Econometric Theory, **20**, 2004, 1094–1139. Printed in the United States of America.
DOI: 10.1017/S026646660420603X

THE LIVE METHOD FOR GENERALIZED ADDITIVE VOLATILITY MODELS

WOOCHEOL KIM
Korea Institute of Public Finance
and
Humboldt University of Berlin

OLIVER LINTON
The London School of Economics

We investigate a new separable nonparametric model for time series, which includes many autoregressive conditional heteroskedastic (ARCH) models and autoregressive (AR) models already discussed in the literature. We also propose a new estimation procedure called LIVE, or local instrumental variable estimation, that is based on a localization of the classical instrumental variable method. Our method has considerable computational advantages over the competing marginal integration or projection method. We also consider a more efficient two-step likelihood-based procedure and show that this yields both asymptotic and finite-sample performance gains.

1. INTRODUCTION

Volatility models are of considerable interest in empirical finance. There are many types of parametric volatility models, following the seminal work of Engle (1982). These models are typically nonlinear, which poses difficulties both in computation and in deriving useful tools for statistical inference. Parametric models are prone to misspecification, especially when there is no theoretical reason to prefer one specification over another. Nonparametric models can provide greater flexibility. However, the greater generality of these models comes at a cost—including a large number of lags requires estimation of a high-dimensional smooth, which is known to behave very badly (Silverman, 1986). The "curse of dimensionality" puts severe limits on the dynamic flexibility of

This paper is based on Chapter 2 of the first author's Ph.D. dissertation from Yale University. We thank Wolfgang Härdle, Joel Horowitz, Peter Phillips, and Dag Tjøstheim for helpful discussions. We are also grateful to Donald Andrews and two anonymous referees for valuable comments. The second author thanks the National Science Foundation and the ESRC for financial support. Address correspondence to Woocheol Kim, Korea Institute of Public Finance, 79-6 Garak-Dong, Songpa-Gu, Seoul, Republic of Korea, 138-774; e-mail: wkim@kipf.re.kr. Oliver Linton, Department of Economics, The London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom; e-mail: lintono@lse.ac.uk.

  0266-4666/04 $12.00

nonparametric models. Separable models offer an intermediate position between the complete generality of nonparametric models and the restrictiveness of parametric models. These models have been investigated in cross-sectional settings and also in time series settings.

In this paper, we investigate a generalized additive nonlinear autoregressive conditional heteroskedastic model (GANARCH):

$$y_t = m(y_{t-1}, y_{t-2}, \ldots, y_{t-d}) + u_t,$$
$$u_t = v^{1/2}(y_{t-1}, y_{t-2}, \ldots, y_{t-d})\varepsilon_t, \tag{1.1}$$

$$m(y_{t-1}, y_{t-2}, \ldots, y_{t-d}) = F_m\left(c_m + \sum_{\alpha=1}^{d} m_\alpha(y_{t-\alpha})\right), \tag{1.2}$$

$$v(y_{t-1}, y_{t-2}, \ldots, y_{t-d}) = F_v\left(c_v + \sum_{\alpha=1}^{d} v_\alpha(y_{t-\alpha})\right), \tag{1.3}$$

where $m_\alpha(\cdot)$ and $v_\alpha(\cdot)$ are smooth but unknown functions and $F_m(\cdot)$ and $F_v(\cdot)$ are known monotone transformations (whose inverses are $G_m(\cdot)$ and $G_v(\cdot)$, respectively).[1] The error process, $\{\varepsilon_t\}$, is assumed to be a martingale difference with unit scale, that is, $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ and $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = 1$, where $\mathcal{F}_t$ is the σ-algebra of events generated by $\{y_k\}_{k=-\infty}^{t}$. Under some weak assumptions, the time series of nonlinear autoregressive models can be shown to be stationary and strongly mixing with mixing coefficients decaying exponentially fast. Auestadt and Tjøstheim (1990) use α-mixing or geometric ergodicity to identify their nonlinear time series model. Similar results are obtained for the additive nonlinear autoregressive conditional heteroskedastic (ARCH) process by Masry and Tjøstheim (1997); see also Cai and Masry (2000) and Carrasco and Chen (2002). We follow the same argument as Masry and Tjøstheim (1997) and will assume all the necessary conditions for stationarity and mixing property of the process $\{y_t\}_{t=1}^{n}$ in (1.1). The standard identification for the components of the mean and variance is made by

$$E[m_\alpha(y_{t-\alpha})] = 0 \quad \text{and} \quad E[v_\alpha(y_{t-\alpha})] = 0 \tag{1.4}$$

for all $\alpha = 1, \ldots, d$. The notable aspect of the model is additivity via known links for conditional mean and volatility functions. As will be shown later, (1.1)–(1.3) include a wide variety of time series models in the literature. See Horowitz (2001) for a discussion of generalized additive models in a cross-section context.

In a much simpler univariate setup, Robinson (1983), Auestadt and Tjøstheim (1990), and Härdle and Vieu (1992) study the kernel estimation of the conditional mean function $m(\cdot)$ in (1.1). The so-called CHARN (conditionally heteroskedastic autoregressive nonlinear) model is the same as (1.1) except that $m(\cdot)$ and $v(\cdot)$ are univariate functions of y_{t-1}. Masry and Tjøstheim (1995) and Härdle and Tsybakov (1997) apply the Nadaraya–Watson and local linear

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

smoothing methods, respectively, to jointly estimate $v(\cdot)$ together with $m(\cdot)$. Alternatively, Fan and Yao (1996) and Ziegelmann (2002) propose local linear least square estimation for the volatility function, with the extension given by Avramidis (2002) based on local linear maximum likelihood estimation. Also, in a nonlinear vector autoregressive (VAR) context, Härdle, Tsybakov, and Yang (1998) deal with the estimation of conditional mean in a multilagged extension similar to (1.1). Unfortunately, however, introducing more lags in nonparametric time series models has unpleasant consequences, more so than in the parametric approach. As is well known, smoothing methods in high dimensions suffer from a slower convergence rate—the "curse of dimensionality." Under twice differentiability of $m(\cdot)$, the optimal rate is $n^{-2/(4+d)}$, which gets rapidly worse with dimension. In high dimensions it is also difficult to describe graphically the function m.

The additive structure has been proposed as a useful way to circumvent these problems in multivariate smoothing. By assuming the target function to be a sum of functions of covariates, say, $m(y_{t-1}, y_{t-2}, \ldots, y_{t-d}) = c_m + \sum_{\alpha=1}^{d} m_\alpha(y_{t-\alpha})$, we can effectively reduce the dimensionality of a regression problem and improve the implementability of multivariate smoothing up to that of the one-dimensional case. Stone (1985, 1986) shows that it is possible to estimate $m_\alpha(\cdot)$ and $m(\cdot)$ with the one-dimensional optimal rate of convergence—for example, $n^{2/5}$ for twice differentiable functions—regardless of d. The estimates are easily illustrated and interpreted. For these reasons, since the 1980s, additive models have been fundamental to nonparametric regression among both econometricians and statisticians. Regarding the estimation method for achieving the one-dimensional optimal rate, the literature suggests two different approaches: backfitting and marginal integration. The former, originally suggested by Breiman and Friedman (1985), Buja, Hastie, and Tibshirani (1989), and Hastie and Tibshirani (1987, 1990), is to execute iterative calculations of one-dimensional smoothing until some convergence criterion is satisfied. Though appealing to our intuition, the statistical properties of backfitting algorithm were not clearly understood until the very recent works by Opsomer and Ruppert (1997) and Mammen, Linton, and Nielsen (1999). They develop specific (linear) backfitting procedures and establish the geometric convergence of their algorithms and the pointwise asymptotic distributions under some conditions. However, one disadvantage of these procedures is the time-consuming iterations required for implementation. Also, the proofs for the linear case cannot be easily generalized to nonlinear cases such as generalized additive models.

A more recent approach, called marginal integration (MI), is theoretically more manipulable—its statistical properties are easy to derive, because it simply uses averaging of multivariate kernel estimates. Developed independently by Newey (1994), Tjøstheim and Auestadt (1994), and Linton and Nielsen (1995), its simplicity inspired subsequent applications such as Linton, Wang, Chen, and Härdle (1995) for transformation models and Linton, Nielsen, and van de Geer (2003) for hazard models with censoring. In the time series mod-

els that are special cases of (1.1) and (1.2) with F_m being the identity, Chen and Tsay (1993a, 1993b) and Masry and Tjøstheim (1997) apply backfitting and MI, respectively, to estimate the conditional mean function. Mammen et al. (1999) provide useful results for the same type of models by improving the previous backfitting method with some modification and successfully deriving the asymptotic properties under weak conditions. The separability assumption is also used in volatility estimation by Yang, Härdle, and Nielsen (1999), where the nonlinear ARCH model is of additive mean and multiplicative volatility in the form of

$$y_t = c_m + \sum_{\alpha=1}^{d} m_\alpha(y_{t-\alpha}) + \left(c_v \prod_{\alpha=1}^{d} v_\alpha(y_{t-\alpha}) \right)^{1/2} \varepsilon_t. \tag{1.5}$$

To estimate (1.5), they rely on marginal integration with local linear fits as a pilot estimate and derive asymptotic properties.

This paper features two contributions to the additive literature. The first concerns theoretical development of a new estimation tool called the local instrumental variable estimator for the components of additive models (LIVE for CAM), which was outlined for simple additive cross-sectional regression in Kim, Linton, and Hengartner (1999). The novelty of the procedure lies in the simple definition of the estimator based on univariate smoothing combined with new kernel weights. That is, adjusting kernel weights via conditional density of the covariate enables a univariate kernel smoother to estimate consistently the corresponding additive component function. In many respects, the new estimator preserves the good properties of univariate smoothers. The instrumental variable method is analytically tractable for asymptotic theory: it is shown to attain the optimal one-dimensional rate. Furthermore, it is computationally more efficient than the two existing methods (backfitting and MI) in the sense that it reduces the computations by a factor of n smoothings. The other contribution relates to the general coverage of the model we work with. The model in (1.1)–(1.3) extends ARCH models to a generalized additive framework where both the mean and variance functions are additive after some known transformation (see Hastie and Tibshirani, 1990). All the time series models in our previous discussion are regarded as a subclass of the data generating process for $\{y_t\}$ in (1.1)–(1.3). For example, setting G_m to be an identity and G_v a logarithmic function reduces our model to (1.5). Similar efforts to apply transformation have been made in parametric ARCH models. Nelson (1991) considers a model for the log of the conditional variance—the exponential (G)ARCH class—to embody the multiplicative effects of volatility. It has also been argued to use the Box–Cox transformation for volatility, which is intermediate between linear and logarithm and which allows nonseparable news impact curves. Because it is hard to tell a priori which structure of volatility is more realistic and it should be determined by real data, our generalized additive model provides useful flexible specifications for empirical work. Addi-

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

tionally, from the perspective of potential misspecification problems, the transformation used here alleviates the restriction imposed by the additivity assumption, which increases the approximating power of our model. Note that when the lagged variables in (1.1)–(1.3) are replaced by different covariates and the observations are independent and identically distributed (i.i.d.), the model becomes the cross-sectional additive model studied by Linton and Härdle (1996). Finally, we also consider more efficient estimation along the lines of Linton (1996, 2000).

The rest of the paper is organized as follows. Section 2 describes the main estimation idea in a simple setting. In Section 3, we define the estimator for the full model. In Section 4 we give our main results, including the asymptotic normality of our estimators. Section 5 discusses more efficient estimation. Section 6 reports a small Monte Carlo study. The proofs are contained in the Appendix.

2. NONPARAMETRIC INSTRUMENTAL VARIABLES: THE MAIN IDEA

This section explains the basic idea behind the instrumental variable method and defines the estimation procedure. For ease of exposition, this will be carried out using an example of simple additive models with i.i.d. data. We then extend the definition to the generalized additive ARCH case in (1.1)–(1.3).

Consider a bivariate additive regression model for i.i.d. data (y, X_1, X_2),

$$y = m_1(X_1) + m_2(X_2) + \varepsilon,$$

where $E(\varepsilon|X) = 0$ with $X = (X_1, X_2)$ and the components satisfy the identification conditions $E[m_\alpha(X_\alpha)] = 0$, for $\alpha = 1,2$ (the constant term is assumed to be zero, for simplicity). Letting $\eta = m_2(X_2) + \varepsilon$, we rewrite the model as

$$y = m_1(X_1) + \eta, \tag{2.6}$$

which is a classical example of "omitted variable" regression. That is, although (2.6) appears to take the form of a univariate nonparametric regression model, smoothing y on X_1 will incur a bias due to the omitted variable η, because η contains X_2, which in general depends on X_1. One solution to this is suggested by the classical econometric notion of instrumental variable. That is, we look for an instrument W such that

$$E(W|X_1) \neq 0; \qquad E(W\eta|X_1) = 0 \tag{2.7}$$

with probability one.[2] If such a random variable exists, we can write

$$m_1(x_1) = \frac{E(Wy|X_1 = x_1)}{E(W|X_1 = x_1)}. \tag{2.8}$$

This suggests that we estimate the function $m_1(\cdot)$ by nonparametric smoothing of Wy on X_1 and W on X_1. In parametric models the choice of instrument is

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

usually not obvious and requires some caution. However, our additive model has a natural class of instruments—$p_2(X_2)/p(X)$ times any measurable function of X_1 will do, where $p(\cdot)$, $p_1(\cdot)$, and $p_2(\cdot)$ are the density functions of the covariates X, X_1, and X_2, respectively. It follows that

$$\frac{E(Wy|X_1)}{E(W|X_1)} = \frac{\int W(X)m(X)\,\dfrac{p(X)}{p_1(X_1)}\,dX_2}{\int W(X)\,\dfrac{p(X)}{p_1(X_1)}\,dX_2} = \frac{\int W(X)m(X)p(X)\,dX_2}{\int W(X)p(X)\,dX_2}$$

$$= \frac{\int m(X)p_2(X_2)\,dX_2}{\int p_2(X_2)\,dX_2} = \int m(X)p_2(X_2)\,dX_2$$

as required. This formula shows what the instrumental variable estimator is estimating when m is not additive—an average of the regression function over the X_2 direction, exactly the same as the target of the marginal integration estimator. For simplicity we will take

$$W(X) = \frac{p_2(X_2)}{p(X)} \tag{2.9}$$

throughout.[3]

Up to now, it was implicitly assumed that the distributions of the covariates are known a priori. In practice, this is rarely true, and we have to rely on estimates of these quantities. Let $\hat{p}(\cdot), \hat{p}_1(\cdot)$, and $\hat{p}_2(\cdot)$ be kernel estimates of the densities $p(\cdot), p_1(\cdot)$, and $p_2(\cdot)$, respectively. Then the feasible procedure is defined with a replacement of the instrumental variable W by $\hat{W} = \hat{p}_2(X_2)/\hat{p}(X)$ and taking sample averages instead of population expectations. Section 3 provides a rigorous statistical treatment for feasible instrumental variable estimators based on local linear estimation. See Kim et al. (1999) for a slightly different approach.

Next, we come to the main advantage that the local instrumental variable method has. This is in terms of the computational cost. The marginal integration method actually needs n^2 regression smoothings evaluated at the pairs (X_{1i}, X_{2j}), for $i, j = 1, \ldots, n$, whereas the backfitting method requires nr operations—where r is the number of iterations to achieve convergence. The instrumental variable procedure, in contrast, takes at most $2n$ operations of kernel smoothings in a preliminary step for estimating the instrumental variable and another n operations for the regressions. Thus, it can be easily combined with the bootstrap method whose computational costs often become prohibitive in the case of marginal integration (see Kim et al., 1999).

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Finally, we show how the instrumental variable approach can be applied to generalized additive models. Let $F(\cdot)$ be the inverse of a known link function $G(\cdot)$ and let $m(X) = E(y|X)$. The model is defined as

$$y = F(m_1(X_1) + m_2(X_2)) + \varepsilon, \tag{2.10}$$

or equivalently $G(m(X)) = m_1(X_1) + m_2(X_2)$. We maintain the same identification condition, $E[m_\alpha(X_\alpha)] = 0$. Unlike in the simple additive model, there is no direct way to relate Wy to $m_1(X_1)$ here, so (2.8) cannot be implemented. However, under additivity

$$m_1(X_1) = \frac{E[WG(m(X))|X_1]}{E[W|X_1]} \tag{2.11}$$

for the W defined in (2.9). Because $m(\cdot)$ is unknown, we need consistent estimates of $m(X)$ in a preliminary step, and then the calculation in (2.11) is feasible. In the next section we show how these ideas are translated into estimators for the general time series setting.

3. INSTRUMENTAL VARIABLE PROCEDURE FOR GANARCH

We start with some simplifying notations that will be used repeatedly in the discussion that follows. Let x_t be the vector of d lagged variables until $t-1$, that is, $x_t = (y_{t-1},\ldots,y_{t-d})$, or concisely, $x_t = (y_{t-\alpha}, \underline{y}_{t-\alpha})$, where $\underline{y}_{t-\alpha} = (y_{t-1},\ldots,y_{t-\alpha-1},y_{t-\alpha+1},\ldots,y_{t-d})$. Defining $m_{\underline{\alpha}}(\underline{y}_{t-\alpha}) = \sum_{\beta=1,\neq\alpha}^{d} m_\beta(y_{t-\beta})$ and $v_{\underline{\alpha}}(\underline{y}_{t-\alpha}) = \sum_{\beta=1,\neq\alpha}^{d} v_\beta(y_{t-\beta})$, we can reformulate (1.1)–(1.3) with a focus on the αth components of the mean and variance as

$$\begin{aligned} y_t &= m(x_t) + v^{1/2}(x_t)\varepsilon_t, \\ m(x_t) &= F_m(c_m + m_\alpha(y_{t-\alpha}) + m_{\underline{\alpha}}(\underline{y}_{t-\alpha})), \\ v(x_t) &= F_v(c_v + v_\alpha(y_{t-\alpha}) + v_{\underline{\alpha}}(\underline{y}_{t-\alpha})). \end{aligned}$$

To save space we will use the following abbreviations for the functions to be estimated:

$$\begin{aligned} H_\alpha(y_{t-\alpha}) &\equiv [m_\alpha(y_{t-\alpha}), v_\alpha(y_{t-\alpha})]^\top, \qquad H_{\underline{\alpha}}(\underline{y}_{t-\alpha}) \equiv [m_{\underline{\alpha}}(\underline{y}_{t-\alpha}), v_{\underline{\alpha}}(\underline{y}_{t-\alpha})]^\top, \\ c &\equiv [c_m, c_v]^\top, \qquad r_t \equiv H(x_t) = [G_m(m(x_t)), G_v(v(x_t))]^\top, \\ \varphi_\alpha(y_\alpha) &= c + H_\alpha(y_\alpha). \end{aligned}$$

Note that the components $[m_\alpha(\cdot), v_\alpha(\cdot)]^\top$ are identified, up to constant c, by $\varphi_\alpha(\cdot)$, which will be our major interest in estimation. Subsequently, we examine in some detail each relevant step for computing the feasible nonparametric instrumental variable estimator of $\varphi_\alpha(\cdot)$. The set of observations is given by $\mathcal{Y} = \{y_t\}_{t=1}^{n'}$, where $n' = n + d$.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

3.1. Step I. Preliminary Estimation of $r_t = H(x_t)$

Because r_t is unknown, we start with computing the pilot estimates of the regression surface by a local linear smoother. Let $\tilde{m}(x)$ be the first component of $(\tilde{a}, \tilde{b}^\top)^\top$ that solves

$$\min_{a,b} \sum_{t=d+1}^{n'} K_h(x_t - x)\{y_t - a - b^\top(x_t - x)\}^2, \tag{3.12}$$

where $K_h(x) = \Pi_{i=1}^d K(x_i/h)/h^d$, K is a one-dimensional kernel function, and $h = h(n)$ is a bandwidth sequence. In a similar way, we get the estimate of the volatility surface, $\tilde{v}(\cdot)$, from (3.12) by replacing y_t with the squared residuals, $\tilde{\varepsilon}_t^2 = (y_t - \tilde{m}(x_t))^2$. Then, transforming $\tilde{m}$ and $\tilde{v}$ by the known links will lead to consistent estimates of $\tilde{r}_t$,

$$\tilde{r}_t = \tilde{H}(x_t) = [G_m(\tilde{m}(x_t)), G_v(\tilde{v}(x_t))]^\top.$$

3.2. Step II: Instrumental Variable Estimation of Additive Components

This step involves the estimation of $\varphi_\alpha(\cdot)$, which is equivalent to $[m_\alpha(\cdot), v_\alpha(\cdot)]^\top$, up to the constant c. Let $p(\cdot)$ and $p_{\underline{\alpha}}(\cdot)$ denote the density functions of the random variables $(y_{t-\alpha}, \underline{y}_{t-\alpha})$ and $\underline{y}_{t-\alpha}$, respectively. Define the feasible instrument as

$$\hat{W}_t = \frac{\hat{p}_{\underline{\alpha}}(\underline{y}_{t-\alpha})}{\hat{p}(y_{t-\alpha}, \underline{y}_{t-\alpha})},$$

where $\hat{p}_{\underline{\alpha}}(\cdot)$ and $\hat{p}(\cdot)$ are computed using the kernel function $L(\cdot)$, for example, $\hat{p}(x) = \sum_{t=1}^n \prod_{i=1}^d L_g(x_{it} - x_i)/n$ with $L_g(\cdot) \equiv L(\cdot/g)/g$ and $g = g(n)$ is a bandwidth sequence. The instrumental variable local linear estimates $\hat{\varphi}_\alpha(y_\alpha)$ are given as $(a_1, a_2)^\top$ through minimizing the localized squared errors elementwise:

$$\min_{a_j, b_j} \sum_{t=d+1}^{n'} K_h(y_{t-\alpha} - y_\alpha)\hat{W}_t\{\tilde{r}_{jt} - a_j - b_j(y_{t-\alpha} - y_\alpha)\}^2, \tag{3.13}$$

where $\tilde{r}_{jt}$ is the jth element of $\tilde{r}_t$.[4] The closed form of the solution is

$$\hat{\varphi}_\alpha(y_\alpha)^\top = e_1^\top(\mathbf{Y}_-^\top \mathbf{K} \mathbf{Y}_-)^{-1} \mathbf{Y}_-^\top \mathbf{K} \tilde{\mathbf{R}}, \tag{3.14}$$

where $e_1 = (1,0)^\top$, $\mathbf{Y}_- = [\iota, Y_-]$, $\mathbf{K} = \mathrm{diag}[K_h(y_{d+1-\alpha} - y_\alpha)\hat{W}_{d+1}, \ldots, K_h(y_{n'-\alpha} - y_\alpha)\hat{W}_{n'}]$, and $\tilde{\mathbf{R}} = (\tilde{r}_{d+1}, \ldots, \tilde{r}_{n'})^\top$, with $\iota = (1, \ldots, 1)^\top$ and $Y_- = (y_{d+1-\alpha} - y_\alpha, \ldots, y_{n'-\alpha} - y_\alpha)^\top$.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

4. MAIN RESULTS

Let $\mathcal{F}_b^a$ be the σ-algebra of events generated by $\{y_t\}_a^b$ and $\alpha(k)$ the strong mixing coefficient of $\{y_t\}$ that is defined by

$$\alpha(k) \equiv \sup_{A \in \mathcal{F}_{-\infty}^0,\, B \in \mathcal{F}_k^\infty} |P(A \cap B) - P(A)P(B)|.$$

Throughout the paper, we make the following assumptions.

Assumption A.

A1. $\{y_t\}_{t=1}^\infty$ is a stationary and strongly mixing process generated by (1.1)–(1.3), with a mixing coefficient such that $\sum_{k=0}^\infty k^a \{\alpha(k)\}^{1-2/\nu} < \infty$, for some $\nu > 2$ and $0 < a < (1 - 2/\nu)$.

As pointed out by Masry and Tjøstheim (1997), the condition on the mixing coefficient in A1 is milder than assumed on the standard mixing process where the coefficient decreases at a geometric rate, that is, $\alpha(k) = \rho^{-\beta k}$ (for some $\beta > 0$). Some technical conditions for regularity are stated here. For simplicity, we assume that the process $\{y_t\}_{t=1}^\infty$ has a compact support.

A2. The additive component functions, $m_\alpha(\cdot)$ and $v_\alpha(\cdot)$, for $\alpha = 1, \ldots, d$, are continuous and twice differentiable on the compact support.

A3. The link functions, G_m and G_v, have bounded continuous second-order derivatives over any compact interval.

A4. The joint and marginal density functions, $p(\cdot)$, $p_\alpha(\cdot)$, and $p_\alpha(\cdot)$, for $\alpha = 1, \ldots, d$, are continuous, twice differentiable with bounded (partial) derivatives, and bounded away from zero on the compact support.

A5. The kernel functions, $K(\cdot)$ and $L(\cdot)$, are a real bounded nonnegative symmetric (around zero) function on a compact support satisfying $\int K(u)\,du = \int L(u)\,du = 1$, $\int uK(u)\,du = \int uL(u)\,du = 0$. Also, assume that the kernel functions are Lipschitz-continuous, $|K(u) - K(v)| \leq C|u - v|$.

A6. (i) $g \to 0$, $ng^d \to \infty$, $(\log n)^2 \sqrt{h}/\sqrt{n} g^d \to 0$. (ii) $h \to 0$, $(\log n)^2/\sqrt{nh^{2d-1}} \to 0$. (iii) The bandwidth satisfies $\sqrt{n/h}\,\alpha(t(n)) \to 0$, where $\{t(n)\}$ is a sequence of positive integers, $t(n) \to \infty$, such that $t(n) = o(\sqrt{nh})$.

Conditions A2–A5 are standard in kernel estimation. The continuity assumption in A2 and A4, together with the compact support, implies that the functions are bounded. The bandwidth conditions in A6(i) and A6(ii) are necessary for showing negligibility of the stochastic error terms arising from the preliminary estimation of m, v, and $p_\alpha(\cdot)$. Under twice-differentiability of these functions as in A2–A4, the given side conditions are satisfied when $d \leq 4$. Our asymptotic results that follow can be extended into a more general case of

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

$d > 4$, although we do not prove it in the paper. One way of extension to higher dimensions is to strengthen the differentiability conditions in A2–A4 and use higher order polynomials (see Kim et al., 1999). The additional bandwidth condition in A6(iii) is necessary to control the effects from the dependence of the mixing processes in showing the asymptotic normality of instrumental variable estimates. The proof of consistency, however, does not require this condition. Define $D^2 f(x_1,\ldots,x_d) = \sum_{l=1}^{d} \partial^2 f(x_l)/\partial^2 x$ and $[\nabla G_m(t), \nabla G_v(t)] = [dG_m(t)/dt, dG_v(t)/dt]$. Let $(K * K)_i(u) = \int K(w)K(w+u) \times w^i\, dw$, a convolution of kernel functions, $\mu^2_{K*K} = \int (K*K)_0(u)u^2\, du$, and $\|K\|_2^2$ denote $\int K^2(u)\, du$. The asymptotic properties of the feasible instrumental variable estimates in (3.14) are summarized in the following theorem, whose proof is in the Appendix. Let $\kappa_3(y_\alpha, z_{\underline{\alpha}}) = E[\varepsilon_t^3 | x_t = (y_\alpha, z_{\underline{\alpha}})]$ and $\kappa_4(y_\alpha, z_{\underline{\alpha}}) = E[(\varepsilon_t^2 - 1)^2 | x_t = (y_\alpha, z_{\underline{\alpha}})]$. $A \odot B$ denotes the matrix Hadamard product.

THEOREM 1. *Assume that conditions A1–A6 hold. Then,*

$$\sqrt{nh}[\hat{\varphi}_\alpha(y_\alpha) - \varphi_\alpha(y_\alpha) - B_\alpha] \xrightarrow{d} N[0, \Sigma^*_\alpha(y_\alpha)], \quad \textit{where}$$

$$\begin{aligned} B_\alpha(y_\alpha) &= \frac{h^2}{2}\mu_K^2 D^2\varphi_\alpha(y_\alpha) \\ &\quad + \frac{h^2}{2}\int [\mu^2_{K*K} D^2\varphi_\alpha(y_\alpha) + \mu_K^2 D^2\varphi_{\underline{\alpha}}(z_{\underline{\alpha}})] \\ &\quad \odot [\nabla G_m(m(y_\alpha, z_{\underline{\alpha}})), \nabla G_v(v(y_\alpha, z_{\underline{\alpha}}))]^\top p_{\underline{\alpha}}(z_{\underline{\alpha}})\, dz_{\underline{\alpha}} \\ &\quad + \frac{g^2}{2}\mu_K^2 \int \left[D^2 p_{\underline{\alpha}}(z_{\underline{\alpha}}) - \frac{p_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} D^2 p(y_\alpha, z_{\underline{\alpha}}) \right] H_{\underline{\alpha}}(z_{\underline{\alpha}})\, dz_{\underline{\alpha}}, \end{aligned}$$

$$\begin{aligned} \Sigma^*_\alpha(y_\alpha) &= \|K\|_2^2 \int \frac{p^2_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} \begin{bmatrix} m^2_{\underline{\alpha}}(z_{\underline{\alpha}}) & m_{\underline{\alpha}}(z_{\underline{\alpha}})v_{\underline{\alpha}}(z_{\underline{\alpha}}) \\ m_{\underline{\alpha}}(z_{\underline{\alpha}})v_{\underline{\alpha}}(z_{\underline{\alpha}}) & v^2_{\underline{\alpha}}(z_{\underline{\alpha}}) \end{bmatrix} dz_{\underline{\alpha}} \\ &\quad + \|(K*K)_0\|_2^2 \int \frac{p^2_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} \\ &\quad \times \begin{bmatrix} \nabla G_m(m)^2 v & (\nabla G_m \nabla G_v)(\kappa_3 v^{3/2}) \\ (\nabla G_m \nabla G_v)(\kappa_3 v^{3/2}) & \nabla G_v(v)^2 \kappa_4 v^2 \end{bmatrix} (y_\alpha, z_{\underline{\alpha}})\, dz_{\underline{\alpha}}. \end{aligned}$$

Remarks.

1. To estimate $[m_\alpha(y_\alpha), v_\alpha(y_\alpha)]^\top$ we can use the following recentered estimates: $\hat{\varphi}_\alpha(y_\alpha) - \hat{c}$, where $\hat{c} = [\hat{c}_m, \hat{c}_v] = (1/n)[\sum_t y_t, \sum_t \tilde{\varepsilon}_t^2]^\top$ and

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

$\tilde{\varepsilon}_t = y_t - \tilde{m}(x_t)$. Because $\hat{c} = c + O_p(1/\sqrt{n})$, the bias and variance of $[\hat{m}_\alpha(y_\alpha), \hat{v}_\alpha(y_\alpha)]^\top$ are the same as those of $\hat{\varphi}_\alpha(y_\alpha)$. For $y = (y_1, \ldots, y_d)$, the estimates for the conditional mean and volatility are defined by

$$[\hat{m}(y), \hat{v}(y)] \equiv \left[F_m\left[-(d-1)\hat{c}_m + \sum_{\alpha=1}^{d} \hat{\varphi}_{\alpha 1}(y_\alpha) \right], \right.$$
$$\left. F_v\left[-(d-1)\hat{c}_v + \sum_{\alpha=1}^{d} \hat{\varphi}_{\alpha 2}(y_\alpha) \right] \right].$$

Let $\nabla F(y) \equiv [\nabla F_m(m(y)), \nabla F_v(v(y))]^\top$. Then, by Theorem 1 and the delta method, their asymptotic distribution satisfies

$$\sqrt{nh}[\hat{m}(y) - m(y) - b_m(y), \hat{v}(y) - v(y) - b_v(y)]^\top \xrightarrow{d} N[0, \Sigma^*(y)],$$

where $[b_m(y), b_v(y)]^\top = \nabla F(y) \odot \sum_{\alpha=1}^{d} B_\alpha(y_\alpha)$ and $\Sigma^*(y) = [\nabla F(y) \times \nabla F(y)^\top] \odot [\Sigma_1^*(y_1) + \ldots + \Sigma_d^*(y_d)]$. It is easy to see that $\hat{\varphi}_\alpha(y_\alpha)$ and $\hat{\varphi}_\beta(y_\beta)$ are asymptotically uncorrelated for any α and β and that the asymptotic variance of their sum is also the sum of the variances of $\hat{\varphi}_\alpha(y_\alpha)$ and $\hat{\varphi}_\beta(y_\beta)$.

2. The first term of the bias is of standard form, depending only on the second derivatives as in other local linear smoothing. The last term reflects the biases from using estimates for density functions to construct the feasible instrumental variable, $\hat{p}_{\underline{\alpha}}(\underline{y}_{t-\alpha})/\hat{p}(x_t)$. When the instrument consisting of known density functions, $p_{\underline{\alpha}}(\underline{y}_{t-\alpha})/p(x_t)$, is used in (3.13), the asymptotic properties of instrumental variable estimates are the same as those from Theorem 1 except that the new asymptotic bias now includes only the first two terms of $B_\alpha(y_\alpha)$.
3. The convolution kernel $(K * K)(\cdot)$ is the legacy of double smoothing in the instrumental variable estimation of "generalized" additive models because we smooth $[G_m(\tilde{m}(\cdot)), G_v(\tilde{v}(\cdot))]$ with $\tilde{m}(\cdot)$ and $\tilde{v}(\cdot)$ given by (multivariate) local linear fits. When $G_m(\cdot)$ is the identity, we can directly smooth y instead of $G_m(\tilde{m}(x_t))$ to estimate the components of the conditional mean function. Then, as the following theorem shows, the second term of the bias of B_α does not arise, and the convolution kernel in the variance is replaced by a usual kernel function.

Suppose that $F_m(t) = F_v(t) = t$ in (1.2) and (1.3). In this case, the instrumental variable estimates of $\varphi_\alpha(y_\alpha)$ can be defined in a simpler way. For $\varphi_\alpha(y_\alpha) = [M_\alpha(y_\alpha), V_\alpha(y_\alpha)] = [c_m + m_\alpha(y_\alpha), c_v + v_\alpha(y_\alpha)]$, we define $[\hat{M}_\alpha(y_\alpha), \hat{V}_\alpha(y_\alpha)]$ by the solution to the adjusted-kernel least squares in (3.13) with the modification that the (2×1) vector $\tilde{z}_t$ is replaced by $[y_t, \tilde{\varepsilon}_t^2]^\top$, where $\tilde{\varepsilon}_t$ is given in step I in Section 3.1. Theorem 2 shows the asymptotic normality of these estimates. The proof is almost the same as that of Theorem 1 and thus is omitted.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

THEOREM 2. *Under the same conditions as Theorem 1,*

(i) $\sqrt{nh}[\hat{M}_\alpha(y_\alpha) - M_\alpha(y_\alpha) - b_\alpha^m] \xrightarrow{d} N[0, \sigma_\alpha^m(y_\alpha)]$, *where*

$$b_\alpha^m(y_\alpha) = \frac{h^2}{2}\mu_K^2 D^2 m_\alpha(y_\alpha) + \frac{g^2}{2}\mu_K^2 \times \int \left[D^2 p_{\underline{\alpha}}(z_{\underline{\alpha}}) - \frac{p_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} D^2 p(y_\alpha, z_{\underline{\alpha}})\right] m_{\underline{\alpha}}(z_{\underline{\alpha}})\, dz_{\underline{\alpha}},$$

$$\sigma_\alpha^m(y_\alpha) = \|K\|_2^2 \int \frac{p_{\underline{\alpha}}^2(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})}[m_{\underline{\alpha}}^2(z_{\underline{\alpha}}) + v(y_\alpha, z_{\underline{\alpha}})]\, dz_{\underline{\alpha}}$$

and

(ii) $\sqrt{nh}[\hat{V}_\alpha(y_\alpha) - V_\alpha(y_\alpha) - b_\alpha^v] \xrightarrow{d} N[0, \sigma_\alpha^v(y_\alpha)]$, *where*

$$b_\alpha^v(y_\alpha) = \frac{h^2}{2}\mu_K^2 D^2 v_\alpha(y_\alpha) + \frac{g^2}{2}\mu_K^2 \times \int \left[D^2 p_{\underline{\alpha}}(z_{\underline{\alpha}}) - \frac{p_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} D^2 p(y_\alpha, z_{\underline{\alpha}})\right] v_{\underline{\alpha}}(z_{\underline{\alpha}})\, dz_{\underline{\alpha}},$$

$$\Sigma_\alpha^v(y_\alpha) = \|K\|_2^2 \int \frac{p_{\underline{\alpha}}^2(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})}[v_{\underline{\alpha}}^2(z_{\underline{\alpha}}) + \kappa_4(y_\alpha, z_{\underline{\alpha}})v^2(y_\alpha, z_{\underline{\alpha}})]\, dz_{\underline{\alpha}}.$$

Although the instrumental variable estimators achieve the one-dimensional optimal convergence rate, there is room for improvement in terms of variance. For example, compared with the marginal integration estimators of Linton and Härdle (1996) or Linton and Nielsen (1995), the asymptotic variances of the instrumental variable estimates for $m_1(\cdot)$ in Theorems 1 and 2 include an additional factor of $m_2^2(\cdot)$. This is because the instrumental variable approach treats $\eta = m_2(X_2) + \varepsilon$ in (2.6) as if it were the error term of the regression equation for $m_1(\cdot)$. Note that the second term of the asymptotic covariance in Theorem 2 is the same as that in Yang et al. (1999), where the authors only considered the case with additive mean and multiplicative volatility functions. The issue of efficiency in estimating an additive component was first addressed by Linton (1996) based on "oracle efficiency" bounds of infeasible estimators under the knowledge of other components. According to this, both instrumental variable and marginal integration estimators are inefficient, but they can attain the efficiency bounds through one simple additional step, following Linton (1996, 2000) and Kim et al. (1999).

5. MORE EFFICIENT ESTIMATION

5.1. Oracle Standard

In this section we define a standard of efficiency that could be achieved in the presence of certain information, and then we show how to achieve this in prac-

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

tice. There are several routes to efficiency here, depending on the assumptions one is willing to make about ε_t. We shall take an approach based on likelihood, that is, we shall assume that ε_t is i.i.d. with known density function f like the normal or t with given degrees of freedom. It is easy to generalize this to the case where f contains unknown parameters, but we shall not do so here. It is also possible to build an efficiency standard based on the moment conditions in (1.1)–(1.3). We choose the likelihood approach because it leads to easy calculations and links with existing work and is the most common method for estimating parametric ARCH/GARCH models in applied work.

There are several standards that we could apply here. First, suppose that we know $(c_m, \{m_\beta(\cdot) : \beta \neq \alpha\})$ and $(c_v, \{v_\alpha(\cdot) : \alpha\})$; then what is the best estimator we can obtain for the function m_α within the local polynomial paradigm? Similarly, suppose that we know $(c_m, \{m_\alpha(\cdot) : \alpha\})$ and $(c_v, \{v_\beta(\cdot) : \beta \neq \alpha\})$; then what is the best estimator we can obtain for the function v_α? It turns out that this standard is very high and cannot be achieved in practice. Instead we ask: suppose that we know $(c_m, \{m_\beta(\cdot) : \beta \neq \alpha\})$ and $(c_v, \{v_\beta(\cdot) : \beta \neq \alpha\})$; then what is the best estimator we can obtain for the functions (m_α, v_α)? It turns out that this standard can be achieved in practice. Let π denote $-\log f(\cdot)$, where $f(\cdot)$ is the density function of ε_t. We use z_t to denote (x_t, y_t), where $x_t = (y_{t-1}, \ldots, y_{t-d}) = (y_{t-\alpha}, \underline{y}_{t-\alpha})$. For $\theta = (\theta_a, \theta_b) = (a_m, a_v, b_m, b_v)$, we define

$$l_t^*(\theta, \gamma_\alpha) = l^*(z_t; \theta, \gamma_\alpha) = \pi\left(\frac{y_t - F_m(\gamma_{1\alpha}(\underline{y}_{t-\alpha}) + a_m + b_m(y_{t-\alpha} - y_\alpha))}{F_v^{1/2}(\gamma_{2\alpha}(\underline{y}_{t-\alpha}) + a_v + b_v(y_{t-\alpha} - y_\alpha))} \right)$$

$$+ \frac{1}{2} \log F_v(\gamma_{2\alpha}(\underline{y}_{t-\alpha}) + a_v + b_v(y_{t-\alpha} - y_\alpha)),$$

$$l_t(\theta, \gamma_\alpha) = l(z_t; \theta, \gamma_\alpha) = K_h(y_{t-\alpha} - y_\alpha) l^*(z_t; \theta, \gamma_\alpha), \tag{5.15}$$

where $\gamma_\alpha(\underline{y}_{t-\alpha}) = (\gamma_{1\alpha}(\underline{y}_{t-\alpha}), \gamma_{2\alpha}(\underline{y}_{t-\alpha})) = (c_m + m_{\underline{\alpha}}(\underline{y}_{t-\alpha}), c_v + v_{\underline{\alpha}}(\underline{y}_{t-\alpha})) = (c_m + \sum_{\beta \neq \alpha}^{d} m_\beta(y_{t-\beta}), c_v + \sum_{\beta \neq \alpha}^{d} v_\beta(y_{t-\beta}))$. With $l_t(\theta, \gamma_\alpha)$ being the (negative) conditional local log likelihood, the infeasible local likelihood estimator $\hat{\theta} = (\hat{a}_m, \hat{a}_v, \hat{b}_m, \hat{b}_v)$ is defined by the minimizer of

$$Q_n(\theta) = \sum_{t=d+1}^{n'} l_t(\theta, \gamma_\alpha^0),$$

where $\gamma_\alpha^0(\cdot) = (\gamma_{1\alpha}^0(\cdot), \gamma_{2\alpha}^0(\cdot)) = (c_m^0 + m_{\underline{\alpha}}^0(\cdot), c_v^0 + v_{\underline{\alpha}}^0(\cdot))$. From the definition for the score function

$$s_t^*(\theta, \gamma_\alpha) = s^*(z_t; \theta, \gamma_\alpha) = \frac{\partial l^*(z_t; \theta, \gamma_\alpha)}{\partial \theta},$$

$$s_t(\theta, \gamma_\alpha) = s(z_t; \theta, \gamma_\alpha) = \frac{\partial l(z_t; \theta, \gamma_\alpha)}{\partial \theta},$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

the first-order condition for $\hat{\theta}$ is given by

$$0 = \bar{s}_n(\hat{\theta}, \gamma_\alpha^0) = \frac{1}{n} \sum_{t=d+1}^{n'} s_t(\hat{\theta}, \gamma_\alpha^0).$$

The asymptotic distribution of the local maximum likelihood estimator has been studied by Avramidis (2002). For $y = (y_1, \ldots, y_d) = (y_\alpha, \underset{\sim}{y}_\alpha)$, define

$$V_\alpha = V_\alpha(y_\alpha) = \int V(y; \theta_0, \gamma_\alpha^0) p(y)\, d\underset{\sim}{y}_\alpha;$$

$$D_\alpha = D(y_\alpha) = \int D(y; \theta_0, \gamma_\alpha^0) p(y)\, d\underset{\sim}{y}_\alpha,$$

where

$$V(y; \theta, \gamma_\alpha) = E[s^*(z_t; \theta, \gamma_\alpha) s^*(z_t; \theta, \gamma_\alpha)^\top | x_t = y];$$
$$D(y; \theta, \gamma_\alpha) = E(\nabla_\theta s_t^*(z_t; \theta, \gamma_\alpha) | x_t = y).$$

With a minor generalization of the results by Avramidis (2002, Theorem 2), we obtain the following asymptotic properties for the infeasible estimators: $\hat{\varphi}_\alpha^{inf}(y_\alpha) = [\hat{m}_\alpha^{inf}(y_\alpha), \hat{v}_\alpha^{inf}(y_\alpha)]^\top = [\hat{a}_m, \hat{a}_v]^\top$. Let $\varphi_\alpha^c(y_\alpha) \equiv [m_\alpha(y_\alpha), v_\alpha(y_\alpha)]^\top$, that is, $\varphi_\alpha^c(y_\alpha) = \varphi_\alpha(y_\alpha) - c$, where $c = (c_m, c_v)$.

THEOREM 3. *Under Assumption C in the Appendix, it holds that*

$$\sqrt{nh}[\hat{\varphi}_\alpha^{inf}(y_\alpha) - \varphi_\alpha^c(y_\alpha) - B_\alpha] \xrightarrow{d} N[0, \Omega_\alpha^*(y_\alpha)],$$

where $B_\alpha = \frac{1}{2} h^2 \mu_K^2 [m_\alpha''(y_\alpha), v_\alpha''(y_\alpha)]^\top$ *and* $\Omega_\alpha^*(y_\alpha) = \|K\|_2^2 D_\alpha^{-1} V_\alpha D_\alpha^{-1}$.

A more specific form for the asymptotic variance can be calculated. For example, suppose that the error density function, $f(\cdot)$, is symmetric. Then, the asymptotic variance of the volatility function is given by

$$\omega_{22}(y_\alpha) = \frac{\int \left\{ \int g^2(y) f(y)\, dy \right\} (\nabla F_v / F_v)^2 (G_v(v(y))) p(y)\, d\underset{\sim}{y}_\alpha}{\left[\int \left\{ \int q(y) f(y)\, dy \right\} (\nabla F_v / F_v)^2 (G_v(v(y))) p(y)\, d\underset{\sim}{y}_\alpha \right]^2},$$

where $g(y) = f'(y) f^{-1}(y) y + 1$ and $q(y) = [y^2 f''(y) f(y) + y f'(y) f(y) - y^2 f'(y)^2] f^{-2}(y)$.

When the error distribution is Gaussian, we can further simplify the asymptotic variance; that is,

$$\omega_{11}(y_\alpha) = \left[\int v^{-1}(y) \nabla F_m^2 (G_m(m(y))) p(y)\, d\underset{\sim}{y}_\alpha \right]^{-1}; \qquad \omega_{12} = \omega_{21} = 0;$$

$$\omega_{22}(y_\alpha) = 2 \left[\int v^{-2}(y) \nabla F_v^2 (G_v(v(y))) p(y)\, d\underset{\sim}{y}_\alpha \right]^{-1}.$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

In this case, one can easily find the infeasible estimator to have lower asymptotic variance than the instrumental variable estimator. To see this, we note that $\nabla G_m = 1/\nabla F_m$ and $\|K\|_2^2 \le \|(K * K)_0\|_2^2$ and apply the Cauchy–Schwarz inequality to get

$$\|(K * K)_0\|_2^2 \int \frac{p_{\underline{\alpha}}^2(\underline{y}_\alpha)}{p(y_\alpha, \underline{y}_\alpha)} \nabla G_m(m)^2 v(y_\alpha, \underline{y}_\alpha)\, d\underline{y}_\alpha$$
$$\ge \|K\|_2^2 \left[\int v^{-1}(y_\alpha, \underline{y}_\alpha) \nabla F_m^2(G_m(m)) p(y_\alpha, \underline{y}_\alpha)\, d\underline{y}_\alpha \right]^{-1}.$$

In a similar way, from $\kappa_4 = 3$ due to the Gaussianity assumption on ε, it follows that

$$\|(K * K)_0\|_2^2 \kappa_4 \int \frac{p_{\underline{\alpha}}^2(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} \nabla G_v(v)^2 v^2(y_\alpha, y_{\underline{\alpha}})\, dy_{\underline{\alpha}}$$
$$\ge 2 \left[\int v^{-2}(y) \nabla F_v^2(G_v(v(y))) p(y)\, d\underline{y}_\alpha \right]^{-1}.$$

These, together with $\kappa_3 = 0$, imply that the second term of $\Sigma_\alpha^*(y_\alpha)$ in Theorem 1 is greater than $\Omega_\alpha^*(y_\alpha)$ in the sense of positive definiteness, and hence $\Sigma_\alpha^*(y_\alpha) \ge \Omega_\alpha^*(y_\alpha)$, because the first term of $\Sigma_\alpha^*(y_\alpha)$ is a nonnegative matrix. The infeasible estimator is more efficient than the instrumental variable estimator because the former uses more information concerning the mean-variance structure. We finally remark that the infeasible estimator is also more efficient that the marginal integration estimator in Yang et al. (1999) whose asymptotic variance corresponds to the second term of $\Sigma_\alpha^*(y_\alpha)$; see the discussion following Theorem 2.

5.2. Feasible Estimation

Let $(\tilde{c}_m, \{\tilde{m}_\beta(\cdot) : \beta \ne \alpha\})$ and $(\tilde{c}_v, \{\tilde{v}_\beta(\cdot) : \beta \ne \alpha\})$ be the estimators from (3.12) and (3.13) in Section 3, with the common bandwidth parameter h_0 chosen for the kernel function $K(\cdot)$. We define the feasible local likelihood estimator $\hat{\theta}^* = (\hat{a}_m^*, \hat{a}_v^*, \hat{b}_m^*, \hat{b}_v^*)$ as the minimizers of

$$\tilde{Q}_n(\theta) = \sum_{t=d+1}^{n'} l_t(\theta, \tilde{\gamma}_\alpha),$$

where $\tilde{\gamma}_\alpha(\cdot) = (\tilde{\gamma}_{1\alpha}(\cdot), \tilde{\gamma}_{2\alpha}(\cdot)) = (\tilde{c}_m + \tilde{m}_{\underline{\alpha}}(\cdot), \tilde{c}_v + \tilde{v}_{\underline{\alpha}}(\cdot))$ and $l_t(\cdot)$ is given by (5.15), with the additional bandwidth parameter h, possibly different from h_0. Then, the first-order condition for $\hat{\theta}^*$ is given by

$$0 = \bar{s}_n(\hat{\theta}^*, \tilde{\gamma}_\alpha) = \frac{1}{n} \sum_{t=d+1}^{n'} s_t(\hat{\theta}^*, \tilde{\gamma}_\alpha). \tag{5.16}$$

Let $\hat{\varphi}_\alpha^*(y_\alpha) = (\hat{m}_\alpha^*(y_\alpha), \hat{v}_\alpha^*(y_\alpha))^\top = (\hat{a}_m^*, \hat{a}_v^*)^\top$. We have the following result.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

THEOREM 4. *Under Assumptions B and C in the Appendix, it holds that*

$\sqrt{nh}[\hat{\varphi}_\alpha^*(y_\alpha) - \hat{\varphi}_\alpha^{inf}(y_\alpha)] \xrightarrow{p} 0.$

This result shows that the oracle efficiency bound is achieved by the two-step estimator.

6. NUMERICAL EXAMPLES

A small-scale simulation is carried out to investigate the finite-sample properties of both the instrumental variable and two-step estimators. The design in our experiment is additive nonlinear ARCH(2):

$$y_t = [0.2 + v_1(y_{t-1}) + v_2(y_{t-2})]\varepsilon_t,$$

$$v_1(y) = 0.4\Phi_N(|2y|)[2 - \Phi_N(y)]y^2,$$

$$v_2(y) = 0.4\{1/\sqrt{1 + 0.1y^2} + \ln(1 + 4y^2) - 1\},$$

where $\Phi_N(\cdot)$ is the (cumulative) standard normal distribution function and ε_t is i.i.d. with $N(0,1)$. Figure 1 (solid lines) depicts the shapes of the volatility functions defined by $v_1(\cdot)$ and $v_2(\cdot)$. Based on the preceding model, we simulate 500 samples of ARCH processes with sample size $n = 500$. For each realization of the ARCH process, we apply the instrumental variable estimation procedure in (3.13) with $\tilde{r}_t = y_t^2$ to get preliminary estimates of $v_1(\cdot)$ and $v_2(\cdot)$. Those estimates then are used to compute the two-step estimates of volatility functions based on the feasible local maximum likelihood estimator in Section 5.2, under the normality assumption for the errors. The infeasible oracle estimates are also provided for comparisons. The Gaussian kernel is used for all the nonparametric estimates, and bandwidths are chosen according to the rule of thumb (Härdle, 1990), $h = c_h std(y_t)n^{-1/(4+d)}$, where $std(y_t)$ is the standard deviation of y_t. We fix $c_h = 1$ for both the density estimates (for computing the instruments, W) and instrumental variable estimates in (3.13) and $c_h = 1.5$ for the (feasible and infeasible) local maximum likelihood estimator. To evaluate the performance of the estimators, we calculate the mean squared error, together with the mean absolute deviation error, for each simulated datum; for $\alpha = 1,2$,

$$e_{\alpha,MSE} = \left\{ \frac{1}{50} \sum_{i=1}^{50} [v_\alpha(y_i) - \hat{v}_\alpha(y_i)]^2 \right\}^{1/2},$$

$$e_{\alpha,MAE} = \frac{1}{50} \sum_{i=1}^{50} |v_\alpha(y_i) - \hat{v}_\alpha(y_i)|,$$

where $\{y_1,..,y_{50}\}$ are grid points on $[-1,1)$. The grid range covers about 70% of the observations on average. Table 1 gives averages of $e_{\alpha,MSE}$'s and $e_{\alpha,MAE}$'s from 500 repetitions.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

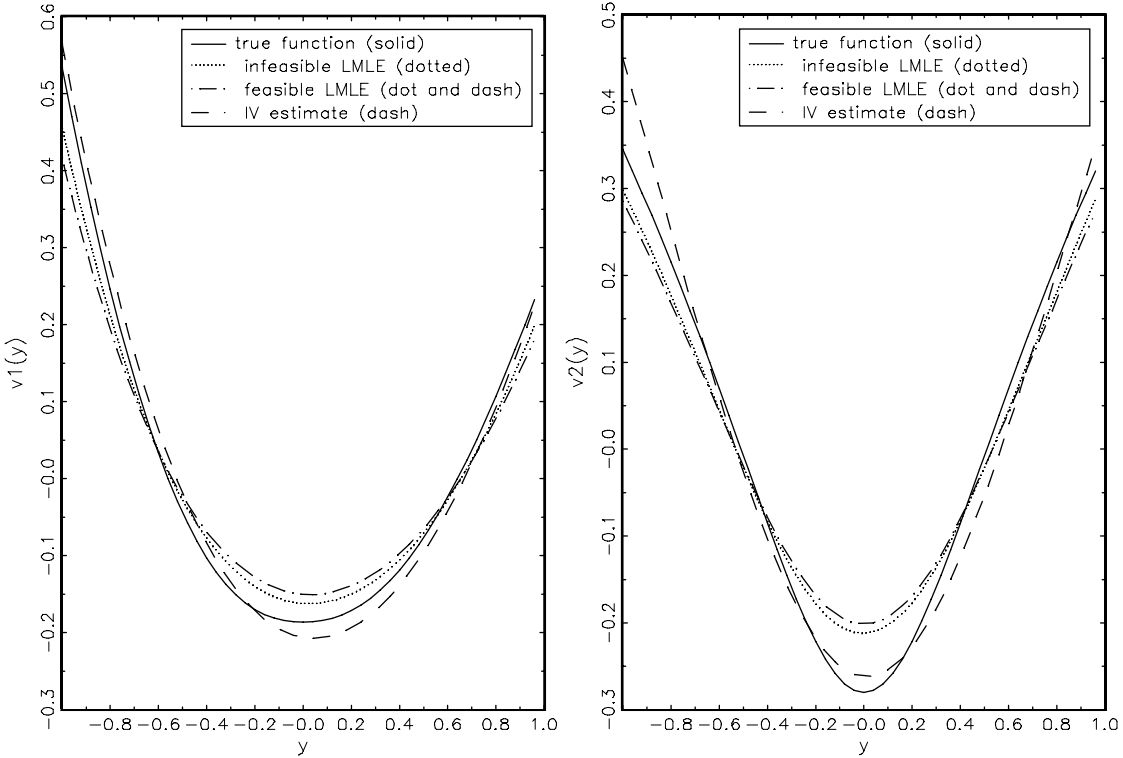

FIGURE 1. Averages of volatility estimates (demeaned): (a) first lag; (b) second lag.

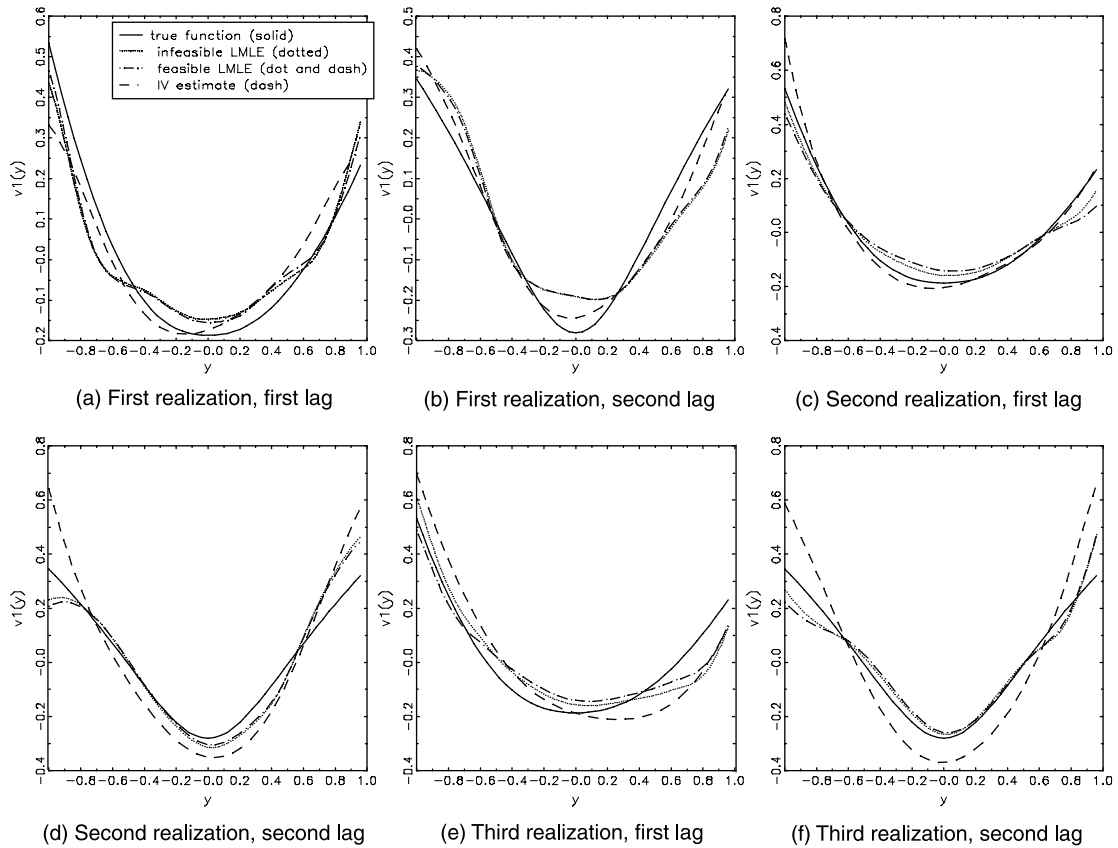

FIGURE 2. Volatility estimates (demeaned).

TABLE 1. Averages MSE and MAE for three volatility estimators

	$e_{1,MSE}$	$e_{2,MSE}$	$e_{1,MAE}$	$e_{2,MAE}$
Oracle est.	0.07636	0.08310	0.06049	0.06816
IV est.	0.08017	0.11704	0.06660	0.09725
Two-step	0.08028	0.08524	0.06372	0.07026

Table 1 shows that the infeasible oracle estimator is the best out of the three, as would be expected. The performance of the instrumental variable estimator seems to be reasonably good, compared to the local maximum likelihood estimators, at least in estimating the volatility function of the first lagged variable. However, the overall accuracy of the instrumental variable estimates is improved by the two-step procedure, which behaves almost as well as the infeasible one, confirming our theoretical results in Theorem 4. For more comparisons, Figure 1 shows the averaged estimates of volatility functions, where the averages are made, at each grid, over 500 simulations. In Figure 2, we also illustrate the estimates for three typical (consecutive) realizations of ARCH processes.

NOTES

1. The extension to allow the F transformations to be of unknown functional form is considerably more complicated; see Horowitz (2001).

2. Note the contrast with the marginal integration or projection method. In this approach one defines m_1 by some unconditional expectation

$$m_1(x_1) = E[m(x_1, X_2)W(X_2)]$$

for some weighting function W that depends only on X_2 and that satisfies

$$E[W(X_2)] = 1; \qquad E[W(X_2)m_2(X_2)] = 0.$$

3. If instead we take

$$W(X) = \frac{p_1(X_1)p_2(X_2)}{p(X)},$$

this satisfies $E(W|X_1) = 1$ and $E(W\eta|X_1) = 0$. However, the term $p_1(X_1)$ cancels out of the expression and is redundant.

4. For simplicity, we choose the common bandwidth parameter for the kernel function $K(\cdot)$ in (3.12) and (3.13), which amounts to undersmoothing (for our choice of h) for the purposes of estimating m. Undersmoothing in the preliminary estimation of step I allows us control over the biases from estimating m and v. In addition, the convolution kernel function in the asymptotic variance of Theorem 1 relies on the condition of the same bandwidth for $K(\cdot)$.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

REFERENCES

Andrews, D.W.K. (1994) Empirical process methods in econometrics. In R.F. Engle & D. McFadden (eds.), *Handbook of Econometrics*, vol. IV, pp. 2247–2294. North-Holland.

Auestadt, B. & D. Tjøstheim (1990) Identification of nonlinear time series: First order characterization and order estimation. *Biometrika* 77, 669–687.

Avramidis, P. (2002) Local maximum likelihood estimation of volatility function. Manuscript, LSE.

Breiman, L. & J.H. Friedman (1985) Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association* 80, 580–619.

Buja, A., T. Hastie, & R. Tibshirani (1989) Linear smoothers and additive models (with discussion). *Annals of Statistics* 17, 453–555.

Cai, Z. & E. Masry (2000) Nonparametric estimation of additive nonlinear ARX time series: Local linear fitting and projections. *Econometric Theory* 16, 465–501.

Carrasco, M. & X. Chen (2002) Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* 18, 17–39.

Chen, R. (1996) A nonparametric multi-step prediction estimator in Markovian structures. *Statistical Sinica* 6, 603–615.

Chen, R. & R.S. Tsay (1993a) Nonlinear additive ARX models. *Journal of the American Statistical Association* 88, 955–967.

Chen, R. & R.S. Tsay (1993b) Functional-coefficient autoregressive models. *Journal of the American Statistical Association* 88, 298–308.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987–1008.

Fan, J. & Q. Yao (1996) Efficient estimation of conditional variance functions in stochastic regression. *Biometrica* 85, 645–660.

Gozalo, P. & O. Linton (2000) Local nonlinear least squares: Using parametric information in nonparametric regression. *Journal of Econometrics* 99(1), 63–106.

Hall, P. & C. Heyde (1980) *Martingale Limit Theory and Its Application.* Academic Press.

Härdle, W. (1990) *Applied Nonparametric Regression.* Econometric Monograph Series 19. Cambridge University Press.

Härdle, W. & A.B. Tsybakov (1997) Locally polynomial estimators of the volatility function. *Journal of Econometrics* 81, 223–242.

Härdle, W., A.B. Tsybakov, & L. Yang (1998) Nonparametric vector autoregression. *Journal of Statistical Planning and Inference* 68(2), 221–245.

Härdle, W. & P. Vieu (1992) Kernel regression smoothing of time series. *Journal of Time Series Analysis* 13, 209–232.

Hastie, T. & R. Tibshirani (1990) *Generalized Additive Models.* Chapman and Hall.

Hastie, T. & R. Tibshirani (1987) Generalized additive models: Some applications. *Journal of the American Statistical Association* 82, 371–386.

Horowitz, J. (2001) Estimating generalized additive models. *Econometrica* 69, 499–513.

Jones, M.C., S.J. Davies, & B.U. Park (1994) Versions of kernel-type regression estimators. *Journal of the American Statistical Association* 89, 825–832.

Kim, W., O. Linton, & N. Hengartner (1999) A computationally efficient oracle estimator of additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics* 8, 1–20.

Linton, O.B. (1996) Efficient estimation of additive nonparametric regression models. *Biometrika* 84, 469–474.

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

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

Linton, O.B. & J. Nielsen (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.
Linton, O.B., J. Nielsen, & S. van de Geer (2003) Estimating multiplicative and additive hazard functions by kernel methods. *Annals of Statistics* 31, 464–492.
Linton, O.B., N. Wang, R. Chen, & W. Härdle (1995) An analysis of transformation for additive nonparametric regression. *Journal of the American Statistical Association* 92, 1512–1521.
Mammen, E., O.B. Linton, & J. Nielsen (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* 27, 1443–1490.
Masry, E. (1996) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17, 571–599.
Masry, E. & D. Tjøstheim (1995) Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory* 11, 258–289.
Masry, E. & D. Tjøstheim (1997) Additive nonlinear ARX time series and projection estimates. *Econometric Theory* 13, 214–252.
Nelson, D.B. (1991) Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59, 347–370.
Newey, W.K. (1994) Kernel estimation of partial means. *Econometric Theory* 10, 233–253.
Opsomer, J.D. & D. Ruppert (1997) Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* 25, 186–211.
Pollard, D. (1990) *Empirical Processes: Theory and Applications.* CBMS Conference Series in Probability and Statistics, vol. 2. Institute of Mathematical Statistics.
Robinson, P.M. (1983) Nonparametric estimation for time series models. *Journal of Time Series Analysis* 4, 185–208.
Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.
Stein, E.M. (1970) *Singular Integrals and Differentiability Properties of Functions.* Princeton University Press.
Stone, C.J. (1985) Additive regression and other nonparametric models. *Annals of Statistics* 13, 685–705.
Stone, C.J. (1986) The dimensionality reduction principle for generalized additive models. *Annals of Statistics* 14, 592–606.
Tjøstheim, D. & B. Auestad (1994) Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association* 89, 1398–1409.
Volkonskii, V. & Y. Rozanov (1959) Some limit theorems for random functions. *Theory of Probability and Applications* 4, 178–197.
Yang, L., W. Härdle, & J. Nielsen (1999) Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis* 20, 579–604.
Ziegelmann, F. (2002) Nonparametric estimation of volatility functions: The local exponential estimator. *Econometric Theory* 18, 985–992.

APPENDIX

A.1. Proofs for Section 4. The proof of Theorem 1 consists of three steps. Without loss of generality we deal with the case $\alpha = 1$; here we will use the subscript 2, for expositional convenience, to denote the nuisance direction. That is, $p_2(\underline{y}_{k-1}) = p_{\underline{1}}(\underline{y}_{k-1})$ in the case of density function. For component functions, $m_2(\underline{y}_{k-1})$, $v_2(\underline{y}_{k-1})$, and $H_2(\underline{y}_{k-1})$ will be used instead of $m_{\underline{1}}(\underline{y}_{k-1})$, $v_{\underline{1}}(\underline{y}_{k-1})$, and $H_{\underline{1}}(\underline{y}_{k-1})$, respectively. We start by decomposing the estimation errors, $\hat{\varphi}_1(y_1) - \varphi_1(y_1)$, into the main stochastic term and bias. Use $X_n \simeq Y_n$ to denote $X_n = Y_n\{1 + o_p(1)\}$ in the following. Let $\mathrm{vec}(X)$ denote the vectorization of the elements of the matrix X along with columns.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Proof of Theorem 1.

Step I. Decompositions and Approximations. Because $\hat{\varphi}_1(y_1)$ is a column vector, the vectorization of equation (3.14) gives

$$\hat{\varphi}_1(y_1) = [I_2 \otimes e_1^\top (\mathbf{Y}_-^\top \mathbf{K} \mathbf{Y}_-)^{-1}](I_2 \otimes \mathbf{Y}_-^\top \mathbf{K}) \mathrm{vec}(\tilde{\mathbf{R}}).$$

A similar form is obtained for the true function, $\varphi_1(y_1)$,

$$[I_2 \otimes e_1^\top (\mathbf{Y}_-^\top \mathbf{K} \mathbf{Y}_-)^{-1}](I_2 \otimes \mathbf{Y}_-^\top \mathbf{K}) \mathrm{vec}(\iota \varphi_1^\top(y_1) + Y_- \nabla \varphi_1^\top(y_1)),$$

by the identity

$$\varphi_1(y_1) = \mathrm{vec}\{e_1^\top (\mathbf{Y}_-^\top \mathbf{K} \mathbf{Y}_-)^{-1} \mathbf{Y}_-^\top \mathbf{K}[\iota \varphi_1^\top(y_1) + Y_- \nabla \varphi_1^\top(y_1)]\},$$

because

$$e_1^\top (\mathbf{Y}_-^\top \mathbf{K} \mathbf{Y}_-)^{-1} \mathbf{Y}_-^\top \mathbf{K} \iota = 1, \qquad e_1^\top (\mathbf{Y}_-^\top \mathbf{K} \mathbf{Y}_-)^{-1} \mathbf{Y}_-^\top \mathbf{K} Y_- = 0.$$

By defining $D_h = \mathrm{diag}(1, h)$ and $Q_n = D_h^{-1} \mathbf{Y}_-^\top \mathbf{K} \mathbf{Y}_- D_h^{-1}$, the estimation errors are

$$\hat{\varphi}_1(y_1) - \varphi_1(y_1) = [I_2 \otimes e_1^\top Q_n^{-1}] \tau_n,$$

where

$$\tau_n = (I_2 \otimes D_h^{-1} \mathbf{Y}_-^\top \mathbf{K}) \mathrm{vec}[\tilde{\mathbf{R}} - \iota \varphi_1^\top(y_1) - Y_- \nabla \varphi_1^\top(y_1)].$$

Observing

$$\tau_n = \frac{1}{n} \sum_{k=d+1}^{n'} K_h^{\hat{W}_k}(y_{k-1} - y_1)[\tilde{r}_k - \varphi_1(y_1) - (y_{k-1} - y_1) \nabla \varphi_1(y_1)] \otimes \left(1, \frac{y_{k-1} - y_1}{h}\right)^\top,$$

where $K_h^{\hat{W}_k}(y) = K_h(y) \hat{W}_k$, it follows by adding and subtracting $r_k = \varphi_1(y_{k-1}) + H_2(\underline{y}_{k-1})$ that

$$\begin{aligned} \tau_n = {} & \frac{1}{n} \sum_{k=d+1}^{n'} K_h^{\hat{W}_k}(y_{k-1} - y_1)[\tilde{r}_k - r_k + H_2(\underline{y}_{k-1})] \otimes \left(1, \frac{y_{k-1} - y_1}{h}\right)^\top \\ & + \frac{1}{n} \sum_{k=d+1}^{n'} K_h^{\hat{W}_k}(y_{k-1} - y_1)[\varphi_1(y_{k-1}) - \varphi_1(y_1) - (y_{k-1} - y_1) \nabla \varphi_1(y_1)] \\ & \otimes \left(1, \frac{y_{k-1} - y_1}{h}\right)^\top. \end{aligned}$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

As a result of the boundedness condition in Assumption A2, the Taylor expansion applied to $[G_m(\tilde{m}(x_k)), G_v(\tilde{v}(x_k))]$ at $[m(x_k), v(x_k)]$ yields the first term of τ_n as

$$\tilde{\tau}_n \equiv \frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\hat{W}_k}(y_{k-1}-y_1)\left[\tilde{u}_k \otimes \left(1, \frac{y_{k-1}-y_1}{h}\right)^{\top}\right],$$

where $\tilde{u}_k \equiv \tilde{r}_k^1 + \tilde{r}_k^2 + H_2(\underline{y}_{k-1})$,

$$\tilde{r}_k^1 \equiv \{\nabla G_m(m(x_k))[\tilde{m}(x_k)-m(x_k)], \nabla G_v(v(x_k))[\tilde{v}(x_k)-v(x_k)]\}^{\top},$$

$$\tilde{r}_k^2 \equiv \frac{1}{2}\{D^2 G_m(m^*(x_k))[\tilde{m}(x_k)-m(x_k)]^2, D^2 G_v(v^*(x_k))[\tilde{v}(x_k)-v(x_k)]^2\}^{\top},$$

and $m^*(x_k)[v^*(x_k)]$ is between $\tilde{m}(x_k)[\tilde{v}(x_k)]$ and $m(x_k)[v(x_k)]$, respectively]. In a similar way, the Taylor expansion of $\varphi_1(y_{k-1})$ at y_1 gives the second term of τ_n as

$$s_{0n} = \frac{h^2}{2}\frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\hat{W}_k}(y_{k-1}-y_1)\left(\frac{y_{k-1}-y_1}{h}\right)^2\left[D^2\varphi_1(y_1)\otimes\left(1,\frac{y_{k-1}-y_1}{h}\right)^{\top}\right] \times (1+o_p(1)).$$

The term $\tilde{\tau}_n$ continues to be simplified by some further approximations. Define the marginal expectation of estimated density functions $\hat{p}_2(\cdot)$ and $\hat{p}(\cdot)$ as follows:

$$\bar{p}(y_{k-1}, \underline{y}_{k-2}) \equiv \int L_g(z_1-y_{k-1})L_g(z_2-\underline{y}_{k-2})p(z_1,z_2)\,dz_1\,dz_2,$$

$$\bar{p}_2(\underline{y}_{k-2}) \equiv \int L_g(z_2-\underline{y}_{k-2})p_2(z_2)\,dz_2.$$

In the first approximation, we replace the estimated instrument, $\hat{W}$, by the ratio of the expectations of the kernel density estimates, $\bar{p}_2(\underline{y}_{k-1})/\bar{p}(x_k)$ and deal with the linear terms in the Taylor expansions. That is, $\tilde{\tau}_n$ is approximated with an error of $o_p(1/\sqrt{nh})$ by $t_{1n}+t_{2n}$:

$$t_{1n} \equiv \frac{1}{n}\sum_{k=d+1}^{n'} K_h(y_{k-1}-y_1)\frac{\bar{p}_2(\underline{y}_{k-1})}{\bar{p}(x_k)}\left[\tilde{r}_k^1 \otimes \left(1,\frac{y_{k-1}-y_1}{h}\right)^{\top}\right],$$

$$t_{2n} \equiv \frac{1}{n}\sum_{k=d+1}^{n'} K_h(y_{k-1}-y_1)\frac{\bar{p}_2(\underline{y}_{k-1})}{\bar{p}(x_k)}\left[H_2(\underline{y}_{k-1}) \otimes \left(1,\frac{y_{k-1}-y_1}{h}\right)^{\top}\right],$$

based on the following results:

(i) $(1/n)\sum_{k=d+1}^{n'} K_h(y_{k-1}-y_1)[\hat{p}_2(\underline{y}_{k-1})/\hat{p}(x_k)][\tilde{r}_k^2 \otimes (1,(y_{k-1}-y_1)/h)^{\top}] = o_p(1/\sqrt{nh})$,

(ii) $(1/n)\sum_{k=d+1}^{n'} K_h(y_{k-1}-y_1)[\hat{p}_2(\underline{y}_{k-1})/\hat{p}(x_k) - \bar{p}_2(\underline{y}_{k-1})/\bar{p}(x_k)][H_2(\underline{y}_{k-1}) \otimes (1,(y_{k-1}-y_1)/h)^{\top}] = o_p(1/\sqrt{nh})$,

(iii) $(1/n)\sum_{k=d+1}^{n'} K_h(y_{k-1}-y_1)[\hat{p}_2(\underline{y}_{k-1})/\hat{p}(x_k) - \bar{p}_2(\underline{y}_{k-1})/\bar{p}(x_k)][\tilde{r}_k^1 \otimes (1,(y_{k-1}-y_1)/h)^{\top}] = o_p(1/\sqrt{nh})$.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

To show (i), consider the first two elements of the term, for example, which are bounded elementwise by

$$\max_k |\tilde{m}(x_k) - m(x_k)|^2 \frac{1}{2}\frac{1}{n}\sum_k K_h(y_{k-1} - y_1)\frac{\hat{p}_2(\underline{y}_{k-1})}{\hat{p}(x_k)} D^2 G_m(m(x_k))\left(1, \frac{y_{k-1} - y_1}{h}\right)^\top$$
$$= o_p(1/\sqrt{nh}).$$

The last equality is direct from the uniform convergence theorems in Masry (1996) that

$$\max_t |\tilde{m}(x_t) - m(x_t)| = O_p(\log n/\sqrt{nh^d}) \tag{A.1}$$

and $(1/n)\sum_k K_h(y_{k-1} - y_1)[\hat{p}_2(\underline{y}_{k-1})/\hat{p}(x_k)]D^2 G_m(m(x_k))(1, (y_{k-1} - y_1)/h)^\top = O_p(1)$. The proof for (ii) is shown by applying Lemma A.1, which follows. The negligibility of (iii) follows in a similar way from (ii), considering (A.1). Although the asymptotic properties of s_{0n} and t_{2n} are relatively easy to derive, additional approximation is necessary to make t_{1n} more tractable. Note that the estimation errors of the local linear fits, $\tilde{m}(x_k) - m(x_k)$, are decomposed into

$$\frac{1}{n}\sum_l \frac{K_h(x_l - x_k)}{p(x_l)} v^{1/2}(x_l)\varepsilon_l + \text{the remaining bias}$$

from the approximation results for the local linear smoother in Jones, Davies, and Park (1994). A similar expression holds for volatility estimates, $\tilde{v}(x_k) - v(x_k)$, with a stochastic term of $(1/n)\sum_l [K_h(x_l - x_k)/p(x_l)]v(x_l)(\varepsilon_l^2 - 1)$. Define

$$J_{k,n}(x_l) \equiv \frac{1}{nh^d}\sum_k \frac{K(y_{k-1} - y_1/h)K(x_l - x_k/h)}{p(x_l)}\frac{p_2(\underline{y}_{k-1})}{p(x_k)}$$
$$\times \left[\text{diag}(\nabla G_m, \nabla G_v) \otimes \left(1, \frac{y_{k-1} - y_1}{h}\right)^\top\right]$$

and let $\bar{J}(x_l)$ denote the marginal expectation of $J_{k,n}$ with respect to x_k. Then, the stochastic term of t_{1n}, after rearranging its the double sums, is approximated by

$$\tilde{t}_{1n} = \frac{1}{nh}\sum_l \bar{J}(x_l)[(v^{1/2}(x_l)\varepsilon_l, v(x_l)(\varepsilon_l^2 - 1))^\top \otimes I_2]$$

because the approximation error from $\bar{J}(X_l)$ negligible, that is,

$$\frac{1}{nh}\sum_l (J_{k,n} - \bar{J})[(v^{1/2}(X_l)\varepsilon_l, v(X_l)(\varepsilon_l^2 - 1))^\top \otimes I_2]^\top = o_p(1/\sqrt{nh}),$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

applying the same method as in Lemma A.1. A straightforward calculation gives

$$\bar{J}(x_l) \simeq \frac{1}{h}\int K(u_1 - y_1/h)K(u_1 - y_{l-1}/h)\int \frac{1}{h^{d-1}}K(\underline{y}_{l-1} - u_2/h)\frac{p_2(u_2)}{p(x_l)}$$

$$\times\left[\operatorname{diag}(\nabla G_m(u),\nabla G_v(u)) \otimes \left(1, \frac{u_1 - y_1}{h}\right)^\top\right] du_2\, du_1$$

$$\simeq \frac{1}{h}\int K(u_1 - y_1/h)K(u_1 - y_{l-1}/h)\frac{p_2(\underline{y}_{l-1})}{p(x_l)}$$

$$\times\left[\operatorname{diag}(\nabla G_m(u_1,\underline{y}_{l-1}),\nabla G_v(u_1,\underline{y}_{l-1})) \otimes \left(1, \frac{u_1 - y_1}{h}\right)^\top\right] du_1$$

$$\simeq \frac{p_2(\underline{y}_{l-1})}{p(x_l)}\left[\operatorname{diag}(\nabla G_m(y_1,\underline{y}_{l-1}),\nabla G_v(y_1,\underline{y}_{l-1}))\right.$$

$$\left.\otimes\left((K*K)_0\left(\frac{y_{l-1} - y_1}{h}\right),(K*K)_1\left(\frac{y_{l-1} - y_1}{h}\right)\right)^\top\right],$$

where

$$(K*K)_i\left(\frac{y_{l-1} - y_1}{h}\right) = \int w_1^i K(w_1)K\left(w_1 + \frac{y_{l-1} - y_1}{h}\right) dw.$$

Observe that $(K*K)_i((y_{l-1} - y_1)/h)$ in $\bar{J}(X_l)$ is actually a convolution kernel and behaves just like a one-dimensional kernel function of y_{l-1}. This means that the standard method (central limit theorem or law of least numbers) for univariate kernel estimates can be applied to show the asymptotics of

$$\tilde{t}_{1n} = \frac{1}{nh}\sum_l \frac{p_2(\underline{y}_{l-1})}{p(x_l)}\left\{\begin{bmatrix}\nabla G_m(y_1,\underline{y}_{l-1})v^{1/2}(X_l)\varepsilon_l \\ \nabla G_v(y_1,\underline{y}_{l-1})v(X_l)(\varepsilon_l^2 - 1)\end{bmatrix} \otimes \begin{bmatrix}(K*K)_0\left(\frac{y_{l-1} - y_1}{h}\right) \\ (K*K)_1\left(\frac{y_{l-1} - y_1}{h}\right)\end{bmatrix}\right\}.$$

If we define s_{1n} as the remaining bias term of t_{1n}, the estimation errors of $\hat{\varphi}_1(y_1) - \varphi_1(y_1)$ consist of two stochastic terms, $[I_2 \otimes e_1^\top Q_n^{-1}](\tilde{t}_{1n} + \tilde{t}_{2n})$, and three bias terms, $[I_2 \otimes e_1^\top Q_n^{-1}](s_{0n} + s_{1n} + s_{2n})$, where

$$\tilde{t}_{2n} = \frac{1}{n}\sum_{k=d+1}^{n'} K_h(y_{k-1} - y_1)\frac{p_2(\underline{y}_{k-1})}{p(X_k)}\left[H_2(\underline{y}_{k-1}) \otimes \left(1, \frac{Y_{k-1} - y_1}{h}\right)^\top\right],$$

$$s_{2n} = t_{2n} - \tilde{t}_{2n}.$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Step II. Computation of Variance and Bias. We start with showing the order of the main stochastic term,

$$\tilde{t}_n^* = \tilde{t}_{1n} + \tilde{t}_{2n} = \frac{1}{n}\sum_k \xi_k,$$

where $\xi_k = \xi_{1k} + \xi_{2k}$,

$$\xi_{1k} = \frac{p_2(\underline{y}_{k-1})}{p(y_{k-1}, \underline{y}_{k-1})}\left\{\begin{bmatrix} \nabla G_m(y_1, \underline{y}_{k-1}) v^{1/2}(X_k)\varepsilon_k \\ \nabla G_v(y_1, \underline{y}_{k-1}) v(X_k)(\varepsilon_k^2 - 1) \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{h}(K * K)_0\left(\frac{y_{k-1} - y_1}{h}\right) \\ 0 \end{bmatrix}\right\},$$

$$\xi_{2k} = \frac{p_2(\underline{y}_{k-1})}{p(y_{k-1}, \underline{y}_{k-1})}\left\{\begin{bmatrix} m_2(\underline{y}_{k-1}) \\ v_2(\underline{y}_{k-1}) \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{h}K\left(\frac{y_{k-1} - y_1}{h}\right) \\ \frac{1}{h}K\left(\frac{y_{k-1} - y_1}{h}\right)\left(\frac{y_{l-1} - y_1}{h}\right) \end{bmatrix}\right\},$$

by calculating its asymptotic variance. Dividing a normalized variance of $\tilde{t}_n^*$ into the sums of variances and covariances gives

$$\begin{aligned} \operatorname{var}(\sqrt{nh}\tilde{t}_n^*) &= \operatorname{var}\left(\frac{\sqrt{h}}{\sqrt{n}}\sum_k \xi_k\right) = \frac{h}{n}\sum_k \operatorname{var}(\xi_k) + \frac{h}{n}\sum\sum_{k\neq l} \operatorname{cov}(\xi_k, \xi_l) \\ &= h\operatorname{var}(\tilde{\xi}_k) + \sum_k \left[\frac{n-k}{n}\right] h[\operatorname{cov}(\xi_d, \xi_{d+k})], \end{aligned}$$

where the last equality comes from the stationarity assumption.

We claim that

(a) $h\operatorname{var}(\xi_k) \to \Sigma_1(y_1)$,
(b) $\sum_k [1 - (k/n)] h \operatorname{cov}(\xi_d, \xi_{d+k}) = o(1)$, and
(c) $nh\operatorname{var}(\tilde{t}_n^*) \to \Sigma_1(y_1)$,

where

$$\begin{aligned} \Sigma_1(y_1) = &\left\{\int \frac{p_2^2(z_2)}{p(y_1, z_2)}\begin{bmatrix} \nabla G_m(y_1, z_2)^2 v(y_1, z_2) & (\nabla G_m \cdot \nabla G_v)(\kappa_3 \cdot v^{3/2})(y_1, z_2) \\ (\nabla G_m \cdot \nabla G_v)(\kappa_3 \cdot v^{3/2})(y_1, z_2) & \nabla G_v(y_1, z_2)^2 \kappa_4(y_1, z_2) v^2(y_1, z_2) \end{bmatrix} dz_2 \right. \\ &\left. \otimes \begin{bmatrix} \|(K * K)_0\|_2^2 & 0 \\ 0 & 0 \end{bmatrix}\right\} \\ &+ \int \frac{p_2^2(z_2)}{p(y_1, z_2)} H_2(z_2) H_2^\top(z_2)\, dz_2 \otimes \begin{bmatrix} \|K\|_2^2 & 0 \\ 0 & \int K^2(u)u^2\, du \end{bmatrix}. \end{aligned}$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Proof of (a). Noting $E(\xi_{1k}) = E(\xi_{2k}) = 0_{4\times 1}$ and $E(\xi_{1k}\xi_{2k}^\top) = 0_{4\times 4}$,

$$h\,\mathrm{var}(\xi_k) = hE(\xi_{1k}\xi_{1k}^\top) + hE(\xi_{2k}\xi_{2k}^\top)$$

by the stationarity assumption. Applying the integration with substitution of variable and Taylor expansion, the expectation term is

$$hE(\xi_{1k}\xi_{1k}^\top) = \left\{ \int \frac{p_2^2(z_2)}{p(y_1,z_2)} \begin{bmatrix} \nabla G_m(y_1,z_2)^2 v(y_1,z_2) & (\nabla G_m \cdot \nabla G_v)(\kappa_3 \cdot v^{3/2})(y_1,z_2) \\ (\nabla G_m \cdot \nabla G_v)(\kappa_3 \cdot v^{3/2})(y_1,z_2) & \nabla G_v(y_1,z_2)^2 \kappa_4(y_1,z_2) v^2(y_1,z_2) \end{bmatrix} dz_2 \right.$$

$$\left. \otimes \begin{bmatrix} \|(K*K)_0\|_2^2 & 0 \\ 0 & 0 \end{bmatrix} \right\}$$

and

$$hE(\xi_{2k}\xi_{2k}^\top) = \int \frac{p_2^2(z_2)}{p(y_1,z)} \begin{bmatrix} m_2^2(z_2) & m_2(z_2)v_2(z_2) \\ m_2(z_2)v_2(z_2) & v_2^2(z_2) \end{bmatrix} dz_2$$

$$\otimes \begin{bmatrix} \|K\|_2^2 & 0 \\ 0 & \int K^2(u)u^2\,du \end{bmatrix} + o(1),$$

where $\kappa_3(y_1,z_2) = E[\varepsilon_t^3 | x_t = (y_1,z_2)]$ and $\kappa_4(y_1,z_2) = E[(\varepsilon_t^2 - 1)^2 | x_t = (y_1,z_2)]$. ■

Proof of (b). Because $E(\xi_{1k}\xi_{1j}^\top)|_{j\neq k} = E(\xi_{1k}\xi_{2j}^\top)|_{j\neq k} = 0$, $\mathrm{cov}(\xi_{d+1},\xi_{d+1+k}) = \mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})$. By setting $c(n)h \to 0$, as $n \to \infty$, we separate the covariance terms into two parts:

$$\sum_{k=1}^{c(n)} \left[1 - \frac{k}{n}\right] h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k}) + \sum_{k=c(n)+1}^{n'} \left[1 - \frac{k}{n}\right] h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k}).$$

To show the negligibility of the first part of the covariances, consider that the dominated convergence theorem used after Taylor expansion and the integration with substitution of variables gives

$$|\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})|$$

$$\simeq \left| \int H_2(\underline{y}_d) H_2^\top(\underline{y}_{d+k}) \frac{p(y_1,\underline{y}_d,y_1,\underline{y}_{d+k})}{p_{1|2}(y_1|\underline{y}_d)p_{1|2}(y_1|\underline{y}_{d+k})}\, d(\underline{y}_d,\underline{y}_{d+k}) \right| \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Therefore, it follows from the assumption on the boundedness condition in Assumption A2 that

$$|\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})| \le E|H_2(\underline{y}_d)|E|H_2^{\top}(\underline{y}_{d+k})|$$

$$\times \int \frac{p(y_1,\underline{y}_d,y_1,\underline{y}_{d+k})}{p_{1|2}(y_1|\underline{y}_d)p_{1|2}(y_1|\underline{y}_{d+k})}\, d(\underline{y}_d,\underline{y}_{d+k}) \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\equiv A^*,$$

where $A \le B$ means $a_{ij} \le b_{ij}$, for all element of matrices A and B. By the construction of $c(n)$,

$$\sum_{k=1}^{c(n)} \left[1 - \frac{k}{n}\right] h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})$$

$$\le 2c(n)|h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})| \le 2c(n)hA^* \to 0, \quad \text{as } n \to \infty.$$

Next, we turn to the negligibility of the second part of the covariances:

$$\sum_{k=c(n)+1}^{n'} \left[1 - \frac{k}{n}\right] h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k}).$$

Let ξ_{2k}^i be the ith element of ξ_{2k}, for $i = 1,\ldots,4$. Using Davydov's lemma (in Hall and Heyde, 1980, Theorem A.5), we obtain

$$|h\,\mathrm{cov}(\xi_{2d+1}^i,\xi_{2d+1+k}^j)| = |\mathrm{cov}(\sqrt{h}\xi_{2d+1}^i,\sqrt{h}\xi_{2d+1+k}^j)|$$

$$\le 8[\alpha(k)^{1-2/\upsilon}]\left[\max_{i=1,\ldots,4} E(\sqrt{h}|\xi_{2k}^i|^{\upsilon})\right]^{2/\upsilon}$$

for some $\upsilon > 2$. The boundedness of $E(\sqrt{h}|\xi_{2k}^1|^{\upsilon})$, for example, is evident from the direct calculation that

$$\xi_{2k} = \frac{p_2(\underline{y}_k)}{p(x_k)} \left\{ \begin{bmatrix} m_2(\underline{y}_d) \\ v_2(\underline{y}_d) \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{h} K\left(\frac{y_{k-1}-y_1}{h}\right) \\ \frac{1}{h} K\left(\frac{y_{k-1}-y_1}{h}\right)\left(\frac{y_{k-1}-y_1}{h}\right) \end{bmatrix} \right\},$$

$$E(|\sqrt{h}\xi_{2k}^1|^{\upsilon}) \simeq \frac{h^{\upsilon/2}}{h^{\upsilon-1}} \int \frac{p_2^{\upsilon}(z_2)}{p^{\upsilon-1}(y_1,z_2)}\, |m_2^{\upsilon}(z_2)|\, dz_2$$

$$= O\left(\frac{h^{\upsilon/2}}{h^{\upsilon-1}}\right) = O\left(\frac{1}{h^{\upsilon/2-1}}\right).$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Thus, the covariance is bounded by

$$|h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})| \le C\left[\frac{1}{h^{v/2-1}}\right]^{2/v}[\alpha(k)^{1-2/v}].$$

This implies

$$\sum_{k=c(n)+1}^{n'}\left[1-\frac{k}{n}\right]h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})$$

$$\le 2\sum_{k=c(n)+1}^{\infty}|h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k})| \le C'\left[\frac{1}{h^{1-2/v}}\right]\sum_{k=c(n)+1}^{\infty}[\alpha(k)^{1-2/v}]$$

$$= C'\sum_{k=c(n)+1}^{\infty}\left[\frac{1}{h^{1-2/v}}\right][\alpha(k)^{1-2/v}] \le C'\sum_{k=c(n)+1}^{\infty}k^a[\alpha(k)^{1-2/v}],$$

if a is such that

$$k^a \ge (c(n)+1)^a \ge c(n)^a = \frac{1}{h^{1-2/v}},$$

for example, $c(n)^a h^{1-2/v} = 1$, which implies $c(n) \to \infty$. If we further restrict a such that

$$0 < a < 1-\frac{2}{v},$$

then

$$c(n)^a h^{1-2/v} = 1 \quad \text{implies } c(n)^a h^{1-2/v} = [c(n)h]^{1-2/v}c(n)^{-\delta} = 1, \quad \text{for } \delta > 0.$$

Thus, $c(n)h \to 0$ as required. Therefore,

$$\sum_{k=c(n)+1}^{n'}\left[1-\frac{k}{n}\right]h\,\mathrm{cov}(\xi_{2d+1},\xi_{2d+1+k}) \le C'\sum_{k=c(n)+1}^{\infty}k^a[\alpha(k)^{1-2/v}] \to 0,$$

as n goes to ∞. ■

The proof of (c) is immediate from (a) and (b).

Next, we consider the asymptotic bias. Using the standard result on the kernel weighted sum of the stationary series, we first get

$$s_{0n} \xrightarrow{p} \frac{h^2}{2}[D^2\varphi_1(y_1) \otimes (\mu_K^2,0)^\top],$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

because

$$\begin{aligned}
&\frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\hat{W}}(y_{k-1}-y_1)\left(\frac{y_{k-1}-y_1}{h}\right)^2\left[D^2\varphi_1(y_1)\otimes\left(1,\frac{y_{k-1}-y_1}{h}\right)^\top\right]\\
&\quad\xrightarrow{p}\int K_h^{\hat{W}}(z_1-y_1)\left(\frac{z_1-y_1}{h}\right)^2\left[D^2\varphi_1(y_1)\otimes\left(1,\frac{z_1-y_1}{h}\right)^\top\right]p(z)\,dz\\
&\quad\simeq\int K_h(z_1-y_1)p_2(z_2)\left(\frac{z_1-y_1}{h}\right)^2\left[D^2\varphi_1(y_1)\otimes\left(1,\frac{z_1-y_1}{h}\right)^\top\right]dz\\
&\quad=\int K_h(z_1-y_1)\left(\frac{z_1-y_1}{h}\right)^2\left[D^2\varphi_1(y_1)\otimes\left(1,\frac{z_1-y_1}{h}\right)^\top\right]dz_1\\
&\quad=\left[D^2\varphi_1(y_1)\otimes\int K_h(z_1-y_1)\left(\frac{z_1-y_1}{h}\right)^2\left(1,\frac{z_1-y_1}{h}\right)^\top dz_1\right]\\
&\quad=[D^2\varphi_1(y_1)\otimes(\mu_K^2,0)^\top].
\end{aligned}$$

For the asymptotic bias of s_{1n}, we again use the approximation results in Jones et al. (1994). Then, the first component of s_{1n}, for example, is

$$\begin{aligned}
&\frac{1}{n}\sum_k K_h(y_{k-1}-y_1)\,\frac{p_2(\underline{y}_{k-1})}{p(x_k)}\nabla G_m(m(x_k))\\
&\quad\times\left\{\frac{1}{2}\,\frac{1}{n}\sum_l\frac{K_h(x_l-x_k)}{p(x_l)}\sum_{\alpha=1}^{d}(y_{l-\alpha}-y_{k-\alpha})^2\,\frac{\partial^2 m(x_k)}{\partial y_{k-\alpha}^2}\right\}
\end{aligned}$$

and converges to

$$\frac{h^2}{2}\int p_2(z_2)\nabla G_m(m(y_1,z_2))[\mu_{K*K}^2 D^2 m_1(y_1)+\mu_K^2 D^2 m_2(z_2)]\,dz_2,$$

based on the argument for the convolution kernel given previously. A convolution of symmetric kernels is symmetric, so that $\int(K*K)_0(u)u\,du=0$ and $\int(K*K)_1(u)u^2\,du=\int\int wK(w)K(w+u)u^2\,dw\,du=0$. This implies that

$$\begin{aligned}
\tilde{s}_{1n}\xrightarrow{p}\frac{h^2}{2}\int p_2(z_2)\{&[\nabla G_m(m(y_1,z_2)),\nabla G_v(v(y_1,z_2))]^\top\\
&\odot[\mu_{K*K}^2 D^2\varphi_1(y_1)+\mu_K^2 D^2\varphi_2(z_2)]\}\otimes(1,0)^\top dz_2.
\end{aligned}$$

To calculate s_{2n}, we use the Taylor series expansion of $\bar{p}_2(\underline{y}_{k-1})/\bar{p}(X_k)$:

$$\begin{aligned}
&\left[\bar{p}_2(\underline{y}_{k-1})-\frac{p_2(\underline{y}_{k-1})\bar{p}(X_k)}{p(X_k)}\right]\frac{1}{\bar{p}(X_k)}\\
&\quad=\left[\bar{p}_2(\underline{y}_{k-1})-\frac{p_2(\underline{y}_{k-1})\bar{p}(X_k)}{p(X_k)}\right]\frac{1}{p(X_k)}\times\left[1-\frac{\bar{p}(X_k)-p(X_k)}{p^2(X_k)}+\cdots\right]\\
&\quad=\frac{\bar{p}_2(\underline{y}_{k-1})}{p(X_k)}-\frac{p_2(\underline{y}_{k-1})\bar{p}(X_k)}{p^2(X_k)}+o_p(1).
\end{aligned}$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Thus,

$$s_{2n} = \frac{1}{n}\sum_{k=d+1}^{n'} K_h(y_{k-1}-y_1)\left[\frac{\bar{p}_2(\underline{y}_{k-1})}{\bar{p}(X_k)} - \frac{p_2(\underline{y}_{k-1})}{p(X_k)}\right]$$

$$\times\left[H_2(\underline{y}_{k-1})\otimes\left(1,\frac{y_{k-1}-y_1}{h}\right)^\top\right]$$

$$\xrightarrow{p}\int K_h(z_1-y_1)\left[\frac{\bar{p}_2(z_2)}{\bar{p}(z)} - \frac{p_2(z_2)}{p(z)}\right]\left[H_2(z_2)\otimes\left(1,\frac{z_1-y_1}{h}\right)^\top\right]p(z)\,dz$$

$$\simeq\int K_h(z_1-y_1)\left[\frac{\bar{p}_2(z_2)}{p(z)} - \frac{p_2(z_2)\bar{p}(z)}{p^2(z)}\right]\left[H_2(z_2)\otimes\left(1,\frac{z_1-y_1}{h}\right)^\top\right]p(z)\,dz$$

$$=\int K_h(z_1-y_1)\left[\frac{\bar{p}_2(z_2)}{p(z)} - \frac{p_2(z_2)}{p(z)}\right]\left[H_2(z_2)\otimes\left(1,\frac{z_1-y_1}{h}\right)^\top\right]p(z)\,dz$$

$$+\int K_h(z_1-y_1)\left[\frac{p_2(z_2)p(z)}{p^2(z)} - \frac{p_2(z_2)\bar{p}(z)}{p^2(z)}\right]$$

$$\times\left[H_2(z_2)\otimes\left(1,\frac{z_1-y_1}{h}\right)^\top\right]p(z)\,dz$$

$$\simeq\frac{g^2}{2}\left[\int D^2p_2(z_2)H_2(z_2)\,dz_2\otimes(\mu_K^2,0)^\top\right]$$

$$-\frac{g^2}{2}\left[\int\frac{p_2(z_2)}{p(y_1,z_2)}D^2p(y_1,z_2)H_2(z_2)\,dz_2\otimes(\mu_K^2,0)^\top\right].$$

Finally, for the probability limit of $[I_2\otimes e_1^\top Q_n^{-1}]$, we note that

$$Q_n = D_h^{-1}\mathbf{Y}_-^\top\mathbf{K}\mathbf{Y}_-D_h^{-1} = [\hat{q}_{ni+j-2}(y_1;h)]_{(i,j)=1,2}$$

with $\hat{q}_{ni} = (1/n)\sum_{k=d}^n K_h^{\hat{W}}(Y_{k-1}-y_1)((y_{k-1}-y_1)/h)^i$, for $i=0,1,2$, and

$$\hat{q}_{ni}\xrightarrow{p}\int K_h(z_1-y_1)\left(\frac{z_1-y_1}{h}\right)^i p_2(z_2)\,dz = \int K(u_1)u_1^i\,du_1\int p_2(z_2)\,dz_2$$

$$=\int K(u_1)u_1^i\,du_1\equiv q_i,$$

where $q_0=1$, $q_1=0$, and $q_2=\mu_K^2$.

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

Thus, $Q_n \to \begin{bmatrix} 1 & 0 \\ 0 & \mu_K^2 \end{bmatrix}$, $Q_n^{-1} \to (1/\mu_K^2)\begin{bmatrix} \mu_K^2 & 0 \\ 0 & 1 \end{bmatrix}$, and $e_1^\top Q_n^{-1} \to e_1^\top$. Therefore,

$$
\begin{aligned}
B_{1n}(y_1) &= [I_2 \otimes e_1^\top Q_n^{-1}](s_{0n} + s_{1n} + s_{2n}) \\
&= \frac{h^2}{2}\,\mu_K^2 D^2\varphi_1(y_1) \\
&\quad \times \frac{h^2}{2}\int [\mu_{K*K}^2 D^2\varphi_1(y_1) + \mu_K^2 D^2\varphi_2(z_2)] \\
&\quad \odot [\nabla G_m(m(y_1,z_2)), \nabla G_v(v(y_1,z_2))]^\top p_2(z_2)\,dz_2 \\
&\quad + \frac{g^2}{2}\,\mu_K^2 \int Dp_2(z_2)H_2(z_2)\,dz_2 - \frac{g^2}{2}\,\mu_K^2 \\
&\quad \times \int \frac{p_2(z_2)}{p(y_1,z_2)}\,D^2 p(y_1,z_2)H_2(z_2)\,dz_2 \\
&\quad + o_p(h^2) + o_p(g^2).
\end{aligned}
$$

Step III. Asymptotic Normality of $\tilde{t}_n^*$. Applying the Cramer–Wold device, it is sufficient to show

$$
D_n \equiv \frac{1}{\sqrt{n}} \sum_k \sqrt{h}\tilde{\xi}_k \xrightarrow{\mathcal{D}} N(0, \beta^\top \Sigma_1 \beta),
$$

for all $\beta \in \mathbb{R}^4$, where $\tilde{\xi}_k = \beta^\top \xi_k$. We use the small block–large block argument (see Masry and Tjøstheim, 1997). Partition the set $\{d, d+1, \ldots, n\}$ into $2k+1$ subsets with large blocks of size $r = r_n$ and small blocks of size $s = s_n$ where

$$
k = \left[\frac{n_1}{r_n + s_n}\right]
$$

and $[x]$ denotes the integer part of x. Define

$$
\eta_j = \sum_{t=j(r+s)}^{j(r+s)+r-1} \sqrt{h}\tilde{\xi}_t, \qquad \omega_j = \sum_{t=j(r+s)+r}^{(j+1)(r+s)-1} \sqrt{h}\tilde{\xi}_t, \qquad 0 \le j \le k-1,
$$

$$
\varsigma_k = \sum_{t=k(r+s)}^{n} \sqrt{h}\tilde{\xi}_t.
$$

Then,

$$
D_n = \frac{1}{\sqrt{n}}\left(\sum_{j=0}^{k-1}\eta_j + \sum_{j=0}^{k-1}\omega_j + \varsigma_k\right) \equiv \frac{1}{\sqrt{n}}\,(S_n' + S_n'' + S_n''').
$$

Because of Assumption A6, there exists a sequence $a_n \to \infty$ such that

$$
a_n s_n = o(\sqrt{nh}) \quad \text{and} \quad a_n\sqrt{n/h}\,\alpha(s_n) \to 0, \quad \text{as } n \to \infty, \tag{A.2}
$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

defining the large block size as

$$r_n = \left[\frac{\sqrt{nh}}{a_n} \right]. \tag{A.3}$$

It is easy to show by (A.2) and (A.3) that as $n \to \infty$

$$\frac{r_n}{n} \to 0, \qquad \frac{s_n}{r_n} \to 0, \qquad \frac{r_n}{\sqrt{nh}} \to 0, \tag{A.4}$$

and

$$\frac{n}{r_n} \alpha(s_n) \to 0.$$

We first show that S_n'' and S_n''' are asymptotically negligible. The same argument used in step II yields

$$\begin{aligned} \operatorname{var}(\omega_j) &= s \times \operatorname{var}(\sqrt{h}\tilde{\xi}_t) + 2s \sum_{k=1}^{s-1} \left(1 - \frac{k}{s}\right) \operatorname{cov}(\sqrt{h}\tilde{\xi}_{d+1}, \sqrt{h}\tilde{\xi}_{d+1+k}) \\ &= s\beta^\top \Sigma_1 \beta (1 + o(1)), \end{aligned} \tag{A.5}$$

which implies

$$\sum_{j=0}^{k-1} \operatorname{var}(\omega_j) = O(ks) \sim \frac{ns_n}{r_n + s_n} \sim \frac{ns_n}{r_n} = o(n),$$

from the condition (A.4). Next, consider

$$\sum_{\substack{i,j=0,\\ i \neq j}}^{k-1} \operatorname{cov}(\omega_i, \omega_j) = \sum_{\substack{i,j=0,\\ i \neq j}}^{k-1} \sum_{k_1=1}^{s} \sum_{k_2=1}^{s} \operatorname{cov}(\sqrt{h}\tilde{\xi}_{N_i+k_1}, \sqrt{h}\tilde{\xi}_{N_j+k_2}),$$

where $N_j = j(r+s) + r$. Because $|N_i - N_j + k_1 - k_2| \geq r$, for $i \neq j$, the covariance term is bounded by

$$\begin{aligned} &2 \sum_{k_1=1}^{n-r} \sum_{k_2=k_1+r}^{n} |\operatorname{cov}(\sqrt{h}\tilde{\xi}_{k_1}, \sqrt{h}\tilde{\xi}_{k_2})| \\ &\leq 2n \sum_{j=r+1}^{n} |\operatorname{cov}(\sqrt{h}\tilde{\xi}_{d+1}, \sqrt{h}\tilde{\xi}_{d+1+j})| = o(n). \end{aligned}$$

The last equality also follows from step II. Hence, $(1/n)E\{(S_n'')^2\} \to 0$, as $n \to \infty$. Repeating a similar argument for S_n''', we get

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

$$\frac{1}{n}E\{(S_n''')^2\} \le \frac{1}{n}[n-k(r+s)]\mathrm{var}(\sqrt{h}\tilde{\xi}_{d+1}) + 2\frac{n-k(r+s)}{n}\sum_{j=1}^{n-k(r+s)}\mathrm{cov}(\sqrt{h}\tilde{\xi}_{d+1},\sqrt{h}\tilde{\xi}_{d+1+j})$$

$$\le \frac{r_n+s_n}{n}\beta^\top\Sigma_1\beta + o(1)$$

$$\to 0, \quad \text{as } n\to\infty.$$

Now, it remains to show $(1/\sqrt{n})S_n' = (1/\sqrt{n})\sum_{j=0}^{k-1}\eta_j \xrightarrow{\mathcal{D}} N(0,\beta^\top\Sigma_1\beta)$.

Because η_j is a function of $\{\tilde{\xi}_t\}_{t=j(r+s)+1}^{j(r+s)+r-1}$ that is $\mathcal{F}_{j(r+s)+1-d}^{j(r+s)+r-1}$-measurable, the Volkonskii and Rozanov's lemma (1959) in the appendix of Masry and Tjøstheim (1997) implies that, with $\tilde{s}_n = s_n - d + 1$,

$$\left|E\left[\exp\left(it\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j\right)\right] - \prod_{j=0}^{k-1}E(\exp(it\eta_j))\right| \le 16k\alpha(\tilde{s}_n - d + 1) \simeq \frac{n}{r_n+s_n}\alpha(\tilde{s}_n) \simeq \frac{n}{r_n}\alpha(\tilde{s}_n) \simeq o(1),$$

where the last two equalities follow from (A.4). Thus, the summands $\{\eta_j\}$ in S_n' are asymptotically independent, because an operation similar to (A.5) yields

$$\mathrm{var}(\eta_j) = r_n\beta^\top\Sigma_1\beta(1+o(1))$$

and hence

$$\mathrm{var}\left(\frac{1}{\sqrt{n}}S_n'\right) = \frac{1}{n}\sum_{j=0}^{k-1}E(\eta_j^2) = \frac{k_n r_n}{n}\beta^\top\Sigma_1\beta(1+o(1)) \to \beta^\top\Sigma^*\beta.$$

Finally, because of the boundedness of density and kernel functions, the Lindeberg–Feller condition for the asymptotic normality of S_n' holds:

$$\frac{1}{n}\sum_{j=0}^{k-1}E[\eta_j^2 I\{|\eta_j| > \sqrt{n}\delta\sqrt{\beta^\top\Sigma_1\beta}\}] \to 0$$

for every $\delta > 0$. This completes the proof of step III.

From $e_1^\top Q_n^{-1} \xrightarrow{p} e_1^\top$, the Slutzky theorem implies $\sqrt{nh}[I_2 \otimes e_1^\top Q_n^{-1}]\tilde{t}_n^* \xrightarrow{d} N(0,\Sigma_1^*)$, where $\Sigma_1^* = [I_2 \otimes e_1^\top]\Sigma_1[I_2 \otimes e_1]$. In sum, $\sqrt{nh}(\hat{\varphi}_1(y_1) - \varphi_1(y_1) - B_n) \xrightarrow{d} N(0,\Sigma_1^*)$, with $\Sigma_1^*(y_1)$ given by

$$\int \frac{p_2^2(z_2)}{p(y_1,z_2)}\|(K*K)_0\|_2^2 \times \begin{bmatrix} \nabla G_m(y_1,z_2)^2 v(y_1,z_2) & (\nabla G_m\cdot\nabla G_v)(\kappa_3\cdot v^{3/2})(y_1,z_2) \\ (\nabla G_m\cdot\nabla G_v)(\kappa_3\cdot v^{3/2})(y_1,z_2) & \nabla G_v(y_1,z_2)^2\kappa_4(y_1,z_2)v^2(y_1,z_2) \end{bmatrix} dz_2 + \int \frac{p_2^2(z_2)}{p(y_1,z_2)}\|K\|_2^2 H_2(z_2)H_2^\top(z_2)\,dz_2. \quad \blacksquare$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

LEMMA A.1. *Assume the conditions in Assumptions A1 and A4–A6. For a bounded function, $F(\cdot)$, it holds that*

(a) $r_{1n} = (\sqrt{h}/\sqrt{n})\sum_{k=d}^{n} K_h(y_{k-1} - y_1)(\hat{p}_2(\underline{y}_{k-2}) - \bar{p}_2(\underline{y}_{k-2}))F(x_k) = o_p(1)$,
(b) $r_{2n} = (\sqrt{h}/\sqrt{n})\sum_{k=d}^{n} K_h(y_{k-1} - y_1)(\hat{p}(x_k) - p(x_k))F(x_k) = o_p(1)$.

Proof. The proof of (b) is almost the same as (a). Therefore we only show (a). By adding and subtracting $\bar{L}_{l|k}(y_{l-2}|y_{k-2})$, the conditional expectation of $L_g(\underline{y}_{l-2} - \underline{y}_{k-2})$ given $\underline{y}_{k-2}$ in r_{1n}, we get $r_{1n} = \xi_{1n} + \xi_{2n}$, where

$$\xi_{1n} = \frac{1}{n^2}\sum_{k=d}^{n}\sum_{l=d}^{n} K_h(y_{k-1} - y_1)F(x_k)[L_g(\underline{y}_{l-2} - \underline{y}_{k-2}) - \bar{L}_{l|k}(y_{l-2}|y_{k-2})],$$

$$\xi_{2n} = \frac{1}{n^2}\sum_{k}\sum_{l} K_h(y_{k-1} - y_1)F(x_k)[\bar{L}_{l|k}(y_{l-2}|y_{k-2}) - \bar{p}_2(\underline{y}_{k-2})].$$

Rewrite ξ_{2n} as

$$\frac{1}{n^2}\sum_{k}\sum_{s<k^*(n)} K_h(y_{k-1} - y_1)F(x_k)[\bar{L}_{k+s|k}(y_{k+s-2}|y_{k-2}) - \bar{p}_2(\underline{y}_{k-2})],$$

$$+ \frac{1}{n^2}\sum_{k}\sum_{s\geq k^*(n)} K_h(y_{k-1} - y_1)F(x_k)[\bar{L}_{k+s|k}(y_{k+s-2}|y_{k-2}) - \bar{p}_2(\underline{y}_{k-2})],$$

where $k^*(n)$ is increasing to infinity as $n \to \infty$. Let

$$B = E\{K_h(y_{k-1} - y_1)F(x_k)[\bar{L}_{k+s|k}(y_{k+s-2}|y_{k-2}) - \bar{p}_2(\underline{y}_{k-2})]\},$$

which exists as a result of the boundedness of $F(x_k)$. Then, for a large n, the first part of ξ_{2n} is asymptotically equivalent to $(1/n)k^*(n)B$. The second part of ξ_{2n} is bounded by

$$\sup_{s\geq k^*(n)} |p_{k+s|k}(y_{k+s-2}|y_{k-2}) - p(y_{k-2})| \frac{1}{n}\sum_{k}^{n} K_h(y_{k-1} - y_1)|F(x_k)|$$

$$\leq \rho^{k(n)}O_p(1).$$

Therefore, $\sqrt{nh}\xi_{2n} \leq O_p((\sqrt{h}/\sqrt{n})k^*(n)) + O_p(\rho^{-k^*(n)}\sqrt{nh}) = o_p(1)$, for $k(n) = \log n$, for example.

It remains to show $\xi_{1n} = o_p(1/\sqrt{nh})$. Because $E(\xi_{1n}) = 0$ from the law of iteration, we just compute

$$E(\xi_{1n}^2) = \frac{1}{n^4}\sum_{k\neq l}^{n}\sum^{n}\sum_{i\neq j}^{n}\sum^{n} E\{K_h(y_{k-1} - y)K_h(y_{i-1} - y)F(x_k)$$

$$\times F(X_l)[L_g(\underline{y}_{l-2} - \underline{y}_{k-2}) - \bar{L}_{l|k}(\underline{y}_{k-2})]$$

$$\times [L_h(\underline{y}_{j-2} - \underline{y}_{i-2}) - \bar{L}_{j|i}(\underline{y}_{i-2})]\}.$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

(1) Consider the case $k = i$ and $l \neq j$.

$$\frac{1}{n^4}\sum_{k}^{n}\sum_{l\neq j}^{n}\sum^{n} E\{K_h^2(y_{k-1}-y)F^2(x_k) \times [L_g(\underline{y}_{l-2}-\underline{y}_{k-2})-\bar{L}_{l|k}(\underline{y}_{k-2})][L_g(\underline{y}_{j-2}-\underline{y}_{k-2})-\bar{L}_{j|k}(\underline{y}_{k-2})]\} = 0,$$

because, by the law of iteration and the definition of $\bar{L}_{j|k}(\underline{y}_{k-2})$,

$$E_{|k,l}[L_g(\underline{y}_{j-2}-\underline{y}_{k-2})-\bar{L}_{j|k}(\underline{y}_{k-2})] = E_{|k}[L_g(\underline{y}_{j-2}-\underline{y}_{k-2})-\bar{L}_{j|k}(\underline{y}_{k-2})] = E_{|k}[L_g(\underline{y}_{j-2}-\underline{y}_{k-2})]-\bar{L}_{j|k}(\underline{y}_{k-2}) = 0.$$

(2) Consider the case $l = j$ and $k \neq i$.

$$\frac{1}{n^4}\sum_{k\neq i}^{n}\sum^{n}\sum_{l}^{n} E\{K_h(y_{k-1}-y)K_h(y_{i-1}-y)F(x_k)F(x_i) \times [L_g(\underline{y}_{l-2}-\underline{y}_{k-2})-\bar{L}_{l|k}(\underline{y}_{k-2})][L_g(\underline{y}_{l-2}-\underline{y}_{i-2})-\bar{L}_{l|i}(\underline{y}_{i-2})]\}.$$

We only calculate

$$\frac{1}{n^4}\sum_{k\neq i}^{n}\sum^{n}\sum_{l}^{n} E\{K_h(y_{k-1}-y)K_h(y_{i-1}-y)L_g(\underline{y}_{l-2}-\underline{y}_{k-2}) \times L_g(\underline{y}_{l-2}-\underline{y}_{i-2})F(x_k)F(x_i)\} \tag{A.6}$$

because the rest of the triple sum consists of expectations of standard kernel estimates and is $O(1/n)$. Note that

$$E_{|(i,k)}L_g(\underline{y}_{l-2}-\underline{y}_{k-2})L_g(\underline{y}_{l-2}-\underline{y}_{i-2}) \simeq (L*L)_g(\underline{y}_{k-2}-\underline{y}_{i-2})p_{l|(k,i)}(\underline{y}_{k-2}|\underline{y}_{k-2},\underline{y}_{i-2}),$$

where $(L*L)_g(\cdot) = (1/g)\int L(u)L(u+\cdot/g)$ is a convolution kernel. Thus, (A.6) is

$$\frac{1}{n^4}\sum_{k\neq i}^{n}\sum^{n}\sum_{l}^{n} E[K_h(y_{k-1}-y)K_h(y_{i-1}-y)(L*L)_g(\underline{y}_{k-2}-\underline{y}_{i-2}) \times F(x_k)F(x_i)p_{l|(k,i)}(\underline{y}_{k-2}|\underline{y}_{k-2},\underline{y}_{i-2})] = O\left(\frac{1}{n}\right).$$

(3) Consider the case with $i = k$, $j = m$:

$$\frac{1}{n^4}\sum_{k\neq l}^{n}\sum^{n} E\{K_h^2(y_{k-1}-y)F^2(x_k)[L_g(y_{l-2}-y_{k-2})-\bar{L}_{l|k}(\underline{y}_{k-2})]^2\} = O\left(\frac{1}{n^2hg}\right) = o\left(\frac{1}{nh}\right).$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

(4) Consider the case $k \neq i$, $l \neq j$:

$$\frac{1}{n^4}\sum_{k\neq l}^{n}\sum^{n}\sum_{i\neq j}^{n}\sum^{n} E\{K_h(y_{k-1}-y)K_h(y_{i-1}-y)F(x_k)F(x_i) \times [L_g(\underline{y}_{l-2}-\underline{y}_{k-2})-\bar{L}_{l|k}(\underline{y}_{k-2})][L_g(\underline{y}_{j-2}-\underline{y}_{i-2})-\bar{L}_{j|i}(\underline{y}_{i-2})]\} = 0,$$

for the same reason as in (1). ■

A.2. Proofs for Section 5. Recall that $x_t = (y_{t-1},\ldots,y_{t-d}) = (y_{t-\alpha},\underline{y}_{t-\alpha})$ and $z_t = (x_t, y_t)$. In a similar context, let $x = (y_1,..,y_d) = (y_\alpha,\underline{y}_\alpha)$ and $z = (x, y_0)$. For the score function $s^*(z,\theta,\gamma_\alpha) = s^*(z,\theta,\gamma_\alpha(\underline{y}_\alpha))$, we define its first derivative with respect to the parameter θ by

$$\nabla_\theta s^*(z,\theta,\gamma_\alpha) = \frac{\partial s^*(z,\theta,\gamma_\alpha)}{\partial\theta}$$

and use $\overline{s^*}(\theta,\gamma_\alpha)$ and $\overline{\nabla_\theta s^*}(\theta,\gamma_\alpha)$ to denote $E[s^*(z_t,\theta,\gamma_\alpha)]$ and $E[\nabla_\theta s^*(z_t,\theta,\gamma_\alpha)]$, respectively. Also, the score function $s^*(z,\theta,\cdot)$ is said to be Frechet differentiable (with respect to the sup norm $\|\cdot\|_\infty$) if there is $S^*(z,\theta,\gamma_\alpha)$ such that for all γ_α with $\|\gamma_\alpha - \gamma_\alpha^0\|_\infty$ small enough,

$$\|s^*(z,\theta,\gamma_\alpha) - s^*(z,\theta,\gamma_\alpha^0) - S^*(z,\theta,\gamma_\alpha^0(\underline{y}_\alpha))(\gamma_\alpha-\gamma_\alpha^0)\| \leq b(z)\|\gamma_\alpha-\gamma_\alpha^0\|^2, \qquad \textbf{(A.7)}$$

for some bounded function $b(\cdot)$. The term $S^*(z,\theta,\gamma_\alpha^0)$ is called the functional derivative of $s^*(z,\theta,\gamma_\alpha)$ with respect to γ_α. In a similar way, we define $\nabla_{\boldsymbol{\gamma}} S^*(z,\theta,\gamma_\alpha)$ to be the functional derivative of $S^*(z,\theta,\gamma_\alpha)$ with respect to γ_α.

Assumption B. Suppose that (i) $\overline{\nabla_\theta s^*}(\theta_0)$ is nonsingular; (ii) $S^*(z,\theta,\gamma_\alpha(\underline{y}_\alpha))$ and $\nabla_{\boldsymbol{\gamma}} S^*(z,\theta,\gamma_\alpha(\underline{y}_\alpha))$ exist and have square integrable envelopes $\bar{S}^*(\cdot)$ and $\bar{\nabla}_{\boldsymbol{\gamma}} S^*(\cdot)$, satisfying

$$\|S^*(z,\theta,\gamma_\alpha(\underline{y}_\alpha))\| \leq \bar{S}^*(z), \qquad \|\nabla_{\boldsymbol{\gamma}} S^*(z,\theta,\gamma_\alpha(\underline{y}_\alpha))\| \leq \bar{\nabla}_{\boldsymbol{\gamma}} S^*(z);$$

and (iii) both $s^*(z,\theta,\gamma_\alpha)$ and $S^*(z,\theta,\gamma_\alpha)$ are continuously differentiable in θ, with derivatives bounded by square integrable envelopes.

Note that the first condition is related to the identification condition of component functions, whereas the second concerns Frechet differentiability (up to the second order) of the score function and uniform boundedness of the functional derivatives. For the main results in Section 5, we need the following conditions. Some of the assumptions are stronger than their counterparts in Assumption A in Section 4. Let h_0 and h denote the bandwidth parameter used for the preliminary instrumental variable and the two-step estimates, respectively, and g denote the bandwidth parameter for the kernel density.

Assumption C.

1. $\{y_t\}_{t=1}^{\infty}$ is stationary and strongly mixing with a mixing coefficient $\alpha(k) = \rho^{-\beta k}$, for some $\beta > 0$, and $E(\varepsilon_t^4 x_t) < \infty$, where $\varepsilon_t = y_t - E(y_t|x_t)$.
2. The joint density function, $p(\cdot)$, is bounded away from zero and q-times continuously differentiable on the compact supports $\mathcal{X} = \mathcal{X}_\alpha \times \mathcal{X}_{\bar{\alpha}}$, with Lipschitz

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

continuous remainders, that is, there exists $C < \infty$ such that for all $x, x' \in \mathcal{X}$, $|D_x^\mu p(x) - D_x^\mu p(x')| \leq C\|x - x'\|$, for all vectors $\mu = (\mu_1, \ldots, \mu_d)$ with $\sum_{i=1}^d \mu_i \leq q$.

3. The component functions, $m_\alpha(\cdot)$ and $v_\alpha(\cdot)$, for $\alpha = 1, \ldots, d$, are q-times continuously differentiable on $\mathcal{X}_\alpha$ with Lipschitz continuous qth derivative.
4. The link functions, G_m and G_v, are q-times continuously differentiable over any compact interval of the real line.
5. The kernel functions, $K(\cdot)$ and $L(\cdot)$, are of bounded support, symmetric about zero, satisfying $\int K(u)\,du = \int L(u)\,du = 1$, and of order q, that is, $\int u^i K(u)\,du = \int u^i L(u)\,du = 0$, for $i = 1, \ldots, q-1$. Also, the kernel functions are q-times differentiable with Lipschitz continuous qth derivative.
6. The true parameters $\theta_0 = (m_\alpha(y_\alpha), v_\alpha(y_\alpha), m'_\alpha(y_\alpha), v'_\alpha(y_\alpha))$ lie in the interior of the compact parameter space Θ.
7. (i) $g \to 0$, $ng^d \to \infty$ and (ii) $h_0 \to 0$, $nh_0 \to \infty$.
8. (i) $nh_0^2/(\log n)^2 h \to \infty$ and $\sqrt{nh}\,h_0^q \to 0$; and for some integer $\omega > d/2$,
   (ii) $n(h_0 h)^{2\omega+1}/\log n \to \infty$; $h_0^{q-\omega} h^{-\omega-1/2} \to 0$;
   (iii) $nh_0^{d+(4\omega+1)}/\log n \to \infty$; $q \geq 2\omega + 1$.

Some facts about empirical processes will be useful in the discussion that follows. Define the L^2-Sobolev norm (of order q) on the class of real-valued function with domain $\mathcal{W}_0$:

$$\|\tau\|_{q,2,\mathcal{W}_0} = \left(\sum_{\mu \leq q} \int_{\mathcal{W}_0} (D_x^\mu \tau(x))^2\,dx \right)^{1/2},$$

where, for $x \in \mathcal{W}_0 \subset R^k$ and a k-vector $\mu = (\mu_1, \ldots, \mu_k)$ of nonnegative integers,

$$D^\mu \tau(x) = \frac{\partial^{\Sigma_{i=1}^k \mu_i} \tau(x)}{\partial^{\mu_1} x_1 \cdots \partial^{\mu_k} x_k}$$

and $q \geq 1$ is some positive integer. Let $\mathcal{X}_\alpha$ be an open set in $\mathbb{R}^1$ with minimally smooth boundary as defined by, for example, Stein (1970), and $\mathcal{X} = \times_{\beta=1}^d \mathcal{X}_\beta$, with $\mathcal{X}_{\bar{\alpha}} = \times_{\beta=1(\neq\alpha)}^d \mathcal{X}_\beta$. Define $\mathcal{T}_1$ as a class of smooth functions on $\mathcal{X}_{\bar{\alpha}} = \times_{\beta=1(\neq\alpha)}^d \mathcal{X}_\beta$ whose L^2-Sobolev norm is bounded by some constant $\mathcal{T}_1 = \{\tau : \|\tau\|_{q,2,\mathcal{X}_{\bar{\alpha}}} \leq C\}$. In a similar way, $\mathcal{T}_2 = \{\tau : \|\tau\|_{q,2,\mathcal{X}} \leq C\}$.

Define (i) an empirical process, $v_{1n}(\cdot)$, indexed by $\tau \in \mathcal{T}_1$:

$$v_{1n}(\tau_1) = \frac{1}{\sqrt{n}} \sum_{t=1}^n [\, f_1(x_t; \tau_1) - E f_1(x_t; \tau_1)], \tag{A.8}$$

with pseudometric $\rho_1(\cdot,\cdot)$ on $\mathcal{T}_1$:

$$\rho_1(\tau, \tau') = \left[\int_{\mathcal{X}} (\, f_1(w; \tau(\underline{w}_\alpha)) - f_1(w; \tau'(\underline{w}_\alpha)))^2 p(w)\,dw \right]^{1/2},$$

where $f_1(w;\tau) = h^{-1/2} K((w_\alpha - y_\alpha)/h) \underline{S}^*(w, \gamma_\alpha^0) \tau_1(\underline{w}_\alpha)$; and (ii) an empirical process, $v_{2n}(\cdot,\cdot)$, indexed by $(y_\alpha, \tau_2) \in \mathcal{X}_\alpha \times \mathcal{T}_2$:

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

$$v_{2n}(y_\alpha,\tau_2) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}[\, f_2(x_t;y_\alpha,\tau_2) - Ef_2(x_t;y_\alpha,\tau_2)], \tag{A.9}$$

with pseudometric $\rho_2(\cdot,\cdot)$ on $\mathcal{T}_2$:

$$\rho_2((y_\alpha,\tau_2),(y'_\alpha,\tau'_2)) = \left[\int_{\mathcal{X}} (\, f_2(w;y_\alpha,\tau_2) - f_2(w;y'_\alpha,\tau'_2))^2 p(w)\, dw\right]^{1/2},$$

where $f_2(w;y_\alpha,\tau_2) = h_0^{-1/2}K[(w_\alpha - y_\alpha)/h_0][\, p_{\bar{\alpha}}(\underline{w}_\alpha)/p(w)]G'_m(m(w))\tau_2(w)$.

We say that the processes $\{\nu_{1n}(\cdot)\}$ and $\{\nu_{2n}(\cdot,\cdot)\}$ are stochastically equicontinuous at τ_1^0 and (y_α^0,τ_2^0), respectively (with respect to the pseudometric $\rho_1(\cdot,\cdot)$ and $\rho_2(\cdot,\cdot)$, respectively), if

$$\forall\, \varepsilon,\eta > 0, \qquad \exists\, \delta > 0 \; s.t.$$

$$\overline{\lim_{T\to\infty}}\, P^*\left[\sup_{\rho_1(\tau,\tau_0)<\delta} |\nu_{1n}(\tau_1) - \nu_{1n}(\tau_1^0)| > \eta\right] < \varepsilon, \tag{A.10}$$

and

$$\overline{\lim_{T\to\infty}}\, P^*\left[\sup_{\rho_2((y_\alpha,\tau_2),(y_\alpha^0,\tau_2^0))<\delta} |\nu_{2n}(y_\alpha,\tau_2) - \nu_{2n}(y_\alpha^0,\tau_2^0)| > \eta\right] < \varepsilon, \tag{A.11}$$

respectively, where P^* denotes the outer measure of the corresponding probability measure.

Let $\mathcal{F}_1$ be the class of functions such as $f_1(\cdot)$ defined previously. Note that (A.10) follows, if Pollard's entropy condition is satisfied by $\mathcal{F}_1$ with some square integrable envelope $\bar{F}_1$; see Pollard (1990) for more details. Because $f_1(w;\tau_1) = c_1(w)\tau_1(\underline{w}_\alpha)$ is the product of smooth functions τ_1 from an infinite-dimensional class (with uniformly bounded partial derivatives up to order q) and a single unbounded function $c(w) = [h^{-1/2}K((w_\alpha - y_\alpha)/h)\underline{S}^*(w,\gamma_\alpha^0)]$, the entropy condition is verified by Theorem 2 in Andrews (1994) on a class of functions of type III. Square integrability of the envelope $\bar{F}_1$ comes from Assumption B(ii). In a similar way, we can show (A.11) by applying the "mix and match" argument of Theorem 3 in Andrews (1994) to $f_2(w;y_\alpha,\tau_2) = c_2(w)h^{-1/2}K((w_\alpha - y_\alpha)/h_0)\tau_2(w)$, where $K(\cdot)$ is Lipschitz continuous in y_α, that is, a function of type II.

Proof of Theorem 4. We only give a sketch, because the whole proof is lengthy and relies on arguments similar to Andrews (1994) or Gozalo and Linton (2000) for the i.i.d. case. Expanding the first-order condition in (5.16) and solving for $(\hat{\theta}^* - \theta_0)$ yields

$$\hat{\theta}^* - \theta_0 = -\left[\frac{1}{n}\sum_{t=d+1}^{n'} \nabla_\theta s(z_t,\bar{\theta},\tilde{\gamma}_\alpha)\right]^{-1}\frac{1}{n}\sum_{t=d+1}^{n'} s(z_t,\tilde{\gamma}_\alpha),$$

where $\bar{\theta}$ is the mean value between $\hat{\theta}$ and θ_0 and $s(z_t,\tilde{\gamma}_\alpha) = s(z_t,\theta_0,\tilde{\gamma}_\alpha)$. By the uniform law of large numbers in Gozalo and Linton (1995), we have $\sup_{\theta\in\Theta}|Q_n(\theta) - E(Q_n(\theta))| \xrightarrow{p} 0$, which, together with (i) uniform convergence of $\tilde{\gamma}_\alpha$ by Lemma A.3 and (ii) uniform continuity of the localized likelihood function, $Q_n(\theta,\gamma_\alpha)$ over $\Theta\times\Gamma_\alpha$, yields $\sup_{\theta\in\Theta}|\tilde{Q}_n(\theta) - E(Q_n(\theta))| \xrightarrow{p} 0$ and thus consistency of $\hat{\theta}^*$. Based on the ergodic theo-

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

rem on the stationary time series and a similar argument to Theorem 1 in Andrews (1994), consistency of $\hat{\theta}^*$ and uniform convergence of $\tilde{\gamma}_\alpha$ imply

$$\frac{1}{n}\sum_{t=d+1}^{n'} \nabla_\theta s(z_t, \bar{\theta}, \tilde{\gamma}_\alpha) \xrightarrow{p} E[\nabla_\theta s(z_t, \theta_0, \gamma_\alpha^0)] \equiv D_\alpha(y_\alpha). \tag{A.12}$$

For the numerator, we first linearize the score function. Under Assumption B(ii), $s^*(z, \theta, \gamma_\alpha)$ is Frechet differentiable and (A.7) holds, which, because of $\sqrt{nh}\|\tilde{\gamma}_\alpha - \gamma_\alpha^0\|_\infty^2 \xrightarrow{p} 0$ (by Lemma A.3 and Assumption C.8(i)), yields a proper linearization of the score term:

$$\begin{aligned}\frac{1}{n}\sum_{t=d+1}^{n'} s(z_t, \tilde{\gamma}_\alpha) = {} & \frac{1}{n}\sum_{t=d+1}^{n'} K_h(y_{t-\alpha} - y_\alpha^0) s^*(z_t, \gamma_\alpha^0) \\ & + \frac{1}{n}\sum_{t=d+1}^{n'} K_h(y_{t-\alpha} - y_\alpha^0) S^*(z_t, \gamma_\alpha^0(\underline{y}_{t-\alpha}))[\tilde{\gamma}_\alpha(\underline{y}_{t-\alpha}) - \gamma_\alpha^0(\underline{y}_{t-\alpha})] \\ & + o_p(1/\sqrt{nh}),\end{aligned}$$

where $S^*(z_t, \gamma_\alpha^0(\underline{y}_{t-\alpha})) = S^*(z_t, \theta_0, \gamma_\alpha^0(\underline{y}_{t-\alpha}))$. Or equivalently, by letting

$$\underline{S}^*(y, \gamma_\alpha^0(\underline{y}_\alpha)) = E[S^*(z_t, \gamma_\alpha^0(\underline{y}_{t-\alpha}))x_t = y]$$

and $u_t = \underline{S}^*(x_t, \gamma_\alpha^0(\underline{y}_{t-\alpha})) - E[\underline{S}^*(x_t, \gamma_\alpha^0(\underline{y}_{t-\alpha}))|x_t = y]$, we have

$$\begin{aligned}&\frac{\sqrt{nh}}{n}\sum_{t=d+1}^{n'} s(z_t, \tilde{\gamma}_\alpha) \\ &\quad = \frac{\sqrt{h}}{\sqrt{n}}\sum_{t=d+1}^{n'} K_h(y_{t-\alpha} - y_\alpha^0) s^*(z_t, \gamma_\alpha^0) \\ &\qquad + \frac{\sqrt{h}}{\sqrt{n}}\sum_{t=d+1}^{n'} K_h(y_{t-\alpha} - y_\alpha^0) \underline{S}^*(x_t, \gamma_\alpha^0(\underline{y}_{t-\alpha}))[\tilde{\gamma}_\alpha(\underline{y}_{t-\alpha}) - \gamma_\alpha^0(\underline{y}_{t-\alpha})] \\ &\qquad + \frac{\sqrt{h}}{\sqrt{n}}\sum_{t=d+1}^{n'} K_h(y_{t-\alpha} - y_\alpha^0) u_t[\tilde{\gamma}_\alpha(\underline{y}_{t-\alpha}) - \gamma_\alpha^0(\underline{y}_{t-\alpha})] + o_p(1) \\ &\quad \equiv T_{1n} + T_{2n} + T_{3n} + o_p(1).\end{aligned}$$

Note that the asymptotic expansion of the infeasible estimator is equivalent to the first term of the linearized score function premultiplied by the inverse Hessian matrix in (A.12). Because of the asymptotic boundedness of (A.12), it suffices to show the negligibility of the second and third terms.

To calculate the asymptotic order of T_{2n}, we make use of the preceding stochastic equicontinuity results. For a real-valued function $\delta(\cdot)$ on $\mathcal{X}_{\bar{\alpha}}$ and $\mathcal{T} = \{\delta : \|\delta\|_{\omega,2,\mathcal{X}_{\bar{\alpha}}} \leq C\}$, we define an empirical process

$$v_n(y_\alpha, \delta) = \frac{1}{\sqrt{n}}\sum_{t=d+1}^{n'} [\, f(x_t; y_\alpha, \delta) - E(\, f(x_t; y_\alpha, \delta))],$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

where $f(x_t;y_\alpha,\delta) = K((y_{t-\alpha} - y_\alpha)/h)h^\omega \underline{S}^*(x_t,\gamma_\alpha^0(\underline{y}_{t-\alpha}))\delta(\underline{y}_{t-\alpha})$, for some integer $\omega > d/2$. Let $\tilde{\delta} = h^{-\omega-1/2}[\tilde{\gamma}_\alpha(\underline{y}_{t-\alpha}) - \gamma_\alpha^0(\underline{y}_{t-\alpha})]$. From the uniform convergence rate in Lemma A.3 and the bandwidth condition C.8(ii), it follows that

$$\|\tilde{\delta}\|_{\omega,2,\mathcal{X}_{\bar{\alpha}}} = O_p\left(h^{-\omega-1/2}\left[\sqrt{\frac{\log n}{nh_0^{(2\omega+1)}}} + h_0^{q-\omega}\right]\right) = o_p(1).$$

Because $\hat{\delta}$ is bounded uniformly over $\mathcal{X}_{\bar{\alpha}}$, with probability approaching one, it holds that $\Pr(\hat{\delta} \in \mathcal{T}) \to 1$. Also, because, for some positive constant $C < \infty$,

$$\rho^2((y_\alpha,\hat{\delta}),(y_\alpha,0)) \le Ch^{-(2\omega+1)}\|\tilde{\gamma}_\alpha - \gamma_\alpha^0\|^2_{\omega,2,\mathcal{X}_{\bar{\alpha}}} = o_p(1),$$

we have $\rho((y_\alpha,\hat{\delta}),(y_\alpha,0)) \xrightarrow{p} 0$. Hence, following Andrews (1994, p. 2257), the stochastic equicontinuity condition of $v_n(y_\alpha,\cdot)$ at $\delta^0 = 0$ implies that $|v_n(y_\alpha,\hat{\delta}) - \nu_n(y_\alpha,\delta^0)| = |v_n(y_\alpha,\hat{\delta})| = o_p(1)$; that is, T_{2n} is approximated (with an $o_p(1)$ error) by

$$T_{2n}^* = \sqrt{nh}\int K_h(y_\alpha - y_\alpha^0)\underline{S}^*(x,\gamma_\alpha^0)[\tilde{\gamma}_\alpha(\underline{y}_\alpha) - \gamma_\alpha^0(\underline{y}_\alpha)]p(x)\,dx.$$

We proceed to show negligibility of T_{n2}^*. From the integrability condition on $S^*(z,\gamma_\alpha^0(\underline{y}_\alpha))$, it follows, by change of variables and the dominated convergence theorem, that $\int K_h(y_\alpha - y_\alpha^0)S^*(z,\gamma_\alpha^0(\underline{y}_\alpha))\,dF_0(z) = \int S^*[(y,y_\alpha^0,\underline{y}_\alpha),\gamma_\alpha^0(\underline{y}_\alpha)] \times p(y,y_\alpha^0,\underline{y}_\alpha)\,d(y,\underline{y}_\alpha) < \infty$, which, together with $\sqrt{n}$-consistency of $\hat{c} = (\hat{c}_m,\hat{c}_v)^\top$, means that $(\hat{c} - c)\sqrt{nh}\int K_h(y_\alpha - y_\alpha^0)S^*(z,\gamma_\alpha^0(\underline{y}_\alpha))\,dF_0(z) = o_p(1)$. Because

$$\tilde{\gamma}_\alpha(\underline{y}_\alpha) - \gamma_\alpha^0(\underline{y}_\alpha) = \sum_{\beta=1,\neq\alpha}^{d}(\hat{\varphi}_\beta(y_\beta) - \varphi_\beta^0(y_\beta)) - (d-2)(\hat{c} - c),$$

this yields

$$T_{2n}^* = \sum_{\beta=1,\neq\alpha}^{d}\sqrt{nh}\int K_h(y_\alpha - y_\alpha^0)\underline{S}^*(x,\gamma_\alpha^0)(\hat{\varphi}_\beta(y_\beta) - \varphi_\beta^0(y_\beta))p(x)\,d(x) + o_p(1).$$

From Lemma A.3,

$$\begin{aligned}\hat{\varphi}_\beta(y_\beta) - \varphi_\beta(y_\beta) = {} & h_0^q\bar{b}_\beta(y_\beta) + \frac{1}{n}\sum_t (K*K)_{h_0}(y_{t-\beta} - y_\beta)\frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} \\ & \times (\nabla G(x_\beta,\underline{y}_{t-\beta}) \odot \xi_t) \\ & + \frac{1}{n}\sum_t K_{h_0}(y_{t-\beta} - y_\beta)\frac{p_2(\underline{y}_{t-\beta})}{p(x_t)}\gamma_\alpha^{*0}(\underline{y}_{t-\beta}) + O_p(\rho_n^2) + o_p(n^{-1/2}),\end{aligned}$$

where $\xi_t = (\varepsilon_t,(\varepsilon_t^2 - 1))^\top$, $\nabla G(x_t) = [\nabla G_m(y_\beta,\underline{y}_{t-\beta})v(x_t)^{1/2}, \nabla G_v(y_\beta,\underline{y}_{t-\beta})v(x_t)]^\top$, and $\gamma_\alpha^{*0}(\underline{y}_{t-\beta}) = \gamma_\alpha^0(\underline{y}_{t-\beta}) - c_0$. Under the condition C.8(i), $\sqrt{nh}h_0^q = o(1)$, integrability of the bias function $\bar{b}_\beta(y_\beta)$ and $S^*(z,\theta_0,\gamma_\alpha^0(\underline{y}_\alpha))$ imply

$$T_{2n}^* = \mathcal{S}_{1n} + \mathcal{S}_{2n} + o_p(1),$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

where

$$\mathcal{S}_{1n} = \sum_{\beta=1,\neq\alpha}^{d} \sqrt{nh} \int K_h(y_\alpha - y_\alpha^0) \underline{S}^*(x, \gamma_\alpha^0) \frac{1}{n}$$

$$\times \sum_t (K * K)_{h_0}(y_{t-\beta} - y_\beta) \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} (\nabla G(x_\beta, \underline{y}_{t-\beta}) \odot \xi_t) p(x)\, dx,$$

and

$$\mathcal{S}_{2n} = \sum_{\beta=1,\neq\alpha}^{d} \sqrt{nh} \int K_h(y_\alpha - y_\alpha^0) \underline{S}^*(x, \gamma_\alpha^0) \frac{1}{n}$$

$$\times \sum_t K_{h_0}(y_{t-\beta} - y_\beta) \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} \gamma_\alpha^{*0}(\underline{y}_{t-\beta}) p(x)\, dx.$$

Let $\mathcal{S}_{1n}^i$ and $\mathcal{S}_{2n}^i$ be the ith elements of $\mathcal{S}_{1n}$ and $\mathcal{S}_{2n}$, respectively, with $\underline{S}^{*ij}(\cdot)$ being the (i,j) element of $\underline{S}^*(\cdot)$. By the dominated convergence theorem and the integrability condition, we have

$$\mathcal{S}_{1n}^i = \frac{\sqrt{h}}{\sqrt{n}} \sum_t \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} v(x_t)^{1/2} \varepsilon_t$$

$$\times \left[\int K_h(y_\alpha - y_\alpha^0) \underline{S}^{i1*}(x, \gamma_\alpha^0) \sum_{\beta=1,\neq\alpha}^{d} (K * K)_{h_0}(y_\beta - y_{t-\beta}) \nabla G_m(y_\beta, \underline{y}_{t-\beta}) p(x)\, dx \right]$$

$$+ \frac{\sqrt{h}}{\sqrt{n}} \sum_t \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} v(x_t)(\varepsilon_t^2 - 1)$$

$$\times \left[\int K_h(y_\alpha - y_\alpha^0) \underline{S}^{i2*}(x, \gamma_\alpha^0) \sum_{\beta=1,\neq\alpha}^{d} (K * K)_{h_0}(y_\beta - y_{t-\beta}) \nabla G_v(y_\beta, \underline{y}_{t-\beta}) p(x)\, dx \right]$$

$$= \frac{\sqrt{h}}{\sqrt{n}} \sum_t \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} [v(x_t)^{1/2} \varpi_{i1}^1(x_t) \varepsilon_t + v(x_t) \varpi_{i2}^1(x_t)(\varepsilon_t^2 - 1)] + o_p(1),$$

where

$$\varpi_{ij}^1(x_t) = \nabla G^j(y_\alpha^0, \underline{y}_{t-\alpha}) \sum_{\beta=1,\neq\alpha}^{d} \int \underline{S}^{ij*}[(y_\alpha^0, y_{t-\beta}, \underline{y}_{(\alpha,\beta)}), \gamma_\alpha^0] p(y_\alpha^0, y_{t-\beta}, \underline{y}_{(\alpha,\beta)})\, d\underline{y}_{(\alpha,\beta)}$$

and $\nabla G^j(\cdot) = \nabla G_m(\cdot)$, for $j = 1$; $\nabla G_v(\cdot)$, for $j = 2$. Because $p_2(\cdot)/p(\cdot)$ and $\varpi_{ij}(\cdot)$ are bounded under the condition of compact support, applying the law of large numbers for i.i.d. errors $\xi_t = (\varepsilon_t, (\varepsilon_t^2 - 1))^\top$ leads to $\mathcal{S}_{1n}^i = o_p(1)$ and consequently $\mathcal{S}_{1n} = o_p(1)$. Likewise,

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

$$\mathcal{S}_{2n}^{i} = \frac{\sqrt{h}}{\sqrt{n}} \sum_{t} \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} \gamma_{1\alpha}^{*0}(\underline{y}_{t-\beta})$$
$$\times \left[\int K_h(y_\alpha - y_\alpha^0) \underline{S}^{i1*}(x, \gamma_\alpha^0) \sum_{\beta=1,\neq\alpha}^{d} K_h(y_{t-\beta} - y_\beta) p(x)\, dx \right]$$
$$+ \frac{\sqrt{h}}{\sqrt{n}} \sum_{t} \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} \gamma_{2\alpha}^{*0}(\underline{y}_{t-\beta})$$
$$\times \left[\int K_h(y_\alpha - y_\alpha^0) \underline{S}^{i2*}(x, \gamma_\alpha^0) \sum_{\beta=1,\neq\alpha}^{d} K_h(y_{t-\beta} - y_\beta) p(x)\, dx \right]$$
$$= \frac{\sqrt{h}}{\sqrt{n}} \sum_{t} \frac{p_2(\underline{y}_{t-\beta})}{p(x_t)} [\varpi_{i1}^2(x_t) m_{\underline{\alpha}}(\underline{y}_{t-\alpha}) + \varpi_{i1}^2(x_t) v_{\underline{\alpha}}(\underline{y}_{t-\alpha})] + o_p(1),$$

where

$$\varpi_{ij}^2(x_t) = \sum_{\beta=1,\neq\alpha}^{d} \int \underline{S}^{ij*}[(y_\alpha^0, y_{t-\beta}, \underline{y}_{(\alpha,\beta)}), \gamma_\alpha^0] p(y_\alpha^0, y_{t-\beta}, \underline{y}_{(\alpha,\beta)})\, d\underline{y}_{(\alpha,\beta)},$$

and, for the same reason as before, we get $\mathcal{S}_{2n}^{i} = o_p(1)$ and $\mathcal{S}_{2n} = o_p(1)$, because $E(m_{\underline{\alpha}}(\underline{y}_{t-\alpha})) = E(v_{\underline{\alpha}}(\underline{y}_{t-\alpha})) = 0$.

We finally show negligibility of the last term:

$$T_{3n} = \frac{\sqrt{h}}{\sqrt{n}} \sum_{t=d+1}^{n'} K_h(y_{t-\alpha} - y_\alpha^0) u_t [\tilde{\gamma}_\alpha(\underline{y}_{t-\alpha}) - \gamma_\alpha^0(\underline{y}_{t-\alpha})].$$

Substituting the error decomposition for $\tilde{\gamma}_\alpha(\underline{y}_{t-\alpha}) - \gamma_\alpha^0(\underline{y}_{t-\alpha})$ and interchanging the summations gives

$$T_{3n} = \sum_{\beta=1,\neq\alpha}^{d} \frac{\sqrt{h}}{n\sqrt{n}} \sum_{t} \sum_{s(\neq t)} K_h(y_{t-\alpha} - y_\alpha^0) K_{h_0}(y_{s-\beta} - y_\beta)$$
$$\times \frac{p_2(\underline{y}_{s-\beta})}{p(x_s)} \gamma_\alpha^{*0}(\underline{y}_{s-\beta}) u_t \gamma_\alpha^{*0}(\underline{y}_{s-\beta})$$
$$+ \sum_{\beta=1,\neq\alpha}^{d} \frac{\sqrt{h}}{n\sqrt{n}} \sum_{t} \sum_{s(\neq t)} K_h(y_{t-\alpha} - y_\alpha^0)(K * K)_{h_0}(y_{s-\beta} - y_\beta)$$
$$\times \frac{p_2(\underline{y}_{s-\beta})}{p(x_s)} (\nabla G(x_\beta, \underline{y}_{s-\beta}) \odot u_t \xi_s)$$
$$+ o_p(1),$$

where the $o_p(1)$ errors for the remaining bias terms hold under the assumption that $\sqrt{nh}h_0^2 = o(1)$. For

$$\pi_n^{i,\beta}(z_t, z_s) = K_h(y_{t-\alpha t} - y_\alpha^0) K_{h_0}(y_{s-\beta} - y_\beta) \frac{p_2(\underline{y}_{s-\beta})}{p(x_s)} u_t \gamma_\alpha^{*0}(\underline{y}_{s-\beta}) \sqrt{h}/(n\sqrt{n}),$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

we can easily check that $E(\pi_{1n}^{i,\beta}(z_t,z_s)|z_t) = E(\pi_{1n}^{i,\beta}(z_t,z_s)|z_s) = 0$, for $t \neq s$, implying that $\sum\sum_{t\neq s}\ \pi_n^{i,\beta}(z_t,z_s)$ is a degenerate second-order U-statistic. The same conclusion also holds for the second term. Hence, the two double sums are mean zero and have variance of the same order as

$$n^2 \times \{E\pi_n^{i,\beta}(z_t,z_s)^2 + E\pi_n^{i,\beta}(z_t,z_s)E\pi_n^{i,\beta}(z_s,z_t)\},$$

which is of order $n^{-1}h^{-1}$. Therefore, $T_{3n} = o_p(1)$. ■

LEMMA A.2. (Masry, 1996). *Suppose that Assumption C holds. Then, for any vector* $\mu = (\mu_1,\ldots,\mu_d)^\top$ *with* $|\mu| = \Sigma_j \mu_j \leq \omega$,

(a) $\sup_{x\in\mathcal{X}}|D_x^\mu \hat{p}(x) - D_x^\mu p(x)| = O_p(\sqrt{\log n/ng^{(2|\mu|+d)}}) + O_p(g^{q-\mu})$,
(b) $\sup_{x\in\mathcal{X}}|D_x^\mu \tilde{m}(x) - D_x^\mu m(x)| = O_p(\sqrt{\log n/nh_0^{(2|\mu|+d)}}) + O_p(h_0^{q-\mu}) \equiv \rho_n(\mu)$,
(c) $\sup_{x\in\mathcal{X}}|\tilde{m}(x) - m(x) - \tilde{L}(x)| = O_p(\rho_n^2)$, *where,*

$$\tilde{L}(x) = \frac{1}{n}\sum_{s\neq t}\frac{K_{h_0}(x_s - x)}{p(x_s)}v^{1/2}(x_s)\varepsilon_s + h_0^q b_n(x).$$

LEMMA A.3. *Suppose that Assumption C holds. Then, for any vector* $\mu = (\mu_1,..,\mu_d)^\top$ *with* $|\mu| = \Sigma_j \mu_j \leq \omega$,

(a) $\sup_{x_\alpha\in\mathcal{X}_\alpha}|D^\mu\hat{\varphi}_\alpha(y_\alpha) - D^\mu\varphi_\alpha(y_\alpha)| = O_p(\sqrt{\log n/nh^{(2|\mu|+1)}}) + O_p(h^{q-\mu}) + O_p(\rho_n^2(\mu))$,
(b) $\sup_{x_\alpha\in\mathcal{X}_\alpha}|\hat{\varphi}_\alpha(y_\alpha) - \varphi_\alpha(y_\alpha) - \hat{L}_\varphi(y_\alpha)| = O_p(\rho_n^2) + o_p(n^{-1/2})$,

where

$$\hat{L}_\varphi(y_\alpha) = \frac{1}{n}\sum_t (K*K)_h(y_{t-\alpha} - y_\alpha)\frac{p_{\bar{\alpha}}(\underline{y}_{t-\alpha})}{p(x_t)}\begin{bmatrix} G_m'(m(y_\alpha,\underline{y}_{t-\alpha}))v(x_t)^{1/2} \\ G_v'(v(y_\alpha,\underline{y}_{t-\alpha})v(x_t)) \end{bmatrix}\begin{bmatrix} \varepsilon_t, \\ \varepsilon_t^2 - 1 \end{bmatrix}$$

$$+ \frac{1}{n}\sum_t K_h(y_{t-\alpha} - y_\alpha)\frac{p_2(\underline{y}_{t-\alpha})}{p(x_t)}[m_\alpha(\underline{y}_{t-\alpha}),v_\alpha(\underline{y}_{t-\alpha})]^\top + h^q\bar{b}_\alpha(y_\alpha).$$

Proof. We first show (b). For notational simplicity, the bandwidth parameter h (only in this proof) abbreviates h_0. From the decomposition results for the instrumental variable estimates,

$$\hat{\varphi}_\alpha(y_\alpha) - \varphi_\alpha(y_\alpha) = [I_2 \otimes e_1^\top Q_n^{-1}]\tau_n,$$

where $Q_n = [\hat{q}_{ni+j-2}(y_\alpha)]_{(i,j)=1,2}$, with $\hat{q}_{ni} = (1/n)\sum_{t=d}^n K_h(y_{t-\alpha} - y_\alpha)[\hat{p}_{\bar{\alpha}}(\underline{y}_{t-\alpha})/\hat{p}(x_t)][(y_{t-\alpha} - y_\alpha)/h]^i$, for $i = 0,1,2$, and $\tau_n = (1/n)\sum_t K_h(y_{t-\alpha} - y_\alpha)[\hat{p}_{\bar{\alpha}}(\underline{y}_{t-\alpha})/\hat{p}(x_t)][\tilde{r}_t - \varphi_\alpha(y_\alpha) - (y_{t-\alpha} - y_\alpha)\nabla\varphi_\alpha(y_\alpha)] \otimes (1,(y_{t-\alpha} - y_\alpha)/h)^\top$. By the Cauchy–Schwarz inequality and Lemma A.2 applied with Taylor expansion, it holds that

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

$$\sup_{x_\alpha \in \mathcal{X}_\alpha} \frac{1}{n}\sum_{t=d}^{n} K_h(y_{t-\alpha}-y_\alpha)\left[\frac{\hat{p}_{\bar{\alpha}}(\underline{y}_{t-\alpha})}{\hat{p}(x_t)}-\frac{p_{\bar{\alpha}}(\underline{y}_{t-\alpha})}{p(x_t)}\right]\left(\frac{y_{t-\alpha}-y_\alpha}{h}\right)^i$$

$$\le \sup_{x\in\mathcal{X}}\left|\frac{\hat{p}_{\bar{\alpha}}(\underline{y}_\alpha)}{\hat{p}(x)}-\frac{p_{\bar{\alpha}}(\underline{y}_\alpha)}{p(x)}\right| \sup_{x_\alpha\in\mathcal{X}_\alpha}\frac{1}{n}\sum_{t=d}^{n} K_h(y_{t-\alpha}-y_\alpha)\left|\frac{y_{t-\alpha}-y_\alpha}{h}\right|^i$$

$$= O_p\left(\sup_{x\in\mathcal{X}}|\hat{p}(x)-p(x)|\right)\equiv O_p(\rho_n),$$

where the boundedness condition of C.2 is used for the last line. Hence, the standard argument of Masry (1996) implies that $\sup_{x_\alpha\in\mathcal{X}_\alpha}|\hat{q}_{ni}-q_i| = o_p(1)$, where $q_i = \int K(u_1)u_1^i\,du_1$. From $q_0=1$, $q_1=0$, and $q_2=\mu_K^2$, we get the following uniform convergence result for the denominator term; that is, $e_1^\top Q_n^{-1}\xrightarrow{p} e_1^\top$, uniformly in $y_\alpha\in\mathcal{X}_\alpha$. For the numerator, we show the uniform convergence rate of the first element of τ_n because the other terms can be treated in the same way. Let τ_n^1 denote the first element of τ_n, that is,

$$\tau_n^1 = \frac{1}{n}\sum_t K_h(y_{t-\alpha}-y_\alpha)\frac{\hat{p}_{\bar{\alpha}}(\underline{y}_{t-\alpha})}{\hat{p}(x_t)}[G_m(\tilde{m}(x_t))-M_\alpha(y_\alpha)-(y_{t-\alpha}-y_\alpha)m'_\alpha(y_\alpha)],$$

or alternatively,

$$\tau_n^1 = \frac{1}{n}\sum_t K_h(y_{t-\alpha}-y_\alpha)r(x_t;\hat{\mathbf{g}}),$$

where

$$r(x_t;\mathbf{g}) = \frac{g_2(\underline{y}_{t-\alpha})}{g_3(x_t)}[G_m(g_1(x_t))-M_\alpha(y_\alpha)-(y_{t-\alpha}-y_\alpha)m'_\alpha(y_\alpha)],$$

$$\mathbf{g}(x_t) = [g_1(x_t),g_2(\underline{y}_{t-\alpha}),g_3(x_t)] = [m(x_t),p_{\bar{\alpha}}(\underline{y}_{t-\alpha}),p(x_t)],$$

$$\hat{\mathbf{g}} = \hat{\mathbf{g}}(x_t) = [\tilde{m}(x_t),\hat{p}_{\bar{\alpha}}(\underline{y}_{t-\alpha}),\hat{p}(x_t)].$$

Because $p_{\bar{\alpha}}(\cdot)/p(\cdot)$ is bounded away from zero and G_m has a bounded second-order derivative, the functional $r(x_t;\mathbf{g})$ is Frechet differentiable in $\mathbf{g}$, with respect to the sup norm $\|\cdot\|_\infty$, with the (bounded) functional derivative $R(x_t;\mathbf{g}) = [\partial r(x_t;\mathbf{g})/\partial\mathbf{g}]\downarrow_{\mathbf{g}=\mathbf{g}(x_t)}$. This implies that for all $\mathbf{g}$ with $\|\mathbf{g}-\mathbf{g}^0\|_\infty$ small enough, there exists some bounded function $b(\cdot)$ such that

$$\|r(x_t;\mathbf{g})-r(x_t;\mathbf{g}^0)-R(x_t;\mathbf{g}^0)(\mathbf{g}-\mathbf{g}^0)\|_\infty \le b(x_t)\|\mathbf{g}-\mathbf{g}^0\|_\infty^2.$$

By Lemma A.2, $\|\hat{\mathbf{g}}-\mathbf{g}^0\|_\infty^2 = O_p(\rho_n^2)$, and consequently, we can properly linearize τ_n^1 as

$$\tau_n^1 = \frac{1}{n}\sum_t K_h(y_{t-\alpha}-y_\alpha)r(x_t;\mathbf{g}^0)+\frac{1}{n}\sum_t K_h(y_{t-\alpha}-y_\alpha)R(x_t;\mathbf{g}^0)(\hat{\mathbf{g}}-\mathbf{g}^0)+O_p(\rho_n^2),$$

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press

where the $O_p(\rho_n^2)$ error term is uniformly in x_α. After plugging $G_m(m(x_t)) = c_m + \Sigma_{1\le\beta\le d} m_\beta(y_{t-\beta})$ into $r(x_t;\mathbf{g}^0)$, a straightforward calculation shows that

$$\tau_n^1 = \frac{1}{n}\sum_t K_h(y_{t-\alpha}-y_\alpha)\varsigma_t[1+O_p(\rho_n)] + \frac{1}{n}\sum_t K_h(y_{t-\alpha}-y_\alpha)\frac{p_{\bar\alpha}(\underline{y}_{t-\alpha})}{p(x_t)}G_m'(m(x_t))[\tilde m(x_t)-m(x_t)] + \frac{h^q}{q!}\mu_q(k)b_{1\alpha}(y_\alpha)+o_p(h^q), \tag{A.13}$$

where $\varsigma_t = [p_2(\underline{y}_{t-\alpha})/p(x_t)]M_{\bar\alpha}(\underline{y}_{t-\alpha})$ and $M_{\bar\alpha}(\underline{y}_{t-\alpha}) = \Sigma_{1\le\beta\le d,(\ne\alpha)} m_\beta(\underline{y}_{t-\alpha})$. Note that as a result of the identification condition $E[\varsigma_t|\underline{y}_{t-\alpha}]=0$ and consequently the first term is a standard stochastic term appearing in kernel estimates. For a further asymptotic expansion of the second term of τ_n^1, we use the stochastic equicontinuity argument to the empirical process $\{v_n(\cdot,\cdot)\}$, indexed by $(y_\alpha,\delta)\in\mathcal{X}_\alpha\times\mathcal{T}$, with $\mathcal{T}=\{\delta:\|\delta\|_{\omega,2,\mathcal{X}_\alpha}\le C\}$, such that

$$v_n(y_\alpha,\delta) = \frac{1}{\sqrt{n}}\sum_{t=d+1}^{n'}[f(x_t;y_\alpha,\delta)-E(f(x_t;y_\alpha,\delta))],$$

where $f(x_t;y_\alpha,\delta) = K[(y_{t-\alpha}-y_\alpha)/h]h^\omega[p_{\bar\alpha}(\underline{y}_{t-\alpha})/p(x_t)]G_m'(m(x_t))\delta(y_{t-\alpha})$, for some positive integer $\omega > d/2$. Let $\tilde\delta = h^{-\omega-1/2}[\tilde m(x_t)-m(x_t)]$. From the uniform convergence rate in Lemma A.2 and the bandwidth condition in C.8(iii), it follows that $\|\tilde\delta\|_{\omega,2,\mathcal{X}} = O_p(h^{-\omega-1/2}[\sqrt{\log n/nh^{(2\omega+d)}}+h^{q-\omega}]) = o_p(1)$, leading to (i) $\Pr(\hat\delta\in\mathcal{T})\to 1$ and (ii) $\rho((y_\alpha,\hat\delta),(y_\alpha,\delta^0))\xrightarrow{p}0$, where $\delta^0=0$. These conditions and stochastic equicontinuity of $v_n(\cdot,\cdot)$ at (y_α,δ^0) yield $\sup_{y_\alpha\in\mathcal{X}_\alpha}|v_n(y_\alpha,\hat\delta)-\nu_n(y_\alpha,\delta^0)| = \sup_{x_\alpha\in\mathcal{X}_\alpha}|v_n(y_\alpha,\hat\delta)| = o_p(1)$. Thus, the second term of τ_n^1 is approximated with an $o_p(1/\sqrt{n})$ error (uniform in y_α) by

$$\int K_h(y_{t-\alpha}-y_\alpha)\frac{p_{\bar\alpha}(\underline{y}_{t-\alpha})}{p(x_t)}G_m'(m(x_t))[\tilde m(x_t)-m(x_t)]p(x_t)\,dx_t,$$

which, by substituting $\tilde L(x_t)$ for $\tilde m(x_t)-m(x_t)$, is given by

$$\frac{1}{n}\sum_s (K*K)_h\left(\frac{y_{s-\alpha}-y_\alpha}{h}\right)p_{\bar\alpha}(\underline{y}_{s-\alpha})G_m'(m(y_\alpha,\underline{y}_{s-\alpha}))\frac{v^{1/2}(x_s)\varepsilon_s}{p(x_s)} + \frac{h^q}{q!}\mu_q(k)b_{2\alpha}(y_\alpha), \tag{A.14}$$

where $(K*K)(\cdot)$ is actually a convolution kernel as defined before. Hence, by letting $\bar b_\alpha(y_\alpha)$ summarize two bias terms appearing in (A.13) and (A.14), Lemma A.3(b) is shown. The uniform convergence results in part (a) then follow by the standard arguments of Masry (1996), because two stochastic terms in the asymptotic expansion of $\hat\varphi_\alpha(y_\alpha)-\varphi_\alpha(y_\alpha)$ consist only of univariate kernels. ■

https://doi.org/10.1017/S026646660420603X Published online by Cambridge University Press