# Nonparametric Nonstationary Autoregression and Nonparametric Cointegrating Regression: Automated Bandwidth Selection[*]

Federico M. Bandi,[*] Valentina Corradi,[**] Daniel Wilhelm[***]

[*]Johns Hopkins University and Edhec-Risk,[**]University of Warwick, and [***]University of Chicago

January 2011

### Abstract

We propose an automated bandwidth selection procedure for the nonparametric estimation of conditional moments, focusing on nonparametric nonstationary autoregressions and nonparametric cointegrating regressions. The methods apply to both $\beta$-recurrent Markov chains and nonlinear functions of integrated processes, the stationary short-memory case being a sub-case of the former. The procedure consists in choosing the relevant bandwidth(s) by virtue of the minimization of a set of moment conditions constructed using nonparametric residuals. 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. We provide an initial solution to a largely open problem, that of bandwidth selection in nonstationary models, rather than an alternative solution to cross-validation, which is solely justified in stationary environments. However, in light of the widespread use of cross-validation in empirical work, the finite sample behavior of our proposed bandwidth selection method, and that of its subsequent bias correction, are analyzed in a Monte Carlo exercise and compared to cross-validation. We find that our combined procedure fares favorably with respect to it and delivers conditional moment estimates conforming accurately with their limiting normal laws.

*Keywords:* Automated Bandwidth Selection, Nonstationary Autoregression, Nonstationary Cointegration, Recurrence.

[*] 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 University of Manchester, the University of Maryland, the 2010 Vilnius Conference in Probability Theory, and the Monash University Time Series Conference at Caulfield campus, for many helpful comments and suggestions.

# 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 $\beta$-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 Tjostheim, 2001, Moloche, 2001, Gao, Li, and Tjostheim, 2009). 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 emerged. The first is based on a multidimensional extension of $\beta$-recurrent Markov chains and, again, heavily employs the notion of regeneration time (e.g., Karlsen, Myklebust and Tjostheim, 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, 2009b).[1] 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, 2004). 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, King, Lu, and Tjostheim (2009) suggest a bootstrap procedure to select the bandwidth parameter which maximizes the local

[1] If $X_t$ is a (near) integrated process, then the dependence of $X_t$ on $X_{t-1}$ is (nearly) linear. For this reason, we are considering this second approach only in the nonparametric cointegrating regression case.

power function, while controlling for size.[2] Their approach, however, may not be employed to find optimal bandwidths for conditional moment kernel estimators.

This paper aims at filling an important gap in the existing literature by suggesting a procedure for automated bandwidth selection in the context of nonparametric autoregressions and nonparametric cointegrating regressions. The proposed method applies to both $\beta$-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 automated bandwidth choice.

We offer three contributions. The rate conditions on the bandwidth sequence for asymptotic mixed normality depend on $\beta$, the generally unknown regularity of the chain. Although $\beta$ can be estimated, its estimator converges only at a logarithmic rate (see, e.g., Karlsen and Tjostheim, 2010). 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 methods 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., Lepski, 1990). 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 stop-

[2] In the stationary case, the same bootstrap approach to bandwidth selection has been suggested by Gao and Gijbels (2008).

ping rule ensuring that the selected bandwidths are the largest ones 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 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 automated 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). In Section 6 we report the findings of a Monte Carlo study. Section 7 concludes. All proofs are collected in the Appendix.

## 2 Nonparametric Nonstationary Autoregression

Intuitively, one can estimate conditional moments, evaluated at a given point, only if 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 Tjostheim (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 $\mu(X_{t-1}) = \mathrm{E}\left(X_t|X_{t-1}\right)$ and $\sigma^2(X_{t-1}) = \mathrm{var}\left(X_t|X_{t-1}\right) = \mathrm{E}\left(\epsilon_t^2|X_{t-1}\right)$ so that $X_t$ can be written as

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

where $u_t$ is a martingale difference sequence with respect to the filtration generated by $X_{t-1}$ and $\mathrm{E}\left(u_t^2|X_{t-1}\right) = 1$. Now, define

$$\widehat{\mu}_{n,h_n^\mu}(x) = \frac{\sum_{j=1}^n X_j K\left(\frac{X_{j-1}-x}{h_n^\mu}\right)}{\sum_{j=1}^n K\left(\frac{X_{j-1}-x}{h_n^\mu}\right)} \tag{1}$$

$$\widehat{\mu}^{(2)}_{n,h_n^\sigma}(x) = \frac{\sum_{j=1}^n X_j^2 K\left(\frac{X_{j-1}-x}{h_n^\sigma}\right)}{\sum_{j=1}^n K\left(\frac{X_{j-1}-x}{h_n^\sigma}\right)}, \tag{2}$$

and $\widehat{\sigma}^2_{h_n}(x) = \widehat{\mu}^{(2)}_{n,h_n^\sigma}(x) - \left(\widehat{\mu}_{n,h_n^\mu}(x)\right)^2$. We rely on the following Assumption which largely corresponds to Assumption B$_0$-B$_4$ in Karlsen and Tjostheim (2001).

**Assumption A.**

$(i)$ Let $\{X_t,\ t \geq 0\}$ be a $\beta-$recurrent, $\phi-$irreducible Markov chain on a general state space $(\mathbf{E}, \mathcal{E})$ with transition probability $P$. Let $\beta \in (0, 1]$.[3]

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

[3] As said, the case $\beta = 1$, with the addition of some innocuous regularity conditions, corresponds to the case of positive recurrent chains.

$(iii)$ The kernel function $K$ is a bounded density with compact support satisfying $\int uK(u)\mathrm{d}u = 0$. Write $K_h(y-x) = \frac{1}{h}K\left(\frac{y-x}{h}\right)$. The set $\mathcal{N}_x = \{y : K_{h=1}(y-x) \neq 0\}$ is small for all $x \in \mathcal{D}_x$, where $\mathcal{D}_x$ is a compact set in $\mathcal{R}$ so that $\mathcal{D}_x = \{x : p_s(x) > \delta\}$ with $\delta$ arbitrarily small and independent of $x$.[4] In what follows, $K_2 = \int K^2(u)\mathrm{d}u < \infty$.
$(iv)$ For all sets $A_h \in \mathcal{E}$ so that $A_h \downarrow \varnothing$ when $h \downarrow 0$, we have $\lim_{h\downarrow 0} \overline{\lim}_{y\to x} P(y, A_h) = 0$.[5]
$(v)$ The functions $\mu(x)$ and $\sigma^2(x)$ are locally twice continuously differentiable for all $x \in \mathcal{D}_x$.

Define

$$\widehat{L}_{n,h_n^i}(x) = \frac{1}{h_n^i}\sum_{j=1}^{n} K\left(\frac{X_{j-1}-x}{h_n^i}\right) \text{ with } i = \mu, \sigma. \tag{3}$$

In the positive recurrent case ($\beta = 1$), as $n \to \infty$ and $h_n \to 0$ with $nh_n \to \infty$, $\widehat{L}_{n,h_n}(x)/n \overset{a.s.}{\to} \varphi(x)$, where $\varphi(x)$ is the density associated with the time-invariant probability measure.[6] Whenever $0 < \beta < 1$, under Assumption A$(i)$-$(iii)$ and provided $n \to \infty$ and $h_n \to 0$ with $h_n n^\beta u(n) \to \infty$, $\widehat{L}_{n,h_n}(x)/\left(n^\beta u(n)\right) \overset{d}{\to} c_X \mathcal{M}_\beta$, where $c_X$ is a process-specific constant, $\mathcal{M}_\beta$ is the Mittag-Leffler density with parameter $\beta$, and the positive function $u(.)$ 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 $\widehat{L}_{n,h_n}(x)$, namely $n^\beta u(n)$, and the features of the asymptotic distribution, $\mathcal{M}_\beta$, depend on the degree of recurrence $\beta$. Similarly, $T_n/n^\beta u(n) \overset{d}{\to} \mathcal{M}_\beta$, where $T_n$ is, as earlier, the number of complete regenerations. [7]

**Proposition 1.** Let Assumption A hold and let $\left(\mathrm{E}\left(X_t|X_{t-1}\right)\right)^{2m} < \infty$ and $\left(\mathrm{E}\left(X_t^2|X_{t-1}\right)\right)^{2m} < \infty$, for $X_{t-1}$ in a neighborhood of $x$ and for $m \geq 2$.
(a) If $(i)$ $h_n^\mu \widehat{L}_{n,h_n^\mu}(x) \overset{a.s.}{\to} \infty$ and $(ii)$ $h_n^{\mu 5}\widehat{L}_{n,h_n^\mu}(x) \overset{a.s.}{\to} 0$, then[8]

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

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

$$\sqrt{h_n^\sigma \widehat{L}_{n,h_n^\sigma}(x)}\left(\widehat{\mu}^{(2)}_{n,h_n^\sigma}(x) - \mu^{(2)}(x)\right) \overset{d}{\to} N\left(0, \left(\mu^{(4)}(x) - \left(\mu^{(2)}(x)\right)^2\right)K_2\right). \tag{5}$$

[4] For a definition of "small" set, we refer the reader to Karlsen and Tjostheim (2001).

[5] This is a continuity assumption of the process' transition density.

[6] We have suppressed the superscripts $\mu$ or $\sigma$ since the same result applies to any sequence $h_n$ with similar vanishing properties.

[7] Write $\widehat{L}_{n,h_n}(x) = \frac{1}{h_n}\sum_{k=1}^{T_n}\sum_{t=\tau_{k-1}+1}^{\tau_k} K\left(\frac{X_t - x}{h_n}\right)$. If the random sums $\sum_{t=\tau_{k-1}+1}^{\tau_k} K\left(\frac{X_t-x}{h_n}\right)$ with $k = 1, \ldots$ are independent random variables, then $T_n$ is the number of complete regenerations and the $\tau_k$'s are the regeneration time points.

[8] Here, and in similar results below, the condition $h_n^5 \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} C$, 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 Proposition above is similar to that in Theorem 5.4 in Karlsen and Tjostheim (2001). However, Karlsen and Tjostheim (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, $\beta$ 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 Tjostheim, 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 Proposition 1. The key argument used in the proof of Proposition 1 is that $h_n \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ and $h_n^5 \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ if, and only if, $h_n a(n) \to \infty$ and $h_n^5 a(n) \to 0$ respectively, with $a(n) = n^\beta \left(\log\log\left(n^\beta u(n)\right)\right)^{1-\beta} u(n \log\log n^\beta u(n))$ and $u(.)$ 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 $\frac{\widehat{L}_{n,h_n}(x)}{a(n)}$, under mild regularity conditions, satisfies a strong law of large numbers, and thus $\widehat{L}_{n,h_n}(x) = O_{a.s.}\left(a(n)\right)$. As for the less intuitive "only if" part, it follows from the fact that, as shown in the Appendix, $\widehat{L}_{n,h_n}(x) = O_{a.s.}\left(a(n)\right) + O_p\left(\sqrt{\frac{a(n)}{h_n}}\right)$ and so $h_n \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ only if $h_n a(n) \to \infty$.

# 3 Nonparametric Cointegrating Regression

We now consider the following data generating process:

$$Y_t = f(X_t) + \alpha(X_t)\epsilon_t. \tag{6}$$

It is immediate to see that, whenever $X_t$ is a null recurrent Markov process or, using more traditional jargon, an integrated processes, and $\epsilon_t$ is short-memory, the data generating process in Eq. (6) can be viewed as a nonlinear generalization of the classical cointegrating equation.[9] In general, $Y_t$ and $X_t$ are jointly dependent, as they both belong to a larger structural model, and, consequently, $\epsilon_t$ 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 Tjostheim (2007) assume that $X_t$ is a $\beta$-recurrent Markov chain and extend the methodology outlined in the previous section to the multivariate case and to the possible

[9] The case of spurious nonparametric cointegration, occurring when $\epsilon_t$ is an integrated process, is studied in Phillips (2009).

endogeneity of $\epsilon_t$. Bandi (2004) and Wang and Phillips (2009a and 2009b), 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).

The key difference between the two cases lies in the different, but ultimately equivalent, representation of the asymptotic behavior of the estimated occupation density $\widehat{L}_{n,h_n}(x)$, as defined in Eq. (3). As mentioned, if $X_t$ is a $\beta$-recurrent Markov chain, then, as $n \to \infty$ and $h_n \to 0$ with $h_n n^\beta u(n) \to \infty$, $\widehat{L}_{n,h_n}(x)/n^\beta u(n) \xrightarrow{d} c_X \mathcal{M}_\beta$, where $c_X$ is a process-specific constant and $\mathcal{M}_\beta$ is the Mittag-Leffler density. When, instead, $X_t$ is an integrated process, then, as $n \to \infty$ and $h_n \to 0$ so that $h_n\sqrt{n} \to \infty$, $\widehat{L}_{n,h_n}(x)/\sqrt{n} \xrightarrow{d} L_0(0,1)$, 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.[10] The more explicit representation in the second case derives, of course, from the stronger (but more conventional in nonstationary econometrics) I(1) structure of the underlying process. Clearly, when setting $\beta = \frac{1}{2}$ (the Brownian motion case) in the first approach, we obtain $\mathcal{M}_{\frac{1}{2}} \overset{d}{=} L_0(0,1)$, where $\overset{d}{=}$ denotes equivalence in distribution. The common distribution is that of a truncated Gaussian random variable on a positive support.

Write now

$$\widehat{f}_{n,h_n^\mu}(x) = \frac{\sum_{j=1}^n Y_j K\left(\frac{X_j - x}{h_n^\mu}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h_n^\mu}\right)} \tag{7}$$

$$\widehat{f}^{(2)}_{n,h_n^\sigma}(x) = \frac{\sum_{j=1}^n Y_j^2 K\left(\frac{X_j - x}{h_n^\sigma}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h_n^\sigma}\right)}, \tag{8}$$

and $\widehat{\alpha}^2_{n,h_n}(x) = \widehat{f}^{(2)}_{n,h_n^\sigma}(x) - \left(\widehat{f}_{n,h_n^\mu}(x)\right)^2$.

When $X_t$ is $\beta$-recurrent and $\mathrm{E}(\epsilon_t|\mathcal{F}_t) = 0$ with $\mathcal{F}_t = \sigma(X_{t-1}, X_{t-2}, ...)$, the statement in Proposition 1 extends rather straightforwardly to the cointegrating regression case. In fact, if $\mathrm{E}(\epsilon_t|X_t) = 0$ and $\epsilon_t$ is geometrically strong mixing, given Assumption A$(i)$-$(v)$, consistency and asymptotic mixed normality follow directly from Theorem 3.5 in Karlsen, Myklebust, and Tjostheim (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 $\mathrm{E}(\epsilon_t|\mathcal{F}_t) = 0$, consistency and mixed asymptotic normality are established in Wang and Phillips (2009a).

[10] If $X_t$ is a near to integrated process, i.e., $X_t = \exp(c/n)X_{t-1} + v_t$ with $c < 0$ and $v_t$ strong mixing, then as $n \to \infty$ and $h_n \to 0$, $\widehat{L}_{n,h_n}(x)/\sqrt{n} \xrightarrow{d} L_c(0,1)$, where $L_c(0,1)$ is instead the local time of an Ornstein-Uhlenbeck process.

We now turn to the endogenous case in which $\epsilon_t$ 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 $\beta$-recurrent Markov chain. We then focus on the integrated (or near integrated) case.

In what follows, we will make use of Assumption B which largely corresponds to Assumptions D$_1$-D$_5$ in Karlsen, Myklebust, and Tjostheim (2007) and builds on Assumption A.

**Assumption B.**
$(i)$ The joint process $\{(X_t, \epsilon_t),\ t \geq 0\}$ is a $\phi-$irreducible Harris recurrent Markov chain on the state space $\left(\widetilde{\mathbf{E}}, \widetilde{\mathcal{E}}\right) = (\mathbf{E}_1 \times \mathbf{E}_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ 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 $\int p_s(x,\epsilon)\mathrm{d}\epsilon > 0$, $\lim_{\delta\downarrow 0}\int |p_s(x+\delta,\epsilon) - p_s(x,\epsilon)|\, \mathrm{d}\epsilon = 0$ and, for all $A_h \in \widetilde{\mathcal{E}}^{\infty}$ such that $A_h \downarrow \varnothing$, $\lim_{h\downarrow 0}\overline{\lim}_{y\to x}\int_{\epsilon} P\left((y,\epsilon), A_h\right)|\epsilon|\, \mathrm{d}\epsilon = 0$.
$(ii)$ The marginal process $X_t$ satisfies Assumption A$(i)$ and Assumption A$(iii)$-$(iv)$. In addition, the marginal transition probability function $P_1$ is independent of any initial distribution $\lambda$. The kernel function satisfies Assumption A$(iii)$.
$(iii)$ The residual $\epsilon$ has bounded support.
$(iv)$ $\int \epsilon p_{\epsilon|X}\left(\epsilon|x\right)\mathrm{d}\epsilon = 0$
$(v)$ The functions $f(x)$ and $\alpha(x)$ are locally twice continuously differentiable for all $x \in \mathcal{D}_x$.

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

**Proposition 2.** Let Assumption B be satisfied. Also, let $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(\mathrm{E}\left(Y_t|X_t\right)\right)^{2m} < \infty$ and $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(\mathrm{E}\left(Y_t^2|X_t\right)\right)^{2m} < \infty$, for $X_t$ in a neighborhood of $x$ and for some $m \geq 2$. Furthermore, assume that $\int \overline{\lim}_{y\to x}\left|\frac{\partial}{\partial^2 y}p_s(y,\epsilon)\right||\epsilon|\mathrm{d}\epsilon < \infty$.

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

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

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

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

Proposition 2(a) is adapted from Theorem 4.1 in Karlsen, Myklebust, and Tjostheim (2007). As earlier, in order to provide a feasible bandwidth selection procedure, we show that our rate conditions $h_n\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ and $h_n^5\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ 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 $\epsilon_t$ 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 B$(iv)$, one can interpret $f(x)$ as $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n \mathrm{E}\left(Y_t|X_t = x\right)$ and $f^{(2)}(x)$ as $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n \mathrm{E}\left(Y_t^2|X_t = x\right)$, with probability one.

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 corresponding result is contained in the following theorem.

**Proposition 3.** Let Assumption B hold and let $\left(\mathrm{E}\left(Y_t|X_{t-1}\right)\right)^{2m} < \infty$ and $\left(\mathrm{E}\left(Y_t^2|X_{t-1}\right)\right)^{2m} < \infty$ for $X_{t-1}$ in a neighborhood of $x$ for all $x \in \mathcal{D}_x$ and for $m \geq 2$.

If $\sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)^{1/2}h^{1/2}\ln^{1/2}(n)}\sum_{t=1}^n \mathrm{E}\left(K\left(\frac{X_t - x}{h}\right)\epsilon_t\alpha(X_t)\right)\right| = O(1)$ and $\inf_{x\in\mathcal{D}_x} p_s(x) \geq \delta > 0$, then:
(a)

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

(b) If, in addition, $\sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)^{1/2}h^{1/2}\ln^{1/2}(n)}\sum_{t=1}^n \mathrm{E}\left(K\left(\frac{X_t - x}{h}\right)\alpha(X_t)\left(\epsilon_t^2 - 1\right)\right)\right| = O(1)$,

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

The statement in Proposition 2 is similar to that in Theorem 4.2 in Gao, Li and Tjostheim (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 B$(iv)$. We simply require the dependence between $X_t$ and $\epsilon_t$ to go to zero fast enough.

We now turn to the case in which $X_t$ is an integrated process, not necessarily Markov, and $\epsilon_t$ in Eq. (6) is not independent of $X_t$. Assumption C(*ii*)-(*iv*) below corresponds to Assumptions 2-4 in Wang and Phillips (2009b) while Assumption C(*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 C.**
(*i*) $X_t = X_{t-1} + \xi_t$, $\xi_t = \sum_{k=0}^{\infty} \phi_k \eta_{t-k}$, where (a) $\mathrm{E}\left(|\xi_t|^{2(4+\gamma)}\right) \leq C_1 < \infty$ for $\gamma > 0$, (b) $\eta_k$ is iid, (c) $\phi_k$ decays fast enough, as $k \to \infty$, as to ensure that $\xi_t$ is $\alpha-$mixing with size $-(4(4+\gamma))/\gamma$, and (d) there exists $0 < \omega_0^2 < \infty$ so that $\left|T^{-1}\mathrm{E}\left(\left(\sum_{k=m+1}^{m+T} \xi_k\right)^2\right) - \omega_0^2\right| \leq C_2 T^{-\psi}$, with $\psi > 0$ and $C_2$ independent of $m$.
(*ii*) $K$ is a second-order kernel, bounded and with bounded support, and $\int \left|e^{ixt}K(t)\mathrm{d}t\right| \mathrm{d}x < \infty$
(*iii*) $\epsilon_t$ as defined in Eq. (6) writes as $\epsilon_t = g\left(\eta_t, ..., \eta_{t-m_0}\right)$, where $g$ is a measurable function on $\mathcal{R}^{m_0}$ and $m_0 < \infty$. In addition, $\eta_t = 0$ for $t = 1, ..., m_0 - 1$, $\mathrm{E}(\epsilon_t) = 0$, $\mathrm{E}\left(\epsilon_t^4\right) < \infty$.[11]
(*iv*) The functions $f(x)$ and $\alpha(x)$ are locally twice continuously differentiable for all $x \in \mathcal{D}_x$.

**Proposition 4.** Let Assumption C hold.
(a) If (*i*) $h_n^{\mu}\widehat{L}_{n,h_n^{\mu}}(x) \overset{a.s.}{\to} \infty$ and (*ii*) $h_n^{\mu 5}\widehat{L}_{n,h_n^{\mu}}(x) \overset{a.s.}{\to} 0$, then

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

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

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

The statement in Proposition 4(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 $h_n\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ if, and only if, $\sqrt{n}h_n \to \infty$ and that $h_n^5\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ only if $\sqrt{n}h_n^5 \to 0$. On the other hand, $\sqrt{n}h_n^5 \to 0$ implies $h_n^5\widehat{L}_{n,h_n}(x) \xrightarrow{p} 0$, not necessarily $h_n^5\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$. Proposition 4(b) follows naturally.

Contrary to the general $\beta$-recurrent case, for which $\beta$ 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 $\varepsilon > 0$

[11] As in Assumption 2 in Wang and Phillips (2009b), $\epsilon_t$ may also depend on a finite number of lags of another iid process, say $\lambda_t$, which is independent of $\eta_t$.

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 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 $\widehat{L}_{n,h_n}(x)$ diverges at rate $\sqrt{n}$. If the null is rejected in favor of stationarity, however, then $\widehat{L}_{n,h_n}(x)$ 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 Proposition 4 hold even if $X_t$ in Assumption C($i$) is replaced by $X_t = \alpha X_{t-1} + \xi_t$ with $|a| \leq 1$. Hence, Proposition 4, like Proposition 1-3 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.[12]

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 $c \leq 0$. In our context, we could allow for $c < 0$ at the cost of stating our rate conditions as $h_n\widehat{L}_{n,h_n}(x) \xrightarrow{p} \infty$ and $h_n^5\widehat{L}_{n,h_n}(x) \xrightarrow{p} 0$, i.e., by weakening the almost-sure rates to rates in probability. Thus, in practical applications, we can employ $\widehat{L}_{n,h_n}(x)$, instead of $\sqrt{n}$, even in the case of near-integrated processes. Finally, we establish a uniform consistency result, which will be needed in the next Section.

**Proposition 5.** Let Assumption C hold. Furthermore, assume that $\sup_{x\in\mathcal{D}_x}\left|\frac{1}{n^{1/4}h^{1/2}\ln^{1/2}(n)}\sum_{t=1}^{n}\mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\alpha(X_t)\epsilon_t\,|\mathcal{F}_t\right)\right| = O_p(1)$, where $\mathcal{F}_t = \sigma\left(X_1, ..., X_t\right)$, and that $\mathrm{E}\left(\exp\left(K\left(\frac{X_t-x}{h}\right)\alpha(X_t)\epsilon_t\right)\right) \leq \Delta < \infty$. Then:
(a)

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

(b) If, in addition, $\sup_{x\in\mathcal{D}_x}\left|\frac{1}{n^{1/4}h^{1/2}\ln^{1/2}(n)}\sum_{t=1}^{n}\mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\alpha(X_t)\left(\epsilon_t^2-1\right)|\mathcal{F}_t\right)\right| = O_p(1)$, and $\mathrm{E}\left(\exp\left(K\left(\frac{X_t-x}{h}\right)\alpha(X_t)\left(\epsilon_t^2-1\right)\right)\right) \leq \Delta < \infty$,

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

[12] Though, in the stationary case, Assumption C($iii$) no longer suffices and one needs stronger exogeneity conditions.

The uniform rate in Proposition 5(a) requires two additional conditions. The first condition, controlling the rate at which the dependence between $\epsilon_t$ and $(X_1, ..., X_t)$ approaches zero, allows us to treat the term $K\left(\frac{X_t - x}{h}\right)\alpha(X_t)\epsilon_t$ as a martingale difference sequence. The second is a Cramèr-type condition permitting the use of exponential inequalities for unbounded martingales, e.g., Lesigné and Volny (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 automated 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. An important contribution in this area is, however, the recent work by Guerre (2004), in which a bandwidth based on the minimization of the empirical bias-variance trade-off is suggested. In terms of our notation, Guerre's adaptive bandwidth is defined as

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

where $L$ is the Lipschitz constant characterizing the conditional expectation function, i.e. $|\mu(x) - \mu(x')| \leq L|x - x'|$ and $\sigma^2$ is so that $\mathrm{E}\left(u_i^2|X_i\right) \leq \sigma^2$. The selected bandwidth is a function of two constants, $L$ and $\sigma^2$, which are, in general, unknown.[13] It is, therefore, not automated.

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 $\sigma^2$, and is, therefore, fully automated. We begin by outlining the case in which we select a local bandwidth which depends on the evaluation point.

Let

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

Let, also, $w_{i,h_n^\mu}(x) = 1\left\{|X_i - x| < h_n^\mu\right\} / \sum_{i=1}^n 1\left\{|X_i - x| < h_n^\mu\right\}$ and define

$$\widehat{m}_{n,h_n}^u(x) = \left(\begin{array}{c} \sum_{i=1}^n \widehat{u}_{i,h_n} w_{i-1,h_n^\mu}(x) \\ \sum_{i=1}^n \widehat{u}_{i,h_n}^2 w_{i-1,h_n^\mu}(x) - 1 \end{array}\right),$$

[13] Guerre (2004) assumes only a first order Lipschitz condition for $\mu$. Under our assumption of twice continuous differentiability around $x$, we would have $L^4 h^4$ instead of $L^2 h^2$.

and

$$\widehat{m}^{\epsilon}_{n,h_n}(x) = \left( \begin{array}{c} \sum_{i=1}^{n} \widehat{\epsilon}_{i,h_n} w_{i,h_n^{\mu}}(x) \\ \sum_{i=1}^{n} \widehat{\epsilon}^{2}_{i,h_n} w_{i,h_n^{\mu}}(x) - 1 \end{array} \right),$$

where $h_n = \left(h_n^{\mu}, h_n^{\sigma}\right)$. Needless to say, one could employ a larger and/or different set of moment conditions. Here, we limit our attention to the first two conditional moments for conciseness but show, in Section 6, that this choice translates into satisfactory finite-sample performance.

We begin with the case of nonparametric autoregression. The bandwidth vector $\widehat{h}_n$ is selected as:

$$\widehat{h}_n(x) = \left(\widehat{h}_n^{\mu}(x), \widehat{h}_n^{\sigma}(x)\right) = \arg\inf_{h_n} \left\| \widehat{m}^{u}_{n,h_n}(x) \right\|, \tag{9}$$

where $\|.\|$ denotes the Euclidean norm. It is immediate to see that $\sum_{i=1}^{n} \widehat{u}_{i,h_n} w_{i-1,h_n^{\mu}}(x) = \sum_{i=1}^{n} u_{i,h_n} w_{i-1,h_n^{\mu}}(x) +$ $o_p(1)$ if, and only if, $\left|\widehat{\mu}_{n,h_n^{\mu}}(x) - \mu(x)\right| = o_p(1)$ and, analogously, $\sum_{i=1}^{n} \widehat{u}^{2}_{i,h_n} w_{i-1,h_n^{\mu}}(x) = \sum_{i=1}^{n} u^{2}_{i,h_n} w_{i-1,h_n^{\mu}}(x) +$ $o_p(1)$ if, and only if, $\left|\widehat{\mu}_{n,h_n^{\mu}}(x) - \mu(x)\right| = o_p(1)$ and $\left|\widehat{\mu}^{(2)}_{n,h_n^{\sigma}}(x) - \mu^{(2)}(x)\right| = o_p(1)$. Also, $\sum_{i=1}^{n} u_{i,h_n} w_{i-1,h_n^{\mu}}(x) =$ $o_p(1)$ and $\sum_{i=1}^{n} u^{2}_{i,h_n} w_{i-1,h_n^{\mu}}(x) = 1 + o_p(1)$ since, by construction, $\mathrm{E}\left(u_i | X_{i-1}\right) = 0$ and $\mathrm{E}\left(u_i^2 | X_{i-1}\right) = 1$. Thus, the bandwidth vector selected according to Eq. (9) ensures the consistency of the first two conditional moment estimators. Given Assumption A, 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 defined analogously. Specifically,

$$\widetilde{h}_n(x) = \left(\widetilde{h}_n^{\mu}(x), \widetilde{h}_n^{\sigma}(x)\right) = \arg\inf_{h_n} \left\| \widehat{m}^{\epsilon}_{n,h_n}(x) \right\| \tag{10}$$

As already pointed out, we wish to allow for $\mathrm{E}\left(\epsilon_i | X_i\right) \neq 0$. Nonetheless, under either Assumption B($iv$) in the $\beta$-recurrent case, or Assumption C($i$) and Assumption C($iii$) in the case of integrated processes, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left(\epsilon_i | X_i\right) \to 0$ and $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left(\left(\epsilon_i^2 - 1\right) | X_i\right) \to 0$. In these cases, therefore, $\sum_{i=1}^{n} \widehat{\epsilon}_{i,h_n} w_{i,h_n^{\mu}}(x) = o_p(1)$ and $\sum_{i=1}^{n} \widehat{\epsilon}^{2}_{i,h_n} w_{i,h_n^{\mu}}(x) = 1 + o_p(1)$ if, and only if, $\left|\widehat{f}_{n,h_n^{\mu}}(x) - f(x)\right| =$ $o_p(1)$ and $\left|\widehat{f}^{(2)}_{n,h_n^{\sigma}}(x) - f^{(2)}(x)\right| = o_p(1)$. Moreover, under additional conditions (in Propositions 3 and 5) on the rate at which $\mathrm{E}\left(\epsilon_i | \mathcal{F}_i\right)$ and $\mathrm{E}\left(\left(\epsilon_i^2 - 1\right) | \mathcal{F}_i\right)$ approach zero, $\widetilde{h}_n(x)$ is also rate optimal up to a logarithmic factor.

It is evident from the definition of $\widehat{h}_n(x)$ and $\widetilde{h}_n(x)$ 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 6.** Assume that the kernel $K$ is twice continuously differentiable on the interior of its support.

(a) *Nonparametric Autoregression.* Let the assumptions in Proposition 1 hold. Then, for $i = \mu, \sigma$, $\widehat{h}_n^i(x)$, as defined in Eq. (9), 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, $\beta < 1$ and

$$\gamma(n) = a(n) = n^{\beta} \left(\log\log\left(n^{\beta}u(n)\right)\right)^{1-\beta} u(n \log\log n^{\beta}u(n)), \tag{11}$$

where $u(b(n))$ denotes a slowly-varying function as $b(n) \to \infty$.

(b) *Nonparametric cointegration.* Either (b1) the assumptions in Proposition 3 hold or (b2) the assumptions in Proposition 5 hold. Then, for $i = \mu, \sigma$, $\widetilde{h}_n^i(x)$, as defined in Eq. (10), 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 (a) if (b1) holds or is $\gamma(n) = n^{1/2}$ if (b2) holds.

**Remark 1.** As established in Theorem 6, 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 $\beta$-recurrent processes. The logarithmic factor is the same cost of adaptation as in, e.g., Lepski (1990), 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 6 proposes an automated procedure for selecting a variable bandwidth vector ensuring pointwise consistent estimation. The theorem below establishes that there exist rate-optimal (again, up to a logarithmic factor) uniform bandwidths. Let

$$\widehat{h}_n = \left(\widehat{h}_n^{\mu}, \widehat{h}_n^{\sigma}\right) = \arg\inf_{h_n} \sup_{x \in \mathcal{D}_x} \left\| \widehat{m}_{n,h_n}^{u}(x) \right\|, \tag{12}$$

and

$$\widetilde{h}_n = \left(\widetilde{h}_n^{\mu}, \widetilde{h}_n^{\sigma}\right) = \arg\inf_{h_n} \sup_{x \in \mathcal{D}_x} \left\| \widehat{m}_{n,h_n}^{\epsilon}(x) \right\|. \tag{13}$$

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

(a) *Nonparametric autoregression.* Let the assumptions in Proposition 1 hold. Then, for $i = \mu, \sigma$, $\widehat{h}_n^i$, as defined in Eq. (12), is of probability order $\log^{1/5}(n)\gamma(n)^{-1/5}$, where $\gamma(n)$ is defined as in Part (a) of Theorem 6.

(b) *Nonparametric cointegration.* Either (b1) the assumptions in Proposition 3 hold or (b2) the assumptions in Proposition 5 hold. Then, for $i = \mu, \sigma$, $\widetilde{h}_n^i$, as defined in Eq. (13), is of probability order

$\log^{1/5}(n)\gamma(n)^{-1/5}$, where $\gamma(n)$ is defined as in Part $(a)$ of Theorem 6 if (b1) holds or is $\gamma(n) = n^{1/2}$ if (b2) holds.

# 5 Bias correction

## 5.1 The point-wise test

The previously-discussed adaptive bandwidths are large enough as to ensure the consistency of the estimators of the first two conditional moments. However, in light of their minimax optimality, they are too large to satisfy the condition for zero-mean asymptotic (mixed) normality. The purpose of the bias correction procedure introduced in this section is to select the largest bandwidth for which the bias approaches zero. The outcome of the procedure will tell us whether to keep the bandwidth originally selected or whether to search for a smaller one. In the latter case, a simple stopping rule will guarantee that the final bandwidth is the largest bandwidth leading to a zero-mean asymptotic normal distribution. We emphasize that the suggested bias correction is made possible by our representation of the bandwidth conditions as functions of the process' occupation density (as in Sections 2 and 3).

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

The hypothesis of interest is

$$H_0^{\mu}(x): \ \widehat{h}_n^{\mu(5-\varepsilon)}(x) \sum_{j=1}^{n} K_{\widehat{h}_n^{\mu}(x)}\left(X_j - x\right) \overset{a.s.}{\to} \infty \tag{14}$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$, $x \in \mathcal{D}_x$, and $\varepsilon > 0$ arbitrarily small, versus[14]

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

The role of $\varepsilon > 0$ is to ensure that rejection of the null implies $\widehat{h}_n^{\mu 5}\widehat{L}_{n,h_n^{\mu}}(x) \overset{a.s.}{\to} 0$. It is immediate to see that if we reject the null the selected bandwidth satisfies the required rate condition for a vanishing asymptotic bias and it should be kept. If we fail to reject, then we need to search for a smaller bandwidth.

[14] $H_0^{\sigma}(x)$ and $H_A^{\sigma}(x)$ are defined in an analogous way, simply replacing $\widehat{h}_n^{\mu}(x)$ with $\widehat{h}_n^{\sigma}(x)$, of course.

**Remark 2.** We note, crucially, that the same bandwidth $\widehat{h}^{\mu}_{n}(x)$ should appear outside and inside of the kernel. Failure to do so would result in fundamental inconsistencies. An obvious implication of this observation is that setting the bandwidth as a function of the estimated occupation density, i.e., $\widehat{h}^{\mu}_{n}(x) \propto \widehat{L}^{\theta}_{n,h^{\mu}_{n}}(x)$ for some $\theta$, (in just the same way as we set it as a function of the number of observations in stationary frameworks) is not possible since the same bandwidth would appear on the left-hand side and on the right-hand side of the equation. Plug-in procedures based on asymptotic MSEs and classical rules-of-thumb would therefore be even less operational in nonstationary environments than they are in stationary environments.

Following Bandi, Corradi, and Moloche (2009), we define

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

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

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

with $\boldsymbol{\eta}$ ∼iid$N(0, I_R)$.

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 $\boldsymbol{\eta}$, i.e., a standard normal, conditional on the sample. 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 samples but a set of measure 0".

Suppose that $\widehat{h}^{\mu 5}_{n}(x)\sum_{j=1}^{n} K_{\widehat{h}^{\mu}_{n}}\left(X_j - x\right) \overset{a.s.}{\to} \infty$. Then, conditionally on the sample and $a.s. - P$, $v_{j,n}$ diverges to $\infty$ with probability 1/2 and to $-\infty$ with probability 1/2. Thus, as $n \to \infty$, for any $u \in U$, $1\left\{v_{j,n} \leq u\right\}$ will be distributed as a Bernoulli random variable with parameter 1/2. Furthermore, note that, as $n \to \infty$, for any $u \in U$, $1\left\{v_{j,n} \leq u\right\}$ is equal to either 1 or 0 regardless of the evaluation point $u$. In consequence, as $n, R \to \infty$, for all $u, u' \in U$, $\frac{2}{\sqrt{R}}\sum_{j=1}^{R}\left(1\left\{v_{j,n} \leq u\right\} - \frac{1}{2}\right)$ and $\frac{2}{\sqrt{R}}\sum_{j=1}^{R}\left(1\left\{v_{j,n} \leq u'\right\} - \frac{1}{2}\right)$ will converge in $d^*-$distribution to the same standard normal random variable. Thus, $\widetilde{V}_{R,n} \overset{d^*}{\to} \chi^2_1$ $a.s. - P$. It is now immediate to notice that, for all $u \in U$, $V_{R,n}(u)$ and $\widetilde{V}_{R,n}$ 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 $\widehat{h}^{\mu 5}_{n}(x)\sum_{j=1}^{n} K_{\widehat{h}^{\mu}_{n}}\left(X_j - x\right) \overset{a.s.}{\to} 0$, (or, if $\widehat{h}^{\mu 5}_{n}(x)\sum_{j=1}^{n} K_{\widehat{h}^{\mu}_{n}}\left(X_j - x\right) = O_{a.s.}(1)$), then $v_{j,n}$, as $n \to \infty$, conditionally on the sample and $a.s. - P$, will converge to a (mixed) zero-mean normal random variable. Thus, $\frac{2}{\sqrt{R}}\sum_{j=1}^{R}\left(1\left\{v_{j,n}\leq u\right\}-\frac{1}{2}\right)$ will diverge to infinity at speed $\sqrt{R}$ whenever $u\neq 0$ $a.s.-P$.

**Theorem 8.** Let Assumption A, B, or C hold. As $R,n\to\infty$,[15]
(a) Under $H_0^{\mu}(x)$,

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

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

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

If we fail to reject $H_0^{\mu}(x)$ because $V_{R,n}$ is smaller than, say, the 95% percentile of a chi-squared 1 random variable, then we should choose a smaller bandwidth until rejection is reached. Specifically, we should proceed by searching on a grid until $H_0^{\mu}(x)$ is rejected, i.e., until reaching

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

It is immediate to see that the suggested stopping rule leads to the choice of the largest bandwidth ensuring a zero asymptotic bias.

## 5.2 The uniform test

Let $\widehat{h}_n=\left(\widehat{h}_n^{\mu},\widehat{h}_n^{\sigma}\right)$ be the uniform bandwidth vector previously chosen (c.f., Theorem 6). 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:

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

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 $\widehat{L}_{n,\widehat{h}_n^{\mu}}(x)$ over evaluation points, i.e.,

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

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

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.

## 6 Simulations

We now report a simulation experiment which applies our bandwidth selection procedure, as well as the proposed bias correction, and illustrates their finite sample performance. Three different data generating processes are considered:

**Model I** As an example of a nonstationary autoregression we simulate a simple unit root process ($\mu(x) = x$ and $\sigma(x) = 1$), viz.

$$X_t = X_{t-1} + u_t.$$

We choose $x_0 = 0$, $\mathcal{D}_x = [-5, 5]$ and let $u_t$ be i.i.d. $N(0, 1)$.

**Model II** The discrete-time square-root process is an autoregression with $\mu(x) = (1 - \phi)\theta + \phi x$ and $\sigma(x) = \sigma\sqrt{|x|}$, viz.

$$X_t = (1 - \phi)\theta + \phi X_{t-1} + \sigma\sqrt{|X_{t-1}|}u_t$$

whose parameters are chosen to be $\theta = 1$, $\phi = 0.8$, $\sigma = 1$ and $\mathcal{D}_x = [0, 4]$. We start the process at its unconditional mean $x_0 = \theta$ and, again, $u_t$ is i.i.d. $N(0, 1)$.

**Model III** To illustrate our procedure in the case of cointegrating regressions, we consider a simulation design similar[16] to the one in Hall and Horowitz (2005) and Wang and Phillips (2009b) with

$$f(x) = \sum_{j=1}^{4} \frac{(-1)^{j+1}\sin(j\pi x)}{j^2}$$

and $a(x) = 1$, viz.

$$\begin{aligned} Y_t &= f(X_t) + \epsilon_t \\ X_t &= X_{t-1} + u_t \\ \epsilon_t &= \frac{\eta_t + \theta u_t}{\sqrt{1 + \theta^2}} \end{aligned}$$

where $(u_t, \epsilon_t, \eta_t)'$ i.i.d. $N(0, I_3)$, $I_3$ a diagonal matrix of ones, $x_0 = 0$ and $\mathcal{D}_x = [0, 1]$. We consider two scenarios: no endogeneity ($\theta = 0$) and strong endogeneity ($\theta = 2$).

[16] We only deviate in our specification of the function $\alpha(\cdot)$.

To summarize, there are four simulation scenarios: model I, model II, and two versions of model III, each of which is estimated using our point-wise and uniform criteria for selecting the bandwidths. Even though cross-validation has not been formally justified in a nonstationary framework, it is the classical paradigm in empirical work and we therefore consider it here as an important benchmark.

## 6.1 Implementation details

The conditional moments impose the same requirements on the rate of divergence of the relevant bandwidth sequences. However, the optimization to find

$$h_n(x) = (h_n^\mu(x), h_n^\sigma(x)) = \arg\inf_{h_n} \|\hat{m}_{n,h_n}(x)\|\,,$$

in the point-wise case and

$$h_n = (h_n^\mu, h_n^\sigma) = \arg\inf_{h_n} \sup_{x\in\mathcal{D}_x} \|\hat{m}_{n,h_n}(x)\|\,,$$

in the uniform case is performed with separate bandwidths for the first and the second conditional moment in order to improve finite sample accuracy. Specifically, we implement a search over a grid of $5 \times 5$ bandwidths on $[0.01, 1]^2$. The bias correction is instead implemented by virtue of a search over a $100 \times 100$ grid on $[0.01, 10]^2$. The supremum over $x$ in the uniform criterion is calculated over a grid of five equally spaced points in $\mathcal{D}_x$. For the point-wise criterion, $\mathcal{D}_x$ is partitioned into five parts of equal size. Five bandwidths are calculated at the center of each of the five subsets of $\mathcal{D}_x$. Since determining the partition depends on the path and introduces extraneous randomness, we choose it to be the same for every simulated path which, in turn, creates issues which are, admittedly, little understood in the literature. For example, it could be the case that a given simulated path does not visit a certain region of the domain at all, or only very few times, so that estimation of a function in that region can be based only on certain paths, but not on all. To minimize these effects, we restrict estimation of the various functions to areas near the processes' points of initialization and, thus, all paths take values in at least some portion of those regions.

The remaining parameters are $R = 200$ and uniform weights $\pi(u) = 1$ over the interval $U = [2, 3]$. Throughout the experiment we use the Tukey-Hanning kernel. The second-stage tests are performed at the 95% confidence level. All results are based on $1,000$ Monte Carlo samples of length 500.

## 6.2 Results

Tables 1 – 3 report the selected bandwidths for models I – III calculated using our point-wise and uniform procedures as well as cross-validation ("CV"). We emphasize that the second-step cross-validated bandwidths have been obtained by applying our bias correction to the original cross-validated bandwidths.

Thus, importantly, when comparing classical cross-validation to our combined methods one should compare the first-step cross-validated bandwidths to our final bandwidths (inclusive of the bias correction).

Tables 4 – 6 present the bias, standard deviation ("SD") and root mean square error ("RMSE") of the estimated functions, averaged over 20 equally-spaced points in their respective domains $\mathcal{D}_x$.

Figures 1 – 4 show the corresponding estimates of the first and second conditional moment functions, $\mu(x)$ and $\sigma(x)$ or $f(x)$ and $\alpha(x)$, respectively. Included in the graphs are the true line (thick blue), the line based on our uniform criterion (blue circles) and the cross-validated estimates (red squares) as well as empirical (point-wise) 95% confidence bands. The graphs corresponding to the point-wise criterion, which are similar to the reported ones, are not shown to save space.

Figures 5 – 8 depict the kernel density estimates of the first conditional moment estimates at the fixed points $x = 0$ for model I, $x = 2$ for model II, and $x = 0.5$ for model III. Specifically, we estimated the density of the centered and re-scaled quantity

$$\sqrt{\frac{\hat{h}_n^{\mu}\hat{L}_{n,\hat{h}_n^{\mu}}(x)}{K_2\sigma(x)^2}}\,\left(\hat{\mu}_{n,\hat{h}_n^{\mu}}(x)-\mu(x)\right)$$

or

$$\sqrt{\frac{\hat{h}_n^{f}\hat{L}_{n,\hat{h}_n^{f}}(x)}{K_2\alpha(x)^2}}\,\left(\hat{f}_{n,\hat{h}_n^{f}}(x)-f(x)\right),$$

respectively, where $\mu$, $\sigma$, $f$ and $\alpha$ are the true functions. Again, for brevity, graphs are only shown for the uniform criterion and the first conditional moment.

The findings can be summarized as follows:

1. Our combined procedure outperforms cross-validation. In the first and in the third model, the point-wise and the uniform criteria produce comparable (relative to cross-validation), or slightly lower, RMSEs in both stages. In these two specifications, the second conditional moment is flat and, since cross-validation tends to oversmooth in these models, these are scenarios in favor of a uniform criterion like cross-validation. In model II, however, the nonlinear second conditional moment of the process reveals a dramatic difference in relative performance. The bandwidths selected by cross-validation are much too small leading to a large variance of the resulting estimates and an RMSE which is more than twice as large as the ones produced by our combined procedure.

2. As discussed, the proposed bandwidth procedure optimally balances the estimators' biases and variances. This may, of course, be achieved by choosing relatively large bandwidths $h_n^{\mu}$ and $h_n^{\sigma}$ which have the potential to cause some oversmoothing (see, e.g., model III). The reported bias correction is designed to address this issue explicitly since it forces the bandwidths to also satisfy the

conditions $(h_n^\mu)^5 \hat{L}_{n,h_n^\mu}(x) \overset{a.s.}{\to} 0$ and $(h_n^\sigma)^5 \hat{L}_{n,h_n^\sigma}(x) \overset{a.s.}{\to} 0$, which are necessary for a vanishing limiting bias. These conditions require both bandwidths to be small enough. Tables 1 – 3 show significant reductions in the size of the bandwidths after the second-stage procedure is applied.[17] This effect can also be seen by inspecting Figures 5 – 8 which show that the bias correction successfully re-adjust the distribution of the first moment estimator towards the normal distribution – in model III strikingly so.

3. The properties of cross-validated bandwidths in nonstationary frameworks are unknown. However, the results in this section suggest that they may not necessarily perform poorly in such scenarios (see models I and III). Importantly, however, if cross-validated smoothing sequences are used in practice, in light of their tendency to oversmooth, we find that their performance can be further enhanced by applying to them our proposed bias correction.

4. Table 3, Table 6 and Figures 3 – 4, 7 – 8 all confirm our theoretical results on cointegrating regressions, namely that – whether the regressor and the error are independent or not – the distributions of the first and second conditional moment estimates conform with a zero-mean normal distribution after applying our combined procedure. The results show no difference in performance with or without dependence. Since the presence of this type of endogeneity is common in empirical work, this is an important feature of our proposed method.

## 7 Conclusions

In nonstationary frameworks, the rate conditions which ought to be satisfied by the smoothing parameter are, in general, not operational in that they depend on the unknown regularity of the chain. In stationary frameworks, these conditions are known to depend on the divergence rate of the number of observations but this is, of course, a purely theoretical statement having little to do with the actual dynamic properties of the series in any given sample. Our representation of the rate conditions in terms of the process' occupation density contributes to making existing functional theories more operational. It also clarifies

[17] The fact that, in model II (Table 2), the second stage adjusts the average cross-validation bandwidths from about 0.3 down to about 0.15 while some of the average pointwise and uniform bandwidths are not rejected at levels of about 0.3 may seem puzzling at first glance. This effect is due to the large variability in the bandwidths selected by cross-validation: it mostly chooses bandwidths much smaller than 0.3, but also some huge ones (reflected in the large standard deviation of the first stage). The second stage does not reject the former, but adjusts downwards the latter to values around 0.3, which in turn yields an average bandwidth smaller than 0.3. On the other hand, the uniform and pointwise criteria tend to select bandwidths between 0.4 and 0.7 with a small standard deviation so that the second step decreases most of them down to values near 0.3 leading to an average of that order of magnitude.

that the rate of divergence of the smoothing parameter should be tailored to the "in-sample" divergence properties of the number of visits at points around which estimation is performed.

The core of our contribution is to provide a data-driven, minimax optimal up to a logarithmic term, method of bandwidth choice which does not hinge on the "notional" divergence rates of the number of observations (as is the case for plug-in procedures, classical rules-of-thumb, and - in a sense - cross-validation), but explicitly adapts to the point-wise and uniform divergence rates of the occupation density of the conditioning variable(s). In light of the absence of fully automated (parameter-free) bandwidth selection procedures in the nonstationary case, the methods are presented in this context. However, they are equally applicable to stationary models. Importantly, in our view, they are naturally suited to provide guidance in selecting the smoothing sequence when one is unwilling to make assumptions on the dynamic properties of the underlying series while, at the same time, being true to the in-sample information at individual estimation levels, as represented by the data's occupation densities at these levels.

Since minimax optimality is generally the final objective, we emphasize that one could stop here. It is, however, a theoretical fact - one which we confirm through simulations - that, due to their optimality, minimax optimal bandwidths may lead to finite sample biases and coverage which might be improved upon. Selecting the largest bandwidth for which the bias approaches zero may therefore be beneficial. We exploit our representation of the bandwidth conditions in terms of functionals of occupation densities to propose a solution to this issue.

# 8 Appendix

**Proof of Proposition 1.** (a) Hereafter, for notational simplicity we omit the superscript $\mu$, i.e. we write $h_n$ instead of $h_n^{\mu}$. We first need to show that $h_n\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ if, and only if, $h_n a(n) \to \infty$, where $a(n)$

$$a(n) = n^{\beta}\left(\log\log\left(n^{\beta}u(n)\right)\right)^{1-\beta} u(n \log\log n^{\beta}u(n)), \tag{15}$$

where $u(b(n))$ denotes a slowly varying function as $b(n) \to \infty$. We begin with the "if" part. Given Assumption A($i$), following Karlsen and Tjostheim (2001, KT01 hereafter), $\widehat{L}_{n,h_n}(x)$ in Eq. (3) can be re-written as a split chain, i.e.,

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

where

$$U_{k,x,h_n} = \begin{cases} \frac{1}{h_n}\sum_{j=1}^{\tau_0} K\left(\frac{X_{j-1}-x}{h_n}\right) \text{ when } k=0 \\ \frac{1}{h_n}\sum_{j=\tau_{k-1}+1}^{\tau_k} K\left(\frac{X_{j-1}-x}{h_n}\right) \text{ for } 1 \leq k < n \\ \frac{1}{h_n}\sum_{\tau_{T_n}+1}^{n} K\left(\frac{X_{j-1}-x}{h_n}\right) \text{ for } k = n. \end{cases}$$

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 $\tau_k$'s, with $k = 0, ..., n$, 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

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

where $\mu_{x,h_n} = \mathrm{E}\left(U_{k,x,h_n}\right)$. 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 $T_C(n) = \sum_{t=1}^{n} 1\left\{X_t \in C\right\}$. From Lemma 3.5 in KT01, it follows that $T_n$ and $T_C(n)$ are of the same almost-sure order. Furthermore, given A($i$)-($iii$), from Theorem 2 in Chen (1999), it follows that $T_C(n)$ is of almost-sure order $a(n)$, where $a(n)$ is defined in Eq. (15). Hence, both $T_n$ and $T_C(n)$ are of almost-sure order $a(n)$. Let, now,

$$\Omega_n = \left\{\omega : \underline{\Delta} \leq \lim_{n\to\infty} \frac{T_n}{a(n)} \leq \overline{\Delta}\right\},$$

with $0 < \underline{\Delta} \leq \overline{\Delta} < \infty$, and note that, because of Lemma 3.5 in KT01 and Theorem 2 in Chen (1999), $P\left(\lim_{n\to\infty}\Omega_n\right) = 1$. We can then proceed conditionally on $\omega \in \Omega_n$. Assume, without loss 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:

$$\begin{aligned} &\mathrm{E}\left(\left(\frac{1}{a(n)}\sum_{k=1}^{a(n)}\left(U_{k,x,h_n} - \mu_{x,h_n}\right)\right)^{2m}\right) \\ \simeq\ & \frac{1}{a(n)^{2m}}\sum_{k_1=1}^{a(n)} ... \sum_{k_m=1}^{a(n)} \mathrm{E}\left(U_{k_1,x,h_n}^2\right) ... \mathrm{E}\left(U_{k_m,x,h_n}^2\right) \\ \simeq\ & \frac{1}{a(n)^m} h_n^{-m}, \qquad (16) \end{aligned}$$

where $\simeq$ means "of the same order as", and the last term on the right-hand side of Eq. (16) comes from the fact that given A($iii$), by Lemma 5.2 in KT01, $\mathrm{E}\left(U_{k,x,h_n}^{2m}\right) \leq ch_n^{-2m+1}$. Thus, by Borel-Cantelli, letting $h_n = a(n)^{-\psi}$,

$$\begin{aligned}
&\limsup_n P\left(\left|\frac{1}{a(n)}\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)\right|>\delta\right)\\
\leq\ & a(n)P\left(\left|\frac{1}{a(n)}\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)\right|>\delta\right)\\
\leq\ & \frac{a(n)}{a(n)^{2m}\delta^{2m}}\mathrm{E}\left(\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)\right)^{2m}\\
\leq\ & c_m\frac{1}{\delta^{2m}}a(n)^{-m+1}h_n^{-m}\leq c_m\frac{1}{\delta^{2m}}a(n)^{-m+1+\psi m}, \qquad (17)
\end{aligned}$$

and $\frac{1}{a(n)}\sum_{k=1}^{a(n)}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)=o_{a.s.}(1)$, 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,

$$\begin{aligned}
\widehat{L}_{n,h_n}(x) &= U_{0,x,h_n}+\sum_{k=1}^{T_n}U_{k,x,h_n}+U_{n,x,h_n}\\
&= \sum_{k=1}^{T_n}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)+T_n\mu_{x,h_n}+o_{a.s.}\left(T_n\right)\\
&= \sum_{k=1}^{a_n}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)+a_n\mu_{x,h_n}+o_{a.s.}\left(a_n\right)\\
&= o_{a.s}(a_n)+O_{a.s.}(a_n), \qquad (18)
\end{aligned}$$

where the first term in the last equality in Eq. (18) holds when $h_na_n\to\infty$. Thus, as $h_na_n\to\infty$, we obtain $h_n\widehat{L}_{n,h_n}(x)\overset{a.s.}{\to}\infty$. This concludes the proof of the "if" part. As for the "only if" part, note that the first three equalities in Eq. (18) hold regardless of the speed at which $h_n$ approaches zero, hence

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

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

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

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

$$h_n^5\widehat{L}_{n,h_n}(x)=h_n^5\sum_{k=1}^{a_n}\left(U_{k,x,h_n}-\mu_{x,h_n}\right)+h_n^5a_n\mu_{x,h_n}+o_{a.s.}\left(h_n^5a_n\right), \qquad (19)$$

and it is immediate to see that $h_n^5 \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ 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, $\sum_{k=1}^{a_n} \left(U_{k,x,h_n} - \mu_{x,h_n}\right) = o_{a.s.}(a(n))$, and thus, from Eq. (19), we observe that if $h_n^5 a_n \to 0$, then $h_n^5 \widehat{L}_{n,h_n}(x)_n \overset{a.s.}{\to} 0$. On the other hand, if $h_n a(n) = O(1)$ or $o(1)$, by the same steps used in Eq. (17):

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

as $a(n) h_n^5 \to 0$, for $m \geq 2$. Thus, the second term on the right-hand side of Eq. (19) is $o_{a.s.}(h_n^5 a(n))$. Hence, if $h_n^5 a_n \to 0$, then $h_n^5 \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$. The statement in the Proposition now follows from Theorem 5.4 in KT01 by noting that their conditions $h_n^{-1} = o\left(n^{\beta-\varepsilon}\right)$ and $h_n^{-1} = o\left(n^{\beta/5+\varepsilon}\right)$ 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).

**Proof of Proposition 2.** (a) By the same argument as in the proof of Proposition 1, $h_n \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$ and $h_n^5 \widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} 0$ if, and only if, $h_n a(n) \to \infty$ and $h_n^5 a(n) \to 0$, respectively, since the divergence rate of $\widehat{L}_{n,h_n}(x)$ 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 Tjostheim (2007). (b) By the same argument as in (a).

**Proof of Proposition 3.** (a) As shown in Proposition 1, $h_n \widehat{L}_{n,h_n}(x) = h_n a(n) \left(1 + o_{a.s.}(1)\right)$. Hence, we need to show that

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

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 Tjostheim (2009), it suffices to focus on the variance term and show that

$$\sup_{x\in\mathcal{D}_x} \left|\frac{1}{h_n^\mu a(n)} \sum_{t=1}^{n} \alpha(X_t)\epsilon_t K\left(\frac{X_t - x}{h_n^\mu}\right)\right| = O_p\left(\sqrt{\frac{\log(n)}{h_n^\mu a(n)}}\right). \tag{20}$$

Now notice that

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\frac{1}{h_n^\mu a(n)}\sum_{t=1}^n \alpha(X_t)\epsilon_t K\left(\frac{X_t-x}{h_n^\mu}\right)\right| \\
\leq\ &\sup_{x\in\mathcal{D}_x}\left|\frac{1}{h_n^\mu a(n)}\sum_{t=1}^n \alpha(X_t)\epsilon_t K\left(\frac{X_t-x}{h_n^\mu}\right)-\frac{1}{h_n^\mu a(n)}\sum_{t=1}^n \mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\epsilon_t\alpha(X_t)\right)\right| \\
&+\sup_{x\in\mathcal{D}_x}\left|\frac{1}{h_n^\mu a(n)}\sum_{t=1}^n \mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\epsilon_t\alpha(X_t)\right)\right|.
\end{aligned}$$

Given the condition $\sup_{x\in\mathcal{D}_x}\left|\frac{1}{a(n)^{1/2}h^{1/2}\ln^{1/2}(n)}\sum_{t=1}^n \mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\epsilon_t\alpha(X_t)\right)\right| = O(1)$, the bound becomes

$$\sup_{x\in\mathcal{D}_x}\left|\frac{1}{h_n^\mu a(n)}\sum_{t=1}^n \alpha(X_t)\epsilon_t K\left(\frac{X_t-x}{h_n^\mu}\right)-\frac{1}{h_n^\mu a(n)}\sum_{t=1}^n \mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\epsilon_t\alpha(X_t)\right)\right|+O_p\left(\sqrt{\frac{\log(n)}{h_n^\mu a(n)}}\right)$$

and we can proceed as if $\mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\epsilon_t\alpha(X_t)\right)=0$ 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 $Q_n=\frac{n}{a(n)^{1/2}h_n^{\mu 3/2}}$ balls $S_i$, centered at $s_i$, of radius $\frac{a(n)^{1/2}h_n^{\mu 3/2}}{n}$, $i=1,...,Q_n$. Now,

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

Given Assumption B$(ii)$-$(iv)$-$(v)$, 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 $I_{n,h_n^\mu}$, given Assumption B$(i)$-$(iv)$,

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

where $Z_k(s_j)=\sum_{t=\tau_k-1}^{\tau_k}\frac{1}{h_n^\mu}\alpha(X_t)\epsilon_t K\left(\frac{X_t-s_j}{h_n^\mu}\right)$, $\tau_k$, $k=1,...,T_n$, are the regeneration times, and $T_n$ is the number of complete regenerations. For each $j$, $Z_k(s_j)$, $k=1,...,T_n$, is a sequence of iid random variables so that $\max_{j=1,...,Q_n}\mathrm{E}\left(|Z_k(s_j)|^{2m}\right)=O\left(h_n^{\mu(-2m+1)}\right)$ (KT01, Lemma 5.2), with $m$ defined in the statement of the Theorem. As shown in the proof of Proposition 1, with probability one, $\underline{\Delta}\leq\lim_{n\to\infty}\frac{T_n}{a(n)}\leq\overline{\Delta}$, hence we can replace $a(n)$ with $T_n$. Now, given Assumption B($v$), by the same argument used in Hansen (2004, proof of Theorem 2), it follows that for some constant $C$,

$$\lim_{n\to 0}\Pr\left(\max_{j=1,\ldots,Q_n}\left|\frac{1}{a(n)}\sum_{k=1}^{a(n)}Z_k(s_j)1\left\{\left|Z_k(s_j)>a(n)^{1/2}h_n^{\mu(-1/2)}\right|\right\}\right|>C\sqrt{\frac{\log(n)}{h_n^{\mu}a(n)}}\right)=0.$$

Let $\widetilde{Z}_k(s_j)=Z_k(s_j)1\left\{\left|Z_k(s_j)\le a(n)^{1/2}h_n^{\mu(-1/2)}\right|\right\}$, given Assumption B($iv$), by Bernstein inequality for zero mean iid sequences (e.g., Theorem 2.18 in Fan and Yao, 2005), letting $\eta=C\sqrt{\frac{\log(n)}{h_n^{\mu}a(n)}}$,

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

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

**Proof of Proposition 4.** (a) As in the case of previous propositions, we only prove Part (a). We need to show that $h_n\widehat{L}_{n,h_n}(x)\overset{a.s.}{\to}\infty$ only if $h_n\sqrt{n}\to\infty$ and, analogously, $h_n^5\widehat{L}_{n,h_n}(x)\overset{a.s.}{\to}0$ only if $h_n^5\sqrt{n}\to 0$. Given Assumption C, the statement then follows from Theorem 3.1 in Wang and Phillips (2009b). Write

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

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

$$\begin{aligned}
\frac{1}{\sqrt{n}}\widehat{L}_{n,h_n}(x)=&\left(\frac{c_n}{n}\sum_{j=1}^{n}g\left(c_n x_{j,n}\right)-\frac{c_n}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty}g\left(c_n(x_{j,n}+z\epsilon)\right)\phi(z)\mathrm{d}z\right)\\
&+\left(\frac{c_n}{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty}g\left(c_n(x_{j,n}+z\epsilon)\right)\phi(z)\mathrm{d}z-\frac{1}{n}\sum_{j=1}^{n}\phi_\epsilon(x_{j,n})\right)+\frac{1}{n}\sum_{j=1}^{n}\phi_\epsilon(x_{j,n}).
\end{aligned}$$

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

$$\begin{aligned}
& \sup_{0\leq r\leq 1}\left|\frac{1}{n}\sum_{j=1}^{[nr]}\phi_\epsilon(x_{j,n}) - \int_0^r \phi_\epsilon(G\left(t\right))\mathrm{d}t\right| \\
\leq\ & \int_0^1 \left|\phi_\epsilon(x_{nt,n}) - \int_0^r \phi_\epsilon(G\left(t\right))\right|\mathrm{d}t + \frac{2}{n} \\
\leq\ & A_\epsilon \sup_{0\leq t\leq 1}\left|x_{[nt],n} - G\left(t\right)\right| + 2/n = o_{a.s.}\left(\sqrt{2\log\log n}\right), \qquad (21)
\end{aligned}$$

where $A_\epsilon$ is a term depending on $\epsilon$, and the last equality on the right-hand side of Eq. (21) follows from the fact that, given Assumption C($i$), by Lemma 2.1($i$) in Corradi (1999), uniformly in $t \in [0,1]$, $\left|x_{[nt],n} - G\left(t\right)\right| = o_{a.s.}\left(\sqrt{\log\log n}\right)$. Now, as $\epsilon \to 0$,

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

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 Jegannathan (2004), for any $\epsilon > 0$, and recalling that $\int K(u)\mathrm{d}u = 1$,

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

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

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

Thus,

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

that is

$$h_n\widehat{L}_{n,h_n}(x) = \sqrt{n}h_n L(0,1) + o_{a.s.}\left(\sqrt{n}h_n\sqrt{\log\log n}\right) + o_p(\sqrt{n}h_n). \qquad (22)$$

Because $L(0,1)$ is a continuous random variable, and so it is equal to zero with probability zero, it is immediate to see that, whenever $\sqrt{n}h_n \to \infty$, then $h_n\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$. Similarly, if $h_n\widehat{L}_{n,h_n}(x) \overset{a.s.}{\to} \infty$, then $\sqrt{n}h_n \to \infty$. Also,

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

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

**Proof of Proposition 5.** (a) As shown in Proposition 4, $h_n\widehat{L}_{n,h_n}(x) = h_n n^{1/2}\left(1 + o_{a.s.}(1)\right)$. Hence, it suffices to show that

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

Given the condition $\sup_{x\in\mathcal{D}_x}\left|\frac{1}{n^{1/4}h^{1/2}\ln^{1/2}(n)}\sum_{t=1}^n \mathrm{E}\left(K\left(\frac{X_t-x}{h}\right)\epsilon_t\alpha(X_t)\,|\mathcal{F}_t\right)\right| = O_p(1)$, we can proceed as if $K\left(\frac{X_t-x}{h}\right)\epsilon_t\alpha(X_t)$ were a martingale difference sequence. By the same argument used in the proof of Proposition 3, we simply need to show that

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

where $\widetilde{Q}_n = \frac{n^{3/4}}{h_n^{\mu 3/2}}$. Given the condition $\mathrm{E}\left(\exp\left(K\left(\frac{X_t-x}{h}\right)\alpha(X_t)\epsilon_t\right)\right) \le \Delta < \infty$, by Theorem 3.2 in Lesigné and Volny (2001), letting $\eta = C\sqrt{\frac{\log(n)}{h_n^\mu n^{1/2}}}$,

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

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

**Proof of Theorem 6.** Below, we prove Part (b). Part (a) follows by the same argument. We begin with the first moment condition:

$$\begin{aligned}
&\sum_{i=1}^n \widehat{\epsilon}_{i,h_n} w_{i,h_n^\mu}(x)\\
=\ &\sum_{i=1}^n \epsilon_i w_{i,h_n^\mu}(x) - \sum_{i=1}^n \frac{\widehat{f}_{n,h_n^\mu}(X_i) - f(X_i)}{\alpha(X_i)} w_{i,h_n^\mu}(x) + \sum_{i=1}^n \frac{\alpha(X_i) - \widehat{\alpha}_{n,h_n}(X_i)}{\widehat{\alpha}_{n,h_n}(X_i)}\epsilon_i w_{i,h_n^\mu}(x)\\
&-\sum_{i=1}^n \frac{\alpha(X_i) - \widehat{\alpha}_{n,h_n}(X_i)}{\alpha(X_i)\widehat{\alpha}_{n,h_n}(X_i)}\left(\widehat{f}_{n,h_n}(X_i) - f(X_i)\right) w_{i,h_n^\mu}(x) \qquad (23)
\end{aligned}$$

It is immediate to see that the third term is of a smaller order than the first, and the fourth term is of a smaller order than the second. Now, write

$$\sum_{i=1}^{n} \epsilon_i w_{i,h_n^\mu}(x) = \frac{\frac{1}{\gamma(n)h_n^\mu}\sum_{i=1}^{n} \epsilon_i 1\left\{|X_i - x| < h_n^\mu\right\}}{\frac{1}{\gamma(n)h_n^\mu}\sum_{i=1}^{n} 1\left\{|X_i - x| < h_n^\mu\right\}}, \tag{24}$$

where the denominator in Eq. (24) is $O_{a.s.}(1)$, and bounded away from zero, for $\gamma(n) = a(n)$, under Assumption B, by Proposition 2, for $\gamma(n) = n^{1/2}$, under Assumption C, by Proposition 4, and for $\gamma(n) = n$ in the stationary case, by the strong law of large numbers. As for the numerator in Eq. (24), recalling that $\mathrm{E}\left(\epsilon_i 1\left\{|X_i - x| < h_n^\mu\right\}\right) = o(1)$, the contribution of the bias term is negligible, and thus it is $O_p\left(\frac{1}{\sqrt{\gamma(n)h_n^\mu}}\right)$. As for the second term in Eq. (23), by either Proposition 3 or Proposition 5:

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

As shown in the proof of Proposition 1 and 4 respectively, for all $x \in \mathcal{D}_x$, $\widehat{L}_{n,h_n^\mu}(x)h_n^\mu$ is $O_{a.s.}\left(a(n)h_n^\mu\right)$, in the $\beta$-recurrent case, $O_{a.s.}\left(nh_n^\mu\right)$ when $\beta = 1$, and $O_{a.s.}\left(\sqrt{n}h_n^\mu\right)$ in the integrated case. Hence, $\sum_{i=1}^{n} \widehat{\epsilon}_{i,h_n} w_{i,h_n^\mu}(x)$ is at least of probability order $\left(h_n^{\mu 2} + \frac{1}{\sqrt{\gamma(n)h_n^\mu}}\right)$ and at most of probability order $\left(h_n^{\mu 2} + \frac{\sqrt{\log n}}{\sqrt{\gamma(n)h_n^\mu}}\right)$. We now turn to the second moment condition.

$$\begin{aligned}
& \sum_{i=1}^{n} \left(\widehat{\epsilon}_{i,h_n}^2 w_{i,h_n^\mu}(x) - 1\right) \\
= & \sum_{i=1}^{n} \left(\epsilon_{i,h_n}^2 w_{i,h_n^\mu}(x) - 1\right) - \sum_{i=1}^{n} \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\widehat{\alpha}_{n,h_n}^2(X_i)} \left(\widehat{f}^{(2)}_{n,h_n^\sigma}(X_i) - f^{(2)}(X_i)\right) \\
& + \sum_{i=1}^{n} \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\widehat{\alpha}_{n,h_n}^2(X_i)} \left(\widehat{f}_{n,h_n^\mu}(X_i)^2 - f(X_i)^2\right) \\
& + \sum_{i=1}^{n} \frac{\left(\widehat{f}_{n,h_n^\mu}(X_i) - f(X_i)\right)^2}{\widehat{\alpha}_{n,h_n}^2(X_i)} w_{i,h_n^\mu}(x) \\
& + 2\sum_{i=1}^{n} \frac{\left(\widehat{f}_{n,h_n^\mu}(X_i) - f(X_i)\right)\left(Y_i - f(X_i)\right)}{\widehat{\alpha}_{n,h_n}^2(X_i)} w_{i,h_n^\mu}(x).
\end{aligned} \tag{25}$$

It is immediate to see the fourth term in Eq. (25) 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 Eq. (25) can be neglected. Since $\mathrm{E}\Big(\epsilon^2_{i,h_n}|X_i = x\Big) - \sigma^2(x) = o(1)$, the bias component is zero and $\sum_{i=1}^n \left(\epsilon^2_{i,h_n} w_{i,h_n^\mu}(x) - 1\right) = O_p\left(\frac{1}{\sqrt{\gamma(n)h_n^\mu}}\right)$, where, again, $\gamma(n)$ differs across the various cases. As for the second term on the right-hand side of Eq. (25), because of either Proposition 3 or Proposition 5,

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

By the same argument, $\left|\sum_{i=1}^n \frac{\epsilon_i^2 w_{i,h_n^\mu}(x)}{\widehat{\alpha}^2_{n,h_n}(X_i)}\left(\widehat{f}_{n,h_n^\mu}(X_i)^2 - f(X_i)^2\right)\right|$ is majorized by a $\left(O_p\left(\sqrt{\frac{\log(n)}{\gamma(n)h_n^\mu}}\right) + O_p\left(h_n^{\mu 2}\right)\right)$ term. Thus, $\sum_{i=1}^n \widehat{\epsilon}_{i,h_n} w_{i,h_n^\mu}(x)$ is at least of probability order $\left(h_n^{\mu 2} + \frac{1}{\sqrt{\gamma(n)h_n^\mu}} + h_n^{\sigma 2} + \frac{1}{\sqrt{\gamma(n)h_n^\sigma}}\right)$ and at most of probability order $\left(h_n^{\mu 2} + \frac{\sqrt{\log n}}{\sqrt{\gamma(n)h_n^\mu}} + h_n^{\sigma 2} + \frac{\sqrt{\log n}}{\sqrt{\gamma(n)h_n^\sigma}}\right)$. The statement then follows directly from the definition of $\widetilde{h}_n(x)$.

**Proof of Theorem 7.** We prove Part (b) as earlier. Part (a) follows analogously. As for the first moment condition, by triangle inequality, we note that

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^n \widehat{\epsilon}_{i,h_n} w_{i,h}(x)\right| \\
\leq& \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^n \epsilon_{i,h_n} w_{i,h_n^\mu}(x)\right| + \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^n \frac{\widehat{f}_{n,h_n^\mu}(X_i) - f(X_i)}{\alpha(X_i)} w_{i,h_n^\mu}(x)\right| \\
&+ \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^n \frac{\alpha(X_i) - \widehat{\alpha}_{n,h_n}(X_i)}{\widehat{\alpha}_{n,h_n}(X_i)\alpha(X_i)} \epsilon_i w_{i,h_n^\mu}(x)\right| \\
&+ \sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^n \frac{\alpha(X_i) - \widehat{\alpha}_{n,h_n}(X_i)}{\alpha(X_i)\widehat{\alpha}_{n,h_n}(X_i)}\left(\widehat{f}_{n,h_n^\mu}(X_i) - f(X_i)\right) w_{h_n^\mu}(x)\right|. \qquad (26)
\end{aligned}$$

By the same argument used in either Proposition 3 or Proposition 5, $\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^n \epsilon_{i,h_n} w_{i,h_n^\mu}(x)\right| =$

$O_p\left(\frac{\sqrt{\log n}}{\sqrt{\gamma(n)h_n^{\mu}}}\right)$. Because the last two terms on the right-hand side of Eq. (26) are of smaller probability order than the first two terms, and because

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

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

$$\begin{aligned}
&\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\widehat{\epsilon}_{i,h_n}w_{i,h}(x)\right| \\
\geq\ &\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\epsilon_{i,h_n}w_{i,h_n^{\mu}}(x)-\sum_{i=1}^{n}\frac{\widehat{f}_{n,h_n^{\mu}}(X_i)-f(X_i)}{\alpha(X_i)}w_{i,h_n^{\mu}}(x)\right| \\
&-\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\frac{\alpha(X_i)-\widehat{\alpha}_{n,h_n}(X_i)}{\widehat{a}_{n,h_n}(X_i)a(X_i)}\epsilon_i w_{i,h_n^{\mu}}(x)\right| \\
&-\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\frac{\alpha(X_i)-\widehat{\alpha}_{n,h_n}(X_i)}{\alpha(X_i)\widehat{\alpha}_{n,h_n}(X_i)}\left(\widehat{f}_{n,h_n^{\mu}}(X_i)-f(X_i)\right)w_{h_n^{\mu}}(x)\right|, \qquad (27)
\end{aligned}$$

and, thus, $\sup_{x\in\mathcal{D}_x}|\sum_{i=1}^{n}\widehat{\epsilon}_{i,h_n}w_{i,h}(x)|$ is at least $O_p\left(\frac{\sqrt{\log n}}{\sqrt{\gamma(n)h_n^{\mu}}}\right)+O_p\left(h_n^{\mu 2}\right)$. By the same argument used in the proof of Theorem 6, it is immediate to see that $\sup_{x\in\mathcal{D}_x}\left|\sum_{i=1}^{n}\left(\widehat{\epsilon}_{i,h_n}^{2}w_{i,h_n^{\mu}}(x)-1\right)\right|$ is (at most and at least) $O_p\left(h_n^{\mu 2}+\frac{\sqrt{\log n}}{\sqrt{\gamma(n)h_n^{\mu}}}+h_n^{\sigma 2}+\frac{\sqrt{\log n}}{\sqrt{\gamma(n)h_n^{\sigma}}}\right)$. The statement then follows from the definition of $\widetilde{h}_n$ in Eq. (13).

**Proof of Theorem 8.** The methods used to prove Theorem 3 in Bandi, Corradi, and Moloche (2009) yield the result.

## 9 References

BANDI, F.M. (2004). On Persistence and Nonparametric Estimation (With an Application to Stock Return Predictability). *Working paper.*

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

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

BANDI, F.M., and G. MOLOCHE (2004). On the Functional Estimation of Multivariate Diffusion Processes. *Working paper.*

BANDI, F.M., V. CORRADI and G. MOLOCHE (2004). Bandwidth Selection for Continuous Time Markov Processes. *Working paper.*

CHEN, X. (1999). How often does a Harris recurrent Markov Chain recur? *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 Q. YAO (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer. Berlin.

GAO, J. and I. GIJBELS (2008). Bandwidth Selection in Nonparametric Kernel Testing. *Journal of the American Statistical Association* 103, 1584-1594.

GAO, J., M. KING, Z. LU and D. TJOSTHEIM (2009). Nonparametric Specification Testing for Nonlinear Time Series With Nonstationarity. *Econometric Theory* 25, 1869-1892.

GAO, J., D. LI and D. TJOSTHEIM (2009). Uniform Consistency for Nonparametric Estimators in Null Recurrent Time Series. *Working Paper.*

GUERRE, E. (2004). Design-Adaptive Pointwise Nonparametric Regression Estimation for Recurrent Markov Time Series. *Working Paper.*

HALL, P. and J.L. HOROWITZ (2005). Nonparametric Methods for Inference in the Presence of Instrumental Variables. *Annals of Statistics* 33, 2904-2929.

HANSEN, B.E. (2004). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory* 24, 726-748.

JEGANNATHAN, P. (2004). Convergence of Functionals of Sums of r.v.s to Local Times of Fractional Stable Motions, *Annals of Probability* 32, 17717-1795.

KARLSEN, H.A. and V. TJOSTHEIM (2001). Nonparametric Estimation in Null Recurrent Time Series. *Annals of Statistics* 29, 372-416.

KARLSEN, H.A., T. MYKLEBUST, and V. TJOSTHEIM (2007), Nonparametric Estimation in Nonlinear Cointegration Type Model. *Annals of Statistics* 29, 372-416.

LEPSKI, O.V. (1990). Asymptotically Minimax Adaptive Estimation I: Upper Bounds, Optimally Adaptive Estimates. *Theory of Probability and Applications* 36, 682-697.

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

LEPSKI, O.V., E. MAMMEN and V.G. SPOKOINY (1997). Ideal Spatial Adaptation to In-homogeneous Smoothness: an Approach Based on Kernel Estimates With Variable Bandwidth Selection. *Annals of Statistics* 25, 929-947.

LESIGNÈ, E. and D. VOLNÝ (2001). Large Deviations for Martingales. *Stochastic Processes and Their Applications* 96, 143-159.

MOLOCHE, G. (2001). Kernel Regression for Nonstationary Harris-Recurrent Processes. *Working paper.*

PHILLIPS, P.C.B. (2009). Local Limit Theory and Spurious Nonparametric Regression. *Econometric Theory* 25, 1466-1497.

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. *Working Paper.*

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

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

|  |  | MODEL I |  |  |  |
|---|---|---|---|---|---|
|  |  | optimal |  | bias correction |  |
|  |  | bandwidth | SD | bandwidth | SD |
| pointwise | $h^{\mu}$ | 0.5817 | 0.3585 | 0.3253 | 0.1793 |
|  |  | 0.6396 | 0.3063 | 0.3624 | 0.1393 |
|  |  | 0.6216 | 0.3198 | 0.3617 | 0.1571 |
|  |  | 0.6510 | 0.3191 | 0.3666 | 0.1632 |
|  |  | 0.5736 | 0.3597 | 0.3210 | 0.1755 |
|  | $h^{\sigma}$ | 0.6008 | 0.3541 | 0.3323 | 0.1752 |
|  |  | 0.6654 | 0.3051 | 0.3689 | 0.1370 |
|  |  | 0.5622 | 0.2974 | 0.3451 | 0.1276 |
|  |  | 0.6728 | 0.3111 | 0.3690 | 0.1503 |
|  |  | 0.5916 | 0.3547 | 0.3349 | 0.1850 |
| uniform | $h^{\mu}$ | 0.6971 | 0.3167 | 0.3705 | 0.1604 |
|  | $h^{\sigma}$ | 0.6495 | 0.3541 | 0.3463 | 0.1782 |
| CV | $h^{\mu}$ | 0.7634 | 0.4197 | 0.3333 | 0.2199 |
|  | $h^{\sigma}$ | 0.7582 | 0.4186 | 0.3295 | 0.2131 |

Table 1: Selected bandwidths and their standard deviation (“SD”).

|  |  | MODEL II |  |  |  |
|---|---|---|---|---|---|
|  |  | optimal |  | bias correction |  |
|  |  | bandwidth | SD | bandwidth | SD |
| pointwise | $h^\mu$ | 0.4236 | 0.4429 | 0.2115 | 0.2182 |
|  |  | 0.4399 | 0.2109 | 0.3349 | 0.1186 |
|  |  | 0.6159 | 0.2429 | 0.3849 | 0.1073 |
|  |  | 0.7008 | 0.2626 | 0.3911 | 0.1206 |
|  |  | 0.7208 | 0.2609 | 0.3997 | 0.1314 |
|  | $h^\sigma$ | 0.3956 | 0.2849 | 0.2849 | 0.1340 |
|  |  | 0.4372 | 0.2304 | 0.3295 | 0.1214 |
|  |  | 0.6010 | 0.2499 | 0.3778 | 0.0988 |
|  |  | 0.6976 | 0.2587 | 0.3915 | 0.1005 |
|  |  | 0.7334 | 0.2542 | 0.3970 | 0.1039 |
| uniform | $h^\mu$ | 0.7567 | 0.2931 | 0.3933 | 0.1449 |
|  | $h^\sigma$ | 0.5565 | 0.2344 | 0.3695 | 0.0958 |
| CV | $h^\mu$ | 0.3367 | 0.4592 | 0.1494 | 0.2068 |
|  | $h^\sigma$ | 0.3360 | 0.4584 | 0.1487 | 0.2057 |

Table 2: Selected bandwidths and their standard deviation ("SD").

| | | MODEL III ($\theta = 0$) | | | | MODEL III ($\theta = 2$) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | optimal | | bias correction | | optimal | | bias correction | |
| | | bandwidth | SD | bandwidth | SD | bandwidth | SD | bandwidth | SD |
| pointwise | $h^f$ | 0.5612 | 0.3124 | 0.3468 | 0.1419 | 0.5372 | 0.3198 | 0.3323 | 0.1383 |
| | | 0.5345 | 0.2678 | 0.3619 | 0.1219 | 0.5218 | 0.2651 | 0.3602 | 0.1310 |
| | | 0.5072 | 0.2812 | 0.3463 | 0.1176 | 0.5236 | 0.2817 | 0.3493 | 0.1129 |
| | | 0.5637 | 0.3149 | 0.3477 | 0.1198 | 0.5701 | 0.3134 | 0.3520 | 0.1216 |
| | | 0.6050 | 0.3000 | 0.3628 | 0.1294 | 0.6161 | 0.3087 | 0.3631 | 0.1364 |
| | $h^a$ | 0.5815 | 0.3183 | 0.3516 | 0.1394 | 0.5491 | 0.3233 | 0.3394 | 0.1452 |
| | | 0.5939 | 0.2931 | 0.3574 | 0.0956 | 0.6030 | 0.2921 | 0.3616 | 0.1004 |
| | | 0.5602 | 0.2793 | 0.3617 | 0.1216 | 0.5580 | 0.2781 | 0.3637 | 0.1242 |
| | | 0.6189 | 0.2948 | 0.3701 | 0.1155 | 0.5981 | 0.2933 | 0.3614 | 0.1049 |
| | | 0.6929 | 0.3167 | 0.3689 | 0.1289 | 0.6842 | 0.3127 | 0.3736 | 0.1369 |
| uniform | $h^f$ | 0.6292 | 0.3033 | 0.3731 | 0.1419 | 0.6280 | 0.3137 | 0.3658 | 0.1448 |
| | $h^a$ | 0.7389 | 0.3200 | 0.3790 | 0.1693 | 0.7218 | 0.3293 | 0.3703 | 0.1687 |
| CV | $h^f$ | 0.5622 | 0.3533 | 0.3165 | 0.1897 | 0.5637 | 0.3537 | 0.3168 | 0.1901 |
| | $h^a$ | 0.7493 | 0.4218 | 0.3293 | 0.2176 | 0.7409 | 0.4207 | 0.3293 | 0.2178 |

Table 3: Selected bandwidths and their standard deviation (“SD”).

|  |  | MODEL I |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | optimal |  |  | bias correction |  |  |
|  |  | bias | SD | RMSE | bias | SD | RMSE |
| pointwise | $\mu(x)$ | -0.0216 | 0.5101 | 0.5110 | -0.0166 | 0.5575 | 0.5577 |
|  | $\mu^{(2)}(x)$ | 0.1227 | 2.7806 | 2.7886 | 0.1186 | 3.1204 | 3.1251 |
| uniform | $\mu(x)$ | -0.0096 | 0.5126 | 0.5141 | -0.0091 | 0.5650 | 0.5652 |
|  | $\mu^{(2)}(x)$ | -0.0249 | 2.6401 | 2.6446 | -0.0020 | 3.0054 | 3.0065 |
| CV | $\mu(x)$ | -0.0108 | 0.5167 | 0.5202 | -0.0136 | 0.5651 | 0.5658 |
|  | $\mu^{(2)}(x)$ | -0.0956 | 2.7128 | 2.7204 | -0.0341 | 3.0783 | 3.0801 |

Table 4: Average bias, standard deviation ("SD") and root mean square error ("RMSE") of the respective estimated functions.

|  |  | MODEL II |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | optimal |  |  | bias correction |  |  |
|  |  | bias | SD | RMSE | bias | SD | RMSE |
| pointwise | $\mu(x)$ | -0.0302 | 0.3498 | 0.3518 | -0.0151 | 0.3993 | 0.3996 |
|  | $\mu^{(2)}(x)$ | -0.1035 | 1.5338 | 1.5433 | -0.0391 | 1.8316 | 1.8349 |
| uniform | $\mu(x)$ | -0.0423 | 0.2775 | 0.2838 | -0.0150 | 0.3386 | 0.3394 |
|  | $\mu^{(2)}(x)$ | -0.0756 | 1.5944 | 1.6054 | -0.0278 | 1.8256 | 1.8307 |
| CV | $\mu(x)$ | -0.0247 | 0.8542 | 0.8547 | -0.0053 | 0.8725 | 0.8724 |
|  | $\mu^{(2)}(x)$ | -0.0703 | 4.2487 | 4.2506 | 0.0107 | 4.3597 | 4.3593 |

Table 5: Average bias, standard deviation ("SD") and root mean square error ("RMSE") of the respective estimated functions.

|  |  | MODEL III ($\theta = 0$) |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | optimal |  |  | bias correction |  |  |
|  |  | bias | SD | RMSE | bias | SD | RMSE |
| pointwise | $f(x)$ | -0.1665 | 0.4886 | 0.5269 | -0.0790 | 0.5120 | 0.5226 |
|  | $f^{(2)}(x)$ | 0.0022 | 0.7907 | 0.8241 | -0.0036 | 0.9280 | 0.9420 |
| uniform | $f(x)$ | -0.2027 | 0.4420 | 0.5028 | -0.0887 | 0.4938 | 0.5077 |
|  | $f^{(2)}(x)$ | -0.0193 | 0.6776 | 0.7407 | -0.0070 | 0.8837 | 0.9022 |
| CV | $f(x)$ | -0.2072 | 0.4830 | 0.5445 | -0.0787 | 0.5245 | 0.5357 |
|  | $f^{(2)}(x)$ | -0.0068 | 0.8164 | 0.8592 | 0.0100 | 0.9862 | 1.0009 |

|  |  | MODEL III ($\theta = 2$) |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | optimal |  |  | bias correction |  |  |
|  |  | bias | SD | RMSE | bias | SD | RMSE |
| pointwise | $f(x)$ | -0.1321 | 0.4927 | 0.5182 | -0.0419 | 0.5243 | 0.5286 |
|  | $f^{(2)}(x)$ | 0.0582 | 0.8864 | 0.9109 | 0.0753 | 1.0145 | 1.0268 |
| uniform | $f(x)$ | -0.1818 | 0.4573 | 0.5062 | -0.0539 | 0.5139 | 0.5202 |
|  | $f^{(2)}(x)$ | 0.0347 | 0.7721 | 0.8226 | 0.0682 | 0.9620 | 0.9774 |
| CV | $f(x)$ | -0.1845 | 0.4804 | 0.5292 | -0.0523 | 0.5356 | 0.5411 |
|  | $f^{(2)}(x)$ | 0.0382 | 0.9077 | 0.9427 | 0.0778 | 1.0869 | 1.1009 |

Table 6: Average bias, standard deviation ("SD") and root mean square error ("RMSE") of the respective estimated functions.

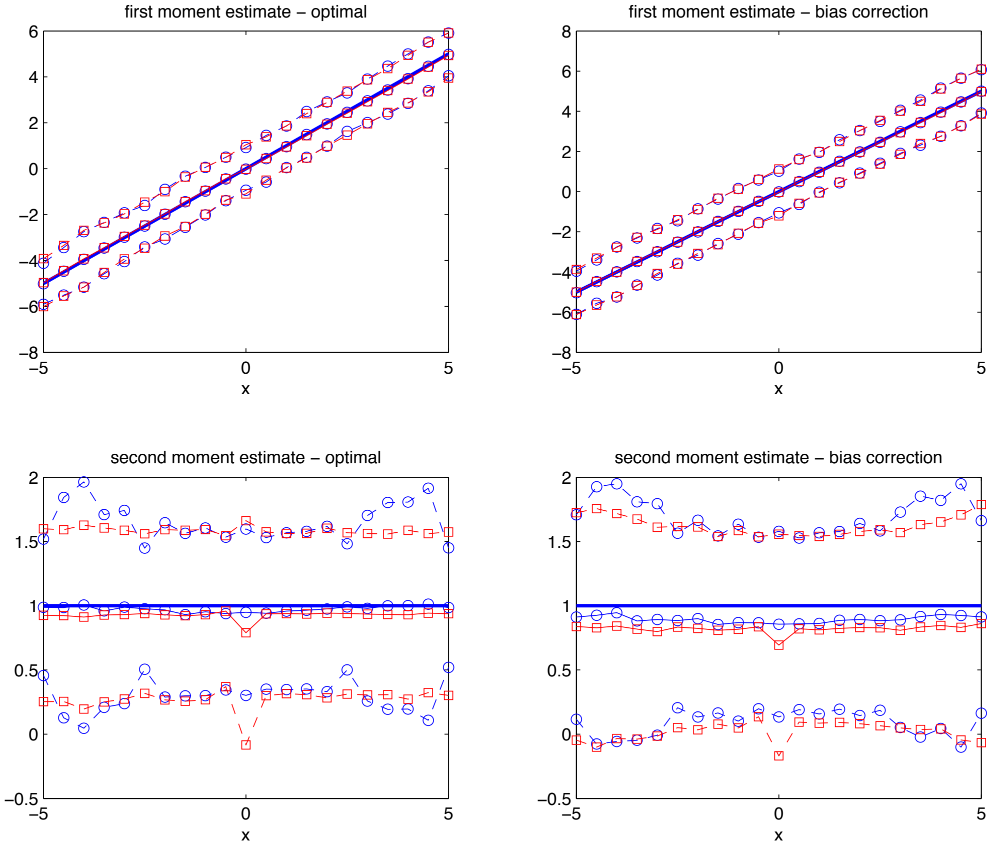

Figure 1: Model I, estimated moments based on uniform criterion (blue circles), CV (red squares) and the true moments (thick blue lines).

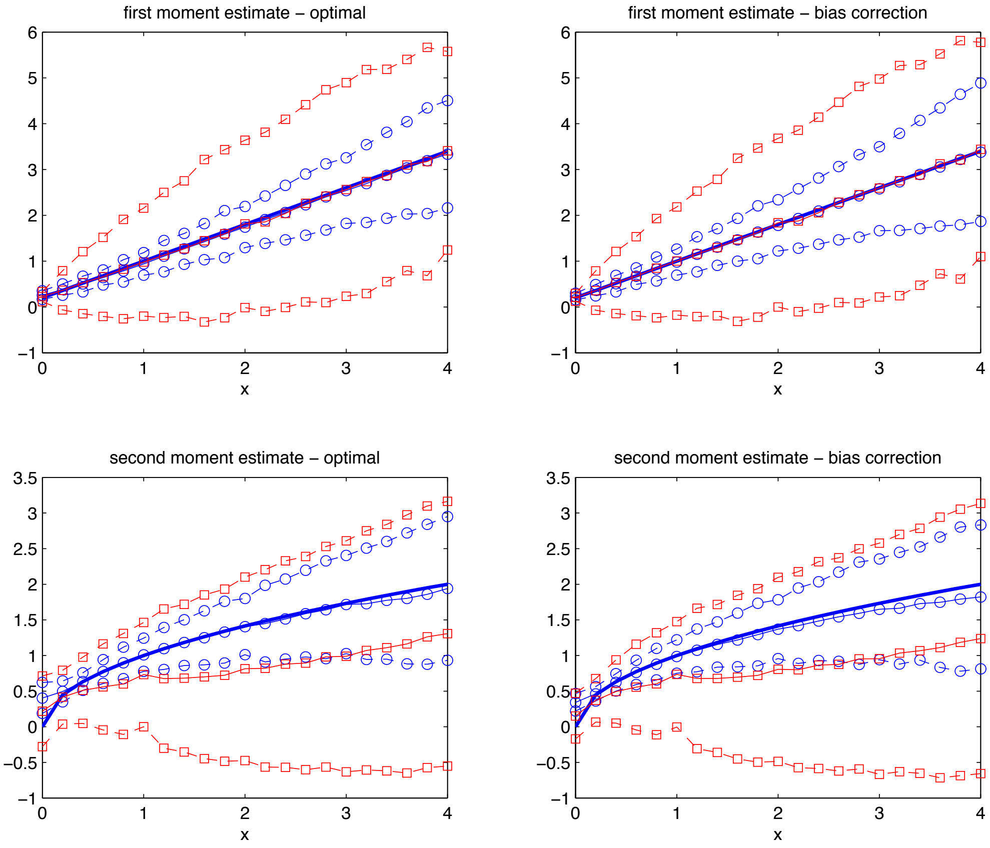

Figure 2: Model II, estimated moments based on uniform criterion (blue circles), CV (red squares) and the true moments (thick blue lines).

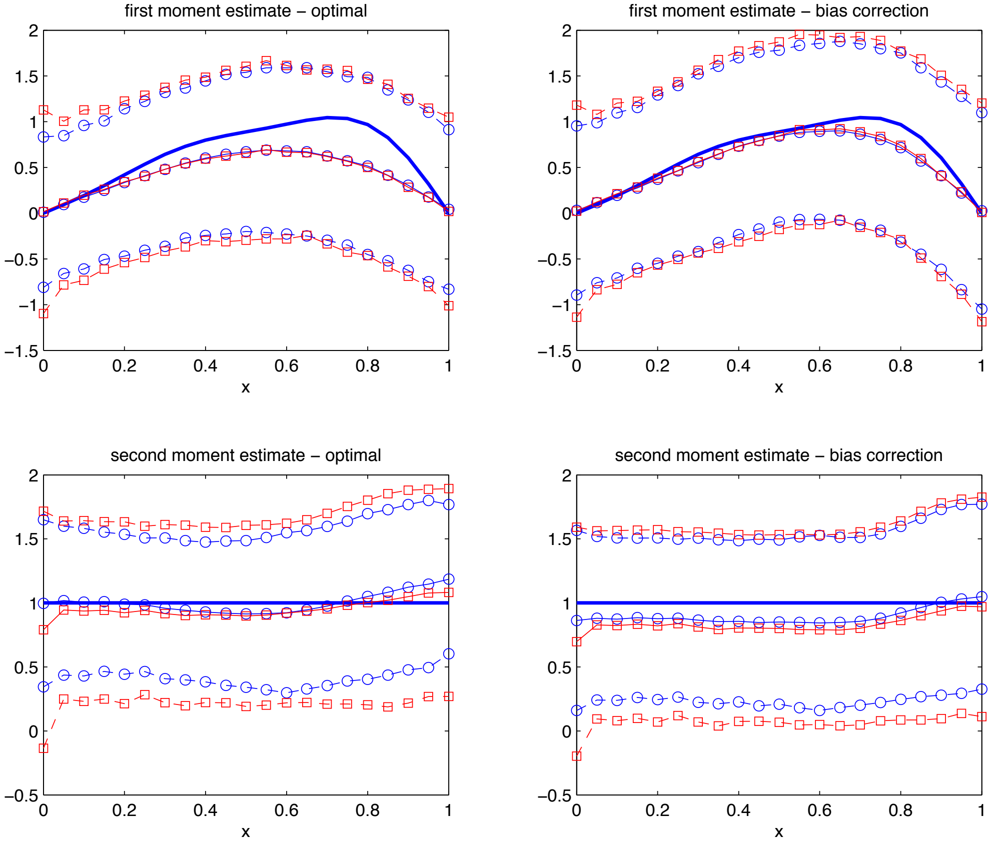

Figure 3: Model III ($\theta = 0$), estimated moments based on uniform criterion (blue circles), CV (red squares) and the true moments (thick blue lines).

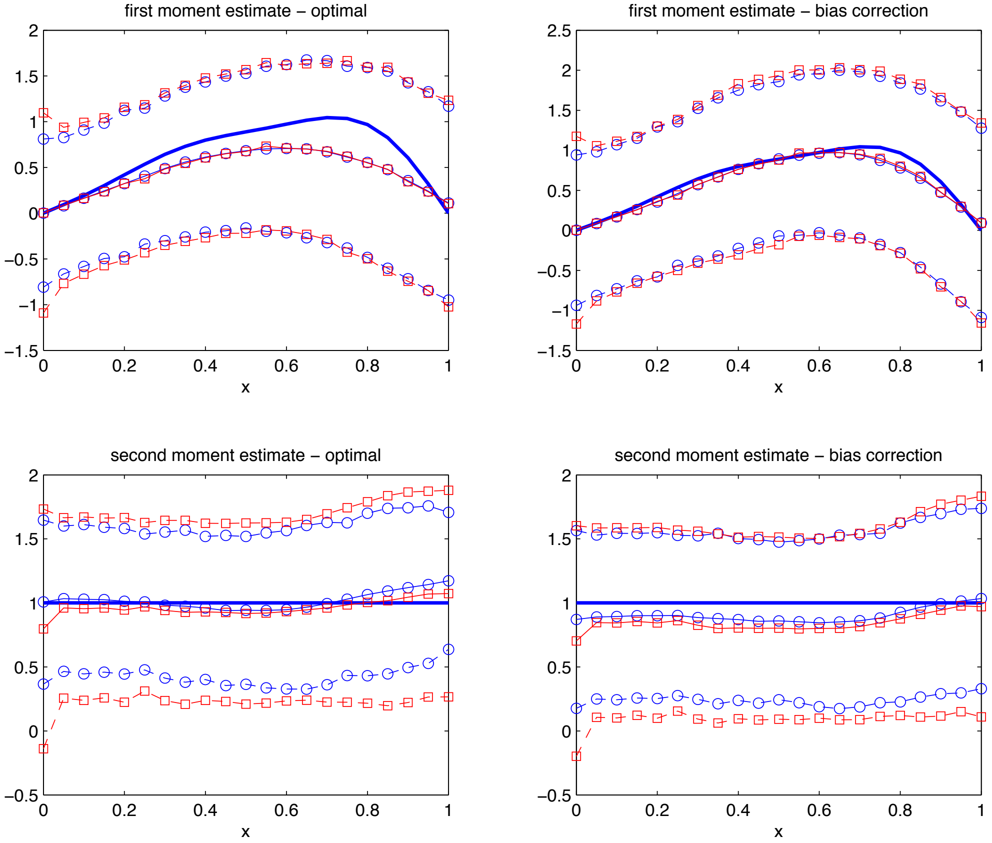

Figure 4: Model III ($\theta = 2$), estimated moments based on uniform criterion (blue circles), CV (red squares) and the true moments (thick blue lines).

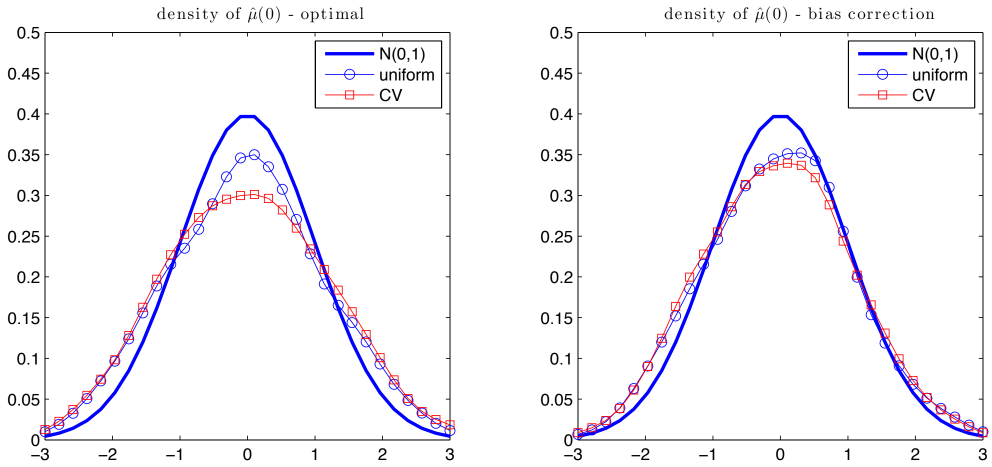

Figure 5: Model I, distribution of the first moment estimator at $x = 0$, based on uniform bandwidths.

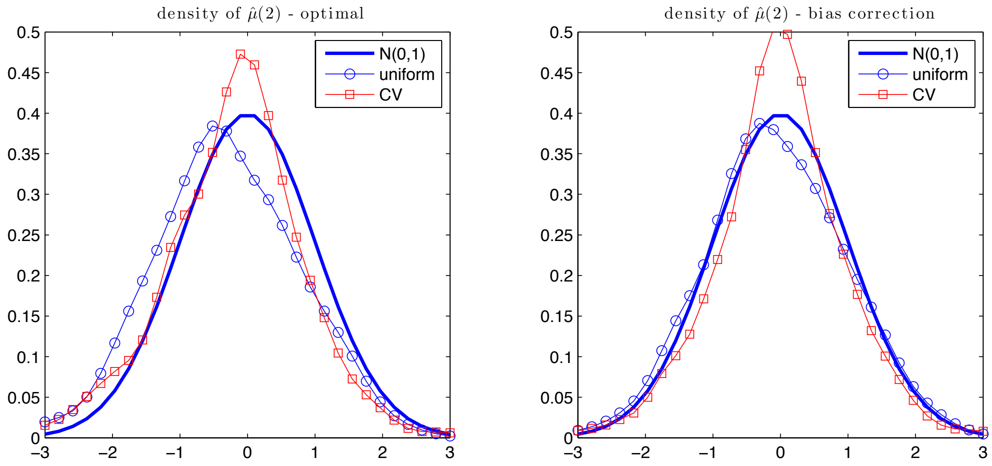

Figure 6: Model II, distribution of the first moment estimator at $x = 2$, based on uniform bandwidths.

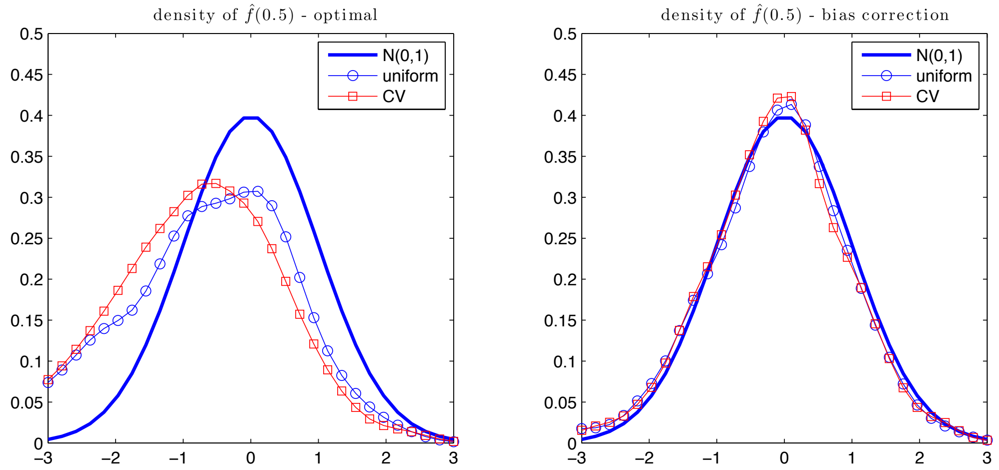

Figure 7: Model III ($\theta = 0$), distribution of the first moment estimator at $x = 0.5$, based on uniform bandwidths.

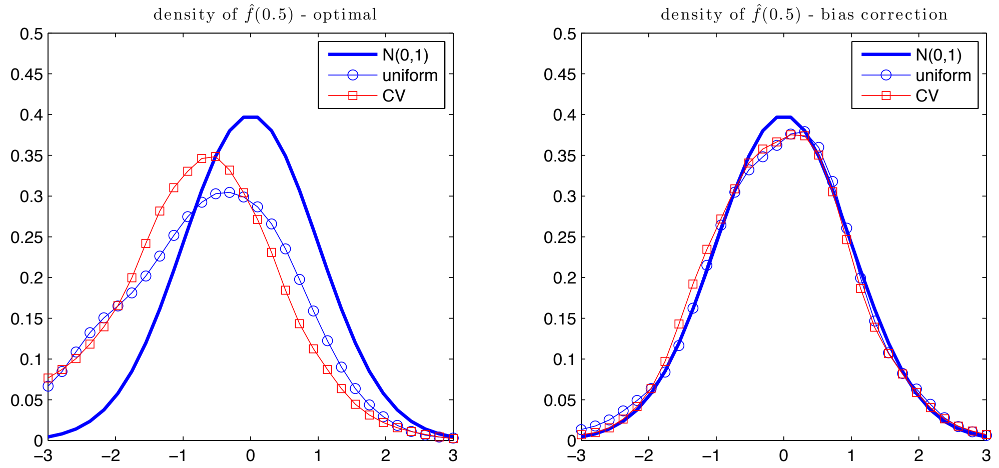

Figure 8: Model III ($\theta = 2$), distribution of the first moment estimator at $x = 0.5$, based on uniform bandwidths.