DATA-DRIVEN BANDWIDTH SELECTION FOR NONPARAMETRIC NONSTATIONARY REGRESSIONS

By Federico M. Bandi, Valentina Corradi and Daniel Wilhelm

Johns Hopkins University and Edhec-Risk, University of Warwick and University of Chicago

We provide a solution to the open problem of bandwidth selection for the nonparametric estimation of potentially nonstationary regressions, a setting in which the popular method of cross-validation has not been justified theoretically. Our procedure is based on minimizing moment conditions involving nonparametric residuals and applies to β-recurrent Markov chains, stationary processes being a special case, as well as nonlinear functions of integrated processes. Local and uniform versions of the criterion are proposed. The selected bandwidths are rate-optimal up to a logarithmic factor, a typical cost of adaptation in other contexts. We further show that the bias induced by (near-)minimax optimality can be removed by virtue of a simple randomized procedure. In a Monte Carlo exercise, we find that our proposed bandwidth selection method, and its subsequent bias correction, fare favorably relative to cross-validation, even in stationary environments.

1. Introduction. The vast literature on unit root and cointegration has largely focused on linear models. While it is well-known that the limiting behavior of partial sums, and affine functionals of them, can be approximated by Gaussian processes, much less is known about the asymptotic behavior of functional estimators of nonstationary time series.

Nonparametric regression with nonstationary discrete-time processes has been receiving attention only in recent years. The literature on nonparametric autoregression mainly focuses on β-recurrent Markov chains and heavily uses the number of regenerations of recurrent Markov chains to derive the limiting behavior of the number of visits around a given point (see, e.g., Karlsen and Tjøstheim (2001) and Moloche (2001)). Schienle (2010) considers the case of many regressors and addresses the issue of the curse of dimensionality in the nonstationary case. Guerre (2004) derives convergence rates for a somewhat more general class of recurrent Markov chains. As for nonparametric cointegrating regression, two influential approaches have

AMS 2000 subject classifications: Primary 62M05, 62M10; secondary 62G05

Keywords and phrases: Data-driven Bandwidth Selection, Nonstationary Autoregression, Nonparametric Cointegration, Recurrence

emerged. The first is based on a multidimensional extension of β-recurrent Markov chains and, again, employs the notion of regeneration time (e.g., Karlsen, Myklebust and Tjøstheim (2007)). The second considers nonparametric transformations of integrated and near-integrated processes and uses the occupation density (local time) of partial sums to derive the estimators' asymptotic behavior (e.g., Bandi (2004), Wang and Phillips (2009a,b). There is indeed a parallel literature on the nonparametric estimation of the infinitesimal moment functionals of recurrent diffusion processes (see, e.g., Bandi and Phillips (2003, 2007), and Bandi and Moloche (2008)). On the one hand, in this case, one can possibly exploit the local Gaussianity property of a diffusion processes for the purpose of statistical inference. On the other hand, contrary to the corresponding estimation problem in discrete time, one has to control the rate at which the discrete time interval between adjacent observations goes to zero. Conditions on this rate are needed to approximate the continuous sample path of the underlying process and yield consistency (see, e.g., Bandi, Corradi and Moloche (2009)).

The papers cited above establish consistency and asymptotic mixed normality for kernel estimators of nonstationary autoregressions and cointegrating regressions but provide little practical guidance on bandwidth selection. Guerre (2004) proposes useful adaptive rates (guaranteeing that the bias and variance are of the same order) but does not provide a rule to select the "constant" term and, ultimately, the numerical value of the smoothing sequence. In the context of kernel-based tests for the correct specification of the functional form in a nonstationary environment, Gao et al. (2009) suggest a bootstrap procedure to select the bandwidth parameter which maximizes the local power function, while controlling for size. Their approach, however, may not be employed to find optimal bandwidths for conditional moment kernel estimators.

This paper aims at filling an important – in our opinion – gap in the existing literature by suggesting a procedure for data-driven bandwidth selection in the context of nonparametric autoregressions and nonparametric cointegrating regressions. The proposed method applies to both β-recurrent Markov chains and nonlinear functions of integrated (and stationary) processes. Importantly, while we emphasize the nonstationary (null recurrent) case ($\beta < 1$) for which automated bandwidth procedures have – to the best of our knowledge – not been proposed, the methods are readily applicable to stationary (or positive recurrent) models ($\beta = 1$) for which cross-validation continues to be the most widely-used method of data-driven bandwidth choice.

We offer three contributions. The rate conditions on the bandwidth sequence for asymptotic mixed normality depend on β, the generally unknown regularity of the chain. Although β can be estimated, its estimator converges only at a logarithmic rate (see, e.g., Karlsen and Tjøstheim (2001)). First, we establish that the (generally unknown and process-specific) rate conditions for consistency and asymptotic mixed normality in nonparametric nonstationary autoregressions and nonparametric cointegrating regressions, respectively, can be expressed in terms of the almost-sure rates of divergence of the empirical occupation densities. This set of results provides us with a useful framework to verify the relevant rate conditions empirically and guarantee that they are satisfied in any given sample. Second, we discuss a fully automated method of bandwidth choice. The method consists in selecting the bandwidth vector minimizing a set of sample moment conditions constructed using nonparametric residuals. Even though the limiting rate conditions for mixed asymptotic normality are the same for first and second conditional moment estimation, we allow the search to be over two distinct bandwidth parameters in order to improve finite-sample performance. We show that the resulting adaptive bandwidths are rate-optimal – in the sense of optimally balancing the rates of the asymptotic bias and variance term of the estimator(s) – up to a logarithmic factor, a traditional cost of adaptation in other contexts (see, e.g., Lepskii (1991)). One would generally stop here. However, minimax optimality is, of course, such that the rate condition for zero-mean asymptotic normality will not be satisfied. The presence of an asymptotic bias, as yielded by minimax optimality, may unduly affect statistical inference, something that one might want to rectify for the purpose of superior finite-sample performance. To this extent, third, we propose a simple bias correction relying on a randomized procedure based on conditional inference. The outcome of the latter indicates whether the selected bandwidths satisfy all rate conditions for zero-mean mixed normality or whether, more likely, one should search for smaller bandwidths. We suggest an easy-to-implement stopping rule ensuring that the selected bandwidths are the largest ones, starting from the minimax solution, for which the asymptotic biases are zero.

Two versions of our methods are discussed. The first version selects adaptive bandwidths guaranteeing consistency and mixed normality at a given point of the function of interest and is, therefore, *point-wise* in nature. The second version selects *uniform* bandwidths yielding consistency and mixed normality regardless of the evaluation point.

Finite-sample behavior is analyzed in a Monte Carlo exercise and compared to cross-validation. We show that our methods fare favorably with respect to cross-validation. We view this result as being important. Cross-validation continues to be the most widely-employed approach in empirical work but has not been justified theoretically in the context of nonstationary models. Contrary to cross-validation, which is uniform in nature, the method we provide has a point-wise version leading to local adaptation of the smoothing parameter(s). In its uniform version, our method outperforms cross-validation and applies to nonstationary and stationary models alike, thereby allowing the user to be agnostic about the stationarity feature of the underlying process.

The paper is organized as follows. Section 2 and 3 present asymptotic mixed normality results for nonparametric nonstationary autoregressions and nonparametric cointegrating regressions, respectively. We show how the bandwidth conditions which the extant literature has expressed as functions of the unknown regularity of the chain can be suitably expressed in terms of the almost-sure rate of divergence of the chain's empirical occupation density. Section 4 contains the substantive core of our work and discusses data-driven bandwidth choice in nonstationary, as well as stationary, environments and its minimax optimality properties. Finally, Section 5 provides a simple randomized procedure to adjust the adaptive optimal bandwidths in order to reduce the biases induced by minimax optimality, when it is deemed appropriate to do so. We stress that the suggested bias correction is made possible by our representation of the bandwidth conditions as functions of the process' occupation density (as in Section 2 and 3). All proofs are collected in Appendix A. The supplementary document Bandi, Corradi and Wilhelm (2011) reports the findings of a Monte Carlo study.

2. Nonparametric Nonstationary Autoregression. Intuitively, one can estimate conditional moments, evaluated at a given point, only if a neighborhood of that point is visited infinitely often as time grows. Otherwise, not enough information is gathered. For this reason, it is natural to focus attention on irreducible recurrent chains, i.e., chains satisfying the property that, at any point in time, the neighborhood of each point has a strictly positive probability of being visited and, eventually, it will be visited an infinite number of times. For positive recurrent chains, the expected time between two consecutive visits is finite. Hence, the time spent in the neighborhood of a point grows linearly with the sample size, n say. For null recurrent chains, the expected time between two consecutive visits is infinite. Therefore, the time spent in the neighborhood of a point grows at a rate, possibly random, which is slower than n. Since, up to some mild regularity conditions, positive recurrent chains are strongly mixing, consistency and asymptotic normality follow by, e.g., Robinson (1983) and bandwidth selection may be implemented, as is customary in much empirical work, by virtue of cross-validation. Nonparametric regression with null recurrent chains, however, poses substantial theoretical challenges since the amount of time spent in the neighborhood of a point is not only unknown but also random.

In an important contribution, Karlsen and Tjøstheim (2001) derive consistency and mixed asymptotic normality for conditional moment estimators in the case of null recurrent Markov chains. This is accomplished via split chains, i.e., by splitting the chain into identically and independently distributed components. The number of these iid components, i.e. the number of complete regenerations, T_n say, is of the same almost-sure order as the time spent in the neighborhood of each point.

Let $\mu(X_{t-1}) = E[X_t|X_{t-1}]$ and $\sigma^2(X_{t-1}) = Var(X_t|X_{t-1}) = E[\tilde{u}_t^2|X_{t-1}]$ so that X_t can be written as

$$X_t = \mu(X_{t-1}) + \tilde{u}_t = \mu(X_{t-1}) + \sigma(X_{t-1})u_t,$$

where u_t is such that $E[u_t|\mathcal{F}_t] = 0$ and $E[u_t^2|\mathcal{F}_t] = 1$ given the filtration $\mathcal{F}_t = \sigma(X_{t-1}, X_{t-2}, \ldots)$. Now, define the two estimators

$$\hat{\mu}_{n,h_n^\mu}(x) = \frac{\sum_{j=1}^n X_j K_{h_n^\mu}(X_{j-1} - x)}{\sum_{j=1}^n K_{h_n^\mu}(X_{j-1} - x)},$$
$$\hat{\mu}^{(2)}_{n,h_n^\sigma}(x) = \frac{\sum_{j=1}^n X_j^2 K_{h_n^\sigma}(X_{j-1} - x)}{\sum_{j=1}^n K_{h_n^\sigma}(X_{j-1} - x)},$$

and $\hat{\sigma}^2_{h_n}(x) = \hat{\mu}^{(2)}_{n,h_n^\sigma}(x) - (\hat{\mu}_{n,h_n^\mu}(x))^2$. Here, K is some kernel function and $K_h(x) = \frac{1}{h}K(\frac{x}{h})$. We rely on the following Assumption which largely corresponds to Assumption B_0-B_4 in Karlsen and Tjøstheim (2001).

Assumption 1. (i) Let $\{X_t, t \geq 0\}$ be a β-recurrent, ϕ-irreducible Markov chain on a general state space $(\mathbf{E}, \mathcal{E})$ with transition probability P. Let $\beta \in (0, 1]$.

(ii) The invariant measure π_s has a locally twice continuously differentiable density p_s which is locally strictly positive, i.e., $p_s(x) > 0$.

(iii) The kernel function K is a bounded density with compact support satisfying $\int uK(u)du = 0$ and $K_2 = \int K^2(u)du < \infty$. The set $\mathcal{N}_x = \{y : K_{h=1}(y - x) \neq 0\}$ is a small set (see Karlsen and Tjøstheim (2001)) for all $x \in \mathcal{D}_x$, where $\mathcal{D}_x$ is a compact set in $\mathbb{R}$ so that $\mathcal{D}_x = \{x : p_s(x) > \delta\}$ with $\delta > 0$ arbitrarily small and independent of x.

(iv) We have $\lim_{h\downarrow 0} \overline{\lim}_{y\to x} P(y, A_h) = 0$ for all sets $A_h \in \mathcal{E}$ so that $A_h \downarrow \emptyset$ when $h \downarrow 0$.

(v) The functions $\mu(x)$ and $\sigma^2(x)$ are locally twice continuously differentiable for all $x \in \mathcal{D}_x$.

For $i = \mu, \sigma$, define the estimator of the occupation density as

$$\hat{L}_{n,h_n^i}(x) = \frac{1}{h_n^i} \sum_{j=1}^{n} K_{h_n^i}\left(X_{j-1} - x\right). \tag{1}$$

Here as well as below, we occasionally omit the superscripts μ and σ of the bandwidths when the discussion applies to both or if it is clear from the context which of the two is referred to. In the positive recurrent case ($\beta = 1$), as $n \to \infty$ and $h_n \to 0$ with $nh_n \to \infty$, $\hat{L}_{n,h_n}(x)/n \overset{a.s.}{\to} \varphi(x)$, where $\varphi(x)$ is the density associated with the time-invariant probability measure. Whenever $0 < \beta < 1$ under Assumption 1(i)-(iii) and provided $n \to \infty$ and $h_n \to 0$ with $h_n n^\beta u(n) \to \infty$, $\hat{L}_{n,h_n}(x)/(n^\beta u(n)) \overset{d}{\to} c_X \mathcal{M}_\beta$, where c_X is a process-specific constant, $\mathcal{M}_\beta$ is the Mittag-Leffler density with parameter β, and the positive function $u(\cdot)$ defined on $[b, \infty)$, with $b \geq 0$, is a slowly-varying function at infinity. In this case, both the rate of divergence of the occupation density $\hat{L}_{n,h_n}(x)$, namely $n^\beta u(n)$, and the features of the asymptotic distribution, $\mathcal{M}_\beta$, depend on the degree of recurrence β. Similarly, $T_n/(n^\beta u(n)) \overset{d}{\to} \mathcal{M}_\beta$, where T_n is, as earlier, the number of complete regenerations.

THEOREM 1. *Let Assumption 1 hold and let* $(E[X_t^2|X_{t-1}])^{2m} < \infty$ *for some* $m \geq 2$ *and* X_{t-1} *in a neighborhood of* x *with* $x \in \mathcal{D}_x$.

(a) If (i) $h_n^\mu \hat{L}_{n,h_n^\mu}(x) \overset{a.s.}{\to} \infty$ *and (ii)* $(h_n^\mu)^5 \hat{L}_{n,h_n^\mu}(x) \overset{a.s.}{\to} 0$, *then*

$$\sqrt{h_n^\mu \hat{L}_{n,h_n^\mu}(x)}\left(\hat{\mu}_{n,h_n^\mu}(x) - \mu(x)\right) \overset{d}{\to} N\left(0, \sigma^2(x) K_2\right).$$

(b) If (i) $h_n^\sigma \hat{L}_{n,h_n^\sigma}(x) \overset{a.s.}{\to} \infty$ *and (ii)* $(h_n^\sigma)^5 \hat{L}_{n,h_n^\sigma}(x) \overset{a.s.}{\to} 0$, *then*

$$\sqrt{h_n^\sigma \hat{L}_{n,h_n^\sigma}(x)}\left(\hat{\mu}^{(2)}_{n,h_n^\sigma}(x) - \mu^{(2)}(x)\right) \overset{d}{\to} N\left(0, \left(\mu^{(4)}(x) - \left(\mu^{(2)}(x)\right)^2\right) K_2\right).$$

REMARK 1. *In Theorem 1, and in analogous results below, the condition* $h_n^5 \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} C$, *where* C *is a constant, would give rise to an asymptotic bias which is a function of the process' invariant measure as well as a function of the moment being estimated.*

The statement in the theorem above is similar to that in Theorem 5.4 in Karlsen and Tjøstheim (2001). However, Karlsen and Tjøstheim (2001) state the bandwidth conditions as $h_n n^{\beta-\varepsilon} \to \infty$ and $h_n^5 n^{\beta+\varepsilon} \to 0$. Their rate conditions are sufficient, not necessary. In fact, as is clear from their proofs, they require $h_n T_n \overset{a.s.}{\to} \infty$ and $h_n^5 T_n \overset{a.s.}{\to} 0$, where the number of regenerations T_n is at least of almost-sure order $n^{\beta-\varepsilon}$ and at most of almost-sure order $n^{\beta+\varepsilon}$. Now, in general, β is unknown and, although it can be estimated, its proposed estimator only converges at a logarithmic rate and thus may not be overly useful in practice (Karlsen and Tjøstheim (2001), Remark 3.7). Having made these points, it is empirically important to express the rate conditions on the smoothing sequences in terms of estimated occupation densities, as we do in Theorem 1. The key argument used in the proof of Theorem 1 is that $h_n \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ and $h_n^5 \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ if, and only if, $h_n a(n) \to \infty$ and $h_n^5 a(n) \to 0$ respectively, with

$$a(n) = n^\beta (\log\log n^\beta u(n))^{1-\beta} u(n \log\log n^\beta u(n))$$

and $u(\cdot)$ denoting a slowly-varying function at infinity. Since $a(n)$ defines the almost-sure rate of the number of regenerations, the argument implies that our assumptions are equivalent to expressing the rates in terms of the (random) number of regenerations. The "if" part is somewhat more intuitive. In essence, if $h_n a(n) \to \infty$, then $\hat{L}_{n,h_n}(x)/a(n)$, under mild regularity conditions, satisfies a strong law of large numbers, and thus $\hat{L}_{n,h_n}(x) = O_{a.s.}(a(n))$. As for the less intuitive "only if" part, it follows from the fact that, as shown in the Appendix, $\hat{L}_{n,h_n}(x) = O_{a.s.}(a(n)) + O_p(\sqrt{a(n)/h_n})$ and so $h_n \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ only if $h_n a(n) \to \infty$.

In Section 4, in order to show selection of a (local or global) near rate-optimal bandwidth, we require uniform consistency of the first two conditional moment estimators. The needed result is contained in the following theorem.

THEOREM 2. *Let Assumption 1 hold and let* $(E[X_t^2|X_{t-1}])^{2m} < \infty$ *for some* $m \geq 2$ *and* X_{t-1} *in a neighborhood of* x *for all* $x \in \mathcal{D}_x$. *Then:*

(a)

$$\sup_{x\in\mathcal{D}_x} \left|\hat{\mu}_{n,h_n^\mu}(x) - \mu(x)\right| = O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\mu}(x) h_n^\mu}}\right) + O\left((h_n^\mu)^2\right),$$

(b)

$$\sup_{x\in\mathcal{D}_x} \left|\hat{\mu}^{(2)}_{n,h_n^\sigma}(x) - \mu^{(2)}(x)\right| = O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\sigma}(x) h_n^\sigma}}\right) + O\left((h_n^\sigma)^2\right).$$

3. Nonparametric Cointegrating Regression. We now consider the following data generating process:

$$Y_t = f(X_t) + \alpha(X_t)\epsilon_t. \tag{2}$$

It is immediate to see that, whenever X_t is a null recurrent Markov process or, using a more traditional modeling approach, an integrated processes, and ϵ_t is short-memory, the data generating process in (2) can be viewed as a nonlinear generalization of the classical cointegrating equation. In general, Y_t and X_t are jointly dependent, as they both belong to a larger structural model, and, consequently, ϵ_t is not independent of X_t. In this sense, nonparametric estimation of nonlinear cointegrating regressions is a somewhat more complicated task than nonparametric nonstationary autoregression.

As mentioned, there are two main approaches to nonparametric cointegrating regression. In the first approach, Karlsen, Myklebust and Tjøstheim (2007) assume that X_t is a β-recurrent Markov chain and extend the methodology outlined in the previous section to the multivariate case and to the possible endogeneity of ϵ_t. Bandi (2004) and Wang and Phillips (2009a,b), instead, work under the assumption that X_t is an integrated or a near-integrated process. The interplay between the two methods is discussed in Bandi (2004).

If X_t is an integrated process, then, as $n \to \infty$ and $h_n \to 0$ so that $h_n n^{1/2} \to \infty$, $\hat{L}_{n,h_n}(x)/n^{1/2} \xrightarrow{d} L_0(0,1)$, where $L_0(0,1)$ is the local time of a Brownian motion at 0 between 0 and 1, i.e. the amount of time spent by the process around zero between time 0 and time 1. Compared to the β-recurrent case, this is a more explicit representation of the empirical occupation density's limiting behavior, which results from the stronger (but, in nonstationary econometrics, more conventional) I(1) structure of the underlying process. Clearly, when setting $\beta = 1/2$ (the Brownian motion case) in the first approach, we obtain $\mathcal{M}_{1/2} \stackrel{d}{=} L_0(0,1)$, where $\stackrel{d}{=}$ denotes equivalence in distribution. The common distribution is that of a truncated Gaussian random variable on a positive support.

Now, define the two estimators

$$\hat{f}_{n,h_n^\mu}(x) = \frac{\sum_{j=1}^n Y_j K_{h_n^\mu}(X_j - x)}{\sum_{j=1}^n K_{h_n^\mu}(X_j - x)},$$

$$\hat{f}^{(2)}_{n,h_n^\sigma}(x) = \frac{\sum_{j=1}^n Y_j^2 K_{h_n^\sigma}(X_j - x)}{\sum_{j=1}^n K_{h_n^\sigma}(X_j - x)}$$

and $\hat{\alpha}^2_{n,h_n}(x) = \hat{f}^{(2)}_{n,h_n^\sigma}(x) - (\hat{f}_{n,h_n^\mu}(x))^2$. When X_t is β-recurrent, $E[\epsilon_t|\mathcal{F}_t] = 0$ and $E[\epsilon_t^2|\mathcal{F}_t] = 1$ with $\mathcal{F}_t = \sigma(X_{t-1}, X_{t-2}, \ldots)$, the statement in Theorem 1 extends rather straightforwardly to the cointegrating regression case. In fact, if $E[\epsilon_t|X_t]=0$ and ϵ_t is geometrically strong mixing, given Assumption 1(i)-(v), consistency and asymptotic mixed normality follow directly from Theorem 3.5 in Karlsen, Myklebust and Tjøstheim (2007) by simply setting "their" k equal to 0. Under analogous assumptions, Moloche (2001) establishes consistency and asymptotic mixed normality for local linear and local polynomial estimators of nonlinear cointegrating regressions driven by recurrent Markov chains. For the case of (near-)integrated processes, whenever $E[\epsilon_t|\mathcal{F}_t]=0$, consistency and mixed asymptotic normality are established in Wang and Phillips (2009a).

We now turn to the endogenous case in which ϵ_t is no longer a martingale difference sequence but is, instead, correlated with X_t. For completeness, we consider both approaches in the extant literature. We begin by evaluating the case in which X_t is a β-recurrent Markov chain. We then focus on the integrated (or near-integrated) case.

In what follows, we make use of Assumption 2 which largely corresponds to Assumptions D_1-D_5 in Karlsen, Myklebust and Tjøstheim (2007) and builds on Assumption 1.

ASSUMPTION 2. (i) The joint process $\{(X_t,\epsilon_t),\, t\geq 0\}$ is a ϕ-irreducible Harris recurrent Markov chain on the state space $(\tilde{\mathbf{E}},\tilde{\mathcal{E}})=(\mathbf{E}_1\times\mathbf{E}_2,\mathcal{E}_1\otimes\mathcal{E}_2)$ with marginal transition probabilities P_1 and P_2. The invariant measure of the joint process $\pi(s)$ has a density p_s with respect to the two-dimensional Lebesgue measure so that $\int p_s(x,\epsilon)d\epsilon>0$, $\lim_{\delta\downarrow 0}\int|p_s(x+\delta,\epsilon)-p_s(x,\epsilon)|d\epsilon=0$ and, for all $A_h\in\tilde{\mathcal{E}}^{\infty}$ such that $A_h\downarrow\emptyset$, $\lim_{h\downarrow 0}\overline{\lim}_{y\to x}\int_{\epsilon}P((y,\epsilon),A_h)|\epsilon|d\epsilon=0$.

(ii) The marginal process X_t satisfies Assumption 1(i) and Assumption 1(iii)-(iv). In addition, the marginal transition probability function P_1 is independent of any initial distribution λ. The kernel function satisfies Assumption 1(iii).

(iii) The residual ϵ_t has bounded support for all t.

(iv) (a) $\int\epsilon p_{\epsilon|X}(\epsilon|x)d\epsilon=0$ and (b) $\int\epsilon^2 p_{\epsilon|X}(\epsilon|x)d\epsilon=1$.

(v) The functions $f(x)$ and $\alpha(x)$ are locally twice continuously differentiable for all $x\in\mathcal{D}_x$.

Assumptions 2(i)-(ii) are a multivariate extension of Assumption 1. Assumption 2(iii) – bounded support of ϵ – is used in the proof of Theorem 4.1 in Karlsen, Myklebust and Tjøstheim (2007), a result which we will refer to below. Their simulation results, however, indicate that its violation does not have any practical effect. Assumption 2(iv)(a) qualifies the degree of dependence between X_t and ϵ_t. Even though it seems a rather stringent requirement, it is satisfied whenever (i) X_t and ϵ_t are *asymptotically independent*, in the sense that the joint invariant measure of (X_t, ϵ_t) can be factorized into the product of the corresponding two marginal measures, and (ii) the integral of ϵ with respect to the invariant measure is equal to zero. In this case, in fact, $\int \epsilon p_{\epsilon|X}(\epsilon|x)d\epsilon = \int \epsilon p_s(x,\epsilon)/p_s(x)d\epsilon = \int \epsilon p_s(\epsilon)d\epsilon = 0$. Clearly, asymptotic independence does not imply independence. One important implication of asymptotic independence is the following. Since X_t is null recurrent and, loosely speaking, its variability increases with t, while ϵ_t is short-memory and its variability does not depend on t, we allow for a situation where, analogously to the linear case, $E[\epsilon_t|X_t] \neq 0$ but is a decreasing function of t, so that $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n E[\epsilon_t|X_t] = 0$ a.s.. Similarly, $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n E[\epsilon_t^2|X_t] = 1$ a.s..

THEOREM 3. *Let Assumption 2 be satisfied. Further, assume that we have* $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n (E[Y_t^2|X_t])^{2m} < \infty$ *for some* $m \geq 2$ *and* X_t *in a neighborhood of* x *with* $x \in \mathcal{D}_x$, *and* $\int \overline{\lim}_{y\to x} |\partial^2 p_s(y,\epsilon)/\partial y^2||\epsilon| d\epsilon < \infty$.

(a) If (i) $h_n^\mu \hat{L}_{n,h_n^\mu}(x) \overset{a.s.}{\to} \infty$ *and (ii)* $(h_n^\mu)^5 \hat{L}_{n,h_n^\mu}(x) \overset{a.s.}{\to} 0$ *then*

$$\sqrt{h_n^\mu \hat{L}_{n,h_n^\mu}(x)}\left(\hat{f}_{n,h_n^\mu}(x) - f(x)\right) \overset{d}{\to} N\left(0, \alpha^2(x) K_2\right).$$

(b) If (i) $h_n^\sigma \hat{L}_{n,h_n^\sigma}(x) \overset{a.s.}{\to} \infty$ *and (ii)* $(h_n^\sigma)^5 \hat{L}_{n,h_n^\sigma}(x) \overset{a.s.}{\to} 0$, *then*

$$\sqrt{h_n^\sigma \hat{L}_{n,h_n^\sigma}(x)}\left(\hat{f}^{(2)}_{n,h_n^\sigma}(x) - f^{(2)}(x)\right) \overset{d}{\to} N\left(0, \left(f^{(4)}(x) - \left(f^{(2)}(x)\right)^2\right) K_2\right).$$

Theorem 3(a) is adapted from Theorem 4.1 in Karlsen, Myklebust and Tjøstheim (2007). As earlier, to provide a feasible bandwidth selection procedure, we show that our rate conditions $h_n \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ and $h_n^5 \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ are almost-surely equivalent to $h_n a(n) \to \infty$ and $h_n^5 a(n) \to 0$.

It should be pointed out that, whenever ϵ_t is not a martingale difference sequence, one can no longer interpret $f(x)$ and $f^{(2)}(x)$ as conditional (on x) first and second moments. However, under Assumption 2(iv), one can interpret $f(x)$ as $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n E[Y_t|X_t = x]$ and, similarly, $f^{(2)}(x)$ as $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n E[Y_t^2|X_t = x]$, with probability one.

The corresponding uniform result, needed in the next section, is contained in the following theorem.

THEOREM 4. *Let Assumption 2 and* $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n (E[Y_t^2|X_t])^{2m} < \infty$ *hold for some* $m \geq 2$ *and* X_t *in a neighborhood of* x *for all* $x \in \mathcal{D}_x$. *If*

$$\sup_{x\in\mathcal{D}_x}\left|\sqrt{\frac{h_n^\mu}{a(n)\log(n)}}\sum_{t=1}^n E\left[K_{h_n^\mu}(X_t - x)\epsilon_t \alpha(X_t)\right]\right| = O(1) \tag{3}$$

as well as $\inf_{x\in\mathcal{D}_x} p_s(x) \geq \delta > 0$*, then:*

(a)

$$\sup_{x\in\mathcal{D}_x}\left|\hat{f}_{n,h_n^\mu}(x) - f(x)\right| = O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\mu}(x)h_n^\mu}}\right) + O\left((h_n^\mu)^2\right).$$

(b) If, in addition,

$$\sup_{x\in\mathcal{D}_x}\left|\sqrt{\frac{h_n^\sigma}{a(n)\log(n)}}\sum_{t=1}^{n} E\left[K_{h_n^\sigma}(X_t - x)\,\alpha(X_t)\left(\epsilon_t^2 - 1\right)\right]\right| = O(1), \tag{4}$$

then

$$\sup_{x\in\mathcal{D}_x}\left|\hat{f}^{(2)}_{n,h_n^\sigma}(x) - f^{(2)}(x)\right| = O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\sigma}(x)h_n^\sigma}}\right) + O\left((h_n^\sigma)^2\right).$$

The statement in Theorem 4 is similar to that in Theorem 4.2 in Gao, Li and Tjøstheim (2009). We, however, show how the rates can be stated in terms of estimated occupation densities. Further, we establish sharper rates, but only in probability, and over a compact set, while they establish almost-sure rates over an increasing set. The uniform rate result above relies on a strengthening of Assumption 2(iv). We simply require the dependence between X_t and ϵ_t to go to zero fast enough.

We now turn to the case in which X_t is an integrated process, not necessarily Markov, and ϵ_t in (2) is not independent of X_t. Assumption 3(ii)-(iv) below corresponds to Assumptions 2-4 in Wang and Phillips (2009b) while Assumption 3(i) is a strengthened version of their Assumption 1. We explain below why we use this stronger version and outline what would happen if, instead, we were to use their Assumption 1.

Assumption 3. (i) $X_t = X_{t-1} + \xi_t$, $\xi_t = \sum_{k=0}^{\infty}\phi_k\eta_{t-k}$, where (a) $E[|\xi_t|^{2(4+\gamma)}] \leq C_1 < \infty$ for $\gamma > 0$, (b) η_k is iid, (c) ϕ_k decays fast enough, as $k \to \infty$, as to ensure that ξ_t is α-mixing with size $-(4(4+\gamma))/\gamma$, and (d) there exists $0 < \omega_0^2 < \infty$ so that $|T^{-1}E[(\sum_{k=m+1}^{m+T}\xi_k)^2] - \omega_0^2| \leq C_2T^{-\psi}$, with $\psi > 0$ and C_2 independent of m.

(ii) K is a second-order kernel, bounded and with bounded support, and $\int|e^{ixt}K(t)dt|dx < \infty$.

(iii) ϵ_t as defined in (2) writes as $\epsilon_t = g(\eta_t, \ldots, \eta_{t-m_0})$, where g is a measurable function on $\mathbb{R}^{m_0}$ and $m_0 < \infty$. In addition, $\eta_t = 0$ for $t = 1, \ldots, m_0 - 1$, $E[\epsilon_t] = 0$, $E[\epsilon_t^2] = 1$ and $E[\epsilon_t^4] < \infty$.

(iv) The functions $f(x)$ and $\alpha(x)$ are locally twice continuously differentiable for all $x \in \mathcal{D}_x$.

THEOREM 5. *Let Assumption 3 hold.*

(a) If (i) $h_n^\mu \hat{L}_{n,h_n^\mu}(x) \xrightarrow{a.s.} \infty$ *and (ii)* $(h_n^\mu)^5 \hat{L}_{n,h_n^\mu}(x) \xrightarrow{a.s.} 0$, *then*

$$\sqrt{h_n^\mu \hat{L}_{n,h_n^\mu}(x)} \left(\hat{f}_{n,h_n^\mu}(x) - f(x)\right) \xrightarrow{d} N\left(0, \sigma^2(x) K_2\right).$$

(b) If (i) $h_n^\sigma \hat{L}_{n,h_n^\sigma}(x) \xrightarrow{a.s.} \infty$ *and (ii)* $(h_n^\sigma)^5 \hat{L}_{n,h_n^\sigma}(x) \xrightarrow{a.s.} 0$, *then*

$$\sqrt{h_n^\sigma \hat{L}_{n,h_n^\sigma}(x)} \left(\hat{f}^{(2)}_{n,h_n^\sigma}(x) - f^{(2)}(x)\right) \xrightarrow{d} N\left(0, \left(f^{(4)}(x) - \left(f^{(2)}(x)\right)^2\right) K_2\right).$$

The statement in Theorem 5(a) builds on that in Theorem 3.1 in Wang and Phillips (2009b). Again, their bandwidth conditions are stated in the somewhat more familiar form $\sqrt{n} h_n \to \infty$ and $\sqrt{n} h_n^5 \to 0$. In the proof, we show that $h_n \hat{L}_{n,h_n}(x) \xrightarrow{a.s.} \infty$ if, and only if, $\sqrt{n} h_n \to \infty$ and that $h_n^5 \hat{L}_{n,h_n}(x) \xrightarrow{a.s.} 0$ only if $\sqrt{n} h_n^5 \to 0$. On the other hand, $\sqrt{n} h_n^5 \to 0$ implies $h_n^5 \hat{L}_{n,h_n}(x) \xrightarrow{p} 0$, not necessarily $h_n^5 \hat{L}_{n,h_n}(x) \xrightarrow{a.s.} 0$. Theorem 5(b) follows naturally.

Contrary to the general β-recurrent case, for which β is unknown, in the I(1) case ($\beta = \frac{1}{2}$) one could in principle set the bandwidth parameter equal to $h_n = cn^{-1/10}$ in order to balance the variance and the squared bias term (see, e.g., Bandi (2004)). Alternatively, one could set $h_n = cn^{-(1/10+\varepsilon)}$, with $\varepsilon > 0$ arbitrarily small, to ensure that the bias is asymptotically negligible. Several issues, however, arise. First, choosing the constant term c appropriately is a non-trivial applied problem. Classical rules-of-thumb may, for instance, be imprecise and cross-validation has not been justified for this type of problems. Second, for empirically reasonable sample sizes n, it may be better to set the bandwidth parameter as a function of the occupation density rather than as a function of n. In other words, it may be better to rely on the effective number of visits the process makes at a point, rather than on the notional divergence rate of the occupation density ($\sqrt{n}$). Lastly, in general, one does not know whether X_t is I(1) rather than I(0). If a preliminary unit-root test is run, and the null of a unit root is not rejected, then one may assume that $\hat{L}_{n,h_n}(x)$ diverges at rate $\sqrt{n}$. If the null is rejected in favor of stationarity, however, then $\hat{L}_{n,h_n}(x)$ diverges at rate n. Now, it is well known that unit-root tests have little power against I(0) alternatives characterized by a root close to, but strictly below, one. Importantly, under our rate conditions, the statements in Theorem 5 hold even if X_t in Assumption 3(i) is replaced by $X_t = \alpha X_{t-1} + \xi_t$ with $|a| \leq 1$. Hence, Theorem 5, like Theorems 1-4 above, applies to both the stationary and nonstationary case. We believe that avoiding pre-testing for a unit root and/or stationarity may be empirically useful.

It should be pointed out that Assumption 1 in Wang and Phillips (2009b) allows for near-integrated processes, i.e., $X_t = \exp(c/n)X_{t-1} + \xi_t$ with $c \le 0$. In our context, we could allow for $c < 0$ at the cost of stating our rate conditions as $h_n \hat{L}_{n,h_n}(x) \xrightarrow{p} \infty$ and $h_n^5 \hat{L}_{n,h_n}(x) \xrightarrow{p} 0$, i.e. by weakening the almost-sure rates to rates in probability. Thus, in practical applications, we can employ $\hat{L}_{n,h_n}(x)$, instead of $\sqrt{n}$, even in the case of near-integrated processes. Finally, we establish uniform consistency.

THEOREM 6. *Let Assumption 3 and (3) hold with* $a(n) = n^{1/2}$, *and suppose* $h_n^\mu E[\exp\{K_{h_n^\mu}(X_t - x)\alpha(X_t)\epsilon_t\}] \le \Delta < \infty$. *Then*

(a)

$$\sup_{x \in \mathcal{D}_x} \left| \hat{f}_{n,h_n^\mu}(x) - f(x) \right| = O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\mu}(x) h_n^\mu}} \right) + O\left((h_n^\mu)^2 \right).$$

(b) If, in addition, (4) holds with $a(n) = n^{1/2}$ *and* $h_n^\sigma E[\exp\{K_{h_n^\sigma}(X_t - x)\alpha(X_t)(\epsilon_t^2 - 1)\}] \le \Delta < \infty$, *then*

$$\sup_{x \in \mathcal{D}_x} \left| \hat{f}^{(2)}_{n,h_n^\sigma}(x) - f^{(2)}(x) \right| = O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\sigma}(x) h_n^\sigma}} \right) + O\left((h_n^\sigma)^2 \right).$$

The uniform rate in Theorem 6(a) requires two additional conditions. The first condition, controlling the rate at which the dependence between ϵ_t and $(X_1, \ldots, X_t)$ approaches zero, allows us to treat the term $K_{h_n}(X_t - x)\alpha(X_t)\epsilon_t$ as a martingale difference sequence. The second is a Cramèr-type condition permitting the use of exponential inequalities for unbounded martingales, e.g., Lesigne and Volný (2001). If either condition fails to hold, we would have a less sharp uniform rate. Analogous additional conditions are required for the uniform consistency of the conditional second moment.

4. Adaptive Bandwidth Selection. To the best of our knowledge, there are no data-driven procedures for choosing the bandwidth in the case of nonparametric nonstationary autoregressions or nonparametric cointegrating regressions. In spite of being used widely in empirical work, cross-validation, or suitable modifications of cross-validation, have not been formally justified in a nonstationary framework. However, an important contribution in this area is the work by Guerre (2004) who, under slightly different assumptions, suggests a bandwidth based on the minimization of the empirical bias-variance trade-off. In terms of our notation, Guerre's adaptive bandwidth is defined as

$$\hat{h}_n(x; L, \sigma^2) = \min\left\{ h \geq 0 \text{ s.t. } L^2 h^2 \sum_{j=1}^{n} 1\left\{|X_j - x| \leq h\right\} \geq \sigma^2 \right\},$$

where L is the Lipschitz constant characterizing the conditional expectation function, i.e. $|\mu(x) - \mu(x')| \leq L|x - x'|$ and σ^2 is so that $E[u_i^2|X_i] \leq \sigma^2$. The selected bandwidth is a function of two constants, L and σ^2, which are, in general, unknown. It is, therefore, not feasible in practice.

Our goal is to select a bandwidth which may or may not depend on the evaluation point (and, hence, is point-wise or uniform in nature) but does not require the choice of unknown quantities, such as L and σ^2, and is, in that sense, fully data-driven. We begin by outlining the case in which we select a local bandwidth which depends on the evaluation point. Let

$$\hat{u}_{i,h_n} = \frac{X_i - \hat{\mu}_{n,h_n^\mu}(X_{i-1})}{\hat{\sigma}_{n,h_n}(X_{i-1})} \text{ and } \hat{\epsilon}_{i,h_n} = \frac{Y_i - \hat{f}_{n,h_n^\mu}(X_i)}{\hat{\alpha}_{n,h_n}(X_i)},$$

$w_{i,h_n^\mu}(x) = 1\{|X_i - x| < h_n^\mu\} / \sum_{i=1}^n 1\{|X_i - x| < h_n^\mu\}$ as well as

$$\hat{m}^u_{n,h_n}(x) = \begin{pmatrix} \sum_{i=1}^n \hat{u}_{i,h_n} w_{i-1,h_n^\mu}(x) \\ \sum_{i=1}^n \hat{u}^2_{i,h_n} w_{i-1,h_n^\mu}(x) - 1 \end{pmatrix},$$
$$\hat{m}^\epsilon_{n,h_n}(x) = \begin{pmatrix} \sum_{i=1}^n \hat{\epsilon}_{i,h_n} w_{i,h_n^\mu}(x) \\ \sum_{i=1}^n \hat{\epsilon}^2_{i,h_n} w_{i,h_n^\mu}(x) - 1 \end{pmatrix},$$

where $h_n = (h_n^\mu, h_n^\sigma)$. We begin with the case of nonparametric autoregression. The bandwidth vector $\hat{h}_n$ is selected as:

$$\hat{h}_n(x) = \left(\hat{h}_n^\mu(x), \hat{h}_n^\sigma(x)\right) = \arg \inf_{h_n(x) \in \mathcal{H}_n(x,\varsigma)} \left\| \hat{m}^u_{n,h_n}(x) \right\|, \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm and we define $\mathcal{H}_n(x,\varsigma) = \{h_n(x) : \log(n)/(\hat{L}_{n,h_n^{\mu,\sigma}}(x)\varsigma) < h_n^{\mu,\sigma}(x) < \varsigma\}$ for some $\varsigma > 0$. It is easy to see that $\sum_{i=1}^n \hat{u}_{i,h_n} w_{i-1,h_n^\mu}(x) = \sum_{i=1}^n u_{i,h_n} w_{i-1,h_n^\mu}(x) + o_p(1)$ if, and only if, $|\hat{\mu}_{n,h_n^\mu}(x) - \mu(x)| = o_p(1)$. Similarly, for the second moment condition, we have $\sum_{i=1}^n \hat{u}^2_{i,h_n} w_{i-1,h_n^\mu}(x) = \sum_{i=1}^n u^2_{i,h_n} w_{i-1,h_n^\mu}(x) + o_p(1)$ if, and only if, $|\hat{\mu}_{n,h_n^\mu}(x) - \mu(x)| = o_p(1)$ and $|\hat{\mu}^{(2)}_{n,h_n^\sigma}(x) - \mu^{(2)}(x)| = o_p(1)$. One can also show that $\sum_{i=1}^n u_{i,h_n} w_{i-1,h_n^\mu}(x) = o_p(1)$ and $\sum_{i=1}^n u^2_{i,h_n} w_{i-1,h_n^\mu}(x) = 1 + o_p(1)$ since $E[u_i|X_{i-1}] = 0$ and $E[u_i^2|X_{i-1}] = 1$. Thus, the bandwidth vector selected according to (5) ensures the consistency of the first two conditional moment estimators. Given Assumption 1, such a bandwidth vector exists. Furthermore, we will show that the selected bandwidth vector is rate-optimal, in the sense of optimally balancing the rates of the asymptotic bias and variance terms of the estimator(s), up to a logarithmic factor.

The nonparametric cointegrating case is treated analogously, defining

$$\tilde{h}_n(x) = \left(\tilde{h}_n^{\mu}(x), \tilde{h}_n^{\sigma}(x)\right) = \arg \inf_{h_n(x) \in \mathcal{H}_n(x,\varsigma)} \left\| \hat{m}^{\epsilon}_{n,h_n}(x) \right\|. \tag{6}$$

As already pointed out, we wish to allow for $E[\epsilon_i|X_i] \neq 0$. Nonetheless, under either Assumption 2(iv) in the β-recurrent case, or Assumption 3(i) and Assumption 3(iii) in the case of integrated processes, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n E[\epsilon_i|X_i] \to 0$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n E[(\epsilon_i^2 - 1)|X_i] \to 0$. These conditions ensure that $\sum_{i=1}^n \hat{\epsilon}_{i,h_n} w_{i,h_n^{\mu}}(x) = o_p(1)$ and $\sum_{i=1}^n \hat{\epsilon}^2_{i,h_n} w_{i,h_n^{\mu}}(x) = 1 + o_p(1)$ if, and only if, $|\hat{f}_{n,h_n^{\mu}}(x) - f(x)| = o_p(1)$ and $|\hat{f}^{(2)}_{n,h_n^{\sigma}}(x) - f^{(2)}(x)| = o_p(1)$. Moreover, under additional conditions (in Theorems 4 and 6) on the rate at which $E[\epsilon_i|\mathcal{F}_i]$ and $E[(\epsilon_i^2-1)|\mathcal{F}_i]$ approach zero, $\tilde{h}_n(x)$ is also rate-optimal up to a logarithmic factor.

It is evident from the definition of $\hat{h}_n(x)$ and $\tilde{h}_n(x)$ that we can be silent about stationarity or the degree of recurrence of the process. The criteria to be minimized, in fact, simply depend on the estimated occupation densities.

Theorem 7. *Assume that the kernel K is twice continuously differentiable on the interior of its support.*

(a) Nonparametric Autoregression: Under the assumptions of Theorem 2, $\hat{h}_n^i(x)$, $i = \mu, \sigma$, as defined in (5), is at least of probability order $\gamma(n)^{-1/5}$ and at most of probability order $\log^{1/5}(n)\gamma(n)^{-1/5}$. In the positive recurrent (ergodic) case, $\beta = 1$ and $\gamma(n) = n$, while in the null recurrent case, $\gamma(n) = a(n)$ with

$$a(n) = n^{\beta}(\log\log n^{\beta}u(n))^{1-\beta}u(n\log\log n^{\beta}u(n)) \tag{7}$$

and $u(\cdot)$ a slowly-varying function at infinity.

(b) Nonparametric cointegration: Either (b1) the assumptions in Theorem 4 hold or (b2) the assumptions in Theorem 6 hold. Then, $\tilde{h}_n^i(x)$, $i = \mu, \sigma$, as defined in (6), is at least of probability order $\gamma(n)^{-1/5}$ and at most of probability order $\log^{1/5}(n)\gamma(n)^{-1/5}$, where $\gamma(n)$ is defined as in part (a) if (b1) holds or is $\gamma(n) = n^{1/2}$ if (b2) holds.

Remark 2. *As established in Theorem 7, the adaptive bandwidths obtained by the minimization of the above moment conditions are rate-optimal up to a logarithmic factor. This result holds for stationary processes, integrated processes, and general β-recurrent processes. The logarithmic factor is* *the same cost of adaptation as in, e.g., Lepskii (1991), Lepski, Mammen and Spokoiny (1997) and Lepski and Spokoiny (1997) in other contexts. These methods generally lead to the choice of the largest bandwidth for which the bias is sufficiently small. Their criteria require a choice of threshold, something that is not needed in our framework.*

Theorem 7 proposes a data-driven procedure for selecting a variable bandwidth vector ensuring point-wise consistent estimation. The theorem below establishes that there exist rate-optimal (again, up to a logarithmic factor) uniform bandwidths. Let $\mathcal{H}_n(\varsigma) = \{h_n : \log(n)/(\inf_{x\in\mathcal{D}_x} \hat{L}_{n,h_n^{\mu,\sigma}}(x)\varsigma) < h_n^{\mu,\sigma} < \varsigma\}$ with some $\varsigma > 0$ and

$$\hat{h}_n = \left(\hat{h}_n^{\mu}, \hat{h}_n^{\sigma}\right) = \arg \inf_{h_n \in \mathcal{H}_n(\varsigma)} \sup_{x\in\mathcal{D}_x} \left\| \hat{m}^u_{n,h_n}(x) \right\|, \tag{8}$$

and

$$\tilde{h}_n = \left(\tilde{h}_n^{\mu}, \tilde{h}_n^{\sigma}\right) = \arg \inf_{h_n \in \mathcal{H}_n(\varsigma)} \sup_{x\in\mathcal{D}_x} \left\| \hat{m}^\epsilon_{n,h_n}(x) \right\|. \tag{9}$$

THEOREM 8. *Assume that the kernel K is twice continuously differentiable on the interior of its support.*

(a) Nonparametric autoregression: Under the assumptions of Theorem 2, $\hat{h}_n^i$, $i = \mu, \sigma$, as defined in (8), is of probability order $\log^{1/5}(n)\gamma(n)^{-1/5}$, where $\gamma(n)$ is defined as in Part (a) of Theorem 7.

(b) Nonparametric cointegration: Either (b1) the assumptions in Theorem 4 hold or (b2) the assumptions in Theorem 6 hold. Then, $\tilde{h}_n^i$, $i = \mu, \sigma$, as defined in (9), is of probability order $\log^{1/5}(n)\gamma(n)^{-1/5}$, where $\gamma(n)$ is defined as in Part (a) of Theorem 7 if (b1) holds or is $\gamma(n) = n^{1/2}$ if (b2) holds.

5. Bias Correction.

5.1. *A point-wise test.* The proposed adaptive bandwidths are large enough as to ensure the consistency of the estimators of the first two conditional moments. In light of their minimax optimality, they are too large to satisfy the condition for zero-mean asymptotic normality. If centering of the estimator's asymptotic distribution is of interest, the a researcher may be willing to endure a larger, than optimal, mean-squared error in order to achieve some bias reduction. To this extent, we now propose a theoretical approach to bias reduction which effectively evaluates the magnitude of the minimax bandwidths and, if needed, reduces them to the point where the estimators' biases are zero. Said differently, starting from the minimax bandwidths, the procedure selects the largest bandwidth for which the estimator's asymptotic bias is zero.

We note that, in nonlinear cointegration, for example, the condition for a zero bias is $h_n^5 a(n) \to 0$. This condition is clearly not operational in that the regularity of the chain is, in general, unknown. However, we have shown in Theorems 1 and 3 that $h_n^5 a(n) \to 0$ if, and only if, $h_n^5 \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$. The mapping between bandwidth conditions written in terms of the almost-sure divergence rate of number of regenerations and bandwidth conditions based on the divergence rate of the empirical occupation density (as implied by Theorems 1, 3 and 5) is not just of theoretical interest. It provides empirical content to otherwise theoretical statements. In this section, for example, we show that a bias correction is made possible by our representation of the bandwidths as functions of the process' occupation density (as in Sections 2 and 3). The goal is to verify if $h_n^5 \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ and, consequently by Theorems 1 and 3, $h_n^5 a(n) \to 0$.

We begin with the point-wise bandwidths. Let $\hat{h}_n(x) = (\hat{h}_n^{\mu}(x), \hat{h}_n^{\sigma}(x))$ be the bandwidth vector previously selected. Because the bandwidth rate conditions are the same for both conditional moments, we only consider $\hat{h}_n^{\mu}(x)$ (expressed as $\hat{h}_n^{\mu}$) for conciseness. This said, the procedure outlined below should be separately applied to both bandwidth sequences for bias reduction. In addition, the method works in the same manner for both nonparametric autoregressions and nonparametric cointegrating regressions.

The hypothesis of interest is

$$H_0^{\mu}(x): \ (\hat{h}_n^{\mu})^{5-\varepsilon} \hat{L}_{n,\hat{h}_n^{\mu}}(x) \overset{a.s.}{\to} \infty$$

where $x \in \mathcal{D}_x$, and $\varepsilon > 0$ arbitrarily small, versus

$$H_A^{\mu}(x): \ \text{negation of } H_0.$$

$H_0^{\sigma}(x)$ and $H_A^{\sigma}(x)$ are defined in an analogous way, simply replacing $\hat{h}_n^{\mu}$ by $\hat{h}_n^{\sigma}$. The role of $\varepsilon > 0$ is to ensure that rejection of the null implies $(\hat{h}_n^{\mu})^5 \hat{L}_{n,\hat{h}_n^{\mu}}(x) \overset{a.s.}{\to} 0$. Without $\varepsilon > 0$, $(\hat{h}_n^{\mu})^5 \hat{L}_{n,\hat{h}_n^{\mu}}(x)$ could be $O_{a.s.}(1)$ under the alternative, which would not guarantee a zero asymptotic bias. It is immediate to see that, if we reject the null, the selected bandwidth satisfies the required rate condition for a vanishing asymptotic bias ($(\hat{h}_n^{\mu})^5 \hat{L}_{n,\hat{h}_n^{\mu}}(x) \overset{a.s.}{\to} 0$) and it should be kept. If we fail to reject, then we need to select a smaller bandwidth.

REMARK 3. *The notation* $h_n^5 \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ *clearly highlights that the same bandwidth* $\hat{h}_n^{\mu}(x)$ *should appear outside and inside of the kernel. As* *implied by the proof of Theorems 1, 3, and 5, failure to do so would result in fundamental theoretical inconsistencies. Again, this theoretical statement has important empirical implications. In classical stationary regression models, one could set the bandwidth as being proportional to $n^{-1/5}$, where n is the number of observations ($h_n \propto n^{-1/5}$). This is what plug-in procedures and rules-of-thumb do, for example. The analogous procedure in our framework, unless one is willing to assume a value for the regularity of the chain, would be to set the bandwidth as being proportional to the estimated occupation density $\hat{L}^{\theta}_{n,h_n^{\mu}}(x)$ for some θ ($h_n \propto \hat{L}^{\theta}_{n,h_n^{\mu}}(x)$). However, this is not immediately possible since the bandwidth used to define $\hat{L}^{\theta}_{n,h_n^{\mu}}(x)$ is the same bandwidth that one is aiming to choose in the first place. Said differently, the same bandwidth appears on both sides of the expression $h_n \propto \hat{L}^{\theta}_{n,h_n^{\mu}}(x)$ resulting in an unavoidable circularity. Plug-in procedures based on asymptotic MSEs and classical rules-of-thumb are therefore even less operational in nonstationary environments than they are in stationary environments. Our methods are intended to offer a solution to these issues.*

We construct a test of the above null hypothesis similarly as Bandi, Corradi and Moloche (2009). To that end, define

$$\tilde{V}_{R,n} = \int_U V_{R,n}^2(u)\pi(u)du,$$

with $U = [\underline{u}, \overline{u}]$ being a compact set, $\int_U \pi(u)du = 1$, $\pi(u) \geq 0$ for all $u \in U$,

$$V_{R,n}(u) = \frac{2}{\sqrt{R}} \sum_{j=1}^{R} \left(1\left\{v_{j,n} \leq u\right\} - \frac{1}{2}\right)$$

and

$$v_{j,n} = \left(\exp\left\{(\hat{h}_n^{\mu})^{5-\varepsilon} \sum_{j=1}^{n} K_{\hat{h}_n^{\mu}}\left(X_j - x\right)\right\}\right)^{1/2} \eta_j,$$

with $\eta \sim N(0, I_R)$ iid. The compact set U and the weight function $\pi(\cdot)$ ought to be chosen by the user. Examples are provided in the Monte Carlo section in the Supplementary Material (Bandi, Corradi and Wilhelm (2011)).

In what follows, let the symbols P^* and d^* denote convergence in probability and in distribution under P^*, which is the probability law governing the simulated random variables η, i.e., a standard normal, *conditional* on the data. Also, let E^* and Var^* denote the mean and variance operators under P^*. Furthermore, the notation $a.s. - P$ is used to mean "for all data but a set of measure 0".

Suppose that $(\hat{h}_n^\mu)^5 \sum_{j=1}^n K_{\hat{h}_n^\mu}(X_j - x) \overset{a.s.}{\to} \infty$. Then, conditionally on the sample and $a.s. - P$, $v_{j,n}$ diverges to ∞ with probability $1/2$ and to $-\infty$ with probability $1/2$. Thus, as $n \to \infty$, for any $u \in U$, $1\{v_{j,n} \leq u\}$ is distributed as a Bernoulli random variable with parameter $1/2$. Furthermore, note that, as $n \to \infty$, for any $u \in U$, $1\{v_{j,n} \leq u\}$ is equal to either 1 or 0 regardless of the evaluation point u. In consequence, as $n, R \to \infty$, for all $u, u' \in U$, $\frac{2}{\sqrt{R}} \sum_{j=1}^R (1\{v_{j,n} \leq u\} - \frac{1}{2})$ and $\frac{2}{\sqrt{R}} \sum_{j=1}^R (1\{v_{j,n} \leq u'\} - \frac{1}{2})$ converges in d^*-distribution to the same standard normal random variable. Thus, $\tilde{V}_{R,n} \overset{d^*}{\to} \chi_1^2 \; a.s. - P$. It is now immediate to notice that, for all $u \in U$, $V_{R,n}(u)$ and $\tilde{V}_{R,n}$ have the same limiting distribution. The reason why we are averaging over U is simply because the finite sample type I and type II errors may indeed depend on the particular evaluation point. As for the alternative, if $(\hat{h}_n^\mu)^5 \sum_{j=1}^n K_{\hat{h}_n^\mu}(X_j - x) \overset{a.s.}{\to} 0$, (or, if $(\hat{h}_n^\mu)^5 \sum_{j=1}^n K_{\hat{h}_n^\mu}(X_j - x) = O_{a.s.}(1)$), then $v_{j,n}$, as $n \to \infty$, conditionally on the sample and $a.s. - P$, converges to a (mixed) zero-mean normal random variable. Thus, $\frac{2}{\sqrt{R}} \sum_{j=1}^R (1\{v_{j,n} \leq u\} - \frac{1}{2})$ diverges to infinity at speed $\sqrt{R}$ whenever $u \neq 0 \; a.s. - P$.

THEOREM 9. *Let Assumption 1, 2, or 3 hold. As $R, n \to \infty$,*

(a) Under $H_0^\mu(x)$,

$$V_{R,n} \overset{d^*}{\to} \chi_1^2 \quad a.s. - P.$$

(b) Under $H_A^\mu(x)$, there are $\varepsilon_1, \varepsilon_2 > 0$ so that

$$P^* \left(R^{-1+\varepsilon_1} V_{R,n} > \varepsilon_2 \right) \to 1 \quad a.s. - P.$$

REMARK 4. *In general, R can grow at a faster rate than n. Only, in the case in which $h_n^\mu \sum_{j=1}^n K_{h_n^\mu}(X_j - x)$ diverges at a logarithmic rate, then $R/n \to 0$.*

If one fails to reject $H_0^\mu(x)$ because $V_{R,n}$ is smaller than, say, the 95% percentile of a χ_1^2 random variable, then we suggest choosing a smaller bandwidth until rejection is reached. Specifically, we suggest searching on a grid until $H_0^\mu(x)$ is rejected, i.e., until reaching

$$\hat{\hat{h}}_n^\mu(x) = \max \left\{ h < \hat{h}_n^\mu : \text{ s.t. } H_0^\mu(x) \text{ is rejected} \right\}.$$

Starting from the minimax bandwidth, the proposed rule leads to the choice of the largest bandwidth ensuring a zero asymptotic bias. We stop searching as soon as we reject the null and the likelihood of rejecting the null is 5% at every step. In this sense, sequential testing issues are not a concern.

5.2. *A uniform test.* Let $\hat{h}_n = (\hat{h}_n^\mu, \hat{h}_n^\sigma)$ be the uniform bandwidth vector previously chosen (c.f., Theorem 8). In this case, we need to guarantee that the rate condition for a zero asymptotic bias is satisfied for all $x \in \mathcal{A} \subseteq \mathcal{D}_x$. We formalize the hypotheses as follows:

$$H_0^\mu : (\hat{h}_n^\mu)^{5-\varepsilon} \int_{\mathcal{A}} \sum_{j=1}^{n} K_{\hat{h}_n^\mu}(X_j - x)\, dx \overset{a.s.}{\to} \infty$$

versus

$$H_A^\mu : \text{negation of } H_0.$$

The test statistic $V_{R,n}$ is defined as in the point-wise case except $v_{j,n}$ now integrates the occupation density $\hat{L}_{n,\hat{h}_n^\mu}(x)$ over evaluation points, i.e.

$$v_{j,n} = \left(\exp\left\{(\hat{h}_n^\mu)^{5-\varepsilon} \int_{\mathcal{A}} \sum_{j=1}^{n} K_{\hat{h}_n^\mu}(X_j - x)\, dx\right\}\right)^{1/2} \eta_j$$

with $\mathcal{A} \subseteq \mathcal{D}_x$. The final bandwidth is, as earlier, the bandwidth selected by the moment-based criterion (if the test rejects), or the largest bandwidth for which the test rejects, respectively.

We emphasize again that, should rate-optimality in an MSE sense be the only criterion of interest, one could safely avoid the bias correction. In the Monte Carlo section in the Supplementary Material (Bandi, Corradi and Wilhelm (2011)), however, we show that, whenever re-centering of the estimators asymptotic distribution is indeed needed or desired, the method in this section performs very satisfactorily.

APPENDIX A: PROOFS

PROOF OF THEOREM 1. (a) Hereafter, for notational simplicity we omit the superscript μ, i.e. we write h_n instead of h_n^μ. We first need to show that $h_n \hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ if, and only if, $h_n a(n) \to \infty$, where $a(n)$ is as defined in (7). We begin with the "if" part. Given Assumption 1(i), following Karlsen and Tjøstheim (2001) (KT01 hereafter), $\hat{L}_{n,h_n}(x)$ in (1) can be re-written as a split chain, i.e.,

$$\hat{L}_{n,h_n}(x) = U_{0,x,h_n} + \sum_{k=1}^{T_n} U_{k,x,h_n} + U_{n,x,h_n},$$

where

$$U_{k,x,h_n} = \begin{cases} \sum_{j=1}^{\tau_0} K_{h_n}(X_{j-1} - x), & k = 0 \\ \sum_{j=\tau_{k-1}+1}^{\tau_k} K_{h_n}(X_{j-1} - x), & 1 \leq k < n \\ \sum_{\tau_{T_n}+1}^{n} K_{h_n}(X_{j-1} - x), & k = n \end{cases}.$$

For any given h_n, the U_{k,x,h_n}'s are identically distributed and independent random variables. The quantity T_n denotes the number of complete regenerations from time 0 to time n, and the τ_k's, with $k = 0, \ldots, n$, are the regeneration time points. Thus, T_n is a random variable playing the same role as the sample size. By the same argument as that in the proof of Theorem 5.1 in KT01, U_{0,x,h_n} and U_{n,x,h_n} are of a smaller almost sure order than $\sum_{k=1}^{T_n} U_{k,x,h_n}$. Thus, it suffices to study the asymptotic behavior of

$$\sum_{k=1}^{T_n} U_{k,x,h_n} = \sum_{k=1}^{T_n} (U_{k,x,h_n} - \mu_{x,h_n}) + \sum_{k=1}^{T_n} \mu_{x,h_n},$$

where $\mu_{x,h_n} = E[U_{k,x,h_n}]$. The difficulty is that T_n is a random variable, possibly dependent on U_{k,x,h_n}. Now, define the number of visits to a compact set C as $T_C(n) = \sum_{t=1}^{n} 1\{X_t \in C\}$. From Lemma 3.5 in KT01, it follows that T_n and $T_C(n)$ are of the same almost-sure order. Furthermore, given Assumption 1(i)-(iii), from Theorem 2 in Chen (1999), it follows that $T_C(n)$ is of almost-sure order $a(n)$. Hence, both T_n and $T_C(n)$ are of almost-sure order $a(n)$. Let, now, $\Omega_n = \{\omega : \underline{\Delta} \leq \lim_{n\to\infty} T_n/a(n) \leq \overline{\Delta}\}$ with $0 < \underline{\Delta} \leq \overline{\Delta} < \infty$, and note that, because of Lemma 3.5 in KT01 and Theorem 2 in Chen (1999), $P(\lim_{n\to\infty} \Omega_n) = 1$. We can then proceed conditionally on $\omega \in \Omega_n$. Assume, without loss of generality, that $a(n)$ is an integer or, equivalently, interpret $a(n)$ as $[a(n)]$. Given the independence of the U_{k,x,h_n}'s:

$$E\left[\left(\frac{1}{a(n)}\sum_{k=1}^{a(n)} (U_{k,x,h_n} - \mu_{x,h_n})\right)^{2m}\right]$$
$$\simeq \frac{1}{a(n)^{2m}} \sum_{k_1=1}^{a(n)} \cdots \sum_{k_m=1}^{a(n)} E\left[U_{k_1,x,h_n}^2\right] \ldots E\left[U_{k_m,x,h_n}^2\right] \simeq \frac{1}{a(n)^m} h_n^{-m}$$

where $\simeq$ means "of the same order as". The second "$\simeq$" in the above equation follows from the fact that, given 1(iii), Lemma 5.2 in KT01 implies

$E[U_{k,x,h_n}^{2m}] \leq ch_n^{-2m+1}$. Thus, by Borel-Cantelli, letting $h_n = a(n)^{-\psi}$,

$$\begin{aligned}
&\limsup_n P\left(\left|\frac{1}{a(n)}\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)\right| > \delta\right) \\
&\leq a(n)P\left(\left|\frac{1}{a(n)}\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)\right| > \delta\right) \\
&\leq \frac{a(n)}{a(n)^{2m}\delta^{2m}}E\left[\left(\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)\right)^{2m}\right] \\
&\leq c_m\delta^{-2m}a(n)^{-m+1}h_n^{-m} \leq c_m\delta^{-2m}a(n)^{-m+1+\psi m}, \qquad (10)
\end{aligned}$$

and $a(n)^{-1}\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right) = o_{a.s.}(1)$, provided $-m+\psi m < -1$, i.e., $\psi < \frac{m-1}{m}$. Given A(iii), m can be set arbitrarily large, and then it just suffices that $h_n^{-1} = o(a(n))$. Thus,

$$\begin{aligned}
\hat{L}_{n,h_n}(x) &= U_{0,x,h_n} + \sum_{k=1}^{T_n} U_{k,x,h_n} + U_{n,x,h_n} \\
&= \sum_{k=1}^{T_n}\left(U_{k,x,h_n}-\mu_{x,h_n}\right) + T_n\mu_{x,h_n} + o_{a.s.}\left(T_n\right) \\
&= \sum_{k=1}^{a_n}\left(U_{k,x,h_n}-\mu_{x,h_n}\right) + a_n\mu_{x,h_n} + o_{a.s.}\left(a_n\right) \\
&= o_{a.s}(a_n) + O_{a.s.}(a_n), \qquad (11)
\end{aligned}$$

where the first term in the last equality in (11) holds when $h_na_n \to \infty$. Thus, as $h_na_n \to \infty$, we obtain $h_n\hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$. This concludes the proof of the "if" part. As for the "only if" part, note that the first three equalities in (11) hold regardless of the speed at which h_n approaches zero, hence

$$\hat{L}_{n,h_n}(x) = \sum_{k=1}^{a_n}\left(U_{k,x,h_n}-\mu_{x,h_n}\right) + a_n\mu_{x,h_n} + o_{a.s.}\left(a_n\right).$$

Now, given the independence of the U_{k,x,h_n}'s, $Var(\sum_{k=1}^{a_n}(U_{k,x,h_n}-\mu_{x,h_n})) = O(a_n/h_n)$, and so $\sum_{k=1}^{a_n}(U_{k,x,h_n}-\mu_{x,h_n}) = O_p(\sqrt{a_n/h_n})$. Thus,

$$h_n\hat{L}_{n,h_n}(x) = O_p\left(\sqrt{a_nh_n}\right) + O\left(a_nh_n\right) + o_{a.s.}\left(a_nh_n\right)$$

and $h_n\hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ only if $h_na_n \to \infty$. In fact, if $h_na_n \to 0$, then $h_n\hat{L}_{n,h_n}(x) \overset{p}{\to} 0$. It remains to show that $h_n^5\hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ if, and only if, $h_n^5 a_n \to 0$. Now,

$$(12)\qquad h_n^5 \hat{L}_{n,h_n}(x) = h_n^5 \sum_{k=1}^{a_n} (U_{k,x,h_n} - \mu_{x,h_n}) + h_n^5 a_n \mu_{x,h_n} + o_{a.s.}\left(h_n^5 a_n\right),$$

and it is immediate to see that $h_n^5 \hat{L}_{n,h_n}(x) \stackrel{a.s.}{\to} 0$ only if $h_n^5 a_n \to 0$. As for the "if" part, whenever $h_n a(n) \to \infty$, by the strong law of large numbers, $\sum_{k=1}^{a_n} (U_{k,x,h_n} - \mu_{x,h_n}) = o_{a.s.}(a(n))$, and thus, from (12), we observe that if $h_n^5 a_n \to 0$, then $h_n^5 \hat{L}_{n,h_n}(x)_n \stackrel{a.s.}{\to} 0$. On the other hand, if $h_n a(n) = O(1)$ or $o(1)$, by the same steps used in (10):

$$\begin{aligned}
&\limsup_n P\left(\left| h_n^5 \sum_{k=1}^{a(n)} (U_{k,x,h_n} - \mu_{x,h_n})\right| > \delta\right) \\
&\quad \le \frac{a(n) h_n^{5\times 2m}}{\delta^{2m}} E\left[\left(\sum_{k=1}^{a(n)} (U_{k,x,h_n} - \mu_{x,h_n})\right)^{2m}\right] \\
&\quad \le c a(n)^{m+1} h_n^{9m} = o_{a.s.}(a(n) h_n^5),
\end{aligned}$$

as $a(n) h_n^5 \to 0$, for $m \ge 2$. Thus, the second term on the right-hand side of (12) is $o_{a.s.}(h_n^5 a(n))$. Hence, if $h_n^5 a_n \to 0$, then $h_n^5 \hat{L}_{n,h_n}(x) \stackrel{a.s.}{\to} 0$. The statement in the theorem now follows from Theorem 5.4 in KT01 by noting that their conditions $h_n^{-1} = o(n^{\beta-\varepsilon})$ and $h_n^{-1} = o(n^{\beta/5+\varepsilon})$ are sufficient but not necessary. In effect, their proof relies on the divergence rate of T_n, which is almost-surely $a(n)$. (b) By the same argument as in (a). Q.E.D.

Proof of Theorem 2. In the light of the stronger assumptions on the residuals needed for autoregressions, the same arguments as those in the proof of Theorem 4 imply the result. Q.E.D.

Proof of Theorem 3. (a) By the same argument as in the proof of Theorem 1, $h_n \hat{L}_{n,h_n}(x) \stackrel{a.s.}{\to} \infty$ and $h_n^5 \hat{L}_{n,h_n}(x) \stackrel{a.s.}{\to} 0$ if, and only if, $h_n a(n) \to \infty$ and $h_n^5 a(n) \to 0$, respectively, since the divergence rate of $\hat{L}_{n,h_n}(x)$ depends only on the behavior of the marginal process X_t. The statement of the theorem then follows from Theorem 4.1 in Karlsen, Myklebust and Tjøstheim (2007). (b) By the same argument as in (a). Q.E.D.

Proof of Theorem 4. (a) In the proof of Theorem 1, we have seen that $h_n \hat{L}_{n,h_n}(x) = h_n a(n)(1 + o_{a.s.}(1))$. Hence, we need to show that

$$\sup_{x\in\mathcal{D}_x} \left|\hat{f}_{n,h_n^\mu}(x) - f(x)\right| = O_p\left(\sqrt{\frac{\log(n)}{h_n^\mu a(n)}}\right) + O\left((h_n^\mu)^2\right).$$

Recalling that $\inf_{x\in\mathcal{D}_x} p_s(x) \geq \delta > 0$, by the same argument used in the proof of Theorem 4.2 in Gao, Li and Tjøstheim (2009), it suffices to focus on the variance term and show that

$$\sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)}\sum_{t=1}^{n}\alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t-x)\right| = O_p\left(\sqrt{\frac{\log(n)}{h_n^\mu a(n)}}\right). \tag{13}$$

Now notice that, by condition (3),

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)}\sum_{t=1}^{n}\alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t-x)\right| \\
&\quad\leq \sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)}\sum_{t=1}^{n}\left(\alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t-x) - E\left[K_{h_n^\mu}(X_t-x)\,\epsilon_t\alpha(X_t)\right]\right)\right| \\
&\qquad+\sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)}\sum_{t=1}^{n}E\left[K_{h_n^\mu}(X_t-x)\,\epsilon_t\alpha(X_t)\right]\right|. \\
&\quad= \sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)}\sum_{t=1}^{n}\left(\alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t-x) - E\left[K_{h_n^\mu}(X_t-x)\,\epsilon_t\alpha(X_t)\right]\right)\right| \\
&\qquad+O_p\left(\sqrt{\frac{\log(n)}{h_n^\mu a(n)}}\right).
\end{aligned}$$

Therefore, we can proceed as if $E[K_{h_n^\mu}(X_t-x)\epsilon_t\alpha(X_t)] = 0$ for all $x\in\mathcal{D}_x$. Without loss of generality, assume that $\mathcal{D}_x$ is an interval of length one. We cover $\mathcal{D}_x$ with $Q_n = n/(a(n)^{1/2}(h_n^\mu)^{3/2})$ balls S_i, centered at s_i, of radius $a(n)^{1/2}(h_n^\mu)^{3/2}/n$, $i=1,\dots,Q_n$. Now,

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)}\sum_{t=1}^{n}\alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t-x)\right| \\
&\quad\leq \max_{j=1,\dots,Q_n}\left|\frac{1}{a(n)}\sum_{t=1}^{n}\alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t-s_j)\right| \\
&\qquad+\max_{j=1,\dots,Q_n}\sup_{x\in S_j}\left|\frac{1}{a(n)}\sum_{t=1}^{n}\alpha(X_t)\epsilon_t\left(K_{h_n^\mu}(X_t-x)-K_{h_n^\mu}(X_t-s_j)\right)\right| \\
&\quad= I_{n,h_n^\mu}+II_{n,h_n^\mu}.
\end{aligned}$$

Given Assumption 2(ii)-(iv) and (v), it is immediate to see that

$$II_{n,h_n^\mu} = O_p\left(\frac{n}{h_n^\mu a(n)}\frac{a(n)^{1/2}(h_n^\mu)^{3/2}}{nh_n^\mu}\right) = O_p\left(\frac{1}{\sqrt{h_n^\mu a(n)}}\right) = o_p\left(\sqrt{\frac{\log(n)}{h_n^\mu a(n)}}\right).$$

As for I_{n,h_n^μ}, given Assumption 2(i)-(iv),

$$I_{n,h_n^\mu} = \max_{j=1,\dots,Q_n} \left| \frac{1}{a(n)} \sum_{k=1}^{T_n} Z_k(s_j) \right| (1 + o_{a.s.}(1)),$$

where $Z_k(s_j) = \sum_{t=\tau_k-1}^{\tau_k} \alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t - s_j)$, τ_k, $k = 1, \dots, T_n$, are the regeneration times, and T_n is the number of complete regenerations. For each j, $Z_k(s_j)$, $k = 1, \dots, T_n$, is a sequence of iid random variables so that $\max_{j=1,\dots,Q_n} E[|Z_k(s_j)|^{2m}) = O((h_n^\mu)^{-2m+1})$ (KT01, Lemma 5.2), with m defined in the statement of the theorem. As shown in the proof of Theorem 1, with probability one, $\underline{\Delta} \le \lim_{n\to\infty} T_n/a(n) \le \overline{\Delta}$, hence we can replace $a(n)$ with T_n. Now, given Assumption 2(v), by the same argument used in Hansen (2008), proof of Theorem 2), it follows that for some constant C,

$$\lim_{n\to\infty} \Pr\left(\max_{j=1,\dots,Q_n} \left| \frac{1}{a(n)} \sum_{k=1}^{a(n)} Z_k(s_j) 1\left\{ |Z_k(s_j)| > \sqrt{\frac{a(n)}{h_n^\mu}} \right\} \right| > \eta \right) = 0.$$

where $\eta = C\sqrt{\log(n)/(h_n^\mu a(n))}$. Further, define $\tilde{Z}_k(s_j) = Z_k(s_j)1\{|Z_k(s_j)| \le \sqrt{a(n)/h_n^\mu}\}$ so that, given Assumption 2(iv), by the Bernstein inequality for zero mean iid sequences (e.g., Theorem 2.18 in Fan and Yao (2005)),

$$\begin{aligned}
&\Pr\left(\max_{j=1,\dots,Q_n} \left| \frac{1}{a(n)} \sum_{k=1}^{a(n)} \tilde{Z}_k(s_j) \right| > \eta \right) \\
&\le Q_n \exp\left\{ -\frac{\eta^2 a(n)}{Var\left(\tilde{Z}_k(s_j)\right) + \eta \sup_k \left|\tilde{Z}_k(s_j)\right|} \right\} \\
&\le Q_n \exp\left\{ -\frac{C^2 \frac{\log(n)}{h_n^\mu a(n)} a(n)}{c\frac{1}{h_n^\mu} + C\sqrt{\frac{\log(n)}{h_n^\mu a(n)}}\sqrt{\frac{a(n)}{h_n^\mu}}} \right\} \\
&= \frac{n}{a(n)^{1/2}(h_n^\mu)^{3/2}} n^{-f_C} \to 0,
\end{aligned}$$

with f_C, an increasing function of C, and C finite but sufficiently large. (b) By the same argument used to show (a). Q.E.D.

PROOF OF THEOREM 5. (a) As in the case of previous theorems, we only prove Part (a). We need to show that $h_n \hat{L}_{n,h_n}(x) \stackrel{a.s.}{\to} \infty$ only if $h_n\sqrt{n} \to \infty$ and, analogously, $h_n^5 \hat{L}_{n,h_n}(x) \stackrel{a.s.}{\to} 0$ only if $h_n^5\sqrt{n} \to 0$. Given Assumption 3, the statement then follows from Theorem 3.1 in Wang and Phillips (2009b). Write

$$\frac{1}{\sqrt{n}}\hat{L}_{n,h_n}(x) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} K_{h_n}\left(\frac{\frac{\sum_{i=1}^{j}\xi_i}{\sqrt{n}} - \frac{x}{\sqrt{n}}}{n^{-1/2}}\right) = \frac{c_n}{n}\sum_{j=1}^{n} g\left(c_n x_{j,n}\right),$$

where $g(s) = K(s)$ and $c_n = \sqrt{n}/h_n$. Hereafter, let $\phi_\epsilon(x) = \frac{1}{\epsilon\sqrt{2\pi}}e^{-x^2/(2\epsilon^2)}$ for $\epsilon > 0$. Along the lines of the proof of Theorem 2.1 in Wang and Phillips (2009b):

$$\frac{1}{\sqrt{n}}\hat{L}_{n,h_n}(x) = \left(\frac{c_n}{n}\sum_{j=1}^{n} g\left(c_n x_{j,n}\right) - \frac{c_n}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} g\left(c_n(x_{j,n}+z\epsilon)\right)\phi(z)dz\right)$$
$$+\left(\frac{c_n}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} g\left(c_n(x_{j,n}+z\epsilon)\right)\phi(z)dz - \frac{1}{n}\sum_{j=1}^{n}\phi_\epsilon(x_{j,n})\right) + \frac{1}{n}\sum_{j=1}^{n}\phi_\epsilon(x_{j,n}).$$

Let $G(s) = \omega_0 W_s$, where W_s is a standard Brownian motion, and notice that

$$\begin{aligned}
&\sup_{0\le r\le 1}\left|\frac{1}{n}\sum_{j=1}^{[nr]}\phi_\epsilon(x_{j,n}) - \int_0^r \phi_\epsilon(G(t))dt\right| \\
&\quad\le \int_0^1 \left|\phi_\epsilon(x_{nt,n}) - \int_0^r \phi_\epsilon(G(t))\right| dt + \frac{2}{n} \\
&\quad\le A_\epsilon \sup_{0\le t\le 1}\left|x_{[nt],n} - G(t)\right| + 2/n = o_{a.s.}\left(\sqrt{2\log\log n}\right),
\end{aligned} \tag{14}$$

where A_ϵ is a term depending on ϵ, and the last equality on the right-hand side of (14) follows from the fact that, given Assumption 3(i), by Lemma 2.(i) in Corradi (1999), uniformly in $t \in [0,1]$, $|x_{[nt],n} - G(t)| = o_{a.s.}(\sqrt{\log\log n})$. Now, as $\epsilon \to 0$,

$$\int_0^r \phi_\epsilon(G(t))dt = \int_{-\infty}^{\infty}\phi_\epsilon(x)L(r,\epsilon x)dx = L(0,r) + o_{a.s.}(1),$$

where $L(0,r)$ is the local time of $G(t)$ at spatial point 0 between time 0 and time r. By Lemma 7 in Jeganathan (2004), for any $\epsilon > 0$, and recalling that $\int K(u)du = 1$,

$$\frac{c_n}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} g\left(c_n(x_{j,n}+z\epsilon)\right)\phi(z)dz - \frac{1}{n}\sum_{j=1}^{n}\phi_\epsilon(x_{j,n}) = o_{a.s.}(1).$$

Finally, from the proof of Theorem 2.1 in (Wang and Phillips, 2009a, pp.726-728),

$$\frac{c_n}{n}\sum_{j=1}^{n} g\left(c_n x_{j,n}\right) - \frac{c_n}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} g\left(c_n(x_{j,n}+z\epsilon)\right)\phi(z)dz = o_p(1).$$

Thus,

$$\frac{1}{\sqrt{n}}\hat{L}_{n,h_n}(x) = L(0,1) + o_{a.s.}\left(\sqrt{\log\log n}\right) + o_p(1),$$

that is

$$h_n\hat{L}_{n,h_n}(x) = \sqrt{n}h_n L(0,1) + o_{a.s.}\left(\sqrt{n}h_n\sqrt{\log\log n}\right) + o_p(\sqrt{n}h_n). \tag{15}$$

Because $L(0,1) > 0$ a.s., it is immediate to see that, whenever $\sqrt{n}h_n \to \infty$, then $h_n\hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$. Similarly, if $h_n\hat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$, then $\sqrt{n}h_n \to \infty$. Also,

$$h_n^5\hat{L}_{n,h_n}(x) = \sqrt{n}h_n^5 L(0,1) + o_{a.s.}\left(\sqrt{n}h_n^5\sqrt{\log\log n}\right) + o_p(\sqrt{n}h_n^5).$$

Thus, if $h_n^5\hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$, then $h_n^5\sqrt{n} \to 0$. On the other hand, $h_n^5\sqrt{n} \to 0$ implies $h_n^5\hat{L}_{n,h_n}(x) \overset{p}{\to} 0$, though it does not necessarily imply that $h_n^5\hat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$. Q.E.D.

Proof of Theorem 6. (a) As shown in Theorem 5, we have $h_n\hat{L}_{n,h_n}(x) = h_n n^{1/2}(1 + o_{a.s.}(1))$. Hence, it suffices to show that

$$\sup_{x\in\mathcal{D}_x}\left|\hat{f}_{n,h_n^\mu}(x) - f(x)\right| = O_p\left(\sqrt{\frac{\log(n)}{h_n^\mu n^{1/2}}}\right) + O\left((h_n^\mu)^2\right).$$

Given the condition (3) with $a(n) = n^{1/2}$, we can proceed as if $K_{h_n^\mu}(X_t - x)\epsilon_t\alpha(X_t)$ were a martingale difference sequence. By the same argument used in the proof of Theorem 4, we simply need to show that

$$\max_{j=1,\dots,\tilde{Q}_n}\left|\frac{1}{n^{1/2}}\sum_{t=1}^{n}\alpha(X_t)\epsilon_t K_{h_n^\mu}\left(X_t - s_j\right)\right| = O_p\left(\sqrt{\frac{\log(n)}{h_n^\mu n^{1/2}}}\right),$$

where $\tilde{Q}_n = n^{3/4}(h_n^\mu)^{-3/2}$. Given $h_n^\mu E[\exp\{K_{h_n^\mu}(X_t - x)\alpha(X_t)\epsilon_t\}] \leq \Delta < \infty$, by Theorem 3.2 in Lesigne and Volný (2001), letting $\eta = C\sqrt{\log(n)/(h_n^\mu n^{1/2})}$,

$$\begin{aligned}
&\Pr\left(\max_{j=1,\dots,Q_n}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\alpha(X_t)\epsilon_t K_{h_n^\mu}(X_t - s_j)\right| > \eta\right)\\
&\quad\le \tilde{Q}_n \Pr\left(\left|\sum_{t=1}^{n}\alpha(X_t)\epsilon_t K\left(\frac{X_t - s_j}{h_n^\mu}\right)\right| > \eta h_n^\mu \sqrt{n}\right)\\
&\quad\le \tilde{Q}_n \exp\left(-\frac{1}{2}C_\Delta \eta^{2/3}\left(h_n^\mu\sqrt{n}\right)^{1/3}\right) = \frac{n^{3/4}}{(h_n^\mu)^{3/2}} n^{-C_\Delta f_C} \to 0,
\end{aligned}$$

where $0 < C_\Delta < 1$, with C_Δ a decreasing function of Δ and f_C an increasing function of C. The statement then follows for C large enough. Q.E.D.

PROOF OF THEOREM 7. Below, we prove Part (b). Part (a) follows by the same argument, given Theorem 2. We begin with the first moment condition. For any sequence $h_n(x) \in \mathcal{H}_n(x, \varsigma)$,

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\epsilon}_{i,h_n} w_{i,h_n^\mu}(x) &= \sum_{i=1}^{n}\epsilon_i w_{i,h_n^\mu}(x) - \sum_{i=1}^{n}\frac{\hat{f}_{n,h_n^\mu}(X_i) - f(X_i)}{\alpha(X_i)} w_{i,h_n^\mu}(x)\\
&\quad + \sum_{i=1}^{n}\frac{\alpha(X_i) - \hat{\alpha}_{n,h_n}(X_i)}{\hat{\alpha}_{n,h_n}(X_i)}\epsilon_i w_{i,h_n^\mu}(x)\\
&\quad - \sum_{i=1}^{n}\frac{\alpha(X_i) - \hat{\alpha}_{n,h_n}(X_i)}{\alpha(X_i)\hat{\alpha}_{n,h_n}(X_i)}\left(\hat{f}_{n,h_n^\mu}(X_i) - f(X_i)\right) w_{i,h_n^\mu}(x).
\end{aligned} \tag{16}$$

It is immediate to see that the third term cannot be of larger order than the first term and the fourth term cannot be of larger order than the second term for all $h_n(x) \in \mathcal{H}_n(x, \varsigma)$. For bandwidth sequences in the set $\mathcal{H}_n(x, \varsigma)$, in fact, $\sup_{x\in\mathcal{D}_x}\{\hat{f}_{n,h_n^\mu}(x) - f(x)\}$ and $\sup_{x\in\mathcal{D}_x}\{\hat{\alpha}_{n,h_n}(X_i) - \alpha(X_i)\}$ are bounded in probability given Theorem 4 or 6. Now, write

$$\sum_{i=1}^{n}\epsilon_i w_{i,h_n^\mu}(x) = \frac{\frac{1}{\gamma(n)h_n^\mu}\sum_{i=1}^{n}\epsilon_i 1\{|X_i - x| < h_n^\mu\}}{\frac{1}{\gamma(n)h_n^\mu}\sum_{i=1}^{n} 1\{|X_i - x| < h_n^\mu\}}, \tag{17}$$

where the denominator in (17) is $O_{a.s.}(1)$, and bounded away from zero, for $\gamma(n) = a(n)$, under Assumption 2, by Theorem 3, for $\gamma(n) = n^{1/2}$, under Assumption 3, by Theorem 5, and for $\gamma(n) = n$ in the stationary case, by the strong law of large numbers. As for the numerator in (17), recalling that $E[\epsilon_i 1\{|X_i - x| < h_n^\mu\}] = o(1)$, the contribution of the bias term is negligible, and thus it is $O_p((\gamma(n)h_n^\mu)^{-1/2})$. As for the second term in (16), by either

Theorem 4 or Theorem 6:

$$\begin{aligned}
&\left|\sum_{i=1}^{n} \frac{\hat{f}_{n,h_n^\mu}(X_i) - f(X_i)}{a(X_i)} w_{i,h_n^\mu}(x)\right| \\
&\quad \le \sup_{z:|x-z|\le h_n^\mu, x\in\mathcal{D}_x} \left|\hat{f}_{n,h_n^\mu}(z) - f(z)\right| \sum_{i=1}^{n} \frac{w_{i,h_n^\mu}(x)}{\alpha(X_i)} \\
&\quad = \left(O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\mu}(x)h_n^\mu}}\right) + O\left((h_n^\mu)^2\right)\right) O_p(1).
\end{aligned}$$

As shown in the proof of Theorem 1 and 5 respectively, for all $x \in \mathcal{D}_x$, $\hat{L}_{n,h_n^\mu}(x)h_n^\mu$ is $O_{a.s.}(a(n)h_n^\mu)$, in the β-recurrent case, $O_{a.s.}(nh_n^\mu)$ when $\beta = 1$, and $O_{a.s.}(\sqrt{n}h_n^\mu)$ in the integrated case. Hence, $\sum_{i=1}^n \hat{\epsilon}_{i,h_n} w_{i,h_n^\mu}(x)$ is at least of probability order $(h_n^\mu)^2 + 1/\sqrt{\gamma(n)h_n^\mu}$ and at most of probability order $(h_n^\mu)^2 + \sqrt{\log n/(\gamma(n)h_n^\mu)}$. We now turn to the second moment condition.

$$\begin{aligned}
\sum_{i=1}^{n}\left(\hat{\epsilon}_{i,h_n}^2 w_{i,h_n^\mu}(x) - 1\right) &= \sum_{i=1}^{n}\left(\epsilon_{i,h_n}^2 w_{i,h_n^\mu}(x) - 1\right) \\
&\quad - \sum_{i=1}^{n} \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\hat{\alpha}_{n,h_n}^2(X_i)} \left(\hat{f}_{n,h_n^\sigma}^{(2)}(X_i) - f^{(2)}(X_i)\right) \\
&\quad + \sum_{i=1}^{n} \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\hat{\alpha}_{n,h_n}^2(X_i)} \left(\hat{f}_{n,h_n^\mu}(X_i)^2 - f(X_i)^2\right) \\
&\quad + \sum_{i=1}^{n} \frac{\left(\hat{f}_{n,h_n^\mu}(X_i) - f(X_i)\right)^2}{\hat{\alpha}_{n,h_n}^2(X_i)} w_{i,h_n^\mu}(x) \\
&\quad + 2\sum_{i=1}^{n} \frac{\left(\hat{f}_{n,h_n^\mu}(X_i) - f(X_i)\right)(Y_i - f(X_i))}{\hat{\alpha}_{n,h_n}^2(X_i)} w_{i,h_n^\mu}(x).
\end{aligned} \tag{18}$$

It is immediate to see that, for all $h_n(x) \in \mathcal{H}_n(x,\varsigma)$, the fourth term in (18) cannot be of larger probability order than the third term, while the fifth term cannot be of larger probability order than the first and third terms. Hence, the last two terms in (18) can be neglected. Since $E[\epsilon_{i,h_n}^2|X_i = x] - \sigma^2(x) = o(1)$, the bias component is asymptotically zero and $\sum_{i=1}^n(\epsilon_{i,h_n}^2 w_{i,h_n^\mu}(x) - 1) = O_p((\gamma(n)h_n^\mu)^{-1/2})$, where, again, $\gamma(n)$ differs across the various cases. As for the second term on the right-hand side of (18), because of either

Theorem 4 or Theorem 6,

$$\begin{aligned}
&\left|\sum_{i=1}^{n} \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\hat{\alpha}_{n,h_n}^2(X_i)} \left(\hat{f}_{n,h_n^\sigma}^{(2)}(X_i) - f^{(2)}(X_i)\right)\right| \\
&\quad \leq \sup_{z:|x-z|\leq h_n^\mu, x\in\mathcal{D}_x} \left|\hat{f}_{n,h_n^\sigma}^{(2)}(z) - f^{(2)}(z)\right| \sum_{i=1}^{n} \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\hat{\alpha}_{n,h_n}^2(X_i)} \\
&\quad = \left(O_p\left(\sqrt{\frac{\log(n)}{\hat{L}_{n,h_n^\sigma}(x)h_n^\sigma}}\right) + O\left((h_n^\sigma)^2\right)\right) O_p(1) \\
&\quad = \left(O_p\left(\sqrt{\frac{\log(n)}{\gamma(n)h_n^\sigma}}\right) + O\left((h_n^\sigma)^2\right)\right) O_p(1).
\end{aligned}$$

By the same argument,

$$\left|\sum_{i=1}^{n} \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\hat{\alpha}_{n,h_n}^2(X_i)} \left(\hat{f}_{n,h_n^\mu}(X_i)^2 - f(X_i)^2\right)\right|$$

is majorized by a $O_p(\sqrt{\log(n)/(\gamma(n)h_n^\mu)}) + O_p((h_n^\mu)^2)$ term. Thus, the term $\sum_{i=1}^n \hat{\epsilon}_{i,h_n} w_{i,h_n^\mu}(x)$ is at least of probability order $(h_n^\mu)^2 + 1/\sqrt{\gamma(n)h_n^\mu} + h_n^{\sigma 2} + 1/\sqrt{\gamma(n)h_n^\sigma}$ and at most of probability order $(h_n^\mu)^2 + \sqrt{\log n/(\gamma(n)h_n^\mu)} + (h_n^\sigma)^2 + \sqrt{\log n/(\gamma(n)h_n^\sigma)}$. The statement then follows directly from the definition of $\tilde{h}_n(x)$. Q.E.D.

PROOF OF THEOREM 8. We prove Part (b) as earlier. Part (a) follows from the same argument, given Theorem 2. As for the first moment condition, by the triangle inequality, we note that, for all $h_n \in \mathcal{H}_n(\varsigma)$,

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n} \hat{\epsilon}_{i,h_n} w_{i,h}(x)\right| \\
&\quad \leq \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n} \epsilon_{i,h_n} w_{i,h_n^\mu}(x)\right| + \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n} \frac{\hat{f}_{n,h_n^\mu}(X_i) - f(X_i)}{\alpha(X_i)} w_{i,h_n^\mu}(x)\right| \\
&\qquad + \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n} \frac{\alpha(X_i) - \hat{\alpha}_{n,h_n}(X_i)}{\hat{\alpha}_{n,h_n}(X_i)\alpha(X_i)} \epsilon_i w_{i,h_n^\mu}(x)\right| \\
&\qquad + \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n} \frac{\alpha(X_i) - \hat{\alpha}_{n,h_n}(X_i)}{\alpha(X_i)\hat{\alpha}_{n,h_n}(X_i)} \left(\hat{f}_{n,h_n^\mu}(X_i) - f(X_i)\right) w_{h_n^\mu}(x)\right|. \qquad (19)
\end{aligned}$$

By the same argument used in either Theorem 4 or Theorem 6, we have $\sup_{x\in\mathcal{D}_x} |\sum_{i=1}^n \epsilon_{i,h_n} w_{i,h_n^\mu}(x)| = O_p(\sqrt{\log n/(\gamma(n)h_n^\mu)})$. Because the last two terms on the right-hand side of (19) cannot be of larger probability order than the first two terms on $\mathcal{H}_n(\varsigma)$, and because

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\frac{\hat{f}_{n,h_n^\mu}(X_i)-f(X_i)}{\alpha(X_i)}w_{i,h_n^\mu}(x)\right| \\
&\quad\le \sup_{z:|x-z|\le h_n^\mu, x\in\mathcal{D}_x}\left|\hat{f}_{n,h_n^\mu}(z)-f(z)\right|\sup_{x\in\mathcal{D}_x}\sum_{i=1}^{n}\frac{w_{i,h_n^\mu}(x)}{\alpha(X_i)} \\
&\quad=\left(O_p\left(\sqrt{\frac{\log n}{\gamma(n)h_n^\mu}}\right)+O\left((h_n^\mu)^2\right)\right)O_p(1)
\end{aligned}$$

it follows that the left-hand side sup is at most $O_p(\sqrt{\log n/(\gamma(n)h_n^\mu)}) + O_p((h_n^\mu)^2)$. Also,

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\hat{\epsilon}_{i,h_n}w_{i,h}(x)\right| \\
&\quad\ge \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\epsilon_{i,h_n}w_{i,h_n^\mu}(x)-\sum_{i=1}^{n}\frac{\hat{f}_{n,h_n^\mu}(X_i)-f(X_i)}{\alpha(X_i)}w_{i,h_n^\mu}(x)\right| \\
&\quad-\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\frac{\alpha(X_i)-\hat{\alpha}_{n,h_n}(X_i)}{\hat{a}_{n,h_n}(X_i)a(X_i)}\epsilon_i w_{i,h_n^\mu}(x)\right| \\
&\quad-\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\frac{\alpha(X_i)-\hat{\alpha}_{n,h_n}(X_i)}{\alpha(X_i)\hat{\alpha}_{n,h_n}(X_i)}\left(\hat{f}_{n,h_n^\mu}(X_i)-f(X_i)\right)w_{h_n^\mu}(x)\right|
\end{aligned}$$

and, thus, $\sup_{x\in\mathcal{D}_x}|\sum_{i=1}^{n}\hat{\epsilon}_{i,h_n}w_{i,h}(x)|$ is at least $O_p(\sqrt{\log n/(\gamma(n)h_n^\mu)}) + O_p((h_n^\mu)^2)$. By the same argument used in the proof of Theorem 7, it is immediate to see that $\sup_{x\in\mathcal{D}_x}|\sum_{i=1}^{n}(\hat{\epsilon}_{i,h_n}^2 w_{i,h_n^\mu}(x)-1)|$ is (at most and at least) $O_p((h_n^\mu)^2+\sqrt{\log n/(\gamma(n)h_n^\mu)}+(h_n^\sigma)^2+\sqrt{\log n/(\gamma(n)h_n^\sigma)})$. The statement then follows from the definition of $\tilde{h}_n$ in (9). Q.E.D.

PROOF OF THEOREM 9. The methods used to prove Theorem 3 in Bandi, Corradi and Moloche (2009) yield the result. Q.E.D.

ACKNOWLEDGEMENTS

We thank Jiti Gao, Emmanuel Guerre, Melanie Schienle, Mervyn Silvapulle, Vladimir Spokoiny, as well as seminar participants at Humboldt University, the LSE conference on Nonparametric and Semiparametric Methods, the 2010 Vilnius Conference in Probability Theory, the Monash University

Time Series Conference at Caulfield campus, the University of Manchester, the University of Maryland, the University of Montreal, the University of Western Ontario, Rice University, Rutgers University, and Texas A&M University for many helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement: Simulation Experiment
(). This supplementary document provides the details and results of a simulation experiment illustrating the performance of the bandwidth selection procedure proposed in the main text.

REFERENCES

Bandi, F. (2004). On Persistence and Nonparametric Estimation (With an Application to Stock Return Predictability) Technical Report.

Bandi, F. M., Corradi, V. and Moloche, G. (2009). Bandwidth Selection for Continuous Time Markov Processes Technical Report.

Bandi, F. M., Corradi, V. and Wilhelm, D. (2011). Supplement to "Data-driven Bandwidth Selection for Nonparametric Nonstationary Regressions" Technical Report.

Bandi, F. M. and Moloche, G. (2008). On the Functional Estimation of Multivariate Diffusion Processes Technical Report.

Bandi, F. M. and Phillips, P. C. B. (2003). Fully Nonparametric Estimation of Scalar Diffusion Models. *Econometrica* **71** 241-283.

Bandi, F. M. and Phillips, P. C. B. (2007). A Simple Approach to the Parametric Estimation of Potentially Nonstationary Diffusions. *Journal of Econometrics* **137** 354-395.

Chen, X. (1999). How Often Does a Harris Recurrent Markov Chain Recur? *The Annals of Probability* **27** 1324-1346.

Corradi, V. (1999). Deciding Between I(0) and I(1) via FLIL-Based Bounds. *Econometric Theory* **15** 643-663.

Fan, J. and Yao, Q. (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, Berlin.

Gao, J., Li, D. and Tjøstheim, D. (2009). Uniform Consistency for Nonparametric Estimators in Null Recurrent Time Series Technical Report No. 0085, University of Adelaide.

Gao, J., King, M., Lu, Z. and Tjøstheim, D. (2009). Nonparametric Specification Testing for Nonlinear Time Series with Nonstationarity. *Econometric Theory* **25** 1869-1892.

Guerre, E. (2004). Design-Adaptive Pointwise Nonparametric Regression Estimation for Recurrent Markov Time Series Technical Report.

Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory* **24** 726-748.

Jeganathan, P. (2004). Convergence of Functionals of Sums of r.v.s to Local Times of Fractional Stable Motions. *Annals of Probability* **32** 1717-1795.

Karlsen, H. A., Myklebust, T. and Tjøstheim, D. (2007). Nonparametric Estimation in a Nonlinear Cointegration Type Model. *The Annals of Statistics* **35** 252-299.

Karlsen, H. A. and Tjøstheim, D. (2001). Nonparametric Estimation in Null Recurrent Time Series. *The Annals of Statistics* **29** 372-416.

Lepski, O. V., Mammen, E. and Spokoiny, V. G. (1997). Optimal Spatial Adaptation to Inhomogeneous Smoothness: An Approach Based on Kernel Estimates with Variable Bandwidth Selectors. *The Annals of Statistics* **25** 929-947.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal Pointwise Adaptive Methods in Nonparametric Estimation. *The Annals of Statistics* **25** 2512-2546.

Lepskii, O. V. (1991). Asymptotically Minimax Adaptive Estimation. I: Upper Bounds. Optimally Adaptive Estimates. *Theory of Probability and its Applications* **36** 682-697.

Lesigne, E. and Volný, D. (2001). Large Deviations for Martingales. *Stochastic Processes and their Applications* **96** 143-159.

Moloche, G. (2001). Kernel Regression for Nonstationary Harris-Recurrent Processes Technical Report.

Robinson, P. M. (1983). Nonparametric Estimators of Times Series. *Journal of Time Series Analysis* **4** 185-207.

Schienle, M. (2010). Nonparametric Nonstationary Regression with Many Covariates Technical Report.

Wang, Q. and Phillips, P. C. B. (2009a). Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegrating Regression. *Econometric Theory* **25** 710-738.

Wang, Q. and Phillips, P. C. B. (2009b). Structural Nonparametric Cointegrating Regression. *Econometrica* **77** 1901-1948.

Carey Business School
Johns Hopkins University
E-mail: fbandi1@jhu.edu

Department of Economics
University of Warwick
E-mail: v.corradi@warwick.ac.uk

Booth School of Business
University of Chicago
E-mail: daniel.wilhelm@uchicago.edu