1 Introduction

Stochastic volatility models are of considerable interest in empirical finance. There are many types [illegible]

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

In this paper, we investigate a *generalized additive nonlinear* ARCH model (GANARCH);

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

$$\begin{aligned} m\left(y_{t-1}, y_{t-2}, \ldots, y_{t-d}\right) &= F_m\left(c_m + \sum_{\alpha=1}^{d} m_\alpha(y_{t-\alpha})\right), & (1.2)\\ v\left(y_{t-1}, y_{t-2}, \ldots, y_{t-d}\right) &= F_v\left(c_v + \sum_{\alpha=1}^{d} v_\alpha(y_{t-\alpha})\right), & (1.3)\end{aligned}$$

where $m_\alpha(\cdot)$ and $v_\alpha(\cdot)$ are any smooth but unknown function, while $F_m(\cdot)$ 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, i.e., $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ 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) used α-mixing or geometric ergodicity to identify the nonlinear time series model. Similar results were obtained for the additive nonlinear 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

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

[1] The extension to allow the F transformations to be of unknown functional form is considerably more complicated, but see Horowitz (1999).

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 below, (1.1)-(1.3) includes 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) studied the kernel estimation of conditional mean function, $m(\cdot)$ in (1.1). The so-called CHARN (Conditionally Heteroscedastic 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) applied the Nadaraya-Watson and local linear smoothing methods, respectively, to jointly estimate $v(\cdot)$ together with $m(\cdot)$. Alternatively, Fan and Yao (1996) and Ziegelmann (2002) proposed local linear least square estimation for volatility function, with the extension given by Avramidis (2002) based on local linear maximum likelihood estimation. Also, in a nonlinear VAR context, Härdle, Tsybakov and Yang (1996) dealt 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 method in high dimensions suffers 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.

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) showed that it is possible to estimate $m_\alpha(\cdot)$ and $m(\cdot)$ with the one-dimensional optimal rate of convergence - e.g., $n^{2/5}$ for twice differentiable functions - regardless of d. The estimates are now easily illustrated and interpreted. For these reasons, since the eighties, 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 (1987,1991) 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 developed specific (linear) backfitting procedures and established 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 can't be easily generalized to nonlinear cases like Generalized Additive Models.

A more recent approach, called marginal integration (MI), is theoretically more manipulable - its statistical properties are easy to derive, since it simply uses averaging of multivariate kernel estimates. Developed independently by Newey (1994), Tjøstheim and Auestadt (1994a), and Linton and Nielsen (1995), its advantage of theoretical convenience inspired the subsequent applications such as Linton, Wang, Chen, and Härdle (1997) for transformation models and Linton, Nielsen, and van de Geer (2003) for hazard models with censoring. In the time series models that are special cases of (1.1) and (1.2) with F_m being the identity, Chen and Tsay (1993 a,b) and Masry and Tjøstheim (1997) applied backfitting and MI, respectively, to estimate the conditional mean function. Mammen, Linton, and Nielsen (1999) provided 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 was 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 relied on marginal integration with local linear fits as a pilot estimate, and derived asymptotic properties.

This paper features two contributions to the additive literature. The first concerns theoretical development of a new estimation tool called local instrumental variable method for additive models, which was outlined for simple additive cross-sectional regression in the paper 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 an 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 and can be easily shown to attain the optimal one-dimensional rate as required. Furthermore, it is computationally more efficient than the two existing methods (backfitting and MI), in the sense that it reduces the computations up to a factor of n smoothings. The other contribution relates to the general coverage of the model we work with. The model in (1.1) through (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 discussion above are regarded as a subclass of the data generating process for $\{y_t\}$ in (1.1) through (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 were made in a parametric ARCH models. Nelson (1991) considered a model for the log of the conditional variance - the Exponential (G)ARCH class, to embody the multiplicative effects of volatility. It was also argued to use the

Box-Cox transformation for volatility which is intermediate between linear and logarithm. Since 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. Additionally, from the perspective of potential misspecification problems, the transformation used here alleviates the restriction imposed by additivity assumption, which increases the approximating power of our model. Note that when the lagged variables in (1.1) through (1.3) are replaced by different covariates and the observations are 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) through (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

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

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 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

$$\begin{aligned}\frac{E(Wy|X_1)}{E(W|X_1)} &= \frac{\int W(X)m(X)\frac{p(X)}{p_1(X_1)}dX_2}{\int W(X)\frac{p(X)}{p_1(X_1)}dX_2} = \frac{\int W(X)m(X)p(X)dX_2}{\int W(X)p(X)dX_2} \\ &= \frac{\int m(X)p_2(X_2)dX_2}{\int p_2(X_2)dX_2} = \int m(X)p_2(X_2)dX_2\end{aligned}$$

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 $\widehat{p}(\cdot), \widehat{p}_1(\cdot)$, and $\widehat{p}_2(\cdot)$ 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 $\widehat{W} = \widehat{p}_2(X_2)/\widehat{p}(X)$ and taking sample averages instead of population expectations. Section 3 provides a rigorous statistical

[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 which satisfies

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

[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.

treatment for feasible instrumental variable estimators based on local linear estimation. See Kim, Linton, and Hengartner (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 $i, j = 1, \ldots, n$, while 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 instrumental variable, and another n operations for regressions. Thus, it can be easily combined with bootstrap method whose computational costs often becomes prohibitive in the case of marginal integration.

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

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

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

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

for the W defined in (2.9). Since $m(\cdot)$ 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 throughout the paper. 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) through (1.3) with a focus on the αth components of mean and variance as

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

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

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

Note that the components $[m_\alpha(\cdot), v_\alpha(\cdot)]^T$ are identified, up to constant, c, by $\varphi_\alpha(\cdot)$, which will be our major interest in estimation. Below, we examine some details in 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 $n' = n + d$.

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

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

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

where $K_h(x) = \Pi_{i=1}^d K(x_i/h)/h^d$ and 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, $\widetilde{v}(\cdot)$, from (3.12) by replacing y_t with the squared residuals, $\widetilde{\varepsilon}_t^2 = (y_t - \widetilde{m}(x_t))^2$. Then, transforming $\widetilde{m}$ and $\widetilde{v}$ by the known links will leads to consistent estimates of $\widetilde{z}_t$,

$$\widetilde{z}_t = \widetilde{H}\left(x_t\right) = \left[G_m\left(\widetilde{m}\left(x_t\right)\right), G_v\left(\widetilde{v}\left(x_t\right)\right)\right]^T.$$

3.2 Step II: Instrumental Variable Estimation of Additive Components

This step involves the estimation of $\varphi_\alpha(\cdot)$, which is equivalent to $[m_\alpha(\cdot), v_\alpha(\cdot)]^T$, 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

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

where $\widehat{p}_{\underline{\alpha}}(\cdot)$ and $\widehat{p}(\cdot)$ are computed using the kernel function $L(\cdot)$, e.g., $\widehat{p}(x) = \sum_{t=1}^n \prod_{i=1}^d L_g(x_{it} - x_i)/n$ with $L_g(\cdot) \equiv L(\cdot/g)/g$. The instrumental variable local linear estimates $\widehat{\varphi}_\alpha(y_\alpha)$ are given as

$(a_1, a_2)^T$ through minimizing the localized squared errors elementwise

$$\min_{a_j, b_j} \sum_{t=d+1}^{n'} K_h \left(y_{t-\alpha} - y_\alpha\right) \widehat{W}_t \left\{\widetilde{z}_{jt} - a_j - b_j \left(y_{t-\alpha} - y_\alpha\right)\right\}^2, \tag{3.13}$$

where $\widetilde{z}_{jt}$ is the j-th element of $\widetilde{z}_t$. The closed form of the solution is

$$\widehat{\varphi}_\alpha(y_\alpha)^T = e_1^T \left(\mathbf{Y}_-^T \mathbf{K} \mathbf{Y}_-\right)^{-1} \mathbf{Y}_-^T \mathbf{K} \widetilde{\mathbf{Z}}, \tag{3.14}$$

where $e_1 = (1, 0)^T$, $\mathbf{Y}_- = [\iota, Y_-]$, $\mathbf{K} = \text{diag}[K_h(y_{d+1-\alpha} - y_\alpha)\widehat{W}_{d+1}, \ldots, K_h(y_{n'-\alpha} - y_\alpha)\widehat{W}_{n'}]$, and $\widetilde{\mathbf{Z}}$ $= (\widetilde{z}_{d+1}, \ldots, \widetilde{z}_{n'})^T$, with $\iota = (1, \ldots, 1)^T$ and $Y_- = (y_{d+1-\alpha} - y_\alpha, \ldots, y_{n'-\alpha} - y_\alpha)^T$.

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 $\{y_t\}$ which is defined by

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

Throughout the paper, we assume

C1. $\{y_t\}_{t=1}^\infty$ is stationary and strongly mixing with a mixing coefficient, $\alpha(k) = \rho^{-\beta k}$, for some $\beta > 0$.

C.1 is a standard mixing condition with a geometrically decreasing rate. However, the asymptotic theory for the instrumental variable estimator is developed based on a milder condition on the mixing coefficient - as was pointed out by Masry and Tjøstheim (1997), $\sum_{k=0}^\infty k^a \left\{\alpha(k)\right\}^{1-2/\nu} < \infty$, for some $\nu > 2$ and $0 < a < (1-2/\nu)$. It is easy to verify that this condition holds under C.1. Some technical conditions for regularity are stated.

C2. The additive component functions, $m_\alpha(\cdot)$, and $v_\alpha(\cdot)$, for $\alpha = 1, \ldots, d$, are continuous and twice differentiable on their compact supports.

C.3 The link functions, G_m and G_v, have bounded continuous second order derivatives over any compact interval.

C4. The joint and marginal density functions, $p(\cdot)$, $p_{\underline{\alpha}}(\cdot)$, 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.

C5. The kernel functions, $K(\cdot)$ and $L(\cdot)$, are a real bounded nonnegative symmetric function on 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, $|K(u) - K(v)| \leq C|u - v|$.

C6. (i) $g \to 0$, $ng^d \to \infty$, and (ii) $h \to 0$, $nh \to \infty$. (iii) The bandwidth satisfies $\sqrt{\frac{n}{h}}\alpha(t(n)) \to 0$,where $\{t(n)\}$ be a sequence of positive integers, $t(n) \to \infty$ such that $t(n) = o(\sqrt{nh})$.

Conditions C.2 through C.5 are standard in kernel estimation. The continuity assumption in C2 and C4, together with the compact support, implies that the functions are bounded. The additional bandwidth condition in C.6(iii) is necessary to control the effects from the dependence of mixing processes in showing the asymptotic normality of instrumental variable estimates. The proof of consistency, however, does not require this condition for bandwidths. 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 $(K*K)_i(u) = \int K(w)K(w+u)w^i dw$, a convolution of kernel functions, and $\mu^2_{K*K} = \int (K*K)_0(u)u^2 du$, while $||K||_2^2$ denotes $\int K^2(u)\,du$. 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 C.1 through C.6 hold. Then,*

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

where

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

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

Remarks. 1. To estimate $[m_\alpha(y_\alpha), v_\alpha(y_\alpha)]^T$, we can use the following recentered estimates, $\widehat{\varphi}_\alpha(y_\alpha) - \widehat{c}$, where $\widehat{c} = [\widehat{c}_m, \widehat{c}_v] = \frac{1}{n}[\sum_t y_t, \sum_t \widetilde{\varepsilon}_t^2]^T$ and $\widetilde{\varepsilon}_t = y_t - \widetilde{m}(x_t)$. Since $\widehat{c} = c + O_p(1/\sqrt{n})$, the bias and variance of $[\widehat{m}_\alpha(y_\alpha), \widehat{v}_\alpha(y_\alpha)]^T$ are the same as those of $\widehat{\varphi}_\alpha(y_\alpha)$. For $y = (y_1, \ldots, y_d)$, the estimates for the conditional mean and volatility are defined by

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

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

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

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

2. The first term of the bias is of the 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, $\widehat{p}_{\underline{\alpha}}(\underline{y}_{t-\alpha})/\widehat{p}\,(x_t)$. 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 IV 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, since we smooth $[G_m(\widetilde{m}\,(\cdot)),\ G_v(\widetilde{v}\,(\cdot))]$ with $\widetilde{m}\,(\cdot)$ and $\widetilde{v}\,(\cdot)$ given by (multivariate) local linear fits. When $G_m(\cdot)$ is the identity, we can directly smooth y instead of $G_m(\widetilde{m}\,(x_t))$ 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). The instrumental variable estimate of the α-th component, $[\widehat{M}_\alpha(y_\alpha), \widehat{V}_\alpha(y_\alpha)]$, is now the solution to the adjusted-kernel least squares in (3.13) with a modification that the (2×1) vector $\widetilde{z}_t$ is replaced by $[y_t, \widetilde{\varepsilon}_t^2]^T$ with $\widetilde{\varepsilon}_t$ defined in step I of section 2.2. Theorem 2 shows the asymptotic normality of these instrumental variable estimates. The proof is almost the same as that of Theorem 1 and is thus omitted.

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

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

where

$$\begin{aligned} b_\alpha^m(y_\alpha) &= \frac{h^2}{2}\mu_K^2 D^2 m_\alpha(y_\alpha) + \frac{g^2}{2}\mu_K^2 \int [D^2 p_{\underline{\alpha}}(z_{\underline{\alpha}}) - \frac{p_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} D^2 p(y_\alpha, z_{\underline{\alpha}})] m_{\underline{\alpha}}(z_{\underline{\alpha}}) dz_{\underline{\alpha}}, \\ \sigma_\alpha^m(y_\alpha) &= ||K||_2^2 \int \frac{p_{\underline{\alpha}}^2(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} [m_{\underline{\alpha}}^2(z_{\underline{\alpha}}) + v(y_\alpha, z_{\underline{\alpha}})] dz_{\underline{\alpha}}, \end{aligned}$$

and

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

where

$$\begin{aligned} b^v_\alpha(y_\alpha) &= \frac{h^2}{2}\mu_K^2 D^2 v_\alpha(y_\alpha) + \frac{g^2}{2}\mu_K^2 \int [D^2 p_{\underline{\alpha}}(z_{\underline{\alpha}}) - \frac{p_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})} D^2 p(y_\alpha, z_{\underline{\alpha}})] v_{\underline{\alpha}}(z_{\underline{\alpha}}) dz_{\underline{\alpha}}, \\ \Sigma^v_\alpha(y_\alpha) &= ||K||_2^2 \int \frac{p^2_{\underline{\alpha}}(z_{\underline{\alpha}})}{p(y_\alpha, z_{\underline{\alpha}})}[v^2_{\underline{\alpha}}(z_{\underline{\alpha}}) + \kappa_4(y_\alpha, z_{\underline{\alpha}}) v^2(y_\alpha, z_{\underline{\alpha}})] dz_{\underline{\alpha}}. \end{aligned}$$

Although the instrumental variable estimators achieve the one-dimensional optimal convergence rate, there is room for improvement in terms of variance. For example, compared to 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 Theorem 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 asymptotic covariance in Theorem 1 is the same as that in Yang, Härdle, and Nielsen (1999), where they 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, Linton, and Hengartner (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 show how to achieve this in practice. 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 $(c_v, \{v_\alpha(\cdot) : \alpha\})$, 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 $(c_v, \{v_\beta(\cdot) : \beta \neq \alpha\})$, what is the best estimator we can obtain for the function v_α? It turns out that this standard is very high and can't be achieved in practice. Instead we ask: suppose that we know $(c_m, \{m_\beta(\cdot) : \beta \neq \alpha\})$ and $(c_v, \{v_\beta(\cdot) : \beta \neq \alpha\})$, 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 $x_t = (y_{t-1}, \ldots, y_{t-d})$, or more concisely, $x_t = (y_{t-\alpha}, \underline{y}_{t-\alpha})$. For $\theta = (\theta_a, \theta_b) = (a_m, a_v, b_m, b_v)$, we define

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

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 $\widehat{\theta} = (\widehat{a}_m, \widehat{a}_v, \widehat{b}_m, \widehat{b}_v)$ 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;

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

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

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

The asymptotic distribution of local MLE has been studied by Avramidis (2002). For $y = (y_1, \ldots, y_d)$ $= (y_\alpha, \underline{y}_\alpha)$, define

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

where

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

With a minor generalization of the results by Avramidis (2002, Theorem 2), we obtain the following asymptotic properties for the infeasible estimators, $\widehat{\varphi}_\alpha(y_\alpha) = [\widehat{m}_\alpha(y_\alpha), \widehat{v}_\alpha(y_\alpha)] = [\widehat{a}_m, \widehat{a}_v]$.

Theorem 3 *Under Assumption B in the appendix, it holds that*

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

where $B_\alpha = \frac{1}{2}h^2\mu_K^2[m''_\alpha(y_\alpha), v''_\alpha(y_\alpha)]^T$, *and* ${}^*_\alpha(y_\alpha) = ||K||_2^2 D_\alpha^{-1} V_\alpha D_\alpha^{-1}$.

Remarks.

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

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

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

When the error distribution is Gaussian, we can further simplify the asymptotic variance; i.e.,

$$\begin{aligned}\omega_{11}(y_\alpha) &= \left[\int v^{-1}(y)\nabla F_m^2(G_m(m(y))p(y)d\underline{y}_\alpha\right]^{-1};\ \omega_{12} = \omega_{21} = 0;\\ \omega_{22}(y_\alpha) &= 2\left[\int v^{-2}(y)\nabla F_v^2(G_v(v(y))p(y)d\underline{y}_\alpha\right]^{-1}.\end{aligned}$$

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

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

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

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

These, together with $\kappa_3 = 0$, imply that the second term of $\Sigma^*_\alpha(y_\alpha)$ in Theorem 1 is greater than ${}^*_\alpha(y_\alpha)$ in the sense of positive definiteness, and hence $\Sigma^*_\alpha(y_\alpha) \geq {}^*_\alpha(y_\alpha)$, since the first term of $\Sigma^*_\alpha(y_\alpha)$ is a nonnegative matrix. The infeasible estimator is more efficient than the IV estimator, because the former uses more information concerning the mean-variance structure.

5.2 Feasible Estimation

Let $(\widetilde{c}_m, \{\widetilde{m}_\beta(\cdot) : \beta \neq \alpha\})$ and $(\widetilde{c}_v, \{\widetilde{v}_\beta(\cdot) : \beta \neq \alpha\})]$ be the estimators defined in the previous sections. Define the feasible local likelihood estimator $\widehat{\theta}^* = (\widehat{a}_m^*, \widehat{a}_v^*, \widehat{b}_m^*, \widehat{b}_v^*)$ as the minimizers of

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

where $\widetilde{\gamma}_\alpha(\cdot) = (\widetilde{\gamma}_{1\alpha}(\cdot), \widetilde{\gamma}_{2\alpha}(\cdot)) = (\widetilde{c}_m + \widetilde{m}_{\underline{\alpha}}(\cdot), \widetilde{c}_v + \widetilde{v}_{\underline{\alpha}}(\cdot))$. Then, the first order condition for $\widehat{\theta}^*$ is given by

$$0 = \overline{s}_n(\widehat{\theta}^*, \widetilde{\gamma}_\alpha) = \frac{1}{n} \sum_{t=d+1}^{n'} s_t(\widehat{\theta}^*, \widetilde{\gamma}_\alpha). \tag{5.15}$$

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

Theorem 4 *Under Assumption A and B in the appendix, it holds that*

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

This results 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 IV and two-step estimators. The design in our experiment is Additive Nonlinear ARCH(2)

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

$$\begin{aligned} v_1(y) &= 0.4\Phi_N(|2y|)[2 - \Phi_N(y)]y^2, \\ v_2(y) &= 0.4\left\{1/\sqrt{1 + 0.1y^2} + ln(1 + 4y^2) - 1\right\}, \end{aligned}$$

where $\Phi_N(\cdot)$ is the (cumulative) standard normal distribution function, and ε_t is i.i.d. with $N(0, 1)$. Fig.1(solid lines) depicts the shapes of the volatility functions defined by $v_1(\cdot)$ and $v_2(\cdot)$. Based on the above model, we simulate 500 samples of ARCH processes with sample size $n = 500$. For each realization of the ARCH process, we apply the IV estimation procedure in (3.13) with $\widetilde{z}_t = y_t^2$ 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 MLE 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 IV estimates in (3.13), and $c_h = 1.5$ for the (feasible and infeasible) local MLE. To evaluate the performance of the estimators, we calculate the mean squared error, together with the mean absolute deviation error, for each simulated data ; for $\alpha = 1, 2$,

$$\begin{aligned} e_{\alpha,MSE} &= \left\{ \frac{1}{50} \sum_{i=1}^{50} [v_\alpha(y_i) - \widehat{v}_\alpha(y_i)]^2 \right\}^{1/2}, \\ e_{\alpha,MAE} &= \frac{1}{50} \sum_{i=1}^{50} |v_\alpha(y_i) - \widehat{v}_\alpha(y_i)|, \end{aligned}$$

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

Table 1: Averages MSE and MAE for three volatility estimators

	$e_{1,MSE}$	$e_{2,MSE}$	$e_{1,MAE}$	$e_{2,MAE}$
oracle est.	.07636	.08310	.06049	.06816
IV est.	.08017	.11704	.06660	.09725
two-step	.08028	.08524	.06372	.07026

Table 1 shows that the infeasible oracle estimator is the best out of the three, to one's expectation. The performance of the IV estimator seems to be reasonably good, compared to the local MLE's, at least in estimating the volatility function of the first lagged variable. However, the overall accuracy of the IV 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, Fig.1 shows the averaged estimates of volatility functions, where the averages are made, at each grid, over 500 simulations. In Fig.2, we also illustrate the estimates for three typical (consecutive) realizations of ARCH processes.

A 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 $\alpha = 1$; below 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, $\widehat{\varphi}_1(y_1)-\varphi_1(y_1)$, into the main stochastic term and bias. Use $X_n \simeq Y_n$ to mean $X_n = Y_n \{1+o_p(1)\}$ in the following. Let $vec(X)$ denote the vectorization of the elements of the matrix X along with columns.

Step I: **Decompositions and Approximations**

Since $\widehat{\varphi}_1(y_1)$ is a column vector, the vectorization of eq. (3.14) gives

$$\widehat{\varphi}_1(y_1) = [I_2 \quad e_1^T \left(\mathbf{Y}_-^T\mathbf{K}\mathbf{Y}_-\right)^{-1}] \left(I_2 \quad \mathbf{Y}_-^T\mathbf{K}\right) \mathrm{vec}\left(\widetilde{\mathbf{Z}}\right).$$

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

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

by the identity,

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

since

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

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

$$\widehat{\varphi}_1(y_1) - \varphi_1(y_1) = [I_2 \quad e_1^T Q_n^{-1}]\tau_n,$$

where

$$\tau_n = \left(I_2 \quad D_h^{-1}\mathbf{Y}_-^T\mathbf{K}\right) \mathrm{vec}[\widetilde{\mathbf{Z}} - \iota\varphi_1^T(y_1) - Y_-\nabla\varphi_1^T(y_1)].$$

Observing

$$\tau_n = \frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\widehat{W}_k}(y_{k-1}-y_1)\,[\widetilde{z}_k - \varphi_1(y_1) - (y_{k-1}-y_1)\nabla\varphi_1(y_1)] \quad (1, \frac{y_{k-1}-y_1}{h})^T,$$

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

$$\begin{aligned} \tau_n &= \frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\widehat{W}_k}(y_{k-1}-y_1)\,[\widetilde{z}_k - z_k + H_2\left(\underline{y}_{k-1}\right)] \quad (1, \frac{y_{k-1}-y_1}{h})^T \\ &\quad + \frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\widehat{W}_k}(y_{k-1}-y_1)\,[\varphi_1(y_{k-1}) - \varphi_1(y_1) - (y_{k-1}-y_1)\nabla\varphi_1(y_1)] \quad (1, \frac{y_{k-1}-y_1}{h})^T. \end{aligned}$$

Due to the boundedness condition in C.2, the Taylor expansion applied to $[G_m(\widetilde{m}(x_k)), \quad G_v(\widetilde{v}(x_k))]$ at $[m(x_k), v(x_k)]$ yields the first term of τ_n as

$$\widetilde{\tau}_n \equiv \frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\widehat{W}_k}(y_{k-1}-y_1)\,[\widetilde{u}_k \quad (1, \frac{y_{k-1}-y_1}{h})^T],$$

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

$$\begin{aligned}
\widetilde{z}_k^1 &\equiv \{\nabla G_m\left(m\left(x_k\right)\right)\left[\widetilde{m}\left(x_k\right) - m\left(x_k\right)\right], \nabla G_v\left(v\left(x_k\right)\right)\left[\widetilde{v}\left(x_k\right) - v\left(x_k\right)\right]\}^T \\
\widetilde{z}_k^2 &\equiv \frac{1}{2}\{D^2 G_m\left(m^*\left(x_k\right)\right)\left[\widetilde{m}\left(x_k\right) - m\left(x_k\right)\right]^2, D^2 G_v\left(v^*\left(x_k\right)\right)\left[\widetilde{v}\left(x_k\right) - v\left(x_k\right)\right]^2\}^T,
\end{aligned}$$

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

$$s_{0n} = \frac{h^2}{2}\frac{1}{n}\sum_{k=d+1}^{n'} K_h^{\widehat{W}_k}\left(y_{k-1} - y_1\right)\left(\frac{y_{k-1} - y_1}{h}\right)^2 [D^2\varphi_1(y_1) \quad (1, \frac{y_{k-1} - y_1}{h})^T](1 + o_p\left(1\right)).$$

$\widetilde{\tau}_n$ continues to be simplified by some further approximations. Define the marginal expectation of estimated density functions, $\widehat{p}_2(\cdot)$ and $\widehat{p}(\cdot)$ as follows

$$\begin{aligned}
\overline{p}(y_{k-1}, \underline{y}_{k-2}) &\equiv \int L_g(z_1 - y_{k-1}) L_g(z_2 - \underline{y}_{k-2}) p(z_1, z_2) dz_1 dz_2, \\
\overline{p}_2(\underline{y}_{k-2}) &\equiv \int L_g(z_2 - \underline{y}_{k-2}) p_2(z_2) dz_2.
\end{aligned}$$

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

$$\begin{aligned}
t_{1n} &\equiv \frac{1}{n}\sum_{k=d+1}^{n'} K_h\left(y_{k-1} - y_1\right)\frac{\overline{p}_2(\underline{y}_{k-1})}{\overline{p}(x_k)}[\widetilde{z}_k^1 \quad (1, \frac{y_{k-1} - y_1}{h})^T], \\
t_{2n} &\equiv \frac{1}{n}\sum_{k=d+1}^{n'} K_h\left(y_{k-1} - y_1\right)\frac{\overline{p}_2(\underline{y}_{k-1})}{\overline{p}(x_k)}[H_2(\underline{y}_{k-1}) \quad (1, \frac{y_{k-1} - y_1}{h})^T],
\end{aligned}$$

based on the following results:

$(i)\frac{1}{n}\sum_{k=d+1}^{n'} K_h\left(y_{k-1} - y_1\right)\frac{\widehat{p}_2(\underline{y}_{k-1})}{\widehat{p}(x_k)}[\widetilde{z}_k^2 \quad (1, \frac{y_{k-1}-y_1}{h})^T] = o_p\left(\frac{1}{\sqrt{nh}}\right)$,

$(ii)\ \frac{1}{n}\sum_{k=d+1}^{n'} K_h\left(y_{k-1} - y_1\right)[\frac{\widehat{p}_2(\underline{y}_{k-1})}{\widehat{p}(x_k)} - \frac{\overline{p}_2(\underline{y}_{k-1})}{\overline{p}(x_k)}][H_2(\underline{y}_{k-1}) \quad (1, \frac{y_{k-1}-y_1}{h})^T] = o_p\left(\frac{1}{\sqrt{nh}}\right)$,

$(iii)\ \frac{1}{n}\sum_{k=d+1}^{n'} K_h\left(y_{k-1} - y_1\right)[\frac{\widehat{p}_2(\underline{y}_{k-1})}{\widehat{p}(x_k)} - \frac{\overline{p}_2(\underline{y}_{k-1})}{\overline{p}(x_k)}][\widetilde{z}_k^1 \quad (1, \frac{y_{k-1}-y_1}{h})^T] = o_p\left(\frac{1}{\sqrt{nh}}\right)$.

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

$$\begin{aligned}
&\max_k \left|\widetilde{m}\left(x_k\right) - m\left(x_k\right)\right|^2 \times \\
&\frac{1}{2}\frac{1}{n}\sum_k K_h\left(y_{k-1} - y_1\right)\frac{\widehat{p}_2(\underline{y}_{k-1})}{\widehat{p}(x_k)} D^2 G_m\left(m\left(x_k\right)\right)(1, \frac{y_{k-1} - y_1}{h})^T \\
= \ &o_p\left(1/\sqrt{nh}\right).
\end{aligned}$$

The last equality is direct from the uniform convergence theorems in Masry (1992) that

$$\max_t |\widetilde{m}(x_t) - m(x_t)| = O_p\left(\log n/\sqrt{nh^d}\right), \tag{A.16}$$

and $\frac{1}{n}\sum_k K_h(y_{k-1} - y_1)\frac{\widehat{p}_2(\underline{y}_{k-1})}{\widehat{p}(x_k)} D^2 G_m(m(x_k))(1, \frac{y_{k-1}-y_1}{h})^T = O_p(1)$. The proof for (ii) is given in Lemma A.1. The negligibility of (iii) follows in a similar way from (ii), considering (). While 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 local linear fits, $\widetilde{m}(x_k) - m(x_k)$ of $\widetilde{z}_k^1$, 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 local linear smoother in Jones, Davies and Park(1994). A similar expression holds for volatility estimates, $\widetilde{v}(x_k) - v(x_k)$, with a stochastic term of $\frac{1}{n}\sum_l \frac{K_h(x_l-x_k)}{p(x_l)} v(x_l)(\varepsilon_l^2$ -1). Define

$$\begin{aligned} & J_{k,n}(x_l) \\ \equiv\ & \frac{1}{nh^d}\sum_k \frac{K(y_{k-1} - y_1/h)\,K(x_l - x_k/h)}{p(x_l)}\frac{p_2(\underline{y}_{k-1})}{p(x_k)}[\mathrm{diag}(\nabla G_m, \nabla G_v) \quad (1, \frac{y_{k-1}-y_1}{h})^T], \end{aligned}$$

and let $\overline{J}(x_l)$ 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

$$\widetilde{t}_{1n} = \frac{1}{nh}\sum_l \overline{J}(x_l)\,[(v^{1/2}(x_l)\,\varepsilon_l, v(x_l)\,(\varepsilon_l^2 - 1))^T \quad I_2],$$

since the approximation errors from $\overline{J}(X_l)$ is negligible, i.e.,

$$\frac{1}{nh}\sum_l (J_{k,n} - \overline{J})[(v^{1/2}(X_l)\,\varepsilon_l, v(X_l)\,(\varepsilon_l^2 - 1))^T \quad I_2]^T = o_p\left(1/\sqrt{nh}\right),$$

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

$$\begin{aligned} \overline{J}(X_l) \ \simeq\ & \frac{1}{h}\int K(u_1 - y_1/h)\,K(u_1 - y_{l-1}/h)\int \frac{1}{h^{d-1}}K\left(\underline{y}_{l-1} - u_2/h\right)\frac{p_2(u_2)}{p(x_l)} \times \\ & [\mathrm{diag}(\nabla G_m(u), \nabla G_v(u)) \quad (1, \frac{u_1 - y_1}{h})^T]du_2 du_1 \\ \simeq\ & \frac{1}{h}\int K(u_1 - y_1/h)\,K(u_1 - y_{l-1}/h)\,\frac{p_2(\underline{y}_{l-1})}{p(x_l)} \times \\ & [\mathrm{diag}(\nabla G_m(u_1, \underline{y}_{l-1}), \nabla G_v(u_1, \underline{y}_{l-1})) \quad (1, \frac{u_1 - y_1}{h})^T]du_1 \\ \simeq\ & \frac{p_2(\underline{y}_{l-1})}{p(x_l)}[\mathrm{diag}(\nabla G_m(y_1, \underline{y}_{l-1}), \nabla G_v(y_1, \underline{y}_{l-1})) \\ & ((K*K)_0\left(\frac{y_{l-1} - y_1}{h}\right), (K*K)_1\left(\frac{y_{l-1} - y_1}{h}\right))^T], \end{aligned}$$

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 $(K*K)_i\left(\frac{y_{l-1}-y_1}{h}\right)$ in $\overline{J}(X_l)$ is actually a convolution kernel and behaves just like a one dimensional kernel function of y_{l-1}. This means that the standard method (CLT, or LLN) for univariate kernel estimates can be applied to show the asymptotics of

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

If we define $\widetilde{s}_{1n}$ as the remaining bias term of t_{1n}, the estimation errors of $\widehat{\varphi}_1(y_1)-\varphi_1(y_1)$, consist of two stochastic terms, $[I_2\quad e_1^TQ_n^{-1}]\left(\widetilde{t}_{1n}+\widetilde{t}_{2n}\right)$, and three bias terms, $[I_2\quad e_1^TQ_n^{-1}](s_{0n}+s_{1n}+s_{2n})$, where

$$\begin{aligned}\widetilde{t}_{2n}&=\frac{1}{n}\sum_{k=d+1}^{n'}K_h(y_{k-1}-y_1)\frac{p_2(\underline{y}_{k-1})}{p(X_k)}[H_2(\underline{y}_{k-1})\quad(1,\frac{Y_{k-1}-y_1}{h})^T],\\ s_{2n}&=t_{2n}-\widetilde{t}_{2n}.\end{aligned}$$

Step II: Computation of Variance and Bias

We start with showing the order of the main stochastic term,

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

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

$$\begin{aligned}\xi_{1k}&=\frac{p_2(\underline{y}_{k-1})}{p(y_{k-1},\underline{y}_{k-1})}\left\{\left[\begin{array}{c}\nabla G_m(y_1,\underline{y}_{k-1})v^{1/2}(X_k)\,\varepsilon_k\\ \nabla G_v(y_1,\underline{y}_{k-1})v(X_k)\,(\varepsilon_k^2-1))\end{array}\right]\quad\left[\begin{array}{c}\frac{1}{h}(K*K)_0\left(\frac{y_{k-1}-y_1}{h}\right)\\ 0\end{array}\right]\right\}\\ \xi_{2k}&=\frac{p_2(\underline{y}_{k-1})}{p(y_{k-1},\underline{y}_{k-1})}\left\{\left[\begin{array}{c}m_2\left(\underline{y}_{k-1}\right)\\ v_2\left(\underline{y}_{k-1}\right)\end{array}\right]\quad\left[\begin{array}{c}\frac{1}{h}K\left(\frac{y_{k-1}-y_1}{h}\right)\\ \frac{1}{h}K\left(\frac{y_{k-1}-y_1}{h}\right)\left(\frac{y_{l-1}-y_1}{h}\right)\end{array}\right]\right\},\end{aligned}$$

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

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

where the last equality comes from the stationarity assumption.

We claim that

(a) $h\text{var}\left(\xi_k\right) \longrightarrow \Sigma_1(y_1)$,

(b) $\sum_k \left[1-\frac{k}{n}\right] h\text{cov}\left(\xi_d, \xi_{d+k}\right) = o(1)$, and

(c) $nh\text{var}\left(\widetilde{t}_n^*\right) \longrightarrow \Sigma_1(y_1)$,

where

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

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

$$h\text{var}\left(\xi_k\right) = hE\left(\xi_{1k}\xi_{1k}^T\right) + hE\left(\xi_{2k}\xi_{2k}^T\right),$$

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

$$\begin{aligned}hE\left(\xi_{1k}\xi_{1k}^T\right) \quad = \quad & \{\int \frac{p_2^2(z_2)}{p(y_1,z_2)} \left[\begin{array}{cc} \nabla G_m(y_1,z_2)^2 v(y_1,z_2) & (\nabla G_m \cdot \nabla G_v)(\kappa_3 \cdot v^{3/2})(y_1,z_2) \\ (\nabla G_m \cdot \nabla G_v)(\kappa_3 \cdot v^{3/2})(y_1,z_2) & \nabla G_v(y_1,z_2)^2 \kappa_4(y_1,z_2) v^2(y_1,z_2) \end{array}\right] dz_2 \\ & \left[\begin{array}{cc} ||(K*K)_0||_2^2 & 0 \\ 0 & 0 \end{array}\right]\},\end{aligned}$$

and

$$hE\left(\xi_{2k}\xi_{2k}^T\right) = \int \frac{p_2^2(z_2)}{p(y_1,z)} \left[\begin{array}{cc} m_2^2\left(z_2\right) & m_2\left(z_2\right) v_2\left(z_2\right) \\ m_2\left(z_2\right) v_2\left(z_2\right) & v_2^2\left(z_2\right) \end{array}\right] dz_2 \quad \left[\begin{array}{cc} ||K||_2^2 & 0 \\ 0 & \int K^2\left(u\right) u^2 du \end{array}\right] + o\left(1\right),$$

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). Since $E\left(\xi_{1k}\xi_{1j}^T\right)|_{j\neq k} = E\left(\xi_{1k}\xi_{2j}^T\right)|_{j\neq k} = 0$, $\text{cov}\left(\xi_{d+1}, \xi_{d+1+k}\right) = \text{cov}\left(\xi_{2d+1}, \xi_{2d+1+k}\right)$. By setting $c(n)h \to 0$, as $n \to \infty$, we separate the covariance terms into two parts:

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

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

$$\begin{aligned} & \left|\text{cov}\left(\xi_{2d+1}, \xi_{2d+1+k}\right)\right| \\ \simeq \quad & |\int H_2\left(\underline{y}_d\right) H_2^T\left(\underline{y}_{d+k}\right) \frac{p(y_1, \underline{y}_d, y_1, \underline{y}_{d+k})}{p_{1|2}(y_1|\underline{y}_d) p_{1|2}(y_1|\underline{y}_{d+k})} d(\underline{y}_d, \underline{y}_{d+k})| \quad \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right].\end{aligned}$$

Therefore, it follows from the assumption on the boundedness condition in C.2 that

$$\begin{aligned}\left|\operatorname{cov}\left(\xi_{2d+1}, \xi_{2d+1+k}\right)\right| &\leq E\left|H_2\left(\underline{y}_d\right)\right| E\left|H_2^T\left(\underline{y}_{d+k}\right)\right| \int \frac{p(y_1, \underline{y}_d, y_1, \underline{y}_{d+k})}{p_{1|2}(y_1|\underline{y}_d)p_{1|2}(y_1|\underline{y}_{d+k})} d(\underline{y}_d, \underline{y}_{d+k}) \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\ &\equiv A^*.\end{aligned}$$

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

$$\begin{aligned}&\sum_{k=1}^{c(n)} \left[1 - \frac{k}{n}\right] h\operatorname{cov}\left(\xi_{2d+1}, \xi_{2d+1+k}\right) \\ \leq\ & 2c(n)\left|h\operatorname{cov}\left(\xi_{2d+1}, \xi_{2d+1+k}\right)\right| \leq 2c(n)hA^* \longrightarrow 0, \text{ as } n \to \infty.\end{aligned}$$

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

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

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

$$\left|h\operatorname{cov}\left(\xi_{2d+1}^i, \xi_{2d+1+k}^j\right)\right| = \left|\operatorname{cov}\left(\sqrt{h}\xi_{2d+1}^i, \sqrt{h}\xi_{2d+1+k}^j\right)\right| \leq 8\left[\alpha(k)^{1-2/v}\right]\left[\max_{i=1,\ldots,4} E\left(\sqrt{h}\left|\xi_{2k}^i\right|^v\right)\right]^{2/v},$$

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

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

$$\begin{aligned}E\left(\left|\sqrt{h}\xi_{2k}^1\right|^v\right) &\simeq \frac{h^{v/2}}{h^{v-1}} \int \frac{p_2^v(z_2)}{p^{v-1}(y_1, z_2)} \left|m_2^v\left(z_2\right)\right| dz_2 \\ &= O\left(\frac{h^{v/2}}{h^{v-1}}\right) = O\left(\frac{1}{h^{v/2-1}}\right).\end{aligned}$$

Thus, the covariance is bounded by

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

This implies

$$\begin{aligned}&\sum_{k=c(n)+1}^{n'} \left[1 - \frac{k}{n}\right] h\operatorname{cov}\left(\xi_{2d+1}, \xi_{2d+1+k}\right) \\ \leq\ & 2\sum_{k=c(n)+1}^{\infty} \left|h\operatorname{cov}\left(\xi_{2d+1}, \xi_{2d+1+k}\right)\right| \leq C'\left[\frac{1}{h^{1-2/v}}\right] \sum_{k=c(n)+1}^{\infty} \left[\alpha(k)^{1-2/v}\right] \\ =\ & C' \sum_{k=c(n)+1}^{\infty} \left[\frac{1}{h^{1-2/v}}\right]\left[\alpha(k)^{1-2/v}\right] \leq C' \sum_{k=c(n)+1}^{\infty} k^a \left[\alpha(k)^{1-2/v}\right],\end{aligned}$$

if a is such that

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

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,

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

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

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

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 kernel weighted sum of stationary series, we first get,

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

since

$$\begin{aligned}
& \frac{1}{n} \sum_{k=d+1}^{n'} K_h^{\widehat{W}}(y_{k-1} - y_1)\,(\frac{y_{k-1} - y_1}{h})^2[D^2\varphi_1(y_1) \quad (1, \frac{y_{k-1} - y_1}{h})^T] \\
\to & {}^{p} \int K_h^{\widehat{W}}(z_1 - y_1)\,(\frac{z_1 - y_1}{h})^2[D^2\varphi_1(y_1) \quad (1, \frac{z_1 - y_1}{h})^T] p\,(z)\,dz \\
\simeq & \int K_h\,(z_1 - y_1)\, p_2\,(z_2)\,(\frac{z_1 - y_1}{h})^2[D^2\varphi_1(y_1) \quad (1, \frac{z_1 - y_1}{h})^T]dz \\
= & \int K_h\,(z_1 - y_1)\,(\frac{z_1 - y_1}{h})^2[D^2\varphi_1(y_1) \quad (1, \frac{z_1 - y_1}{h})^T]dz_1 \\
= & [D^2\varphi_1(y_1) \quad \int K_h\,(z_1 - y_1)\,(\frac{z_1 - y_1}{h})^2(1, \frac{z_1 - y_1}{h})^T dz_1] \\
= & [D^2\varphi_1(y_1) \quad (\mu_K^2, 0)^T].
\end{aligned}$$

For the asymptotic bias of $\widetilde{s}_{1n}$, we again use the approximation results in Jones, Davies and Park(1994). Then, the first component of $\widetilde{s}_{1n}$, for example, is

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

and converges to

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

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

$$\widetilde{s}_{1n}\xrightarrow{p}\frac{h^2}{2}\int p_2(z_2)\{[\nabla G_m(m(y_1,z_2)),\nabla G_v(v(y_1,z_2))]^T\odot[\mu^2_{K*K}D^2\varphi_1(y_1)+\mu^2_K D^2\varphi_2(z_2)]\}\quad(1,0)^T dz_2.$$

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

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

Thus,

$$\widetilde{s}_{2n}=\frac{1}{n}\sum_{k=d+1}^{n'}K_h\left(y_{k-1}-y_1\right)[\frac{\overline{p}_2(\underline{y}_{k-1})}{\overline{p}(X_k)}-\frac{p_2(\underline{y}_{k-1})}{p(X_k)}][H_2(\underline{y}_{k-1})\quad(1,\frac{y_{k-1}-y_1}{h})^T]$$

$$\begin{aligned}
&\xrightarrow{p}\int K_h\left(z_1-y_1\right)[\frac{\overline{p}_2(z_2)}{\overline{p}(z)}-\frac{p_2(z_2)}{p(z)}][H_2(z_2)\quad(1,\frac{z_1-y_1}{h})^T]p\left(z\right)dz\\
\simeq&\quad\int K_h\left(z_1-y_1\right)\left[\frac{\overline{p}_2(z_2)}{p(z)}-\frac{p_2(z_2)\overline{p}(z)}{p^2\left(z\right)}\right][H_2(z_2)\quad(1,\frac{z_1-y_1}{h})^T]p\left(z\right)dz\\
=&\quad\int K_h\left(z_1-y_1\right)\left[\frac{\overline{p}_2(z_2)}{p(z)}-\frac{p_2(z_2)}{p\left(z\right)}\right][H_2(z_2)\quad(1,\frac{z_1-y_1}{h})^T]p\left(z\right)dz\\
&+\int K_h\left(z_1-y_1\right)\left[\frac{p_2(z_2)p(z)}{p^2(z)}-\frac{p_2(z_2)\overline{p}(z)}{p^2\left(z\right)}\right][H_2(z_2)\quad(1,\frac{z_1-y_1}{h})^T]p\left(z\right)dz\\
\simeq&\quad\frac{g^2}{2}[\int D^2p_2\left(z_2\right)H_2(z_2)dz_2\quad(\mu^2_K,0)^T]\\
&-\frac{g^2}{2}[\int\frac{p_2(z_2)}{p\left(y_1,z_2\right)}D^2p(y_1,z_2)H_2(z_2)dz_2\quad(\mu^2_K,0)^T].
\end{aligned}$$

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

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

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

$$\begin{aligned}\widehat{q}_{ni} \xrightarrow{p} \int K_h\left(z_1-y_1\right)\left(\frac{z_1-y_1}{h}\right)^i p_2\left(z_2\right)dz &= \int K\left(u_1\right)u_1^i du_1 \int p_2\left(z_2\right)dz_2 \\ &= \int K\left(u_1\right)u_1^i du_1 \equiv q_i,\end{aligned}$$

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

Thus, $Q_n \to \begin{bmatrix} 1 & 0 \\ 0 & \mu_K^2 \end{bmatrix}$, $Q_n^{-1} \to \frac{1}{\mu_K^2}\begin{bmatrix} \mu_K^2 & 0 \\ 0 & 1 \end{bmatrix}$, and $e_1^T Q_n^{-1} \to e_1^T$. Therefore,

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

Step III: Asymptotic Normality of $\widetilde{t}_n^*$

Applying the Cramer-Wold device, it is sufficient to show

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

for all $\beta \in \mathbb{R}^4$, where $\widetilde{\xi}_k = \beta^T\xi_k$. 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

$$\begin{aligned}\eta_j &= \sum_{t=j(r+s)}^{j(r+s)+r-1} \sqrt{h}\widetilde{\xi}_t, \quad \omega_j = \sum_{t=j(r+s)+r}^{(j+1)(r+s)-1} \sqrt{h}\widetilde{\xi}_t, \quad 0 \le j \le k-1, \\ \varsigma_k &= \sum_{t=k(r+s)}^{n} \sqrt{h}\widetilde{\xi}_t,\end{aligned}$$

then,

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

Due to C.6., there exist a sequence $a_n \to \infty$ such that

$$a_n s_n = o\left(\sqrt{nh}\right) \text{ and } a_n \sqrt{n/h} \alpha\left(s_n\right) \to 0, \; as \; n \to \infty, \tag{A.18}$$

and define the large block size as

$$r_n = \left[\frac{\sqrt{nh}}{a_n} \right]. \tag{A.19}$$

It is easy to show by (A.18) and (A.19) that as $n \to \infty:$

$$\frac{r_n}{n} \to 0, \; \frac{s_n}{r_n} \to 0, \; \frac{r_n}{\sqrt{nh}} \to 0, \tag{A.20}$$

and

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

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

$$\begin{aligned} \operatorname{var}\left(\omega_j\right) &= s \times \operatorname{var}\left(\sqrt{h}\widetilde{\xi}_t\right) + 2s \sum_{k=1,}^{s-1} \left(1 - \frac{k}{s}\right) \operatorname{cov}\left(\sqrt{h}\widetilde{\xi}_{d+1}, \sqrt{h}\widetilde{\xi}_{d+1+k}\right) \\ &= s\beta^T \Sigma_1 \beta \left(1 + o\left(1\right)\right), \end{aligned} \tag{A.21}$$

which implies

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

from the condition (A.20). Next, consider

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

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

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

The last equality also follows from Step II. Hence, $\frac{1}{n}E\left\{(S''_n)^2\right\} \to 0$, as $n \to \infty$. Repeating a similar argument for S'''_n, we get

$$\begin{aligned} \frac{1}{n}E\left\{(S'''_n)^2\right\} &\le \frac{1}{n}\left[n - k\left(r+s\right)\right]\operatorname{var}\left(\sqrt{h}\widetilde{\xi}_{d+1}\right) \\ &\quad +2\frac{n-k\left(r+s\right)}{n}\sum_{j=1}^{n-k(r+s)}\operatorname{cov}\left(\sqrt{h}\widetilde{\xi}_{d+1}, \sqrt{h}\widetilde{\xi}_{d+1+j}\right) \\ &\le \frac{r_n+s_n}{n}\beta^T\Sigma_1\beta + o\left(1\right) \\ &\to 0, \quad \text{as } n \to \infty. \end{aligned}$$

Now, it remains to show $\frac{1}{\sqrt{n}}S'_n = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j \xrightarrow{\mathcal{D}} N\left(0, \beta^T\Sigma_1\beta\right)$.

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

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

where the last two equalities follows from hold (A.20). Thus, the summands $\left\{\eta_j\right\}$ in S'_n are asymptotically independent. Since the similar operation to (A.21) yields

$$\operatorname{var}\left(\eta_j\right) = r_n\beta^T\Sigma_1\beta\left(1 + o\left(1\right)\right),$$

and hence

$$\operatorname{var}(\frac{1}{\sqrt{n}}S'_n) = \frac{1}{n}\sum_{j=0}^{k-1}E\left(\eta_j^2\right) = \frac{k_nr_n}{n}\beta^T\Sigma_1\beta\left(1 + o\left(1\right)\right) \to \beta'\Sigma^*\beta.$$

Finally, due to the boundedness of density and kernel functions, the Lindeberg-Feller condition for the asymptotic normality of S'_n holds,

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

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

From $e_1^TQ_n^{-1} \xrightarrow{p} e_1^T$, the Slutzky Theorem implies $\sqrt{nh}[I_2 \quad e_1^TQ_n^{-1}]\widetilde{t}_n^* \xrightarrow{d} N\left(0, \Sigma_1^*\right)$, where $\Sigma_1^* = [I_2 \quad e_1^T]\Sigma_1[I_2 \quad e_1]$. In sum, $\sqrt{nh}(\widehat{\varphi}_1(y_1) - \varphi_1(y_1) - B_n) \xrightarrow{d} N\left(0, \Sigma_1^*\right)$, with $\Sigma_1^*(y_1)$ given by

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

■

Lemma A.1 *Assume the conditions in C.1, C.4. through C.6. For a bounded function, $F(\cdot)$, it holds that*

$$
\begin{aligned}
(a)\ r_{1n} &= \frac{\sqrt{h}}{\sqrt{n}} \sum_{k=d}^{n} K_h\left(y_{k-1}-y_1\right)\left(\widehat{p}_2(\underline{y}_{k-2}) - \overline{p}_2(\underline{y}_{k-2})\right)F\left(x_k\right) = o_p\left(1\right), \\
(b)\ r_{2n} &= \frac{\sqrt{h}}{\sqrt{n}} \sum_{k=d}^{n} K_h\left(y_{k-1}-y_1\right)\left(\widehat{p}(x_k) - p(x_k)\right)F\left(x_k\right) = o_p\left(1\right)
\end{aligned}
$$

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

$$
\begin{aligned}
\xi_{1n} &= \frac{1}{n^2} \sum_{k=d}^{n} \sum_{l=d}^{n} K_h\left(y_{k-1}-y_1\right)F\left(x_k\right)\left[L_g\left(\underline{y}_{l-2} - \underline{y}_{k-2}\right) - \overline{L}_{l|k}\left(y_{l-2}|y_{k-2}\right)\right] \\
\xi_{2n} &= \frac{1}{n^2} \sum_{k} \sum_{l} K_h\left(y_{k-1}-y_1\right)F\left(x_k\right)\left[\overline{L}_{l|k}\left(y_{l-2}|y_{k-2}\right) - \overline{p}_2(\underline{y}_{k-2})\right]
\end{aligned}
$$

Rewrite ξ_{2n} as

$$
\begin{aligned}
&\frac{1}{n^2} \sum_{k} \sum_{s<k^*(n)} K_h\left(y_{k-1}-y_1\right)F\left(x_k\right)\left[\overline{L}_{k+s|k}\left(y_{k+s-2}|y_{k-2}\right) - \overline{p}_2(\underline{y}_{k-2})\right] \\
&+\frac{1}{n^2} \sum_{k} \sum_{s\geq k^*(n)} K_h\left(y_{k-1}-y_1\right)F\left(x_k\right)\left[\overline{L}_{k+s|k}\left(y_{k+s-2}|y_{k-2}\right) - \overline{p}_2(\underline{y}_{k-2})\right],
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
&\sup_{s\geq k^*(n)} \left|p_{k+s|k}(y_{k+s-2}|y_{k-2}) - p\left(y_{k-2}\right)\right| \frac{1}{n}\sum_{k}^{n} K_h\left(y_{k-1}-y_1\right)\left|F\left(x_k\right)\right| \\
\leq\ &\rho^{k(n)}O_p(1).
\end{aligned}
$$

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

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

$$
\begin{aligned}
E\left(\xi_{1n}^2\right) &= \frac{1}{n^4} \sum_{k\neq l}^{n} \sum^{n} \sum_{i\neq j}^{n} \sum^{n} E\{K_h\left(y_{k-1}-y\right)K_h\left(y_{i-1}-y\right)F\left(x_k\right) \\
&\quad F\left(X_l\right)\left[L_g\left(\underline{y}_{l-2} - \underline{y}_{k-2}\right) - \overline{L}_{l|k}(\underline{y}_{k-2})\right]\left[L_h\left(\underline{y}_{j-2} - \underline{y}_{i-2}\right) - \overline{L}_{j|i}(\underline{y}_{i-2})\right]\}.
\end{aligned}
$$

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

$$\begin{aligned}
&\frac{1}{n^4}\sum_{k}^{n}\sum_{l\neq j}^{n}\sum^{n} E\{K_h^2\left(y_{k-1}-y\right)F^2\left(x_k\right) \\
&\left[L_g\left(\underline{y}_{l-2}-\underline{y}_{k-2}\right)-\overline{L}_{l|k}(\underline{y}_{k-2})\right]\left[L_g\left(\underline{y}_{j-2}-\underline{y}_{k-2}\right)-\overline{L}_{j|k}(\underline{y}_{k-2})\right]\} \\
= \;& 0,
\end{aligned}$$

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

$$\begin{aligned}
&E_{|k,l}\left[L_g\left(\underline{y}_{j-2}-\underline{y}_{k-2}\right)-\overline{L}_{j|k}(\underline{y}_{k-2})\right] \\
= \;& E_{|k}\left[L_g\left(\underline{y}_{j-2}-\underline{y}_{k-2}\right)-\overline{L}_{j|k}(\underline{y}_{k-2})\right]=E_{|k}\left[L_g\left(\underline{y}_{j-2}-\underline{y}_{k-2}\right)\right]-\overline{L}_{j|k}(\underline{y}_{k-2})=0
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{n^4}\sum_{k\neq i}^{n}\sum^{n}\sum_{l}^{n} E\{K_h\left(y_{k-1}-y\right)K_h\left(y_{i-1}-y\right)F(x_k)F(x_i)\times \\
&\qquad\left[L_g\left(\underline{y}_{l-2}-\underline{y}_{k-2}\right)-\overline{L}_{l|k}(\underline{y}_{k-2})\right]\left[L_g\left(\underline{y}_{l-2}-\underline{y}_{i-2}\right)-\overline{L}_{l|i}(\underline{y}_{i-2})\right]\}
\end{aligned}$$

We only calculate

$$\frac{1}{n^4}\sum_{k\neq i}^{n}\sum^{n}\sum_{l}^{n} E\{K_h\left(y_{k-1}-y\right)K_h\left(y_{i-1}-y\right)L_g\left(\underline{y}_{l-2}-\underline{y}_{k-2}\right)L_g\left(\underline{y}_{l-2}-\underline{y}_{i-2}\right)F(x_k)F(x_i)\} \tag{A.22}$$

since the rest of the triple sum consist of expectations of standard kernel estimates and are $O\left(1/n\right)$. Note that

$$\begin{aligned}
&E_{|(i,k)}L_g\left(\underline{y}_{l-2}-\underline{y}_{k-2}\right)L_g\left(\underline{y}_{l-2}-\underline{y}_{i-2}\right) \\
\simeq \;& (L*L)_g\left(\underline{y}_{k-2}-\underline{y}_{i-2}\right)p_{l|(k,i)}\left(\underline{y}_{k-2}|\underline{y}_{k-2},\underline{y}_{i-2}\right),
\end{aligned}$$

where $(L*L)_g\left(\cdot\right)=(1/g)\int L\left(u\right)L\left(u+\cdot/g\right)$ is a convolution kernel. Thus, (A.22) is

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

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

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

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

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

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,

$$||s^*(z,\theta,\gamma_\alpha)-s^*(z,\theta,\gamma_\alpha^0)-S^*(z,\theta,\gamma_\alpha^0(\underline{y}_\alpha))(\gamma_\alpha-\gamma_\alpha^0)|| \leq b(z)||\gamma_\alpha-\gamma_\alpha^0||^2, \qquad \text{(A.23)}$$

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

Assumption A

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

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

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 identification condition of component functions, while 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 C in section 4. Let h_0 and h denotes the bandwidth parameter used for the preliminary IV and the two step estimates, respectively, while g is the bandwidth parameter for the kernel density.

Assumption B

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

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 q-th 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, i.e., $\int u^i K(u)\,du = \int u^i L(u)\,du = 0$, for $i = 1, \ldots, q-1$. Also, the kernel functions are q-times differentiable with Lipschitz continuous q-th 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)

$$\frac{nh_0^2}{(\log n)^2 h} \to \infty, \text{ and } \sqrt{nh}h_0^q \to 0,$$

and for some integer $\omega > d/2$,

$$\text{(ii) } n(h_0 h)^{2\omega+1}/\log n \to \infty;\ h_0^{q-\omega} h^{-\omega-1/2} \to 0$$

$$\text{(iii) } nh_0^{d+(4\omega+1)}/\log n \to \infty;\ q \ge 2\omega + 1.$$

Some facts about empirical processes are useful in the sequel. Define the L^2-Sobolev norm (of order q) on the class of real-valued function with domain $\mathcal{W}_0$,

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

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, e.g., Stein (1970), and $\mathcal{X} = \times_{\beta=1}^{d}\mathcal{X}_\beta$, with $\mathcal{X}_{\underline{\alpha}} = \times_{\beta=1(\neq\alpha)}^{d}\mathcal{X}_\beta$. Define $\mathcal{T}_1$ be a class of smooth functions on $\mathcal{X}_{\underline{\alpha}} = \times_{\beta=1(\neq\alpha)}^{d}\mathcal{X}_\beta$ whose L^2-Sobolev norm is bounded by some constant; $\mathcal{T}_1 = \{\tau : ||\tau||_{q,2,\mathcal{X}_{\underline{\alpha}}} \leq C\}$. 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 $\tau \in \mathcal{T}_1$:

$$v_{1n}(\tau_1) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[f_1(x_t;\tau_1) - Ef_1(x_t;\tau_1)\right], \tag{A.24}$$

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 $f_1(w;\tau) = h^{-1/2}K(\frac{w_\alpha - y_\alpha}{h})\underline{S}^*(w,\gamma_\alpha^0)\tau_1(\underline{w}_\alpha)$;

and (ii) an empirical process, $v_{2n}(\cdot,\cdot)$, indexed by $(y_\alpha,\tau_2) \in \mathcal{X}_\alpha \times \mathcal{T}_2$:

$$v_{2n}(y_\alpha,\tau_2) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[f_2(x_t;y_\alpha,\tau_2) - Ef_2(x_t;y_\alpha,\tau_2)\right], \tag{A.25}$$

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

We say that the process $\{\nu_{1n}(\cdot)\}$ 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

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

$$\overline{\lim_{T\to\infty}}\, P^*\left[\sup_{\rho_1(\tau,\tau_0)<\delta}|\nu_{1n}(\tau_1) - \nu_{1n}(\tau_1^0)| > \eta\right] < \varepsilon, \tag{A.26}$$

and

$$\overline{\lim_{T\to\infty}}\, P^*\left[\sup_{\rho_2((y_\alpha,\tau_2),(y_\alpha^0,\tau_2^0))<\delta}|\nu_{2n}(y_\alpha,\tau_2) - \nu_{2n}(y_\alpha^0,\tau_2^0)| > \eta\right] < \varepsilon, \tag{A.27}$$

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 above. Note that (A.26) follows, if Pollard's entropy condition is satisfied by $\mathcal{F}_1$ with some square integrable envelope $\overline{F}_1$;see Pollard (1990) for more details. Since $f_1(w;\tau_1) = c_1(w)\tau_1(\underline{w}_\alpha)$ 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 $c(w) = [h^{-1/2}K(\frac{w_\alpha - y_\alpha}{h})\underline{S}^*(w,\gamma_\alpha^0)]$, the entropy condition is verified by Theorem 2 in Andrews (1994) on a class of functions of type III. Square integrability of the envelope $\overline{F}_1$ comes from Assumption A(ii). In a similar way, we can show (A.27), by applying the " mix and match" argument of Theorem 3 in Andrews (1994) to $f_2(w;y_\alpha,\tau_2) = c_2(w)h^{-1/2}K(\frac{w_\alpha - y_\alpha}{h_0})\tau_2(w)$, where $K(\cdot)$ is Lipschitz continuous in y_α, i.e., a function of type II.

Proof of Theorem 4. We only give a sketch, since the whole proof is lengthy and relies on the similar arguments to Andrews (1994) or Gozalo and Linton (1995) for i.i.d case. Expanding the FOC in (5.15) and solving for $(\widehat{\theta}^* - \theta_0)$ yields

$$\widehat{\theta}^* - \theta_0 = -[\frac{1}{n}\sum_{t=d+1}^{n'} \nabla_\theta s(z_t,\overline{\theta},\widetilde{\gamma}_\alpha)]^{-1}\frac{1}{n}\sum_{t=d+1}^{n'} s(z_t,\widetilde{\gamma}_\alpha),$$

where $\overline{\theta}$ is the mean value between $\widehat{\theta}$ and θ_0, and $s(z_t,\widetilde{\gamma}_\alpha) = s(z_t,\theta_0,\widetilde{\gamma}_\alpha)$. 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 $\widetilde{\gamma}_\alpha$ by Lemma A.3, and (ii) uniform continuity of the localized likelihood function, $Q_n(\theta,\gamma_\alpha)$ over $\Theta \times \Gamma_\alpha$, yields $\sup_{\theta\in\Theta}|\widetilde{Q}_n(\theta) - E(Q_n(\theta))| \xrightarrow{p} 0$, and thus consistency of $\widehat{\theta}^*$. Based on the ergodic theorem on the stationary time series and a similar argument to Theorem 1 in Andrews (1994), consistency of $\widehat{\theta}^*$ as well as uniform convergence of $\widetilde{\gamma}_\alpha$ imply

$$\frac{1}{n}\sum_{t=d+1}^{n'} \nabla_\theta s(z_t,\overline{\theta},\widetilde{\gamma}_\alpha) \xrightarrow{p} E[\nabla_\theta s(z_t,\theta_0,\gamma_\alpha^0)] \equiv D_\alpha(y_\alpha), \tag{A.28}$$

For the numerator, we first linearize the score function. Under Assumption A(ii), $s^*(z,\theta,\gamma_\alpha)$ is Frechet differentiable and (A.23) holds, which, due to $\sqrt{nh}||\widetilde{\gamma}_\alpha - \gamma_\alpha^0||_\infty^2 \xrightarrow{p} 0$ (by Lemma A.3 and B.8(i)), yields a proper linearization of the score term;

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

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

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

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.28). Due to the asymptotic boundedness of (A.28), it suffices to show the negligibility of the second and third terms.

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

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

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

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

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

$$\rho^2((y_\alpha, \widehat{\delta}), (y_\alpha, 0)) \leq Ch^{-(2\omega+1)}||\widetilde{\gamma}_\alpha - \gamma^0_\alpha||^2_{\omega,2,\mathcal{X}_{\overline{\alpha}}} = o_p(1),$$

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

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

We proceed to show negligibility of T^*_{n2}. From integrability condition on $S^*(z, \gamma^0_\alpha(\underline{y}_\alpha))$, it follows, by change of variables and the dominated convergence theorem, that $\int K_h(y_\alpha - y^0_\alpha)S^*(z, \gamma^0_\alpha(\underline{y}_\alpha))dF_0(z) =$

$\int S^*[(y,y_\alpha^0,\underline{y}_\alpha),\gamma_\alpha^0(\underline{y}_\alpha)]p(y,y_\alpha^0,\underline{y}_\alpha)d(y,\underline{y}_\alpha)<\infty$, which, together with $\sqrt{n}$-consistency of $\widehat{c}=(\widehat{c}_m,\widehat{c}_v)'$, means that $(\widehat{c}-c)\sqrt{nh}\int K_h(y_\alpha-y_\alpha^0)S^*(z,\gamma_\alpha^0(\underline{y}_\alpha))dF_0(z)=o_p(1)$. Since

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

this yields

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

From Lemma A.3,

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

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

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

where $\mathcal{S}_{1n}=$

$$\sum_{\beta=1,\neq\alpha}^{d}\sqrt{nh}\int K_h(y_\alpha-y_\alpha^0)\underline{S}^*(x,\gamma_\alpha^0)\frac{1}{n}\sum_t (K*K)_{h_0}(y_{t-\beta}-y_\beta)\frac{p_2(\underline{y}_{t-\beta})}{p(x_t)}(\nabla G(x_\beta,\underline{y}_{t-\beta})\odot\xi_t)p(x)dx,$$

and

$$\mathcal{S}_{2n}=\sum_{\beta=1,\neq\alpha}^{d}\sqrt{nh}\int K_h(y_\alpha-y_\alpha^0)\underline{S}^*(x,\gamma_\alpha^0)\frac{1}{n}\sum_t K_{h_0}(y_{t-\beta}-y_\beta)\frac{p_2(\underline{y}_{t-\beta})}{p(x_t)}\gamma_\alpha^{*0}(\underline{y}_{t-\beta})p(x)dx.$$

Let $\mathcal{S}_{1n}^i$ and $\mathcal{S}_{2n}^i$ be the i-th element 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

$$\begin{aligned}
\mathcal{S}_{1n}^{i} &= \frac{\sqrt{h}}{\sqrt{n}}\sum_{t}\frac{p_2(\underline{y}_{t-\beta})}{p(x_t)}v\left(x_t\right)^{1/2}\varepsilon_t \times \\
&\quad \left[\int K_h(y_\alpha-y_\alpha^0)\underline{S}^{i1*}(x,\gamma_\alpha^0)\sum_{\beta=1,\neq\alpha}^{d}\left(K*K\right)_{h0}\left(y_\beta-y_{t-\beta}\right)\nabla G_m(y_\beta,\underline{y}_{t-\beta})p(x)dx\right] \\
&\quad +\frac{\sqrt{h}}{\sqrt{n}}\sum_{t}\frac{p_2(\underline{y}_{t-\beta})}{p(x_t)}v\left(x_t\right)\left(\varepsilon_t^2-1\right)\times \\
&\quad \left[\int K_h(y_\alpha-y_\alpha^0)\underline{S}^{i2*}(x,\gamma_\alpha^0)\sum_{\beta=1,\neq\alpha}^{d}\left(K*K\right)_{h0}\left(y_\beta-y_{t-\beta}\right)\nabla G_v(y_\beta,\underline{y}_{t-\beta})p(x)dx\right] \\
&= \frac{\sqrt{h}}{\sqrt{n}}\sum_{t}\frac{p_2(\underline{y}_{t-\beta})}{p(x_t)}\left[v\left(x_t\right)^{1/2}\varpi_{i1}^{1}(x_t)\varepsilon_t+v\left(x_t\right)\varpi_{i2}^{1}(x_t)(\varepsilon_t^2-1)\right]+o_p(1),
\end{aligned}$$

where

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

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

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

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 above, we get $\mathcal{S}_{2n}^{i}=o_p(1)$, and $\mathcal{S}_{2n}=o_p(1)$, since $E(m_{\underline{\alpha}}(\underline{y}_{t-\alpha}))=E(v_{\underline{\alpha}}(\underline{y}_{t-\alpha}))=0$.

We finally show negligibility of the last term;

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

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

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

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

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

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

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

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

Lemma A.2 (Masry, 1996) Suppose that Assumption B hold. Then, for any vector $\mu = (\mu_1, \ldots, \mu_d)'$ with $|\mu| = \Sigma_j \mu_j \leq \omega$,

$$(a) \;\; \sup_{x \in \mathcal{X}} \left| D_x^\mu \widehat{p}\,(x) - D_x^\mu p\,(x) \right| = O_p\Big(\sqrt{\frac{\log n}{n g^{(2|\mu|+d)}}}\Big) + O_p(g^{q-|\mu|})$$

$$(a) \;\; \sup_{x \in \mathcal{X}} \left| D_x^\mu \widetilde{m}\,(x) - D_x^\mu m\,(x) \right| = O_p\Big(\sqrt{\frac{\log n}{n h_0^{(2|\mu|+d)}}}\Big) + O_p(h_0^{q-|\mu|}) \equiv \rho_n(\mu)$$

$$(c) \;\; \sup_{x \in \mathcal{X}} \left| \widetilde{m}\,(x) - m\,(x) - \widetilde{L}(x) \right| = O_p(\rho_n^2), \text{ where}$$

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

Lemma A.3 Suppose that Assumption B hold. Then, for any vector $\mu = (\mu_1, .., \mu_d)'$ with $|\mu| = \Sigma_j \mu_j \leq \omega$,

$$(a) \;\; \sup_{x_\alpha \in \mathcal{X}_\alpha} \left| D^\mu \widehat{\varphi}_\alpha(y_\alpha) - D^\mu \varphi_\alpha(y_\alpha) \right| = O_p\Big(\sqrt{\frac{\log n}{n h^{(2|\mu|+1)}}}\Big) + O_p(h^{q-|\mu|}) + O_p(\rho_n^2(\mu))$$

$$(b)\ \sup_{x_\alpha \in \mathcal{X}_\alpha} |\widehat{\varphi}_\alpha(y_\alpha) - \varphi_\alpha(y_\alpha) - \widehat{L}_\varphi(y_\alpha)| = O_p(\rho_n^2) + o_p(n^{-1/2}),$$

where

$$\begin{aligned} \widehat{L}_\varphi(y_\alpha) &= \frac{1}{n}\sum_t (K*K)_h\,(y_{t-\alpha} - y_\alpha)\frac{p_{\overline{\alpha}}(\underline{y}_{t-\alpha})}{p(x_t)}\begin{bmatrix} G_m'(m(y_\alpha, \underline{y}_{t-\alpha}))v\,(x_t)^{1/2} \\ G_v'(v(y_\alpha, \underline{y}_{t-\alpha})v\,(x_t)) \end{bmatrix}\begin{bmatrix} \varepsilon_t, \\ \varepsilon_t^2 - 1 \end{bmatrix} \\ &+ \frac{1}{n}\sum_t K_h\,(y_{t-\alpha} - y_\alpha)\frac{p_2(\underline{y}_{t-\alpha})}{p(x_t)}\left[m_{\underline{\alpha}}\left(\underline{y}_{t-\alpha}\right), v_{\underline{\alpha}}\left(\underline{y}_{t-\alpha}\right)\right]^T + h^q\overline{b}_\alpha(y_\alpha) \end{aligned}$$

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 IV estimates,

$$\widehat{\varphi}_\alpha(y_\alpha) - \varphi_\alpha(y_\alpha) = [I_2 \quad e_1^T Q_n^{-1}]\tau_n,$$

where $Q_n = [\widehat{q}_{ni+j-2}(y_\alpha)]_{(i,j)=1,2}$, with $\widehat{q}_{ni} = \frac{1}{n}\sum_{t=d}^n K_h\,(y_{t-\alpha} - y_\alpha)\frac{\widehat{p_{\overline{\alpha}}}(\underline{y}_{t-\alpha})}{\widehat{p}(x_t)}\left(\frac{y_{t-\alpha} - y_\alpha}{h}\right)^i$, for $i = 0, 1, 2$, and $\tau_n = \frac{1}{n}\sum_t K_h\,(y_{t-\alpha} - y_\alpha)\frac{\widehat{p_{\overline{\alpha}}}(\underline{y}_{t-\alpha})}{\widehat{p}(x_t)}\,[\widetilde{z}_t - \varphi_\alpha(y_\alpha) - (y_{t-\alpha} - y_\alpha)\nabla\varphi_\alpha(y_\alpha)] \quad (1, \frac{y_{t-\alpha} - y_\alpha}{h})^T$. By C-S inequality and Lemma A.2. applied with Taylor expansion, it holds that

$$\begin{aligned} &\sup_{x_\alpha \in \mathcal{X}_\alpha} \frac{1}{n}\sum_{t=d}^n K_h\,(y_{t-\alpha} - y_\alpha)\left[\frac{\widehat{p_{\overline{\alpha}}}(\underline{y}_{t-\alpha})}{\widehat{p}(x_t)} - \frac{p_{\overline{\alpha}}(\underline{y}_{t-\alpha})}{p(x_t)}\right]\left(\frac{y_{t-\alpha} - y_\alpha}{h}\right)^i \\ \le\ &\sup_{x \in \mathcal{X}}\left|\frac{\widehat{p_{\overline{\alpha}}}(\underline{y}_\alpha)}{\widehat{p}(x)} - \frac{p_{\overline{\alpha}}(\underline{y}_\alpha)}{p(x)}\right| \sup_{x_\alpha \in \mathcal{X}_\alpha} \frac{1}{n}\sum_{t=d}^n K_h\,(y_{t-\alpha} - y_\alpha)\left|\frac{y_{t-\alpha} - y_\alpha}{h}\right|^i \\ =\ &O_p(\sup_{x \in \mathcal{X}}|\widehat{p}(x) - p(x)|) \equiv O_p(\rho_n), \end{aligned}$$

where the boundedness condition of B.2 is used for the last line. Hence, the standard argument of Masry (1996), implies that $\sup_{x_\alpha \in \mathcal{X}_\alpha}|\widehat{q}_{ni} - q_i| = o_p(1)$, 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, i.e., $e_1^T Q_n^{-1} \xrightarrow{p} e_1^T$, uniformly in $y_\alpha \in \mathcal{X}_\alpha$. For the numerator, we show the uniform convergence rate of the first element of τ_n, since the other terms can be treated in the same way. Let τ_n^1 denote the first element of τ_n, i.e.,

$$\tau_n^1 = \frac{1}{n}\sum_t K_h\,(y_{t-\alpha} - y_\alpha)\frac{\widehat{p_{\overline{\alpha}}}(\underline{y}_{t-\alpha})}{\widehat{p}(x_t)}\left[G_m(\widetilde{m}(x_t)) - M_\alpha(y_\alpha) - (y_{t-\alpha} - y_\alpha)m_\alpha'(y_\alpha)\right],$$

or alternatively,

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

where

$$\begin{aligned} r(x_t; \mathbf{g}) &= \frac{g_2(\underline{y}_{t-\alpha})}{g_3(x_t)}\,[G_m(g_1(x_t)) - M_\alpha(y_\alpha) - (y_{t-\alpha} - y_\alpha)m_\alpha'(y_\alpha)] \\ \mathbf{g}(x_t) &= [g_1(x_t), g_2(\underline{y}_{t-\alpha}), g_3(x_t)] = [m(x_t), p_{\overline{\alpha}}(\underline{y}_{t-\alpha}), p(x_t)] \\ \widehat{\mathbf{g}} &= \widehat{\mathbf{g}}(x_t) = (\widetilde{m}(x_t), \widehat{p_{\overline{\alpha}}}(\underline{y}_{t-\alpha}), \widehat{p}(x_t)] \end{aligned}$$

Since $p_{\overline{\alpha}}(\cdot)/p(\cdot)$ 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 $R(x_t;\mathbf{g}) = \frac{\partial r(x_t;\mathbf{g})}{\partial \mathbf{g}}\downarrow_{\mathbf{g}=\mathbf{g}(x_t)}$. 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

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

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

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

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

$$\begin{aligned}\tau_n^1 &= \frac{1}{n}\sum_t K_h\left(y_{t-\alpha}-y_\alpha\right)\varsigma_t[1+O_p(\rho_n)] \\ &+\frac{1}{n}\sum_t K_h\left(y_{t-\alpha}-y_\alpha\right)\frac{p_{\overline{\alpha}}(\underline{y}_{t-\alpha})}{p(x_t)}G_m'\left(m\left(x_t\right)\right)\left[\widetilde{m}\left(x_t\right)-m\left(x_t\right)\right] \\ &+\frac{h^q}{q!}\mu_q(k)b_{1\alpha}(y_\alpha)+o_p(h^q), \qquad \text{(A.29)}\end{aligned}$$

where $\varsigma_t = \frac{p_2(\underline{y}_{t-\alpha})}{p(x_t)}M_{\overline{\alpha}}(\underline{y}_{t-\alpha})$, and $M_{\overline{\alpha}}(\underline{y}_{t-\alpha}) = \Sigma_{1\leq\beta\leq d,(\neq\alpha)} m_\beta\left(\underline{y}_{t-\alpha}\right)$. Note that due to the identification condition, $E[\varsigma_t|y_{t-\alpha}] = 0$, and consequently, the first term is of 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 equicontinuiuty argument to the empirical process $\{v_n(\cdot,\cdot)\}$, indexed by $(y_\alpha,\delta)\in\mathcal{X}_\alpha\times\mathcal{T}$, with $\mathcal{T}=\{\delta : ||\delta||_{\omega,2,\mathcal{X}_\alpha}\leq C\}$, 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 $f(x_t;y_\alpha,\delta) = K(\frac{y_{t-\alpha}-y_\alpha}{h})h^\omega\frac{p_{\overline{\alpha}}(\underline{y}_{t-\alpha})}{p(x_t)}G_m'\left(m\left(x_t\right)\right)\delta(y_{t-\alpha})$, for some positive integer $\omega > d/2$. Let $\widetilde{\delta} = h^{-\omega-1/2}[\widetilde{m}\left(x_t\right)-m\left(x_t\right)]$. By the same similar argument in the proof of Theorem 5.2, it holds under B.8(iii) and Lemma A.2 that $||\widetilde{\delta}||_{\omega,2,\mathcal{X}} = O_p(\ h^{-\omega-1/2}[\sqrt{\frac{\log n}{nh^{(2\omega+d)}}}+h^{q-\omega}])= o_p(1)$, leading to (i) $\Pr(\widehat{\delta}\in\mathcal{T})\to 1$ and (ii) $\rho((y_\alpha,\widehat{\delta}),(y_\alpha,\delta^0))\xrightarrow{p} 0$, where $\delta^0=0$. These conditions and stochastic equicontinuity of $v_n(\cdot,\cdot)$ at (y_α,δ^0) yields $\sup_{y_{\alpha\in\mathcal{X}_\alpha}}|v_n(y_\alpha,\widehat{\delta})-\nu_n(y_\alpha,\delta^0| = \sup_{x_{\alpha\in\mathcal{X}_\alpha}}|v_n(y_\alpha,\widehat{\delta})| = o_p(1)$. Thus, the second term of τ_n^1 is approximated with an $o_p(1/\sqrt{n})$ error (uniform in y_α) by

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

which, by substituting $\widetilde{L}(x_t)$ for $\widetilde{m}\left(x_t\right) - m\left(x_t\right)$, is given by

$$\begin{aligned}&\frac{1}{n}\sum_{s}\left(K*K\right)_h\left(\frac{y_{s-\alpha}-y_\alpha}{h}\right)p_{\overline{\alpha}}(\underline{y}_{s-\alpha})G_m'(m(y_\alpha,\underline{y}_{s-\alpha}))\frac{v^{1/2}\left(x_s\right)\varepsilon_s}{p(x_s)} \\ &+\frac{h^q}{q!}\mu_q(k)b_{2\alpha}(y_\alpha),\end{aligned} \tag{A.30}$$

where $(K*K)(\cdot)$ is actually a convolution kernel as defined before. Hence, by letting $\overline{b}_\alpha(y_\alpha)$ summarize two bias terms appearing in (A.29) and (A.30), Lemma A.3(b) is shown. The uniform convergence results in part (a) then follow by the standard arguments of Masry(1996), since two stochastic terms in the asymptotic expansion of $\widehat{\varphi}_\alpha(y_\alpha) - \varphi_\alpha(y_\alpha)$ consist of only univariate kernels. ■

References

[1] 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, New York: North Holland.

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

[3] Auestadt, B. and D. Tjøstheim, (1991). Functional identification in nonlinear time series. In *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, Kluwer Academic: Amsterdam. pp 493–507.

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

[5] Black, F. (1989). Studies of stock price volatility changes. *proceedings from the American Association, Business and Economics Section,* 177-181.

[6] Buja, A., T. Hastie, and R. Tibshirani, (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453-555.

[7] Cai, Z., and E. Masry (2000). Nonparametric Estimation of Additive Nonlinear ARX Time Series: Local Linear Fitting and Projections. *Econometric Theory* 16, 465-501.

[8] Carrasco, M., and X. Chen (2002). Mixing and moment properties of various GARCH and Stochastic Volatility Models. Econometric Theory 18, 17-39.

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

[10] Chen, R., and R.S. Tsay, (1993a). Nonlinear additive ARX models, *Journal of the American Statistical Association* **88**, 955-967.

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

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

[13] Fan, J. (1992). Design-adaptive nonparametric regression. *J. Am. Statist Soc.* **82**, 998-1004.

[14] Fan, J. and Q. Yao (1996). Efficient estimation of conditional variance functions in stochastic regression. *Biometrica* **85**, 645-660.

[15] Granger, C. and T. Teräsvirta, (1993). *Modeling Nonlinear Economic Relationships,* Oxford University Press, Oxford.

[16] Hall, P. and C. Heyde (1980). *Martingale Limit Theory and Its Application.* New York: Academic Press.

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

[18] Härdle, W. and A.B. Tsybakov, (1997). Locally polynomial estimators of the volatility function. *Journal of Econometrics* , **81**, 223-242.

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

[20] Härdle, W. and P. Vieu, (1991). Kernel regression smoothing of time seires. *Journal of Time Series Analysis, **13**, 209-232.*

[21] Hastie, T. and R. Tibshirani, (1990). *Generalized Additive Models.* Chapman and Hall, London.

[22] Hastie, T. and R. Tibshirani, (1987) Generalized additive models: Some applications. *J. Am. Statist Soc.* **82** 371- 386.

[23] Horowitz, J., (2001). Estimating generalized additive models. *Econometrica* 69, 499-513.

[24] Jones, M.C., S.J. Davies, and B.U. Park, (1994). Versions of kernel-type regression estimators. *J. Am. Statist Soc.* **89**, 825-832.

[25] Kim, w. and O. Linton, and 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.

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

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

[28] Linton, O.B. and W. Härdle, (1996). Estimating additive regression models with known links. *Biometrika* **83**, 529-540.

[29] Linton, O.B. and J.B. Nielsen, (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-100.

[30] Linton, O.B., Nielsen, J.P., and S. van de Geer, (2003). Estimating Multiplicative and Additive Hazard functions by Kernel Methods. Annals of Statistics 31, 2.

[31] Linton, O.B., N. Wang, R. Chen, and W. Härdle, (1995). An analysis of transformation for additive nonparametric regression. *Journal of the American Statistical Association **92,*** 1512-21.

[32] Mammen, E., O.B. Linton, and J. Nielsen, (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics,* **27(**5), 1443-1490.

[33] Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571-599.

[34] Masry, E., and D. Tjøstheim, (1995). Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality. *Econometric Theory* **11**, 258-289.

[35] Masry, E., and D. Tjøstheim, (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory* 13, 214-252.

[36] Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**, 347-370.

[37] Newey, W.K. (1994). Kernel estimation of partial means. *Econometric Theory.* **10**, 233-253.

[38] Opsomer, J. D., and D. Ruppert, (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics, **25**, 186-211*

[39] Pollard, D. (1990). *Empirical Processes: Theory and Applications.* CBMS Conference Series in Probability and Statistics, Vol. 2. Hayward, CA: Institute of Mathematical Statistics.

[40] Robinson, P.M. (1983). Nonparametric estimation for time series models, *Journal of Time Series Analysis, 4, 185-208.*

[41] Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

[42] Stein, E.M. (1970). *Singular Integrals and Differentiability Properties of Functions*, Princeton, NJ: Priniceton University Press.

[43] Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 685-705.

[44] Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14**, 592-606.

[45] Teräsvirta, T., D. Tjøstheim, and C.W.J. Granger, (1994). Aspects of Modelling Nonlinear Time Series in *The Handbook of Econometrics*, vol. IV, eds. D.L. McFadden and R.F. Engle, 2919-2960, Amsterdam: Elsevier.

[46] Tjøstheim, D., and B. Auestad, (1994). Nonparametric identification of nonlinear time series: projections. *J. Am. Stat. Assoc.* **89**, 1398-1409.

[47] Tong, H. (1990). *Nonlinear Time Series Analysis: A dynamic Approach*, Oxford University Press, Oxford.

[48] Volkoniskii and Y.U. Rozanov, (1959). Some limit theorems for random functions. *Theory of Probability and Applications,* **4**, 178-197.

[49] Yang, L., W. Härdle, and J. Nielsen, (1999). Nonparametric autoregression with multiplicative volatility and additive mean. Journal of Time Series Analysis. **20**(5): 579-604.

[50] Ziegelmann, F. (2002). Nonparametric estimation of volatility functions: the local exponential estimator. *Econometric Theory **18**,* 985-992.

Figure 1(a)
averages of volatility estimates (demeaned) : the 1st lag

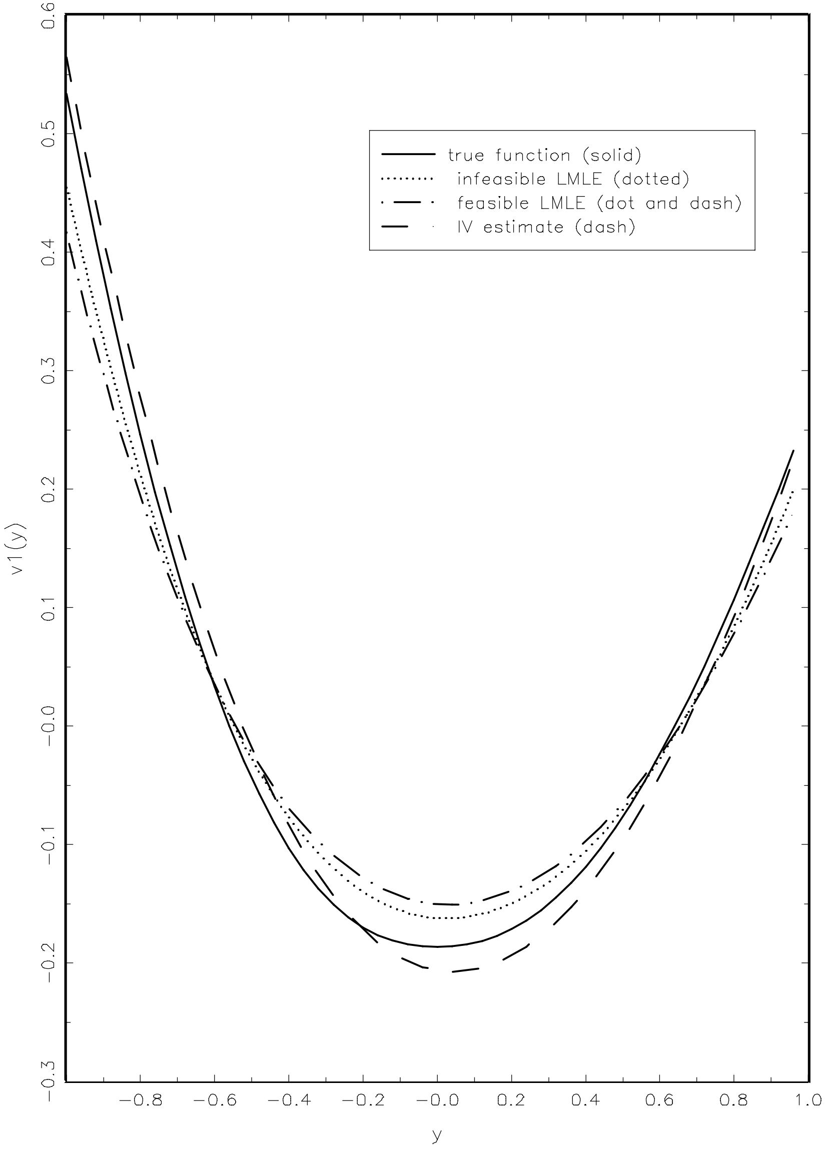


Figure 1(b)
averages of volatility estimates (demeaned) : the 2nd lag

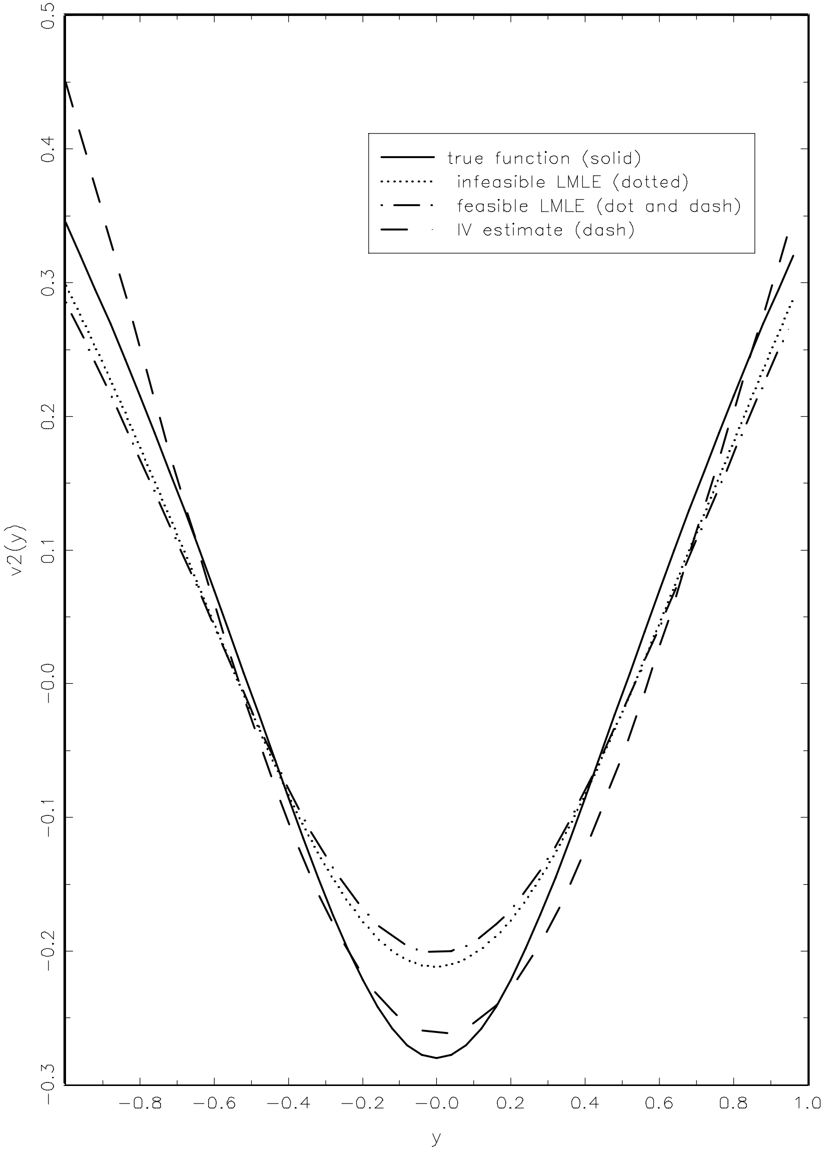


Figure 2-1(a)

volatility estimates (demeaned) : the 1st lag

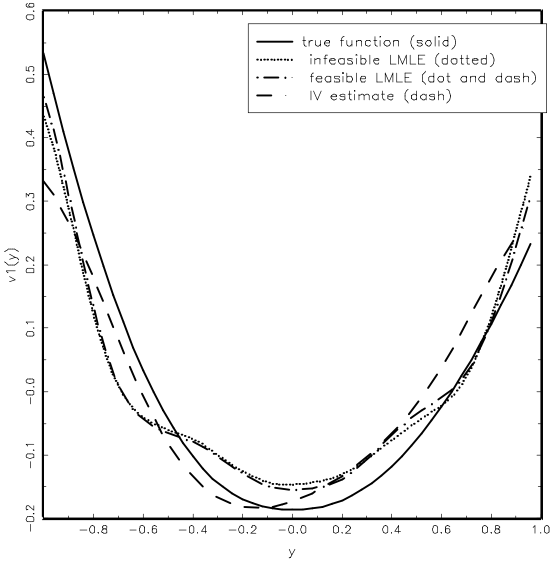


Figure 2-1(b)

volatility estimates (demeaned) : the 2nd lag

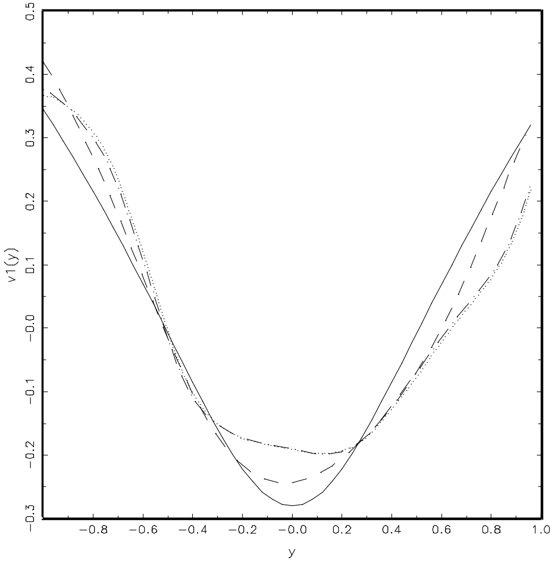


Figure 2-2(a)

volatility estimates (demeaned) : the 1st lag

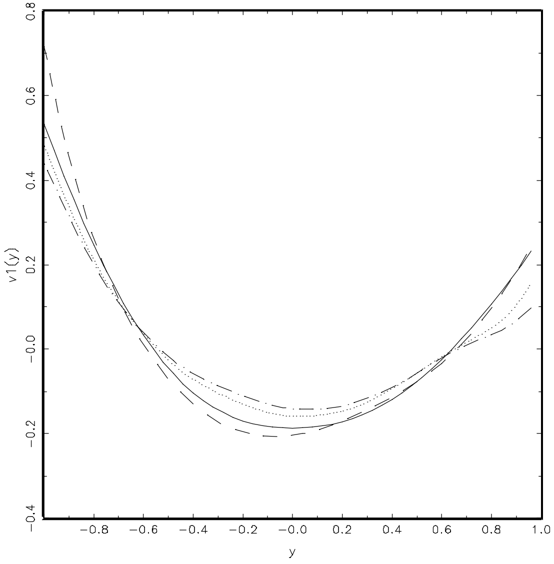


Figure 2-2(b)

volatility estimates (demeaned) : the 2nd lag

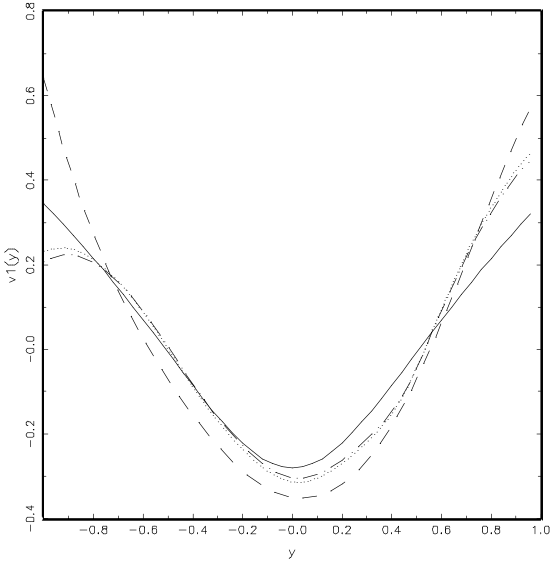


Figure 2-3(a)

volatility estimates (demeaned) : the 1st lag

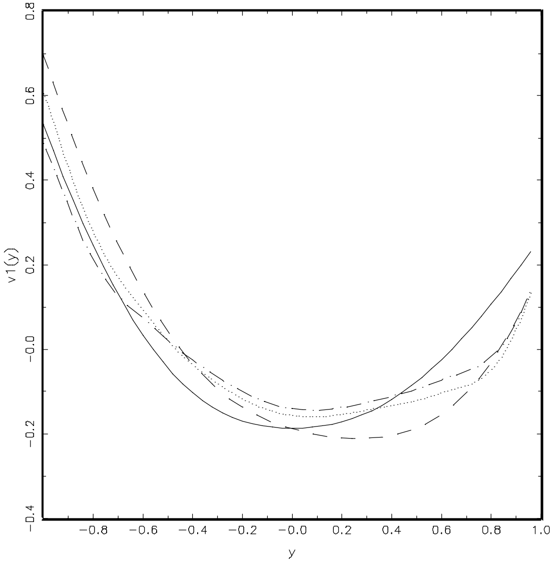
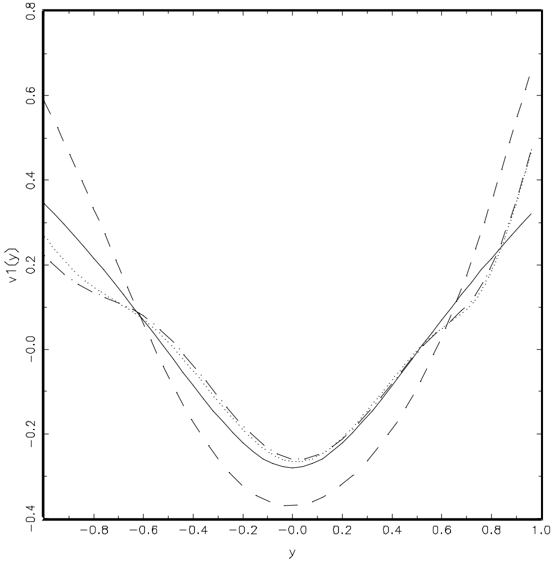


Figure 2-3(b)

volatility estimates (demeaned) : the 2nd lag


v1(y)
0.8
0.6
0.4
0.2
-0.0
-0.2
-0.4
-0.8
-0.6
-0.4
-0.2
-0.0
0.2
0.4
0.6
0.8
1.0
y