Nonparametric Transformation to White Noise

Oliver B. Linton[*]
London School of Economics

Enno Mammen[†]
Universität Mannheim

Abstract

We consider a semiparametric distributed lag model in which the "news impact curve" m is nonparametric but the response is dynamic through some linear filters. A special case of this is a nonparametric regression with serially correlated errors. We propose an estimator of the news impact curve based on a dynamic transformation that produces white noise errors. This yields an estimating equation for m that is a type two linear integral equation. We investigate both the stationary case and the case where the error has a unit root. In the stationary case we establish the pointwise asymptotic normality. In the special case of a nonparametric regression subject to time series errors our estimator achieves efficiency improvements over the usual estimators, see Xiao, Linton, Carroll, and Mammen (2003). In the unit root case our procedure is consistent and asymptotically normal unlike the standard regression smoother. We also present the distribution theory for the parameter estimates, which is non-standard in the unit root case. We also investigate its finite sample performance through simulation experiments.

Keywords: Efficiency; Inverse Problem; Kernel Estimation; Nonparametric regression; Time Series; Unit Roots.

JEL Nos.:   C14

1 Introduction

In this paper we discuss the estimation of the unknown quantities in the model

$$B(L)Y_t = A(L)m(X_t) + \varepsilon_t, \tag{1}$$

where ε_t is a martingale difference sequence with respect to the past of Y_t and current and past regressors X_t, while $A(L) = \sum_{j=0}^{\infty} a_j L^j$ and $B(L) = \sum_{j=0}^{\infty} b_j L^j$ are lag polynomial operators with $a_0 = b_0 = 1$ for identification, where $Lx_t = x_{t-1}$. The function $m(.)$ is assumed to be unknown but smooth, and is the object of central interest, although the dynamics of the model represented by $A(L), B(L)$ are also fundamental to the interpretation.

We first discuss a special case of central interest, the nonparametric regression model

$$Y_t = m(X_t) + u_t, \, t = 1, \ldots, T, \tag{2}$$

where the covariates follow some stationary mixing process, while the residual process u_t satisfies

$$A(L)u_t = \varepsilon_t = \sum_{j=0}^{\infty} a_j u_{t-j}. \tag{3}$$

In this case, $A(L)Y_t = A(L)m(X_t) + \varepsilon_t$, which is a special case of (1) with $A(L) = B(L)$. The parametric version of the regression model (2) and (3) is a standard teaching topic in graduate econometrics, Harvey (1981, Chapter 6). In the semiparametric model there are many standard estimators of m and of the parameters of $A(L)$ that are consistent under summability conditions on A, see for example Robinson (1983), Bierens (1983), Masry and Fan (1997), Hidalgo (1997), and Fan and Yao (2003)). However, unlike in the parametric case, the standard kernel regression smoothers do not take account of the correlation structure in X_t or u_t and estimate the regression function in the same way as if these processes were independent. Furthermore, the variance of such estimators is proportional to the short run variance of u_t, $\sigma_u^2 = \mathrm{var}(u_t)$ and does not depend on the regressor or error covariance functions $\mathrm{cov}(X_t, X_{t-j})$, $\mathrm{cov}(u_t, u_{t-j})$, $j \neq 0$. This is a bit surprising in comparison with the parametric case. One might think that there is useful information in the autocorrelation structure for estimation of the mean. This point has been addressed recently by Xiao, Linton, Carroll, and Mammen (2003) who proposed a more efficient estimator of m based on a prewhitening transformation

$$Y_t - \sum_{j=1}^{\infty} a_j(Y_{t-j} - m(X_{t-j})) = m(X_t) + \varepsilon_t, \tag{4}$$

where the right hand side is now a standard nonparametric regression with whitened errors. The transform implicitly takes account of the autocorrelation structure. In practice they replaced the

unknown quantities on the left hand side by preliminary estimates of m and $a_j(\alpha)$. Their procedure improves in terms of variance over the usual kernel smoothers.

The model (1) is more general than nonparametric regression with autocorrelated errors and is perhaps more rightly viewed as a generalization of the distributed lag model. The traditional distributed lag model (with $m(x) = x$) has been very popular in economics, Dhrymes (1971).[1] More recently, Hendry, Pagan, and Sargan (1984) reviewed the specification of such models and gave a taxonomy of special cases. It can be motivated from some simple economic relationships being distorted by adaptive expectations, partial adjustment, etc., see Harvey (1981, Chapter 7). Suppose there is a latent variable Y^* that has some equilibrium relationship with covariate X, which in general can be nonlinear so that $Y_t^* = m(X_t)$. Then suppose that actual Y only responds to Y^* with some lagging mechanism, for example, $Y_t - Y_{t-1} = \gamma[Y_t^* - Y_{t-1}] + \varepsilon_t$ for some $\gamma \in (0,1)$, then we obtain a special case of (1).[2] The lags arise because production takes time or because agents take time to respond to a signal or because there are institutional constraints. The traditional applications were in for example production studies where Y_t is output and X_t is the capital/labour ratio of a given firm or industry observed over time. More recent applications have been in rational expectations models where the data are at different frequencies, Hansen and Hodrick (1980). The issues concerning formulation and estimation of the lag polynomials A, B are pretty much resolved in the linear case, see Hannan and Deistler (1988) for a more recent discussion in the multivariate case. Linearity of m is just a convenience and was adopted many years ago when computational and technical issues were binding. We allow for nonlinear m because for some problems linear m is not well motivated and at odds with the data. Note that model (1) includes as a special case the so-called NARMAX model introduced in Chen and Billings (1989) and used frequently by systems engineers in which the function m is approximated by some polynomial with unknown coefficients.

Finally, we remark that the ARCH(∞) model of Linton and Mammen (2005) is a special case when $Y_t = y_t^2$ and $X_t = y_{t-1}$, while $B(L) = 1$. This model has been treated elsewhere.

We treat only the case where $A(L), B(L)$ are described by a finite dimensional parameter $\theta = (\alpha, \beta) \in \mathbb{R}^p$ with $\alpha \in \mathbb{R}^{p_a}$ parameterizing A and $\beta \in \mathbb{R}^{p_b}$ parameterizing B. We propose a strategy for estimation of m along with the parameters of $A(L)$ in (2), (3). This is essentially to estimate the

---

[1]Sims (1971) and Geweke (1978) consider a continuous time distributed lag model where $Y(t) = \int_{-\infty}^{\infty} a(s)X(t-s)ds + \varepsilon(t)$ and the data are observed at discrete time intervals in which case the (high frequency) discrete time approximation to this is like (1) with $B(L) = 1$ and $A(L) = \sum_{j=-\infty}^{\infty} a_j L^j$ for some a_j related to the function $a(.)$ under some conditions.

[2]The usual properties of linear dynamic regression models can be extended to the nonlinear case. Thus for example we can define the average instantaneous impact $E[\partial Y_t/\partial X_t]$ as equal to the average derivative of the function m, $= E[m'(X_t)]$, a quantity that has been investigated elsewhere. The total dynamic average impact $\sum_{j=0}^{\infty} E[\partial Y_{t+j}/\partial X_t] = E[m'(X_t)] \sum_{j=0}^{\infty} (B(L)/A(L))_j$ is proportional to the instantaneous impact.

transformed model (1) as an additive (possibly infinite order) nonparametric regression, see Hastie and Tibshirani (1991). Recently, Linton and Mammen (2005) have shown how to estimate similar models using the theory of linear integral equations of the second kind; see also Carrasco, Florens and Renault (2002). We obtain an estimating equation for m that is a type two linear integral equation for each parameter value θ. To obtain the parameters θ we optimize a profile likelihood criterion. We show that our method has attractive theoretical and finite sample properties. In particular, in the special case of nonparametric regression with autocorrelated error it has smaller asymptotic variance than the main method of Xiao, Linton, Carroll, and Mammen (2003). Furthermore, the asymptotics require weaker conditions with regard to the memory properties of the error terms. We define our method in the general model (1). In that case there is not such an obvious alternative estimator of the function m. We mostly consider the case where both X_t, Y_t are stationary and mixing processes in which case the main statistical issue is efficiency. We also consider the case where some of the variables are nonstationary. This could arise for example from a unit root in the residual u_t or in X_t or in both, see Phillips and Park (1998). In this case, estimating in the original data (2) may lead to inconsistency, whereas the transformation involved in (1) yields error terms with a lower order of nonstationarity/persistence and hence consistency can be obtained, see Marinucci and Robinson (2003). The estimation method is more or less the same as in the stationary case although the justification of it differs. The distribution theory for the parametric part though is non standard in this case: in fact we obtain T convergence to the Dickey-Fuller distribution under the unit root.

2 The Stationary Case

In this section we suppose that (Y_t, X_t) are jointly stationary and weakly dependent mixing processes and describe our estimation methods and their properties for this case.

2.1 Estimation Method

2.1.1 Population Characterization

We first suppose that $A(L), B(L)$ are known. Letting $Z_t = B(L)Y_t$ we have

$$Z_t = A(L)m(X_t) + \varepsilon_t = \sum_{j=0}^{\infty} a_j m(X_{t-j}) + \varepsilon_t,$$

which is an additive autoregression with i.i.d. errors where the additive components are subject to the restriction that they all share a common function m. In view of the assumed stationarity, define

the function m as the minimizer of the criterion

$$Q(\theta_0, m) = E\left[\left\{Z_0 - \sum_{j=0}^{\infty} a_j m(X_{-j})\right\}^2\right]. \tag{5}$$

This problem can be viewed as a projection in a suitable Hilbert space. Let $L_2(f_0)$ be the Hilbert space of square integrable functions with respect to the marginal density of X_t denoted f_0. For the stationary mixing process $\{X_t\}_{t=-\infty}^{\infty}$, provided $\sum_{j=0}^{\infty} |a_j| < \infty$, the random variable $\sum_{j=0}^{\infty} a_j m(X_{-j})$ is square integrable for any function $m \in L_2(f_0)$. The set $\mathcal{G} = \{\sum_{j=0}^{\infty} a_j m(X_{-j}) : m \in L_2(f_0)\}$ can be viewed as a subspace of the Hilbert space of square integrable functions defined on the infinite product of random variables $\underline{X} = (X_0, X_{-1}, \ldots)$. By the projection theorem there exists a unique member of $\mathcal{G}$ closest to the random variable Z_0.

A necessary condition for m to be the minimizer of (5) is that it satisfies the first order condition

$$E\left[\left\{Z_0 - \sum_{j=0}^{\infty} a_j m(X_{-j})\right\} \sum_{k=0}^{\infty} a_k h(X_{-k})\right] = 0 \tag{6}$$

for any measurable function h for which this expectation is well-defined. See Sagan (1969), Theorem 1.7 for example. The second order condition is $-E[\{\sum_{k=0}^{\infty} a_k h(X_{-k})\}^2]$ which is negative implying that the solution of the first order condition does indeed (locally) minimize the criterion. Taking $h(.)$ to be the Dirac delta function, we have that

$$\sum_{j=0}^{\infty} a_j E[Z_0 | X_{-j} = x] = \sum_{j=0}^{\infty} a_j^2 m(x) + \sum\sum_{j \neq k} a_j a_k E[m(X_{-j}) | X_{-k} = x] \tag{7}$$

for each x.[3] This is an implicit equation for $m(.)$. It can be re-expressed as a linear type two integral equation in $L_2(f_0)$. Define $a_j^{\dagger} = a_j / \sum_{j=0}^{\infty} a_j^2$ and $a_j^* = \sum_{k \neq 0} a_{j+k} a_j / \sum_{l=0}^{\infty} a_l^2$, and let $f_{0,j}$ be the joint density of (X_t, X_{t-j}). Then

$$m(x) = m^*(x) + \int \mathcal{H}(x,y) m(y) f_0(y) dy, \text{ or } m = m^* + \mathcal{H}m, \tag{8}$$

$$m^*(x) = \sum_{j=0}^{\infty} a_j^{\dagger} E[Z_0 | X_{-j} = x]$$

$$\mathcal{H}(x,y) = -\sum_{j=\pm 1}^{\pm\infty} a_j^* \frac{f_{0,j}(y,x)}{f_0(y) f_0(x)}.$$

---

[3]This equation can also be derived at by directly taking conditional expectations of Z_t given each X_{t-k}, multiplying by a_k, and then summing over k.

This is similar to the equation derived in Linton and Mammen (2005) with the exception that there X_t was lagged values of Y_t. Equation (8) is an implicit equation in m and we need some conditions on the operator $\mathcal{H}(x, y)$ to guarantee that there exists a unique solution.

ASSUMPTION A1. *The operator $\mathcal{H}(x, y)$ satisfies the Hilbert-Schmidt condition i.e.,*

$$\int \int \mathcal{H}(x, y)^2 f_0(x) f_0(y) dx dy < \infty.$$

A sufficient condition for A1 is that the joint densities $f_{0,j}(y, x)$ have compact support and $f_0(x)$ is bounded away from zero on this support, which we shall assume below. However, this is not necessary and condition A1 can hold for many covariate processes with unbounded support. We shall however restrict attention to the case where the support of the marginal covariate density f_0 is a compact set $[\underline{x}, \overline{x}]$. Then the operator $\mathcal{H}$ is a bounded compact linear operator on the Hilbert space of functions $L_2(f_0)$. It is also self-adjoint, see Linton and Mammen (2005). It therefore has a countable number of eigenvalues[4]:

$$\infty > |\lambda_1| \geq |\lambda_2| \geq \ldots,$$

with $\sum_{j=0}^{\infty} \lambda_j^2 < \infty$. The spectral radius of the operator $r(\mathcal{H}) = \sup_j |\lambda_j| < \infty$. Also, the value 0 is a cluster point of the set $\{\lambda_j\}_{j=1}^{\infty}$ and 0 is the only cluster point, see Kress (1999, Theorem 3.9).

ASSUMPTION A2. *There exist no measurable function $m(.)$ with $\int m(x)^2 f_0(x) dx = 1$ such that* $\sum_{j=0}^{\infty} a_j m(X_{t-j}) = 0$ *with probability one.*

This condition rules out a certain 'concurvity' in the stochastic process $\{X_t\}$. That is, the data cannot be functionally related in this particular way. In the AR(1) case this says that there are no nontrivial functions m that satisfy $m(X_t) - \rho m(X_{t-1}) = 0$ with probability one.[5] A consequence of A2 is that $\sup_j \lambda_j < 1$ and therefore the operator $I - \mathcal{H}$ is strictly positive definite. Therefore, there exists a unique solution to (8) that satisfies

$$m = (I - \mathcal{H})^{-1} m^*. \tag{9}$$

This is the main characterization used for estimation, although we must first extend this to the case where a general θ is used not necessarily the true θ_0.

For each $\theta = (\alpha, \beta) \in \Theta$, define $Z_t(\beta) = \sum_{j=0}^{\infty} b_j(\beta) Y_{t-j}$ and $g_j(x; \beta) = E[Z_t(\beta)|X_{t-j} = x]$, $j = 0, \pm 1, \ldots$

$$m_\theta^*(x) = \sum_{j=0}^{\infty} a_j^\dagger(\alpha) g_j(x; \beta)$$

---

[4]These are real numbers for which there exists functions $e_j(.)$ such that $\mathcal{H} e_j = \lambda_j e_j$.

[5]One example where this condition is not satisfied is when $X_t = t/T$.

$$\mathcal{H}_\theta(x,y) = -\sum_{j=\pm 1}^{\pm\infty} a_j^*(\alpha) \frac{f_{0,j}(y,x)}{f_0(y)f_0(x)}, \tag{10}$$

where $a_j^\dagger(\alpha) = a_j(\alpha)/\sum_{j=0}^\infty a_j^2(\alpha)$ and $a_j^*(\alpha) = \sum_{k\neq 0} a_{j+k}(\alpha)a_j(\alpha)/\sum_{l=0}^\infty a_l^2(\alpha)$. We now let m vary with θ, that is, (5) is defined for any θ, and let m_θ be the function that minimizes (5); this satisfies $m_\theta = (I - \mathcal{H}_\theta)^{-1}m_\theta^*$ for all θ provided the conditions A1 and A2 hold uniformly over the parameter space Θ. Furthermore, we can define $\theta = \theta_0$ as the minimizer of

$$Q(\theta, m_\theta) = E\left[\left\{Z_0(\beta) - \sum_{j=0}^\infty a_j(\alpha)m_\theta(X_{-j})\right\}^2\right] \tag{11}$$

with respect to $\theta \in \Theta$. Let $m_0 = m_{\theta_0}$. We adopt this profiling approach to defining θ_0, m_0 as this is the way our estimation strategy works. We suppose that assumptions A1 and A2 hold uniformly over the parameter space Θ so that for each $\theta \in \Theta$, $m_\theta = (I - \mathcal{H}_\theta)^{-1}m_\theta^*$ is well-defined. Note that the operator $\mathcal{H}_\theta$ is not necessarily a contraction, i.e., it may hold that $r(\mathcal{H}_\theta) > 1$ for some $\theta \in \Theta$. Therefore, one cannot guarantee that the infinite sum $\sum_{j=0}^\infty \mathcal{H}_\theta^j$ exists for all $\theta \in \Theta$.

In practice one has to replace m_θ^* and $\mathcal{H}_\theta$ by estimators. Furthermore, one has also to estimate the parameters of the filters A, B. In the sequel we provide some details on this.

2.1.2 Further Details

Suppose we have a sample $\{(Y_1, X_1), \ldots, (Y_T, X_T)\}$. The general estimation strategy is

1. For each θ compute estimators of $\widehat{m}_\theta^*, \widehat{\mathcal{H}}_\theta$ of $m_\theta^*, \mathcal{H}_\theta$

2. Solve an empirical version of the equation (8) to obtain an estimator $\widehat{m}_\theta$ of m_θ

3. Choose $\widehat{\theta}$ to minimize the profiled least squares criterion with respect to θ. Let $\widehat{m}(x) = \widehat{m}_{\widehat{\theta}}(x)$.

Let $\tau = \tau(T)$ be some truncation parameter and define $Z_t^\tau(\beta) = \sum_{j=0}^\tau b_j(\beta)Y_{t-j}$. The choice of truncation depends on the dependence model $A(L), B(L)$. For geometrically declining parameters (as we shall assume) one can work with logarithmic truncation. There are many suitable estimators of the regression functions and density functions in our estimator; we shall use local linear regression estimators for m^* and a fairly standard density estimator for $\mathcal{H}$ but other choices are possible.

For any sequence $\{Z_t^\tau(\beta)\}$ and any lag j define the estimator $\widehat{g}_j(x; \beta) = \widehat{c}_0$, where $(\widehat{c}_0, \widehat{c}_1)$ are the minimizers of the weighted sums of squares criterion

$$\sum_{t=j+1}^T \left\{Z_t^\tau(\beta) - c_0 - c_1(X_{t-j} - x)\right\}^2 K_h(X_{t-j} - x) \tag{12}$$

6

with respect to (c_0, c_1), where K is a symmetric probability density function, h is a positive bandwidth, and $K_h(.) = K(./h)/h$. Further define

$$\widehat{f}_{0,j}(y,x) = \frac{1}{T - |j|} \sum_{t=|j|+1}^{T} K_h(y, X_t) K_h(x, X_{t-j}) \quad ; \quad \widehat{f}_0(x) = \frac{1}{T} \sum_{t=1}^{T} K_h(x, X_t).$$

$$\widehat{m}_\theta^*(x) = \sum_{j=0}^{\tau} a_j^\dagger(\alpha) \widehat{g}_j(x; \beta) \quad ; \quad \widehat{\mathcal{H}}_\theta(x,y) = -\sum_{j=\pm 1}^{\pm \tau} a_j^*(\alpha) \frac{\widehat{f}_{0,j}(y,x)}{\widehat{f}_0(y) \widehat{f}_0(x)},$$

$$\widehat{\mathcal{H}}_\theta m(x) = \int \widehat{\mathcal{H}}_\theta(x,y) m(y) \widehat{f}_0(y) dy.$$

Here, for each x in the support of X_t, $K_h(x,y) = K_h^x(x-y)$ for some kernel K^x such that $K_h^x(u) = h^{-1} K^x(h^{-1}u)$ and $K_h^x(u) = K_h(u)$ for all x in the interior of the support of X_t. We shall assume that the covariate is supported on $[\underline{x}, \overline{x}]$ for some known $\underline{x}, \overline{x}$ and that the covariate density is bounded away from zero on this support. We need to make a boundary adjustment to the kernel K in $\widehat{\mathcal{H}}_\theta$ by using the boundary kernels $K_h^x(y-x)$ to ensure that the bias is the same magnitude everywhere.

Then define $\widehat{m}_\theta$ as any solution to the equation

$$m = \widehat{m}_\theta^* + \widehat{\mathcal{H}}_\theta m, \tag{13}$$

in $L_2(\widehat{f}_0)$. We discuss the computation of this solution in the appendix. Let $\widehat{\theta} = \arg\min_{\theta \in \Theta} \widehat{Q}_T(\theta)$, where

$$\widehat{Q}_T(\theta) = \frac{1}{T} \sum_{t=\tau+1}^{T} \left\{ Z_t^\tau(\beta) - \sum_{j=0}^{\tau} a_j(\alpha) \widehat{m}_\theta(X_{t-j}) \right\}^2.$$

Finally, let $\widehat{m}(x) = \widehat{m}_{\widehat{\theta}}(x)$.

3 Asymptotic Properties

Let $\mathcal{F}_a^b$ be the σ-algebra of events generated by the random variables $\{Y_t, X_t; \quad a \leq j \leq b\}$. A stationary processes $\{Y_t, X_t\}$ is called strongly mixing [Rosenblatt (1956)] if

$$\sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} |\Pr(A \cap B) - \Pr(A)\Pr(B)| \equiv s(k) \to 0 \quad \text{as } k \to \infty. \tag{14}$$

We shall consider two cases. First, the 'weak form case' where we do not maintain that model (1) holds only that $\{Y_t, X_t\}$ is a stationary strong mixing process. Second, we maintain that in additional model (1) holds with a martingale difference error sequence ε_t. To facilitate the asymptotic analysis,

we make the following assumptions on the residuals and regressors, the kernel function $k(\cdot)$, and the bandwidth parameter h. Let $\eta_{t,j}(\beta) = Z_{t+j}(\beta) - E[Z_{t+j}(\beta)|X_t]$, $\zeta_{t,j}(\theta) = m_\theta(X_{t+j}) - E[m_\theta(X_{t+j})|X_t]$,

$$\eta^1_{\theta,t} = \sum_{j=0}^{\infty} a_j^\dagger(\alpha)\eta_{t,j}(\beta) \text{ and } \eta^2_{\theta,t} = -\sum_{j=\pm 1}^{\pm\infty} a_j^*(\alpha)\zeta_{t,j}(\theta). \tag{15}$$

B1 *The process $\{X_t, Y_t\}_{t=-\infty}^{\infty}$ is stationary and alpha mixing with a mixing coefficient, $s(k)$ such that for some $C \geq 0$ and some $\overline{s} < 1$, $s(k) \leq C\overline{s}^k$.*

B2 *$E\left(|Y_t|^{2\rho}\right) < \infty$ for some $\rho > 2$.*

B3 *The covariate process $\{X_t\}_{t=-\infty}^{\infty}$ has absolutely continuous density f_0 supported on $[\underline{x}, \overline{x}]$ for some $-\infty < \underline{x} < \overline{x} < \infty$ and the bivariate densities $f_{0,j}(\cdot)$ are supported on $[\underline{x}, \overline{x}]^2$. The function $m(\cdot)$ together with the densities $f_0(\cdot)$ and $f_{0,j}(\cdot)$ are continuous and twice continuously differentiable over $(\underline{x}, \overline{x})$ [and $(\underline{x}, \overline{x})^2$], and are uniformly bounded. $f_0(\cdot)$ is bounded away from zero on $[\underline{x}, \overline{x}]$, i.e., $\inf_{\underline{x} \leq w \leq \overline{x}} f_0(w) > 0$.*

B4 *The parameter space Θ is a compact subset of $\mathbb{R}^p$, and the value θ_0 is an interior point of Θ. Also, A2 holds, and for any $\epsilon > 0$*

$$\inf_{||\theta - \theta_0|| > \epsilon} Q(\theta, m_\theta) > Q(\theta_0, m_{\theta_0}).$$

B5 *The density function μ of $(\eta^1_{t,j}(\beta), \eta^2_{t,j}(\beta))$ is Lipschitz continuous on its domain. The joint densities $\mu_{0,j}, j = 1, 2, \ldots,$ of $(\eta^1_{t,0}(\beta), \eta^2_{t,0}(\beta)), (\eta^1_{t,j}(\beta), \eta^2_{t,j}(\beta))$ are uniformly bounded.*

B6 *The parameters $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$ compact subsets of $\mathbb{R}^{p_a}$ and $\mathbb{R}^{p_b}$ respectively. The coefficients satisfy $\sup_{\alpha \in A, k=0,1,2} ||\partial^k a_j(\alpha)/\partial\alpha^k|| \leq C\overline{a}^j$ for some $\overline{a} < 1$ and some finite constant C, while $\inf_{\alpha \in \mathcal{A}} \sum_{j=0}^{\infty} a_j^2(\alpha) > 0$. Likewise, $\sup_{\beta \in B, k=0,1,2} ||\partial^k b_j(\beta)/\partial\beta^k|| \leq C\overline{b}^j$ for some $\overline{b} < 1$ and some finite constant C.*

B7 *The truncation sequence τ_T satisfies $\tau_T = C \log T$ for some constant $C > (-2\log\overline{b})^{-1}$.*

B8 *The bandwidth sequence $h(T)$ satisfies $T^{1/5}h(T) \to \gamma$ as $T \to \infty$ with γ bounded away from zero and infinity.*

B9 *For each $x \in [\underline{x}, \overline{x}]$ the kernel function K^x has support $[-1, 1]$ and satisfies $\int K^x(u)du = 1$ and $\int K^x(u)udu = 0$, such that for some constant C, $\sup_{x \in [\underline{x}, \overline{x}]} |K^x(u) - K^x(v)| \leq C|u-v|$ for all $u, v \in [-1, 1]$. Define $\mu_j(K) = \int u^j K(u)du$ and $||K||_2^2 = \int K^2(u)du$.*

B10 *ε_t satisfies $E\left[\varepsilon_t | \{X_{t-j}\}_{j=0}^{\infty}, \{\varepsilon_{t-j}\}_{j=1}^{\infty}\right] = 0$ a.s.*

8

B11 (a) ε_t is i.i.d. and independent of the process $\{X_t\}$; (b) ε_t is also normally distributed.

These conditions are similar to Linton and Mammen (2005) but we also need conditions on the $b_j(\beta)$ coefficients and separate conditions on X and Y.

Note that B1-B6 imply the uniform version of conditions A1-A2. Condition B1 rules out long memory but allows a wide range of processes used in practice. We will make use of the mixing property to apply the exponential inequality of Bosq (1998) and to establish a central limit theorem for $\widehat{m}_\theta$ in the weak form case. In this weak form case we can't apply martingale limit theory. We need to apply a central limit theorem to (local) averages of the processes $\eta^1_{\theta,t}$ and $\eta^2_{\theta,t}$ defined above. These processes need not be mixing but are near epoch dependent processes on the strong mixing bases Y_t, X_t with exponentially declining weights under our conditions on B, A; we apply a CLT due to Lu (2001) for such processes using conditions B1 and B5,B6.

Condition B3 is quite standard in the nonparametric regression literature. Note that we only assume twice continuously differentiable m.

In B4 we explicitly assume the identification of the parametric part. We make this high level assumption for three reasons. First, we need identification in the weak case, and this seems like a natural assumption to make in view of our definition of the weak form process. Second, we allow the coefficients $a_j(\theta), b_j(\theta)$ to depend on θ in a complicated way. Third, the mapping $\theta \longmapsto m_\theta$ may be quite complicated to analyze. Hannan (1973) used high level conditions [c.f. his condition (4)] similar to ours.

The truncation rate assumed in B7 is consistent with the exponential decaying mixing coefficients. It can be weakened at the expense of more detailed argumentation. In B8 we are anticipating a rate of convergence of $T^{-2/5}$ for $\widehat{m}_\theta$, which is consistent with second order smoothness on the function m. The assumptions B10 are expressed in terms of the unobserved $\{\varepsilon_{t-j}\}_{j=1}^\infty$ and are equivalent to assumptions on $\{y_{t-j}\}_{j=1}^\infty$ under an invertibility condition. Assumption B10 is needed for the consistency of the parameter estimates $\widehat{\theta}$. In the pure regression model (2, 3) one only needs a weaker assumption $E\left[\varepsilon_t | \{X_{t-j}\}_{j=0}^\infty\right] = 0$ a.s. for consistent estimation of m and θ as is known from the parametric case.

Define the functions $\beta_\theta = (I - \mathcal{H}_\theta)^{-1}\beta_\theta^*$ as solutions to the integral equations $\beta_\theta = \beta_\theta^* + \mathcal{H}_\theta\beta_\theta$, in which:

$$\beta_\theta^*(x) = \frac{f_0'(x)}{f_0(x)}\frac{\partial}{\partial x}\mathcal{H}_\theta m_\theta(x) + \mathcal{H}_\theta m_\theta''(x).$$

Then define

$$\omega_\theta(x) = \frac{||K||_2^2}{f_0(x)}\mathrm{var}[\eta^1_{\theta,t} + \eta^2_{\theta,t}],$$
$$b_\theta(x) = \frac{1}{2}\mu_2(K)\beta_\theta(x),$$

where $\eta_{\theta,t}^j$, $j = 1, 2$ were defined above in (15). We prove the following theorem in the appendix.

THEOREM 1. *Suppose that B1-B9 hold. Then for each $\theta \in \Theta$ and $x \in (\underline{x}, \overline{x})$*

$$\sqrt{Th} \left[\widehat{m}_\theta(x) - m_\theta(x) - h^2 b_\theta(x) \right] \Longrightarrow N\left(0, \omega_\theta(x)\right), \tag{16}$$

Both the bias and variance in this result are quite complicated even though a local linear smoother has been used in estimating g_j. This is a 'weak form' result, where the model (1) is not assumed.

We next maintain a 'semi-strong form' assumption B10, which requires the filters to be correctly specified. Under this assumption we can apply a CLT for martingale difference sequences. We obtain the properties of $\widehat{\theta}$ by an application of the asymptotic theory for semiparametric profiled estimators, see Severini and Wong (1992) and Newey (1994). This requires a uniform expansion for $\widehat{m}_\theta(x)$ and for the derivatives (with respect to θ) of $\widehat{m}_\theta(x)$. Under B10, we get that

$$\eta_{\theta_0,t}^1 + \eta_{\theta_0,t}^2 = \sum_{j=0}^{\infty} a_j^\dagger \varepsilon_{t+j}.$$

Thus:

$$\omega_{\theta_0}(x) = \frac{||K||_2^2 \sum_{j=0}^{\infty} a_j^2(\alpha_0) E\left[\varepsilon_t^2 | X_{t-j} = x\right]}{f_0(x) \left[\sum_{j=0}^{\infty} a_j^2(\alpha_0)\right]^2} \tag{17}$$

Let $\varepsilon_t(\theta) = Z_t(\beta) - \sum_{j=0}^{\infty} a_j(\alpha) m_\theta(X_{t-j})$, and let

$$\mathcal{J} = E\left[\frac{\partial^2 \varepsilon_t}{\partial\theta\partial\theta^\top}(\theta_0)\right] \text{ and } \mathcal{I} = E\left[\frac{\partial\varepsilon_t}{\partial\theta}\frac{\partial\varepsilon_t}{\partial\theta^\top}\varepsilon_t^2(\theta_0)\right].$$

THEOREM 2. *Suppose that Assumptions B1 to B10 hold. Then,*

$$\sqrt{T}(\widehat{\theta} - \theta_0) \Longrightarrow N(0, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}).$$

Furthermore, for $x \in (\underline{x}, \overline{x})$

$$\sqrt{Th} \left(\widehat{m}(x) - m(x) - h^2 b_{\theta_0}(x)\right) \Longrightarrow N\left(0, \omega_{\theta_0}(x)\right).$$

Under the 'strong form' special case B11(a), $\omega(x) = ||K||_2^2 \sigma_\varepsilon^2 / f_0(x) \sum_{j=0}^{\infty} a_j^2$. In the nonparametric regression case we can compare the efficiency of our procedure with that of alternative estimators like the usual kernel regression estimator, which has asymptotic variance $\omega_{Ker}(x) = ||K||_2^2 \sigma_\varepsilon^2 \sum_{j=0}^{\infty} c_j^2 / f_0(x)$, where $C(L) = A(L)^{-1}$. Compare also with the estimator of Xiao, Linton, Carroll, and Mammen (2003), which has variance $\omega_{XLCM}(x) = ||K||_2^2 \sigma_\varepsilon^2 / f_0(x)$. In this case, $\omega(x) \leq \omega_{XLCM}(x) \leq \omega_{Ker}(x)$.

Our estimator can be modified to get an asymptotic bias expression of the form that is asymptotically equivalent to

$$b(x) = \frac{1}{2}\mu_2(K) m''(x), \tag{18}$$

which is as for a standard local linear estimator in regression. Then we get a straight mean squared error reduction over the local linear regression estimator. There exist two proposals for additive models to achieve a bias term of the form (18). The first approach is local linear smooth backfitting of Mammen, Linton, and Nielson (1999). This estimator has the same asymptotic distribution as a theoretical oracle estimator. The oracle estimator of one additive component uses the knowledge of the other components and is based on standard local linear smoothing of the differences of the observations minus the other components. In the smooth backfitting approach the backfitting algorithm updates estimates of the functions and its derivatives. This approach could be implemented in our setting by using an appropriate integral operator that acts on tuples of two functions (fitting m and its derivative). A modification of local linear smooth backfitting has been proposed in Mammen and Park (2006). Their backfitting only uses one dimensional operators but achieves the same asymptotic behaviour as local linear smooth backfitting. In our setting their approach works as follows. One replaces $\widehat{\mathcal{H}}_\theta$ by

$$\widehat{\mathcal{H}}_\theta^{mod}(y,x) = -\sum_{j=\pm 1}^{\pm \tau_T} a_j^*(\alpha)\widehat{\pi}_{0j}(y,x),$$

where

$$\widehat{\pi}_{jk}(x_j, x_k) = \frac{\widetilde{p}_{jk}(x_j, x_k)}{\widetilde{p}_j(x_j)} - \frac{\int \widetilde{p}_{jk}(u, x_k)du}{\int \widetilde{p}_j(u)du}$$

with

$$\widetilde{p}_j(x_j) = \widehat{p}_j(x_j) - \frac{\widehat{p}_j^*(x_j)^2}{\widehat{p}_j^{**}(x_j)},$$

$$\widehat{p}_j^*(x_j) = \frac{1}{n}\sum_{i=1}^n K_{h_j}(X_j^i - x_j)(X_j^i - x_j),$$

$$\widehat{p}_j^{**}(x_j) = \frac{1}{n}\sum_{i=1}^n K_{h_j}(X_j^i - x_j)(X_j^i - x_j)^2,$$

$$\widetilde{p}_{jk}(x_j, x_k) = \widehat{p}_{jk}(x_j, x_k) - \frac{\widehat{p}_{jk}^*(x_j, x_k)\widehat{p}_j^*(x_j)}{\widehat{p}_j^{**}(x_j)},$$

$$\widehat{p}_{jk}(x_j, x_k) = \frac{1}{n}\sum_{i=1}^n K_{h_j}(X_j^i - x_j)L_{h_k}(x_k, X_k^i),$$

$$\widehat{p}_{jk}^*(x_j, x_k) = \frac{1}{n}\sum_{i=1}^n K_{h_j}(X_j^i - x_j)L_{h_k}(x_k, X_k^i)(X_j^i - x_j).$$

Here, the kernel L is defined as $L(u) = 2K_{1/\sqrt{2}}(u) - K_{\sqrt{2}}(u)$. Furthermore, L_h is defined as

$$L_h(u, v) = \{a(v, h)u + b(v, h)\}L\left[h^{-1}(v - u)\right]$$

with a and b chosen so that

$$\int_{\underline{x}}^{\overline{x}} L_h(u,v)du = 1,$$

$$\int_{\underline{x}}^{\overline{x}} (u-v)L_h(u,v)du = 0.$$

Note that the integration runs over u and not over v. Thus the kernel is not a boundary corrected kernel in the usual sense. A similar proposal has been made in Linton and Mammen (2005, p789) but their proposal could be not directly used here because there no corrections at the boundary are needed. For a discussion of oracle efficiency in additive models see also Horowitz, Klemelä, and Mammen (2006). One can also replace the standard kernel density estimators by other suitable density estimators like the Jones, Linton and Nielsen (1995) procedure, but it is not clear if such estimators would achieve the IMSE performance of the two just discussed modified estimators.

The asymptotic distribution can be used to guide bandwidth selection. The IMSE optimal bandwidth is

$$h = \left[\frac{||K||_2^2}{\mu_2^2(K)}\right]^{1/5} \left[\frac{\sigma_\varepsilon^2 (\overline{x} - \underline{x})}{\sum_{j=0}^{\infty} a_j^2(\alpha_0)E\left[m''(X_t)^2\right]}\right]^{1/5} T^{-1/5}$$

for the modified estimator under homoskedasticity, where σ_ε^2 is the variance of ε_t. In practice one must replace these quantities by estimates based on a parametric or nonparametric scheme.

Under the 'strong form' assumption B11 the parametric estimator is semiparametrically efficient, see Linton and Mammen (2005). There is generally an information loss from the necessity of estimating the function m.

4   A Nonstationary Case

In this section we investigate the case where Y_t can be nonstationary but X_t is stationary mixing as before. We wish to allow for the possibility of unit roots even if they might be quite rare in practical applications of this technology.

The most general case would be where both A, B contained unit roots either simple or complex. For expositional reason we shall focus on the special case where $B(L) = A(L) = 1 - L$. Consider the model

$$(1 - \rho L)Y_t = (1 - \rho L)m(X_t) + \varepsilon_t, \tag{19}$$

where in fact $\rho_0 = 1$ and ε_t obeys B11. In this case,

$$Y_t = m(X_t) + u_t, \tag{20}$$

where $u_t = u_{t-1} + \varepsilon_t$ is a unit root process, Phillips (1987). We suppose that $u_0 = 0$.

Direct estimation of Y_t on X_t will produce inconsistent estimates of m. The Xiao, Linton, Carroll, and Mammen (2003) procedure is also inconsistent in this unit root case because it relies on the initial standard nonparametric regression estimator that is inconsistent. On the other hand our estimation of the additive model

$$Y_t - Y_{t-1} = m(X_t) - m(X_{t-1}) + \varepsilon_t$$

with white noise errors will produce consistent estimates of m. In fact, the theory for m_{ρ_0} is exactly as in Theorem 1. The task here is to determine that we can estimate the parameter ρ in (19) consistently and thence estimate m consistently.[6]

One issue is that for $\rho \neq 1$, the process $(1 - \rho L)Y_t$ is non-stationary and so some of the definitions of the previous section do not make sense. Instead we define $m_{T\rho}$ to be the potentially time varying minimizer of

$$Q_T(m) = \frac{1}{T} \sum_{t=1}^{T} E\left[\{Y_t - \rho Y_{t-1} - m(X_t) + \rho m(X_{t-1})\}^2\right].$$

A necessary condition for m to be the minimizer is that it satisfies the first order condition

$$\frac{1}{T} \sum_{t=1}^{T} E[Y_t - \rho Y_{t-1} | X_t = x] - \rho E[Y_t - \rho Y_{t-1} | X_{t-1} = x] \tag{21}$$

$$= (1 + \rho^2) m_{T\rho}(x) + \rho \left(E[m_{T\rho}(X_t) | X_{t-1} = x] + E[m_{T\rho}(X_{t-1}) | X_t = x]\right).$$

Then note that $Y_t - \rho Y_{t-1} = m(X_t) - \rho m(X_{t-1}) + \varepsilon_t + (1 - \rho)u_{t-1}$, and so $E[Y_t - \rho Y_{t-1} | X_t = x]$ and $E[Y_t - \rho Y_{t-1} | X_{t-1} = x]$ are time invariant. Furthermore, we have assumed that X_t is stationary. Therefore, there exists a time invariant solution to equation (21) as in the purely stationary case.[7] Furthermore, the solution is characterized by the integral equation (8) with in this special case:

$$m_\rho^*(x) = \frac{1}{1 + \rho^2} \left(E[Y_t - \rho Y_{t-1} | X_t = x] - \rho E[Y_t - \rho Y_{t-1} | X_{t-1} = x]\right)$$

$$\mathcal{H}_\rho(x, y) = -\frac{\rho}{1 + \rho^2} \left(\frac{f_{0,1}(y, x)}{f_0(y) f_0(x)} + \frac{f_{0,1}(x, y)}{f_0(y) f_0(x)}\right).$$

What is different here is the error in estimating $E[Y_t - \rho Y_{t-1} | X_{t-1} = x]$ for example can be large unless ρ is close to one in which case the term $(1 - \rho)u_{t-1}$ is small and the process $Y_t - \rho Y_{t-1}$ is almost stationary. The difference in behaviour of the resulting $\widehat{m}_\rho$ for $\rho = 1$ and $\rho \neq 1$ is what drives the faster rate of convergence for $\widehat{\rho}$.

---

[6]Differencing can be expected to eliminate unit roots so long as enough differencing is undertaken. However, differencing produces additive models for which the optimal estimation strategy is a similar type of method to ours.

[7]Note also that $m_\rho = m$ for all ρ.

Define

$$\widehat{Q}_T(\rho) = \frac{1}{T}\sum_{t=2}^{T}\{Y_t - \rho Y_{t-1} - \widehat{m}_\rho(X_t) + \rho\widehat{m}_\rho(X_{t-1})\}^2$$

and let $\widehat{\rho} = \arg\min_\rho \widehat{Q}_T(\rho)$. We use a subset of the regularity conditions B that are relevant. Let B denote the standard Brownian Motion on $[0, 1]$.

THEOREM 3. *Suppose that assumption B1 holds for X_t, that B2 holds for ε_t, that B3, B7-B9 and B11 hold. Then*

$$T(\widehat{\rho} - 1) \Longrightarrow \frac{\int_0^1 B(s)dB(s)}{\int_0^1 B^2(s)ds}.$$

Furthermore,

$$\sqrt{Th}\left(\widehat{m}(x) - m(x) - h^2 b(x)\right) \Longrightarrow N\left(0, \omega(x)\right),$$

where

$$b(x) = (I - \mathcal{H}_1)^{-1}\left[\frac{f_0'}{f_0}\frac{\partial}{\partial x}\mathcal{H}_1 m + \mathcal{H}_1 m''\right](x)$$

and

$$\omega(x) = ||K||_2^2 \frac{E\left[\varepsilon_t^2\right]}{2f_0(x)}.$$

Note that the asymptotics for $\widehat{\rho}$ are the same as those of the infeasible least squares estimator $\overline{\rho} = \sum_{t=2}^{T} u_t u_{t-1}/\sum_{t=2}^{T} u_{t-1}^2$, so that estimation of m has no effect on the limiting distribution. One can also obtain local to unity asymptotics which are the same as those of $\overline{\rho}$. The distribution theory can be used to perform a test of the null hypothesis of a unit root.

This can be generalized easily to allow for short run dynamics in addition to the unit root. Suppose that in (20), $(1 - L)u_t = C(L)\varepsilon_t$, where $C(L) = \sum_{j=0}^{\infty} c_j L^j$ and $\sum_{j=0}^{\infty} j|c_j| < \infty$. Then by the Beveridge-Nelson decomposition we have $u_t = C(1)\sum_{s=1}^{t}\varepsilon_s + C^*(L)\varepsilon_t$ under our assumptions, where $C^*(L) = \sum_{j=0}^{\infty} c_j^* L^j$ with $c_j^* = -\sum_{i=j+1}^{\infty} c_i$ being summable. Then the result in Theorem 3 follows (for the corresponding estimator) with the correction factor $C(1)$ in the variance.

5   Numerical Results

We investigate the performance of our procedure on simulated data in the context of a nonparametric regression with correlated errors. Our purpose is to compare the performance of our estimator to the natural competitor for that case, the local linear estimator. We focus on the relative performance of two optimally implemented methods to dispense with issues about bandwidth selection and the small sample performance of the benchmark estimator.

We suppose that

$$Y_t = m(X_t) + u_t, \ u_t = \rho_0 u_{t-1} + \varepsilon_t$$

with $m(x) = \beta_0 x^2/2$, where $X_t \sim N(0,1)$, and $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. We take $\beta_0 = 1$ and $\sigma_\varepsilon^2 = 1$. We examine the cases $T \in \{800, 400, 200\}$ and $\rho_0 \in \{0, 0.05, 0.1, \ldots, 0.95, 1.0\}$, and use $ns = 1000$ replications. We compute our estimator $\widehat{m}$ using 200 grid points and use a grid search method to select $\rho \in [\rho_0 - \delta, \rho_0 + \delta]$ for $\delta = 0.2$. We also compute the standard local linear estimator $\widetilde{m}$, in both cases the Gaussian kernel was used.

We chose bandwidth to be optimal according to (asymptotic) weighted mean squared error

$$P_\infty^c(\widehat{m}) = \plim_{T \to \infty} T^{4/5} \int_{-c}^{c} [\widehat{m}(x) - m(x)]^2 f_0(x) dx,$$

which gives $h_{opt} = c_K c_M T^{-1/5}$, where $c_K = (2c\|K\|_2^2/\mu_2^2(K))^{1/5}$ is to do with the kernel and $c_M = (\sigma_\varepsilon^2/(1+\rho_0^2)\beta_0^2(F_0(c) - F_0(-c)))^{1/5}$, where $F_0(x)$ is the c.d.f. of the covariate, is to do with the model. We have taken $c = 2$, which corresponds to an interval containing almost 95% of the covariate distribution. For the standard local linear estimator the optimal bandwidth is $c_K c_M^* T^{-1/5}$ with $c_M^* = (\sigma_\varepsilon^2/(1-\rho_0^2)\beta_0^2(F_0(c) - F_0(-c))^{1/5}$ provided $\rho_0 \neq 1$ (when $\rho_0 = 1$ we set ρ_0 in the formula arbitrarily to 0.95).

In Figure 1 below we report the relative value of the performance measure $P_T(\widehat{m})/P_T(\widetilde{m})$, where

$$P_T(\widehat{m}) = E \int_{-c}^{c} [\widehat{m}(x) - m(x)]^2 f_0(x) dx$$

and where E is computed by the mean or median over Monte Carlo simulations.[8] Both estimators use their optimal bandwidths, and consequently their theoretical relative efficiency is $((1-\rho_0^2)/(1+\rho_0^2))^{4/5}$, which is independent of the other parameters. This is plotted below along with the simulation average value for the different sample sizes against ρ values. The results indicate that $\widehat{m}$ is indeed more efficient than $\widetilde{m}$ and that the advantage takes off after $\rho_0 = 0.8$; until this value the advantage is less than 20% in MSE terms. For small values of ρ_0 the finite sample performance ratio is actually better than predicted, although this is partly because $\widetilde{m}$ performs worse than predicted by its asymptotic theory. Note that when $\rho_0 = 1$ the standard local linear estimator is inconsistent. The relative performance seems to get slightly worse with sample size. The absolute performance of both estimators improves with sample size but the MSE of $\widetilde{m}$ improves more rapidly in the relevant range of sample sizes than does the MSE of $\widehat{m}$.

---

[8]We also examined the IMAE performance measure, but the results are similar.
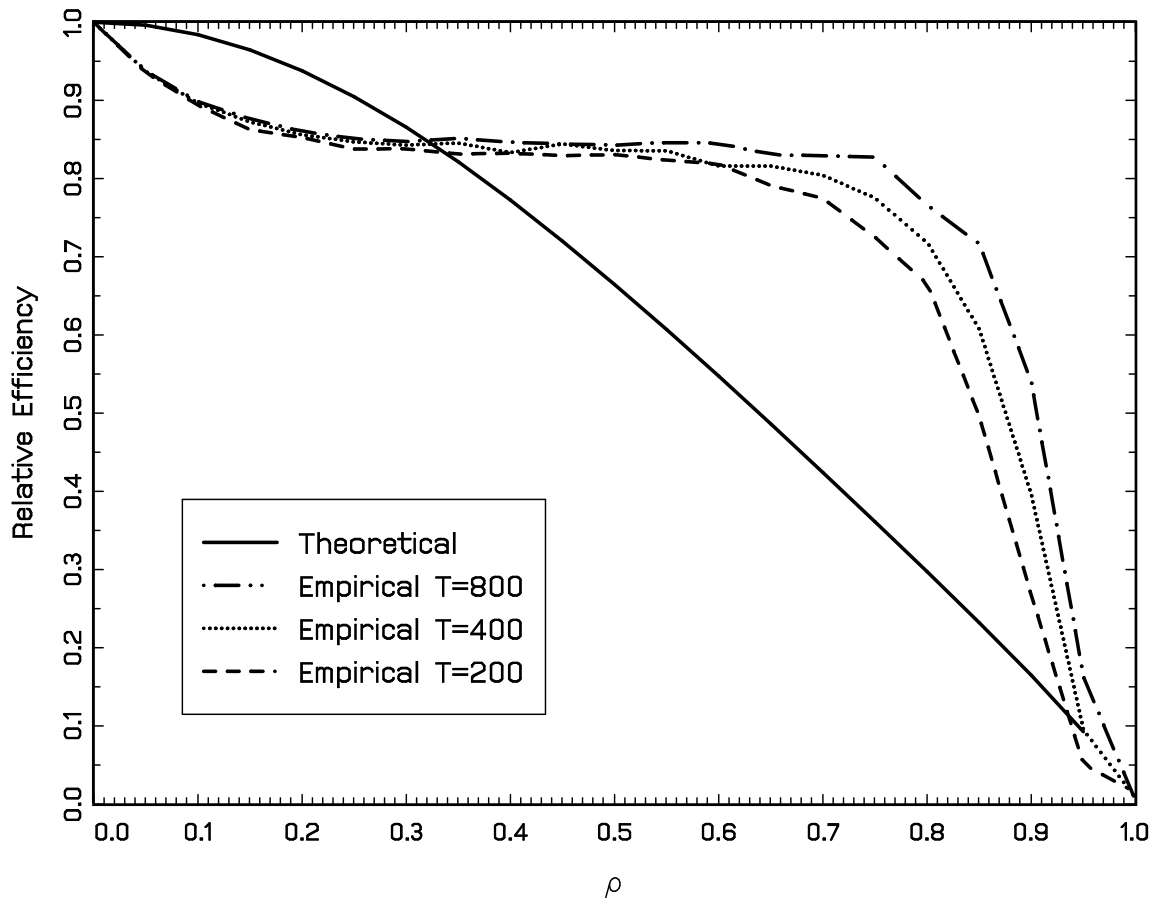
Figure 1. Shows the empirical performance ratio $P_T(\widehat{m})/P_T(\widetilde{m})$ for different sample sizes along with the asymptotic value $P_\infty(\widehat{m})/P_\infty(\widetilde{m})$ predicted from the asymptotic theory. X_t iid $N(0,1)$.

We also looked at the case where X_t is autocorrelated, specifically, $X_t = 0.95X_{t-1} + u_t$, where u_t is normally distributed such that X_t is marginally $N(0,1)$. Theoretically, this does not make any difference, and in practice if anything relative performance is improved for this case. The results are not shown for brevity.

We next examine the performance of $\widehat{\rho}$. When $\rho < 1$ the MSE decreases pretty much as predicted and the distribution approximates a normal for the larger sample size. When $\rho_0 = 1$, our simulations show that the variance of $\widehat{\rho}$ decreases rapidly with sample size with standard deviation being 0.0161, 0.00896, and 0.00458 for $T = 200, 400,$ and 800 respectively, which is consistent with superconsistency. Below we show the qq plots of the empirical quantiles against those of the Dicky-Fuller density in this unit root case. As the sample size increases the distribution approaches the asymptotic distribution.
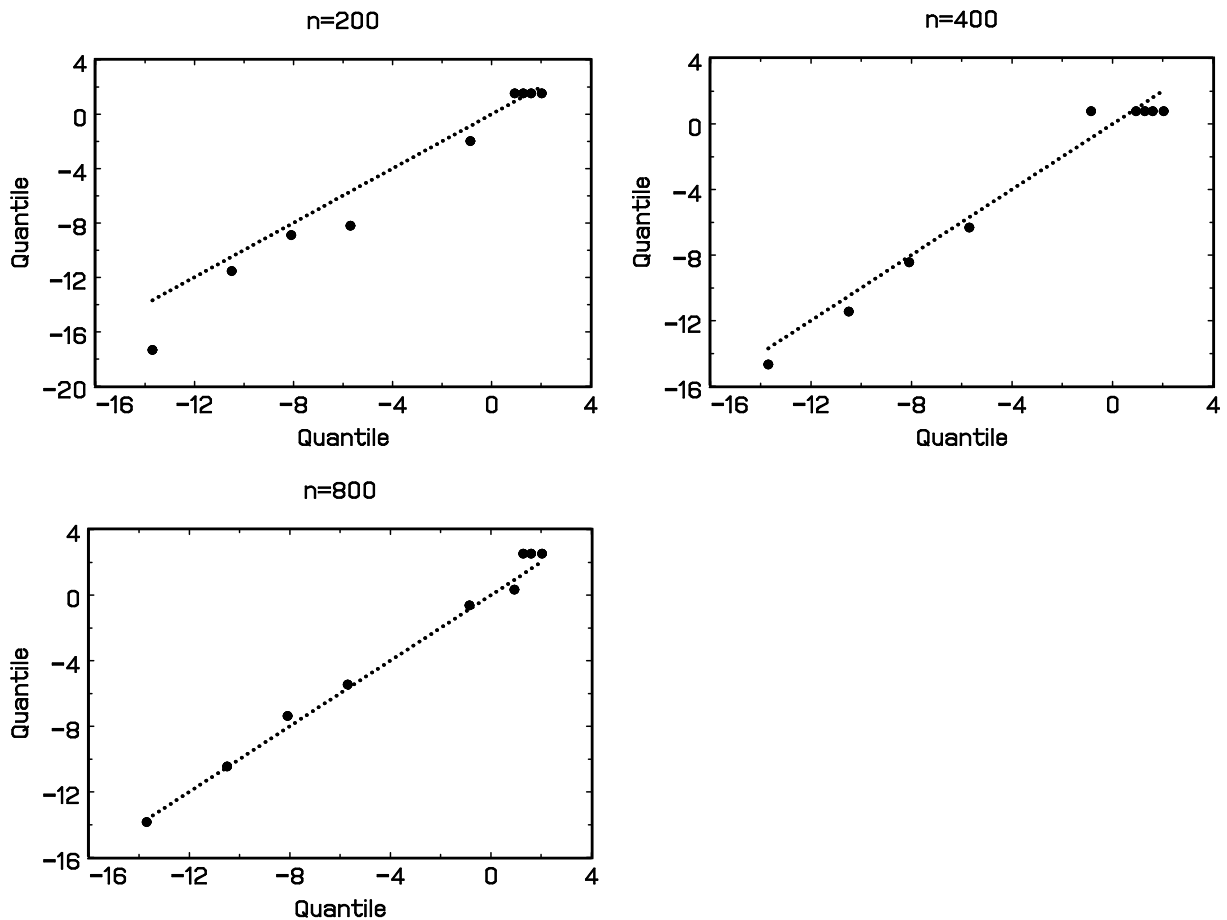
16

Figure 2. Shows the q-q plots of $\widehat{\rho}$ against the Dicky-Fuller density for three different sample sizes.

$$X_t = 0.95 X_{t-1} + u_t \text{ with } X_t \sim N(0,1).$$

Overall these results are much better than obtained in Xiao et al. (2003) in terms of the small sample relative performance, and show in some cases substantial gains over the standard smoothing methods widely employed in practice. However, we acknowledge that in more complicated settings where the order of the polynomials A, B is higher and perhaps has to be determined that the results will worsen.

6 Extensions

We conclude the paper with a discussion of two important extensions.

6.1 Nonstationary X

Suppose that

$$X_t = X_{t-1} + \eta_t$$

17

with η_t also white noise and uncorrelated with ε_t. Thus X_t is a unit root process. This makes a substantial difference to the asymptotics since the corresponding operator $\mathcal{H}_\theta(x, y)$ is random. Provided X_t is null recurrent, we might expect consistency (Phillips and Park (1998)) but the rates of convergence are slower and the asymptotic distributions change. Simulation results support the consistency of $\widehat{m}$. In particular, the corresponding graphic to Figure 1 is almost identical.

An alternative type of nonstationarity for X_t is deterministic trend. Suppose that

$$X_t = \mu(t/T) + \sigma(t/T)\eta_t, \tag{22}$$

where η_t is a stationary mixing process, see Dahlhaus (1997). If $\sigma \equiv 0$, X_t is purely deterministic. In this case, the asymptotics of kernel regression smoother are different and reflect the autocorrelation in u_t, see Hart (1991) and Fan and Yao (2003, Theorem 6.1). Also, there is a problem applying our method because of concurvity. Specifically, we have for any j, $m((t-j)/T) = m(t/T) + O(j/T)$ and so assumption A2 is violated. In this case we have $B(L)Y_t \simeq A(1)m(t/T) + \varepsilon_t$ and there appears to be no estimator that improves over the standard nonparametric regression. This is a bit like the well known result that OLS=GLS when the regressors are polynomial or trigonometric time trends. See Opsomer, Wang, and Yang (2001) for a review of nonparametric methods and results in this case. In the more general locally stationary case, our method may work due to the stochasticness of η_t.

6.2 Multivariate X, Y

When X_t is multivariate the above method can be applied with obvious changes in the dimensionality of various quantities. However, it may be appealing in that case to consider the following model

$$B(L)Y_t = \sum_{j=1}^{d} A_j(L)m_j(X_{jt}) + \varepsilon_t,$$

where the functions $m_1(.), \ldots, m_d(.)$ are unknown and the filters $A_j(L) = \sum_{k=0}^{\infty} a_{jk}$, $j = 1, \ldots, d$. The estimation strategy involves a combination of Mammen, Linton, and Nielsen (1999) and the methods above. Instead one might want to make the function $m(X_{1t}, \ldots, X_{dt})$ obey some other dimensionality reducing restrictions.

A Appendix

A.1 Computational Appendix

We discuss briefly how we solve the equation (13) in practice. Note that one can rewrite (9) as an integral equation on $[0, 1]^2$ as $m_\theta^\dagger(s) = m_\theta^{*\dagger}(s) + \int_0^1 \mathcal{H}_\theta^\dagger(s, t)m_\theta(t)dt$, where $\mathcal{H}_\theta^\dagger(s, t) = \mathcal{H}_\theta(F_0^{-1}(s), F_0^{-1}(t))$

with $y = F_0^{-1}(s)$, $x = F_0^{-1}(t)$ and $m_\theta^\dagger(t) = m_\theta(F_0^{-1}(t))$ and $m_\theta^{*\dagger}(t) = m_\theta^*(F_0^{-1}(t))$ and F_0 is the c.d.f. of X_t. For simplicity we drop the superfluous $\dagger$ superscript in the sequel. Let $\{t_{j,n}, j = 1, \ldots, n\}$ be some equally spaced grid of points in $[0, 1]$, and let $q_{j,n} = \widehat{F}_0^{-1}(t_{j,n})$ be the empirical $t_{j,n}$ quantile of X_t. Now approximate (13) by

$$\widehat{m}_\theta(q_{i,n}) = \widehat{m}_\theta^*(q_{i,n}) + \sum_{j=1}^n \widehat{\mathcal{H}}_\theta(q_{i,n}, q_{j,n})\widehat{m}_\theta(q_{j,n}), \ \ i = 1, \ldots, n. \tag{23}$$

The linear system (23) can be written in matrix notation

$$(I_n - \widehat{\mathbf{H}}_\theta)\widehat{\mathbf{m}}_\theta = \widehat{\mathbf{m}}_\theta^*, \tag{24}$$

where I_n is the $n \times n$ identity, $\widehat{\mathbf{m}}_\theta = (\widehat{m}_\theta(q_{1,n}), \ldots, \widehat{m}_\theta(q_{n,n}))^\top$ and $\widehat{\mathbf{m}}_\theta^* = (\widehat{m}_\theta^*(q_{1,n}), \ldots, \widehat{m}_\theta^*(q_{n,n}))^\top$, while

$$\widehat{\mathbf{H}}_\theta = -\sum_{k=\pm 1}^{\pm\tau} a_k^+(\alpha) \left[\frac{\widehat{f}_{0,k}(q_{i,n}, q_{j,n})}{\widehat{f}_0(q_{i,n})\widehat{f}_0(q_{j,n})} \right]_{i,j=1}^n$$

is an $n \times n$ matrix. We then find the solution values $\widehat{\mathbf{m}}_\theta = (\widehat{m}_\theta(q_{1,n}), \ldots, \widehat{m}_\theta(q_{n,n}))^\top$ to this system (24) by direct inversion when n is less than say 2000.

A.2 Proof of Theorems

A.2.1 Stationary Case

PROOF OF THEOREM 1. The proof strategy uses the general result in Linton and Mammen (2005) for the treatment of empirical integral equations. First, for general θ we apply Linton and Mammen (2005, Proposition 1). Thus we write

$$\widehat{m}_\theta^*(x) - m_\theta^*(x) = \widehat{m}_\theta^{*,B}(x) + \widehat{m}_\theta^{*,C}(x) + \widehat{m}_\theta^{*,D}(x) \tag{25}$$

$$(\widehat{\mathcal{H}}_\theta - \mathcal{H}_\theta)m_\theta(x) = \widehat{m}_\theta^{*,E}(x) + \widehat{m}_\theta^{*,F}(x) + \widehat{m}_\theta^{*,G}(x), \tag{26}$$

where $\widehat{m}_\theta^{*,B}(x)$ and $\widehat{m}_\theta^{*,E}(x)$ are deterministic and $O(T^{-2/5})$,

$$\widehat{m}_\theta^{*,B}(x) = \frac{h^2}{2}\mu_2(K)\left[m_\theta''(x) - \frac{\partial^2}{\partial x^2}(\mathcal{H}_\theta m_\theta)(x)\right]$$

$$\widehat{m}_\theta^{*,E}(x) = \frac{h^2}{2}\mu_2(K)\left[\mathcal{H}_\theta m_\theta''(x) + \frac{\partial^2}{\partial x^2}(\mathcal{H}_\theta m_\theta)(x) + 2\frac{f_0'(x)}{f_0(x)}\frac{\partial}{\partial x}(\mathcal{H}_\theta m_\theta)(x)\right],$$

while:

$$\widehat{m}_\theta^{*,C}(x) = \frac{1}{Tf_0(x)}\sum_t K_h(x, X_t)\eta_{\theta,t}^1$$

19

$$\widehat{m}_\theta^{*,F}(x) = \frac{1}{T f_0(x)} \sum_t K_h(x, X_t) \eta_{\theta,t}^2,$$

and the remainder terms $\widehat{m}_\theta^{*,D}(x)$ and $\widehat{m}_\theta^{*,G}(x)$ satisfy

$$\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}_{int}} \left| \widehat{m}_\theta^{*,j}(x) \right| = o_p(T^{-2/5}), \quad j = D, G,$$

$$\sup_{\theta \in \Theta} \sup_{x \in [\underline{x}, \overline{x}]} \left| \widehat{m}_\theta^{*,j}(x) \right| = O_p(T^{-2/5}), \quad j = D, G,$$

where $\mathcal{X}_{int} = [\underline{x} + h, \overline{x} - h]$.

From this one obtains an expansion

$$\widehat{m}_\theta(x) - m_\theta(x) - \left[m_\theta^B(x) + m_\theta^E(x) + \widehat{m}_\theta^{*,C}(x) + \widehat{m}_\theta^{*,F}(x) \right] = o_p(T^{-2/5}), \tag{27}$$

uniformly for $\theta \in \Theta$ and for $x \in \mathcal{X}_{int}$ and $= O_p(T^{-2/5})$, elsewhere. Here $m_\theta^B = (I - \mathcal{H}_\theta)^{-1} \widehat{m}_\theta^{*,B}$ and $m_\theta^E = (I - \mathcal{H}_\theta)^{-1} \widehat{m}_\theta^{*,E}$. From this expansion we obtain the main result. Specifically, $\widehat{m}_\theta^{*,C}(x) + \widehat{m}_\theta^{*,F}(x)$ is asymptotically normal with zero mean and the stated variance after applying a CLT for near epoch dependent functions of mixing processes. The asymptotic bias comes from $m_\theta^B(x) + m_\theta^E(x)$. Note that because of the boundary modification to the kernel we have $E\widehat{f}_0(x) = f_0(x) + O(h^2)$ and $E\widehat{f}_{0,j}(x,y) = f_{0,j}(x,y) + O(h^2)$ for all x, y.

Our proof below make use of the following results. For $\delta_T = T^{-3/10+\xi}$ with $\xi > 0$ small enough,

$$\max_{1 \le j \le \tau_T} \sup_{x,y \in [\underline{x}, \overline{x}]} \left| \widehat{f}_{0,j}(x,y) - f_{0,j}(x,y) \right| = o_p(\delta_T) \tag{28}$$

$$\sup_{x \in [\underline{x}, \overline{x}]} \left| \widehat{f}_0(x) - f_0(x) \right| = o_p(\delta_T). \tag{29}$$

This follows by the exponential inequality of Bosq (1998, Theorem 1.3), see Linton and Mammen (2005, p817).

PROOF OF (25). Write

$$Z_t(\beta) - Z_t^\tau(\beta) = \sum_{j=\tau+1}^\infty b_j(\beta) Y_{t-j}.$$

We have $E[Z_t(\beta) - Z_t^\tau(\beta)] = E[Y_t] \sum_{j=\tau+1}^\infty b_j(\beta) = O(\overline{b}^\tau)$ and

$$\begin{aligned}
\text{var}\left[Z_t(\beta) - Z_t^\tau(\beta) \right] &= \sum_{j=\tau+1}^\infty \sum_{j'=\tau+1}^\infty b_j(\beta) b_{j'}(\beta) \text{cov}(Y_{t-j}, Y_{t-j'}) \\
&\le \sum_{j=\tau+1}^\infty \sum_{j'=\tau+1}^\infty |b_j(\beta)||b_{j'}(\beta)||\gamma_Y(j - j')| \\
&\le \sup_u |\gamma_Y(u)| \left(\sum_{j=\tau+1}^\infty |b_j(\beta)| \right)^2 = O(\overline{b}^{2\tau}) = o(T^{-1})
\end{aligned}$$

for each β. Similar bounds can be obtained for the covariance $\operatorname{cov}\left[Z_t(\beta) - Z_t^{\tau}(\beta), Z_s(\beta) - Z_s^{\tau}(\beta)\right]$. Let $\widetilde{g}_j(x; \beta)$ denote (12) with $Z_t(\beta)$ replacing $Z_t^{\tau}(\beta)$. Then

$$\max_{1 \leq j \leq \tau} \sup_{\beta \in \mathcal{B}} \sup_{x \in [\underline{x}, \overline{x}]} |\widehat{g}_j(x; \beta) - \widetilde{g}_j(x; \beta)| = o_p(T^{-1/2}). \tag{30}$$

This follows using the above moment bounds and because of the assumed uniform decay rates on $b_j(\beta)$ and its derivatives and the moment condition on Y. See Xiao et al. (2003) for a similar argument.

Then for each $j \geq 0$,

$$\widetilde{g}_j(x; \beta) - g_j(x; \beta) = \frac{1}{T f_0(x)} \sum_{t=1}^{T-j} K_h\left(x, X_t\right) \eta_{t,j}(\beta) + \frac{h^2}{2} \mu_2(K) \mathbf{b}_j(x; \beta) + R_{Tj}(x; \beta),$$

where $\mathbf{b}_j(x; \beta)$ is the bias function and $R_{Tj}(x; \beta)$ is the remainder term, which is $o_p(T^{-2/5})$ uniformly over $j \leq \tau_T$, $x \in [\underline{x}, \overline{x}]$ and $\beta \in \mathcal{B}$. By interchanging the order of summation we obtain for $x \in \mathcal{X}_{int}$

$$
\begin{aligned}
\sum_{j=0}^{\tau} a_j^{\dagger}(\alpha) \sum_{t=1}^{T-j} K_h\left(x, X_t\right) \eta_{t,j}(\beta) &= \sum_{t=1}^{T-\tau} \sum_{j=0}^{T-t} K_h\left(x, X_t\right) a_j^{\dagger}(\alpha) \eta_{t,j}(\beta) \\
&= \sum_{t=1}^{T-\tau} K_h\left(x, X_t\right) \sum_{j=0}^{\infty} a_j^{\dagger}(\alpha) \eta_{t,j}(\beta) \\
&\quad - \sum_{t=1}^{T-\tau} K_h\left(x, X_t\right) \sum_{j=T-t+1}^{\infty} a_j^{\dagger}(\alpha) \eta_{t,j}(\beta) \\
&= \sum_{t=1}^{T-\tau} K_h\left(x, X_t\right) \sum_{j=0}^{\infty} a_j^{\dagger}(\alpha) \eta_{t,j}(\beta) \\
&\quad - \sum_{j=\tau+1}^{T} a_j^{\dagger}(\alpha) \sum_{t=T-j+1}^{T-\tau} K_h\left(x, X_t\right) \eta_{t,j}(\beta) \\
&\quad - \sum_{j=T+1}^{\infty} a_j^{\dagger}(\alpha) \sum_{t=1}^{T-\tau} K_h\left(x, X_t\right) \eta_{t,j}(\beta),
\end{aligned}
$$

where the terms apart from the first are of smaller order. Specifically,

$$\max_{1 \leq j \leq \tau} \sup_{\theta \in \Theta} \sup_{x \in [\underline{x}, \overline{x}]} \left| \frac{1}{T f_0(x)} \sum_{j=\tau+1}^{T} a_j^{\dagger}(\alpha) \sum_{t=T-j+1}^{T-\tau} K_h\left(x, X_t\right) \eta_{t,j}(\beta) \right| = o_p(T^{-2/5}) \tag{31}$$

$$\max_{1 \leq j \leq \tau} \sup_{\theta \in \Theta} \sup_{x \in [\underline{x}, \overline{x}]} \left| \frac{1}{T f_0(x)} \sum_{j=T+1}^{\infty} a_j^{\dagger}(\alpha) \sum_{t=1}^{T-\tau} K_h\left(x, X_t\right) \eta_{t,j}(\beta) \right| = o_p(T^{-2/5}). \tag{32}$$

21

These follow by standard arguments. Therefore,

$$\sum_{j=0}^{\tau} a_j^{\dagger}(\alpha)[\widehat{g}_j(x;\beta) - g_j(x;\beta)] = \frac{1}{Tf_0(x)} \sum_{t=1}^{T-\tau} K_h(x, X_t) \sum_{j=0}^{\infty} a_j^{\dagger}(\alpha)\eta_{t,j}(\beta)$$
$$+ \frac{h^2}{2}\mu_2(K) \sum_{j=0}^{\tau} a_j^{\dagger}(\alpha)\mathbf{b}_j(x;\beta) + o_p(T^{-2/5})$$

uniformly over $x \in [\underline{x}, \overline{x}]$. Claim (25) now follows from

$$\widehat{m}_{\theta}^{*,B}(x) = \frac{h^2}{2}\mu_2(K) \sum_{j=0}^{\tau} a_j^{\dagger}(\alpha)\mathbf{b}_j(x;\beta) + o_p(T^{-2/5}) \text{ uniformly for } x \in \mathcal{X}_{int} \text{ and } \theta \in \Theta,$$

$$\widehat{m}_{\theta}^{*,B}(x) = \frac{h^2}{2}\mu_2(K) \sum_{j=0}^{\tau} a_j^{\dagger}(\alpha)\mathbf{b}_j(x;\beta) + O_p(T^{-2/5}) \text{ uniformly for } x \in x \in [\underline{x}, \overline{x}] \setminus \mathcal{X}_{int} \text{ and } \theta \in \Theta.$$

PROOF OF (26). We have

$$\int \widehat{\mathcal{H}}_{\theta}(x, y)m_{\theta}(y)\widehat{f}_0(y)dy - \int \mathcal{H}_{\theta}(x, y)m_{\theta}(y)f_0(y)dy$$

$$= -\sum_{j=\pm 1}^{\pm \tau} a_j^*(\alpha) \int \left[\frac{\widehat{f}_{0,j}(x, y)}{\widehat{f}_0(x)} - \frac{f_{0,j}(x, y)}{f_0(x)}\right] m_{\theta}(y)dy + \sum_{j=\pm\tau\pm 1}^{\pm\infty} a_j^*(\alpha) \int \frac{f_{0,j}(x, y)}{f_0(x)}m_{\theta}(y)dy$$

$$= -\sum_{j=\pm 1}^{\pm \tau} a_j^*(\alpha) \int \left[\frac{\widehat{f}_{0,j}(x, y)}{\widehat{f}_0(x)} - \frac{f_{0,j}(x, y)}{f_0(x)}\right] m_{\theta}(y)dy + o(T^{-2/5})$$

uniformly over x, θ due to the uniform decay rates on $a_j(\alpha)$. Specifically,

$$\sup_{\theta \in \Theta} \sup_{x \in [\underline{x}, \overline{x}]} \left| \sum_{j=\pm\tau\pm 1}^{\pm\infty} a_j^*(\alpha) \int \frac{f_{0,j}(x, y)}{f_0(x)}m_{\theta}(y)dy \right| \leq C\overline{a}^{\tau} \times \overline{m} = o(T^{-2/5}),$$

where $\sup_{\theta \in \Theta} \sup_{y \in [\underline{x}, \overline{x}]} |m_{\theta}(y)| = \overline{m} < \infty$.

Denote by

$$\int \frac{f_{0,j}(x, y)}{f_0(x)}m_{\theta}(y)dy = E[m(X_{t-j})|X_t = x] \equiv r_j(x).$$

Then write

$$\frac{\int \widehat{f}_{0,j}(x, y)m_{\theta}(y)dy}{\widehat{f}_0(x)} = \frac{\frac{1}{T}\sum_t K_h(x, X_t) m_{t-j}^*}{\frac{1}{T}\sum_t K_h(x, X_t)}, \tag{33}$$

where

$$m_t^* = \int K_h^y(y - X_t)m_{\theta}(y)dy.$$

Then note that for $X_t \in \mathcal{X}_{int}$

$$\int K_h^y(y - X_t) m_\theta(y) dy - m_\theta(X_t) = \int K_h^y(y - X_t) \left[m_\theta(y) - m_\theta(X_t) \right] dy$$

$$= m_\theta'(X_t) \int K_h^y(y - X_t)(y - X_t) dy$$

$$+ \frac{1}{2} \int K_h^y(y - X_t)(y - X_t)^2 m_\theta''(X_t^*(y)) dy$$

$$= \frac{h^2}{2} \mu_2(K) m_\theta''(X_t) + o(h^2)$$

by a second order Taylor expansion, a change of variables and property B9 of the kernels. The error is uniformly $o(h^2)$ over t with $X_t \in \mathcal{X}_{int}$ and θ. Note that (33) is just like a local constant smoother of m_{t-j}^* on X_t and can be analyzed in the same way.

Using $\widehat{a}/\widehat{b} - c = (\widehat{a} - \widehat{b}c)/\widehat{b}$, we have

$$\frac{\int \widehat{f}_{0,j}(x, y) m_\theta(y) dy}{\widehat{f}_0(x)} - \int \frac{f_{0,j}(x, y)}{f_0(x)} m_\theta(y) dy$$

$$= \frac{\frac{1}{T} \sum_t K_h(x, X_t) \left[m_{t-j}^* - r_j(x) \right]}{\frac{1}{T} \sum_t K_h(x, X_t)}$$

$$= \frac{\frac{1}{T} \sum_t K_h(x, X_t) \left[m_\theta(X_{t-j}) - r_j(x) \right]}{\frac{1}{T} \sum_t K_h(x, X_t)} + \frac{\frac{1}{T} \sum_t K_h(x, X_t) \left[m_{t-j}^* - m_\theta(X_{t-j}) \right]}{\frac{1}{T} \sum_t K_h(x, X_t)} \quad (34)$$

$$\simeq \frac{\frac{1}{T} \sum_t K_h(x, X_t) \left[m_\theta(X_{t-j}) - r_j(X_t) \right]}{\frac{1}{T} \sum_t K_h(x, X_t)} + \frac{\frac{1}{Th} \sum_t K_h(x, X_t) \left[r_j(X_t) - r_j(x) \right]}{\frac{1}{Th} \sum_t K_h(x, X_t)}$$

$$+ \frac{h^2}{2} \mu_2(K) E[m_\theta''(X_{t-j}) | X_t = x]$$

$$\simeq \frac{1}{Th} \frac{1}{f_0(x)} \sum_t K_h(x - X_t) \zeta_{t,j} + \frac{h^2}{2} \mu_2(K) \left[r_j''(x) + \frac{2r_j'(x) f_0'(x)}{f_0(x)} + E[m_\theta''(X_{t-j}) | X_t = x] \right] \quad (35)$$

by standard arguments for Nadaraya-Watson smoothers. The approximation is of order $o(T^{-2/5})$, uniformly over $j \leq \tau_T$, over x in $\mathcal{X}_{int}$, and over $\theta \in \Theta$. Summing this up, gives (26) for $x \in \mathcal{X}_{int}$. The proof for the boundary follows by standard arguments. $\qquad \square$

PROOF OF THEOREM 2. The consistency of $\widehat{\theta}$ follows along the lines of Linton and Mammen (2005) using the expansions obtained above uniform over θ. Note that the solution value m_θ is twice continuously differentiable in θ under our assumptions and

$$\frac{\partial m_\theta}{\partial \theta} = \left(\frac{\partial m_\theta^*}{\partial \theta} + \frac{\partial \mathcal{H}_\theta}{\partial \theta} m_\theta\right) + \mathcal{H}_\theta \frac{\partial m_\theta}{\partial \theta} \tag{36}$$

$$\frac{\partial^2 m_\theta}{\partial \theta \partial \theta^\top} = \left(\frac{\partial^2 m_\theta^*}{\partial \theta \partial \theta^\top} + \frac{\partial^2 \mathcal{H}_\theta}{\partial \theta \partial \theta^\top} m_\theta + \frac{\partial \mathcal{H}_\theta}{\partial \theta} \frac{\partial m_\theta}{\partial \theta^\top}\right) + \mathcal{H}_\theta \frac{\partial^2 m_\theta}{\partial \theta \partial \theta^\top}. \tag{37}$$

These define $\partial m_\theta/\partial\theta$ and $\partial^2 m_\theta/\partial\theta\partial\theta^\top$ as solutions to integral equations with different intercepts but the same operator $\mathcal{H}_\theta$ as (10), so the solution to these equations exists and is unique by the arguments given above.

Let $Q(\theta) = Q(\theta, m_\theta)$ with $Q(\theta, m_\theta)$ defined in (11). We first show that

$$\sup_{\theta \in \Theta} \left|\widehat{Q}_T(\theta) - Q(\theta)\right| \xrightarrow{P} 0, \tag{38}$$

which follows from $\sup_{\theta \in \Theta} \sup_{x \in [\underline{x}, \overline{x}]} |\widehat{m}_\theta(x) - m_\theta(x)| \xrightarrow{P} 0$ given the moment and mixing conditions etc. This follows from the expansions in Theorem 1 and standard uniform convergence arguments for kernel smoothers. Specifically, $\sup_{\theta \in \Theta} \sup_{x \in [\underline{x}, \overline{x}]} |\widehat{m}_\theta^{*,j}(x)| = o_p(1), \quad j = B, C$. The uniformity over θ comes from analysis of $\partial m_\theta^*(x)/\partial\theta$ and $\partial \widehat{m}_\theta^{*,j}(x)/\partial\theta$. Then apply assumption B4 to yield consistency of $\widehat{\theta}$.

Define the score function and Hessian

$$\frac{\partial \widehat{Q}_T(\theta)}{\partial \theta} = \frac{1}{T} \sum_{t=2}^{T} \widehat{\varepsilon}_t^\tau(\theta) \frac{\partial \widehat{\varepsilon}_t^\tau(\theta)}{\partial \theta}$$

$$\frac{\partial^2 \widehat{Q}_T(\theta)}{\partial \theta \partial \theta^\top} = \frac{1}{T} \sum_{t=2}^{T} \frac{\partial \widehat{\varepsilon}_t^\tau(\theta)}{\partial \theta} \frac{\partial \widehat{\varepsilon}_t^\tau(\theta)}{\partial \theta^\top} + \widehat{\varepsilon}_t^\tau(\theta) \frac{\partial^2 \widehat{\varepsilon}_t^\tau(\theta)}{\partial \theta \partial \theta^\top},$$

where $\widehat{\varepsilon}_t^\tau(\theta) = Z_t^\tau(\beta) - \sum_{j=0}^{\tau} a_j(\alpha) \widehat{m}_\theta(X_{t-j})$. One then establishes a CLT for the score function at $\theta = \theta_0$ and a local uniform law of large numbers for the Hessian, which establish the CLT for $\widehat{\theta}$.

We can now effectively take $\theta = \theta_0$ in Theorem 1. The asymptotic statement on the distribution of $\widehat{m}_{\theta_0}(x) - m(x)$ directly follows from Theorem 1. Note that $\eta_{\theta,t}^1 + \eta_{\theta,t}^2 = \sum_{j=0}^{\infty} a_j^* \varepsilon_{t+j}$. $\qquad\square$

A.2.2 Nonstationary Case

PROOF OF THEOREM 3. Let

$$\varepsilon_t(\rho) = Y_t - \rho Y_{t-1} - m_\rho(X_t) + \rho m_\rho(X_{t-1}) = Y_t - \rho Y_{t-1} - m(X_t) + \rho m(X_{t-1}) = \varepsilon_t + (1-\rho)u_{t-1}$$

$$\widehat{\varepsilon}_t(\rho) = Y_t - \rho Y_{t-1} - \widehat{m}_\rho(X_t) + \rho \widehat{m}_\rho(X_{t-1}).$$

We have

$$Q_T(\rho) = \frac{1}{T}\sum_{t=2}^{T}\varepsilon_t^2(\rho) = \frac{1}{T}\sum_{t=2}^{T}\varepsilon_t^2 + T(1-\rho)^2\frac{1}{T^2}\sum_{t=2}^{T}u_{t-1}^2 + 2(1-\rho)\frac{1}{T}\sum_{t=2}^{T}\varepsilon_t u_{t-1}$$

$$\simeq \sigma_\varepsilon^2 + T(1-\rho)^2\sigma_\varepsilon^2\int B^2(s)ds + 2(1-\rho)\sigma_\varepsilon^2\int B(s)dB(s).$$

The least squares estimator that minimizes $Q_T(\rho)$, denoted $\bar\rho$, has closed form expression $\bar\rho = \sum_{t=2}^{T}u_t u_{t-1}/\sum_{t=2}^{T}u_{t-1}^2$. It is consistent at rate T and furthermore

$$T(\bar\rho - 1) \Longrightarrow \frac{\int B(s)dB(s)}{\int B^2(s)ds}. \tag{39}$$

We next consider the difference between $\widehat{Q}_T(\rho)$ and $Q_T(\rho)$. We have

$$\widehat{Q}_T(\rho) = Q_T(\rho) + \frac{1}{T}\sum_{t=2}^{T}\{\widehat\varepsilon_t(\rho) - \varepsilon_t(\rho)\}^2 + 2\frac{1}{T}\sum_{t=2}^{T}\{\widehat\varepsilon_t(\rho) - \varepsilon_t(\rho)\}\varepsilon_t(\rho), \tag{40}$$

$$\widehat\varepsilon_t(\rho) - \varepsilon_t(\rho) = -(\widehat{m}_\rho(X_t) - m_\rho(X_t)) + \rho(\widehat{m}_\rho(X_{t-1}) - m_\rho(X_{t-1})).$$

PROOF OF CONSISTENCY. We prove:

$$\widehat{Q}_T(1) \to^p q \tag{41}$$

for some $q > 0$ (hence $\widehat{Q}_T(1)/T \to^p 0$), and

$$\lim_{T\to\infty}\inf_{|\rho-1|>\delta}\frac{1}{T}\widehat{Q}_T(\rho) > 0. \tag{42}$$

Combine (41) and (42) yields $\widehat\rho \xrightarrow{P} 1$.

PROOF OF (41). The properties of $\widehat{Q}_T(1)$ can be derived using the expansion of Theorem 1, and specifically the uniform over x consistency of $\widehat{m}_1(x)$. We have $\widehat{Q}_T(1) \xrightarrow{P} E(\varepsilon_t^2) > 0$.

PROOF OF (42). We first derive the properties of $\widehat{m}_\rho - m_\rho$ for $\rho \neq 1$. As in the stationary case we can approximate $\widehat{m}_\rho - m_\rho$ in terms of $\widehat{m}_\rho^* - m_\rho^*$ and $(\widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho)m_\rho$. The expansion for $(\widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho)m_\rho$ is as above. The main difference concerns the fact that the expansion for $\widehat{m}_\rho^* - m_\rho^*$ contains a term that is large when $\rho \neq 1$ and indeed $\widehat{m}_\rho^*$ does not consistently estimate m_ρ^* unless $\rho = 1$. Therefore, $\widehat{m}_\rho - m_\rho$ is dominated by the large term in $\widehat{m}_\rho^* - m_\rho^*$. Specifically, (26) holds but (25) needs to be modified.

The intercept function m_ρ^* is

$$m_\rho^*(x) = \frac{1}{1+\rho^2}\left(E[Y_t - \rho Y_{t-1}|X_t = x] - \rho E[Y_t - \rho Y_{t-1}|X_{t-1} = x]\right) = \frac{1}{1+\rho^2}\left[g_{0\rho}(x) - \rho g_{1\rho}(x)\right],$$

a linear combination of $g_{0\rho}(x) = E[Y_t - \rho Y_{t-1}|X_t = x]$ and $g_{1\rho}(x) = E[Y_t - \rho Y_{t-1}|X_{t-1} = x]$. Therefore, we must establish the properties of $\widehat{g}_{j\rho}(x) - g_{j\rho}(x)$, $j = 0, 1$, where $\widehat{g}_{j\rho}(x)$ are the estimates of $g_{j\rho}(x)$ when $\rho \neq 1$. We have

$$Y_t - \rho Y_{t-1} - E\left[Y_t - \rho Y_{t-1}|X_t = x\right] = m(X_t) - m(x) - \rho\left(m(X_{t-1}) - E[m(X_{t-1})|X_t = x]\right) + \varepsilon_t + (1-\rho)u_{t-1}.$$

$$Y_t - \rho Y_{t-1} - E\left[Y_t - \rho Y_{t-1}|X_{t-1} = x\right] = m(X_t) - E[m(X_t)|X_{t-1} = x] - \rho\left(m(X_{t-1}) - m(x)\right) + \varepsilon_t + (1-\rho)u_{t-1}.$$

The terms $m(X_t) - m(x)$ and $m(X_{t-1}) - m(x)$ on the rhs contribute to biases; the stationary error terms $-\rho\left(m(X_{t-1}) - E[m(X_{t-1})|X_t = x]\right) + \varepsilon_t$ and $m(X_t) - E[m(X_t)|X_{t-1} = x] + \varepsilon_t$ may contribute to the variance but are standard, it is the term $(1 - \rho)u_{t-1}$ containing the unit root that is different. We have

$$\widehat{g}_{j\rho}(x) - g_{j\rho}(x) = \frac{1}{Tf_0(x)}\sum_{t=j+1}^{T} K_h\left(x, X_{t-j}\right)\varepsilon_t + (1-\rho)\frac{1}{Tf_0(x)}\sum_{t=j+1}^{T} K_h\left(x, X_{t-j}\right)u_{t-1}$$

$$+ \frac{h^2}{2}\mu_2(K)\mathbf{b}_j(x; \rho) + R_T(x; \rho) \equiv \delta_{T1}(x) + \delta_{T2}(x) + \delta_{T3}(x) + R_T(x; \rho),$$

where $\sup_{x \in [\underline{x}, \overline{x}]} \delta_{T1}(x) = O_p(\sqrt{\log T}T^{-2/5})$ and $\sup_{x \in [\underline{x}, \overline{x}]} \delta_{T3}(x) = O_p(T^{-2/5})$ under our bandwidth conditions, while the remainder term is of smaller order than $\delta_{T2}(x)$. This approximation is valid because the X process is stationary so the terms except $\delta_{T2}(x)$ are standard.

We consider the term $\delta_{T2}(x)$ and write $\delta_{T2}(x) = \sqrt{T}(1 - \rho)\xi_T(x) + \sqrt{T}(1 - \rho)\eta_T(x)$ with

$$\xi_T(x) = \frac{1}{T}\sum_{t=1}^{T} E\left[\frac{1}{f_0(x)}K_h\left(x, X_{t-j}\right)\right]\frac{u_{t-1}}{\sqrt{T}}$$

$$\eta_T(x) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{f_0(x)}K_h\left(x, X_{t-j}\right) - E\left[\frac{1}{f_0(x)}K_h\left(x, X_{t-j}\right)\right]\right)\frac{u_{t-1}}{\sqrt{T}}.$$

Clearly, because $E\left[\frac{1}{f_0(x)}K_h\left(x, X_{t-j}\right)\right] = 1 + O(h^2)$ uniformly in x,

$$\xi_T(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{u_{t-1}}{\sqrt{T}} + o_p(1) = O_p(1)$$

uniformly in x.

We argue that $\sup_{x \in [\underline{x}, \overline{x}]} |\eta_T(x)| = o_p(1)$. Note that $E[\eta_T(x)] = 0$ by assumption B11. Define

$$\epsilon_{Tt} = \frac{1}{f_0(x)}K_h\left(x, X_{t-j}\right) - E\left[\frac{1}{f_0(x)}K_h\left(x, X_{t-j}\right)\right]. \tag{43}$$

This has (approximately as $T \to \infty$) covariance function

$$\text{cov}(\epsilon_{Tt}, \epsilon_{Tt-r}) = E\left[\frac{1}{f_0^2(x)}K_h\left(x, X_t\right)K_h\left(x, X_{t-r}\right)\right] - E^2\left[\frac{1}{f_0(x)}K_h\left(x, X_t\right)\right]$$

$$\simeq \frac{f_{0,t-r}(x, x)}{f_0^2(x)} - 1 \equiv \gamma_\epsilon(t - r),$$

by the standard change of variable and dominated convergence argument.

Furthermore,

$$\text{var}\left[\eta_T(x)\right] = \frac{1}{T^3}\sum_{t=j+1}^{T} E\left[\epsilon_{Tt}^2\right] E[u_{t-1}^2] + \frac{1}{T^3}\sum\sum_{t\neq s} E\left[\epsilon_{Tt}\epsilon_{Ts}\right] E[u_t u_s]$$

$$\simeq \frac{\sigma_\varepsilon^2}{T^3}\sum\sum_{t\neq s} \min\{s,t\}\gamma_\epsilon(t-s) \simeq \frac{2\sigma_\varepsilon^2}{T^3}\sum_{s=1}^{T-1} s\sum_{t=s+1}^{T}\gamma_\epsilon(t-s)$$

$$\simeq \frac{2\sigma_\varepsilon^2}{T^2}\sum_{s=1}^{T-1} s(T-s)\sum_{k=1}^{\infty}\gamma_\epsilon(k) = \frac{2\sigma_\varepsilon^2}{3T}\sum_{k=1}^{\infty}\gamma_\epsilon(k),$$

so that $\text{var}\left[\eta_T(x)\right] = O(T^{-1})$ and $\eta_T(x) = O_p(T^{-1/2})$ for each $x \in [\underline{x}, \overline{x}]$. The pointwise result can be extended to uniformity over $x \in [\underline{x}, \overline{x}]$ by standard arguments, so $\sup_{x\in[\underline{x},\overline{x}]}|\eta_T(x)| = o_p(1)$ as required. Therefore

$$\widehat{g}_{j\rho}(x) - g_{j\rho}(x) = \sqrt{T}(1-\rho)\frac{1}{T}\sum_{t=1}^{T}\frac{u_{t-1}}{\sqrt{T}} + o_p(\sqrt{T}). \tag{44}$$

Note that the rhs is the same regardless of location x and j and the error is uniform over these quantities. By the usual arguments (Phillips (1987)), $T^{-3/2}\sum_{t=1}^{T} u_{t-1} \Longrightarrow \sigma_\varepsilon \int_0^1 B(s)ds$. Therefore, $(\widehat{g}_{j\rho}(x) - g_{j\rho}(x))/\sqrt{T} \Longrightarrow (1-\rho)\sigma_\varepsilon \int_0^1 B(s)ds$ for all x and $j = 0, 1$. In fact this convergence is uniform over x.

It holds that:

$$\frac{1}{T^2}\sum_{t=2}^{T}\{\widehat{\varepsilon}_t(\rho) - \varepsilon_t(\rho)\}^2 = \frac{(1-\rho)^6}{(1+\rho^2)^2}\sigma_\varepsilon^2\left(\int_0^1 B(s)ds\right)^2 + o_p(1) \tag{45}$$

$$(1-\rho)\frac{1}{T^2}\sum_{t=2}^{T}\{\widehat{\varepsilon}_t(\rho) - \varepsilon_t(\rho)\}\, u_{t-1} = \frac{-(1-\rho)^4}{1+\rho^2}\sigma_\varepsilon^2\left(\int_0^1 B(s)ds\right)^2 + o_p(1) \tag{46}$$

$$\frac{1}{T}\sum_{t=2}^{T}\{\widehat{\varepsilon}_t(\rho) - \varepsilon_t(\rho)\}\,\varepsilon_t = \frac{-(1-\rho)^3}{1+\rho^2}\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\frac{u_{t-1}}{\sqrt{T}}\varepsilon_t + o_p(1) = O_p(1). \tag{47}$$

We just show the argument for (45). We have

$$\frac{1}{T^2} \sum_{t=2}^{T} \{\widehat{\varepsilon}_t(\rho) - \varepsilon_t(\rho)\}^2$$

$$= \frac{1}{T^2} \sum_{t=2}^{T} \{(\widehat{m}_\rho(X_t) - m(X_t)) - \rho(\widehat{m}_\rho(X_{t-1}) - m(X_{t-1}))\}^2$$

$$= \frac{1}{T^2} \sum_{t=2}^{T} \{(\widehat{m}_\rho^* - m_\rho^*)(X_t) - \rho(\widehat{m}_\rho^* - m_\rho^*)(X_{t-1})\}^2 + o_p(1)$$

$$= \frac{1}{(1+\rho^2)^2} \frac{1}{T^2} \sum_{t=2}^{T} \left\{ [\widehat{g}_{0\rho} - g_{0\rho}](X_t) + \rho^2[\widehat{g}_{1\rho} - g_{1\rho}](X_{t-1}) - \rho[\widehat{g}_{0\rho} - g_{0\rho}](X_{t-1}) - \rho[\widehat{g}_{1\rho} - g_{1\rho}](X_t) \right\}^2$$
$$+ o_p(1)$$

$$= \frac{(1-\rho)^6}{(1+\rho^2)^2} \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{u_{t-1}}{\sqrt{T}} \right\}^2 + o_p(1)$$

by (44). From this (45) follows. The arguments for (46) and (47) are similar.

Then, by (45)-(47) we have

$$\widehat{Q}_T(\rho) \simeq \sigma_\varepsilon^2 + T(1-\rho)^2\sigma_\varepsilon^2 \int_0^1 B^2(s)ds + 2(1-\rho)\sigma_\varepsilon^2 \int_0^1 B(s)dB(s)$$
$$+ \frac{(1-\rho)^6 T}{(1+\rho^2)^2}\sigma_\varepsilon^2 \left(\int_0^1 B(s)ds \right)^2 - \frac{2(1-\rho)^4 T}{1+\rho^2}\sigma_\varepsilon^2 \left(\int_0^1 B(s)ds \right)^2 - \frac{-2(1-\rho)^3}{1+\rho^2}O_p(1).$$

Therefore

$$\frac{1}{T}\widehat{Q}_T(\rho) \simeq (1-\rho)^2\sigma_\varepsilon^2 \int_0^1 B^2(s)ds + \left[\frac{(1-\rho)^6}{(1+\rho^2)^2} - \frac{2(1-\rho)^4}{1+\rho^2} \right] \sigma_\varepsilon^2 \left(\int_0^1 B(s)ds \right)^2$$

$$= (1-\rho)^2\sigma_\varepsilon^2 \int_0^1 B^2(s)ds - \frac{(1-\rho)^4 (\rho+1)^2}{(1+\rho^2)^2}\sigma_\varepsilon^2 \left(\int_0^1 B(s)ds \right)^2.$$

By the Cauchy-Schwarz inequality $\int_0^1 B^2(s)ds \geq \left(\int_0^1 B(s)ds \right)^2$. Therefore, with probability one:

$$\frac{1}{T}\widehat{Q}_T(\rho) \geq 4(1-\rho)^2 \frac{\rho^2}{(1+\rho^2)^2}\sigma_\varepsilon^2 \left(\int_0^1 B(s)ds \right)^2 > 0 \tag{48}$$

for all $\rho \neq 1$. This establishes (42).

28

PROOF OF ASYMPTOTIC DISTRIBUTION. Reparameterizing $\rho \mapsto r = 1 - \rho/T$ we get

$$\widehat{Q}_T(r) \simeq \sigma_\varepsilon^2 + \frac{r^2}{T}\sigma_\varepsilon^2 \int_0^1 B^2(s)ds + 2\frac{r}{T}\sigma_\varepsilon^2 \int_0^1 B(s)dB(s) + o(T^{-1}),$$

so that the terms from the nonparametric estimation drop out. Therefore, the asymptotic distribution is just the Dickey-Fuller, i.e.,

$$T(\widehat{\rho} - 1) \Longrightarrow \frac{\int_0^1 B(s)dB(s)}{\int_0^1 B^2(s)ds}.$$

$\square$

References

[1] BIERENS, H., (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association* 77, 699-707.

[2] BOSQ, D. (1998): *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction.* Springer, Berlin.

[3] CARRASCO, M., J.P. FLORENS, AND E. RENAULT (2002), "Linear Inverse problems in Structural Econometrics," Forthcoming in *Handbook of Econometrics*, volume 6, eds. J.J. Heckman and E. Leamer.

[4] CHEN, AND S.A. BILLINGS (1989). Representation of Nonlinear Systems: The NARMAX model. *International Journal of Control* 56, 319-346.

[5] CONLEY, T.G., L.P. HANSEN, E.G.J. LUTTMER, AND J.A. SCHEINKMAN (1997), "Short-Term Interest Rates as Subordinated Diffusions," *The Review of Financial Studies,* 10, 525-577.

[6] DAHLHAUS, R. (1997): "Fitting time series models to nonstationary processes," Annals of Statistics 25, 1-37.

[7] DHRYMES, P.J. (1971). *Distributed Lags, Problems of Estimation and Formulation.* Holden-Day, San Francisco.

[8] FAN, Y., AND Q. YAO. (2003). *Nonlinear Time Series.* Springer Verlag: Berlin.

[9] GEWEKE, J., (1978). Temporal Aggregation in the Multiple Regression Model," *Econometrica* 46, 643-661.

[10] HANNAN, E.J., AND M. DEISTLER (1988). *The Statistical Theory of Linear Systems.* John Wiley; New York.

[11] HANSEN, L., AND R. HODRICK (1980). Forward exchange rates as optimal predictors of future exchange rates: An economeric analysis. *Journal of Political Economy* 88, 829-853.

[12] HART, J.D. (1991). Kernel Regression Estimation with Time Series Errors. *Journal of the Royal Statistical Society.* 53, 173-187.

[13] HARVEY, A.C. (1981). *The Econometric Analysis of Time Series.* Philip Alan, Oxford.

[14] HASTIE, T. AND R. TIBSHIRANI (1991). *Generalized Additive Models.* Chapman and Hall, London.

[15] HENDRY, D.F., A.R. PAGAN, J.D. SARGAN (1984). Dynamic Specification. in *The Handbook of Econometrics*, vol. 2 Eds. R.F. Engle and D.F. McFadden.

[16] HIDALGO, F.J. (1997). Non-parametric estimation with strongly dependent multivariate time series," Journal of Time Series Analysis 18, 95-122.

[17] HOROWITZ, J.L., J. KLEMELÄ, AND E. MAMMEN (2006): "Optimal Estimation in additive regression,"*Bernoulli*, **12**, 271-298.

[18] JONES, M.C., O.B. LINTON, AND J.P. NIELSEN (1995). A simple bias reduction method for density estimation, *Biometrika* **82**, 327-338.

[19] KRESS, R. (1999). *Linear Integral Equations*, Springer, Berlin.

[20] LIN, X. AND CARROLL, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error," Journal of the American Statistical Association, 95, 520-534.

[21] LINTON, O.B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, **84**, 469-474.

[22] LINTON, O.B., AND E. MAMMEN (2005): "Estimating semiparametric ARCH(∞) models by kernel smoothing methods," *Econometrica* 73, 771-836.

[23] MAMMEN, E., AND B. PARK (2006): A simple smooth backfitting method for additive models. *The Annals of Statistics*, forthcoming

[24] MAMMEN, E., O. LINTON, AND J. P. NIELSON (1999): The existence and asymptotic properties of a backfitting algorithm under weak conditions. *The Annals of Statistics*, **27**, 1443–1490.

[25] MARINUCCI, D., AND P.M. ROBINSON (2003). Semiparametric Frequency Domain Analysis of Fractional Cointegration. in *Time Series with Long Memory*. Oxford University Press. Oxford.

[26] MASRY, E. (1996a). Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and their Applications.* **65**, 81-101.

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

[28] MASRY, E. and Y. FAN (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics* **24**, 165-179.

[29] NEWEY, W.K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica* 62, 1349-1382.

[30] OPSOMER, J., Y. WANG, AND Y. YANG (2001). Nonparametric Regression with Correlated Errors. Manuscript.

[31] PHILLIPS, P.C.B. (1987): "Time Series Regression with a Unit Root," *Econometrica* 55, 277-301.

[32] PHILLIPS, P.C.B., AND J.Y. PARK (1998): "Nonstationary density estimation and kernel autoregression," CFDP no 1181.

[33] ROBINSON, P.M. (1983). Nonparametric Estimators for Time Series. Journal of Time Series Analysis 4, 185-207.

[34] ROBINSON, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56, 931-954.

[35] ROSENBLATT, M. (1956). A central limit theorem and strong mixing conditions, *Proc. Nat. Acad. Sci.* **4**, 43-47.

[36] SAGAN, H. (1969). *Introduction to the Calculus of Variations.* Dover Publications Inc, New York.

[37] SEVERINI, T.A., AND W.H. WONG (1992): "Profile likelihood and conditionally parametric models," *Annals of Statistics* 20, 1768-1802.

[38] SIMS, C.A. (1971). Discrete Approximation to Continuous Time Distributed Lags in Econometrics. *Econometrica* 39, 545-563.

[39] XIAO, Z., O. LINTON, R. J. CARROLL, AND E. MAMMEN (2003). More Efficient Local Polynomial Estimation in Nonparametric Regression with Autocorrelated Errors *Journal of the American Statistical Association* 98, 980-992.