(\underline{U}_{t-1})$$
$$+ o_p(n^{-1/2}h_0^{-d_1/2}) \quad \text{uniformly in } t \tag{A.11}$$

and

$$\max_{d_2+1\le t\le n}|\eta_{nt}I_t| = O_p(\nu_{1n} + b^{-1}\nu_{2n}) = o_p(1). \tag{A.12}$$

Proof of Theorem 4.2. By the assumption on $A(\cdot)$ and $C(\cdot)$, we expand $G(W_t;\hat{Z}_t(\theta))$ and its derivatives about $G(W_t;\bar{Z}_t(\theta))$ in a Taylor series to get (for $j = 0,1$, and $r = 1,2,\ldots,\max(q,2)$),

$$G^{(j)}(W_t;\hat{Z}_t(\theta)) = \sum_{i=0}^{r-1} G^{(j+i)}(W_t;\bar{Z}_t(\theta))[\eta_{nt}]^i + G^{(j+r)}(W_t;\bar{Z}_t^{*j}(\theta))[\eta_{nt}]^r, \qquad \textbf{(A.13)}$$

where $\bar{Z}_t^{*j}(\theta)$ are intermediate between $\hat{Z}_t(\theta)$ and $\bar{Z}_t(\theta)$. By (A.12), we have

$$\sup_{\theta\in\Theta_n(x)} \max_{d_2+1\le t\le n} I_t|G^{(j)}(W_t;\hat{Z}_t(\theta)) - G^{(j)}(W_t;\bar{Z}_t(\theta))| = o_p(1), \qquad j = 0,1.$$

A similar result evidently holds for A and C and their derivatives. Therefore

$$\begin{aligned}
&\sup_{\theta\in\Theta}|\hat{Q}_n(\theta) - Q_n(\theta)| \\
&\quad = \sup_{\theta\in\Theta}\left|\frac{1}{(n-d_2)h_0^{d_1}}\sum_{t=d_2+1}^{n} K_0\left(\frac{x-X_t}{h_0}\right)\{Y_t[\hat{I}_t\hat{C}_t(\theta) - C_t(\theta)] + [\hat{I}_t\hat{A}_t(\theta) - A_t(\theta)]\}\right| \\
&\quad \le \sup_{\theta\in\Theta}\frac{1}{(n-d_2)h_0^{d_1}}\sum_{t=d_2+1}^{n}\left|K_0\left(\frac{x-X_t}{h_0}\right)\varepsilon_t\right||\hat{I}_t\hat{C}_t(\theta) - C_t(\theta)| \\
&\qquad + \sup_{\theta\in\Theta}\frac{1}{(n-d_2)h_0^{d_1}}\sum_{t=d_2+1}^{n}\left|K_0\left(\frac{x-X_t}{h_0}\right)\right||\hat{I}_tG(W_t;\hat{Z}_t(\theta)) - G(W_t;\bar{Z}_t(\theta))| \\
&\quad \le \frac{c}{(n-d_2)h_0^{d_1}}\sum_{t=d_2+1}^{n}\left|K_0\left(\frac{x-X_t}{h_0}\right)\varepsilon_t(1-\hat{I}_t)\right| \\
&\qquad + \frac{c}{(n-d_2)h_0^{d_1}}\sum_{t=d_2+1}^{n}\left|K_0\left(\frac{x-X_t}{h_0}\right)(1-\hat{I}_t)\right| + o_p(1) \\
&\quad = D_{n,1} + D_{n,2} + o_p(1),
\end{aligned}$$

where the second inequality follows from (A.13) and the following facts: (i) K_0 is compactly supported; (ii) $m_2(\cdot)$ is bounded on its support. Recall that $\hat{I}_t = 1\{\hat{f}_{\underline{U}}(\hat{\underline{U}}_{t-1}) \ge b\}$. Note that $P(\hat{f}_{\underline{U}}(\hat{\underline{U}}_{t-1}) < b) \le P(|\hat{f}_{\underline{U}}(\hat{\underline{U}}_{t-1}) - f_{\underline{U}}(\underline{U}_{t-1})| < b) + P(f_{\underline{U}}(\underline{U}_{t-1}) < 2b)$. By Lemma 6 of Robinson (1988), the second term goes to zero as $b \to 0$. For the first term, one can show that it goes to zero by altering the proof of Proposition 4 in Robinson (1988). So $E[\hat{I}_t] \to 1$ as $n \to \infty$. Consequently $D_{n,i} = o_p(1)$, $i = 1,2$, implying $\hat{m}_1^*(x) \xrightarrow{p} m_1(x)$.

The proof is complete if we can show

$$\left\|H_n^{-1}\left(\frac{\partial\hat{Q}_n(\theta^0(x))}{\partial\theta} - \frac{\partial Q_n(\theta^0(x))}{\partial\theta}\right)\right\| = o_p(n^{-1/2}h_0^{-d_1/2}) \quad \text{and} \qquad \textbf{(A.14)}$$

$$\sup_{\theta\in\Theta_n(x)}\left\|H_n^{-1}\left(\frac{\partial^2\hat{Q}_n(\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta^T}\right)H_n^{-1}\right\| = o_p(1). \qquad \textbf{(A.15)}$$

Consider the $N \times 1$ vector $H_n^{-1}\big(\partial\hat{Q}_n(\theta^0(x))/\partial\theta - \partial Q_n(\theta^0(x))/\partial\theta\big) = (n - d_2)^{-1} h_0^{-d_1}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t}(x)\varepsilon_t\{\hat{I}_tC'(\hat{Z}_t) - C'(\bar{Z}_t)\} + (n - d_2)^{-1}h_0^{-d_1}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t}(x)\{\hat{I}_tG'(W_t;\hat{Z}_t) - G'(W_t;\bar{Z}_t)\} \equiv T_{n,5} + T_{n,6}$. In the discussion that follows, we shall restrict our attention to the first components of $T_{n,5}$ and $T_{n,6,}$ which are of interest. The other components behave similarly because the functions $K_0(u)$ and $u^{\mathbf{j}}K_0(u)$, $|\mathbf{j}| = 1,\dots,q$, have similar properties. Recall that $I_t = 1\{f_{\underline{U}}(\underline{U}_{t-1}) > b\}$. Following the strategy of Hidalgo (1992, p. 73), we first establish

$$(n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\varepsilon_t\{I_t C'(\hat{Z}_t) - C'(\bar{Z}_t)\} = o_p(1), \tag{A.16}$$

$$(n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\{I_t G'(W_t;\hat{Z}_t) - G'(W_t;\bar{Z}_t)\} = o_p(1) \tag{A.17}$$

and then show that there is no difference asymptotically when we replace I_t by $\hat{I}_t$ in the preceding two equations.

Write

$$\begin{aligned}
&(n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\varepsilon_t\{I_t C'(\hat{Z}_t) - C'(\bar{Z}_t)\}\\
&\quad = (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\varepsilon_t I_t\{C'(\hat{Z}_t) - C'(\bar{Z}_t)\}\\
&\qquad - (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\varepsilon_t C'(\bar{Z}_t)(1-I_t)\\
&\quad = T_{n,5,1} - T_{n,5,2},
\end{aligned} \tag{A.18}$$

where, e.g., $T_{n,5,2} = (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}S_t$, $S_t = K_{0,t}(x)\varepsilon_t C'(\bar{Z}_t)(1-I_t)$. Clearly, $E[T_{n,5,2}] = 0$, $E[T_{n,5,2}]^2 = h_0^{d_1}E[S_t]^2 + 2(n-d_2)^{-1}h_0^{d_1}\sum_{\tau=d_2+1}^{n-1}\sum_{t=\tau+1}^{n}E[S_{t-\tau}S_t] \equiv I + II$. By the dominated convergence theorem, $I \to 0$. By the Davydov inequality (Bosq, 1996, p. 19), Assumptions A1 and A5, $II \le ch_0^{d_1(1-\delta)/(1+\delta)}\sum_{\tau=d_2+1}^{n-1}(1-\tau/(n-d_2))\alpha(\tau)^{\delta/(1+\delta)} \to 0$. So

$$T_{n,5,2} = o_p(1) \tag{A.19}$$

by the Chebyshev inequality.

$$\begin{aligned}
T_{n,5,1} &= (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\varepsilon_t I_t\{C'(\hat{Z}_t) - C'(\bar{Z}_t)\}\\
&= (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\varepsilon_t I_t\{C''(\bar{Z}_t)(\hat{Z}_t - \bar{Z}_t)\\
&\qquad\qquad + O_p((\nu_{1n} + b^{-1}\nu_{2n})^2)\}\\
&= (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\varepsilon_t I_t\{C''(\bar{Z}_t)(\hat{Z}_t - \bar{Z}_t)\} + o_p(1),
\end{aligned}$$

by (A.12) and Assumption A.5(ii). By (A.11), Proposition A.1, and Proposition A.6, the leading term of $T_{n,5,1}$ equals

$$\begin{aligned}
T_{n,5,1a} &= \frac{1}{(n-d_2)^{1/2}h_0^{d_1/2}} \sum_{t=d_2+1}^{n} \mathcal{K}_{0,t,0}(x)\varepsilon_t I_t C''(\bar{Z}_t)\eta_{nt} \\
&= \frac{1}{(n-d_2)^{3/2}h_0^{d_1/2}h_2^{d_2}} \sum_{t_1=d_2+1}^{n} \sum_{t_2=d_2+1}^{n} \mathcal{K}_{0,t_1,0}(x)\varepsilon_{t_1} I_{t_1} C''(\bar{Z}_{t_1}) \\
&\quad \times [f_{\underline{U}}(\underline{U}_{t_1-1})M(K_2,p)]^{-1}\mathcal{K}_{2,t_2}(\underline{U}_{t_1-1})\varepsilon_{t_2} \\
&\quad - \frac{1}{n(n-d_2)^{1/2}h_0^{d_1/2}h_1^{d_1}} \sum_{t_1=d_2+1}^{n} \sum_{t_2=1}^{n} \mathcal{K}_{0,t_1,0}(x)\varepsilon_{t_1} I_{t_1} C''(\bar{Z}_{t_1}) \\
&\quad \times \sum_{i=1}^{d_2} \frac{\partial m_2(\underline{U}_{t_1-1})}{\partial U_{t_1-i}} e_1'[Mf_X(X_{t_1-i})]^{-1}\mathcal{K}_{1,t_2}(X_{t_1-i})U_{t_2} + o_p(1) \\
&\equiv \sum_{t_1=d_2+1}^{n} \sum_{t_2=d_2+1}^{n} S_{t_1,t_2}^{(1)} - \sum_{t_1=d_2+1}^{n} \sum_{t_2=1}^{n} S_{t_1,t_2}^{(2)} + o_p(1) \\
&\equiv S_{n,1} - S_{n,2} + o_p(1),
\end{aligned}$$

where, e.g., $S_{t_1,t_2}^{(1)} \equiv S^{(1)}(V_{t_1},V_{t_2})$ with $V_t \equiv (X_t', X_{t-1}', \ldots, X_{t-d_2}', \underline{U}_{t-1}', \varepsilon_t)'$. We shall use $\sum_{t_1,t_2}^{n}$ to denote $\sum_{t_1=d_2+1}^{n}\sum_{t_2=d_2+1}^{n}$ and write $\tilde{S}_{t_1,t_2}^{(1)}$ for $S^{(1)}(\widetilde{V}_{t_1},\widetilde{V}_{t_2})$, where $\{\widetilde{V}_t,\ t \geq 1\}$ denotes an i.i.d. sequence with the same marginal distributions as $\{V_t,\ t \geq 1\}$. Obviously, $E[\tilde{S}_{t_1,t_2}^{(1)}] = 0$ for $t_1 \neq t_2$. We can readily apply the Davydov inequality to show $\sum_{t_1,t_2}^{n} E(S_{t_1,t_2}^{(1)}) = o(1)$ and Lemma C.2 in Gao and King (2002) to show $\sum_{t_1,t_2}^{n}\sum_{t_3,t_4}^{n} E(S_{t_1,t_2}^{(1)}S_{t_3,t_4}^{(1)}) = o(1)$, so that $S_{n,1} = o_p(1)$ by the Chebyshev inequality. Similarly, one can show that $S_{n,2} = o_p(1)$ and hence

$$T_{n,5,1} = o_p(1). \tag{A.20}$$

Combining (A.18), (A.19), and (A.20), we have shown (A.16).

For (A.17), $(n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)\{I_t G'(W_t;\hat{Z}_t) - G'(W_t;\bar{Z}_t)\} = (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)I_t\{G'(W_t;\hat{Z}_t) - G'(W_t;\bar{Z}_t)\} - (n-d_2)^{-1/2}h_0^{-d_1/2}\sum_{t=d_2+1}^{n}\mathcal{K}_{0,t,0}(x)G'(W_t;\bar{Z}_t)(1-I_t) = T_{n,6,1} - T_{n,6,2}$. One can analyze the first term $T_{n,6,1}$ in an analogous fashion to $T_{n,5,1}$ to obtain $T_{n,6,1} = o_p(1)$. For the second term, recall that $G'(W_t;Z_t) = 0$ by Property 1 of Gourieroux et al. (1984) and G' has bounded derivatives over any compact interval, so $T_{n,6,2} = O_p((n-d_2)^{1/2}h_0^{d_1/2}h_0^{q+1}) = o_p(1)$. This proves (A.17).

By the triangle inequality, Masry (1996b), and Hansen (2004), $\max_{d_2+1\leq t\leq n}|\hat{f}_{\underline{U}}(\hat{\underline{U}}_{t-1}) - f_{\underline{U}}(\underline{U}_{t-1})| \leq \sup_{\underline{u}}|\hat{f}_{\underline{U}}(\underline{u}) - f_{\underline{U}}(\underline{u})| + \max_{d_2+1\leq t\leq n}|f_{\underline{U}}(\hat{\underline{U}}_{t-1}) - f_{\underline{U}}(\underline{U}_{t-1})| = O_p(\upsilon_{2n} + h_2^2) + O_p(\upsilon_{1n} + h_1^{q+1}) = o_p(b)$. Equations (A.16) and (A.17) are also true if the trimming parameter is replaced by $b - c(\upsilon_{2n} + h_2^2 + \upsilon_{1n} + h_1^{q+1}) \equiv b - c_n$. Let $\gamma_n = (n-d_2)^{-1/2}h_0^{-d_1/2}$. Then for any finite $\epsilon > 0$ and large enough n,

$$
\begin{aligned}
P\Bigg\{\Bigg|\gamma_n \sum_{t=d_2+1}^{n} \mathcal{K}_{0,t,0}(x)\{\hat{I}_t G'(W_t;\hat{Z}_t) - G'(W_t;\bar{Z}_t)\}\Bigg| \\
> \epsilon, |f_{\underline{U}}(\underline{U}_{t-1}) - f_{\underline{U}}(\underline{U}_{t-1})| < c_n \quad \forall t\Bigg\} \\
\leq P\Bigg\{\Bigg|\gamma_n \sum_{t=d_2+1}^{n} \mathcal{K}_{0,t,0}(x)\{I_t G'(W_t;\hat{Z}_t) - G'(W_t;\bar{Z}_t)\}\Bigg| \\
> \epsilon/2, |f_{\underline{U}}(\underline{U}_{t-1}) - f_{\underline{U}}(\underline{U}_{t-1})| < c_n \quad \forall t\Bigg\}.
\end{aligned}
$$

Nevertheless, $P\{\max_{d_2+1\leq t\leq n}|f_{\underline{U}}(\underline{U}_{t-1}) - f_{\underline{U}}(\underline{U}_{t-1})| > c_n \ \forall t\} \to 0$ as $n \to \infty$. So we conclude that (A.16) is also true when we substitute I_t with $\hat{I}_t$. A similar argument holds for (A.17).

This proves (A.14). The proof of (A.15) follows by another application of (A.13) and arguments similar to the preceding discussion. ■