

Local polynomial estimation of nonparametric simultaneous equations models

Liangjun Su^a, Aman Ullah^{b,*}

^a*School of Economics, Singapore Management University, Singapore 178903, Singapore*

^b*Department of Economics, University of California, Riverside, CA 92521-0427, USA*

Available online 26 January 2008

Abstract

We define a new procedure for consistent estimation of nonparametric simultaneous equations models under the conditional mean independence restriction of Newey et al. [1999. Nonparametric estimation of triangular simultaneous equation models. *Econometrica* 67, 565–603]. It is based upon local polynomial regression and marginal integration techniques. We establish the asymptotic distribution of our estimator under weak data dependence conditions. Simulation evidence suggests that our estimator may significantly outperform the estimators of Pinkse [2000. Nonparametric two-step regression estimation when regressors and errors are dependent. *Canadian Journal of Statistics* 28, 289–300] and Newey and Powell [2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565–1578].

© 2008 Elsevier B.V. All rights reserved.

JEL classification: C13; C14; C22

Keywords: Additive nonparametric regression; Instrumental variables; Local polynomial regression; Structural models

1. Introduction

There are many occasions in econometrics where knowledge of the structural relationship among dependent variables is required to answer questions of interest. As Newey et al. (1999) put it, structural estimation is important because we need it to account correctly for endogeneity that comes from individual choice or market equilibrium. Often, economic theory does not imply tight functional form specifications for structural models so that it is useful to consider nonparametric structural models and their estimation.

Nonparametric structural models were first considered in Roehrig (1988), and Newey and Powell (1989), among others. Assuming that the errors are independent of the instruments, Roehrig (1988) gives identification results for a system of equations. Under the weaker condition that the disturbance has conditional mean zero given the instruments, Newey and Powell (1989) consider both identification and estimation problems. Following this latter paper, Brown and Matzkin (1998), Newey et al. (1999), Pinkse (2000), Darolles et al. (2000), Horowitz (2005), and Imbens and Newey (2006) consider identification and

*Corresponding author. Tel.: +1 909 827 1591.

E-mail addresses: lsu@gsm.pku.edu.cn (L. Su), aman.ullah@ucr.edu (A. Ullah).

estimation of different nonparametric models under various restrictions. For example, Pinkse (2000) considers estimation of a structural model by assuming the independence between the instrumental variable and the error terms in both the structural model and reduced model.

In this paper, we consider the regression model of Newey et al. (1999):

$$\begin{cases} Y = g(X, Z_1) + \varepsilon, & Z = (Z'_1, Z'_2)', \\ X = h(Z) + U, & E(U|Z) = 0, \quad E[\varepsilon|Z, U] = E[\varepsilon|U], \end{cases} \quad (1.1)$$

where Y is an observable scalar random variable, g denotes the true, unknown structural function of interest, X is $d_x \times 1$ vector of explanatory variables, Z_1 and Z_2 are $d_1 \times 1$ and $d_2 \times 1$ vectors of instrumental variables, $h \equiv (h_1, \dots, h_{d_x})'$ is a $d_x \times 1$ vector of functions of the instruments Z , and U and ε are disturbances. We are interested in estimating g and its derivatives consistently.

Newey et al. (1999) show that g is identified up to an additive constant if there is no functional relationship between (X, Z_1) and U .¹ They employ series approximations that exploit the additive structure of the model and propose a two-stage estimator of g . They derive consistency and asymptotic normality results for functionals of their estimator. By contrast, Newey and Powell (2003) study the estimation of g in (1.1) under the restrictions that $E[\varepsilon|Z] = 0$ and $E(U|Z) = 0$, and give identification results. Based on sieve approximations, they propose an estimator of g that is a nonparametric analog to the familiar two-stage least squares (2SLS) estimator for linear models with endogenous regressors and prove a consistency result for their estimator. Nevertheless, neither the consistency rate nor the normality of the proposed estimator is obtained.

In this paper, we propose a local polynomial procedure for estimating $g(\cdot)$ in (1.1) that is based on the following observation:

$$\begin{aligned} E[Y|X, Z, U] &= g(X, Z_1) + E[\varepsilon|X, Z, U] \\ &= g(X, Z_1) + E[\varepsilon|X - h(Z), Z, U] \\ &= g(X, Z_1) + E[\varepsilon|Z, U] \\ &= g(X, Z_1) + E[\varepsilon|U]. \end{aligned} \quad (1.2)$$

Thus it follows from the law of iterated expectations that

$$m(X, Z_1, U) \equiv E[Y|X, Z_1, U] = g(X, Z_1) + E[\varepsilon|U]. \quad (1.3)$$

Like Newey et al. (1999), our procedure can estimate $g(\cdot)$ consistently up to an additive constant that explores the additive structure in the above model. If the realizations of U were observable, the model is simply the additive model widely studied in the literature. One can adopt the marginal integration technique (e.g., Linton and Härdle, 1996) to estimate g . Further, Linton (1997, 2000) defines a two-step estimator for generalized additive nonparametric regression models that is more efficient than the marginal integration estimator. Thus one can go one step further to obtain a more efficient estimator of g . Here, because the realizations of U are not observed, we replace them by the residuals obtained by regressing X on Z nonparametrically. We show that such a replacement does not affect the first-order asymptotic property of the resulting estimator.²

Like Pinkse (2000) we will allow for weak data dependence in our estimation procedure. A typical application is the estimation of the Kuznets curve using time series data, which relates economic growth to economic inequality. Lundberg and Squire (2003) emphasize the simultaneous evolution of growth and inequality. Farrell et al. (1999) model the structural relationship between lottery sales and the expected value using time series data. For more studies on the structural time series models, see Zellner and Palm (2004). Given the unknown nonlinear relationship between the variables of interest in all these works, one may like to model them nonparametrically.

The rest of the paper is organized as follows. We introduce our estimator and its asymptotic distribution theory in Section 2. We report some Monte Carlo simulation results in Section 3. Section 4 concludes. All proofs are given in the Appendix.

¹The conditional moment restrictions in (1.1) are much weaker than those imposed in Pinkse (2000).

²In a different but relevant context, Li and Wooldridge (2002) show that \sqrt{n} -consistent estimation results of the finite dimensional parameter in a partially linear model can be generalized to the case of generated regressors with weakly dependent data.

2. A local polynomial estimator and its asymptotic distribution

In this section, we propose an estimator of the nonparametric object $g(\cdot)$ in (1.1) based upon the local polynomial procedure and then study its asymptotic properties.

2.1. A local polynomial estimator

Let Q be a deterministic weighting function with

$$\int_{\mathbb{R}^{d_x}} dQ(u) = 1,$$

where integrals are in the Stieljes sense. We allow for both discrete and continuous Q . Further, let q be the density of Q with respect to either the Lebesgue measure or a counting measure in \mathbb{R}^{d_x} . For simplicity, we assume that q is bounded on its support. We consider

$$g_Q(x, z_1) \equiv \int m(x, z_1, u) dQ(u). \tag{2.1}$$

Given the additive structure in (1.3), $g_Q(x, z_1) = g(x, z_1) + c$, where $c = \int E[\varepsilon|U = u] dQ(u)$. Therefore, $g_Q(x, z_1)$ is, up to an additive constant, the function of our interest. If we assume $E[\varepsilon] = 0$, one potential choice for Q is the distribution function, F_u , of U . Then $c = 0$ so that $g_{F_u}(x, z_1) = g(x, z_1)$.

Suppose that we have a sample $\{(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)\}$, where $X_t \in \mathbb{R}^{d_x}$, $Y_t \in \mathbb{R}$, and $Z_t = (Z_{1t}, Z_{2t}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ from the nonparametric regression model (1.1). The objective is to estimate consistently $g_Q(x, z_1)$ at some interior point (x, z_1) and to provide asymptotic normality result for the estimator. To explore the additive structure of the regression function in Eq. (1.3), we propose the following estimation procedure:

1. Obtain a consistent estimator of $h(Z_t)$ by local p_1 th-order smoothing $X_t \equiv (X_{1t}, \dots, X_{d_x t})'$ on Z_t with kernel K_1 and bandwidth sequence $b_1 = b_1(n)$. Denote the estimates as $\hat{h}(Z_t) \equiv (\hat{h}_1(Z_t), \dots, \hat{h}_{d_x}(Z_t))'$ and calculate the estimated residuals $\hat{U}_t \equiv (\hat{U}_{1t}, \dots, \hat{U}_{d_x t})'$, where $\hat{U}_{it} \equiv X_{it} - \hat{h}_i(Z_t)$ for $i = 1, \dots, d_x$ and $t = 1, \dots, n$.³
2. Obtain a consistent estimator of $m(x, z_1, u)$ by local p_2 th-order smoothing Y_t on X_t, Z_{1t} , and \hat{U}_t with kernel K_2 and bandwidth sequence $b_2 = b_2(n)$. Denote the estimates as $\hat{m}(x, z_1, u)$.
3. Estimate $g(x, z_1)$ consistently up to an additive constant by

$$\hat{g}_Q(x, z_1) = \int \hat{m}(x, z_1, u) dQ(u). \tag{2.2}$$

If we are interested in estimating $g(x, z_1)$ consistently but not its partial derivatives, we can replace Step 2 by:

- 2*. Obtain a consistent estimator of $m(x, z_1, u)$ by the NW kernel smoothing method with corresponding kernel K_2 and bandwidth sequence $b_2 = b_2(n)$. Denote the estimate as $\hat{m}(x, z_1, u)$.

In this paper, we give an asymptotic analysis based on the local polynomial procedure in both the first and second steps. See Fan (1992) and Fan and Gijbels (1996) for discussions on the attractive properties of local polynomials. For the data set $\{X_t, Z_t\}_{t=1}^n$, the p_1 th-order local polynomial regression of X_{it} , $i = 1, \dots, d_x$, on Z_t can be obtained from the multivariate weighted least squared criterion:

$$Q_{n,i}(\theta^{(i)}) \equiv nb_1^{-d_z} \sum_{t=1}^n K_1((Z_t - z)/b_1) \left[X_{it} - \sum_{0 \leq |j| \leq p_1} \theta_j^{(i)}(Z_t - z)^j \right]^2, \tag{2.3}$$

³Noting that we need to obtain $\hat{h}(Z_t)$ and $\hat{U}_t = X_t - \hat{h}(Z_t)$ for all $t = 1, \dots, n$, we can tell that the Nadaraya–Watson (NW) kernel estimator is less desirable than the local polynomial estimator: we need to correct the boundary bias in case $\{Z_t\}$ is compactly supported, or apply some trimming techniques otherwise. See Pagan and Ullah (1999) for more comparisons between the two types of estimators.

where K_1 is a nonnegative kernel function on \mathbb{R}^{d_z} with $d_z = d_1 + d_2$, and $b_1 = b_1(n)$ is a scalar bandwidth sequence. Here, we use the notation of Masry (1996a, b)⁴:

$$\mathbf{j} = (j_1, \dots, j_{d_z})', \quad |\mathbf{j}| = \sum_{i=1}^{d_z} j_i, \quad z^{\mathbf{j}} = \prod_{i=1}^{d_z} z_i^{j_i} \quad \text{and} \quad \sum_{0 \leq |\mathbf{j}| \leq p_1} = \sum_{|\mathbf{j}|=0}^{p_1}. \tag{2.4}$$

The true value of $\theta_{\mathbf{j}}^{(i)} \equiv \theta_{\mathbf{j}}^{(i)}(z)$ corresponds to $D^{\mathbf{j}}h_i(z)/\mathbf{j}!$, where $D^{\mathbf{j}}h_i(z) = \partial^{|\mathbf{j}|}/(\partial z_1^{j_1}, \dots, \partial z_{d_z}^{j_{d_z}})(h_i(z))$, and

$$\mathbf{j}! \equiv \prod_{i=1}^{d_z} j_i!$$

Further, $\theta^{(i)} = \theta^{(i)}(z)$ is a collection of all the parameters $\theta_{\mathbf{j}}^{(i)}$, $0 \leq |\mathbf{j}| \leq p_1$, in a lexicographical order. In particular, the first element in $\theta^{(i)}$ is denoted as $\theta_0^{(i)} = \theta_0^{(i)}(z)$ throughout our presentation. Let $\widehat{h}(z) = (\widehat{\theta}_0^{(1)}, \dots, \widehat{\theta}_0^{(d_x)})'$, where $\widehat{\theta}_0^{(i)}$ is the minimizing intercept in (2.3).

Similarly, $\widehat{m}(x, z_1, u)$ is obtained as the minimizing intercept in the following objective function:

$$Q_n(\theta) \equiv nb_2^{-(2d_x+d_1)} \sum_{t=1}^n K_2((\widehat{W}_t - w))/b_2 \left[Y_t - \sum_{0 \leq |\mathbf{j}| \leq p_2} \theta_{\mathbf{j}}(\widehat{W}_t - w)^{\mathbf{j}} \right]^2, \tag{2.5}$$

where $\widehat{W}_t = (X'_t, Z'_{1t}, \widehat{U}'_t)'$ and $w = (x', z'_1, u)'$.

2.2. Assumptions

To state the main result, we make the following assumptions.

Assumptions. A1. The kernels K_i , $i = 1, 2$, satisfy

$$K_1(u) = \prod_{j=1}^{d_x} k_1(u_j), \quad K_2(w) = \prod_{j=1}^{2d_x+d_1} k_2(w_j),$$

where k_i is bounded, symmetric about zero, has compact support $[-c_i, c_i]$ and integrates to 1. For $i = 1$ and 2, the functions $H_{ij}(u) = u^{\mathbf{j}}K_i(u)$ for all j with $0 \leq |\mathbf{j}| \leq 2p_i + 1$ are Lipschitz continuous. The matrices M and \overline{M} are defined in the Appendix and are nonsingular.

A2. The process $\{(X'_t, Z'_{1t}, Y_t)', t = 0, 1, \dots\}$ is a strictly stationary α -mixing process with mixing coefficients $\alpha(j)$ satisfying

$$\sum_{j=1}^{\infty} j^2 \alpha^{\delta/(2+\delta)}(j) < \infty$$

for some $0 < \delta \leq 1$. The density f_z of Z_t , the density f_w of $W_t \equiv (X'_t, Z'_{1t}, U'_t)'$ and the joint densities $f_{t_1, \dots, t_l}(\cdot, \dots, \cdot)$ of $(W_0, W_{t_1}, \dots, W_{t_l})$ ($1 \leq l \leq 5$) are uniformly bounded and are bounded away from zero on their compact supports.

A3. $E(e_t | X_t, Z_t, U_t) = 0$ and $E[|e_t|^{2+\delta}] < \infty$, where $e_t \equiv Y_t - m(X_t, Z_{1t}, U_t)$. Also, $E[e_t^2 | X_t = x, Z_{1t} = z_1, U_t = u] = \sigma_e^2(x, z_1, u)$.

A4. The vector of functions $h = (h_1, \dots, h_{d_x})'$ are $(p_1 + 1)$ times partially continuously differentiable and the function m is $(p_2 + 1)$ times partially differentiable. The corresponding $(p_1 + 1)$ th- or $(p_2 + 1)$ th-order partial derivatives are Lipschitz continuous on their supports.

A5. The bandwidth sequences b_1 and b_2 go to zero as $n \rightarrow \infty$ and satisfy that

- (i) $nb_1^{2(p_1+1)}b_2^{d_x+d_1} \rightarrow 0$, $nb_1^{d_x(2+2\delta)/(2+\delta)}b_2^{2+(d_x+d_1)\delta/(2+\delta)} \rightarrow \infty$,
- (ii) $n^{1/2}b_1^{d_x}b_2^{2-(d_x+d_1)/2} / \log n \rightarrow \infty$, $n^{1/2}b_1^{2(p_1+1)}b_2^{(d_x+d_1)/2-2} \rightarrow 0$;
- (iii) $nb_2^{d_x+d_1+2(p_2+1)} \rightarrow \bar{c} \in [0, \infty)$.

⁴For random variables like X_t , Z_{1t} and U_t , and their values, x , z_1 , and u , we do not use boldfaced letters to denote them.

Assumptions A1–A4 parallel Conditions 1–4 in Masry (1996a) except that Assumption A4 is assumed but not listed explicitly as a condition in Masry (1996a). The stationarity condition in Assumption A2 rules out time trend in the regressors. Assumption A3 allows for conditional heteroskedasticity. The differentiability of A4 ensures Taylor expansions to appropriate orders.

Assumption A5 looks complicated and deserves some remarks. First, by requiring δ to be sufficiently small, the requirement that $nb_1^{d_x(2+2\delta)/(2+\delta)}b_2^{2+(d_x+d_1)\delta/(2+\delta)} \rightarrow \infty$ effectively reduces to $nb_1^{d_x}b_2^2 \rightarrow \infty$. Let $v_{1n} = n^{-1/2}b_1^{-d_x/2}\sqrt{\log n}$ and $v_{2n} = n^{-1/2}b_2^{-(2d_x+d_1)}\sqrt{\log n}$. Then by Masry (1996b), $\max_{1 \leq t \leq n} \|\widehat{h}(X_t) - h(X_t)\| = O_p(v_{1n} + b_1^{p_1+1})$ and $\max_{1 \leq t \leq n} |\widehat{m}(W_t) - m(W_t)| = O_p(v_{2n} + b_2^{p_2+1})$ if $\{U_1, \dots, U_n\}$ were used in forming $\widehat{m}(\cdot)$, where $\|\cdot\|$ denotes the Euclidean norm. Assumption A5(i) requires that the estimation error from the first stage estimation should be $o_p(n^{-1/2}b_2^{-(d_x+d_1)/2})$. Because $\{U_1, \dots, U_n\}$ is not observed and we use $\{\widehat{U}_1, \dots, \widehat{U}_n\}$ to approximate it in forming $\widehat{m}(\cdot)$, the approximation error is accounted in Assumption A5(ii): $(b_2^{-1}v_{1n})^2 = o(n^{-1/2}b_2^{-(d_x+d_1)/2})$, where the appearance of b_2^{-1} is due to the use of Taylor expansion in our proof. Note that A5(ii) also implies that $b_2^{-1}(v_{1n} + b_1^{p_1+1}) = o(1)$, which is used in several places in the Appendix. Assumption A5(iii) will facilitate the proof and it permits us to choose $b_2 \propto n^{-1/[2(p_2+1)+(d_x+d_1)]}$, the optimal rate of bandwidth in the local polynomial estimation of $g(\cdot)$.

It is worth mentioning that undersmoothing may or may not be required in the first stage estimation as in much of the nonparametric kernel estimation literature when a preliminary kernel estimator is used in the second stage. It depends on the sizes of d_x, d_1, d_2, p_1 , and p_2 as well. For example, when $p_1 = 3$ and $p_2 = 1$, $b_1 \propto n^{-1/[2(p_1+1)+(d_1+d_2)]}$ will suffice for a variety of combinations of d_x, d_1 and d_2 . In contrast, when $p_1 = 3$ and $p_2 = 3$, we can choose the undersmoothing bandwidth $b_1 \propto n^{-1/[2p_1+1+(d_1+d_2)]}$.

2.3. Asymptotic normality

Let $w = (x', z_1', u)'$, $d = 2d_x + d_1$, and $\mathbf{r} = (r_1, \dots, r_d)'$. Define

$$D^{\mathbf{r}}m(w) \equiv \frac{\partial^{|\mathbf{r}|}m(w)}{\partial^{r_1}w_1 \dots \partial^{r_d}w_d}, \quad |\mathbf{r}| \leq p_2 + 1.$$

The total number of derivatives $D^{\mathbf{r}}m(w)$ with $|\mathbf{r}| = r$ is given by $N_r = (r + d - 1)!/(r!(d - 1)!)$. We arrange the N_r derivatives $D^{\mathbf{r}}m(w)/r!$ as an $N_r \times 1$ column vector $m^{(r)}(w)$ in the lexicographical order. The following theorem gives the asymptotic distribution of $\widehat{g}_Q(x, z_1)$.

Theorem 2.1. *Suppose that (1.3) holds. Then under Assumptions A1–A5, we have*

$$\begin{aligned} & \sqrt{nb_2^{d_x+d_1}} \left(\widehat{g}_Q(x, z_1) - g(x, z_1) - c - b_2^{p_2+1} \left[M^{-1}B \int m^{(p_2+1)}(x, z_1, u) dQ(u) \right]_{0,0} \right) \\ & \xrightarrow{d} N(0, a_Q(x, z_1)[M^{-1}\Gamma M^{-1}]_{0,0}), \end{aligned} \tag{2.6}$$

where the matrices M, Γ and B are defined in (B.1) in the Appendix, they only depend on the kernel and not on the design, $[A]_{0,0}$ signifies the upper-left element of matrix A , and $a_Q(x, z_1) = \int q^2(u)\sigma_c^2(x, z_1, u) \times f_w^{-1}(x, z_1, u) du$.

Remark 1. The bias term in (2.6) follows by integrating the bias of the local polynomial estimator for $m(x, z_1, u)$. The fact that $\{U_t\}_{t=1}^n$ has to be estimated from the data does not have any impact on the bias correction for well chosen first stage and second stage bandwidth sequences. If we further restrict $nb_2^{d_x+d_1+2(p_2+1)} \rightarrow 0$, which is possible for sufficiently large p_2 , then there is no need to correct the bias.

Remark 2. Our procedure resembles many other kernel-based multi-stage nonparametric procedures (e.g., Linton, 2000; Xiao et al., 2003; Su and Ullah, 2006) in that the first stage estimation does not contribute to the asymptotic variance of the final stage estimators. Nevertheless, this is not generally the case in parametric estimation problems, unless some orthogonality conditions are satisfied. Nor is it the general case for

two-stage nonparametric series estimators (e.g., Newey et al., 1999), where the first-stage estimation contributes to the asymptotic variance of the second-stage estimators.⁵

Remark 3. For the following reasons, it is not easy to compare theoretically the relative efficiency of our estimator with the earlier estimators proposed by Pinkse (2000), Newey et al. (1999) and Newey and Powell (2003). First, Pinkse assumes that Z_t is independent of ε_t and U_t , which is stronger than the conditional mean independence assumption of Newey et al. (1999) and Newey and Powell (2003). The latter assumption is also imposed in this paper. Second, all the earlier estimators are based upon two-stage series approximations under some smoothness conditions for $m(\cdot)$ and $h(\cdot)$ that control the rate at which polynomials or splines approximate the true unknown function and depend on the dimensions $d_x + d_1$ and d_z of (x, z_1) and z , respectively. As a result, only mean square and uniform consistency of the estimator of $g(\cdot)$ were established and no asymptotic normality result was obtained. In contrast, we have established the asymptotic normality of our estimator under some smoothness conditions for $m(\cdot)$ and $h(\cdot)$ that depend on the orders of local polynomials. Third, in all cases, $(X_t, Z_t, \varepsilon_t, U_t)$ is assumed to be continuously distributed and the density of (X_t, Z_t, U_t) is bounded away from zero on its support. Fourth, like Pinkse (2000), we allow weak data dependence, whereas Newey et al. (1999) and Newey and Powell (2003) do not. In the next section we will provide the comparison of relative efficiency of various estimators through Monte Carlo experiments.

Remark 4. The asymptotic normal distribution given by Theorem 2.1 can be used to calculate pointwise confidence intervals for the estimator described here. To do this we require a consistent estimate of the asymptotic variance. The procedure is standard and we omit it for brevity.

Under the assumption that $E[\varepsilon] = 0$, $g(x, z_1)$ can be fully identified. One can choose $Q(u) = F_u(u)$, the distribution function of U . Since F_u is unknown in practice, we can replace it by its empirical analog,

$$\widehat{F}_u(u) \equiv n^{-1} \sum_{t=1}^n 1\{\widehat{U}_t \leq u\},$$

and define

$$\widehat{g}(x, z_1) = \int \widehat{m}(x, z_1, u) d\widehat{F}_u(u) = n^{-1} \sum_{t=1}^n \widehat{m}(x, z_1, \widehat{U}_t).$$

Here, $1\{\cdot\}$ is the usual indicator function. Even though Theorem 2.1 allows the weighting function $Q(u)$ to be discrete, we cannot apply it directly to obtain the asymptotic distribution of $\widehat{g}(x, z_1)$ because $\widehat{F}_u(u)$ depends on the data. Instead, we prove the following theorem in Appendix B by applying some moment bounds for the third-order U -statistics.

Theorem 2.2. *Suppose that (1.3) holds and $E[\varepsilon] = 0$. Then under Assumptions A1–A5, we have*

$$\begin{aligned} & \sqrt{nb_2^{d_x+d_1}} \left(\widehat{g}(x, z_1) - g(x, z_1) - b_2^{p_2+1} \left[M^{-1} B \int m^{(p_2+1)}(x, z_1, u) dF_u(u) \right]_{0,0} \right) \\ & \xrightarrow{d} N(0, a(x, z_1) [M^{-1} \Gamma M^{-1}]_{0,0}), \end{aligned}$$

where M , Γ and B are defined in (B.1), $[A]_{0,0}$ signifies the upper-left element of matrix A , and $a(x, z_1) = \int f_u^2(u) \sigma_\varepsilon^2(x, z_1, u) f_w^{-1}(x, z_1, u) du$ with f_u being the density of F_u .

⁵In semiparametric models with generated regressors, the first stage estimation also has nonnegligible impact on the asymptotic variance of the parametric estimator. See Li and Wooldridge (2002).

Remark 5. Theorem 2.2 implies that the asymptotically optimal bandwidth b_2 that minimizes the asymptotic, integrated mean squared error (AIMSE) of \hat{g} is given by

$$b_2^* = \left\{ \frac{(d_x + d_1) \int \sigma_e^2(x, z_1, u) \frac{f_u^2(u) f(x, z_1)}{f_w(x, z_1, u)} du dx dz_1 [M^{-1} \Gamma M^{-1}]_{0,0}}{2(p_2 + 1) \int \{ [M^{-1} B \int m^{(p_2+1)}(x, z_1, u) dF_u(u)]_{0,0} \}^2 f(x, z_1) dx dz_1} \right\}^{1/(2(p_2+1)+d_x+d_1)} \times n^{-1/(2(p_2+1)+d_x+d_1)}, \tag{2.7}$$

where $f(x, z_1)$ is the density of (X_i, Z_{1i}) . By Stone (1980), if the $(p_2 + 1)$ th-order derivatives of $g(x, z_1)$ are Lipschitz continuous, the optimal rate of convergence of nonparametric estimators for g is $O(n^{-(p_2+1)/[2(p_2+1)+d_x+d_1]})$ by choosing bandwidth b_2 proportional to $O(n^{-1/[2(p_2+1)+d_x+d_1]})$. Obviously, this optimal rate of convergence tends to zero rather slowly if $d_x + d_1$ is large and p_2 is small. Naturally for large values of $d_x + d_1$, higher-order local polynomial estimation is preferable if g has higher order of smoothness.

If we choose b_2 that has the optimal rate given in (2.7), Assumption A5 implies that we can choose b_1 such that $b_1 \propto n^{-\alpha}$, where

$$\underline{\alpha} \equiv \max \left\{ \frac{p_2 + 1}{p_1 + 1}, \frac{p_2 + 3}{2(p_1 + 1)} \right\} < \alpha \gamma^* < \frac{p_2 + d_x + d_1 - 1}{d_z} \equiv \bar{\alpha} \tag{2.8}$$

and $\gamma^* = 2(p_2 + 1) + d_x + d_1$.

Remark 6. Similarly to the case of Horowitz (1999, 2001), implementing the estimator \hat{g} requires methods for choosing the values of two bandwidth parameters b_1 and b_2 . The bandwidth b_1 does not affect the asymptotic distribution of \hat{g} as long as it satisfies Assumption A5, whereas b_2 does. So in the next subsection we follow Horowitz to describe a systematic method for choosing b_2 and a rule of thumb for choosing b_1 :

$$b_1 = b_2 n^{-(\bar{\alpha} - \underline{\alpha})/(2\gamma^*)}. \tag{2.9}$$

2.4. Bandwidth selection

As is the case for all nonparametric curve estimation, the bandwidth parameter plays an essential role in practice. It is desirable to have a reliable data-driven and yet easily implementable bandwidth selection procedure.

For the estimation of the general additive models, there are several ways to choose the bandwidth parameters. The most commonly used one relies on cross-validation or one of its approximations (Hastie and Tibshirani, 1990). For example, Nielsen and Sperlich (2005) apply the leave-one-out least squares cross-validation (LSCV) in the smooth backfitting context. A second way is to base on penalized sums of squared residuals (e.g., Mammen and Park, 2005 for smooth backfitting), or more generally the nonparametric AIC criterion of Cai (2002) for marginal integration. Nevertheless, all of these methods aim to minimize the mean squared error of estimating the whole regression function ($m(x, z_1, u)$ here) but not any particular component of the regression function ($g(x, z_1)$ in particular). A third way is the subsampling method of Horowitz (2001) who chooses bandwidth sequences by minimizing a sample analog of the AIMSE of the nonparametric estimate. Unfortunately, there is no theory that justifies the choice of subsample size. In addition, we find through simulations that this method is extremely time-consuming so that even Horowitz (1999) did not use this method in his simulations. So we believe that we need to find other method that is appropriate in the current setting.

In this paper we describe a “plug-in” method for selecting the bandwidth b_2 and a rule of thumb for choosing b_1 . Plug-in bandwidth selection methods are well known in kernel smoothing, kernel regression, and local polynomial regression. They have also been widely applied in additive models. See Opsomer and Ruppert (1998) in the back-fitting context, and Severance-Lossin and Sperlich (1999), Sperlich et al. (1999), and Yang et al. (2003) for derivative estimation and testing in the marginal integration context.

In the following, we focus on the case where $p_2 = 1$ so that local linear regression is used in the second stage. In this case, the formula for b_2^* reduces to

$$b_2^* = \left\{ \frac{\gamma_{02}^{d_x+d_1} (d_x + d_1) \int \sigma_e^2(x, z_1, u) \frac{f_u^2(u) f(x, z_1)}{f_w(x, z_1, u)} du dx dz_1}{\gamma_{21}^2 \int \left(\sum_{j=1}^{d_x+d_1} \frac{\partial^2 g(v)}{\partial v_j^2} \right)^2 f(x, z_1) dx dz_1} \right\}^{1/(4+d_x+d_1)} \times n^{-1/(4+d_x+d_1)}, \tag{2.10}$$

where $v = (x', z_1')$ and $\gamma_{ij} = \int_{\mathbb{R}} t^i k_2(t)^j dt$, $i = 0, 1, 2, j = 1, 2$.

To obtain a plug-in estimator of b_2^* , we need to estimate both the denominator and numerator inside the curly bracket. First, we run a preliminary p_1 th-order local polynomial regression of X_i on Z_i with kernel K_1 and bandwidth b_3 , where b_3 is chosen by the leave-one-out LSCV. Denote the residual sequence from this regression as $\{\tilde{U}_i\}$. Second, let $V'_i = (X'_i, Z'_i)$ and $\tilde{W}'_i = (V'_i, \tilde{U}'_i)$. We estimate

$$\int \left(\sum_{j=1}^{d_x+d_1} \frac{\partial^2 g(v)}{\partial v_j^2} \right)^2 f(x, z_1) dx dz_1$$

by

$$\tilde{c}_{n1} = n^{-1} \sum_{i=1}^n \left\{ \sum_{j=1}^{d_x+d_1} \frac{\partial^2 \hat{g}(V_i)}{\partial v_j^2} \right\}^2, \tag{2.11}$$

where $\partial^2 \hat{g}(V_i)/\partial v_j^2$ is a consistent estimator of $\partial^2 g(V_i)/\partial v_j^2$ by the third-order local polynomial regression of Y_i on X_i, Z_{1i} , and \tilde{U}_i with product kernel K_2 and bandwidth parameter b_4 . Denote the residual sequence from this regression as $\{\tilde{\varepsilon}_i\}$. We estimate

$$\int \sigma_e^2(x, z_1, u) \frac{f_u^2(u) f(x, z_1)}{f_w(x, z_1, u)} du dx dz_1 = E \left[\sigma_e^2(X_i, Z_{1i}, U_i) \frac{f_u^2(U_i) f(X_i, Z_{1i})}{f_w^2(X_i, Z_{1i}, U_i)} \right] \tag{2.12}$$

by

$$\tilde{c}_{n2} = \frac{1}{n} \sum_{i=1}^n \tilde{\varepsilon}_i^2 \tilde{a}_i, \tag{2.13}$$

where

$$\tilde{a}_i = \left[\frac{n^{-1} b_5^{-d_x} \sum_{j=1}^n K_3((\tilde{U}_i - \tilde{U}_j)/b_5)}{n^{-1} b_5^{-(2d_x+d_1)} \sum_{j=1}^n K_2((\tilde{W}_i - \tilde{W}_j)/b_5)} \right]^2 \left[n^{-1} b_5^{-(d_x+d_1)} \sum_{j=1}^n K_4((V_i - V_j)/b_5) \right] \tag{2.14}$$

is an estimator of the unknown density quantities in (2.12) evaluated at \tilde{W}_i by using the product kernels K_3 and K_4 and bandwidth b_5 . Consequently, we estimate b_2^* by

$$\hat{b}_2 = \left\{ \frac{\gamma_{02}^{d_x+d_1} (d_x + d_1) \tilde{c}_{n2}}{\gamma_{21}^2 \tilde{c}_{n1}} \right\}^{1/(4+d_x+d_1)} \times n^{-1/(4+d_x+d_1)}. \tag{2.15}$$

In the simulations, we shall choose each kernel to be a multivariate kernel of the same univariate kernel. In addition, to obtain \hat{b}_2 , we use the leave-one-out LSCV method to choose the preliminary bandwidth sequences b_3, b_4 and b_5 . Once we obtain \hat{b}_2 , we follow the lead of Horowitz (2001) and set the data-driven choice of b_1 to be

$$\hat{b}_1 = \hat{b}_2 n^{-(\bar{\alpha}^* - \underline{\alpha}^*)/(2\gamma^*)}, \tag{2.16}$$

where $\bar{\alpha}^*, \underline{\alpha}^*$ and γ^* are defined in (2.8).

3. Monte Carlo simulation

We assume $E[\varepsilon] = 0$ and investigate the proposed estimator $\hat{g}(x, z_1)$ on simulated data and compare it with three other estimators available in the literature. The first one is the naive local linear estimator that ignores the issue of endogeneity, the second one is the two-stage series estimator of Pinkse (2000) (see also Newey et al., 1999) and the last one is that of Newey and Powell (2003).

3.1. Data generating processes

We will consider four data generating processes (DGPs). The first one is adopted from Newey and Powell (2003) who consider a nonparametric simultaneous equations model in the simple i.i.d. setting. DGP1 below specifies how the data are generated:

$$\text{DGP1: } \begin{cases} Y_t = g(X_t) + \varepsilon_t = \log(|X_t - 1| + 1)\text{sgn}(X_t - 1) + \varepsilon_t, \\ X_t = h(Z_t) + U_t = Z_t + U_t, \end{cases}$$

where the errors ε_t and U_t and the instrument Z_t are generated as

$$\begin{pmatrix} \varepsilon_t \\ U_t \\ Z_t \end{pmatrix} \sim \text{i.i.d. N} \left(0, \begin{pmatrix} 1 & \theta & 0 \\ \theta & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right), \tag{3.1}$$

in which $\theta = 0.2, 0.5,$ and 0.8 indicate weak, middle and strong endogeneity, respectively. It is easy to verify that the above design satisfies the identification conditions in Pinkse (2000), Newey and Powell (2003) and this paper as well: $E(\varepsilon_t|Z_t) = 0, E(U_t|Z_t) = 0,$ and $E(\varepsilon_t|U_t, Z_t) = E(\varepsilon_t|U_t) = 0.5U_t.$ Due to the unbounded support of the normal distribution, the regressors X_t and Z_t in the above structural model do not have compact support. Hence Assumption A2 is violated in this case. The simulation result will indicate the robustness of various estimators against noncompact support.

In DGPs 2–4 we generate (Y_t, X_t) according to

$$\text{DGP2: } \begin{cases} Y_t = 1 + 2 \exp(X_t)/(1 + \exp(X_t)) + \varepsilon_t, \\ X_t = Z_t + U_t, \end{cases}$$

$$\text{DGP3: } \begin{cases} Y_t = \cos(X_t) + \varepsilon_t, \\ X_t = Z_t + \sin(0.2Z_t) + U_t, \end{cases}$$

$$\text{DGP4: } \begin{cases} Y_t = 2\Phi(X_t) + \varepsilon_t, \\ X_t = \log(0.1 + Z_t^2) + U_t, \end{cases}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal, and the errors ε_t and U_t and the instrument Z_t are generated as

$$\varepsilon_t = \theta w_t + 0.3v_{Y_t}, \quad U_t = 0.5w_t + 0.2v_{X_t}, \quad Z_t = 1 + 0.5Z_{t-1} + 0.5v_{Z_t}, \tag{3.2}$$

in which $v_{Y_t}, v_{X_t}, v_{Z_t}, w_t$ are i.i.d. sum of 48 independent random variables each uniformly distributed on $[-0.25, 0.25]$. According to the central limit theorem, we can treat $v_{Y_t}, v_{X_t}, v_{Z_t}, w_t$ as being nearly standard normal random variables but with compact support $[-12, 12]$. We consider three configurations: $\theta = 0.2, 0.5, 0.8.$ As θ increases, the correlation between ε_t and X_t increases and the problem of simultaneity is magnified. It is also easy to verify that the above design satisfies the identification conditions in Pinkse (2000), Newey and Powell (2003) and this paper as well.

3.2. Different estimators

We consider four different estimators of g in the simulation.

The first estimator is termed as the naive estimator that is obtained via the local linear regression of Y_t on X_t directly, ignoring the endogeneity problem. The bandwidth is chosen by the leave-one-out LSCV method. Clearly, this estimator serves as the benchmark estimator, which should be outperformed by a reasonably good nonparametric instrumental variable estimator.

The second estimator is the two-stage series estimator of Pinkse (2000). To obtain the estimator, we follow Pinkse (2000) and use Legendre polynomials, which constitute an orthogonal system of functions on $[-1, 1]$. The numbers of terms of approximation for the first step and second step regressions are denoted as L_1 and L_2 , respectively, which are chosen according to Pinkse (2000).

The third estimator is the two-stage nonparametric estimator of Newey and Powell (2003). To obtain their estimator, we follow their recommendations carefully. In particular, we choose the Hermite series approximation with two values of the constraint parameter in their paper: $L = 5$ and 50.

The fourth estimator is our two-stage local polynomial estimator. To obtain our estimator \hat{g} , we set $p_1 = 3$ and 1, and use the second-order Epanechnikov kernel (standardized to have unit variance):

$$k_i(u) = \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5}u^2\right) 1_{\{|u| \leq \sqrt{5}\}}, \quad i = 1, 2.$$

For the bandwidth sequences (b_1, b_2) , we choose b_2 based upon the plug-in formula in (2.15) and b_1 based upon the rule of thumb in (2.16).

3.3. Simulation results

For each DGP and estimator, we consider two sample sizes: $n = 100$ and 400. We consider 1000 and 500 repetitions for $n = 100$ and 400, respectively. We summarize the simulation results for DGPs 1–4 in Tables 1–4, respectively. The third and fifth columns of Tables 1–4 report the means of the root mean squared errors (RMSEs) of different estimators obtained by averaging across the realized values of X and the 1000 or 500 repetitions. The fourth and sixth columns of Tables 1–4 report the medians of the RMSEs of different estimators obtained by averaging across the realized values of X only. The last two columns of Tables 1–4 give the ratios of the RMSEs for $n = 400$ over those for $n = 100$, which gives some hint on the rates of convergence for the corresponding estimators.

We first summarize some main findings from Table 1. (a) For each sample size and endogeneity parameter value (θ) under consideration, the naive estimator is outperformed by all other three estimators in terms of mean or median RMSEs. This indicates the need to account for endogeneity. (b) In terms of mean RMSEs, our estimator outperforms Pinkse's (2000) estimator in most cases and performs almost equally well as Newey and Powell's (2003) estimator when $n = 100$. But as the sample size increases to $n = 400$, our estimator is the best. (c) In terms of median RMSEs, our estimator outperforms all other estimators for both sample sizes. (d) For Pinkse's estimator, the choice of L_2 has an important effect on the performance of the estimator. The larger the L_2 , the worse the performance it has. (e) For each scenario, the RMSE declines as the sample size is quadrupled, at a rate less than the parametric \sqrt{n} -rate, as would be expected. The RMSE of our estimator declines at a faster rate than that of Newey and Powell (2003) and that of Pinkse (2000).

Tables 2–4 indicate some interesting and quite different findings. First, we find that both Pinkse's (2000) and Newey and Powell's (2003) estimators can be outperformed by the naive estimator, whereas our estimator always dominates the naive estimator in terms of RMSE. For DGP2, both Pinkse's and Newey and Powell's estimators are beaten by the naive estimator. For DGP3, Newey and Powell's estimator is also beaten by the naive estimator. For DGP4, in order for Newey and Powell's estimator to outperform the naive estimator, a larger value of the constraint parameter L is demanded. Second, for Pinkse's estimator, the choice of L_2 has an important effect on the performance of the estimator and its effect is not predictable, for example, for DGP3, the larger the L_2 , the better its performance when the endogeneity is weak ($\theta = 0.2$). But this is not true when

Table 1
RMSEs and convergence rate comparison for DGPI

Parameter	Estimators	$n = 100$		$n = 400$		Ratio	
		Mean (1)	Median (2)	Mean (3)	Median (4)	(3)/(1)	(4)/(2)
$\theta = 0.2$	<i>Naive</i>	0.534	0.532	0.524	0.525	0.981	0.987
	<i>Pinkse (2000)</i>						
	$(L_1, L_2) = (2, 2)$	0.274	0.267	0.223	0.221	0.814	0.828
	$(L_1, L_2) = (2, 4)$	0.318	0.306	0.224	0.222	0.704	0.725
	$(L_1, L_2) = (2, 6)$	0.430	0.359	0.240	0.237	0.558	0.660
	$(L_1, L_2) = (4, 2)$	0.273	0.267	0.223	0.220	0.817	0.824
	$(L_1, L_2) = (4, 4)$	0.315	0.304	0.224	0.222	0.711	0.730
	$(L_1, L_2) = (4, 6)$	0.421	0.357	0.241	0.237	0.572	0.664
	$(L_1, L_2) = (6, 2)$	0.273	0.265	0.223	0.220	0.817	0.830
	$(L_1, L_2) = (6, 4)$	0.315	0.303	0.224	0.223	0.711	0.736
	$(L_1, L_2) = (6, 6)$	0.411	0.356	0.241	0.237	0.586	0.666
	<i>Newey and Powell (2003)</i>						
	$L = 5$	0.250	0.224	0.185	0.179	0.740	0.799
	$L = 50$	0.292	0.267	0.215	0.202	0.736	0.757
	<i>Su and Ullah</i>	0.306	0.217	0.150	0.140	0.490	0.645
	$\theta = 0.5$	<i>Naive</i>	0.533	0.533	0.523	0.524	0.981
<i>Pinkse (2000)</i>							
$(L_1, L_2) = (2, 2)$		0.270	0.263	0.221	0.218	0.819	0.829
$(L_1, L_2) = (2, 4)$		0.307	0.300	0.219	0.216	0.713	0.720
$(L_1, L_2) = (2, 6)$		0.430	0.342	0.241	0.232	0.560	0.678
$(L_1, L_2) = (4, 2)$		0.270	0.265	0.221	0.218	0.819	0.823
$(L_1, L_2) = (4, 4)$		0.305	0.300	0.220	0.216	0.721	0.720
$(L_1, L_2) = (4, 6)$		0.414	0.343	0.240	0.232	0.580	0.676
$(L_1, L_2) = (6, 2)$		0.270	0.265	0.222	0.219	0.822	0.826
$(L_1, L_2) = (6, 4)$		0.307	0.298	0.220	0.216	0.717	0.725
$(L_1, L_2) = (6, 6)$		0.425	0.344	0.241	0.233	0.567	0.677
<i>Newey and Powell (2003)</i>							
$L = 5$		0.249	0.226	0.184	0.180	0.739	0.796
$L = 50$		0.288	0.262	0.209	0.198	0.726	0.756
<i>Su and Ullah</i>		0.269	0.211	0.144	0.138	0.535	0.654
$\theta = 0.8$		<i>Naive</i>	0.532	0.534	0.521	0.520	0.979
	<i>Pinkse (2000)</i>						
	$(L_1, L_2) = (2, 2)$	0.261	0.256	0.220	0.218	0.843	0.852
	$(L_1, L_2) = (2, 4)$	0.276	0.269	0.213	0.212	0.772	0.788
	$(L_1, L_2) = (2, 6)$	0.349	0.298	0.222	0.220	0.636	0.738
	$(L_1, L_2) = (4, 2)$	0.262	0.257	0.220	0.218	0.840	0.848
	$(L_1, L_2) = (4, 4)$	0.279	0.272	0.214	0.213	0.767	0.783
	$(L_1, L_2) = (4, 6)$	0.348	0.303	0.223	0.221	0.641	0.729
	$(L_1, L_2) = (6, 2)$	0.264	0.259	0.220	0.220	0.833	0.849
	$(L_1, L_2) = (6, 4)$	0.284	0.275	0.215	0.214	0.757	0.778
	$(L_1, L_2) = (6, 6)$	0.355	0.304	0.224	0.223	0.631	0.734
	<i>Newey and Powell (2003)</i>						
	$L = 5$	0.258	0.230	0.184	0.181	0.713	0.787
	$L = 50$	0.296	0.270	0.206	0.196	0.696	0.726
	<i>Su and Ullah</i>	0.250	0.206	0.144	0.136	0.576	0.660

the endogeneity is large ($\theta = 0.8$). DGP4 is another example. A choice of $L_2 = 4$ tends to have a better performance than $L_2 = 2$ or 6. (c) For each scenario, the RMSE of our estimator declines as the sample size is quadrupled, at a rate somewhat slower than the parametric \sqrt{n} -rate. The RMSE of our estimator declines at a rate much faster than that of Newey and Powell (2003) and that of Pinkse (2000) for DGPs 2–4.

Table 2
RMSEs and convergence rate comparison for DGP2

Parameter	Estimators	$n = 100$		$n = 400$		Ratio	
		Mean (1)	Median (2)	Mean (3)	Median (4)	(3)/(1)	(4)/(2)
$\theta = 0.2$	<i>Naive</i>	0.310	0.309	0.298	0.298	0.961	0.964
	Pinkse (2000)						
	$(L_1, L_2) = (2, 2)$	0.621	0.619	0.611	0.611	0.984	0.987
	$(L_1, L_2) = (2, 4)$	0.616	0.614	0.597	0.597	0.969	0.972
	$(L_1, L_2) = (2, 6)$	0.652	0.624	0.598	0.597	0.917	0.957
	$(L_1, L_2) = (4, 2)$	0.618	0.616	0.610	0.611	0.987	0.992
	$(L_1, L_2) = (4, 4)$	0.615	0.612	0.597	0.597	0.971	0.975
	$(L_1, L_2) = (4, 6)$	0.646	0.622	0.598	0.598	0.926	0.961
	$(L_1, L_2) = (6, 2)$	0.614	0.612	0.610	0.609	0.993	0.995
	$(L_1, L_2) = (6, 4)$	0.613	0.610	0.597	0.596	0.974	0.977
	$(L_1, L_2) = (6, 6)$	0.640	0.621	0.598	0.597	0.934	0.961
	Newey and Powell (2003)						
	$L = 5$	0.957	0.956	0.958	0.959	1.001	1.003
	$L = 50$	0.550	0.550	0.527	0.528	0.958	0.960
Su and Ullah	0.188	0.162	0.115	0.107	0.612	0.660	
$\theta = 0.5$	<i>Naive</i>	0.340	0.334	0.308	0.306	0.906	0.916
	Pinkse (2000)						
	$(L_1, L_2) = (2, 2)$	0.640	0.633	0.617	0.619	0.964	0.978
	$(L_1, L_2) = (2, 4)$	0.638	0.630	0.603	0.605	0.945	0.960
	$(L_1, L_2) = (2, 6)$	0.680	0.645	0.606	0.607	0.891	0.941
	$(L_1, L_2) = (4, 2)$	0.632	0.626	0.615	0.617	0.973	0.986
	$(L_1, L_2) = (4, 4)$	0.639	0.629	0.604	0.608	0.945	0.967
	$(L_1, L_2) = (4, 6)$	0.679	0.648	0.606	0.607	0.892	0.937
	$(L_1, L_2) = (6, 2)$	0.625	0.618	0.613	0.614	0.981	0.994
	$(L_1, L_2) = (6, 4)$	0.635	0.627	0.603	0.606	0.950	0.967
	$(L_1, L_2) = (6, 6)$	0.670	0.644	0.606	0.606	0.904	0.941
	Newey and Powell (2003)						
	$L = 5$	0.969	0.970	0.960	0.960	0.991	0.990
	$L = 50$	0.585	0.581	0.536	0.535	0.916	0.921
Su and Ullah	0.265	0.229	0.159	0.137	0.600	0.598	
$\theta = 0.8$	<i>Naive</i>	0.387	0.372	0.327	0.321	0.845	0.863
	Pinkse (2000)						
	$(L_1, L_2) = (2, 2)$	0.671	0.655	0.626	0.628	0.933	0.959
	$(L_1, L_2) = (2, 4)$	0.674	0.664	0.614	0.618	0.911	0.931
	$(L_1, L_2) = (2, 6)$	0.731	0.684	0.618	0.617	0.845	0.902
	$(L_1, L_2) = (4, 2)$	0.661	0.650	0.623	0.626	0.943	0.963
	$(L_1, L_2) = (4, 4)$	0.679	0.670	0.615	0.619	0.906	0.924
	$(L_1, L_2) = (4, 6)$	0.733	0.695	0.619	0.620	0.844	0.892
	$(L_1, L_2) = (6, 2)$	0.651	0.645	0.620	0.621	0.952	0.963
	$(L_1, L_2) = (6, 4)$	0.677	0.667	0.614	0.617	0.907	0.925
	$(L_1, L_2) = (6, 6)$	0.726	0.691	0.619	0.619	0.853	0.896
	Newey and Powell (2003)						
	$L = 5$	0.985	0.983	0.963	0.961	0.978	0.978
	$L = 50$	0.634	0.624	0.548	0.545	0.864	0.873
Su and Ullah	0.373	0.323	0.197	0.183	0.528	0.567	

4. Conclusion and extensions

In this paper we propose a three-step procedure to estimate the structural equation in nonparametric simultaneous equations models under the conditional mean independence restriction. It is based upon local polynomial regression and marginal integration techniques. We establish the asymptotic normality of our estimator under weak data dependence conditions. Our small set of simulation results suggests that our

Table 3
RMSEs and convergence rate comparison for DGP3

Parameter	Estimators	$n = 100$		$n = 400$		Ratio	
		Mean (1)	Median (2)	Mean (3)	Median (4)	(3)/(1)	(4)/(2)
$\theta = 0.2$	<i>Naive</i>	0.591	0.590	0.589	0.589	0.997	0.998
	<i>Pinkse (2000)</i>						
	$(L_1, L_2) = (2, 2)$	0.548	0.547	0.552	0.556	1.007	1.016
	$(L_1, L_2) = (2, 4)$	0.369	0.363	0.369	0.366	1.000	1.008
	$(L_1, L_2) = (2, 6)$	0.357	0.328	0.295	0.291	0.826	0.887
	$(L_1, L_2) = (4, 2)$	0.547	0.546	0.552	0.555	1.009	1.016
	$(L_1, L_2) = (4, 4)$	0.368	0.363	0.369	0.366	1.003	1.008
	$(L_1, L_2) = (4, 6)$	0.352	0.328	0.295	0.293	0.838	0.893
	$(L_1, L_2) = (6, 2)$	0.546	0.549	0.551	0.554	1.009	1.009
	$(L_1, L_2) = (6, 4)$	0.367	0.362	0.369	0.366	1.005	1.011
	$(L_1, L_2) = (6, 6)$	0.347	0.326	0.294	0.291	0.847	0.893
	<i>Newey and Powell (2003)</i>						
	$L = 5$	0.611	0.611	0.617	0.618	1.010	1.011
	$L = 50$	0.606	0.604	0.613	0.613	1.012	1.015
	<i>Su and Ullah</i>	0.393	0.354	0.316	0.257	0.804	0.726
$\theta = 0.5$	<i>Naive</i>	0.620	0.617	0.600	0.599	0.968	0.971
	<i>Pinkse (2000)</i>						
	$(L_1, L_2) = (2, 2)$	0.572	0.570	0.558	0.562	0.976	0.986
	$(L_1, L_2) = (2, 4)$	0.402	0.392	0.379	0.378	0.943	0.964
	$(L_1, L_2) = (2, 6)$	0.401	0.366	0.307	0.303	0.766	0.828
	$(L_1, L_2) = (4, 2)$	0.571	0.567	0.558	0.561	0.977	0.989
	$(L_1, L_2) = (4, 4)$	0.405	0.395	0.379	0.379	0.936	0.959
	$(L_1, L_2) = (4, 6)$	0.403	0.374	0.308	0.308	0.764	0.824
	$(L_1, L_2) = (6, 2)$	0.569	0.566	0.557	0.560	0.979	0.989
	$(L_1, L_2) = (6, 4)$	0.405	0.396	0.379	0.379	0.936	0.957
	$(L_1, L_2) = (6, 6)$	0.398	0.373	0.309	0.306	0.776	0.820
	<i>Newey and Powell (2003)</i>						
	$L = 5$	0.631	0.625	0.622	0.621	0.986	0.994
	$L = 50$	0.632	0.626	0.619	0.618	0.979	0.987
	<i>Su and Ullah</i>	0.437	0.411	0.332	0.294	0.760	0.715
$\theta = 0.8$	<i>Naive</i>	0.662	0.651	0.616	0.615	0.931	0.945
	<i>Pinkse (2000)</i>						
	$(L_1, L_2) = (2, 2)$	0.610	0.603	0.569	0.571	0.933	0.947
	$(L_1, L_2) = (2, 4)$	0.456	0.441	0.394	0.391	0.864	0.887
	$(L_1, L_2) = (2, 6)$	0.474	0.424	0.327	0.320	0.690	0.755
	$(L_1, L_2) = (4, 2)$	0.609	0.602	0.569	0.570	0.934	0.947
	$(L_1, L_2) = (4, 4)$	0.466	0.545	0.397	0.394	0.852	0.723
	$(L_1, L_2) = (4, 6)$	0.484	0.448	0.332	0.327	0.686	0.730
	$(L_1, L_2) = (6, 2)$	0.607	0.600	0.568	0.569	0.936	0.948
	$(L_1, L_2) = (6, 4)$	0.468	0.454	0.397	0.394	0.848	0.868
	$(L_1, L_2) = (6, 6)$	0.480	0.454	0.334	0.330	0.696	0.727
	<i>Newey and Powell (2003)</i>						
	$L = 5$	0.665	0.651	0.630	0.627	0.947	0.963
	$L = 50$	0.675	0.656	0.628	0.625	0.930	0.953
	<i>Su and Ullah</i>	0.509	0.489	0.359	0.335	0.705	0.685

estimator may significantly outperform the estimators of Pinkse (2000) and Newey and Powell (2003) in finite samples.

Our theoretical results can be extended in three directions. First, one can pursue one step further to obtain a potentially asymptotically more efficient estimator than ours by following the procedure of Linton (1997, 2000). Linton (2000) defines a novel procedure for estimating generalized additive nonparametric regression models that are potentially more efficient than our integration-based method (Step 3). Let $V_t = (X'_t, Z'_t)'$ and

Table 4
RMSEs and convergence rate comparison for DGP4

Parameter	Estimators	$n = 100$		$n = 400$		Ratio	
		Mean (1)	Median (2)	Mean (3)	Median (4)	(3)/(1)	(4)/(2)
$\theta = 0.2$	<i>Naive</i>	0.449	0.449	0.429	0.429	0.955	0.955
	Pinkse (2000)						
	$(L_1, L_2) = (2, 2)$	0.400	0.401	0.390	0.389	0.975	0.970
	$(L_1, L_2) = (2, 4)$	0.371	0.359	0.335	0.333	0.903	0.928
	$(L_1, L_2) = (2, 6)$	0.449	0.374	0.326	0.320	0.726	0.856
	$(L_1, L_2) = (4, 2)$	0.396	0.396	0.386	0.386	0.975	0.975
	$(L_1, L_2) = (4, 4)$	0.356	0.352	0.324	0.324	0.910	0.920
	$(L_1, L_2) = (4, 6)$	0.420	0.360	0.316	0.310	0.752	0.861
	$(L_1, L_2) = (6, 2)$	0.395	0.400	0.387	0.386	0.980	0.965
	$(L_1, L_2) = (6, 4)$	0.356	0.351	0.325	0.325	0.913	0.926
	$(L_1, L_2) = (6, 6)$	0.395	0.358	0.315	0.310	0.797	0.866
	Newey and Powell (2003)						
	$L = 5$	0.460	0.460	0.451	0.449	0.980	0.976
	$L = 50$	0.298	0.292	0.269	0.266	0.903	0.911
Su and Ullah	0.247	0.218	0.215	0.160	0.870	0.734	
$\theta = 0.5$	<i>Naive</i>	0.483	0.480	0.447	0.447	0.925	0.931
	Pinkse (2000)						
	$(L_1, L_2) = (2, 2)$	0.450	0.443	0.428	0.429	0.951	0.968
	$(L_1, L_2) = (2, 4)$	0.441	0.420	0.382	0.377	0.866	0.898
	$(L_1, L_2) = (2, 6)$	0.551	0.446	0.374	0.365	0.679	0.818
	$(L_1, L_2) = (4, 2)$	0.423	0.419	0.396	0.396	0.936	0.945
	$(L_1, L_2) = (4, 4)$	0.393	0.383	0.334	0.335	0.850	0.875
	$(L_1, L_2) = (4, 6)$	0.475	0.404	0.326	0.323	0.686	0.800
	$(L_1, L_2) = (6, 2)$	0.420	0.417	0.394	0.393	0.938	0.942
	$(L_1, L_2) = (6, 4)$	0.392	0.383	0.334	0.333	0.852	0.869
	$(L_1, L_2) = (6, 6)$	0.448	0.402	0.325	0.322	0.725	0.801
	Newey and Powell (2003)						
	$L = 5$	0.487	0.479	0.456	0.453	0.936	0.946
	$L = 50$	0.355	0.340	0.286	0.279	0.806	0.821
Su and Ullah	0.345	0.288	0.228	0.205	0.661	0.712	
$\theta = 0.8$	<i>Naive</i>	0.520	0.512	0.463	0.462	0.890	0.902
	Pinkse (2000)						
	$(L_1, L_2) = (2, 2)$	0.537	0.525	0.499	0.495	0.929	0.943
	$(L_1, L_2) = (2, 4)$	0.549	0.516	0.459	0.450	0.836	0.872
	$(L_1, L_2) = (2, 6)$	0.701	0.551	0.455	0.439	0.649	0.797
	$(L_1, L_2) = (4, 2)$	0.472	0.464	0.416	0.413	0.881	0.890
	$(L_1, L_2) = (4, 4)$	0.458	0.447	0.359	0.358	0.784	0.801
	$(L_1, L_2) = (4, 6)$	0.567	0.481	0.355	0.348	0.626	0.723
	$(L_1, L_2) = (6, 2)$	0.465	0.460	0.409	0.407	0.880	0.885
	$(L_1, L_2) = (6, 4)$	0.456	0.444	0.353	0.351	0.774	0.791
	$(L_1, L_2) = (6, 6)$	0.536	0.474	0.349	0.343	0.651	0.724
	Newey and Powell (2003)						
	$L = 5$	0.533	0.512	0.466	0.458	0.874	0.895
	$L = 50$	0.436	0.411	0.313	0.299	0.718	0.727
Su and Ullah	0.454	0.384	0.333	0.265	0.733	0.690	

$v = (x', z_1')$. For simplicity, suppose $c = 0$ and $\hat{m}_u(u)$ is some initial consistent estimator for $E(\varepsilon|U = u)$ obtained in the same fashion as one obtains $\hat{g}_Q(x, z_1)$. Let $g_Q^*(x, z_1)$ be the minimizing intercept in the objective function

$$Q_n(\beta) \equiv nb^{-(d_x+d_1)} \sum_{t=1}^n K((V_t - v))/b \left[Y_t - \hat{m}_u(U_t) - \sum_{0 \leq |j| \leq p_3} \beta_j (V_t - v)^j \right]^2, \tag{4.1}$$

where β is a collection of the parameters β_j , $0 \leq |j| \leq p_3$, K is a kernel function, and $b = b(n)$ is a bandwidth sequence. Under some suitable conditions, we conjecture that $g_Q^*(x, z_1)$ is more efficient than $\hat{g}_Q(x, z_1)$ asymptotically in the sense of mean squared errors by the argument of Linton (2000). Since this involves further complication, we leave it for future research.

Second, due to the ‘‘curse of dimensionality’’, it is reasonable to extend our results to semiparametric simultaneous equations models. This extension is straightforward given the faster convergence rate of the parametric component than the nonparametric one.

Third, like Pinkse (2000), Newey et al. (1999), and Newey and Powell (2003), we have implicitly assumed that the regressors X_t and Z_t exhibit a density. This assumption may be untenable in practice. It is possible to extend our test to allow (X_t, Z_t) to be a mixture of continuous and discrete variables by using the important tools developed by Racine and Li (2004). For the bandwidth selection problem in the case of mixed categorical and continuous data, see the important work by Li and Racine (2004).

Acknowledgements

We are grateful to Ronald Gallant, the Associate Editor, and two anonymous referees for their very helpful comments and constructive suggestions on an earlier version of this paper. The first author gratefully acknowledges financial support from the NSFC under Grant numbers 70501001 and 70601001. The second author gratefully acknowledges the financial support from the Academic Senate, UCR. The usual disclaimers apply.

Appendix A. Some technical lemmas

This appendix presents some technical lemmas that are used in proving the main results.

Lemma A.1. *Let $\{\xi_i, i \geq 1\}$ be a d -dimensional strong mixing process with the mixing coefficient $\alpha(i)$. For any integer $p > 1$ and integers (i_1, \dots, i_p) such that $1 \leq i_1 < i_2 < \dots < i_p$, let φ be a Borel function defined on \mathbb{R}^{pd} such that*

$$\int |\varphi(v_1, \dots, v_p)|^{1+\vartheta} dF^{(1)}(v_1, \dots, v_j) dF^{(2)}(v_{j+1}, \dots, v_p) \leq M_1$$

for some $\vartheta > 0$ and $M_1 > 0$, where $F^{(1)} = F_{i_1, \dots, i_j}$ and $F^{(2)} = F_{i_{j+1}, \dots, i_p}$ are the distribution functions of $(\xi_{i_1}, \dots, \xi_{i_j})$ and $(\xi_{i_{j+1}}, \dots, \xi_{i_p})$, respectively. Let F denote the distribution function of $(\xi_{i_1}, \dots, \xi_{i_p})$. Then

$$\left| \int \varphi(v_1, \dots, v_p) dF(v_1, \dots, v_p) - \int \varphi(v_1, \dots, v_p) dF^{(1)}(v_1, \dots, v_j) dF^{(2)}(v_{j+1}, \dots, v_p) \right| \leq 4M_1^{1/(1+\vartheta)} \alpha(i_{j+1} - i_j)^{\vartheta/(1+\vartheta)}.$$

Proof. This is Lemma 2.1 of Sun and Chiang (1997). □

Lemma A.2. *Let $\phi(\cdot, \cdot)$ be a symmetric Borel function defined on $\mathbb{R}^d \times \mathbb{R}^d$. Let the strictly stationary process $\{\xi_i, i \geq 1\}$ be defined as in Lemma A.1. Assume that for any fixed $v \in \mathbb{R}^d$, $E[\phi(\xi_1, v)] = 0$. Let*

$$M_2 = \max_{1 \leq t < s \leq n} \max \left\{ E|\phi(\xi_t, \xi_s)|^{2(1+\vartheta)}, \int \phi(\xi_t, \xi_s)^{2(1+\vartheta)} dF(\xi_t) dF(\xi_s) \right\},$$

where $\vartheta > 0$ is a fixed constant, and $F(\xi_t)$ indicates the probability law of ξ_t . Then

$$E \left\{ \sum_{1 \leq t < s \leq n} \phi(\xi_t, \xi_s) \right\}^2 \leq Cn^2 M_2^{1/(1+\vartheta)},$$

where $C > 0$ is a constant independent of n and the function ϕ .

Proof. This is Lemma C.2(ii) of Gao and King (2001). □

Lemma A.3. Let $\psi(\cdot, \cdot, \cdot)$ be a symmetric Borel function defined on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$. Let the strictly stationary process $\{\xi_i, i \geq 1\}$ be defined as in Lemma A.1. Assume that for any fixed $v, v^* \in \mathbb{R}^d$, $E[\phi(\xi_1, v, v^*)] = 0$. Fix small constant $\vartheta > 0$. Let $M_3 = \max\{M_{31}, M_{32}, M_{33}\}$, where

$$M_{31} = \max_{1 \leq t < s \leq n} \max \left\{ E|\psi(\xi_1, \xi_t, \xi_s)|^{2(1+\vartheta)}, \int |\psi(\xi_1, \xi_t, \xi_s)|^{2(1+\vartheta)} dF(\xi_1) dF(\xi_t, \xi_s) \right\},$$

$$M_{32} = \max_{1 \leq t < s \leq n} \int |\psi(\xi_1, \xi_t, \xi_s)|^{2(1+\vartheta)} dF(\xi_s) dF(\xi_1, \xi_t),$$

and

$$M_{33} = \max_{1 \leq t < s \leq n} \int |\psi(\xi_1, \xi_t, \xi_s)|^{2(1+\vartheta)} dF(\xi_1) dF(\xi_t) dF(\xi_s).$$

Then

$$E \left\{ \sum_{1 \leq t_1 < t_2 < t_3 \leq n} \phi(\xi_{t_1}, \xi_{t_2}, \xi_{t_3}) \right\}^2 \leq Cn^3 M_3^{1/(1+\vartheta)},$$

where $C > 0$ is a constant independent of n and the function ψ .

Proof. This is Lemma C.2(i) of Gao and King (2001). \square

Appendix B. Proof of the main theorems

We use C to signify a generic constant whose exact value may vary from case to case. Following the notation of Masry (1996a, b), let $N_l = (l + d - 1)! / (l!(d - 1)!)$ be the number of distinct d -tuples \mathbf{j} with $|\mathbf{j}| = l$, where $d = 2d_x + d_1$. Arrange the N_l d -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 $0 \leq |\mathbf{j}| \leq 2p_2$, let

$$\mu_{\mathbf{j}}(K) = \int_{\mathbb{R}^{2d_x+d_1}} w^{\mathbf{j}} K_2(w) dw.$$

For each \mathbf{k} and \mathbf{l} with

$$0 \leq |\mathbf{k}|, |\mathbf{l}| \leq p_2, \quad v_{\mathbf{k}, \mathbf{l}}(K_2) = \int_{\mathbb{R}^{d_x+d_1}} \int_{\mathbb{R}^{d_x}} \int_{\mathbb{R}^{d_x}} (v, u)^{\mathbf{k}} (v, \tilde{u})^{\mathbf{l}} K_2(v, u) K_2(v, \tilde{u}) du d\tilde{u} dv.$$

Define the $N \times N$ dimensional matrices M and Γ , and the $N \times N_{p_2+1}$ matrix B by

$$M = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p_2} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p_2} \\ \vdots & \vdots & & \vdots \\ M_{p_2,0} & M_{p_2,1} & \cdots & M_{p_2,p_2} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,p_2} \\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,p_2} \\ \vdots & \vdots & & \vdots \\ \Gamma_{p_2,0} & \Gamma_{p_2,1} & \cdots & \Gamma_{p_2,p_2} \end{bmatrix},$$

$$B = \begin{bmatrix} M_{0,p_2+1} \\ M_{1,p_2+1} \\ \vdots \\ M_{p_2,p_2+1} \end{bmatrix}, \tag{B.1}$$

where

$$N = \sum_{l=0}^{p_2} N_l,$$

M_{ij} and Γ_{ij} are $N_i \times N_j$ dimensional matrices whose (l, r) elements are, respectively, $\mu_{\phi_j(l)+\phi_j(r)}$ and $v_{\phi_j(l),\phi_j(r)}$. Note that the elements of the matrix $M = M(K_2, p_2)$ are simply multivariate moments of the kernel K_2 . The elements of the matrix $\Gamma = \Gamma(K_2, p_2)$ look more complicated but can be simplified for given p_2 and K_2 . For example, if $p_2 = 1$ and K_2 is a second-order product kernel, Γ has nonzero elements along the diagonal and 0 everywhere else. All matrices, M , Γ and B , depend on the kernel (K_2) and the order (p_2) of the local polynomial we use.

Let \bar{M} be defined analogously as M but with kernel K_1 and local polynomial order p_1 . Clearly, \bar{M} is also a square matrix and we denote its number of rows as \bar{N} . As in the main text, let $W_t = (X'_t, Z'_t, U'_t)'$, $\widehat{W}_t = (X'_t, Z'_t, \widehat{U}'_t)'$, $w = (x', z'_1, u)'$, and $v_{1n} = n^{-1/2} b_1^{-d_x/2} \sqrt{\log n}$.

Proof of Theorem 2.1. To facilitate the proof, let $\mathcal{K}_{2,t}(w)$ be an $N \times 1$ vector, $\mathcal{K}_{2,t}^{(1)}(w)$ be an $N \times d_x$ matrix, and $M_n(w)$ be a symmetric $N \times N$ matrix:

$$\mathcal{K}_{2,t}(w) = \begin{bmatrix} \mathcal{K}_{2,t,0}(w) \\ \mathcal{K}_{2,t,1}(w) \\ \vdots \\ \mathcal{K}_{2,t,p_2}(w) \end{bmatrix}, \quad \mathcal{K}_{2,t}^{(1)}(w) = \begin{bmatrix} \mathcal{K}_{2,t,0}^{(1)}(w) \\ \mathcal{K}_{2,t,1}^{(1)}(w) \\ \vdots \\ \mathcal{K}_{2,t,p_2}^{(1)}(w) \end{bmatrix},$$

$$M_n(w) = \begin{bmatrix} M_{n,0,0}(w) & M_{n,0,1}(w) & \cdots & M_{n,0,p_2}(w) \\ M_{n,1,0}(w) & M_{n,1,1}(w) & \cdots & M_{n,1,p_2}(w) \\ \vdots & \vdots & & \vdots \\ M_{n,p_2,0}(w) & M_{n,p_2,1}(w) & \cdots & M_{n,p_2,p_2}(w) \end{bmatrix},$$

where $\mathcal{K}_{2,t,j}(w)$ is an N_j dimensional subvector whose r th element is given by

$$[\mathcal{K}_{2,t,j}(w)]_r = \left(\frac{W_t - w}{b_2} \right)^{\phi_j(r)} K_2 \left(\frac{W_t - w}{b_2} \right),$$

$\mathcal{K}_{2,t,j}^{(1)}(w)$ is an $N_j \times d_x$ matrix with the (r, l) element being the partial derivative of $[\mathcal{K}_{2,t,j}(w)]_r = [\mathcal{K}_{2,t,j}(x, z_1, u)]_r$ with respect to its l th element in u , and $M_{n,j,k}(w)$ is an $N_j \times N_k$ dimensional submatrix with the (l, r) element given by

$$[M_{n,j,k}(w)]_{l,r} = \frac{1}{nh_2^{2d_x+d_1}} \sum_{t=1}^n \left(\frac{W_t - w}{b_2} \right)^{\phi_j(l)+\phi_k(r)} K_2 \left(\frac{W_t - w}{b_2} \right).$$

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

Noticing that $A_1^{-1} - A_2^{-1} = A_1^{-1}(A_2 - A_1)A_2^{-1}$ for nonsingular matrices A_1 and A_2 , we write (recall $w = (x', z_1', u')$)

$$\begin{aligned} \widehat{m}(x, z_1, u) - m(x, z_1, u) &= \iota_2' \widehat{M}_n^{-1}(w) \widehat{V}_n(w) + \iota_2' \widehat{M}_n^{-1}(w) \widehat{B}_n(w) \\ &= \iota_2' [f_w(w)M]^{-1} \widehat{V}_n(w) + \iota_2' [f_w(w)M]^{-1} \widehat{B}_n(w) \\ &\quad - \iota_2' [f_w(w)M]^{-1} [\widehat{M}_n(w) - f_w(w)M] \widehat{M}_n^{-1}(w) \widehat{V}_n(w) \\ &\quad - \iota_2' [f_w(w)M]^{-1} [\widehat{M}_n(w) - f_w(w)M] \widehat{M}_n^{-1}(w) \widehat{B}_n(w) \\ &\equiv T_{n,1}(w) + T_{n,2}(w) - T_{n,3}(w) - T_{n,4}(w), \end{aligned} \tag{B.2}$$

where $\iota_2 = (1, 0, \dots, 0)'$ is an N -vector, M is defined in (B.1), the “variance” term $\widehat{V}_n(w)$ and the “bias” term $\widehat{B}_n(w)$ are $N \times 1$ vectors defined by

$$\widehat{V}_n(w) = n^{-1} b_2^{-(2d_x+d_1)} \sum_{t=1}^n \widehat{\mathcal{K}}_{2,t}(w) e_t \tag{B.3}$$

and

$$\widehat{B}_n(w) = n^{-1} b_2^{-(2d_x+d_1)} \sum_{t=1}^n \widehat{\mathcal{K}}_{2,t}(w) \widehat{\Delta}_t(w), \tag{B.4}$$

where $e_t \equiv Y_t - m(X_t, Z_{1t}, U_t)$, and

$$\widehat{\Delta}_t(w) \equiv m(\widehat{W}_t) - \sum_{0 \leq |\mathbf{k}| \leq p_2} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}m)(w) (\widehat{W}_t - w)^{\mathbf{k}}.$$

The presentation of (B.2) assumes that \widehat{M}_n is invertible, which we shall show is true with probability approaching 1 as $n \rightarrow \infty$ (w.p.a.1) in the proof of Lemma B.3. We analyze the properties of $T_{n,i}(w)$, $i = 1, \dots, 4$ in Lemmas B.1–B.4, which shall complete the proof of the theorem. \square

Lemma B.1. Under Assumptions A1–A5,

$$\sqrt{nh_2^{d_x+d_1}} \int T_{n,1}(w) dQ(u) \xrightarrow{d} N(0, a_Q(x, z_1)[M^{-1}\Gamma M^{-1}]_{0,0}),$$

where M and Γ are defined in (B.1).

Proof. Let $V_n(w)$ be defined analogously as $\widehat{V}_n(w)$ in (B.3) but with $\mathcal{K}_{2,t}(w)$ in place of $\widehat{\mathcal{K}}_{2,t}(w)$. By the Taylor expansion and Assumptions A1 and A5, we have

$$\begin{aligned} \widehat{V}_n(w) - V_n(w) &= n^{-1} b_2^{-(2d_x+d_1)} \sum_{t=1}^n (\widehat{\mathcal{K}}_{2,t}(w) - \mathcal{K}_{2,t}(w)) e_t \\ &= n^{-1} b_2^{-(2d_x+d_1)} \sum_{t=1}^n \mathcal{K}_{2,t}^{(1)}(w) (\widehat{U}_t - U_t) e_t + O_p((b_2^{-1}v_{1n} + b_2^{-1}b_1^{p_1+1})^2) \\ &= n^{-1} b_2^{-(2d_x+d_1)} \sum_{t=1}^n \sum_{i=1}^{d_x} \frac{\partial \mathcal{K}_{2,t}(x, z_1, u)}{\partial u_i} (h_i(Z_t) - \widehat{h}_i(Z_t)) e_t + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}) \end{aligned}$$

uniformly in w , where the second equality follows from the property of K_2 and the fact that $\max_{1 \leq t \leq n} \|\widehat{U}_t - U_t\| \leq d_x \max_{1 \leq i \leq d_x} \max_{1 \leq t \leq n} |\widehat{h}_i(Z_t) - h_i(Z_t)| = O_p(v_{1n} + b_1^{p_1+1})$. Define $\mathcal{K}_{1,t}(z)$ analogously to $\mathcal{K}_{2,t}(w)$ with a typical element:

$$[\mathcal{K}_{1,t,j}(z)]_r = \left(\frac{Z_t - z}{b_1}\right)^{\phi_j(r)} K_1\left(\frac{Z_t - z}{b_1}\right).$$

By results in Masry (1996b, Eq. (2.13), Corollary 2(ii)), for $i = 1, \dots, d_x$,

$$\hat{h}_i(z) - h_i(z) = n^{-1} b_1^{-d_z} l_1' [\overline{M}f_z(z)]^{-1} \left\{ \sum_{t=1}^n \mathcal{K}_{1,t}(z) \left[U_{it} + \sum_{|\mathbf{k}|=p_1+1} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}h_i)(z)(Z_t - z)^{\mathbf{k}} \right] + \gamma_n(z) \right\} \times \{1 + o_p(1)\} \text{ uniformly in } z, \tag{B.5}$$

where $l_1 = (1, 0, \dots, 0)'$ be an \overline{N} -vector, and

$$\gamma_n(z) \equiv \frac{p_1 + 1}{n b_1^{d_z}} \sum_{|\mathbf{k}|=p_1+1} \frac{1}{\mathbf{k}!} \sum_{t=1}^n \mathcal{K}_{1,t}(z)(Z_t - z)^{\mathbf{k}} \times \int_0^1 [(D^{\mathbf{k}}h_i)\{z + \tau(Z_t - z)\} - (D^{\mathbf{k}}h_i)(z)](1 - \tau)^{p_1} d\tau.$$

So

$$\widehat{V}_n(w) - V_n(w) = \{V_{n,1}(w) + V_{n,2}(w) + V_{n,3}(w)\} \{1 + o_p(1)\} + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}),$$

in which

$$V_{n,1}(w) \equiv n^{-2} b_1^{-d_z} b_2^{-(2d_x+d_1)} \sum_{t=1}^n \sum_{s=1}^n \alpha_n(\xi_t, \xi_s; w),$$

$$V_{n,2}(w) \equiv n^{-1} b_2^{-(2d_x+d_z)} \sum_{t=1}^n \sum_{i=1}^{d_x} \frac{\partial \mathcal{K}_{2,t}(x, z_1, u)}{\partial u_i} e_t \beta_{ni}(Z_t),$$

and

$$V_{n,3}(w) \equiv n^{-1} b_2^{-(2d_x+d_z)} \sum_{t=1}^n \sum_{i=1}^{d_x} \frac{\partial \mathcal{K}_{2,t}(x, z_1, u)}{\partial u_i} e_t \gamma_n(Z_t),$$

where for $\xi_t \equiv (X'_t, Z'_t, U'_t, e_t)'$,

$$\alpha_n(\xi_t, \xi_s; w) \equiv - \sum_{i=1}^{d_x} \frac{\partial \mathcal{K}_{2,t}(x, z_1, u)}{\partial u_i} e_t l_1' [\overline{M}f_z(Z_t)]^{-1} \mathcal{K}_{1,s}(Z_t) U_{is}$$

and

$$\beta_{ni}(z) \equiv -n^{-1} b_1^{-d_z} l_1' [\overline{M}f_z(z)]^{-1} \sum_{s=1}^n \mathcal{K}_{1,s}(z) \sum_{|\mathbf{k}|=p_1+1} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}h_i)(z)(Z_s - z)^{\mathbf{k}}.$$

Consequently, we have

$$\int T_{n,1}(w) dQ(u) = T_{n,1a} + (T_{n,1b} + T_{n,1c} + T_{n,1d}) \{1 + o_p(1)\} + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}), \tag{B.6}$$

where

$$T_{n,1a} = \frac{1}{n b_2^{2d_x+d_1}} \sum_{t=1}^n e_t l_2' M^{-1} \int \mathcal{K}_{2,t}(w) f_w^{-1}(w) dQ(u),$$

$$T_{n,1b} = \frac{1}{n^2 b_1^{d_z} b_2^{2d_x+d_1}} \sum_{t=1}^n \sum_{s=1}^n l_2' M^{-1} \int \alpha_n(\xi_t, \xi_s; w) f_w^{-1}(w) dQ(u),$$

$$T_{n,1c} = \frac{1}{n b_2^{2d_x+d_1}} \sum_{t=1}^n \sum_{i=1}^{d_x} e_t l_2' M^{-1} \int \sum_{i=1}^{d_x} \frac{\partial \mathcal{K}_{2,t}(x, z_1, u)}{\partial u_i} \beta_{ni}(Z_t) f_w^{-1}(w) dQ(u),$$

and

$$T_{n,1d} = \frac{1}{nb_2^{2d_x+d_1}} \sum_{t=1}^n \sum_{i=1}^{d_x} e_t i_2' M^{-1} \int \sum_{i=1}^{d_x} \frac{\partial \mathcal{K}_{2,t}(x, z_1, u)}{\partial u_i} \gamma_n(Z_t) f_w^{-1}(w) dQ(u).$$

We complete the proof by showing that

- (i) $\sqrt{nb_2^{d_x+d_1}} T_{n,1a} \xrightarrow{d} N(0, a_Q(x, z_1)[M^{-1}\Gamma M^{-1}]_{0,0})$, and
- (ii) $\sqrt{nb_2^{d_x+d_1}} T_{n,1l} = o_p(1)$, $l = b, c, d$.

To show (i), let $\zeta_t = b_2^{-(2d_x+d_1)} e_t i_2' M^{-1} \int \mathcal{K}_{2,t}(w) f_w^{-1}(w) dQ(u)$. By the law of iterated expectations and dominated convergence arguments, $E[\zeta_t] = 0$ and

$$\begin{aligned} E[\zeta_t^2] &= b_2^{-(2d_x+2d_1)} i_2' M^{-1} E\{\sigma_\epsilon^2(X_t, Z_{1t}, U_t) \iint b_2^{-2d_x} \mathcal{K}_{2,t}(x, z_1, u) [\mathcal{K}_{2,t}(x, z_1, \tilde{u})]' \\ &\quad \times f_w^{-1}(x, z_1, u) f_w^{-1}(x, z_1, \tilde{u}) dQ(u) dQ(\tilde{u})\} M^{-1} i_2 \\ &= b_2^{-(d_x+d_1)} a_Q(x, z_1) i_2' M^{-1} \Gamma M^{-1} i_2 \{1 + o(1)\}, \end{aligned}$$

where $a_Q(x, z_1) = \int q^2(u) \sigma_\epsilon^2(x, z_1, u) f_w^{-1}(x, z_1, u) du$ and Γ is defined in (B.1). By Assumptions A1–A5 and the proof of Theorem 4 in Masry (1996a),

$$\sqrt{nb_2^{d_x+d_1}} T_{n,1a} \xrightarrow{d} N(0, a_Q(x, z_1)[M^{-1}\Gamma M^{-1}]_{0,0}). \tag{B.7}$$

To show $\sqrt{nb_2^{d_x+d_1}} T_{n,1b} = o_p(1)$, let $\varphi(\zeta_t, \zeta_s) = i_2' M^{-1} \int \alpha_n(\zeta_t, \zeta_s; w) f_w^{-1}(w) dQ(u)$ and $\bar{\varphi}(\zeta_t, \zeta_s) = \{\varphi(\zeta_t, \zeta_s) + \varphi(\zeta_s, \zeta_t)\}/2$, a symmetric version of $\varphi(\zeta_t, \zeta_s)$. Then $E_t \bar{\varphi}(\zeta_t, \zeta_s) = 0$, where E_t denotes expectation with respect to ζ_t only. Write

$$\begin{aligned} T_{n,1b} &= \frac{1}{n^2 b_1^{d_z} b_2^{2d_x+d_1}} \sum_{t=1}^n \sum_{s=1}^n \varphi(\zeta_t, \zeta_s) \\ &= \frac{1}{n^2 b_1^{d_z} b_2^{2d_x+d_1}} \sum_{t=1}^n \varphi(\zeta_t, \zeta_t) + \frac{2}{n^2 b_1^{d_z} b_2^{2d_x+d_1}} \sum_{1 \leq t < s \leq n} \bar{\varphi}(\zeta_t, \zeta_s) \\ &\equiv T_{n,1b}^{(1)} + 2T_{n,1b}^{(2)}. \end{aligned} \tag{B.8}$$

By the law of iterated expectations, the dominated convergence theorem, and Lemma A.1 (with $\vartheta = \delta/2$), $E T_{n,1b}^{(1)} = 0$, and

$$\begin{aligned} \text{Var}(T_{n,1b}^{(1)}) &= \frac{1}{n^4 b_1^{2d_z} b_2^{4d_x+2d_1}} \sum_{t=1}^n \sum_{s=1}^n E[\varphi_0(\zeta_t, \zeta_t) \varphi_0(\zeta_s, \zeta_s)] \\ &= \frac{1}{n^3 b_1^{2d_z} b_2^{4d_x+2d_1}} E[\varphi_0(\zeta_t, \zeta_t)^2] + \frac{1}{n^4 b_1^{2d_z} b_2^{4d_x+2d_1}} \sum_{\tau=1}^{n-1} \sum_{t=\tau+1}^n E[\varphi_0(\zeta_t, \zeta_t) \varphi_0(\zeta_{t-\tau}, \zeta_{t-\tau})] \\ &\leq Cn^{-3} b_1^{-2d_z} b_2^{-d_x-d_1-2} + Cn^{-3} b_1^{-2d_z} b_2^{-4d_x-2d_1-2} \\ &\quad \times (b_1^{2d_z} b_2^{2(d_x+d_1)+2d_x(1+\delta/2)})^{1/(1+\delta/2)} \sum_{\tau=1}^{n-1} (1 - \tau/n) \alpha^{\delta/(2+\delta)}(\tau) \\ &= O(n^{-3} b_1^{-2d_z} b_2^{-d_x-d_1-2} + n^{-3} b_1^{-2d_z \delta/(2+\delta)} b_2^{-2(d_x+d_1)\delta/(2+\delta)-2}) \\ &= o(n^{-1} b_2^{-(d_x+d_1)}) \text{ by Assumptions A5(i) and A2.} \end{aligned}$$

Thus $T_{n,1b}^{(1)} = o_p(n^{-1/2}b_2^{-(d_x+d_1)/2})$ by the Chebyshev inequality. By Lemma A.1 (with $\vartheta = \delta/2$) and Assumptions A1–A5,

$$\begin{aligned} |ET_{n,1b}^{(2)}| &\leq Cn^{-1}b_1^{-d_z}b_2^{-(2d_x+d_1)-1}(b_1^{d_z}b_2^{(d_x+d_1)+d_x(1+\delta/2)})^{1/(1+\delta/2)} \sum_{\tau=1}^{n-1} \alpha^{\delta/(2+\delta)}(\tau) \\ &= O(n^{-1}b_1^{-d_z\delta/(2+\delta)}b_2^{-(d_x+d_1)\delta/(2+\delta)-1}) = o(n^{-1/2}b_2^{-(d_x+d_1)/2}), \end{aligned}$$

where the last equality follows because Assumption A5(ii) implies that $nb_1^{2d_z\delta/(2+\delta)}b_2^{2-(d_x+d_1)(1-\delta)/(2+\delta)} \rightarrow \infty$ given the facts that $2\delta/(2+\delta) < 1$ and $(1-\delta)/(2+\delta) < \frac{1}{2}$. By Lemma A.2 (with $\vartheta = \delta/2$) and Assumptions A1–A5,

$$\begin{aligned} E(T_{n,1b}^{(2)})^2 &\leq Cn^{-2}b_1^{-2d_z}b_2^{-2(2d_x+d_1)-2}(b_1^{d_z}b_2^{(d_x+d_1)+2d_x(1+\delta/2)})^{1/(1+\delta/2)} \\ &= O(n^{-2}b_1^{-2d_z(1+\delta)/(2+\delta)}b_2^{-2(d_x+d_1)(1+\delta)/(2+\delta)-2}) = o_p(n^{-1}b_2^{-(d_x+d_1)}). \end{aligned}$$

It follows from the Chebyshev inequality that $T_{n,1b}^{(2)} = o_p(n^{-1/2}b_2^{-(d_x+d_1)/2})$. Consequently,

$$\sqrt{nb_2^{d_x+d_1}}T_{n,1b} = o_p(1). \tag{B.9}$$

Now it is straightforward to extend the proof of Theorem 2 in Masry (1996b) to show that $\sup_z |b_1^{-(p_1+1)}\beta_m(z) - \beta_i(z)| = O_p(v_{1n})$, where $\beta_i(z) = i_1'\overline{M}^{-1}\overline{B}h_i^{(p_1+1)}(z)$ and \overline{B} is defined analogously as B but with kernel K_1 and local polynomial order p_1 . So by Assumptions A1–A5,

$$\begin{aligned} T_{n,1c} &= \frac{b_1^{p_1+1}}{nb_2^{2d_x+d_1}} \sum_{i=1}^n \sum_{i=1}^{d_x} e_i i_2' M^{-1} \int \sum_{i=1}^{d_x} \frac{\partial \mathcal{H}_{2,i}(x, z_1, u)}{\partial u_i} \beta_i(Z_i) f_w^{-1}(w) dQ(u) + O_p(b_1^{(p_1+1)}b_2^{-1}v_{1n}) \\ &= b_1^{(p_1+1)} O_p(n^{-1/2}b_2^{-(d_x+d_1)/2-1}) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}), \end{aligned}$$

where the first term follows from the same argument as used in the proof of Theorem 2 in Masry (1996b). Hence

$$\sqrt{nb_2^{d_x+d_1}}T_{n,1c} = O_p(b_1^{p_1+1}b_2^{-1}) + o_p(1) = o_p(1). \tag{B.10}$$

Similarly, one can show

$$\sqrt{nb_2^{d_x+d_1}}T_{n,1d} = o_p(1). \tag{B.11}$$

Eqs. (B.6)–(B.11) complete the proof. \square

Lemma B.2. Under Assumptions A1–A5,

$$\int T_{n,2}(w) dQ(u) = b_2^{p_2+1} i_2' M^{-1} B \int m^{(p_2+1)}(w) dQ(u) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}).$$

Proof. Let

$$\Delta_t(w) = m(W_t) - \sum_{0 \leq |\mathbf{k}| \leq p_2} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}m)(w)(W_t - w)^{\mathbf{k}}.$$

Then by Assumption A4,

$$\Delta_t(w) = \sum_{|\mathbf{k}|=p+1} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}m)(w_t^*)(W_t - w)^{\mathbf{k}}$$

for some w_t^* that lies between W_t and w , and

$$\widehat{\Delta}_t(w) = \sum_{|\mathbf{k}|=p+1} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}m)(\widehat{w}_t^*)(\widehat{W}_t - w)^{\mathbf{k}}$$

for some \widehat{w}_t^* that lies between \widehat{W}_t and w . Clearly $\max_{1 \leq t \leq n} \|\widehat{w}_t^* - w_t^*\| = O_p(v_{1n} + b_1^{p_1+1})$. So by Assumptions A4–A5, uniformly in w and t for $\|W_t - w\| \leq Cb_2$,

$$\begin{aligned} \widehat{\Delta}_t(w) - \Delta_t(w) &= \left\{ \sum_{|\mathbf{k}|=p_2+1} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}m)(\widehat{w}_t^*)[(\widehat{W}_t - w)^{\mathbf{k}} - (W_t - w)^{\mathbf{k}}] \right\} \\ &\quad + \left\{ \sum_{|\mathbf{k}|=p_2+1} \frac{1}{\mathbf{k}!} [(D^{\mathbf{k}}m)(\widehat{w}_t^*) - (D^{\mathbf{k}}m)(w_t^*)](W_t - w)^{\mathbf{k}} \right\} \\ &= O_p(b_2^{p_2}(v_{1n} + b_1^{p_1+1})) = o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}). \end{aligned}$$

Since $|\Delta_t(w)| = O_p(b_2^{p_2+1})$ for $\|W_t - w\| \leq Cb_2$, it follows that $|\widehat{\Delta}_t(w)| = O_p(b_2^{p_2+1}) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}) = O_p(b_2^{p_2+1})$ uniformly in w and t such that $\|W_t - w\| \leq Cb_2$.

Let $\widehat{B}_n(w)$ be defined analogously as $\widehat{B}_n(w)$ in (B.4) but with $\mathcal{K}_{2,t}(w)\Delta_t(w)$ in place of $\widehat{\mathcal{K}}_{2,t}(w)\widehat{\Delta}_t(w)$. Then

$$\begin{aligned} |\widehat{B}_n(w) - B_n(w)| &= \left| \frac{1}{nb_2^{2d_x+d_1}} \sum_{t=1}^n \widehat{\mathcal{K}}_{2,t}(w)\widehat{\Delta}_t(w) - \frac{1}{nh_2^{2d_x+d_1}} \sum_{t=1}^n \mathcal{K}_{2,t}(w)\Delta_t(w) \right| \\ &\leq \frac{1}{nb_2^{2d_x+d_1}} \sum_{t=1}^n |\mathcal{K}_{2,t}(w)| |\widehat{\Delta}_t(w) - \Delta_t(w)| + \frac{1}{nb_2^{2d_x+d_1}} \sum_{t=1}^n |\widehat{\mathcal{K}}_{2,t}(w) - \mathcal{K}_{2,t}(w)| |\widehat{\Delta}_t(w)|. \end{aligned}$$

The sup of the first term is $o_p(n^{-1/2}b_2^{-(d_x+d_1)/2})$. By the Taylor expansion, the second term is

$$\begin{aligned} &\frac{1}{nb_2^{2d_x+d_1}} \sum_{t=d_2+1}^n |\mathcal{K}_{2,t}^{(1)}(w)| |\widehat{U}_t - U_t| |\widehat{\Delta}_t(w)| + O_p((b_2^{-1}v_{1n})^2) |\widehat{\Delta}_t(w)| \\ &= O_p(b_2^{-1}v_{1n})O_p(b_2^{p_2+1}) + O_p((b_2^{-1}v_{1n})^2)O_p(b_2^{p_2+1}) \\ &= o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}) \quad \text{uniformly in } w \text{ by Assumption A5.} \end{aligned}$$

So $\widehat{B}_n(w) = B_n(w) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2})$ uniformly in w , and by dominated convergence arguments and Assumption A5,

$$\begin{aligned} \int T_{n,2}(w) dQ(u) &= \int l_2' f_w(w) M^{-1} B_n(w) dQ(u) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}) \\ &= l_2' M^{-1} \frac{1}{nb_2^{2d_x+d_1}} \sum_{t=1}^n \int \mathcal{K}_{2,t}(w)\Delta_t(w) f_w^{-1}(w) dQ(u) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}) \\ &= l_2' M^{-1} \frac{1}{nb_2^{2d_x+d_1}} \sum_{t=1}^n \int \mathcal{K}_{2,t}(w) \sum_{|\mathbf{k}|=p_2+1} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}}m)(w)(W_t - w)^{\mathbf{k}} f_w^{-1}(w) dQ(u) \\ &\quad + o_p(b_2^{p_2+1}) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}) \\ &= b_2^{p_2+1} l_2' M^{-1} B \int m^{(p_2+1)}(w) dQ(u) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}). \end{aligned}$$

The proof is complete. \square

Lemma B.3. *Under Assumptions A1–A5,*

$$\int T_{n,3}(w) dQ(u) = o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}).$$

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

$$\begin{aligned} &[\widehat{M}_{n,j,k}(w)]_{l,r} - [M_{n,j,k}(w)]_{l,r} \\ &= \frac{1}{nb_2^{2d_x+d_1}} \sum_{t=d_2+1}^n \left[\left(\frac{\widehat{W}_t - w}{b_2} \right)^{\phi_j(l)+\phi_k(r)} K_2 \left(\frac{\widehat{W}_t - w}{b_2} \right) - \left(\frac{W_t - w}{b_2} \right)^{\phi_j(l)+\phi_k(r)} K_2 \left(\frac{W_t - w}{b_2} \right) \right]. \end{aligned}$$

Expanding the last expression at W_t , we can show

$$\sup_w |[\widehat{M}_{n,j,k}(w)]_{l,r} - [M_{n,j,k}(w)]_{l,r}| = O_p(b_2^{-1}(v_{1n} + b_1^{p_1+1})) = o_p(1).$$

By Masry (1996b), $\sup_w \|M_n(w) - f_w(w)M\| = O_p(b_2 + n^{-1/2}b_2^{-(2d_x+d_1)/2}\sqrt{\log n}) = o_p(1)$. By the triangle inequality, $\sup_w |\widehat{M}_n(w) - f_w(w)M| = o_p(1)$, and by Assumption A1, $\widehat{M}_n^{-1}(w) = O_p(1)$ w.p.a.1. Consequently, the result in Lemma B.1 implies

$$\int T_{n,3}(w) dQ(u) = o_p(1)O_p(n^{-1/2}b_2^{-(d_x+d_1)/2}) = o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}).$$

The proof is complete. \square

Lemma B.4. Under Assumptions A1–A5,

$$\int T_{n,4}(w) dQ(u) = o_p(n^{-1/2}b_2^{-(d_x+d_1)/2}).$$

Proof. The lemma follows from the proofs of Lemmas B.2–B.3 and Assumption A5: $\widehat{B}_n(w) = B_n(w) + o_p(n^{-1/2}b_2^{-(d_x+d_1)/2})$ uniformly in w , $\widehat{M}_n^{-1}(w) = O_p(1)$ w.p.a.1, and $\sup_w |\widehat{M}_n(w) - f_w(w)M| = o_p(1)$. \square

Proof of Theorem 2.2. Write

$$\begin{aligned} (\widehat{g}(x, z_1) - g(x, z_1)) &= n^{-1} \sum_{t=1}^n \widehat{m}(x, z_1, \widehat{U}_t) - \int m(x, z_1, u) dF_u(u) \\ &= \left\{ n^{-1} \sum_{t=1}^n m(x, z_1, U_t) - \int m(x, z_1, u) dF_u(u) \right\} \\ &\quad + \left\{ n^{-1} \sum_{t=1}^n m(x, z_1, \widehat{U}_t) - n^{-1} \sum_{t=1}^n m(x, z_1, U_t) \right\} \\ &\quad + \left\{ n^{-1} \sum_{t=1}^n \widehat{m}(x, z_1, \widehat{U}_t) - n^{-1} \sum_{t=1}^n m(x, z_1, \widehat{U}_t) \right\} \\ &\equiv A_{n,1} + A_{n,2} + A_{n,3}. \end{aligned}$$

It suffices to show that

- (i) $\sqrt{nb_2^{d_x+d_1}} A_{n,1} = o_p(1)$,
- (ii) $\sqrt{nb_2^{d_x+d_1}} A_{n,2} = o_p(1)$, and
- (iii) $\sqrt{nb_2^{d_x+d_1}} (A_{n,3} - b_2^{p_2+1} t_2' M^{-1} B \int m^{(p_2+1)}(x, z_1, u) dF_u(u)) \xrightarrow{d} N(0, a(x, z_1) t_2' M^{-1} \Gamma M^{-1} t_2)$,

where $t_2 = (1, 0, \dots, 0)'$ is an N -vector.

(i) follows because by the central limit theorem for α -mixing processes, $A_{n,1} = O_p(n^{-1/2}) = o_p(n^{-1/2}b_2^{-(d_x+d_1)/2})$. To show (ii), noting that under Assumption A5(i), (B.5) implies that uniformly in z :

$$\widehat{h}_i(z) - h_i(z) = n^{-1} b_1^{-d_z} t_1' [\overline{M}f_z(z)]^{-1} \sum_{s=1}^n \mathcal{K}_{1,s}(z) U_{is} + o_p(n^{-1/2}b_2^{-(d_x+d_2)/2}). \tag{B.12}$$

By the second-order Taylor expansion, (B.12), and Assumption A5,

$$\begin{aligned}
 A_{n,2} &= n^{-1} \sum_{t=1}^n m(x, z_1, \widehat{U}_t) - n^{-1} \sum_{t=1}^n m(x, z_1, U_t) \\
 &= -n^{-1} \sum_{t=1}^n \sum_{i=1}^{d_x} \frac{\partial m(x, z_1, U_t)}{\partial u_i} (\widehat{h}_i(Z_t) - h_i(Z_t)) + o_p((v_{1n} + b_1^{p_1+1})^2) \\
 &= -n^{-2} b_1^{-d_z} \sum_{t=1}^n \sum_{s=1}^n \sum_{i=1}^{d_x} \frac{\partial m(x, z_1, U_t)}{\partial u_i} l_1' [\overline{M} f_z(Z_t)]^{-1} \mathcal{K}_{1,s}(Z_t) U_{is} + o_p(n^{-1/2} b_2^{-(d_x+d_z)/2}) \\
 &\equiv \widetilde{A}_{n2} + o_p(n^{-1/2} b_2^{-(d_x+d_z)/2}).
 \end{aligned}$$

Analogous to the proof of $T_{n,1b} = o_p(n^{-1/2} b_2^{-(d_x+d_1)/2})$ (see (B.8)), we can show $\widetilde{A}_{n2} = O_p(n^{-1} b_1^{-d_z/2}) = o_p(n^{-1/2} b_2^{-(d_x+d_1)/2})$ and hence (ii) follows.

To show (iii), first notice that the proof of Theorem 2.1 implies that uniformly in u ,

$$\begin{aligned}
 \widehat{m}(x, z_1, u) - m(x, z_1, u) &= T_{n,1}(w) + T_{n,2}(w) + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}) \\
 &= \frac{1}{n b_2^{2d_x+d_1}} \sum_{s=1}^n e_s l_2' M^{-1} \mathcal{K}_{2,s}(w) f_w^{-1}(w) \\
 &\quad + b_2^{p_2+1} l_2' M^{-1} B m^{(p_2+1)}(x, z_1, u) + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}).
 \end{aligned} \tag{B.13}$$

Hence

$$\begin{aligned}
 A_{n,3} &= n^{-1} \sum_{t=1}^n \{\widehat{m}(x, z_1, \widehat{U}_t) - m(x, z_1, \widehat{U}_t)\} \\
 &= \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{s=1}^n e_s l_2' M^{-1} \sum_{t=1}^n \mathcal{K}_{2,s}(x, z_1, \widehat{U}_t) f_w^{-1}(x, z_1, \widehat{U}_t) \\
 &\quad + \frac{b_2^{p_2+1}}{n} l_2' M^{-1} B \sum_{t=1}^n m^{(p_2+1)}(x, z_1, \widehat{U}_t) + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}) \\
 &= \frac{b_2^{p_2+1}}{n} l_2' M^{-1} B \sum_{t=1}^n m^{(p_2+1)}(x, z_1, U_t) \\
 &\quad + \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{s=1}^n e_s l_2' M^{-1} \sum_{t=1}^n \mathcal{K}_{2,s}(x, z_1, U_t) f_w^{-1}(x, z_1, U_t) \\
 &\quad + \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{s=1}^n e_s l_2' M^{-1} \sum_{t=1}^n \{\mathcal{K}_{2,s}(x, z_1, \widehat{U}_t) f_w^{-1}(x, z_1, \widehat{U}_t) - \mathcal{K}_{2,s}(x, z_1, U_t) f_w^{-1}(x, z_1, U_t)\} \\
 &\quad + \frac{b_2^{p_2+1}}{n} l_2' M^{-1} B \sum_{t=1}^n \{m^{(p_2+1)}(x, z_1, \widehat{U}_t) - m^{(p_2+1)}(x, z_1, U_t)\} + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}) \\
 &\equiv A_{n,3a} + A_{n,3b} + A_{n,3c} + A_{n,3d} + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}).
 \end{aligned} \tag{B.14}$$

By the ergodic theorem,

$$A_{n,3a} = b_2^{p_2+1} l_2' M^{-1} B \int m^{(p_2+1)}(x, z_1, u) dF_u(u) + o_p(b_2^{p_2+1}). \tag{B.15}$$

Let $\zeta_t \equiv (X_t, Z_t', U_t', e_t)'$, $\psi_0(\zeta_t, \zeta_s) = e_s l_2' M^{-1} \mathcal{K}_{2,s}(x, z_1, U_t) f_w^{-1}(x, z_1, U_t)$, $\bar{\psi}_0(\zeta_t, \zeta_s) = \psi_0(\zeta_t, \zeta_s) - E_t[\psi_0(\zeta_t, \zeta_s)]$, and $\widetilde{\psi}(\zeta_t, \zeta_s) = (\bar{\psi}_0(\zeta_t, \zeta_s) + \psi_0(\zeta_s, \zeta_t))/2$, where E_t denotes expectation with respect to ζ_t . By construction,

$\bar{\psi}(\cdot, \cdot)$ is symmetric and $E_t \bar{\psi}(\zeta_t, \zeta_s) = 0$. Then

$$\begin{aligned} A_{n,3b} &= \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{s=1}^n \sum_{t=1}^n \psi_0(\zeta_t, \zeta_s) \\ &= \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{s=1}^n \sum_{t=1}^n E_t[\psi_0(\zeta_t, \zeta_s)] + \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{t=1}^n \bar{\psi}_0(\zeta_t, \zeta_t) + \frac{2}{n^2 b_2^{2d_x+d_1}} \sum_{1 \leq t < s \leq n} \bar{\psi}(\zeta_t, \zeta_s) \\ &\equiv A_{n,3b}^{(1)} + A_{n,3b}^{(2)} + A_{n,3b}^{(3)}. \end{aligned}$$

By the central limit theorem for α -mixing processes,

$$\begin{aligned} \sqrt{nb_2^{d_x+d_1}} A_{n,3b}^{(1)} &= \frac{1}{\sqrt{nb_2^{d_x+d_1}}} \sum_{s=1}^n e_s i_2' M^{-1} \int \mathcal{K}_{2,s}(x, z_1, u) f_w^{-1}(x, z_1, u) dF_u(u) \\ &\xrightarrow{d} N(0, a(x, z_1) i_2' M^{-1} \Gamma M^{-1} i_2). \end{aligned}$$

By the ergodic theorem, $A_{n,3b}^{(2)} = O_p(n^{-1} b_2^{-d_x}) = o_p(n^{-1/2} b_2^{-(d_x+d_1)/2})$. By Lemmas A.1 and A.2, we can show that $A_{n,3b}^{(3)} = o_p(n^{-1/2} b_2^{-(d_x+d_1)/2})$. Consequently

$$\sqrt{nb_2^{d_x+d_1}} A_{n,3b} \xrightarrow{d} N(0, a(x, z_1) i_2' M^{-1} \Gamma M^{-1} i_2). \tag{B.16}$$

Next, by the second-order Taylor expansion and (B.12),

$$\begin{aligned} A_{n,3c} &= \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{s=1}^n e_s i_2' M^{-1} \sum_{t=1}^n \{ \mathcal{K}_{2,s}(x, z_1, \hat{U}_t) f_w^{-1}(x, z_1, \hat{U}_t) - \mathcal{K}_{2,s}(x, z_1, U_t) f_w^{-1}(x, z_1, U_t) \} \\ &= - \frac{1}{n^2 b_2^{2d_x+d_1}} \sum_{s=1}^n e_s i_2' M^{-1} \sum_{t=1}^n \mathcal{K}_{2,s}^{(1)}(x, z_1, U_t) f_w^{-1}(x, z_1, U_t) (\hat{h}(Z_t) - h(Z_t)) \\ &\quad + O_p((b_2^{-1} v_{1n} + b_2^{-1} b_1^{p_1+1})^2) \\ &= - \frac{1}{n^3 b_1^{d_z} b_2^{2d_x+d_1}} \sum_{s=1}^n \sum_{t=1}^n \sum_{l=1}^n \sum_{i=1}^{d_x} e_s i_2' M^{-1} \frac{\partial \mathcal{K}_{2,s}(x, z_1, U_t)}{\partial u_i} f_w^{-1}(x, z_1, U_t) \\ &\quad \times i_1' [\bar{M} f_z(Z_t)]^{-1} \mathcal{K}_{1,l}(z) U_{il} + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}) \\ &\equiv \tilde{A}_{n,3c} + o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}). \end{aligned}$$

Using Lemma A.1 repeatedly together with Lemma A.3, we can show that $E[\tilde{A}_{n,3c}] = o(n^{-1/2} b_2^{-(d_x+d_1)/2})$ and $E[\tilde{A}_{n,3c}^2] = o(n^{-1} b_2^{-(d_x+d_1)})$. It follows from the Chebyshev inequality that

$$\tilde{A}_{n,3c} = o_p(n^{-1/2} b_2^{-(d_x+d_1)/2}). \tag{B.17}$$

Now, by the Lipschitz continuity of $m^{(p_2+1)}(\cdot)$,

$$|A_{n,3d}| \leq C b_2^{p_2+1} \max_{1 \leq t \leq n} \|\hat{U}_t - U_t\| = o_p(b_2^{p_2+1}). \tag{B.18}$$

Combining (B.14)–(B.15), we have proved (iii). \square

References

Brown, D.J., Matzkin, R., 1998. Estimation of nonparametric functions in simultaneous equations models with an application to consumer demand. Cowles Foundation Working Paper, Yale University.
 Cai, Z., 2002. A two-stage approach to additive time series models. *Statistica Neerlandica* 56, 415–433.
 Darolles, S., Florens, J.P., Renault, E., 2000. Nonparametric instrumental regression. Manuscript, GREMAQ, University of Toulouse, April.
 Fan, J., 1992. Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87, 998–1004.

- Fan, J., Gijbels, I., 1996. *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.
- Farrell, L., Morgenroth, E., Walker, I., 1999. A time series analysis of U.K. lottery sales: long and short run price elasticities. *Oxford Bulletin of Economics and Statistics* 61, 513–626.
- Gao, J., King, M., 2001. Estimation and model specification testing in nonparametric and semiparametric regression models. Technical Report, Department of Mathematics and Statistics, University of Western Australia.
- Hastie, T.J., Tibshirani, R.J., 1990. *Generalized Additive Models*. Chapman & Hall, New York.
- Horowitz, J.L., 1999. Nonparametric estimation of a generalized additive model with an unknown link function. Mimeo, Department of Economics, University of Iowa.
- Horowitz, J.L., 2001. Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica* 69, 499–513.
- Horowitz, J.L., 2005. Asymptotic normality of a nonparametric instrumental variables estimator. Working paper, Department of Economics, Northwestern University.
- Imbens, G.W., Newey, W.K., 2006. Identification and estimation of triangular simultaneous equations models without additivity. Preprint, MIT.
- Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. *Statistica Sinica* 14, 485–512.
- Li, Q., Wooldridge, J.M., 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., Härdle, W., 1996. Estimating additive regression models with known links. *Biometrika* 83, 529–540.
- Lundberg, M., Squire, L., 2003. The simultaneous evolution of growth and inequality. *Economic Journal* 113, 326–344.
- Mammen, E., Park, B.U., 2005. Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics* 33, 1260–1294.
- 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.
- Newey, W.K., Powell, J.L., 1989. Instrumental variables estimation for nonparametric models. Manuscript, Department of Economics, Princeton University.
- Newey, W.K., Powell, J.L., 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565–1578.
- Newey, W.K., Powell, J.L., Vella, F., 1999. Nonparametric estimation of triangular simultaneous equation models. *Econometrica* 67, 565–603.
- Nielsen, J.P., Sperlich, S., 2005. Smooth backfitting in practice. *Journal of the Royal Statistical Society B* 67, 43–61.
- Opsomer, J.D., Ruppert, D., 1998. A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association* 93, 605–619.
- Pagan, A., Ullah, A., 1999. *Nonparametric Econometrics*. Cambridge University Press, Cambridge.
- Pinkse, J., 2000. Nonparametric two-step regression estimation when regressors and errors are dependent. *Canadian Journal of Statistics* 28, 289–300.
- Racine, J., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119, 99–130.
- Roehrig, C.S., 1988. Conditions for identification in nonparametric and parametric models. *Econometrica* 56, 433–447.
- Severance-Lossin, E., Sperlich, S., 1999. Estimation of derivatives for additive separable models. *Statistics* 33, 241–265.
- Sperlich, S., Linton, O.B., Härdle, W., 1999. Integration and backfitting methods in additive models—finite sample properties and comparison. *Test* 8, 419–458.
- Stone, C.J., 1980. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* 8, 1348–1360.
- Su, L., Ullah, A., 2006. More efficient estimation in nonparametric regression with nonparametric autocorrelated errors. *Econometric Theory* 22, 98–126.
- Sun, S., Chiang, C.-Y., 1997. Limiting behavior of the perturbed empirical distribution functions evaluated at U-statistics for strongly mixing sequences of random variables. *Journal of Applied Mathematics and Stochastic Analysis* 10, 3–20.
- Xiao, Z., Linton, O.B., Carroll, R.J., Mammen, E., 2003. More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of American Statistical Association* 98, 980–992.
- Yang, L., Sperlich, S., Härdle, W., 2003. Derivative estimation and testing in generalized additive models. *Journal of Statistical Planning and Inference* 115, 521–542.
- Zellner, A., Palm, F.C., 2004. *The Structural Econometric Time Series Analysis Approach*. Cambridge University Press, New York.