Data-Driven Bandwidth Selection for Nonstationary Semiparametric Models

Yiguo Sun\textsuperscript{a} & Qi Li\textsuperscript{b}

\textsuperscript{a} Department of Economics, University of Guelph, Guelph, ON, Canada N1G 2W1
\textsuperscript{b} Department of Economics, Texas A&M University, College Station, TX 77843-4228

Published online: 24 Jan 2012.


To link to this article: http://dx.doi.org/10.1198/jbes.2011.09159

Please scroll down for article.
Data-Driven Bandwidth Selection for Nonstationary Semiparametric Models

Yiguo SUN
Department of Economics, University of Guelph, Guelph, ON, Canada N1G 2W1 (yisun@uoguelph.ca)

Qi Li
Department of Economics, Texas A&M University, College Station, TX 77843-4228 (qi@econmail.tamu.edu)

This article extends the asymptotic results of the traditional least squares cross-validatory (CV) bandwidth selection method to semiparametric regression models with nonstationary data. Two main findings are that (a) the CV-selected bandwidth is stochastic even asymptotically and (b) the selected bandwidth based on the local constant method converges to 0 at a different speed than that based on the local linear method. Both findings are in sharp contrast to existing results when working with weakly dependent or independent data. Monte Carlo simulations confirm our theoretical results and show that the automatic data-driven method works well.

KEY WORDS: Integrated time series; Local constant; Local linear; Semiparametric varying-coefficient model.

1. INTRODUCTION

It is well known that the capability of selecting the optimal smoothing parameter is crucial to the application of nonparametric estimation techniques. For independent and weakly dependent data, plug-in methods and (generalized) cross-validation (CV) methods have been incorporated into many popular software packages, and the corresponding asymptotic theory has been well developed (e.g., Härdle and Marron 1985). However, to the best of our knowledge, no asymptotic analysis exists for studying the performance of bandwidth selection methods when integrated time series data are involved, even though nonparametric/semiparametric estimation of regression models with integrated processes has recently attracted much attention among econometricians and statisticians (see, e.g., Juhl 2005; Karlsen, Myklebust, and Tjostheim 2007; Cai, Li, and Park 2009; Xiao 2009). This article aims to fill this gap.

We focus on a particular type of semiparametric model, the semiparametric varying-coefficient model with integrated regressors. Cai, Li, and Park (2009) and Xiao (2009) derived the asymptotic properties of kernel estimators for this model, but used ad hoc methods for selecting the smoothing parameter. In this article we suggest using the least squares CV (LS–CV) method to choose the smoothing parameter and examine the asymptotic properties of this data-driven method–selected smoothing parameter. We consider the following semiparametric model:

\[ Y_t = X_t^T \beta(Z_t) + u_t, \quad 1 \leq t \leq n, \]  

(1.1)

where \( Y_t, Z_t \), and \( u_t \) are scalars, \( Z_t \) and \( u_t \) are stationary variables with \( E(u_t|Z_t) = 0 \) for all \( t \), \( X_t \) is a \( p \times 1 \) vector containing some nonstationary components, and \( \beta(Z_t) \) is a \( p \times 1 \) vector of unspecified smooth functions. In situations where all of the variables are weakly dependent or iid, model (1.1) was considered by Robinson (1989) and Cai (2007). Model (1.1) adds extra flexibility to a linear cointegrating regression model by allowing the cointegrating vector to be smooth functions of some stationary variables. If some or all of the coefficient functions are constant, then model (1.1) becomes a partially linear or linear cointegrating regression model. Thus our asymptotic result also covers the popular partially linear model; see Section 2.4 for a more detailed discussion of this.

To estimate the unknown coefficient curve \( \beta(z) \) in model (1.1), we apply both the local constant and the local linear regression approaches. The local constant (LC) estimator of \( \beta(z) \) is given by

\[ \hat{\beta}(z) = \left( \sum_{i=1}^{n} X_i X_i^T K_{h,tz} \right)^{-1} \sum_{i=1}^{n} X_i Y_i K_{h,tz}, \]  

(1.2)

where \( K_{h,tz} = h^{-1} K(\frac{Z_t-z}{h}) \), \( K(\cdot) \) is a kernel function, and \( h \) is the smoothing parameter. The local linear estimator of \( \beta(z) \), along with the estimator of its derivative function \( \hat{\beta}^{(1)}(z) = \frac{d\beta(z)}{dz} \), is given by

\[ \left( \begin{array}{c} \hat{\beta}(z) \\ \hat{\beta}^{(1)}(z) \end{array} \right) = \left[ \sum_{i=1}^{n} K_{h,tz} \left( \begin{array}{c} 1 \\ Z_t-z \\ (Z_t-z)^2 \end{array} \right) \otimes (X_i X_i^T) \right]^{-1} \times \sum_{i=1}^{n} K_{h,tz} \left( X_i (Z_t-z) \right) Y_i, \]  

(1.3)

where \( \otimes \) denotes the Kronecker product.

In this article we are interested in studying the asymptotic properties of the LS–CV–selected bandwidth for both the local constant and the local linear estimator defined earlier when \( X_t \) contains integrated covariates. We show that the CV-selected bandwidth is stochastic even asymptotically, and that the CV-selected bandwidth for the local constant estimator and the local...
linear estimator converge to 0 at different speeds. Both findings are in sharp contrast to the existing results obtained for independent and weakly dependent data cases. We show that the local linear estimation method has smaller estimated mean squared error than the local constant estimation method. We also show that the asymptotic properties of the CV-selected bandwidth remain unchanged when the I(1) regressor is correlated with the error term \(u_t\), provided that the correlation between the I(1) regressor and the error term \(u_t\) is not overly persistent [see Equation (2.7) for the detailed condition].

To simplify notation/proofs without affecting the essence of our results, we give our theories and proofs for scalar cases; that is, \(X_t\) is a scalar I(1) variable in Section 2, and \(X_t = (X_{t1}, X_{t2})^T\) is a 2 × 1 vector, with \(X_{t1}\) an I(0) variable and \(X_{t2}\) an I(1) variable in Section 3. Allowing higher dimensions in \(X_t\) will only make the mathematical representation of the asymptotic results more complicated without providing any additional insight into the problem.

The rest of the article is organized as follows. Section 2 describes the cross-validation method and derives asymptotic results when \(X_t\) in model (1.1) is an I(1) variable. Section 3 provides asymptotic analysis when \(X_t = (X_{t1}, X_{t2})^T\), where \(X_{t1}\) is an I(0) variable and \(X_{t2}\) an I(1) variable. Section 4 presents Monte Carlo simulations. We relegate all mathematical proofs to two appendices.

2. CROSS–VALIDATION METHOD WITH AN I(1) COVARIATE

2.1 The CV Function and Regularity Conditions

Let \(\hat{h}\) be the data-driven bandwidth selected to minimize the following LS–CV function:

\[
CV(h) = \frac{1}{n^2} \sum_{t=1}^{n} [Y_t - X_t^T \hat{\beta}_{-t}(Z_t)]^2 M(Z_t),
\]

(2.1)

where \(0 \leq M(\cdot) \leq 1\) is a nonnegative weight function that trims out observations near the boundary of the support of \(Z_t\) and \(\hat{\beta}_{-t}(Z_t)\) is the leave-one-out local constant or local linear estimator of \(\beta(Z_t)\) defined in Section 1. Apparently, scaling the CV function by \(n^{-2}\) rather than the conventional choice of \(n^{-1}\) is introduced purely for theoretical reasons, and this does not affect the value of the selected bandwidth minimizing (2.1). With a scale of \(n^{-2}\), \(CV(h)\) asymptotically has the same order as \(\int (\hat{\beta}(z) - \beta(z))^2 M(z) dz\), because \(\sup_{1 \leq t \leq n} ||X_t||^2/n = O_p(1)\). Thus, \(CV(h)\) with the scale of \(n^{-2}\), can be roughly viewed as a weighted version of the average squared error of the nonparametric estimator \(\hat{\beta}(\cdot)\); we provide a more detailed discussion of this point later.

To simplify notation, we write \(\beta_t = \beta(Z_t)\), \(\hat{\beta}_{-t} = \hat{\beta}_{-t}(Z_t)\), and \(M_t = M(Z_t)\) for all \(t\). Substituting (1.1) into (2.1) gives

\[
CV(h) = \frac{1}{n^2} \sum_{t} [X_t^T (\beta_t - \hat{\beta}_{-t})]^2 M_t + \frac{2}{n^2} \sum_{t} u_t X_t^T (\beta_t - \hat{\beta}_{-t}) M_t + \frac{1}{n^2} \sum_{t} u_t^2 M_t,
\]

(2.2)

where the last term does not depend on \(h\). Thus minimizing \(CV(h)\) over \(h\) is equivalent to minimizing \(CV_0(h)\), where \(CV_0(h)\) consists of the first two terms of \(CV(h)\),

\[
CV_0(h) \overset{\text{def}}{=} n^{-2} \sum_{t} [X_t^T (\beta_t - \hat{\beta}_{-t})]^2 M_t + 2n^{-2} \sum_{t} u_t X_t^T (\beta_t - \hat{\beta}_{-t}) M_t
\]

\[
= CV_{0,1} + 2CV_{0,2},
\]

(2.3)

where the definitions of \(CV_{0,1}\) and \(CV_{0,2}\) should be apparent. In Appendix A we show that \(CV_{0,1}\) is the leading term of \(CV_0(h)\). With \(CV_{0,1} = n^{-1} \sum_{t} [X_t^T (\beta_t - \hat{\beta}_{-t})]^2 M_t\), and by \(\sup_{1 \leq t \leq n} ||X_t||/\sqrt{n} = O_p(1)\), we would expect \(CV_{0,1}\) to have the same order as \(\int (\hat{\beta}(z) - \beta(z))^2 M(z) dz\).

We make the following assumptions:

\(A1\) \(\{Z_t\}\) is a strictly \(\beta\)-mixing stationary sequence of size \(-2 + \delta'/\delta'< \delta < 1\), and \(E[Z_t^2] < \infty\). Define \(M = \{z \in \mathbb{R} : M(z) > 0\}\) [the support of the weight function \(M(\cdot)\)]. We require that \(M\) be a compact subset of \(R\). Let \(f(z)\) be the pdf of \(Z_t\). Then \(f(z)\) has bounded derivatives (uniformly in \(z \in M\)) up to the fourth order, and \(\inf_{z \in M} f(z) > 0\).

\(A2\) \(\beta(z)\) is not a linear function and is continuously differentiable up to the fourth order over \(\mathbb{Z} \in M\).

\(A3\) Let \(F_{\mathbb{Z}} = \sigma(X_{t1}, Z_{t1}, u_i ; i \leq t)\) be the smallest sigma field containing the past history of \((X_{t1}, Z_{t1}, u_i)_{i=0}^t\). Here \((u_t, F_{\mathbb{Z}})\) is a martingale difference sequence with \(E(u_1^2 | F_{t-1} ) = \sigma_u^2 < \infty\) for all \(t\) and \(\sup_{t} E(|u_t|^q | F_{t-1}) < \infty\) for some \(q > 2\). In addition, the error terms \(u_t\) are independent of \((X_t, Z_t)_{i=1}^n\).

\(A4\) The kernel function \(K(\cdot)\) is a symmetric (around 0) bounded probability density function on the interval \([-1, 1]\).

\(A5\) \(nh^3(\ln n)^2 \rightarrow 0\), \(h(\ln n)^2 \rightarrow 0\), and \(n^{1-h} \rightarrow \infty\) for some (arbitrarily) small \(\epsilon > 0\), as \(n \rightarrow \infty\).

\(A6\) Let \(v_t = \Delta X_t = X_t - X_{t-1}\) and \(\eta_t(z) = e_t(z) - E(e_t(z))\), where \(e_t(z) = (\beta(Z_t) - \beta(z)) K_{h,t}\), and \(K_{h,t} = h^{-1} \times K(Z_t - h)/h\). The partial sums of the vector process \((v_t, \eta_t, u_t K_{h,t})\) follow a multivariate invariance principle.

\[
\begin{bmatrix}
B_{n,t}(r) \\
B_{n,\beta,z}(r) \\
B_{n,u,z}(r)
\end{bmatrix} \overset{\text{def}}{=} \begin{bmatrix}
\sum_{i=1}^{[nr]} n^{1/2} [\eta_i(z)] \\
(nh)^{-1/2} \sum_{i=1}^{[nr]} n^{1/2} [\eta_i(z)] \\
\sqrt{\frac{1}{nh}} \sum_{i=1}^{[nr]} n^{1/2} [u_i K_{h,t}]
\end{bmatrix}
\]

\[
\overset{\text{BM}(\Omega, \Omega)}{\rightarrow}
\]

(2.4)

where \(\text{BM}(0, \Omega)\) denotes a Brownian motion with mean 0 and finite nonsingular variance–covariance matrix \(\Omega\). Here \([a]\) is the integer part of \(a\) and \(r \in [0, 1]\).

\(A6^*\) On a suitable probability space, there exists a vector Brownian process, \(\text{BM}(0, \Omega)\), with mean 0 and finite
nonsingular variance–covariance matrix such that

\[ \sup_{r \in [0,1]} |B_{n,\epsilon}(r) - B_{\epsilon}(r)| = o_p(1), \]

\[ \sup_{1 \leq t \leq n} |X_t| = O(\sqrt{n \ln \ln n}) \]

almost surely; \hspace{1cm} (2.5)

\[ \sup \sup_{z \in \mathcal{M}} \left[ \left| B_{n,\beta,z}(r) \right| - \left| B_{\beta,z}(r) \right| \right] = o_p(1). \hspace{1cm} (2.6) \]

Assumptions A1 and A2 impose a \( \beta \)-mixing weak dependence condition on \( Z_t \) and some moments and smoothness conditions on \( f(z) \) and \( \beta(z) \). Assumption A3 assumes that \( u_t \) is a martingale difference process independent of \( (X_t, Z_t)_{t=1}^n \), and this assumption significantly simplifies our proofs. Later we discuss how to relax Assumption A3 in two ways: (a) allowing for \( u_t \) to be a stationary mixing process in Assumption A3', which relaxes the martingale difference assumption, and (b) allowing \( u_t \) to be correlated with \( (X_t, Z_t) \) in Assumption A3", which removes the independence assumption. Assumptions A4 and A5 impose mild conditions on the kernel function and the bandwidth \( h \). The kernel function with a compact support is not essential and can be removed at the cost of a lengthy proof. Assumption A5 implies that \( nh/\ln(n) \) \( \rightarrow \infty \) for any constant \( d > 0 \).

In Assumption A6, the weak convergence of the vector of the partial sums holds under some standard regularity conditions, such as strong mixing of \( \{v_i, Z_i, u_i\} \) with some moment conditions. Assumption A6* imposes strong convergence results on the partial sums. Similar conditions were used by Wang and Phillips (2009a). Note that Equation (2.6) in Assumption A6* requires a uniform convergence result of \( (u_{n,\beta,z}(r), B_{n,\epsilon,z}(r)) \) over \( z \in \mathcal{M} \). Equation (2.5) is a result of the functional law of the iterated logarithm (see Rio 1995).

Under Assumptions A1–A3, \( B_{\beta}(r) \), \( B_{\epsilon,\beta}(r) \), and \( B_{n,\beta}(r) \) are independent of one another because the asymptotic covariances between each pair of the partial sums \( B_{\beta}(r), B_{\epsilon,\beta}(r) \), and \( B_{n,\beta}(r) \) are 0, with the variances of \( B_{\beta}(r) \) and \( B_{\epsilon,\beta}(r) \) given by \( \sigma_\beta^2(z) = \lim_{n \to \infty} \text{var} \left( \sum_{i=1}^{\lfloor n h \rfloor} u_i K_{h,z}(z) \right) = v_2(K) \times [\beta(1)]^2 f(z) \) and \( \sigma_\epsilon^2(z) = \lim_{n \to \infty} \text{var} \left( \sqrt{n} \sum_{i=1}^{\lfloor n h \rfloor} u_i K_{h,z}(z) \right) = \sigma_\epsilon^2 v_0(K) f(z) \), respectively, with \( v_2(K) = \int dK^2(u) \, du \). In Section 2.3, we show that the independence between \( B_{\beta}(r) \) and \( B_{\epsilon,\beta}(r) \) also holds under the “weaker” Assumption A3". Write \( W_\beta(r) = B_{\beta}(r)/\sigma_\beta(z) \) and \( W_\epsilon(r) = B_{\epsilon,\beta}(r)/\sigma_\epsilon(z) \). Then \( (W_\beta(r), W_\epsilon(r)) \) is a bivariate standard Brownian motion vector independent of the stochastic process \( B_{\beta}(r) \).

The foregoing assumptions exclude the cases where the error term \( u_t \) is serially correlated and the integrated variables are endogenous. These assumptions are imposed to simplify the proofs and can be replaced by the weaker Assumption A3', or by Assumption A3".

A3' Same as Assumption A3 with this exception: Instead of assuming that \( u_t \) is a martingale difference process, \( u_t \) is a strictly stationary \( \alpha \)-mixing process with mean 0, variance \( \sigma_u^2 < \infty \), and mixing coefficients \( \alpha(t) \equiv \alpha \), satisfying \( \alpha_t = O(\tau^{-q}) \) for some \( p > \delta_1/\delta_2 - 2 \) and \( \delta_1 > 2 \). Also, \( E(|u_t|^{\delta_1}) \) is finite.

The following (endogeneity) assumption is motivated by Saikkonen’s (1991) idea. Assumption 2 of Wang and Phillips (2009b) is in a similar spirit but more general than ours, because their assumption 2 allows a potential nonlinear relation between \( u_t \) and \( v_t \).

A3" The following representation of \( u_t \) allows \( X_t \) to be endogenous:

\[ u_t = \sum_{j=-k_0}^{k_0} \omega_j v_{t-j} + \epsilon_t, \hspace{1cm} (2.7) \]

where \( \{\epsilon_t\}_{t=1}^\infty \) is independent of \( \{(v_i, Z_i)\}_{i=1}^\infty \) and \( k_0 \) is a nonnegative integer. In addition, \( \omega_j = (\epsilon_j, v_j)^T \) is a strictly stationary \( \beta \)-mixing process with mean 0 and mixing coefficients \( \beta(\tau) \equiv \beta \), satisfying \( \beta_t = O(\tau^{-q}) \) for some \( q > 2(1 + \delta'/\delta') \) and \( 0 < \delta' < \delta < 1 \), and \( \omega_1 \) has a finite fourth moment. In addition, \( \lim \text{var}(n^{-1/2} \sum_{j=1}^{n} \omega_j) = \text{finite, positive definite matrix} \). Moreover, \( E(u_t^2 | Z_t = z) \) is continuously differentiable up to the second order over \( z \in \mathcal{M} \).

Here we mainly use Assumption A3 (along with other assumptions) to prove the main results of the article. In supplemental Appendix B (available from the authors on request), we briefly discuss how the proofs can be modified so that our results remain valid even when Assumption A3 is replaced by Assumption A3' or by Assumption A3".

2.2 Main Results

In this section we present the asymptotic results of \( \hat{h} \) minimizing (2.1) where the leave-one-out estimator can be either the local constant or the local linear estimator. The results are given in Theorem 2.1 for the local constant estimation case and in Theorem 2.2 for the local linear estimation case.

Theorem 2.1. Let \( \hat{h}_{lc} \) denote the cross-validation–selected bandwidth based on the local constant estimation method. Under Assumptions A1–A5 and A6*, and assuming that \( \beta(z) \) is not a constant function, we have

\[ \text{CV}(\hat{h}) - \frac{1}{n^2} \sum_{t=1}^{n} \alpha^2_t M(z_t) - \text{CV}_{lc,\cdot}(\hat{h}) = o_p(h/n + (n^2 h)^{-1}), \]

where

\[ \text{CV}_{lc,\cdot}(\hat{h}) = \frac{\hat{h}}{n^2} v_2(K) B_{x,2}^{-1} \left( \int M(z) \beta(1) f(z) dz \right) z_{\beta,2} \]

\[ + \frac{1}{n^2 h} v_0(K) \alpha_\beta^2 \int M(z) dz B_{x,2}^{-1} z_{\alpha,1}^2, \hspace{1cm} (2.9) \]

\[ B_{x,2}(z) = \int_0^1 B_{x,2}^2(r) \, dr, \ \text{and} \ \int_0^1 B_{x,2}^2(r) \, dW_r(r) \text{ for } i = \beta, u \text{ and } j = 1, 2; \]

\[ \sqrt{n} \bar{h}_{lc} - \sigma_u \left[ \frac{\xi_{\alpha,1} \alpha_\beta(v_k) \int M(z) dz}{\xi_{\beta,2} v_2(K) \int M(z) \beta(1) f(z) dz} \right] - \frac{\alpha_\beta^2}{2} \int_0^1 B_{x,2}^2(r) \, dr. \]

\[ (2.10) \]
The proof of Theorem 2.1 is given in Appendix A. Theorem 2.1(i) states that, apart from a term \( (n^{-2} \sum u_i^2 M_i) \) that does not depend on \( h \), \( CV_{lc,L}(h) \) is the leading term of \( CV(h) \). This leading term consists of two parts: the \( O_p(h/n) \) term corresponding to the leading bias term, and the \( (n^2h)^{-1} \) term from the leading variance term. The selected bandwidth balances these two terms, and we obtain \( \hat{h}_{lc} = O_p(n^{-1/2}) \), as stated in Theorem 2.1(ii).

A fundamental difference between the result presented in Theorem 2.1 and previously reported results (when dealing with independent or weakly dependent data) is that the “optimal” bandwidth is stochastic even asymptotically. More specifically, let the CV-selected bandwidth be \( \hat{h}_{lc} = \hat{c} n^{-\alpha} \). With weakly dependent or independent data, it is well known that \( \alpha = 1/5 \) and \( \hat{c} = c_{opt} \), where \( c_{opt} > 0 \) is a nonstochastic (optimal) constant so that \( \hat{h}_{lc} / h_{opt} \rightarrow 1 \), and \( h_{opt} = c_{opt} n^{-1/5} \) is the nonstochastic benchmark (optimal) bandwidth (see Härdle, Hall, and Marron 1992). In contrast, when \( X_t \) is an \( I(1) \) process, Theorem 2.1 states that \( \alpha = 1/2 \), and that \( \hat{c} \) does not converge to a nonstochastic constant, but instead \( \hat{c} \) has a well-defined nondegenerate limiting distribution. Simulation results in Section 4 confirm the theoretical results of Theorem 2.1. There we show that, as the sample size \( n \) increases, \( \hat{h}_{lc} \) shrinks to 0, whereas \( \hat{c} \) has a stable nondegenerate distribution. The next theorem describes the asymptotic behavior for the LL–CV–selected bandwidth.

**Theorem 2.2.** Let \( \hat{h}_{ll} \) denote the CV-selected bandwidth based on the local linear estimation method. Under Assumptions A1–A5 and A6*, we have

\[
\begin{align*}
(i) \quad CV(h) & = -n^{-2} \sum_{i=1}^{n} u_i^2 M_i - CV_{ll,L}(h) = o_p(h^4 + (n^2h)^{-1}), \\
(ii) \quad CV_{ll,L}(h) & = (4/3) h^4 B_{ls,(2)} \kappa_2^2 E(\{\beta^2\}_2^2 M_1) \\
& \quad + (n^2h)^{-1} \nu_0(K) \sigma_u^2 \mathbb{E}[\hat{c}^{-1} u_{l1,1}^2] \int M(z) dz,
\end{align*}
\]

where

\[
\kappa_2 = \int v^2 K(v) dv, \quad \text{and} \quad \beta_{l1} = d^2 \beta(z)/dz^2|_{z=\nu};
\]

\[
\nu_0(K) \sigma_u^2 \mathbb{E}[\hat{c}^{-1} u_{l1,1}^2] \int M(z) dz \stackrel{p}{\rightarrow} 0.
\]

The proof of Theorem 2.2 is given in supplemental Appendix B, where we show that \( CV_{ll,L}(h) = O_p(h^4 + (n^2h)^{-1}) \). The \( O_p(h^4) \) term corresponds to the leading bias term, and the \( O_p(n^2h)^{-1} \) term is the leading variance term. The “optimal” \( h \) balancing the two terms has an order of \( n^{-2}/5 \). We explain why the leading bias term from the LL estimation method differs from that obtained from the LC estimation method in Section 2.3.

Theorems 2.1 and 2.2 imply that \( CV_{lc,L}(\hat{h}_{lc}) = O_p(n^{-3/2}) \) and \( CV_{ll,L}(\hat{h}_{ll}) = O_p(n^{-8/5}) \), respectively. Thus the LC–CV method gives stochastically a larger average squared error than that obtained from the LL–CV method, indicating that the LL–CV method dominates the LC–CV method. This is in sharp contrast to the existing results obtained for independent or weakly dependent data, because it is well known that for independent or weakly dependent data cases, the CV functions for the local constant and the local linear methods have the same rate of convergence.

Note that the result of Theorem 2.1 requires that \( \beta(z) \) be a nonconstant function, and that Theorem 2.2 assumes that \( \beta(z) \) is nonlinear in \( z \). Here we briefly comment on what happens if these assumptions are violated. First, if \( \text{Pr}(\beta(Z_t) = c) = 1 \) for some constant \( c \), then the true model reduces to a linear cointegration model. Ideally, one would like to select a sufficiently large \( h \) in this case, because when \( h = +\infty \), \( \hat{\beta}(z) \) becomes the least squares estimator of the constant parameter \( c \). However, it can be shown that neither \( \hat{h}_{lc} \) nor \( \hat{h}_{ll} \) will converge \( \infty \) in this case. Moreover, \( h \) will not converge to 0, so our Theorems 2.1 and 2.2 do not cover the case where \( \beta(\cdot) \) is a constant function.

We conjecture that the CV-selected bandwidth has a tendency to take large positive values but will not diverge to infinity even as \( n \rightarrow \infty \). Simulations reported in Section 4 support our conjecture. The asymptotic behavior of the CV-selected bandwidth when the true regression model is linear (\( \beta(\cdot) \) is a constant) is quite complex, and it is beyond our present capabilities to derive the asymptotic distribution of the CV-selected bandwidth in this case.

Next, if \( \text{Pr}(\beta(Z_t) = a + bZ_t) = 1 \), or \( \beta(z) \) is a linear function in \( z \), then it can be shown that in this case, Theorem 2.1 still holds true so that \( \hat{h}_{lc} = O_p(n^{-1/2}) \) (\( \hat{h}_{lc} \) still converges to 0), whereas \( \hat{h}_{ll} \) converges to neither 0 nor \( \infty \). In Section 4 we use simulations to investigate the behavior of \( \hat{h}_{cv} \) when \( \beta(z) = a \) and \( \beta(z) = a + bZ_t \). We also examine a spurious regression case where \( \beta(z) \equiv 0 \) and \( u_t \) is an \( I(1) \) process. The theoretical investigation of \( \hat{h}_{cv} \) under a spurious regression model is quite complicated and is beyond the scope of this article.

### 2.3 Endogenous Regressor Case

For a linear cointegration model, it is well known that when \( X_t \) and \( u_t \) are correlated, the ordinary least squares (OLS) estimator of the cointegrating coefficient is still \( n \)-consistent but has an additional bias term of order \( n^{-1} \) (see Phillips and Hansen 1990; Phillips 1995). Recently, Wang and Phillips (2009b, theorem 3.2) considered a nonparametric cointegration model and showed that the asymptotic analysis remains unchanged even when \( X_t \) is correlated with \( u_t \), provided that the correlation is not overly persistent. However, Wang and Phillips’ framework is quite different from the semiparametric model that we consider here, because in their model, the nonstationary variable enters the model nonparametrically, whereas in our semiparametric model, the nonparametric component \( Z_t \) is a stationary variable. To the best of our knowledge, in the framework of a semiparametric varying-coefficient model with \( I(1) \) regressors, the endogeneity issue has not been addressed. A priori, it is not clear whether an endogenous regressor will lead to a nonnegligible bias term. In this section we show that the asymptotic distribution of \( \hat{\beta}(z) \) remains unchanged when \( X_t \) is correlated with \( u_t \) under some standard regularity conditions. For expositional simplicity, we consider only the local constant estimator in this section; a similar result can be shown to hold true for the local linear estimator.
Let $K_{h,t} = h^{-1}K((Z_t - z)/h)$. Then the local constant estimator of $\beta(z)$ is given by

$$
\hat{\beta}(z) = \beta(z) + \left(\frac{1}{n^2} \sum_{t=1}^{n} X_t^2 K_{h,t}\right)^{-1} \times \left\{ \frac{1}{n^2} \sum_{t=1}^{n} X_t^2 (\beta(Z_t) - \beta(z)) K_{h,t} + \frac{1}{n^2} \sum_{t=1}^{n} X_t u_t K_{h,t} \right\}
$$

$$\equiv \beta(z) + A_{3n}^1 (A_{2n} + A_{3n}), \quad (2.14)$$

where the definitions of $A_{jn}$ should be apparent ($j = 1, 2, 3$). Allowing $X_t$ and $u_t$ to be correlated may affect the asymptotic behavior only of $A_{3n}$ because other terms do not depend on $u_t$. We show that the asymptotic behavior of $A_{3n}$ remains unchanged when Assumption A3 is replaced by Assumption A3'. Therefore, our result remains the same even when $X_t$ is correlated with $u_t$.

Using $X_t = \sum_{s=1}^{t} v_t$, we have $A_{3n} = n^{-2} \sum_{t=2}^{n} X_{t-1} u_t K_{h,t} + n^{-2} \sum_{t=1}^{n} v_t u_t K_{h,t}$, where the stochastic property of $v_t$ is described in Assumption A3'. If $u_t$ is a martingale difference as defined in Assumption A3, or if $X_t$ is strictly exogenous with $E(u_t | Z_t) = 0$, we have $E(A_{3n}) = 0$. However, allowing $X_t$ to be endogenous and $u_t$ to be serially correlated generally leads to $E(A_{3n}) \neq 0$. Assumption A3' ensures that $\sup_{v} E^{(n^{-2})} \sum_{t=1}^{n} |v_t u_t| K_{h,t} = O_p(1)$, and applying Assumption A6 and theorem 4.1 of De Jong and Davidson (2000) gives

$$n^{1/2} A_{3n} = \frac{n^{1/2}}{\sqrt{n}} \sum_{t=1}^{n} X_t u_t \sqrt{h K_{h,t}} \xrightarrow{d} \int_0^1 B_{z}(r) d B_{h,z}(r) + \Lambda, \quad (2.15)$$

where $\Lambda = \lim_{n \to \infty} A_{3n}$ and $\Lambda_n \equiv E(\sqrt{n}/n \sum_{t=1}^{n} X_t u_t K_{h,t})$.

We show $\Lambda = 0$ below. Let $E(|z|) = E(|Z_t|)$. We assume that both $E(v_t^2 | z)$ and $E(u_t^2 | z)$ are bounded by a function of $z$ that has a finite second moment. By Assumptions A1 and A3', $(Z_t, v_t, u_t)_{t=1}^{n}$ is a $\beta$-mixing process. Thus, applying lemma 1 of Yoshihara (1976), we obtain

$$|A_n| \leq \sqrt{n} \sum_{t=1}^{n} \sum_{s=1}^{t-1} \left| E(v_t u_t K_{h,t}) \right|$$

$$\leq M \sqrt{n} \sum_{t=1}^{n} \sum_{s=1}^{t} h^{-d/2(1+\delta)} |t-s|^{-d/2(1+\delta)}$$

$$= O(\sqrt{n} h^{-d/2(1+\delta)} (1 - n^{1-d/2(1+\delta)})) = o(1). \quad (2.16)$$

Hence, $\Lambda_n = o(1)$ and this implies that $\Lambda = \lim_{n \to \infty} \Lambda_n = 0$.

Therefore, the bias term due to the correlation between $X_t$ and $u_t$ is asymptotically negligible, which differs from the linear regression model case. Replacing $K_{h,t}$ by 1 in (2.14), we obtain the OLS estimator of $\beta$. It is easy to see that for a linear cointegration model, the bias term is $A_0 = E(n^{-1} \sum_{t=1}^{n} X_t u_t) \neq 0$. In fact, it is easy to see that if both $v_t$ and $u_t$ are iid series but with $E(v_t u_t) \neq 0$, we have $A_0 = E(v_t u_t) \neq 0$. For our semiparametric model, even if $E(v_t u_t) \neq 0$, we have $A_0 = 0$, because our bias term has an additional factor $\sqrt{h} K_{h,t}$ and $E(\sqrt{h} K_{h,t}) = O(\sqrt{h}) = o(1)$.

We use the foregoing decomposition to provide an intuitive explanation of Theorem 2.1 as to why $CV_{h,L}(h) = O_p(h/n + (n^2 h)^{-1})$. Applying Hansen’s (1992) theorem 3.3 yields

$$A_{1n} = n^{-2} \sum_{t=1}^{n} X_t^2 E(K_{h,t}) + o_p(1)$$

$$d \to f(z) \int_0^1 B_{r}(r) dr = O_c(1).$$

Here the notation $O_c(a_n)$ means an exact probability order of $O_p(a_n)$, but is not $o_p(a_n)$.

Next consider $A_{2n}$. By adding/subtracting terms, we rewrite $A_{2n} = A_{2n,1} + A_{2n,2}$, where

$$A_{2n,1} = n^{-2} \sum_{t=1}^{n} X_t^2 E[(\beta(z_t) - \beta(z)) K_{h,t}]$$

$$= c_{1n} [h^2 \text{Bias}(z) + O(h^3)] \quad (2.17)$$

with $c_{1n} = n^{-2} \sum_{t=1}^{n} X_t^2$. Bias $(z) = (1/2) x z \beta_2^{(2)} f(z) + f^{(1)}(z) \beta_1(z)$, and $K = \int k(v) v^2 dv$. By Assumption A6 and the independence between $B_{z}(r)$ and $W_{h}(r)$, we have

$$A_{2n,2} = \sqrt{n} \sum_{t=1}^{n} X_t^2 \frac{1}{\sqrt{n}} \eta_t(z) + \sqrt{n} \sum_{t=1}^{n} X_t -1 \frac{1}{\sqrt{n}} \eta_t(z)$$

$$+ n^{-2} \sum_{t=1}^{n} \frac{1}{\sqrt{n}} \eta_t(z)$$

$$= \sqrt{n} \sum_{t=1}^{n} X_t^2 \frac{1}{\sqrt{n}} \eta_t(z) + O_p(n^{-1} \sqrt{n} + n^{-1} h^2)$$

$$= \sqrt{n} \int f(z) \beta^{(1)}(z)^2 v_2(K)$$

$$\times \int_0^1 B_{r}(r) d W_{h}(r) + o_p(1). \quad (2.18)$$

where $\eta_t(z) = e_t(z) - E(e_t(z))$ and $e_t(z) = (\beta(Z_t) - \beta(z)) K_{h,t}$.

Equation (2.17) shows that the order of $A_{2n,1}$ is related to $E(e_t(z))$, the mean part of $e_t(z)$, whereas Equation (2.18) tells us that the order of $A_{2n,2}$ is determined by the variance part of $e_t(z)$ as $\eta_t(z)$ has mean 0. With weakly dependent data, it is well known that the mean part of $e_t(z)$ dominates the variance part of $e_t(z)$. However, if the local constant estimation method is used to analyze integrated time series, then the variance part of $e_t(z)$ determines the variance part of $e_t(z)$. To see this, we square each term to get rid of the square root expressions, which gives $A_{2n,1}^2 = O_p(h^4)$, $A_{2n,2}^2 = O_p(h/n)$, and $A_{3n}^2 = O_p(n^2 h^{-1})$. These results imply that $A_{2n,1}^2 + A_{3n}^2 = O_p(h/n + (n^2 h)^{-1})$ has a stochastic order larger than that of $A_{2n,1}^2 = O_p(h^4)$.

The foregoing analysis also provides an explanation as to why the leading term of the CV function, $CV_{h,L}(h)$, contains two terms of orders $O_p(h/n)$ and $O_p((n^2 h)^{-1})$, which are related to the variances of $e_t(z)$ and $u_t K_{h,t}$, respectively. The $O_p(h^4)$ term corresponding to the squared mean of $e_t(z)$ is asymptotically negligible. However, if the local linear estimation method is used instead of the local constant method, $\beta(Z_t) - \beta(z_t)$ must be replaced by $R(Z_t, Z_t) \equiv \beta(z_t) - \beta(z_t)$.
\( \beta(1)(Z_t)(Z_t - Z_t) \). In this case the two leading terms, \( O_p(h^4) \) and \( O_p((n^2h)^{-1}) \) [of \( CV_{\text{LL}}(h) \)], correspond to the (squared) mean of \( R(Z_t, Z_t)K_{h,t,t} \) and the variance of \( u_tK_{h,t,t} \), whereas the term related to the variance of \( R(Z_t, Z_t)K_{h,t,t} \), which has order \( O_p(h^5/n) \) (this term differs from the LC case), is asymptotically negligible. Therefore, the reason for the different convergence rates of \( CV_{\text{LC}}(h) \) and \( CV_{\text{LL}}(h) \) is that, unlike in the weakly dependent data case, the LC and the LL estimator have different stochastic orders in \( A_{\text{LL}} \).

It can be shown that the results of Theorems 2.1 and 2.2 hold true when \( X_t \) and \( u_t \) are correlated, but the proofs will be lengthy and tedious and will provide no additional insight into the problem. Thus to save space, we do not pursue proofs for Theorems 2.1 and 2.2 for the endogenous \( X_t \) case. In Section 4 we report Monte Carlo simulations that allow \( X_t \) to be correlated with \( u_t \). The simulation results reported there support the foregoing theoretical analysis and show that the estimation results are virtually unaffected whether or not \( X_t \) and \( u_t \) are correlated.

### 2.4 A Partially Linear Varying-Coefficient Model

In this section we consider the CV selection of the bandwidth \( h \) when estimating the following partially linear varying-coefficient model:

\[
Y_t = S_t^T \gamma + X_t\beta(Z_t) + u_t, \tag{2.19}
\]

where \( S_t \) is a \((d - 1) \times 1\) vector of \( I(1) \) variables, and \( \gamma \) is a \((d - 1) \times 1\) vector of constant parameters. For expositional simplicity, we assume that \( X_t \) is a scalar \( I(1) \) variable. Note that the foregoing partially linear model is a special case of a general varying-coefficient model \( Y_t = S_t^T \alpha(Z_t) + X_t\beta(Z_t) + u_t \), where the restriction that \( \alpha(z) \equiv \gamma \), a vector of constant parameters, is imposed.

We propose using a profile least squares approach to estimate \( \gamma \). First, we treat \( \gamma \) as if it were known and rewrite (2.19) as \( Y_t = S_t^T \gamma + X_t\beta(Z_t) + u_t \). Then the local constant estimator of \( \beta(Z_t) \) is given by

\[
\hat{\beta}(Z_t) = \left( \sum_s X_s^2 K_{h,s,t} \right)^{-1} \sum_s X_s (Y_s - S_t^T \gamma) K_{h,s,t}.
\]

\( \equiv A_{2t} - A_{1t}^T \gamma \). \tag{2.20}

Replacing \( \gamma \) with \( \hat{\gamma} \) in (2.20), we obtain a feasible leave-one-out estimator of \( \beta(Z_t) \) given by

\[
\hat{\beta}_{-i}(Z_t) = \left( \sum_{s \neq i} X_s^2 K_{h,s,t} \right)^{-1} \sum_{s \neq i} X_s (Y_s - S_t^T \hat{\gamma}) K_{h,s,t}. \tag{2.23}
\]

It can be shown that \( S_t^T (\gamma - \hat{\gamma}) \) has a stochastic order smaller than \( X_t(\beta_t - \hat{\beta}_{-i}) \). Thus the leading term of \( CV_t(\hat{\gamma})(h) = n^{-2} \sum_{s=1}^n \left[ (Y_s - S_t^T \hat{\gamma} - X_s\beta(Z_s)) \right]^2 M_t \) is given by \( CV_t(\gamma)(h) = n^{-2} \sum_{s=1}^n \left[ (Y_s - S_t^T \gamma - X_s\beta(Z_s)) \right]^2 M_t \). Obviously, \( CV_t(\gamma)(h) \) is the same as \( CV(h) \) defined in (2.1). This is because if we define \( \tilde{Y}_t = Y_t - S_t^T \gamma \), then model (2.19) can be written as \( \tilde{Y}_t = X_t\beta_t + u_t \), which is identical to model (1.1). Thus the results of Theorem 2.1 remain valid for the partially linear model (2.19).

The foregoing discussion is based on the local constant estimation method. The local linear estimation method could be used as well. In this case, \( \hat{\beta}(Z_t) \) in (2.20) is replaced by the local linear estimator, which is also linear in \( \gamma \). The remaining estimation steps are similar to the local constant estimation case discussed earlier. The asymptotic behavior of the LS–CV–selected \( h \) is the same as presented in Theorem 2.2.

### 3. \( X_t \) CONTAINS BOTH I(0) AND I(1) COMPONENTS

The results in Section 2 are obtained assuming that \( X_t \) contains only \( I(1) \) components. This assumption simplifies our theoretical analysis and makes it easier to see why the LC and LL estimation methods lead to different rates of convergence. However, in practice, a cointegration model also may include an intercept term and some other stationary variables in addition to some \( I(1) \) regressors. Therefore, in this section we investigate the asymptotic behavior of the LS–CV–selected bandwidth when \( X_t \) contains both \( I(0) \) and \( I(1) \) components. Specifically, we consider \( X_t = (X_{l1}, X_{l2})^T \) with \( X_{l1} \) and \( X_{l2} \) being \( I(0) \) and \( I(1) \) variables, respectively. We replace Assumption A6* by the following assumption:

B Let \( X_t = (X_{l1}, X_{l2})^T \), where \( X_{l2} \) satisfies Assumption A6* given in Section 2 and \( X_{l1} \) is a strictly stationary \( \beta \)-mixing sequence of size \(-2 + \delta'/\delta \) for some \( 0 < \delta < \delta \) with \( E(|X_{l1}^{(2+\delta)}|) < M < \infty \). In addition, both \( E(X_{l1} | Z_t = z) \) and \( E(X_{l2}^2 | Z_t = z) \) are twice continuously differentiable over \( z \in M \).

Given that the local linear method has a smaller asymptotic MSE than the local constant method, we consider only the local linear estimation method in this section.

**Theorem 3.1.** Let \( \hat{h} \) denote the CV-selected bandwidth via the local linear method. Under Assumptions A1–A5 and B, we have

\[
h^{1/2} \hat{h} - c_3 n^{-1/2} \rightarrow_d 0, \tag{3.1}
\]

where \( c_3 \) is a well-defined \( O(n) \) random variable that is determined by (B.1.2) and (B.13) in supplemental Appendix B.

Comparing the result of Theorem 3.1 with that of Cai, Li, and Park (2009), we see that \( \hat{h} \) (defined in Theorem 3.1) is optimal for estimating \( \beta_2(\cdot) \), the coefficient function associated with the integrated variable, because Cai, Li, and Park (2009)
showed that the optimal \( h \) for estimating \( \beta_1(z) \) should have an order of \( n^{-1/5} \). The reason that the CV method selects an optimal \( h \) for estimating \( \beta_2(z) \) is because \( X_t^2, \beta_2(Z_t) \) is the dominant component of \( X_t^2, \beta_1(Z_t) + X_t^2, \beta_2(Z_t) \). Thus, to minimize the CV function, one must choose an \( h \) that is optimal for estimating \( \beta_2(z) \), the coefficient function of the \( I(1) \) covariate. In the next section, we examine a two-step estimation method as suggested by Cai, Li, and Park (2009) via Monte Carlo simulations, where after obtaining \( \hat{\beta}_1(Z_t) \) and \( \hat{\beta}_2(Z_t) \) in the first step, we use \( Y_t - X_t^2, \hat{\beta}_2(Z_t) \) as the new dependent variable and estimate \( \beta_1(Z_t) \) using a new CV-selected smoothing parameter in the second step. We show that this two-step method leads to improved estimation results for \( \beta_1(z) \).

4. MONTE CARLO SIMULATIONS

4.1 Case (a): \( X_t \) Is \( I(1) \)

We consider the following data-generating process (DGP):

\[
Y_t = X_t \beta(Z_t) + u_t, \quad Z_t = 0.5Z_{t-1} + \eta_{3t}, \quad \text{and} \quad u_t = (\eta_{2t} + \eta_{1t})/\sqrt{1 + \theta^2},
\]

where \( X_t = X_{t-1} + \eta_{1t} \), \( (\eta_{1t}, \eta_{2t})^T \) is iid N(0, \( I_2 \)), \( \eta_{3t} \) is iid Uniform[0, 1], and \( X_0 = 0 \). The data-generating mechanism for \( u_t \) is the same as that of Wang and Phillips (2009b). It is easy to show that \( \text{corr}(\Delta X_t, u_t) = \theta/\sqrt{1 + \theta^2} \). We take two values for \( \theta \), \( \theta = 0 \) and \( \theta = 0.2 \), where the former case implies that \( X_t \) and \( u_t \) are independent of one another, whereas the latter case gives that \( \text{corr}(\Delta X_t, u_t) = 0.1961 \). \( Z_t \) is a stationary AR(1) process with the innovation \( \eta_{3t} \) taking values in [0, 1].

We consider four different DGPs and index them by DGP\(_{i,j}\), specifically, \( i = 1 \) if \( \theta = 0 \), \( i = 2 \) if \( \theta = 0.2 \), \( j = 1 \) if \( \beta(z) = 1 + z + 2z^2 \) whose value increases as \( z \) increases, and \( j = 2 \) if \( \beta(z) = \sin(3z) \), which is not monotone and has more curvature than the quadratic function. Comparing DGP\(_{1,1}\) with DGP\(_{2,2}\) for \( j = 1 \) and 2, we aim to show that \( X_t \) is allowed to be correlated with \( u_t \) as discussed in Section 2.3. Comparing DGP\(_{1,1}\) with DGP\(_{2,2}\) in DGP\(_{i,j}\) for \( i = 1, 2 \), we examine how extra curvature of a coefficient function affects the finite-sample performance of our proposed CV method.

The sample sizes are \( n = 100, 200, \) and 400. The number of simulations is \( m = 1000 \). We report the square-root of the average squared error, \( \text{RMSE} = \sqrt{\text{AMSE}} \), and the mean absolute bias, \( \text{MABIAS} \), where for the 4th simulation, we define \( \text{AMSE}_i = n^{-1} \sum_{t=1}^{n} |\hat{\beta}(z_t) - \beta(z_t)|^2 \) and \( \text{MABIAS}_i = n^{-1} \sum_{t=1}^{n} |\hat{\beta}(z_t) - \beta(z_t)| \).

Figure 1 plots the kernel density functions of the CV-selected constant \( \hat{c} \) and the bandwidth \( \hat{h} \), where \( \hat{h} = \hat{c}n^{-\alpha} \) with \( \alpha = 1/2 \) for the LC–CV method and \( \alpha = 2/5 \) for the LL–CV method. Figure 1 is obtained from DGP\(_{2,1}\), where the dotted line represents \( n = 100 \), the dashed line represents \( n = 200 \), and the solid line represents \( n = 400 \). As predicted by Theorems 2.1 and 2.2, we see that the CV-selected bandwidth \( \hat{h} \) becomes smaller as sample size increases, and that \( \hat{c} \) does not converge to a constant as the estimated density function for \( \hat{c} \) is rather stable for different sample sizes.

Table 1 reports the mean and standard deviation (over the 1000 replications) of the RMSE and the MABIAS. Several interesting patterns are observed. First, the LL–CV method has smaller (mean value of) RMSE and MABIAS compared with the LC–CV method. This is consistent with our theory, because the LL–CV method has smaller asymptotic MSE than the LC–CV method, as shown in Theorems 2.1 and 2.2. Second, the estimation efficiency gain of the LL–CV method over the LC–CV method is more pronounced for DGP\(_{1,2}\) than for DGP\(_{2,2}\) for \( i = 1, 2 \) when the unknown curve has more curvature (i.e., more nonlinearity). Third, comparing the results of DGP\(_{1,j}\) with DGP\(_{2,j}\) for \( j = 1 \) and 2 confirms our theoretical analysis presented in Section 2.3 that the CV method is valid even when \( u_t \) and \( X_t \) are contemporaneously correlated.

To show how the CV-selected bandwidth behaves when \( \beta(z) \) is constant, Table 2 reports the first quartile, median, mean, and third quartile of the CV-selected bandwidths, along with the RMSE and the MABIAS for the LC and LL kernel estimators, where \( \beta(z) \equiv 1 \). The results for \( \theta = 0.2 \) and \( \theta = 0 \) are very similar, so we only report the results for \( \theta = 0 \). In addition, the bandwidth has an upper bound of five times the interquartile range of \( Z_t \) (i.e., 9.45), because allowing the bandwidth to increase further will improve the CV value only beyond the sixth decimal point. In this case, with a much larger selected bandwidth, the LC–CV method gives smaller RMSE and MABIAS than the LL–CV method. We also see that the median value of the LC–CV-selected \( h \) is quite stable and does not seem to change as \( n \) increases. This result is consistent with our earlier finding that when \( \beta(z) \) is a constant function, the CV method tends to choose a large value of \( h \), but the CV-selected \( h \) will converge neither to \( \infty \) nor to 0.

Table 3 reports the results when \( X_t \) and \( Y_t \) are two independent random-walk processes without drift (a spurious regression model). In this case the CV\((h)\) objective function is quite flat, and the RMSE and the MABIAS do not change much even when \( h \) changes substantially, resulting in a wide range of selected \( \hat{h}_{cv} \), as shown in Table 3. Taken together, the results in Tables 2 and 3 show that an unusually large LC–CV–selected
̂h may indicate β(z) = c, a constant, or may suggest a spurious relationship between Yt and Xt. Further diagnostics are needed to distinguish these two possibilities.

4.2 Case (b): Xt Contains Both I(0) and I(1) Variables

We consider the following DGP:

\[ Y_t = X_{1t} \beta_1(Z_t) + X_{2t} \beta_2(Z_t) + \epsilon_t, \tag{4.1} \]

where \( X_{2t} = X_{2t,1-t} + \eta_{1t} \), \( X_{1t} = 1 + 0.5X_{1,t-1} + \eta_{4t} \) with \( \eta_{4t} \sim \text{iidN}(0, 1) \), \( \epsilon_t \) and \( Z_t \) are generated as in Section 4.1, \( \beta_1(z) = 1 + z + 2z^2 \), and \( \beta_2(z) = \sin(3z) \). We use DGP3 and DGP4 to denote the cases where \( \theta = 0 \) and \( \theta = 2 \), respectively. For DGP4, \( \text{corr}(\Delta X_{2t}, \epsilon_t) = 0.8944 \), a rather high contemporaneous correlation.

Table 4 presents simulation results for two estimators: \( \hat{\beta}(\cdot) \), the LL–CV estimator using the LL–CV–selected bandwidth \( \hat{h} \), and \( \hat{\beta}_1(\cdot) \), the two-step LL–CV estimator for \( \beta_1(\cdot) \) in which we construct a new dependent variable \( Y_t - X_{2t} \hat{\beta}(Z_t) \) and reestimate \( \beta_1(Z_t) \) in the second stage, where \( \hat{\beta}_2(Z_t) \) is the first-stage estimator of \( \beta_2(Z_t) \). We also selected a new bandwidth via the LL–CV method in the second-stage estimation. We observe the following. First, the coefficient curve \( \beta_2(\cdot) \) associated with the I(1) variable is estimated more accurately than the coefficient curve \( \beta_1(\cdot) \) associated with the I(0) variable. Second, the second-step estimation of \( \beta_1(\cdot) \) by the LL–CV method performs slightly better than the one-step estimation result of \( \beta_1(\cdot) \). Finally, Table 4 also indicates that we do not need the strictly exogenous assumption to validate the CV method.

Appendix A: Proof of Theorem 2.1

We use the notation \( A_n = B_n + (s.o.) \) to denote that \( A_n = B_n + \) terms of smaller order than \( B_n \). We denote \( \beta_1 = \beta(z), \beta_1(\cdot) = \hat{\beta}_1(\cdot)^\circ \) by the LL–CV method per.

| Method | \( n \) | \( h_{0.25} \) | \( h_{0.5} \) | \( h_{0.75} \) | \( \hat{h} \) | RMSE | MABIAS | RMSE | MABIAS |
|--------|--------|----------------|----------------|----------------|--------|--------|--------|--------|
| CV–LC  | 100    | 8.0816         | 8.3528         | 8.3508         | 8.6341 | 0.0166 | 0.0168 | 0.0167 | 0.0166 |
|        | 200    | 8.3893         | 8.6057         | 8.5097         | 8.8187 | 0.008  | 0.0081 | 0.0082 | 0.0077 |
|        | 400    | 8.6099         | 8.7869         | 8.7863         | 8.9654 | 0.0037 | 0.0043 | 0.0041 | 0.0039 |
| CV–LL  | 100    | 8.0823         | 8.3528         | 8.3588         | 8.6346 | 0.0271 | 0.0214 | 0.0234 | 0.0186 |
|        | 200    | 8.3893         | 8.6057         | 8.5091         | 8.8187 | 0.0132 | 0.0104 | 0.0114 | 0.0088 |
|        | 400    | 8.6099         | 8.7869         | 8.7863         | 8.9654 | 0.0066 | 0.0055 | 0.0059 | 0.0046 |

Note that in the cases of independent or weakly dependent data, \( \hat{A}_1^{-1} \hat{B}_t \) and \( \hat{A}_1^{-1} \hat{C}_t \) correspond to the bias and variance terms, respectively. Therefore, for convenience we refer these two terms as bias and variance terms.
Substituting (A.1) into CV_{0,1} defined in (2.3), we obtain

\[ CV_{0,1} = n^{-2} \sum_i (x_i \hat{A}_i^{-1} \hat{B}_i)^2 M_i + n^{-2} \sum_i (x_i \hat{A}_i^{-1} \hat{C}_i)^2 M_i + 2 n^{-2} \sum_i \hat{x}_i^2 \hat{A}_i^{-2} \hat{B}_i \hat{C}_i M_i \]

\[ \equiv CV_1 + CV_2 + 2 CV_3, \quad (A.2) \]

where the definitions of CV_j (j = 1, 2, 3) should be apparent.

In Lemmas A.2–A.4 we show that

\[ CV_1 = (h/n) B_{\|z_1\|^2} v_2(K) E \left[ M_{f,1}^{-1} (\beta_1) \right]^2 \]

\[ \times \left[ \int_0^1 B^2_\beta(r) dW_\beta(r) \right]^2 + o_p(h/n + h^2), \quad (A.3) \]

\[ CV_2 = \frac{v_0(K) \sigma_{\hat{d}}^2}{n^{-h}} B_{\|z_1\|^2}^{-1} \left[ \int_0^1 B_\nu(r) dW_\nu(r) \right]^2 \int M(z) dz \]

\[ + o_p(n^2 h^{-1}), \quad (A.4) \]

\[ CV_3 = o_p(h/n), \quad (A.5) \]

where \( B_{\|z_1\|^2} \) is as defined in (2.3), \( B_\beta(r) \), \( B_\nu(r) \), and \( W_\beta(r) \) and \( W_\nu(r) \) are standard Brownian motions independent of the stochastic process \( B_{\|z_1\|^2}(r) \). Using similar arguments, it can be shown that \([CV_{0,2} = 0 \text{ as defined in (2.3)}] \).

Combining (A.3)–(A.6), we see that the leading term of \( CV_0(h) \) defined in (2.3) is given by

\[ CV_{lc,L}(h) = (h/n) v_2(K) B_{\|z_1\|^2}^{-1} \]

\[ \times E \left[ M_{f,1}^{-1} (\beta_1) \right]^2 \left[ \int_0^1 B^2_\beta(r) dW_\beta(r) \right]^2 \]

\[ + \frac{v_0(K) \sigma_{\hat{d}}^2}{n^{-h}} \left[ \int_0^1 B_\nu(r) dW_\nu(r) \right]^2 \int M(z) dz, \quad (A.7) \]

Obviously, \( CV_{lc,L}(h) \) is minimized at

\[ h_0 = \sigma_n^{-1/2} \left[ \frac{\xi_1^2 v_0(K) \int M(z) dz}{\xi_2^1 v_2(K) \int M(z) (\beta_1)^2 dz} \right], \quad (A.8) \]

where \( \xi_i = \int_0^1 B_i^2(r) dW_i(r) \) for \( i = \beta, u \) and \( j = 1, 2 \).

The foregoing result can be extended to \( \sup_{h \in H_a} |CV_0(h) - CV_{lc,L}(h)| = o_p(n^{-1/2}) \), where \( H_a = (alpha^{-0.6}, beta^{-0.4}) \) for some \( a > 0 \) and \( b > 0 \). This completes the proof of Theorem 2.1.

**Lemma A.1.** Under Assumptions A1, A2, A4, A5, and A6, we have

\[ \sup_{z \in M} \left| \hat{\beta}_1(z) - \tilde{\beta}_1(z) \right| = O_p \left( \frac{(\ln n)^{1/2}}{(mh)^{1/4}} \right), \quad (A.9) \]

where \( \tilde{\beta}_1(z) = (n^{-2} \sum_i \hat{x}_i^2 f(z) \equiv c_{\hat{d}} f(z) \).

**Table 3.** The LC–CV vs. the LL–CV method when \( X_t \) and \( Y_t \) are independent random walk processes

<table>
<thead>
<tr>
<th>Method</th>
<th>( n )</th>
<th>( h_{0.25} )</th>
<th>( h_{0.5} )</th>
<th>( h_{\text{mean}} )</th>
<th>( h_{0.75} )</th>
<th>RMSE Mean</th>
<th>St. dev.</th>
<th>MABIAS Mean</th>
<th>St. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV–LC</td>
<td>100</td>
<td>0.1355</td>
<td>0.3146</td>
<td>2.6275</td>
<td>7.8475</td>
<td>0.7366</td>
<td>0.6379</td>
<td>0.7156</td>
<td>0.6332</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.1347</td>
<td>0.2912</td>
<td>2.5119</td>
<td>7.8823</td>
<td>0.6990</td>
<td>0.6113</td>
<td>0.6863</td>
<td>0.6089</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.1589</td>
<td>0.3265</td>
<td>2.5747</td>
<td>6.3165</td>
<td>0.7014</td>
<td>0.6320</td>
<td>0.6941</td>
<td>0.6307</td>
</tr>
<tr>
<td>CV–LL</td>
<td>100</td>
<td>0.2393</td>
<td>0.7662</td>
<td>3.8878</td>
<td>8.2403</td>
<td>0.7538</td>
<td>0.6425</td>
<td>0.7233</td>
<td>0.6335</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.1943</td>
<td>0.5494</td>
<td>3.5212</td>
<td>8.4263</td>
<td>0.7157</td>
<td>0.6239</td>
<td>0.6923</td>
<td>0.6068</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.2195</td>
<td>0.7223</td>
<td>3.6277</td>
<td>8.5907</td>
<td>0.7133</td>
<td>0.6443</td>
<td>0.6981</td>
<td>0.6344</td>
</tr>
</tbody>
</table>

**Table 4.** The LC–CV vs. the LL–CV method when \( X_t \) is an \( I(0) \) variable and \( Y_t \) is an \( I(1) \) variable

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( \theta = 0 )</th>
<th>( \theta = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_2(z) )</td>
<td>Mean</td>
<td>St. dev.</td>
</tr>
<tr>
<td>100</td>
<td>0.0724</td>
<td>0.0355</td>
</tr>
<tr>
<td>200</td>
<td>0.0395</td>
<td>0.0182</td>
</tr>
<tr>
<td>400</td>
<td>0.0216</td>
<td>0.0093</td>
</tr>
<tr>
<td>( \hat{\beta}_1(z) )</td>
<td>Mean</td>
<td>St. dev.</td>
</tr>
<tr>
<td>100</td>
<td>0.1577</td>
<td>0.0955</td>
</tr>
<tr>
<td>200</td>
<td>0.1061</td>
<td>0.0563</td>
</tr>
<tr>
<td>400</td>
<td>0.0744</td>
<td>0.0349</td>
</tr>
<tr>
<td>( \hat{\tilde{\beta}}_1(z) )</td>
<td>Mean</td>
<td>St. dev.</td>
</tr>
<tr>
<td>100</td>
<td>0.146</td>
<td>0.0685</td>
</tr>
<tr>
<td>200</td>
<td>0.1016</td>
<td>0.0457</td>
</tr>
<tr>
<td>400</td>
<td>0.0717</td>
<td>0.0293</td>
</tr>
</tbody>
</table>
Proof. By (2.5), we have $\Delta_t = \hat{A}_t - \hat{\mu}_t(z_t) = n^{-2} \sum_{s \neq t} x^2_{st} \times e_{st} + O(n^{-1} \ln n)$, where $e_{st} = K_{h,t} - f(z_t)$ and $E[e_{st}] = O(1)$ for all $s \neq t$. We follow the proof technique of Hansen (1992), pp. 497–498. For any small $\alpha \in (0, 1)$, let $N = \lfloor \lambda^{-1} \rfloor \lfloor \lambda^{-1} \rfloor$ denote the integer part of $\lambda^{-1}$, $s_k = [kn/N] + 1$, $s^*_k = s_{k+1} - 1$, $N^* = N - 1$, and $s^*_{N^*} = \min(s^*_1, n)$. Also denote $M_{n,s} = n^{-1} x^2_{st}$ for any $s$ and $M_n(r) = M_n,[nr]$ for $r \in [0, 1]$. By Assumption A1, and applying the continuous mapping theorem, we have $M_n(\cdot) \Rightarrow B^2_\cdot(\cdot)$. We then have

$$n^{-2} \sum_{s \neq t} x^2_{st} e_{st} \leq n^{-1} \sum_{k=0}^{N^*} \sum_{s_k, s_k \neq t} M_{n,s_k} e_{st} + n^{-1} \sum_{k=0}^{N^*} \sum_{s_k, s_k \neq t} (M_{n,s_k} - M_{n,s_k}) e_{st} \leq \sup_{r \in [0, 1]} M_n(r) n^{-1} \sum_{k=0}^{N^*} \sum_{s_k, s_k \neq t} e_{st} + \sup_{|r-r'| \leq \lambda} |M_n(r) - M_n(r')| \sum_{s \neq t} e_{st} = O_p\left(h^2 + \sqrt{\ln(n\lambda)/n^2} h\right) + O_p\left(\sqrt{\ln \lambda^{-1}}\right)$$

where the last equality follows by choosing $\lambda = (nh)^{-1/2}$, and we also use the following results: by Assumption A6, we have

$$\sup_{r \in [0, 1]} M_n(r) = O_p(1), \quad \sup_{|r-r'| \leq \lambda} |M_n(r) - M_n(r')| = O_p\left(\sqrt{\ln \lambda^{-1}}\right).$$

In addition, in deriving the third line of Equation (A.10) we use theorem 6 of Hansen (2008):

$$n^{-1} \sum_{k=0}^{N^*} \sum_{s_k, s_k \neq t} e_{st} \leq \sup_{z \in \mathcal{M}} \left| n^{-1} \sum_{j=s_k}^{s_k + [n\lambda]} e_{jt} \right| = O_p\left(h^2 + \sqrt{\ln(n\lambda)!/(n^2\lambda)}\right).$$


Proof. Define $CV^0_1$ by replacing $\hat{A}_t$ with $\hat{\mu}_t(z_t) = c_{1st}$ in $CV_1$; that is, $CV^0_1 = n^{-2} \sum_{s \neq t} x^2_{st} / \hat{\mu}_t(z_t)^2 M_{t} = n^{-2} c_{1t}^2 \sum_{s \neq t} x^2_{st} \hat{B}_t M_{t}^{-1} - \sum_{s \neq t} x^2_{st} \hat{B}_t M_{t}^{-1} e_{st}$ and $e_{st} = (\hat{B}_t - \beta_t K_{h,t} B_t)$. Equation (A.9) implies that $CV^0_1$ is the leading term of $CV_1$.

We decompose $\hat{B}_t$ into two terms: $\hat{B}_t = n^{-2} \sum_{s \neq t} x^2_{st} E(e_{st}/z_t) + n^{-2} \sum_{s \neq t} x^2_{st} [e_{st} - E(e_{st}/z_t)] = \omega_{t1} + \omega_{t2}$, where the de-
Finally, we consider $\Delta n_3 = n^{-2}e_{1n}^{-2} \sum x_i^2 \sigma_i^2 M_{d_1}f_{d_2}^{-2}$. By Assumption A6*, (2.6) with $\sigma_1(\cdot) = |\beta(1')(\cdot)| \sqrt{\nu_2(K)(\cdot)}$, we have $\frac{1}{\sigma_1(\cdot)} \sum e_i(\cdot) - E(e_i(\cdot)) - \int_0^1 B_2^e(r) dW_2(r) = o_p(1)$ uniformly over $z \in M$. Also note that when $c_{1n} = n^{-2} \sum x_i^2 = B_{x}(2) + o_p(1)$, where $B_{x}(2) = \int_0^1 B_2 x(r)^2 \, dr$, we have

$$\Delta n_3 = (h/n)\nu_2(K)B_{C}(2)E[M_{d_1}f_{d_2}^{-1}(\beta(1'))^2] \left[ \int_0^1 B_2^e(r) dW_2(r) \right]^2 + o_p(h/n). \quad (A.19)$$

Combining (A.14), (A.18), and (A.19) completes the proof of Lemma A.2.

**Lemma A.3.** Under Assumptions A1–A6*, (A.4) holds true.

**Proof.** Lemma A.1 implies that the leading term of CV2, denoted by $CV_2^0$, is obtained from $CV_2$ by replacing $\tilde{A}$ with $\tilde{A}_1(z)$. That is, $CV_2 = n^{-2} \sum x_i^2 \tilde{C}_{i1} / \tilde{C}_{11} M_{1} = n^{-2} \sum x_i^2 \tilde{C}_{i1} \tilde{M}_1 - M_i$ with $\tilde{C}_{i1} = \tilde{C}_{i1} \tilde{M}_1$. By Assumption A6*, we have $n^{-1/2} \tilde{C} = \sigma_n \sqrt{\nu_2(K)(\cdot)} \int_0^1 B_2(r) dW_2(r) + o_p(1)$. Therefore, we obtain

$$n^{-2} h CV_2^0 = \sigma_n^2 \nu_0(K) C^{-1}_{11} \left[ \int_0^1 B_2(r) dW_2(r) \right]^2 \int \{ \nu(z) \} \, dz + o_p(1). \quad (A.20)$$

This completes the proof of Lemma A.3.

**Lemma A.4.** Under Assumptions A1–A6, (A.5) holds true.

**Proof.** By definition, $CV_3 = n^{-4} \sum x_i^2 \tilde{B}_i \tilde{A}_i^{-2} M_i \times \sum x_i \times u_{i k}$. Assumption A3 implies that $E(CV_3) = 0$ and $\{u_i\}_{i=1}^n$ are serially uncorrelated. Letting $\Delta_3 = E(CV_3)^2 = \{x_i \tilde{B}_i \tilde{A}_i^{-2} M_i \times \sum x_i \times u_{i k},\}$ we have $n^{-8} \sum x_i^2 \tilde{B}_i \tilde{A}_i^{-2} M_i \times \sum x_i \times u_{i k} = \Delta_3(n) = n^{-8} \sigma_n^2 \sum x_i^2 \tilde{B}_i \tilde{A}_i^{-2} M_i \times \sum x_i \times u_{i k} = \Delta_3 + \Delta_{3,2n}$.

Lemma A.1 and (A.17) imply that $\sup_{z \in M} |\tilde{A}_1| = O_p(1)$ and that $\sup_{z \in M} |\tilde{B}_i| = O_p(h \delta h)$. Applying the same technique used in the proof of Lemma A.1, we can show that $\sup_{z \in M} n^{-2} \sum x_i \times u_{i k} = o_p(1)$. Therefore, we have $\Delta_{3,1n} = O_p(h^2 \delta h) n^{-8} \sup_{z \in M} \sum x_i \times u_{i k} = O_p(n^{-3} h \delta h)$ and $\Delta_{3,2n} = O_p(h^2 \delta h) n^{-8} \sum x_i \times u_{i k} = O_p(n^{-2} h^2 \delta h)$ by Lemma A.1. Because $n^{-2} \sim o(h)$, $\Delta_{3,2n}$ asymptotically dominates $\Delta_{3,1n}$. Thus we have shown that $\text{var}(CV_3) = (\{x_i \tilde{A}_i^{-2} M_i \times \sum x_i \times u_{i k},\}) = O_p(n^{-2} h^2 \delta h)$. This implies that $CV_3 = O_p(h/n) \delta \delta h = o_p(h/n) \delta \delta h$ by Markov's inequality.

**ACKNOWLEDGMENTS**

The authors thank two referees, an associate editor, former coeditor Serena Ng, and coeditor Jonathan Wright for insightful comments that greatly improved the article. This research was partially supported by the SSHRC Fund of Canada and the National Science Foundation of China (project 70773005).

[Received June 2009. Revised December 2011.]

**REFERENCES**