cemmap

centre for microdata methods and practice

Possibly Nonstationary Cross-Validation

Federico M. Bandi
Valentina Corradi
Daniel Wilhelm

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP11/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Possibly Nonstationary Cross-Validation*

Federico M. Bandi † Valentina Corradi‡ Daniel Wilhelm§

March 12, 2016

Abstract

Cross-validation is the most common data-driven procedure for choosing smoothing parameters in nonparametric regression. For the case of kernel estimators with iid or strong mixing data, it is well-known that the bandwidth chosen by cross-validation is optimal with respect to the average squared error and other performance measures. In this paper, we show that the cross-validated bandwidth continues to be optimal with respect to the average squared error even when the data-generating process is a β-recurrent Markov chain. This general class of processes covers stationary as well as nonstationary Markov chains. Hence, the proposed procedure adapts to the degree of recurrence, thereby freeing the researcher from the need to assume stationary (or nonstationary) before inference begins. We study finite sample performance in a Monte Carlo study. We conclude by demonstrating the practical usefulness of cross-validation in a highly-persistent environment, namely that of nonlinear predictive systems for market returns.

Keywords: Bandwidth Selection, Recurrence, Predictive Regressions.

*We thank Karim Abadir, Peter Phillips, Mervyn Silvapulle, Liangjun Su, Abderahim Taamouti and Julian Williams for comments and discussions. We are also grateful to seminar participants at the 2014 Conference on Financial Econometrics (Brunel University), the 2014 2nd Symposium in Nonparametric Statistics (Cadiz), the 2014 Conference "Nonparametric and Semiparametric Methods" (Cambridge University), the University of Durham, the University of Southampton and Singapore Management University.

†Carey Business School, Johns Hopkins University, and Edhec-Risk

‡School of Economics, University of Surrey

§Department of Economics, University College London, and CeMMAP. The author gratefully acknowledges financial support from the ESRC Centre for Microdata Methods and Practice at IFS (RES-589-28-0001).

1 Introduction

The vast literature on unit roots and cointegration has largely focused on linear models. While it is well-known that the limiting behavior of partial sums, and affine functionals thereof, can be approximated by Gaussian processes, much less is known about the asymptotic behaviour of functional estimators of nonstationary time series. Nonparametric regression with nonstationary discrete-time processes has, in fact, been receiving attention only in recent years.

Several financial time-series central to asset pricing, like the short-term rate, the dividend-to-price ratio or, more generally, any financial ratio whose denominator depends on the market's price level, are highly persistent and often depend on their past in a non-linear fashion. The joint presence of *possibly* nonstationary behaviour and nonlinearities of unknown form provides econometric content to nonparametric regressions with *possibly* nonstationary time series. The word "possibly" is an important qualifier. While strict nonstationarity is, in many cases, economically unpalatable, allowing for stationary as well as for nonstationary behaviour is a satisfactory (theoretical and empirical) way to accommodate a broad range of levels, especially including very high levels, of persistence in one unified framework.

Coherently with this observation, the literature on nonparametric autoregression has focused on β-recurrent (stationary for $\beta = 1$ and nonstationary for $\beta < 1$) Markov chains. In this framework, the number of regenerations of a recurrent chain has heavily been used to derive the limiting behavior of the number of visits around a given point and related estimators (see, e.g., Karlsen and Tjøstheim (2001), Moloche (2001), and Gao et al. (2014)). Schienle (2010) considers the case of many regressors. Guerre (2004) derives convergence rates for a general class of chains.

This succesful literature has established consistency and asymptotic mixed normality for local constant and local linear kernel estimators of nonstationary autoregressions and nonstationary cointegration. However, it has been largely silent about practical guidance on bandwidth selection.

The two most common approaches for bandwidth selection in functional problems are "plug-in" methods and cross-validation. In the β-recurrent case, the bandwidth conditions for consistency and asymptotic mixed normality depend on the generally unknown parameter β. In light of the slow (i.e., logarithmic) convergence rate of existing estimates of β (Karlsen and Tjøstheim (2001)), plug-in methods appear even more ad-hoc than in the more classical stationary environment. On the other hand, bandwidths chosen by cross-validation have been used broadly in empirical work. Nonetheless, their properties have yet to be established in a (possibly) nonstationary context. This is the subject of the current paper.

For the case of identically distributed and independent observations, it has been shown that "leave-one-out" cross-validated bandwidths are optimal with respect to the Mean Integrated Squared Error (MISE), see, e.g., Härdle and Marron (1985) and Härdle (1986). More explicitly, the ratio of the cross-validated bandwidth and the infeasible bandwidth that minimizes the estimator's MISE converges in probability to one. Optimality with respect to MISE has also been derived in the case of strong mixing observations (Härdle and Vieu (1992), Kim and Cox (1996), and Xia and Li (2002)).

This paper shows that, for β-recurrent Markov chains, the bandwidth chosen via "leave-one-out" cross-validation is asymptotically optimal with respect to the Average Squared Error (ASE). In other words, the ratio of the cross-validated bandwidth and the infeasible bandwidth that minimizes the ASE converges in probability to one. Differently from stationary environments, the asymptotic equivalence between the *random* cross-validated bandwidth sequence and a *deterministic* sequence minimizing the estimator's MISE does not hold. This is easily explained. In the β-recurrent case the effective sample size is random, the asymptotic variance of conditional mean estimates is also random, so is the MISE. Of course, should $\beta = 1$, the ASE would converge uniformly to a deterministic limit, i.e., the MISE, and our findings would again deliver equivalence between an adaptive random sequence and a deterministic sequence, optimal with respect to the MISE, as a subcase of a more general result.

After studying the case of bandwidth selection both for the conditional mean function and for the conditional variance function, we apply the methods to a (mean-variance) portfolio allocation problem relying on (possibly) nonlinear stock-return predictability. Even though the nonstationary behaviour of typical predictors, like the dividend-to-price ratio, may be economically unpalatable, nonstationarity tests find it extremely hard to reject the null hypothesis of nonstationarity. The very high persistence of these predictors is a general feature of many macroeconomic and finance time series. It justifies our emphasis on bandwidth selection procedures capable of handling such a persistence, *without* requiring the researcher to assume either stationarity or nonstationarity before inference begins.

The remainder of the paper is organized as follows. Section 2 defines the framework. Section 3 reports our main result establishing optimality of cross-validation with respect to the ASE. Here, we focus on the conditional mean function. Section 4 turns to the conditional variance function. Section 5 reports the findings of a Monte Carlo study in which we analyze the variance and bias of nonparametric kernel estimators based on cross-validated bandwidths. In Section 6 we discuss an empirical illustration in which cross-validated bandwidths are used in predictive systems for market returns and a nonlinear (mean-variance) portfolio allocation problem. All proofs are gathered in the appendix.

2 The Model

Intuitively, one can estimate conditional moments, evaluated at a given point, only if this specific point is visited infinitely often as the sample size grows. 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, will be visited an infinite number of times.

For positive recurrent (ergodic) 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 (nonstationary) chains, the expected time between two consecutive visits is infinite. Therefore, the time spent in the neighborhood of a point grows at a rate, generally unknown, which is slower than n.

Since, up to some mild regularity conditions, positive recurrent chains are strongly mixing, consistency and asymptotic normality of the conditional moment estimates follow by, e.g., the work of Robinson (1983). In this case, bandwidth selection may be implemented by virtue of cross-validation, whose optimality properties have been thoroughly established (Härdle and Vieu (1992) and Kim and Cox (1996), for kernel regressions, and Xia and Li (2002), for local linear estimators).

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 influential contribution, Karlsen and Tjøstheim (2001) derive consistency and mixed asymptotic normality for conditional moment estimators in the case of recurrent Markov chains.

Let $\{X_t,\ t \geq 0\}$ be a Markov chain and define $\mu(X_{t-1}) = \mathrm{E}\left(X_t|X_{t-1}\right)$ and $\sigma^2(X_{t-1}) = \operatorname{var}\left(X_t|X_{t-1}\right)$, so that X_t can be written as

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

where u_t has conditional mean zero and conditional variance one. Eq. (1) allows for nonlinearities of unknown form in both the conditional mean and the conditional variance. When sampled over small time distances, given iid Gaussian shocks, the resulting time series can also be interpreted as a "discretized" diffusion process. Gaussianity is never assumed in this paper.

Consider the Nadaraya-Watson kernel estimator of $\mu(x)$:

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

where K is a kernel function and h_n a bandwidth parameter. The limiting properties of $\widehat{\mu}_{h_n}(x)$ have been established by Karlsen and Tjøstheim (2001, Theorem 5.4) under Assumption 1 below, which largely corresponds to their Assumptions B_0-B_4.

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

2. *For each $y \in \mathbb{R}$, there exists a transition density $\tilde{p}$ so that $P(y, dx) = \tilde{p}(y, x)dx$. Also, for each y, there exist (small enough) constants γ and $\delta \in (0, 1)$ independent of y so that $\int 1_{\{x\in\mathbb{R}:\tilde{p}(y,x)\geq\gamma\}}dx \geq \delta$.*

3. *The invariant measure π_s has a twice continuously differentiable density p_s, which is strictly positive and bounded on every small set[1].*

4. *The kernel function K is a bounded density with compact support satisfying $\int uK(u)du = 0$ and $K_2 = \int K^2(u)du < \infty$. The set $\mathcal{N}_x = \{y : K(y - x) \neq 0\}$ is a small set for all x.*

[1]In the β-recurrent case, under rather general regularity conditions, compact sets are small sets (see, e.g., Feigin and Tweedie (1985)). A set A is small, if there exists a positive measure λ, positive constant b and an integer $m \geq 1$, such that $P^m \geq 1\{A\}b \oplus \lambda$, where P is the measure governing the chain. For a detailed description of small sets and related properties, see Nummelin (1984) or Section 3 in Karlsen and Tjøstheim (2001).

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

6. *The functions* $\mu(x)$ *and* $\sigma^2(x)$ *are twice continuously differentiable.*

Let Assumption 1 hold. For $x\in\mathcal{C}$, where $\mathcal{C}\subseteq\mathbf{E}$ is a compact set, and if $h_n n^{\beta-\varepsilon}\to\infty$ for some $\varepsilon>0$, then

$$\left(h_n\sum_{i=2}^{n}K\left(\frac{X_{i-1}-x}{h_n}\right)\right)^{1/2}\left(\widehat{\mu}_{h_n}(x)-\mu(x)-b_{h_n}(x)\right)\xrightarrow{d}N\left(0,\sigma^2(x)k_2\right), \quad (3)$$

where $k_2=\int K(u)^2du$ and $b_{h_n}(x)$ denotes a bias term that converges to zero under the additional condition $h_n n^{\beta/5+\varepsilon}\to 0$ (Theorem 5.4 in Karlsen and Tjøstheim (2001)).

The bandwidth rate conditions (and, implicitly, the estimator's convergence rate) depend on the generally unknown degree of recurrence β. Although β can be estimated, existing estimators only converge at a logarithmic rate and, thus, may not be overly useful in practice (Remark 3.7 in Karlsen and Tjøstheim (2001)).

In essence, one remains with the problem of adaptively choosing the bandwidth h_n. Because β is unknown, in general, and cannot be estimated reliably, "plug-in" methods cannot be implemented effectively. Cross-validation appears as a viable alternative, one which is broadly employed in empirical work. In what follows, we set the stage for studying the properties of cross-validated bandwidths in a (possibly) nonstationary environment, i.e., for $\beta\le 1$.

Before doing so, we emphasize that our interest is not in the nonstationary ($\beta<1$) case per se. We are, instead, interested in being robust to deviations from stationarity. We are also interested in allowing for potentially high levels of persistence, provided the process is not transient. From an empirical standpoint, being agnostic about the stationarity properties of the data is an important property. From a theoretical standpoint, after excluding transience of the process, we operate under assumptions that are virtually minimal to obtain consistent functional estimates (see, e.g., Yakowitz (1989)[2]). Bandi (2002) employs a similar approach in continuous time.

Asymptotic mixed normality for β-recurrent chains is shown via split chains, i.e., by splitting the chain into identically and independently distributed components (Nummelin

[2]In his 1989 paper, Yakowitz writes: " ... in the Markov case, the mixing assumptions are not essential ... Even in the absence of a stationary distribution, under conditions general enough to include unbounded random walks and ARMA processes, [nonlinear] regression estimation is possible. We require only stationarity of the transition law, not of the process."

(1984), Chapter 4). We will use split chains in what follows. Let Assumption 1 hold. Write the denominator in Eq. (2) as

$$\frac{1}{h_n}\sum_{i=2}^{n} K\left(\frac{X_{i-1}-x}{h_n}\right) = U_{0,x,h_n} + \sum_{k=1}^{T_n} U_{k,x,h_n} + U_{n,x,h_n}.$$

We abbreviate $K_{x,h}(y) = K((y-x)/h)/h$ and $U_{k,x,h} = U_k(K_{x,h})$, where, for any $f \in \mathcal{F} = \{f = hK_{x,h} : x \in \mathcal{C}, 0 < h \leq 1\}$,

$$U_k(f) = \begin{cases} \sum_{j=1}^{\tau_0} f(X_{j-1}), & k = 0 \\ \sum_{j=\tau_{k-1}+1}^{\tau_k} f(X_{j-1}), & 1 \leq k < n \\ \sum_{j=\tau_{T_n}+1}^{n} f(X_{j-1}), & k = n. \end{cases} \tag{4}$$

The random times τ_k, $k = 0, \ldots, n$, are the regeneration times of the Markov chain and T_n denotes the number of complete regenerations from time 0 to time n. T_n is a random quantity playing the analogous role of the sample size in stationary problems. For any h and x, the random variables $U_{k,x,h}$, $k \geq 1$, can be shown to be independent and identically distributed and, therefore, play a key role in the asymptotic theory for recurrent Markov processes. Notice however, that, in general, $U_{k,x,h}$ is not independent of the "effective" sample size T_n.

For a compact set $\mathcal{C}$, define

$$T_n(\mathcal{C}) = \sum_{i=1}^{n} 1\{X_i \in \mathcal{C}\}, \tag{5}$$

which represents the random number of times the Markov process visits the set $\mathcal{C}$. For some constants $0 < \underline{c}, \overline{c} < \infty$, given $\frac{1}{5} < \overline{\eta} < \underline{\eta} < 1$ and $1 - 2\underline{\eta} + \overline{\eta} > 0$, let

$$H_n = \left\{h \in \mathbb{R} : \underline{c}\, T_n(\mathcal{C})^{-\underline{\eta}} \leq h \leq \overline{c}\, T_n(\mathcal{C})^{-\overline{\eta}}\right\} \tag{6}$$

be a set of bandwidths.

The set H_n is the feasible set over which the cross-validated bandwidth will be chosen. This set is random. This is in contrast to the independent and the strong mixing case in which it is a deterministic set that only depends on the sample size n. In both of these cases, the time between two consecutive visits to any compact set is finite, and so the number of visits to $\mathcal{C}$, grows at the same rate as the sample size n. In the general β-recurrent case, the "effective" sample size, given by the number of regenerations T_n, is not observed. However, $T_n(\mathcal{C})$ is observed and, by Remark 3.5 in Karlsen and Tjøstheim

(2001), of the same almost-sure order as T_n (provided in Lemma 3.4 of Karlsen and Tjøstheim (2001)). Therefore, we define the feasible bandwidth set in terms of $T_n(\mathcal{C})$. This definition ensures that any sequence of bandwidths in H_n satisfies the rate restrictions (reported above and below Eq. (3)) sufficient for consistency, asymptotic normality, and zero asymptotic bias of nonparametric conditional mean (and variance) estimators. We note that the definition of H_n is analogous to the bandwidth set in Härdle and Marron (1985) for the iid case when $T_n(\mathcal{C})$ is of order n.[3]

We define the cross-validated bandwidth as

$$\widehat{h}_n = \arg\min_{h \in H_n} CV_n(h), \tag{7}$$

where

$$\mathrm{CV}_n(h) = \frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(X_j - \widehat{\mu}_{j,h}(X_{j-1})\right)^2 1\left\{X_{j-1} \in \mathcal{C}\right\} \tag{8}$$

is the cross-validation (CV) criterion, $1\{A\}$ is the indicator function that equals one if A is true and zero otherwise, and

$$\widehat{\mu}_{j,h}(x) = \frac{\sum_{i=2, i\neq j}^{n} X_i K_h\left(X_{i-1} - x\right)}{\sum_{i=2, i\neq j}^{n} K_h\left(X_{i-1} - x\right)} \qquad \text{for } j = 2, \ldots, n, \tag{9}$$

is the "leave-one-out" estimator of $\mu(x)$. Notice that the cross-validation criterion does not require knowledge of β nor does it require the researcher to know whether the process $\{X_t\}$ is stationary or nonstationary.

Remark 1. We emphasize that the CV-criterion is computed by averaging nonparametric residuals over a compact set. This is analogous to the iid case (Härdle and Marron (1985) or Härdle (1986)). □

3 Optimal Bandwidth for Conditional Mean Estimates

To state the properties of the CV-bandwidth $\widehat{h}_n$ in Eq. (7) we first introduce some definitions and assumptions.

By a slight abuse of notation, we use the symbol P for probabilities calculated on the probability space of the Markov chain as well as those calculated on the extended

[3] Our upper bound ($\overline{\eta} > \frac{1}{5}$) is slightly smaller than that in Härdle and Marron (1985) ($\overline{\eta} > 0$), but the lower bound $\underline{\eta} < 1$ is the same.

probability space of the corresponding split chain (see section 4.4 of Nummelin (1984) for details).

Let E denote expectations with respect to π_s the invariant measure of the Markov chain $\{X_k : k \geq 0\}$, i.e., $\pi_s g = EU_k(g)$ for $g \in L^1(\pi_s)$. For two functions l, u, let $[l, u]$ be the set of all functions f such that $l \leq f \leq u$. $[l, u]$ is called an ε-bracket in $L^p(\pi_s)$, $p \geq 1$, if $l, u \in L^p(\pi_s)$ and $\pi_s|u - l|^p \leq \varepsilon^p$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L^p(\pi_s))$ is the smallest number of ε-brackets in $L^p(\pi_s)$ covering $\mathcal{F}$. Let

$$\overline{\mathcal{F}} = \{f : \mathbb{R}^2 \to \mathbb{R} : f(x_1, x_2) = x_2 h K_{x,h}(x_1),\ x \in \mathcal{C},\ 0 < h \leq 1\}.$$

This set has envelope $\overline{F}(x_1, x_2) = |x_2| F(x_1)$, i.e. $|f| \leq \overline{F}$ for all $f \in \overline{\mathcal{F}}$, where $F(x) = c1\{x \in D\}$ and the constant c and the set $D \in \mathcal{E}$ are chosen such that F is the envelope of $\mathcal{F}$, i.e. $f \leq F$ for all $f \in \mathcal{F}$. For any $f \in \overline{\mathcal{F}}$, define

$$\overline{U}_k(f) = \begin{cases} \sum_{j=1}^{\tau_0} f(X_{j-1}, X_j), & k = 0 \\ \sum_{j=\tau_{k-1}+1}^{\tau_k} f(X_{j-1}, X_j), & 1 \leq k < n \\ \sum_{j=\tau_{T_n}+1}^{n} f(X_{j-1}, X_j), & k = n. \end{cases}$$

Assumption 2. *1. There are constants c_1 and c_2 so that $N_{[]}(\varepsilon, \mathcal{F}, L^1(\pi_s)) \leq c_1 \varepsilon^{-c_2}$ for all $\varepsilon \in (0, 1]$.*

2. The innovations u_t in Eq. (1) are iid with $\mathrm{E}(u_t^{2\kappa}) < \infty$, where $\kappa > \frac{\beta - \epsilon}{(\beta - \varepsilon)\overline{\eta} - 4\varepsilon}$, $\overline{\eta}$ is defined in Eq. (6), and $\epsilon, \varepsilon > 0$ are arbitrarily small.

3. $\sup_{f \in \mathcal{F}} E(U_k(f)^2) < \infty$ and $\sup_{f \in \overline{\mathcal{F}}} E(\overline{U}_k(f)^2) < \infty$.

Assumption 2.1 on the bracketing number is standard and satisfied, for example, by the Euclidean classes discussed in Nolan and Pollard (1987). To see this, consider the stationary case, $\beta = 1$. The assumption is implied by a standard bracketing condition on the space of bounded kernel functions. Let $N_{[]}(\varepsilon, \mathcal{F}, L^1(P)$ be the smallest number of brackets $[l, u] = \{f : l \leq f \leq u\}$ with $E|u(X_k) - l(X_k)| < \varepsilon$ covering the space of bounded kernel functions. Then, by Lemma 2.2 in Nolan and Pollard (1987), $N_{[]}(\varepsilon, \mathcal{F}, L^1(\pi_s)) \leq c_1 \varepsilon^{-c_2}$ for some constants $c_1, c_2 > 0$. Therefore, Assumption 2.1 holds.

Following Härdle and Marron (1985, p. 1466), we define optimality of a bandwidth sequence relative to a given criterion $d(\hat{\mu}_h, \mu)$ that measures the distance between the estimator $\hat{\mu}_h$ and μ. Specifically, we say that $\hat{h}_n$ is optimal with respect to $d(\hat{\mu}_h, \mu)$ if

$$\left| \frac{d\left(\widehat{\mu}_{\widehat{h}_n}, \mu\right)}{\inf_{h \in H_n} d\left(\widehat{\mu}_h, \mu\right)} \right| = o_p(1).$$

Intuitively, $\widehat{h}_n$ is optimal for $d\left(\widehat{\mu}_h, \mu\right)$ if it is equivalent to the argmin of the latter as the sample size gets large.

The criterion we consider is the Average Squared Error ("ASE"):

$$d_{A,h}\left(\widehat{\mu}, \mu\right) = \frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\widehat{\mu}_h(X_{j-1}) - \mu(X_{j-1})\right)^2 1\left\{X_{j-1} \in \mathcal{C}\right\}. \quad (10)$$

This criterion is natural in that it amounts to a least-squares metric, measuring squared differences between the nonparametric estimates and the unknown true function. Its argmin is defined as

$$\widetilde{h}_n = \arg\min_{h \in H_n} d_{A,h}\left(\widehat{\mu}, \mu\right).$$

Remark 2. The cross-validated criterion in Eq. (8) and the ASE in Eq. (10) are defined over a compact set $\mathcal{C}$. This restriction is important in both the stationary and the nonstationary case and analogous to the construction in Härdle and Marron (1985) and Härdle (1986), for instance. In the stationary case, the cross-validated bandwidth has been shown to be asymptotically equivalent to the *random* sequence minimizing the ASE (defined there as $\frac{T_n(\mathcal{C})}{n} d_{A,h}\left(\widehat{\mu}, \mu\right)$) and the *deterministic* sequence minimizing the mean-integrated squared error (MISE):

$$\begin{aligned} \mathrm{MISE}\left(h\right) &\overset{asy}{\sim} \mathrm{E}\left(\int \left(\left(\widehat{\mu}_h(x) - \mu(x)\right)^2\right) f(x) w(x) \mathrm{d}x\right) \\ &= \frac{1}{nh} \int K^2(u) \mathrm{d}u \int \sigma^2(x) \frac{1}{f(x)} w(x) \mathrm{d}x \\ &\quad + \int \left[\int K(u) \left(\mu(x - hu) - \mu(x)\right) f(x - hu) \mathrm{d}u\right]^2 \frac{1}{f(x)} w(x) \mathrm{d}x, \quad (11) \end{aligned}$$

where $f(x)$ is the stationary density of $\{X_t\}$. Under stationarity, if $w(x) = 1$ and the support is the entire real line, $\int_{\mathbb{R}} \frac{1}{f(x)} \mathrm{d}x$ may however not be finite. Hence, the MISE criterion (as well as the ASE criterion) would not be well-defined. On the other hand, the MISE stays finite for weights with compact support. In practice, this restriction is typically ignored and cross-validation is implemented over the whole state space of the process $\{X_t\}$. □

Remark 3. In nonstationary problems, the effective sample size T_n is a random variable. Hence, there is a difference between our assumed bandwidth set H_n, which is random, and the set assumed in stationary problems which is, instead, deterministic. □

Remark 4. Due to the random variance of the kernel estimator (see, e.g., Eq. (3)), the MISE would not have a deterministic form in nonstationary environments. Therefore, in general, one would not be able to show that the cross-validated bandwidth is, as in the stationary case, asymptotically equivalent to a deterministic sequence minimizing the estimator's MISE. □

Before presenting our main result in Theorem 1 below, we introduce a series of four Lemmas.

Lemma 1. *Under Assumptions 1.1–1.4 and 2,*

$$\sup_{x\in\mathcal{C},\, h\in H_n} |\widehat{p}_h(x) - p_s(x)| = o_{a.s.}(1),$$

where $p_s(x)$ is defined in Assumption 1.2 and

$$\widehat{p}_h(x) = \frac{1}{T_n h}\sum_{j=1}^{n} K\left(\frac{X_{j-1}-x}{h}\right). \tag{12}$$

Lemma 1 establishes uniform convergence of $\widehat{p}_h(x)$ to its deterministic counterpart, i.e., the density associated to the invariant measure π_s. This allows us to handle the contribution of the denominator in Eq. (2) and work with $\widehat{\mu}_h\frac{\widehat{p}_h}{p_s}$ rather than with $\widehat{\mu}_h$.

Lemma 2. *Let*

$$b_{\mathcal{C}}(x,h) = \frac{1}{p_s(x)}\left(\int K(u)\left(\mu(x-hu)-\mu(x)\right)p_s(x-hu)\mathrm{d}u\right)$$

and

$$\Omega_n = \left\{\varpi : \inf_{h\in H_n} d_{A,h}(\widehat{\mu},\mu) \geq \alpha C \inf_{h\in H_n}\left(\frac{1}{hn^{\beta+\varepsilon}} + \inf_{x\in\mathcal{C}} b^2_{\mathcal{C}}(x,h)\right)\right\},$$

for some $0<\alpha\leq 1$ and some constant C. Denote by Ω_n^c the complement of Ω_n. Let Assumptions 1 and 2 hold. Then

$$\lim_{n\to\infty}\Pr\{\Omega_n^c\} = 0.$$

Lemma 2 establishes a lower bound in probability for the ASE. This bound will prove useful in handling the denominator in the ratio of the main quantities in Lemma 3 and Lemma 4. Define

$$\overline{d}_{A,h}\left(\widehat{\mu},\mu\right)=\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{j,h}(X_{j-1})-\mu(X_{j-1})\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}, \tag{13}$$

with $\widehat{\mu}_{j,h}(x)$ as in Eq. (9).

Lemma 3. *Let Assumptions 1 and 2 hold. Then,*

$$\sup_{h\in H_n}\left|\frac{d_{A,h}\left(\widehat{\mu},\mu\right)-\overline{d}_{A,h}\left(\widehat{\mu},\mu\right)}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right|=o_p(1),$$

with $d_{A,h}\left(\widehat{\mu},\mu\right)$ and $\overline{d}_{A,h}\left(\widehat{\mu},\mu\right)$ defined as in Eq. (10) and Eq. (13), respectively.

Lemma 3 simply states that the average mean squared error criterion computed using either $\widehat{\mu}_h(X_{j-1})$ or the "leave-one-out" estimator $\widehat{\mu}_{j,h}(X_{j-1})$ are asymptotically equivalent (uniformly in h).

Lemma 4. *Let Assumptions 1 and 2 hold. Define*

$$\text{Cross}(h)=\overline{d}_{A,h}\left(\widehat{\mu},\mu\right)-CV(h)-\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(X_j-\mu(X_{j-1})\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}, \tag{14}$$

with $CV(h)$ as in Eq. (8). Then,

$$\sup_{h\in H_n}\left|\frac{\text{Cross}(h)}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right|=o_p(1).$$

Note that $\text{Cross}(h)=\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\sigma(X_{j-1})u_j\left(\widehat{\mu}_{h,j}(X_{j-1})-\mu(X_{j-1})\right).$ Lemma 4 establishes that this term converges to zero faster than the average squared error criterion.

Lemmas 1 through 4 lead to the following first main result of this paper:

Theorem 1. *Let Assumptions 1 and 2 hold. Then, the cross-validated bandwidth $\widehat{h}_n$ in Eq. (7) is asymptotically optimal with respect to $d_{A,h}\left(\widehat{\mu},\mu\right)$, i.e.,*

$$\left|\frac{d_{A,\widehat{h}_n}\left(\widehat{\mu},\mu\right)}{\inf_{h\in H_n}d_{A,h}\left(\widehat{\mu},\mu\right)}-1\right|=o_p(1).$$

Theorem 1 shows that the ASE based on the cross-validated bandwidth, $\widehat{h}_n$, is asymptotically equivalent to the minimum of the ASE. As mentioned above the convergence rate of the conditional mean estimator $\hat{\mu}_h(x)$ and, thus, that of the ASE itself is random and depends on the degree of recurrence, β. Remarkably, the optimality result of Theorem 1 holds even though the implementation of cross-validated bandwidths does not require knowledge of that convergence rate, knowledge of β, or knowledge of whether X_t is stationary or not.

Corollary 1. *Let Assumptions 1 and 2 hold. Then,*

$$\left| \frac{\widehat{h}_n}{\widetilde{h}_n} - 1 \right| = o_p(1),$$

where $\widetilde{h}_n = \arg\min_{h \in H_n} d_{A,h}(\widehat{\mu}, \mu)$.

Corollary 1 shows that the cross-validated bandwidth not only leads to an ASE that is asymptotically equivalent to its minimum, but also that the cross-validated bandwidth itself is asymptotically equivalent to the infeasible minimizer of the ASE.

3.1 The regression case

The regression case, namely

$$Y_t = g(X_{t-j}) + u_t$$

with $j \geq 0$ and u_t iid and independent of X_{t-j} can be treated theoretically as the autoregressive model presented above. One such an example will be discussed in Section 6 which is devoted to an empirical application in which Y_t is the future market return and X_{t-1} is the value of a persistent predictor over the previous period.

For a strictly nonstationary regressor X_{t-j} and $j = 0$, the regression can be interpreted as a nonlinear cointegrating model (provided u_t is stationary). Again, if the u_t's are iid and independent of X_t, our proposed methods go through unchanged.

If, instead, the u_t's display dependence of X_t, or are only asymptotically uncorrelated with X_t but are not strictly independent, then Lemma 4 would have to be modified accordingly. As a consequence, the statement in Theorem 1 would not apply directly. We leave the endogenous noise case for future work.

4 Optimal Bandwidth for Conditional Variance Estimates

In the previous section we established the optimality of the cross-validated bandwidth for the conditional mean estimator. In practice, one frequently also needs to optimally select the bandwidth for the conditional variance. One such example is provided in Section 6 below.

Even though the rate conditions for consistency and mixed normality of the conditional mean and variance estimates are the same, the optimal bandwidth cannot be the same because of the different functional form of the corresponding true functions. In the case of the conditional variance, the additional difficulty is that the conditional mean is unknown and ought to be estimated.

Define the feasible estimator

$$\widehat{\sigma}^2_\xi(x) = \frac{1}{T_n\xi}\sum_{i=2}^{n}\frac{1}{\widehat{p}_\xi(x)}K\left(\frac{X_{i-1}-x}{\xi}\right)\left(X_i-\widehat{\mu}_{\widehat{h}_n}(X_{i-1})\right)^2$$

and its infeasible counterpart, i.e.,

$$\widetilde{\sigma}^2_\xi(x) = \frac{1}{T_n\xi}\sum_{i=2}^{n}\frac{1}{\widehat{p}_\xi(x)}K\left(\frac{X_{i-1}-x}{\xi}\right)\left(X_i-\mu(X_{i-1})\right)^2,$$

where $\widehat{h}_n$ is the cross-validated bandwidth for the conditional mean and $\widehat{p}_\xi(x)$ is defined in Eq. (12). Similarly, we define the feasible cross-validation criterion

$$\mathrm{CV}_n(\xi) = \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(X_j-\widehat{\mu}_{\widehat{h}_n}(X_{j-1})\right)^2-\widehat{\sigma}^2_{j,\xi}(X_{j-1})\right)1\left\{X_{j-1}\in\mathcal{C}\right\}$$

and its infeasible counterpart

$$\widetilde{\mathrm{CV}}_n(\xi) = \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left((X_j-\mu(X_{j-1}))^2-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)1\left\{X_{j-1}\in\mathcal{C}\right\}, \tag{15}$$

where $\widehat{\sigma}^2_{j,\xi}(x)$ and $\widetilde{\sigma}^2_{j,\xi}(x)$ are the "leave-one-out" versions of $\widehat{\sigma}^2_\xi(x)$ and $\widetilde{\sigma}^2_\xi(x)$.
It is important to note that the estimation error of the conditional mean plays a twofold role. It affects the conditional variance estimator, i.e. $\widehat{\sigma}^2_{j,\xi}(X_{j-1})$ vs. $\widetilde{\sigma}^2_{j,\xi}(X_{j-1})$. It

also directly affects the cross-validation criterion to be minimized through the term $\left(X_j - \widehat{\mu}_{\widehat{h}_n}(X_{j-1})\right)^2$ versus $\left(X_j - \mu(X_{j-1})\right)^2$.

Now, define the relevant cross-validated bandwidths, namely

$$\widehat{\xi}_n = \arg\min_{\xi \in \Xi_n} \mathrm{CV}_n\left(\xi\right)$$

and

$$\widehat{\widetilde{\xi}}_n = \arg\min_{\xi \in \Xi_n} \widetilde{\mathrm{CV}}_n\left(\xi\right),$$

where

$$\Xi_n \equiv \left\{\underline{\xi}_n, \overline{\xi}_n\right\} = \left\{\underline{c}^\sigma T_n(\mathcal{C})^{-\underline{\eta}}, \overline{c}^\sigma T_n(\mathcal{C})^{-\overline{\eta}}\right\},$$

with $0 < \underline{c}^\sigma, \overline{c}^\sigma < \infty$, and $\underline{\eta}, \overline{\eta}$ as defined in Eq. (6), i.e. $\frac{1}{5} < \overline{\eta} < \underline{\eta} < 1$, and $1 - 2\underline{\eta} + \overline{\eta} > 0$. It is immediate to see that Ξ_n is of the same almost-sure order as H_n. We simply allow $\underline{c}^\sigma$ and $\overline{c}^\sigma$ to differ from $\underline{c}$ and $\overline{c}$.

Finally, we define the ASE criterion for the variance estimator, i.e.,

$$d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma\right) = \frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\widetilde{\sigma}^2_\xi(X_{j-1}) - \sigma^2(X_{j-1})\right)^2 1\left\{X_{j-1} \in \mathcal{C}\right\}$$

and its minimizer

$$\widetilde{\xi}_n = \arg\min_{\xi \in \Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma\right).$$

Our objective is to show that $\widehat{\xi}_n$ is asymptotically optimal with respect to $d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma\right)$, thereby yielding $|\widehat{\xi}_n/\widetilde{\xi}_n - 1| = o_p(1)$. This is accomplished in two steps. First, we establish the optimality of $\widehat{\widetilde{\xi}}_n$ for $d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma\right)$, i.e., we derive ASE optimality for the case in which the true conditional mean is assumed known, leading to $|\widehat{\widetilde{\xi}}_n/\widetilde{\xi}_n - 1| = o_p(1)$. Second, we show that conditional mean estimation does not affect this optimality result, i.e., $|\widehat{\xi}_n/\widehat{\widetilde{\xi}}_n - 1| = o_p(1)$.

Let $\tilde{\mathcal{F}} = \{f : \mathbb{R}^2 \to \mathbb{R} : f(x_1, x_2) = x_2^2 h K_{x,h}(x_1),\ x \in \mathcal{C},\ 0 < h \leq 1\}$. This set has envelope $\tilde{F}(x_1, x_2) = x_2^2 F(x_1)$, i.e. $|f| \leq \tilde{F}$ for all $f \in \tilde{\mathcal{F}}$, where F is the envelope of $\mathcal{F}$.

Assumption 3. *1. The innovations u_t in Eq. (1) are iid with $\mathrm{E}\left(u_t^{4\kappa}\right) < \infty$, $\kappa > \frac{\beta - \epsilon}{(\beta - \varepsilon)\overline{\eta} - 4\varepsilon}$, $\overline{\eta}$ defined in Eq. (6), and $\varepsilon, \epsilon > 0$ arbitrarily small.*

2. $\sup_{f \in \tilde{\mathcal{F}}} E(\overline{U}_k(f)^2) < \infty$.

This Assumption is very similar to Assumptions 2.2 and 2.3, and the same comments apply here. Before deriving our main result in Theorem 2 below, we introduce the following four lemmas.

Lemma 5. *Let*

$$b_{\mathcal{C}}^{\sigma}(x,\xi) = \frac{1}{p_s(x)}\left(\int K(u)\left(\sigma^2(x-\xi u) - \sigma^2(x)\right)p_s(x-\xi u)\mathrm{d}u\right)$$

and

$$\Omega_n^{\sigma} = \left\{\varpi : \inf_{\xi\in\Xi_n} d_{A,\xi}(\widehat{\sigma}^2,\sigma^2) \geq \alpha C \inf_{\xi\in\Xi_n}\left(\frac{1}{\xi n^{\beta+\varepsilon}} + \inf_{x\in\mathcal{C}} b_{\mathcal{C}}^{\sigma 2}(x,\xi)\right)\right\},$$

for some $0<\alpha\leq 1$ and some constant $C>0$. Denote by $\Omega_n^{\sigma,c}$ the complement of Ω_n^{σ}. Let Assumptions 1, 2.1, 2.3 and 3 hold. Then

$$\lim_{n\to\infty}\Pr\left\{\Omega_n^{\sigma,c}\right\} = 0.$$

Lemma 5 parallels Lemma 2 establishing a lower bound for our definition of the ASE criterion for the variance estimator.

Lemma 6. *Let Assumptions 1, 2.1, 2.3 and 3 hold. Then, the infeasible cross-validated bandwidth $\widetilde{\widetilde{\xi}}_n$ in Eq. (15) is asymptotically optimal with respect to $d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)$, i.e.,*

$$\left|\frac{d_{A,\widetilde{\widetilde{\xi}}_n}\left(\widetilde{\sigma}^2,\sigma^2\right)}{\inf_{\xi\in\Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)} - 1\right| = o_p(1).$$

Lemma 6 establishes the asymptotic optimality of the bandwidth minimizing the infeasible CV criterion. Next, we show the same result for the feasible CV criterion. We first need to control the probability order of the conditional mean's estimation error. This is done via Lemma 7.

Lemma 7. *Let Assumptions 1, 2.1, 2.3 and 3 hold. Then,*

$$\sup_{x\in\mathcal{C}}\left(\left(\widehat{\mu}_{\widehat{h}_n}(x) - \mu(x)\right)^4\right) = o_p\left(d_{A,\widehat{h}_n}\left(\widehat{\mu},\mu\right)\right).$$

Now, we ought to isolate the component of $\widetilde{\mathrm{CV}}_n(\xi) - \mathrm{CV}_n(\xi)$ which does not depend on ξ and show that the remaining component is of smaller probabilty order than

$\inf_{\xi\in\Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)$. This is accomplished in Lemma 8 below. Let

$$\begin{aligned}
&\mathrm{CV}(\xi)-\widetilde{\mathrm{CV}}(\xi)\\
&\quad=\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{j,\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)^4 1\left\{X_{j-1}\in\mathcal{C}\right\}\\
&\qquad+\frac{4}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left((X_j-\mu(X_{j-1}))\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}\\
&\qquad+\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})(u_j^2-1)\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)^2\right)1\left\{X_{j-1}\in\mathcal{C}\right\}\\
&\qquad-4\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}(X_j-\mu(X_{j-1}))\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)^3 1\left\{X_{j-1}\in\mathcal{C}\right\}\\
&\qquad-4\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})(u_j^2-1)\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\}\\
&\qquad+\widehat{\mathrm{Error}}(\xi),
\end{aligned}\tag{16}$$

where

$$\begin{aligned}
&\widehat{\mathrm{Error}}(\xi)\\
&\quad=\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\sigma}^2_{j,\xi}(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} &(17)\\
&\qquad+\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(\sigma^2(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)^2\right)1\left\{X_{j-1}\in\mathcal{C}\right\} &(18)\\
&\qquad+\frac{4}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(\sigma^2(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)\sigma(X_{j-1})u_j\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\} &(19)\\
&\qquad-\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(\sigma^2(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)\left(\widehat{\sigma}^2_{j,\xi}-\widetilde{\sigma}^2_{j,\xi}\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\} &(20)\\
&\qquad-\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})(u_j^2-1)\left(\widehat{\sigma}^2_{j,\xi}-\widetilde{\sigma}^2_{j,\xi}\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\} &(21)
\end{aligned}$$

$$+\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(\widehat{\sigma}^2_{j,\xi}-\widetilde{\sigma}^2_{j,\xi}\right)\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)^2\right)1\left\{X_{j-1}\in\mathcal{C}\right\} \tag{22}$$

$$+\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(\widehat{\sigma}^2_{j,\xi}-\widetilde{\sigma}^2_{j,\xi}\right)\sigma(X_{j-1})u_j\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\}. \tag{23}$$

Lemma 8. *Let Assumptions 1, 2.1, 2.3 and 3 hold. Then,*

$$\sup_{\xi\in\Xi_n}\left|\frac{\widehat{\text{Error}}(\xi)}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\right|=o_p(1).$$

Lemma 8 shows that the estimation error component of the cross-validation criterion is of sufficiently small probability order. We can finally turn to the optimality of the bandwidth minimizing the *feasible* criterion.

Theorem 2. *Let Assumptions 1, 2.1, 2.3 and 3 hold. Then, the cross-validated bandwidth $\widehat{\xi}_n$ is asymptotically optimal with respect to $d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)$, i.e.,*

$$\left|\frac{d_{A,\widehat{\xi}_n}\left(\widetilde{\sigma}^2,\sigma^2\right)}{\inf_{\xi\in\Xi_n}d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}-1\right|=o_p(1).$$

Corollary 2. *Let Assumptions 1, 2.1, 2.3 and 3 hold. Then,*

$$\left|\frac{\widehat{\xi}_n}{\widetilde{\xi}_n}-1\right|=o_p(1).$$

5 Simulations

In this section, we report the results of a simulation experiment illustrating the finite sample performance of our proposed bandwidth selection procedure. We generate data from four different models: an autoregressive process

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

with a linear, $\mu(x)=\rho x$, or a nonlinear, $\mu(x)=\rho x^2/10$, mean function, and a nonlinear predictive regression,

$$Y_t=\mu(X_t)+u_t$$
$$X_t=\rho X_{t-1}+\varepsilon_t$$

with a linear, $\mu(x) = x$, or a nonlinear[4], $\mu(x) = \sum_{j=1}^{4}(-1)^{j+1}\sin(j\pi x)/j^2$, mean function. For all four models, X_t is initialized at zero ($X_0 = 0$) and (u_t, ε_t) are independent, standard normal random variables, iid across time.

We are interested in nonparametrically estimating the function $\mu(.)$. We vary the sample size, T, and the degree of persistence, ρ, in the X_t process. All results are based on $1,000$ Monte Carlo samples, the standard normal density kernel, and the set $\mathcal{C}$ chosen to be equal to the range of the X_t data. The range of the data is, in general, not compact. This choice is intended to avoid the selection of an additional variable (i.e., a compact set $\mathcal{C}$) and provide evidence for the satisfactory performance of the criterion when such a selection is not made. We view this approach as being informative for applied work.

Since we are not aware of any other data-driven method for choosing the bandwidth in nonstationary settings, we compare our estimator to a linear least-squares estimator ("OLS"). Tables 1–4 present the bias, standard deviation ("stdev") and root mean-square error ("RMSE") of the nonparametric estimator based on cross-validated bandwidth and of the OLS estimator. The biases, standard deviations, and RMSEs are averaged values over a grid of points. Figure 1 plots the cross-validation criterion function and the selected bandwidth, each averaged over all Monte Carlo samples. Figure 2 plots the functional estimates.

In the linear specifications, both the bias and the standard deviation of the nonparametric estimates are larger than what is found for the least-squares estimates. As expected, while the standard deviation of the latter decreases with the level of persistence, the standard deviation of the former increases, due to a reduced number of visits of the process to each evaluation point. In the nonlinear specifications, the reduced biases of the nonparametric estimates more than offset some increases in dispersion with respect to the least-squares counterparts, thereby yielding root mean-squared errors which are considerably smaller than in the least-squares case.

In essence, the interaction of nonlinearities and increased levels of dependence may lead to large biases and sizeable root mean-squared errors in the least-squares case. Not only do cross-validated bandwidths satisfy meaningful optimality criteria in the presence of (potential) nonstationarities, as shown by our theoretical results, they also appear to lead to an empirically meaningful trade-off between bias and variance.

[4]This is the same nonlinear design as in Hall and Horowitz (2005) and Wang and Phillips (2009).

6 Nonlinear Stock-Return Predictability

We apply our proposed bandwidth selection method to a prototypical portfolio allocation problem. The investor computes optimal allocations to stocks (the market, in our case) and bonds from (possibly) nonlinear predictions of the conditional mean and variance of stock returns using the dividend-to-price ratio. We show the economic gains from using our cross-validated nonparametric procedure for prediction as compared to linear regression, something which is typical in the literature.

6.1 A Portfolio Allocation Problem

Consider an investor with quadratic utility

$$U(W_t) = W_t - \frac{\tilde{\lambda}}{2} W_t^2$$

over wealth W_t, who allocates a fraction ω_t of wealth in period t to the market and a fraction $(1-\omega_t)$ to treasury bills. Let the investor's planning horizon be τ periods, denote by $R_{t,t+\tau}$ and $R^f_{t,t+\tau}$ the τ-period returns of the market and treasury bills, respectively. The investor's end-of-horizon wealth $W_{t+\tau}$ is

$$W_{t+\tau} = \omega_t W_t R_{t,t+\tau} + (1-\omega_t) W_t R^f_{t,t+\tau}.$$

Suppose the excess returns $\tilde{R}_{t,t+\tau} = R_{t,t+\tau} - R^f_{t,t+\tau}$ evolve according to

$$\tilde{R}_{t,t+\tau} = \mu(X_t) + \sigma(X_t) u_{t+\tau} \tag{24}$$

where X_t is an assumed predictor and the errors $u_{t+\tau}$ satisfy $E[u_{t+\tau}|\mathcal{F}_t] = 0$ and $E[u^2_{t+\tau}|\mathcal{F}_t] = 1$. Further, suppose the investor's information set $\mathcal{F}_t$ at time t contains $(X_s, W_s, R^f_{s,s+\tau})$ for $s \le t$. Then, the period-t optimal allocation problem can be written as

$$\begin{aligned} \omega &= \text{argmax}_w E[U(W_{t+\tau}(w))|\mathcal{F}_t] \\ &= \text{argmax}_w \left\{ -\frac{\tilde{\lambda}}{2} W_t \left(\sigma^2(X_t) + \mu(X_t)^2\right) w^2 + \left(1 - \tilde{\lambda} W_t R^f_{t,t+\tau}\right) \mu(X_t) w \right\}. \end{aligned} \tag{25}$$

As is common in the literature (e.g. Fleming et al. (2001)) we facilitate comparisons across portfolios by keeping the relative risk aversion $\gamma_t = \tilde{\lambda} W_t / (1 - \tilde{\lambda} W_t)$ constant at a value

γ. Thus, the optimal portfolio weights are independent of wealth. Let $\lambda = \gamma/(1+\gamma)$. Conditional on $X_t = x$ and $R^f_{t,t+\tau} = r$, the optimal allocation of an investor solving Eq. (25), i.e., a "nonlinear investor", is

$$\omega_{non}(x) = \frac{(1-\lambda r)\mu(x)}{\lambda(\sigma^2(x) + \mu(x)^2)}.$$

A "linear investor" is one who assumes $\mu(x) = \alpha + \beta x$ and $\sigma(x) = \sigma$, leading to the optimal allocation

$$\omega_{lin}(x) = \frac{(1-\lambda r)(\alpha + \beta x)}{\lambda(\sigma^2 + (\alpha + \beta x)^2)}.$$

The optimal portfolio allocations, therefore, depend on the way in which the investor implements the predictive regression in Eq. (24), a problem which has received substantial attention in the literature.

6.2 Motivating (possibly) nonstationary cross-validation

Our bandwidth selection procedure appears attractive in the context of predictive regressions such as Eq. (24) because, as we now argue, the economics of stock return predictability suggest (possible) nonlinearities in the conditional mean and variance, extreme persistence of the predictor(s), and correlation between shocks to prices and shocks to the predictor(s).

First, Campbell and Shiller (1988) justify the standard conceptual framework for regressing stock returns on financial ratios. Their traditional log-linearization leads to a linear predictive regression which, more generally, can be viewed as an approximation to a more complex nonlinear model. Little is known about the quality of such an approximation, especially for longer horizons $h > 1$. In evaluating Eq. (24) using cross-validated nonparametric methods, we are therefore robust to deviations from linearity. Furthermore, the early literature generally focused on the conditional mean and set $\sigma(\cdot)$ equal to a constant. Recent implementations, however, treat the conditional variance similarly to the conditional mean and allow it to be a nonlinear function of the predictor. The survey by Brandt (2010) provides a discussion of nonlinear predictive models and their implications for portfolio allocation. Eq. (24) accommodates nonlinear conditional second moments as well.

Second, autoregressive models with local-to-unit roots are known to capture well the persistence properties of financial ratios used for stock return prediction (e.g. Cavanagh et al. (1995), Torous and Valkanov (2000), Valkanov (2003)). We will show that the assumed predictor, i.e., the dividend-to-price ratio, in Eq. (24) is very highly persistent, nonstationarity tests supporting the null of nonstationarity.

Third, if X_t is the dividend-to-price ratio or, alternatively, any financial ratio whose denominator depends on the market's price level, its variation between time $t-1$ and time t is necessarily correlated with the variation in $R_{t-1,t}$, the market return over the same period. A positive shock to prices lowers X_t while increasing $\tilde{R}_{t-1,t}$, thereby inducing a negative correlation between $\tilde{R}_{t-1,t}$ and X_t. Stambaugh (1986, 1999) provides early discussions of the importance of this correlation and its role in predictive models for stock returns. The limiting properties of the cross-validated moment estimates in Eq. (24) are not affected by this correlation.

6.3 Results

The data is obtained from CRSP and is the same as that in Lewellen (2004). We compute excess returns from monthly continuously-compounded returns ($R_{t,t+1}$) on the value-weighted NYSE index net of the one-month treasury bill rate ($R^f_{t,t+1}$). The predictor X_t is the dividend-to-price ratio constructed as dividend paid during the previous year divided by the current value of the value-weighted NYSE index. The τ-period excess returns $\tilde{R}_{t,t+\tau}$ are aggregates of 1-month excess returns: $\tilde{R}_{t,t+\tau} = \sum_{i=1}^{\tau-1} \tilde{R}_{t+i,t+i+1}$.

Figure 3 displays the data. As discussed, the dividend-to-price ratio is very highly persistent. Conventional unit root tests, irrespective of whether the null is nonstationarity, as in the Dickey-Fuller tradition, or stationarity, as proposed by Kwiatkowski et al. (1992), find it hard not to exclude stationarity on purely statistical grounds. For example, an augmented Dickey-Fuller test (implemented using a constant in the regression, a maximum lag length of 20, and an automatic lag length selection using Schwartz information criterion) delivers a t-statistic of -1.47 for a corresponding 10% critical value of -2.57, thereby leading to a failure to reject the nonstationarity null. The KPSS test, instead, rejects the null of stationarity overwhelmingly at the 1% level. The value of the statistic (when using a Newey-West automated bandwidth and a Bartlett kernel) is 1.04 and the 1% critical value is 0.74. While the reduced power of unit-root tests against local-to-unity alternatives is well-known, these results point to the unit-root behavior of the dividend-to-price ratio and its extreme persistence. Regardless of whether one believes in its stationarity (which is more economically plausible) or lack thereof, methods of inference, like the one we propose, which do not have to rely on either appear theoretically and empirically warranted.

We implement the nonparametric kernel estimators introduced above together with our proposed cross-validation criterion for choosing the bandwidth. For investment horizons τ larger than one month, we modify the cross-validation criterion function so as to leave out not only one observation, but also τ before and τ after a particular observation. This modification controls the dependence structure of the regression residuals upon aggregation. All calculations are based on the normal density kernel and, again, a set $\mathcal{C}$ chosen to be equal to the range of the X_t data. We report results for the investment horizons of one month ($\tau = 1$), 3 years ($\tau = 36$), and 5 years ($\tau = 60$), and risk aversion parameters $\gamma \in \{1, 5, 10\}$.

Figure 4 shows the cross-validation objective function, which is very flat near its minimum when $\tau = 1$, leading to large bandwidth choices, and possesses clearly separate minima for horizons $\tau > 1$, generating much smaller bandwidths. Figure 5 provides the resulting nonparametric estimates and the asymptotic (pointwise) standard errors of the conditional mean and variance functions. The conditional means are increasing in the predictor and mildly nonlinear, with the nonlinearities being statistically significant at the two higher horizons. As is well-known, a high dividend-to-price ratio should predict high returns, low dividend growth, or decreasing prices. It generally predicts high returns. It also generally does so strongly over longer horizons. Leaving nonlinearities aside, our estimates exhibit overall shapes that confirm the findings in the literature. Similarly, at the short horizon we find a flat conditional variance which is consistent with classical parametric approaches in the predictability literature. However, variance appears to be nonlinear, and decreasing, at longer horizons. When predicting over 3 and 5 years, a higher conditional variance seems to be associated with higher prices relative to dividends. This effect is interesting. In Figure 6, we report three scatterplots of squared de-meaned excess returns, i.e. $(\sum_{i=1}^{\tau-1} \tilde{R}_{t+i,t+i+1} - \widehat{\mu}_{\widehat{h}_n}(X_t))^2$, against X_t, along with least-squares estimates of the same relation. Again, the long-run conditional variance appears to decrease with increases in the dividend-to-price ratio. A least-squares specification would capture these effects while missing some nonlinearities around the mean dividend-to-price level.

We now evaluate the economic impact of taking into account the nonlinearities found in the estimates of Figure 5. We quantify the economic gains for the investor employing a nonlinear rather than a linear specification by computing an annualized fee Δ that makes the nonlinear investor indifferent between the nonlinear and the linear specification. Specifically, we follow Fleming et al. (2001) and define Δ as the solution[5] to

$$\frac{1}{T}\sum_{t=1}^{T}\left\{R^p_{t,t+\tau}(\omega_{non}(X_t)) - \Delta - \frac{\lambda}{2}\left(R^p_{t,t+\tau}(\omega_{non}(X_t)) - \Delta\right)^2\right\}$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left\{R^p_{t,t+\tau}(\omega_{lin}(X_t)) - \frac{\lambda}{2}R^p_{t,t+\tau}(\omega_{lin}(X_t))^2\right\}, \quad (26)$$

where $R^p_{t,t+\tau}(\omega) = \omega R_{t,t+\tau} + (1-\omega)R^f_{t,t+\tau}$ denotes the portfolio return for a weight ω. Figure 7 shows the realized utilities of the nonlinear investor minus those of the linear investor plotted over time. Δ is the additional annualized return necessary for the nonlinear investor to have the same average realized utility as the linear investor. Table 5 reports the estimates of Δ. As expected, the annualized fees are small for short investment horizons for which we have already seen that the conditional mean and variance are essentially linear. For the longer investment horizons, however, we found significant nonlinearities in the conditional mean and variance, and therefore the nonlinear investor requires large fees Δ to be indifferent between the nonlinear and the linear specification. The fees range from a 1.5% to a 6.2% annualized return in addition to the portfolio return to reach indifference. These fees indicate large economic gains from (cross-validated) nonlinear predictability.

7 Conclusions

Cross-validation is the most widely used method of bandwidth selection in nonparametric econometrics. It is employed routinely in empirical work irrespective of the level of persistence. We show that this common practice is theoretically justified. Even in the nonstationary case, which is a sub-case of our broader framework, the cross-validated bandwidth is optimal with respect to the averaged squared error criterion, i.e., the averaged squared distance between the true function and its nonparametric counterpart. Should stationarity be satisfied, the classical optimality with respect to the mean integrated squared error would be easily re-established. We provide a treatment which

[5] In general, there are two solutions to Eq. (26). We discard the one that moves the investor's return to the decreasing side of the utility function.

covers both the conditional mean and the conditional variance estimator in the context of (positive and null) recurrent Markov chain models.

Even though many economic time series may not be genuinely nonstationary, persistence is a fact of life. While extreme forms of persistence lead us to re-consider the allowed criterion for optimality, this paper shows that a meaningful notion of optimality continues to be a fundamental property of cross-validated bandwidth choices. This is, in our view, important information for nonparametric applied work in economics and finance.

A Proofs

A.1 Proofs of Section 3

The proof of Lemma 1 is based on the following Proposition.

Proposition 1. *Under Assumptions 1 and 2,*

$$\sup_{x\in\mathcal{C},\, h\in H_n} \left| \frac{1}{T_n}\sum_{k=1}^{T_n} U_{k,x,h} - E\left(U_{k,x,h}\right) \right| = o_{a.s.}(1) \tag{27}$$

and

$$\sup_{x\in\mathcal{C},\, h\in H_n} \left| \frac{1}{T_n(\mathcal{C})}\sum_{k=1}^{T_n(\mathcal{C})} U_{k,x,h} - E\left(U_{k,x,h}\right) \right| = o_{a.s.}(1). \tag{28}$$

Proof. Consider some subsequence $\{a_n : n \geq 1\}$ of $\{n\}$ such that $a_n \to \infty$ as $n \to \infty$. Let

$$\tilde{H}_{a_n,n} = \left\{ h \in \mathbb{R} :\ \underline{\tilde{h}}_{a_n,n} \leq h \leq \tilde{\overline{h}}_{a_n,n} \right\}$$

with $\underline{\tilde{h}}_{a_n,n} = \underline{c} a_n^{-\underline{\eta}}$ and $\tilde{\overline{h}}_{a_n,n} = \overline{c} a_n^{-\overline{\eta}}$. We first show that

$$\bar{A}_{a_n} = \sup_{x\in\mathcal{C},\, h\in\tilde{H}_{a_n,n}} \left| \frac{1}{a_n}\sum_{k=1}^{a_n} U_{k,x,h} - E\left(U_{k,x,h}\right) \right| = o_{a.s.}(1). \tag{29}$$

For a sequence of random variables $A_1, A_2, \ldots$, define the empirical average $P_n A_k = \frac{1}{a_n}\sum_{k=1}^{a_n} A_k$. The proof is organized in three steps.

Step (1): Define the truncated version of $U_k(f)$ as $T_k(f) = U_k(f)1\{U_k(F)^2 \leq b_n\}$, where $b_n = a_n^{2(1-\overline{\eta}+\underline{\eta})}$ and $b_n \to \infty$ because $\underline{\eta} > \overline{\eta}$. Also, let R_n be the approximation error of $U_k(f)$ by $T_k(f)$:

$$R_n = \sup_{f\in\mathcal{F}_{a_n,n}} \left\{ |P_n(U_k(f) - T_k(f))| + |E[U_k(f) - T_k(f)]| \right\},$$

where $\mathcal{F}_{a_n,n} = \{f = hK_{x,h} : x \in \mathcal{C}, h \in \tilde{H}_{a_n,n}\}$ defines a sequence of subsets of $\mathcal{F}$ that incorporates rate restrictions on the bandwidth h. Note that $U_k(F)^2 > b_n$ and $U_k(F) \geq 0$ imply $U_k(F) < U_k(F)^2/\sqrt{b_n}$. We have

$$\begin{aligned} |P_n T_k(f) - P_n U_k(f)| &= P_n U_k(f) 1\{U_k(F)^2 > b_n\} \\ &\leq P_n U_k(F) 1\{U_k(F)^2 > b_n\} \\ &\leq b_n^{-1/2} P_n U_k(F)^2 1\{U_k(F)^2 > b_n\} \\ &\leq b_n^{-1/2} P_n U_k(F)^2. \end{aligned} \tag{30}$$

Taking expectations on both sides of Eq. (30) yields

$$|E(T_k(f)) - E(U_k(f))| \leq E|T_k(f) - U_k(f)| \leq b_n^{-1/2} E(U_k(F)^2). \tag{31}$$

Eqs. (30) and (31) together with the fact that $U_1(f), U_2(f), \ldots$ are i.i.d. (p. 135 in Nummelin (1984)), the strong law of large numbers, Assumption 2.3, and the definition of b_n imply

$$R_n = O_{a.s.}(b_n^{-1/2}) = o_{a.s.}(1),$$

i.e. we can replace $U_k(f)$ in Eq. (29) by the truncated variable $T_k(f)$ since $b_n^{-1/2}$ is a negligible higher-order term.

Step (2): We now apply a standard bracketing argument to show that the truncated process $P_n T_k(f)$ converges almost surely to its expectation uniformly over $\mathcal{F}_{a_n,n}$. To this end, notice that, for any $\xi > 0$:

$$\begin{aligned} P&\left(\sup_{x\in\mathcal{C},\, h\in\tilde{H}_{a_n,n}} |P_n T_k(K_{x,h}) - ET_k(K_{x,h})| > \xi\right) \\ &\leq P\left(\sup_{f\in\mathcal{F}_{a_n,n}} \frac{1}{\underline{\tilde{h}}_{a_n,n}} |P_n T_k(f) - ET_k(f)| > \xi\right) \\ &= P\left(\sup_{f\in\mathcal{F}_{a_n,n}} P_n T_k(f) - ET_k(f) > \xi \underline{\tilde{h}}_{a_n,n}\right) \quad (32) \\ &\quad + P\left(\sup_{f\in\mathcal{F}_{a_n,n}} P_n T_k(f) - ET_k(f) < -\xi \underline{\tilde{h}}_{a_n,n}\right). \quad (33) \end{aligned}$$

Consider Eq. (32). By Assumption 2.1, there is a collection of $\varepsilon\underline{\tilde{h}}_{a_n,n}$-brackets $B(\varepsilon\underline{\tilde{h}}_{a_n,n}, \mathcal{F}_{a_n,n}, L^1(\pi_s))$ of cardinality $N_{[\,]}(\varepsilon\underline{\tilde{h}}_{a_n,n}, \mathcal{F}_{a_n,n}, L^1(\pi_s)) \leq N_{[\,]}(\varepsilon\underline{\tilde{h}}_{a_n,n}, \mathcal{F}, L^1(\pi_s))$ so that, for every $f \in \mathcal{F}_{a_n,n}$, there is a bracket $[f^l, f^u]$ such that $f^l \leq f \leq f^u$ and $E[U_k(f^u) - U_k(f^l)] \leq \varepsilon\underline{\tilde{h}}_{a_n,n}$. Therefore, for any $f \in \mathcal{F}_{a_n,n}$,

$$E[T_k(f^u) - T_k(f)] \leq E\left[(U_k(f^u) - U_k(f))1\{U_k(F)^2 \leq b_n\}\right] \leq E\left[(U_k(f^u) - U_k(f))\right] \leq \varepsilon\underline{\tilde{h}}_{a_n,n}$$

or, equivalently,

$$ET_k(f) \leq ET_k(f^u) - \varepsilon\underline{\tilde{h}}_{a_n,n}. \quad (34)$$

By Assumptions 1 and 2, we can apply Lemma 4.1 in Karlsen and Tjøstheim (1998) to get $hE[U_k(K_{x,h})^2] = O(1)$ and thus, for any $f \in \mathcal{F}_{a_n,n}$, $EU_k(f)^2 = O(\bar{\tilde{h}}_{a_n,n})$. We use this fact together with Eq. (34) to bound the term in Eq. (32) as follows: letting c_j denote some constants,

$$
\begin{aligned}
&P\left(\sup_{f\in\mathcal{F}_{a_n,n}} P_n\left[T_k(f)-ET_k(f)\right] > \xi\underline{\tilde{h}}_{a_n,n}\right)\\
&\quad\le P\left(\max_{f^u:\,[f^l,f^u]\in B(\varepsilon\underline{\tilde{h}}_{a_n,n},\mathcal{F}_{a_n,n},L^1(\pi_s))} P_n\left[T_k(f^u)-ET_k(f^u)\right] > (\xi+\varepsilon)\underline{\tilde{h}}_{a_n,n}\right)\\
&\quad\le N_{[\,]}(\varepsilon,\mathcal{F}_{a_n,n},L^1(\pi_s))\\
&\qquad\times \max_{f^u:\,[f^l,f^u]\in B(\varepsilon\underline{\tilde{h}}_{a_n,n},\mathcal{F}_{a_n,n},L^1(\pi_s))} P\left(P_nT_k(f^u)-ET_k(f^u) > (\xi+\varepsilon)\underline{\tilde{h}}_{a_n,n}\right)\\
&\quad\le c_1\varepsilon^{-c_2}\cdot\exp\left\{-\frac{a_n^2(\xi+\varepsilon)^2\underline{\tilde{h}}_{a_n,n}^2}{a_nE[T_k(f^u)^2]+\frac{1}{3}\sqrt{b_n}(\xi+\varepsilon)\underline{\tilde{h}}_{a_n,n}}\right\}\\
&\quad\le c_1\varepsilon^{-c_2}\cdot\exp\left\{-\frac{a_n^2(\xi+\varepsilon)^2\underline{\tilde{h}}_{a_n,n}^2}{a_nE[U_k(f^u)^2]+(\xi+\varepsilon)o(1)}\right\}\\
&\quad\le c_4\cdot\exp\left\{-\frac{a_n\underline{\tilde{h}}_{a_n,n}^2}{\bar{\tilde{h}}_{a_n,n}}\right\}\\
&\quad\le c_5\cdot\exp\left\{-a_n^{1-2\underline{\eta}+\overline{\eta}}\right\}\to 0,
\end{aligned}
$$

where we use $|T_k(f)|\le\sqrt{b_n}$, the Bernstein inequality in Appendix B of Pollard (1984), and $\frac{1}{5}<\overline{\eta}<\underline{\eta}<1$. The term in Eq. (33) is bounded in a similar fashion. Therefore, the Borel-Cantelli Lemma implies $\bar{A}_{a_n}\to 0$ a.s. as $n\to\infty$.

Step (3): By Lemma 3.2 in Karlsen and Tjøstheim (2001), we have that $T_n(\mathcal{C})\to\infty$ a.s. as $n\to\infty$. Fix an ω such that $\bar{A}_{a_n}(\omega)\to 0$. Then $\bar{A}_{T_n(\mathcal{C})}(\omega)\to 0$ so that Eq. (28) follows because this happens for all ω outside a null set. Eq. (27) follows because $T_n(\mathcal{C})/T_n\to\pi_s 1_{\mathcal{C}}$ a.s. as $n\to\infty$. Q.E.D.

Proof of Lemma 1. The estimator $\hat{p}_h(x)$ can be decomposed as

$$\widehat{p}_h(x)=\frac{1}{T_n}U_{0,x,h}+\frac{1}{T_n}\sum_{k=1}^{T_n}U_{k,x,h}+\frac{1}{T_n}U_{n,x,h}, \tag{35}$$

Consider the middle term first. By Proposition 1,

$$\sup_{x\in\mathcal{C},\,h\in H_n}\left|\frac{1}{T_n}\sum_{k=1}^{T_n}U_{k,x,h}-E\left(U_{k,x,h}\right)\right|=o_{a.s.}(1). \tag{36}$$

Following the same steps as in the method of proof of Theorem 4.1 in Gao et al. (2014), we have

$$E\left(U_{k,x,h}\right)=\int\frac{1}{h}K\left(\frac{u-x}{h}\right)d\pi_s(u)=\int K(u)p_s(x+hu)d(u),$$

Therefore, by Assumption 1.2, $\sup_{x\in\mathcal{C}}p_s(x)<\infty$ and, thus,

$$\sup_{x\in\mathcal{C},\,h\in H_n}\left|E\left(U_{k,x,h}\right)-p_s(x)\right|=o(1). \tag{37}$$

By Eq. (36) and Eq. (37), it remains to show that the boundary terms $\sup_{x\in\mathcal{C},h\in H_n}|\frac{1}{T_n}U_{0,x,h}|$ and $\sup_{x\in\mathcal{C},h\in H_n}|\frac{1}{T_n}U_{n,x,h}|$ are negligible. To this end, notice that Proposition 1 holds with T_n replaced by T_n+1 so that $\sup_{x\in\mathcal{C},h\in H_n}|\frac{1}{T_n}U_{n,x,h}| = o_{a.s.}(1)$. The negligibility of $\sup_{x\in\mathcal{C},h\in H_n}|\frac{1}{T_n}U_{0,x,h}|$ follows similarly as in the proof of Theorem 5.1 in Karlsen and Tjøstheim (2001, p. 405-406). Q.E.D.

Let

$$\overline{U}_{k,x,h} = \begin{cases} \frac{1}{h}\sum_{i=1}^{\tau_0}\overline{u}_i K\left(\frac{X_{i-1}-x}{h}\right) \text{ when } k=0 \\ \frac{1}{h}\sum_{i=\tau_{k-1}+1}^{\tau_k}\overline{u}_i K\left(\frac{X_{i-1}-x}{h}\right) \text{ for } 1\le k<n \\ \frac{1}{h}\sum_{\tau_{T_n}+1}^{n}\overline{u}_i K\left(\frac{X_{i-1}-x}{h}\right) \text{ for } k=n \end{cases},$$

where $\overline{u}_i = \sigma(X_i)u_i$.

Proposition 2. *Under Assumptions 1 and 2,*

$$\sup_{x\in\mathcal{C},\,h\in H_n}\left|\frac{1}{T_n}\sum_{k=1}^{T_n}\overline{U}_{k,x,h}\right| = o_{a.s.}(1) \tag{38}$$

and

$$\sup_{x\in\mathcal{C},\,h\in H_n}\left|\frac{1}{T_n(\mathcal{C})}\sum_{k=1}^{T_n(\mathcal{C})}\overline{U}_{k,x,h}\right| = o_{a.s.}(1). \tag{39}$$

Proof. The proof of this proposition is very similar to the one of Proposition 1. By the same argument as in Example (4.3) of Pollard (1986), $\overline{\mathcal{F}}$ is Euclidean for the envelope $\overline{F}$. Therefore, there is a collection of $\varepsilon\underline{\tilde{h}}_{a_n,n}$-brackets $\overline{B}(\varepsilon\underline{\tilde{h}}_{a_n,n},\overline{\mathcal{F}}_{a_n,n},L^1(\pi_s))$ of cardinality

$$N_{[\,]}(\varepsilon\underline{\tilde{h}}_{a_n,n},\overline{\mathcal{F}}_{a_n,n},L^1(\pi_s)) \le N_{[\,]}(\varepsilon\underline{\tilde{h}}_{a_n,n},\mathcal{F}_{a_n,n},L^1(\mu)) \le N_{[\,]}(\varepsilon\underline{\tilde{h}}_{a_n,n},\mathcal{F},L^1(\mu)),$$

where μ is the measure that has density $|x|$ with respect to π_s and

$$\overline{\mathcal{F}}_{a_n,n} = \{f:\mathbb{R}^2\to\mathbb{R}:\ f(x_1,x_2) = x_2 h K_{x,h}(x_1),\, x\in\mathcal{C},\, h\in\tilde{H}_{a_n,n}\}.$$

The truncation and bracketing argument in the proof of Lemma 1 therefore applies here as well. Q.E.D.

Proposition 3. *Under Assumptions 1 and 2:*

$$\sup_{x\in\mathcal{C},\,h\in H_n}|\widehat{\mu}_h(x)-\mu(x)| = o_{a.s.}(1).$$

Proof. The statement follows from Lemma 1 and a similar calculation as in Eq. (37) to show that the bias is negligible. Q.E.D.

Proof of Lemma 2. Since

$$\widehat{\mu}_h(x)-\mu(x) = (\widehat{\mu}_h(x)-\mu(x))\frac{\widehat{p}_h(x)}{p_s(x)} - (\widehat{\mu}_h(x)-\mu(x))\left(\frac{\widehat{p}_h(x)-p_s(x)}{p_s(x)}\right),$$

we have

$$\begin{aligned}
d_{A,h}\left(\widehat{\mu}_h,\mu\right) &= \frac{1}{T_n(\mathcal{C})}\sum_{j=1}^{n}\left(\left(\widehat{\mu}_h\left(X_{j-1}\right)-\mu(X_{j-1})\right)\frac{\widehat{p}_h(X_{j-1})}{p_s(X_{j-1})}\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&+ \frac{1}{T_n(\mathcal{C})}\sum_{j=1}^{n}\left(\left(\widehat{\mu}_h\left(X_{j-1}\right)-\mu(X_{j-1})\right)\frac{\left(\widehat{p}_h(X_{j-1})-p_s(X_{j-1})\right)}{p_s(X_{j-1})}\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&+ \text{cross} \\
&= I_{n,h}+II_{n,h}+III_{n,h}.
\end{aligned} \tag{40}$$

Letting $\widetilde{\mu}_h(x)=\widehat{\mu}_h(x)\frac{\widehat{p}_h(x)}{p_s(x)}$ and $\mu^*(x)=\mu(x)\frac{\widehat{p}_h(x)}{p_s(x)}$, the term $I_{n,h}$ writes

$$\begin{aligned}
I_{n,h} &= \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&+ \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)-\mu^*(X_{j-1})\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&+ \frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)-\mu^*(X_{j-1})\right)\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&= I^A_{n,h}+I^B_{n,h}+I^C_{n,h}.
\end{aligned}$$

Thus,

$$\begin{aligned}
I^A_{n,h} &= \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&+ \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\Big(\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&- \mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}\Big) \\
&= I^{A1}_{n,h}+I^{A2}_{n,h}.
\end{aligned}$$

Define the set $\Omega_{1,n}=\{\omega:\quad n^{\beta-\epsilon}\ll T_n\ll n^{\beta+\epsilon}\}$, where the symbol "$\ll$" is used to denote smaller order. By Lemma 3.4 in Karlsen and Tjøstheim (2001), we have

$$\mathrm{P}\left(\lim_{n\to\infty}\Omega_{1,n}\right)=1.$$

Now, working conditionally on $\Omega_{1,n}$, we have

$$\begin{aligned}
I^{A1}_{n,h} &\geq \inf_{x\in\mathcal{C}}\mathrm{E}\left(\left(\widetilde{\mu}_h(x)-\mathrm{E}(\widetilde{\mu}_h(x))\right)^2\right)\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&= \inf_{x\in\mathcal{C}}\mathrm{E}\left(\left(\widetilde{\mu}_h(x)-\mathrm{E}(\widetilde{\mu}_h(x))\right)^2\right) \\
&\geq \inf_{x\in\mathcal{C}}\sigma^2(x)\frac{1}{\sup_{x\in\mathcal{C}}p_s^2(x)}\frac{1}{n^{\beta+\epsilon}h}\int K^2(u)du\left(1+o(1)\right) \\
&\geq O\left(\frac{1}{n^{\beta+\epsilon}h}\right),
\end{aligned} \tag{41}$$

uniformly in h, where the expression of the variance term derives from Eq. (3) and the final order (driven by $T_n(\mathcal{C})$) derives from the fact that T_n and $T_n(\mathcal{C})$ are of the same almost-sure order (c.f. Remark 3.5 in Karlsen and Tjøstheim (2001)). We now show that $I_{n,h}^{A2}$ is of smaller order of magnitude than $I_{n,h}^{A1}$. Write

$$I_{n,h}^{A2} = \frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{T_n(\mathcal{C})} \left((\widetilde{\mu}_h(X_{j-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{j-1})))^2 - \mathrm{E}(\widetilde{\mu}_h(X_{j-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{j-1})))^2 \right).$$

It suffices to show that $\sqrt{\mathrm{var}\left(I_{n,h}^{A2}\right)} = o\left(\frac{1}{n^{\beta+\epsilon}h}\right)$. We have,

$$\begin{aligned}
& \mathrm{var}\left(I_{n,h}^{A2}\right) \\
\sim \; & \frac{1}{T_n^2(\mathcal{C})} \sum_{j=2}^{T_n(\mathcal{C})} \mathrm{E}\left((\widetilde{\mu}_h(X_{j-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{j-1})))^2 - \mathrm{var}(\widetilde{\mu}_h(X_{j-1})) \right)^2 \\
& + \frac{1}{T_n^2(\mathcal{C})} \sum_{j=2}^{T_n(\mathcal{C})} \sum_{i=2}^{T_n(\mathcal{C})} \mathrm{E}\Big(1\{|X_{i-1} - X_{j-1}| \leq h\} \left((\widetilde{\mu}_h(X_{j-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{j-1})))^2 - \mathrm{var}(\widetilde{\mu}_h(X_{j-1})) \right) \\
& \left((\widetilde{\mu}_h(X_{i-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{i-1})))^2 - \mathrm{var}(\widetilde{\mu}_h(X_{i-1})) \right) \Big) \\
= \; & A_{n,h} + B_{n,h},
\end{aligned}$$

where the symbol "$\sim$" denotes "asymptotic equivalence." Write,

$$A_{n,h} = \frac{1}{T_n^2(\mathcal{C})} \sum_{j=2}^{T_n(\mathcal{C})} \mathrm{E}\left((\widetilde{\mu}_h(X_{j-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{j-1})))^4 \right) - \frac{1}{T_n^2(\mathcal{C})} \sum_{j=2}^{T_n(\mathcal{C})} \mathrm{var}^2(\widetilde{\mu}_h(X_{j-1})).$$

Recalling that $\overline{U}_{k,h,x} = \sum_{j=\tau_{k-1}}^{\tau_k} \frac{1}{h} K\left(\frac{X_{j-1}-x}{h}\right) \overline{u}_j$, we have

$$\begin{aligned}
\frac{1}{T_n^2(\mathcal{C})} \sum_{j=2}^{T_n(\mathcal{C})} \mathrm{E}(\widetilde{\mu}_h(X_{j-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{j-1})))^4 &\leq \frac{1}{T_n(\mathcal{C})} \sup_{x \in \mathcal{C}} \mathrm{E}(\widetilde{\mu}_h(X_{j-1}) - \mathrm{E}(\widetilde{\mu}_h(X_{j-1})))^4 \\
&\leq \frac{1}{T_n(\mathcal{C})} \frac{1}{\inf_{x \in \mathcal{C}} p_s^4(x)} \sup_{x \in \mathcal{C}} \mathrm{E}\left(\frac{1}{T_n} \sum_{k=1}^{T_n} \overline{U}_{k,h,x} \right)^4. \qquad (42)
\end{aligned}$$

Working now with the same subsequence $\{a_n : n \geq 1\}$ as in Proposition 1 and defining h, again, on $\tilde{H}_{a_n,n} = \left\{ h \in \mathbb{R} : \tilde{\underline{h}}_{a_n,n} \leq h \leq \tilde{\overline{h}}_{a_n,n} \right\}$ with $\tilde{\underline{h}}_{a_n,n} = \underline{c} a_n^{-\overline{\eta}}$ and $\tilde{\overline{h}}_{a_n,n} = \overline{c} a_n^{-\overline{\eta}}$, we have

$$\begin{aligned}
\sup_{x \in \mathcal{C}} \left\{ \frac{1}{a_n^4} \sum_{k=1}^{a_n} \mathrm{E}\left(\overline{U}_{k,h,x}^4\right) + \frac{1}{a_n^4} \sum_{k=1}^{a_n} \sum_{i=1}^{a_n} \mathrm{E}\left(\overline{U}_{k,h,x}^2\right) \mathrm{E}\left(\overline{U}_{i,h,x}^2\right) \right\} &= O\left(\frac{1}{a_n^3 h^3}\right) + O\left(\frac{1}{a_n^2 h^2}\right) \\
&= O\left(\frac{1}{a_n^2 h^2}\right)(1 + o(1)), \qquad (43)
\end{aligned}$$

uniformly in $h \in H_n$, with the order terms in the last two lines deriving from Lemma B.3 in Gao et al. (2014). Combining now Eq. (42) and Eq. (43), and using the fact that $n^{\beta-\epsilon} \ll a_n$ and $n^{\beta-\epsilon} \ll T_n(\mathcal{C})$, we have

$$\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^4=O\left(\frac{1}{n^{3(\beta-\epsilon)}h^2}\right)(1+o(1))\,.$$

Similarly, it is now immediate to see that

$$\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\mathrm{var}^2\left(\widetilde{\mu}_h(X_{j-1})\right)\leq\frac{1}{T_n(\mathcal{C})}\sup_{x\in\mathcal{C}}\mathrm{var}^2\left(\widetilde{\mu}_h(x)\right)=O\left(\frac{1}{n^{3(\beta-\epsilon)}h^2}\right)$$

and, thus, $A_{n,h}=O\left(\frac{1}{n^{3(\beta-\epsilon)}h^2}\right)$, uniformly in h. Turning to $B_{n,h}$, write

$$\begin{aligned}
&B_{n,h}\\
=\;&\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\sum_{i=2}^{T_n(\mathcal{C})}\mathrm{E}\Big(1\left\{|X_{i-1}-X_{j-1}|\leq h\right\}\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^2\left(\widetilde{\mu}_h(X_{i-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{i-1})\right)\right)^2\Big)\\
&+\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\sum_{i=2}^{T_n(\mathcal{C})}\mathrm{E}\left(1\left\{|X_{i-1}-X_{j-1}|\leq h\right\}\right)\mathrm{var}\left(\widetilde{\mu}_h(X_{j-1})\right)\mathrm{var}\left(\widetilde{\mu}_h(X_{i-1})\right)+\text{cross terms.}
\end{aligned}\tag{44}$$

Now, using the same methods as for Eq. (42), we obtain

$$\begin{aligned}
&\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\sum_{i=1}^{T_n(\mathcal{C})}\mathrm{E}\Big(1\left\{|X_{i-1}-X_{j-1}|\,1_C\leq h\right\}\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^2\left(\widetilde{\mu}_h(X_{i-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{i-1})\right)\right)^2\Big)\\
\leq\;&\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\sum_{i=1}^{T_n(\mathcal{C})}\sqrt{\mathrm{E}\left(1\left\{|X_{i-1}-X_{j-1}|\,1_C\leq h\right\}\right)}\times\\
&\times\sqrt{\mathrm{E}\Big(\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^4\left(\widetilde{\mu}_h(X_{i-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{i-1})\right)\right)^4\Big)}\\
\leq\;&\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\sum_{i=1}^{T_n(\mathcal{C})}\sqrt{\mathrm{E}\left(1\left\{|X_{i-1}-X_{j-1}|\,1_C\leq h\right\}\right)}\times\\
&\times\mathrm{E}\Big(\left(\widetilde{\mu}_h(X_{j-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{j-1})\right)\right)^8\Big)^{1/4}\times\Big(\mathrm{E}\left(\widetilde{\mu}_h(X_{i-1})-\mathrm{E}\left(\widetilde{\mu}_h(X_{i-1})\right)\right)^8\Big)^{1/4}\\
=\;&O\left(\frac{\sqrt{h}}{n^{2(\beta-\epsilon)}h^2}\right),
\end{aligned}$$

since

$$\mathrm{E}\left(1\left\{|X_{i-1}-X_{j-1}|\leq h\right\}1_C\right)=O(h),$$

uniformly in h. Also, as from Eq. (41), we have

$$\begin{aligned}
&\frac{1}{T_n^2(\mathcal{C})}\sum_{j=2}^{T_n(\mathcal{C})}\sum_{i=2}^{T_n(\mathcal{C})}\mathrm{E}\left(1\left\{|X_{i-1}-X_{j-1}|\leq h\right\}\right)\mathrm{var}\left(\widetilde{\mu}_h(X_{j-1})\right)\mathrm{var}\left(\widetilde{\mu}_h(X_{i-1})\right)\\
\leq\;&Ch\sup_{x_1,x_2\in C}\mathrm{var}\left(\widetilde{\mu}_h(x_1)\right)\mathrm{var}\left(\widetilde{\mu}_h(x_2)\right)=O\left(\frac{h}{n^{2(\beta-\epsilon)}h^2}\right).
\end{aligned}$$

Therefore, $B_{n,h} = O\left(\frac{\sqrt{h}}{n^{2(\beta-\epsilon)}h^2}\right)$, uniformly in h. Finally,

$$I^{A2}_{n,h} = \max\left\{O\left(\frac{h^{1/4}}{n^{(\beta-\epsilon)}h}\right), O\left(\frac{1}{n^{\frac{3}{2}(\beta-\epsilon)}h}\right)\right\} = o\left(\frac{1}{n^{(\beta+\epsilon)}h}\right),$$

uniformly in h, if $\beta > 5\epsilon$ and if

$$\frac{\frac{1}{n^{\beta-\epsilon}h^{3/4}}}{\frac{1}{n^{\beta+\epsilon}h}} = n^{2\epsilon}h^{1/4} \to 0.$$

The latter condition becomes

$$n^{2\epsilon}n^{-\frac{1}{4}(\beta-\epsilon)\overline{\eta}} \to 0$$

which, for uniformity, ought to be satisfied in the worst case scenario, i.e., $\overline{\eta} = \frac{1}{5}$. We have that

$$n^{2\epsilon}n^{-\frac{1}{20}(\beta-\epsilon)} \to 0$$

if $\beta > 41\epsilon$, which is, of course, stronger than $\beta > 5\epsilon$, but it always satisfied for a β bounded away from zero. Now, notice that

$$I^B_{n,h} \geq \inf_{x\in\mathcal{C}} \left(\mathrm{E}(\widetilde{\mu}_h(x) - \mu^*(x))\right).$$

By a similar argument as that in Appendix A in Gao et al. (2014), for any $x \in \mathcal{C}$, $x \in \mathcal{C}$,

$$\begin{aligned}
\mathrm{E}(\widetilde{\mu}_h(x) - \mu^*(x)) &= \mathrm{E}\left(\frac{1}{p_s(x)}\left(\frac{1}{T_n h}\sum_{j=1}^{n} K\left(\frac{X_{j-1}-x}{h}\right)(\mu(X_{j-1}) - \mu(x))\right)\right) \\
&= \mathrm{E}_\upsilon\left(\frac{1}{p_s(x)}\left(\frac{1}{h}\sum_{j=0}^{\tau_0} K\left(\frac{X_{j-1}-x}{h}\right)(\mu(X_{j-1}) - \mu(x))\right)\right) \\
&= \frac{1}{p_s(x)}\int K\left(\frac{u-x}{h}\right)(\mu(x-hu) - \mu(x))\,\upsilon G_{s,\upsilon}\mathrm{d}u \\
&= \frac{1}{p_s(x)}\int K(u)\,(\mu(x-hu) - \mu(x))\,p_s(x-hu)\mathrm{d}u,
\end{aligned}$$

where υ is a probability measure and s is a small function so that $P^t \geq s \otimes \upsilon$ for an integer $t \geq 1$. Also, $G_{s,\upsilon} = \sum_{i=0}^{\infty}(P - s\otimes\upsilon)^i$. Thus, uniformly over $h \in H_n$,

$$I^B_{n,h} \geq \inf_{x\in\mathcal{C}} \frac{1}{p_s^2(x)}\left(\int K(u)\,(\mu(x-hu) - \mu(x))\,p_s(x-hu)\mathrm{d}u\right)^2.$$

Now, because $I_{n,h}$ is positive, the cross product ($I^C_{n,h}$) is either positive (in which case the lower bound is simply given by the sum of $I^A_{n,h}$ and $I^B_{n,h}$) or, if negative, it is bounded by the sum of the other two terms. Hence, there is an α with $0 < \alpha \leq 1$, so that

$$\lim_{n\to\infty} \Pr\left\{\varpi : \inf_{h\in H_n} I_{n,h} \geq \alpha \inf_{h\in H_n}(I^A_{n,h} + I^B_{n,h})\right\}^c = 0.$$

Now, we turn to $II_{n,h}$. Given Lemma 1, it is clear that $II_{n,h} = o_p(I_{n,h})$. Finally, consider the cross-product term $III_{n,h}$. By Cauchy-Schwartz's inequality

$$\begin{aligned}
III_{n,h} &= \frac{1}{T_n(\mathcal{C})}\sum_{j=1}^{n}\left(\left(\widehat{\mu}_h\left(X_{j-1}\right)-\mu(X_{j-1})\right)\frac{\left(\widehat{p}_h(X_{j-1})-p_s(X_{j-1})\right)}{p_s(X_{j-1})}\right)\times \\
&\times\left(\left(\widehat{\mu}_h\left(X_{j-1}\right)-\mu(X_{j-1})\right)\frac{\widehat{p}_h(X_{j-1})}{p_s(X_{j-1})}\right)1\left\{X_{j-1}\in\mathcal{C}\right\} \\
&\leq \sqrt{\frac{1}{T_n(\mathcal{C})}\sum_{j=1}^{n}\left(\left(\widehat{\mu}_h\left(X_{j-1}\right)-\mu(X_{j-1})\right)\frac{\left(\widehat{p}_h(X_{j-1})-p_s(X_{j-1})\right)}{p_s(X_{j-1})}\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}}\times \\
&\times\sqrt{\frac{1}{T_n(\mathcal{C})}\sum_{j=1}^{n}\left(\left(\widehat{\mu}_h\left(X_{j-1}\right)-\mu(X_{j-1})\right)\frac{\widehat{p}_h(X_{j-1})}{p_s(X_{j-1})}\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}} \\
&\leq \sqrt{I_{n,h}}\sqrt{II_{n,h}} \\
&= o_p(I_{n,h}).
\end{aligned}$$

Q.E.D.

Proof of Lemma 3. By simple arithmetic,

$$\begin{aligned}
\overline{d}_{A,h}\left(\widehat{\mu},\mu\right) = d_{A,h}\left(\widehat{\mu},\mu\right) &+ \underbrace{\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)^2 1\{X_{j-1}\in\mathcal{C}\}}_{A_n} \\
&+ 2\underbrace{\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)\left(\widehat{\mu}_h(X_{j-1})-\mu(X_{j-1})\right)1\{X_{j-1}\in\mathcal{C}\}}_{B_n}. \qquad (45)
\end{aligned}$$

Therefore, by Lemma 2, we need to show that

$$\sup_{h\in H_n} l_{n,h}^{-1}|A_n+B_n| = o_p\left(1\right), \qquad (46)$$

with $l_{n,h} = \frac{c}{hn^{\beta+\varepsilon}} + \inf_{x\in\mathcal{C}}|b^2_{\mathcal{C}}(x,h)|$. Consider A_n first. Define

$$\widetilde{\mu}_{h,j}(X_{j-1}) = \frac{\frac{1}{T_n}\sum_{i=2,i\neq j}^{n}X_iK_h\left(X_{i-1}-X_{j-1}\right)}{\frac{1}{T_n}\sum_{i=2}^{n}K_h\left(X_{i-1}-X_{j-1}\right)} = \widehat{\mu}_{h,j}(X_{j-1})\frac{\widehat{p}_{h,j}(X_{j-1})}{\widehat{p}_h(X_{j-1})},$$

with $\widehat{p}_{h,j}(x) = \frac{1}{T_n}\sum_{i=2,i\neq j}^{n}K_h\left(X_{i-1}-x\right)$. Note that

$$\begin{aligned}
\left(\widehat{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)^2 &= \left(\widehat{\mu}_{h,j}(X_{j-1})-\widetilde{\mu}_{h,j}(X_{j-1})\right)^2 \\
&+\left(\widetilde{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)^2 \\
&+2\left(\widehat{\mu}_{h,j}(X_{j-1})-\widetilde{\mu}_{h,j}(X_{j-1})\right)\left(\widetilde{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right).
\end{aligned}$$

We first show that

$$\sup_{h\in H_n}\left|\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\widetilde{\mu}_{h,j}(X_{j-1})\right)^2 w(X_{j-1})\left(\widehat{\mu},\mu\right)\right| = o_p\left(T_n^{-1+\overline{\eta}}+T_n^{-2\underline{\eta}}\right). \tag{47}$$

Write,

$$\begin{aligned}\widehat{\mu}_{h,j}(X_{j-1})-\widetilde{\mu}_{h,j}(X_{j-1}) &= \frac{1}{T_n}\sum_{i=2,i\neq j}^{n} X_i K_h\left(X_{i-1}-X_{j-1}\right)\left(\frac{\widehat{p}_h(X_{j-1})-\widehat{p}_{h,j}(X_{j-1})}{\widehat{p}_{h,j}(X_{j-1})\widehat{p}_h(X_{j-1})}\right)\\ &= \frac{1}{T_n h}K(0)\frac{1}{T_n}\sum_{i=2,i\neq j}^{n}\frac{X_i K_h\left(X_{i-1}-X_{j-1}\right)}{\widehat{p}_{h,j}(X_{j-1})\widehat{p}_h(X_{j-1})}.\end{aligned}$$

By Lemma 1 and Assumption 1.2, we have $\sup_{x\in\mathcal{C}}\widehat{p}_{h,j}(X_j)>0$ and $\sup_{x\in\mathcal{C}}\widehat{p}_h(X_j)>0$. Therefore,

$$\begin{aligned}&\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\widetilde{\mu}_{h,j}(X_{j-1})\right)^2 1\{X_{j-1}\in\mathcal{C}\}\\ =\;& \frac{1}{T_n^2h^2}K^2(0)\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{1}{T_n}\sum_{i=2,i\neq j}^{n}\frac{X_iK_h\left(X_{i-1}-X_{j-1}\right)}{\widehat{p}_{h,j}(X_{j-1})\widehat{p}_h(X_{j-1})}\right)^2 1\{X_{j-1}\in\mathcal{C}\}\\ \leq\;& \frac{1}{T_n^2h^2}K^2(0)\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{\mu(X_{j-1})}{p_s(X_j)}\right)^2 1\{X_{j-1}\in\mathcal{C}\}\\ &+\frac{1}{T_n^2h^2}K^2(0)\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(\frac{1}{T_n}\sum_{i=2,i\neq j}^{n}\frac{X_iK_h\left(X_{i-1}-X_{j-1}\right)}{\widehat{p}_{h,j}(X_{j-1})\widehat{p}_h(X_{j-1})}\right)^2-\left(\frac{\mu(X_{j-1})}{p_s(X_j)}\right)^2\right)1\{X_{j-1}\in\mathcal{C}\}\\ =\;& \frac{1}{T_n^2h^2}K^2(0)\left(O_p(1)+o_p(1)\right),\end{aligned}$$

where the last line uses Proposition 3. Eq. (47) follows since, by Lemma 2, $d_{A,h}\left(\widehat{\mu},\mu\right)$ is of larger probability order than $1/(T_nh)^2$, uniformly in $h\in H_n$. Now,

$$\widetilde{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})=\frac{1}{\widehat{p}_h(X_{j-1})}K(0)\frac{1}{T_nh}X_j \tag{48}$$

and so

$$\begin{aligned}&\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\widetilde{\mu}_{h,j}(X_{j-1})\right)\left(\widetilde{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)1\{X_{j-1}\in\mathcal{C}\}\\ =\;& K^2(0)\frac{1}{T_n^2h^2}\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{1}{T_n}\sum_{i=2,i\neq j}^{n}\frac{X_iK_h\left(X_{i-1}-X_{j-1}\right)}{\widehat{p}_{h,j}(X_{j-1})\widehat{p}_h(X_{j-1})}\right)\frac{X_j}{\widehat{p}_h(X_{j-1})}1\{X_{j-1}\in\mathcal{C}\}\\ =\;& o_p\left(d_{A,h}\left(\widehat{\mu},\mu\right)\right),\end{aligned}$$

uniformly in $h\in H_n$, because of, again, Lemma 2, Proposition 3 and the definition of H_n. From Eq. (48)

it is now immediate to see that

$$
\begin{aligned}
& \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widetilde{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)^2 1\{X_{j-1}\in\mathcal{C}\} \\
= \;& K^2(0)\frac{1}{T_n^2h^2}\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\frac{X_j^2}{\widehat{p}_h^2(X_{j-1})}1\{X_{j-1}\in\mathcal{C}\} \\
= \;& o_p\left(d_{A,h}\left(\widehat{\mu},\mu\right)\right)
\end{aligned}
$$

uniformly in $h\in H_n$. Hence, A_n in Eq. (45) is $o_p\left(d_{A,h}\left(\widehat{\mu},\mu\right)\right)$ uniformly in $h\in H_n$, because of Assumption 2.1 and Lemma 2. As for B_n, by Cauchy-Schwartz's inequality,

$$
\begin{aligned}
& \left|\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)\left(\widehat{\mu}_h(X_{j-1})-\mu(X_{j-1})\right)1\{X_{j-1}\in\mathcal{C}\}\right| \\
\leq \;& \left(\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_h(X_{j-1})-\mu(X_{j-1})\right)^2 1\{X_{j-1}\in\mathcal{C}\}\right)^{1/2} \\
& \times\left(\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\widehat{\mu}_h(X_{j-1})\right)^2 1\{X_{j-1}\in\mathcal{C}\}\right)^{1/2} \\
= \;& \sqrt{d_{A,h}\left(\widehat{\mu},\mu\right)}\,o_p\left(\sqrt{d_{A,h}\left(\widehat{\mu},\mu\right)}\right)=o_p\left(d_{A,h}\left(\widehat{\mu},\mu\right)\right)
\end{aligned}
$$

uniformly in $h\in H_n$ so that (46) follows. Q.E.D.

Proof of Lemma 4. For Cross(h) defined as in Eq. (14) and u_j defined as in Eq. (1), write

$$
\begin{aligned}
& \text{Cross}(h) \\
= \;& \frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{h,j}(X_{j-1})-\mu(X_{j-1})\right)\sigma(X_{j-1})u_j 1\{X_{j-1}\in\mathcal{C}\} \\
= \;& \frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{1}{T_n}\sum_{i=2,i\neq j}^{n}K_h\left(X_{i-1}-X_{j-1}\right)\sigma(X_{i-1})u_i\right)\sigma(X_{j-1})u_j\frac{1\{X_{j-1}\in\mathcal{C}\}}{\widehat{p}_{h,j}(X_{j-1})} \\
& +\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{1}{T_n}\sum_{i=2,i\neq j}^{n}K_h\left(X_{i-1}-X_{j-1}\right)\left(\mu(X_{i-1})-\mu(X_{j-1})\right)\right)\sigma(X_{j-1})u_j\frac{1\{X_{j-1}\in\mathcal{C}\}}{\widehat{p}_{h,j}(X_{j-1})} \\
= \;& A_{n,h}+B_{n,h}.
\end{aligned}
$$

Let $\widetilde{A}_{n,h}$ and $\widetilde{B}_{n,h}$ be defined as $A_{n,h}$ and $B_{n,h}$ but with $\widehat{p}_{h,j}(X_{j-1})$ replaced by $p_s(X_{j-1})$. Given Lemma 1, it is enough to show that

$$
\sup_{h\in H_n}\left|\frac{\widetilde{A}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right|=o_p(1) \tag{49}
$$

and

$$\sup_{h\in H_n}\left|\frac{\widetilde{B}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right| = o_p(1), \tag{50}$$

where $h \in H_n$. Using the split chain decomposition,

$$\begin{aligned}
&\widetilde{A}_{n,h} \\
=\ & \frac{T_n}{T_n(\mathcal{C})}\frac{1}{T_n^2}\sum_{k=1}^{T_n}\sum_{k'=1}^{T_n}\Bigg(\frac{1}{h}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\ \sum_{i_{k'}=\tau_{k'-1}+1, i_{k'}\neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{h}\right) \\
&\sigma(X_{i_{k'}-1})u_{i_{k'}}\sigma(X_{j_k-1})u_{j_k}\frac{1\{X_{j_{k-1}}\in\mathcal{C}\}}{p_s(X_{j_{k-1}})}\Bigg)
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{B}_{n,h} \ =\ & \frac{T_n}{T_n(\mathcal{C})}\frac{1}{T_n^2}\sum_{k=1}^{T_n}\sum_{k'=1}^{T_n}\Bigg(\frac{1}{h}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\ \sum_{i_{k'}=\tau_{k'-1}+1, i_{k'}\neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{h}\right) \\
&\big(\mu(X_{i_{k'}-1})-\mu(X_{j_k-1})\big)\,\sigma(X_{j_k-1})u_{j_k}\frac{1\{X_{j_{k-1}}\in\mathcal{C}\}}{p_s(X_{j_{k-1}})}\Bigg).
\end{aligned}$$

We introduce the same subsequence $\{a_n : n \geq 1\}$ as in Proposition 1 and define h, again, on $\tilde{H}_{a_n,n} = \left\{h\in\mathbb{R}: \ \underline{\tilde{h}}_{a_n,n} \leq h \leq \tilde{\overline{h}}_{a_n,n}\right\}$ with $\underline{\tilde{h}}_{a_n,n} = \underline{c}a_n^{-\eta}$ and $\tilde{\overline{h}}_{a_n,n} = \overline{c}a_n^{-\overline{\eta}}$. Define a_n bandwidths h^* partitioning the space $\tilde{H}_{a_n,n}$ so that $|h^*-h| \leq a_n^{-1}(a_n^{-\overline{\eta}} - a_n^{-\eta})$, for every h and at least one h^*. The partition defines balls $H_{a_n,j}$ so that $\tilde{H}_{a_n,n} = \cup_{j=1}^{a_n} H_{a_n,j}$.
We begin by evaluating $\widetilde{A}_{n,h}$. By Boole's and Markov's inequality,

$$\begin{aligned}
&\Pr\left(\sup_{h\in\tilde{H}_{a_n,n}}\left|\frac{\widetilde{A}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right| > \zeta\right) \leq \Pr\left(\max_{1\leq j\leq a_n}\sup_{h\in H_{a_n,j}}\left|\frac{\widetilde{A}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right| > \zeta\right) \\
\leq\ & a_n \Pr\left(\sup_{h\in H_{a_n,j}}\left|\frac{\widetilde{A}_{n,h_j}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right| > \zeta\right) \leq \zeta^{-2\kappa}a_n \mathrm{E}\left(\sup_{h\in H_{a_n,j}}\left|\frac{\widetilde{A}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right|\right)^{2\kappa}.
\end{aligned} \tag{51}$$

Conditioning now on the ω's in $\Omega_{n,h}$, where $\Omega_{n,h}$ was defined in the statement of Lemma 2, we have

$$\begin{aligned}
&\mathrm{E}\left(\sup_{h\in H_{a_n,j}}\left|\frac{\widetilde{A}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right|\right)^{2\kappa} \\
=\ & \mathrm{E}\left(\sup_{h\in H_{a_n,j}}\left|\widetilde{A}_{n,h}\right|\frac{1}{d_{A,h}\left(\widehat{\mu},\mu\right)}1\left\{\omega\in\Omega_{n,h}\right\}\right)^{2\kappa}\Pr\left(\Omega_{n,h}\right) \\
&+\mathrm{E}\left(\sup_{h\in H_{a_n,j}}\left|\widetilde{A}_{n,h}\right|\frac{1}{d_{A,h}\left(\widehat{\mu},\mu\right)}1\left\{\omega\in\Omega_{n,h}^c\right\}\right)^{2\kappa}\Pr\left(\Omega_{n,h}^c\right) \\
=\ & \mathrm{E}\left(\sup_{h\in H_{a_n,j}}\left|\widetilde{A}_{n,h}\right|\frac{1}{d_{A,h}\left(\widehat{\mu},\mu\right)}1\left\{\omega\in\Omega_{n,h}\right\}\right)^{2\kappa}\Pr\left(\Omega_{n,h}\right)\left(1+o(1)\right).
\end{aligned}$$

Let $l_{n,h} = \frac{c}{hn^{\beta+\varepsilon}} + \inf_{x\in\mathcal{C}} |b^2_{\mathcal{C}}(x,h)|$, as in Lemma 2. Thus,

$$
\begin{aligned}
& \mathrm{E}\left(\sup_{h\in H_{n,j}} \left| \widetilde{A}_{n,h} \right| \frac{1}{d_{A,h}\left(\widehat{\mu},\mu\right)} 1\left\{\omega \in \Omega_{n,h}\right\} \right)^{2\kappa} \Pr\left(\Omega_{n,h}\right) \\
\leq \quad & \sup_{h\in H_{n,j}} \left(\frac{1}{l^{2\kappa}_{n,h}} \right) \mathrm{E}\left(\sup_{h\in H_{n,j}} \left| \widetilde{A}_{n,h} \right| \right)^{2\kappa} (1+o(1)).
\end{aligned}
$$

Letting $u^{\sigma}_{i_{k'}} = \sigma(X_{i_{k'}-1})u_{i_{k'}}$ and $u^{\sigma}_{j_k} = \sigma(X_{j_k-1})u_{j_k}$, write

$$
\widetilde{A}_{n,h} = \frac{1}{a_n^2} \sum_{k=1}^{a_n}\sum_{k'=1}^{a_n} \left(\frac{1}{h} \sum_{j_k=\tau_{k-1}+1}^{\tau_k} \sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1} - X_{j_k-1}}{h} \right) u^{\sigma}_{i_{k'}} u^{\sigma}_{j_k} \frac{1\{X_{j_{k-1}} \in \mathcal{C}\}}{p_s(X_{j_{k-1}})} \right).
$$

As in the proof of Lemma 4 in Härdle and Marron (1985), we employ Theorem 2 of Whittle (1960):

$$
\begin{aligned}
& \mathrm{E}\left(\left| \widetilde{A}_{n,h} \right| \right)^{2\kappa} \\
= \quad & \mathrm{E}\left(\mathrm{E}\left(\left| \widetilde{A}_{n,h} \right| \right)^{2\kappa} \Big| X_{i_{k'-1}}, X_{j_k-1}, \tau_k, \tau_{k'}, k, k' \leq a_n \right) \\
\leq \quad & C\mathrm{E}\left(\frac{1}{a_n^4} \sum_{k=1}^{a_n}\sum_{k'=1}^{a_n} \right. \\
& \left. \left(\frac{1}{h} \sum_{j_k=\tau_{k-1}+1}^{\tau_k} \sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1} - X_{j_k-1}}{h} \right) \sigma(X_{i_{k'}-1})\sigma(X_{j_k-1}) \frac{1\{X_{j_{k-1}} \in \mathcal{C}\}}{p_s(X_{j_{k-1}})} \right)^2 \right)^{\kappa} \\
\leq \quad & C a_n^{-4\kappa} h^{-2\kappa} a_n^{2\kappa} \left(\mathrm{E}\left(\sum_{j_k=\tau_{k-1}+1}^{\tau_k} \sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1} - X_{j_k-1}}{h} \right) \sigma(X_{i_{k'}-1})\sigma(X_{j_k-1}) \frac{1\{X_{j_{k-1}} \in \mathcal{C}\}}{p_s(X_{j_{k-1}})} \right)^2 \right)^{\kappa} \\
\leq \quad & C a_n^{-4\kappa} h^{-2\kappa} a_n^{2\kappa} h^{\kappa} \\
\leq \quad & C a_n^{-2\kappa} h^{-\kappa},
\end{aligned}
$$

where the second to last inequality derives from the fact that

$$
\begin{aligned}
& \left(\mathrm{E}\left(\sum_{j_k=\tau_{k-1}+1}^{\tau_k} \sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1} - X_{j_k-1}}{h} \right) \sigma(X_{i_{k'}-1})\sigma(X_{j_k-1}) \frac{1\{X_{j_{k-1}} \in \mathcal{C}\}}{p_s(X_{j_{k-1}})} \right)^2 \right)^{\kappa} \\
= \quad & h^{2\kappa} \left(\mathrm{E}\left(\sum_{j_k=\tau_{k-1}+1}^{\tau_k} \sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}} h^{-1} K\left(\frac{X_{i_{k'}-1} - X_{j_k-1}}{h} \right) \sigma(X_{i_{k'}-1})\sigma(X_{j_k-1}) \frac{1\{X_{j_{k-1}} \in \mathcal{C}\}}{p_s(X_{j_{k-1}})} \right)^2 \right)^{\kappa} \\
\leq \quad & C h^{\kappa},
\end{aligned}
$$

since, by Lemma B1 (page 935), in Gao et al. (2014), $\mathrm{E}\left(\sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}} h^{-1}K\left(\frac{X_{i_{k'}-1}-x}{h}\right)\right)^2 \leq \widetilde{C}h^{-1}$ where $\widetilde{C}$ does not depend on either x or h and since the number of terms in the sum $\sum_{j_k=\tau_{k-1}+1}^{\tau_k}$ is finite almost surely. We also use the fact that $\sigma(x)$ is bounded for $x \in \mathcal{C}$ given Assumption 1.5. Thus,

$$
\begin{aligned}
& \Pr\left(\sup_{h\in H_{a_n,j}}\left|\frac{\widetilde{A}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right| > \zeta\right) \\
\leq\ & \zeta^{-2\kappa}a_n\mathrm{E}\left(\sup_{h\in H_{n,j}}\left|\frac{\widetilde{A}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right|\right)^{2\kappa} \\
\leq\ & \zeta^{-2\kappa}a_n\frac{a_n^{-2\kappa}h^{-\kappa}}{h^{-2\kappa}n^{-2k(\beta+\epsilon)}} \\
\leq\ & \zeta^{-2\kappa}\frac{a_n^{1-2\kappa}h^{\kappa}}{n^{-2\kappa(\beta+\epsilon)}}.
\end{aligned}
$$

Now, because $a_n >> n^{\beta-\epsilon}$, we have that the bound becomes

$$
\zeta^{-2\kappa}\frac{a_n^{1-2\kappa}h^{\kappa}}{n^{-2\kappa(\beta+\epsilon)}} \leq \zeta^{-2\kappa}\frac{n^{(1-2\kappa)(\beta-\epsilon)}\overline{h}^{\kappa}}{n^{-2\kappa(\beta+\epsilon)}} \leq \zeta^{-2\kappa}n^{\beta+(4\kappa-1)\epsilon}\overline{h}^{\kappa} \leq \zeta^{-2\kappa}n^{\beta+(4\kappa-1)\epsilon}n^{-\kappa(\beta-\epsilon)\overline{\eta}} \to 0
$$

for $\kappa > \frac{\beta-\epsilon}{(\beta-\varepsilon)\overline{\eta}-4\varepsilon}$ (which we assume in Assumption 2.2) and $\overline{\eta}$ defined in Eq. (4). The statement in Eq. (49) then follows. We now turn to $\widetilde{B}_{n,h}$. It suffices to show that

$$
\sup_{h\in H_n}\left|\frac{\widetilde{\widetilde{B}}_{n,h}}{d_{A,h}\left(\widehat{\mu},\mu\right)}\right| = o_p(1),
$$

where

$$
\begin{aligned}
\widetilde{\widetilde{B}}_{n,h} \ =\ & \frac{1}{a_n^2}\sum_{k=1}^{a_n}\sum_{k'=1}^{a_n}\left(\frac{1}{h}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{h}\right)\right. \\
& \left.\left(\mu(X_{i_{k'}-1})-\mu(X_{j_k-1})\right)\sigma(X_{j_k-1})u_{j_k}1\{X_{j_k}\in\mathcal{C}\}\right).
\end{aligned}
$$

Using, again, Theorem 2 of Whittle (1960), we have

$$
\begin{aligned}
& \mathrm{E}\left(\left|\widetilde{\widetilde{B}}_{n,h}\right|\right)^{2\kappa} \\
=\ & \mathrm{E}\left(\mathrm{E}\left(\left|\widetilde{\widetilde{B}}_{n,h}\right|\right)^{2\kappa}\left|X_{i_{k'-1}}, X_{j_k-1}, \tau_k, \tau_{k'}, k, k' \leq a_n\right.\right) \\
\leq\ & C\mathrm{E}\left(\frac{1}{a_n^4}\sum_{k=1}^{a_n}\sum_{k'=1}^{a_n}\left(\frac{1}{h}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}}\right.\right. \\
& \left.\left.K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{h}\right)\left(\mu(X_{i_{k'}-1})-\mu(X_{j_k-1})\right)\sigma(X_{j_k-1})1\{X_{j_k}\in\mathcal{C}\}\right)^2\right)^{\kappa}. \qquad (52)
\end{aligned}
$$

The right hand-side of the inequality in Eq. (52) is majorized by

$$
\begin{aligned}
& a_n^{-4\kappa} h^{-2\kappa} \sum_{l=2}^{2\kappa} a_n^l \mathrm{E} \left(\sum_{j_k=\tau_{k-1}+1}^{\tau_k} \sum_{i_{k'}=\tau_{k'-1}+1, i'_k \neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1} - X_{j_k-1}}{h} \right) \right. \\
& \left. \big(\mu(X_{i_{k'}-1}) - \mu(X_{j_k-1})\big) \sigma(X_{j_k-1}) 1\{X_{j_k} \in \mathcal{C}\} \right)^l \\
\leq \; & C a_n^{-2\kappa} h^{-2\kappa} \\
& \left(\mathrm{E} \left(\sum_{j_k=\tau_{k-1}+1}^{\tau_k} \sum_{i_{k'}=\tau_{k'-1}+1, i_{k'} \neq j_k}^{\tau_{k'}} K\left(\frac{X_{i_{k'}-1} - X_{j_k-1}}{h} \right) \sigma(X_{j_k-1}) \frac{1\{X_{j_{k-1}} \in \mathcal{C}\}}{p_s(X_{j_{k-1}})} \right)^2 \right)^{\kappa},
\end{aligned}
$$

where the last inequality follows from the fact that, for $X_{j_k-1} \in \mathcal{C}$, both $\sigma(X_{j_k-1})$ and $\big(\mu(X_{i_{k'}-1}) - \mu(X_{j_k-1})\big)^l$ are bounded. The statement in Eq. (50) derives from the same argument used to prove Eq. (49).

Q.E.D.

Proof of Theorem 1. Define

$$
\begin{aligned}
\widehat{h} &= \arg\min_h \left[CV(h) \right], \\
\overline{h} &= \arg\min_h \left[\overline{CV}(h) \right] = \arg\min_h \left[CV(h) + \frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(X_j - \mu(X_{j-1}) \right)^2 1\left\{ X_{j-1} \in \mathcal{C} \right\} \right].
\end{aligned}
$$

It follows that $\overline{h} \overset{a.s.}{=} \widehat{h}$, since the second term in the argmin does not depend on the bandwidth. Now, write

$$
\begin{aligned}
\left| \frac{\overline{CV}(\overline{h})}{\inf_{h \in H_n} d_{A,h}(\widehat{\mu}, \mu)} - 1 \right| &= \sup_{h \in H_n} \left| \frac{\overline{d}_{A,\overline{h}}(\widehat{\mu}, \mu) - \mathrm{Cross}(\overline{h})}{d_{A,h}(\widehat{\mu}, \mu)} - 1 \right| \\
&\leq \sup_{h \in H_n} \left| \frac{\overline{d}_{A,\overline{h}}(\widehat{\mu}, \mu)}{d_{A,h}(\widehat{\mu}, \mu)} - 1 \right| + \sup_{h \in H_n} \left| \frac{\mathrm{Cross}(\overline{h})}{d_{A,h}(\widehat{\mu}, \mu)} \right| \\
&= o_p(1), \qquad (53)
\end{aligned}
$$

by Lemmas 3 and 4. Finally,

$$
\begin{aligned}
\left| \frac{d_{A,\overline{h}}(\widehat{\mu}, \mu)}{\inf_{h \in H_n} d_{A,h}(\widehat{\mu}, \mu)} - 1 \right| &= \left| \frac{\overline{CV}(\overline{h}) + \mathrm{Cross}(\overline{h}) - \left(\overline{d}_{A,\overline{h}}(\widehat{\mu}, \mu) - d_{A,\overline{h}}(\widehat{\mu}, \mu) \right)}{\inf_{h \in H_n} d_{A,h}(\widehat{\mu}, \mu)} - 1 \right| \\
&\leq \left| \frac{\overline{CV}(\overline{h})}{\inf_{h \in H_n} d_{A,h}(\widehat{\mu}, \mu)} - 1 \right| + \sup_{h \in H_n} \left| \frac{\mathrm{Cross}(\overline{h})}{d_{A,h}(\widehat{\mu}, \mu)} \right| \\
&\quad + \sup_{h \in H_n} \left| \frac{\overline{d}_{A,\overline{h}}(\widehat{\mu}, \mu) - d_{A,\overline{h}}(\widehat{\mu}, \mu)}{d_{A,h}(\widehat{\mu}, \mu)} \right| \\
&= o_p(1),
\end{aligned}
$$

by Eq. (53) and Lemmas 3–4.

Q.E.D.

A.2 Proofs of Section 4

Proof of Lemma 5. Because of Lemma 1, we can write

$$\sup_{x\in\mathcal{C},\,\xi\in\Xi_n}\left|\widetilde{\sigma}^2_\xi(x)\frac{\widehat{p}_\xi(x)}{p_s(x)}-\widetilde{\sigma}^2_\xi(x)\right|=o_{a.s.}(1)$$

and, thus, we can prove the statement in the lemma by replacing $\widetilde{\sigma}^2_\xi(x)$ with $\widetilde{\widetilde{\sigma}}^2_\xi(x)=\widetilde{\sigma}^2_\xi(x)\frac{\widehat{p}_\xi(x)}{p_s(x)}$ and $\sigma^2(x)$ with $\sigma^{2*}(x)=\sigma^2(x)\frac{\widehat{p}_\xi(x)}{p_s(x)}$. The result then follows from an analogous argument as that in the proof of Lemma 2. Q.E.D.

Proof of Lemma 6. Let

$$d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)=\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^n\left(\widetilde{\sigma}^2_\xi(X_{j-1})-\sigma^2(X_{j-1})\right)1\left\{X_{j-1}\in\mathcal{C}\right\},$$

$$\overline{d}_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)=\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^n\left(\widetilde{\sigma}^2_{\xi,j}(X_{j-1})-\sigma^2(X_{j-1})\right)1\left\{X_{j-1}\in\mathcal{C}\right\}.$$

By a similar argument as that in the proof of Lemma 3,

$$\sup_{\xi\in\Xi_n}\left|\frac{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)-\overline{d}_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\right|=o_p(1). \tag{54}$$

Now,

$$\begin{aligned}\widetilde{\mathrm{CV}}(\xi)&=\overline{d}_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)-\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^n\left(\sigma^2(X_{j-1})\left(u_j^2-1\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}\\&\quad-\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^n\left(\widetilde{\sigma}^2_{j,\xi}(X_{j-1})-\sigma^2(X_{j-1})\right)\left(\sigma^2(X_{j-1})\left(u_j^2-1\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\}\\&=\overline{d}_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)-\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^n\left(\sigma^2(X_{j-1})\left(u_j^2-1\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}+\widetilde{\mathrm{Cross}}(\xi),\end{aligned} \tag{55}$$

and, by a similar argument as that in the proof of Lemma 4,

$$\sup_{\xi\in\Xi_n}\left|\frac{\widetilde{\mathrm{Cross}}(\xi)}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\right|=o_p(1).$$

The statement then follows from the proof of Theorem 1. Q.E.D.

Proof of Lemma 7. By the triangle inequality,

$$\sup_{x\in\mathcal{C}}\left|\widehat{\mu}_{\widehat{h}_n}(x)-\mu(x)\right|^4\le C\left(\sup_{x\in\mathcal{C}}\left|\widehat{\mu}_{\widehat{h}_n}(x)-\mathrm{E}\left(\widehat{\mu}_{\widehat{h}_n}(x)\right)\right|^4+\sup_{x\in\mathcal{C}}\left|\mathrm{E}\left(\widehat{\mu}_{\widehat{h}_n}(x)\right)-\mu(x)\right|^4\right).$$

Following the proof of Lemma 2, we have

$$\left|\mathrm{E}\left(\widehat{\mu}_{\widehat{h}_n}(x)-\mu(x)\right)\right|=\left|\frac{1}{p_s(x)}\int K(u)\left(\mu(x-\widehat{h}_nu)-\mu(x)\right)p_s(x-\widehat{h}_nu)\mathrm{d}u\right|.$$

Finally, using the limiting orders in Lemma 3.4 of Karlsen and Tjøstheim (2001),

$$\sup_{x\in\mathcal{C}}\left|\widehat{\mu}_{\widehat{h}_n}(x)-\mathrm{E}\left(\widehat{\mu}_{\widehat{h}_n}(x)\right)\right|^4=O_p\left(\frac{1}{n^{2(\beta-\varepsilon)}\widehat{h}_n^2}\right).$$

Then, by Lemma 2, we have

$$\sup_{x\in\mathcal{C}}\left|\widehat{\mu}_{\widehat{h}_n}(x)-\mu(x)\right|^4=O\left(\sup_{x\in\mathcal{C}}b_{\mathcal{C}}^4(x,\widehat{h}_n)\right)+O_p\left(\frac{1}{n^{2(\beta-\varepsilon)}\widehat{h}_n^2}\right)=o_p\left(d_{A,\widehat{h}_n}\left(\widehat{\mu},\mu\right)\right).$$

Q.E.D.

Proof of Lemma 8. We need to show that the terms in Eqs. (17)-(23) are $o_p\left(d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)\right)$ uniformly in ξ. We begin with Eq. (17). Because of Lemma 1, we replace the density estimator in the denominators with the true density. Hence, uniformly in $\xi\in\Xi_n$, we have

$$\begin{aligned}&\widehat{\sigma}^2_{j,\xi}(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\\ =&\left(\frac{1}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)^2\right.\\ &\left.-\frac{2}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\sigma(X_{i-1})u_i\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)\right)(1+o_{a.s.}(1))\end{aligned}\qquad(56)$$

Neglecting the smaller order term,

$$\begin{aligned}&\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\sigma}^2_{j,\xi}\left(X_{j-1}\right)-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}\\ =&\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{1}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)^2\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}\\ &+\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{2}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\sigma(X_{i-1})u_i\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}\\ &-\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{1}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)^2\right.\\ &\left.\times\frac{2}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\sigma(X_{i-1})u_i\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\}\\ =&A1_{n,\xi}+B1_{n,\xi}+C1_{n,\xi}.\end{aligned}$$

Now, write

$$A1_{n,\xi}\leq\sup_{x\in\mathcal{C}}\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)^4 O_p(1)=o_p\left(d_{A,\widehat{h}_n}\left(\widehat{\mu},\mu\right)\right)=o_p\left(\inf_{\xi\in\Xi_n}d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)\right),$$

where the first equality follows from Lemma 7 and the last equality follows from Lemma 5. As for $B1_{n,\xi}$, write

$$
\begin{aligned}
B1_{n,\xi} = & \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{2}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\sigma(X_{i-1})u_i\right. \\
& \left.\times\frac{1}{T_n\widehat{h}_n}\sum_{k=1}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{k-1}-X_{i-1}}{\widehat{h}_n}\right)\sigma(X_{k-1})u_k\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
& +\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\frac{2}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\sigma(X_{i-1})u_i\right. \\
& \left.\times\frac{1}{T_n\widehat{h}_n}\sum_{k=1}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{k-1}-X_{i-1}}{\widehat{h}_n}\right)\left(\mu(X_{k-1})-\mu(X_{i-1})\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\
& +\text{cross term.}
\end{aligned}
$$

Now, $B1_{n,\xi}$ is of smaller probability order than Cross(h) as defined at the beginning of the proof of Lemma 4. Hence, it is $o_p\left(\inf_{h\in H_n} d_{A,h}\left(\widehat{\mu},\mu\right)\right) = o_p\left(\inf_{\xi\in\Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)\right)$. Also, $C1_{n,\xi} = o_p\left(\inf_{\xi\in\Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)\right)$ by Cauchy-Schwartz inequality. This proves that the term in Eq. (17) is $o_p\left(\inf_{\xi\in\Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)\right)$. As for the term in Eq. (18), given Lemma 7, we have

$$
\begin{aligned}
& \frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\left(\sigma^2(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)^2\right)1\left\{X_{j-1}\in\mathcal{C}\right\} \\
\leq \quad & \left(\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\}\right)^{1/2} \\
& \times\left(\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\widehat{\mu}_{\widehat{h}_n}(X_{j-1})-\mu(X_{j-1})\right)^4 1\left\{X_{j-1}\in\mathcal{C}\right\}\right)^{1/2} \\
= \quad & O_p\left(\sqrt{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\right)\sup_{x\in\mathcal{C}}\left(\widehat{\mu}_{\widehat{h}_n}(x)-\mu(x)\right)^2 \\
= \quad & o_p\left(d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)\right).
\end{aligned}
$$

Turning to Eq. (20),

$$
\begin{aligned}
& \frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\left(\sigma^2(X_{j-1}) - \widetilde{\sigma}^2_{j,\xi}(X_{j-1}) \right) \left(\widehat{\sigma}^2_{j,\xi} - \widetilde{\sigma}^2_{j,\xi} \right) \right) 1\left\{ X_{j-1} \in \mathcal{C} \right\} \\
\leq \quad & \left(\frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\sigma^2(X_{j-1}) - \widetilde{\sigma}^2_{j,\xi}(X_{j-1}) \right)^2 1\left\{ X_{j-1} \in \mathcal{C} \right\} \right)^{1/2} \\
& \times \left(\frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\widehat{\sigma}^2_{j,\xi} - \widetilde{\sigma}^2_{j,\xi} \right)^2 1\left\{ X_{j-1} \in \mathcal{C} \right\} \right)^{1/2} \\
= \quad & \left(\overline{d}_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right) \right)^{1/2} o_p\left(\sqrt{d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right)} \right) \\
= \quad & O_p\left(d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right)^{1/2} \right) o_p\left(\sqrt{d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right)} \right),
\end{aligned}
$$

where the last equality derives from Eq. (54). As for Eq. (22), recalling Lemma 7, we obtain

$$
\begin{aligned}
& \frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\left(\widehat{\sigma}^2_{j,\xi} - \widetilde{\sigma}^2_{j,\xi} \right) \left(\widehat{\mu}_{j,\widehat{h}_n}(X_{j-1}) - \mu(X_{j-1}) \right)^2 \right) 1\left\{ X_{j-1} \in \mathcal{C} \right\} \\
& \leq \left(\frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\widehat{\sigma}^2_{j,\xi} - \widetilde{\sigma}^2_{j,\xi} \right)^2 1\left\{ X_{j-1} \in \mathcal{C} \right\} \right)^{1/2} \\
& \times \left(\frac{1}{T_n(\mathcal{C})} \sum_{j=2}^{n} \left(\widehat{\mu}_{j,\widehat{h}_n}(X_{j-1}) - \mu(X_{j-1}) \right)^4 1\left\{ X_{j-1} \in \mathcal{C} \right\} \right)^{1/2} \\
& = o_p\left(\sqrt{d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right)} \right) o_p\left(\sqrt{d_{A,\widehat{h}_n}\left(\widehat{\mu}, \mu \right)} \right) \\
& = o_p\left(\sqrt{d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right)} \right) o_p\left(\sqrt{\inf_{\xi \in \Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right)} \right) \\
& = o_p\left(d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2 \right) \right).
\end{aligned}
$$

We now turn to Eq. (21), which, given Eq. (56), writes as

$$
\begin{aligned}
& \frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})(u_j^2-1)\left(\widehat{\sigma}^2_{j,\xi}(X_{j-1})-\widetilde{\sigma}^2_{j,\xi}(X_{j-1})\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\} \\
= & \frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})(u_j^2-1)\right. \\
& \left.\left(\frac{1}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)^2\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\} \\
& +\frac{2}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})(u_j^2-1)\frac{1}{T_n\xi}\sum_{i\neq j}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\right. \\
& \left.\sigma(X_{i-1})u_i\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\} \\
= & I_{n,\xi}+II_{n,\xi}.
\end{aligned}
$$

Now, using the split chain decomposition, we have

$$
\begin{aligned}
II_{n,\xi} = & \frac{2}{T_n(\mathcal{C})}\sum_{i\neq j}^{n}\left(\frac{1}{T_n\xi}\sum_{j=2}^{n}\frac{1}{p_s(X_{j-1})}K\left(\frac{X_{i-1}-X_{j-1}}{\xi}\right)\right. \\
& \left.\sigma^2(X_{j-1})(u_j^2-1)\sigma(X_{i-1})u_i\left(\widehat{\mu}_{\widehat{h}_n}(X_{i-1})-\mu(X_{i-1})\right)1\left\{X_{j-1}\in\mathcal{C}\right\}\right) \\
= & \frac{2T_n}{T_n(\mathcal{C})}\frac{1}{T_n^2}\sum_{k=1}^{T_n}\sum_{k'=1}^{T_n}\left(\frac{1}{\xi}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\sum_{i_{k'}=\tau_{k'-1}+1,i_{k'}\neq j_k}^{\tau_{k'}}\frac{1}{p_s(X_{j_k-1})}K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{\xi}\right)\right. \\
& \left.\sigma^2(X_{j_k-1})(u_{j_k}^2-1)\sigma(X_{i_{k'}-1})u_{i_{k'}}\left(\widehat{\mu}_{\widehat{h}_n}\left(X_{i_{k'}-1}\right)-\mu\left(X_{i_{k'}-1}\right)\right)1\left\{X_{j_k-1}\in\mathcal{C}\right\}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& \sup_{\xi\in\Xi_n}\mathrm{E}\left(\left(\frac{II_{n,\xi}}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\right)^{2\kappa}\right) \\
\leq & \sup_{\xi\in\Xi_n}\mathrm{E}\left(\frac{\sup_{x\in\mathcal{C}}\left(\widehat{\mu}_{\widehat{h}_n}(x)-\mu(x)\right)}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\left(\frac{2T_n}{T_n(\mathcal{C})}\frac{1}{T_n^2}\sum_{k=1}^{T_n}\sum_{k'=1}^{T_n}\left(\frac{1}{\xi}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\sum_{i_{k'}=\tau_{k'-1}+1,i'_k\neq j_k}^{\tau_{k'}}\frac{1}{p_s(X_{j_k-1})}\right.\right.\right. \\
& \left.\left.\left.K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{\xi}\right)\sigma^2(X_{j-1})(u_j^2-1)\sigma(X_{i-1})u_i\right)\right)1\left\{X_{j-1}\in\mathcal{C}\right\}\right)^{2\kappa} \\
= & o_p(1),
\end{aligned}
$$

since the term

$$
\begin{aligned}
& \frac{2T_n}{T_n(\mathcal{C})}\frac{1}{T_n^2}\sum_{k=1}^{T_n}\sum_{k'=1}^{T_n}\frac{1}{\xi}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\sum_{i_{k'}=\tau_{k'-1}+1,i'_k\neq j_k}^{\tau_{k'}}\frac{1}{p_s(X_{j_k-1})}K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{\xi}\right)\times \\
& \times\sigma^2(X_{j-1})(u_j^2-1)\sigma(X_{i-1})u_i
\end{aligned}
$$

is of the same order as $A_{n,h}$ in the proof of that Lemma 4 and

$$\sup_{x\in\mathcal{C}}\left(\widehat{\mu}_{\widehat{h}_n}(x)-\mu(x)\right)=o_p\left(\left(d_{A,\widehat{h}_n}\left(\widehat{\mu},\mu\right)\right)^{1/4}\right)$$

by Lemma 7. As for $I_{n,\xi}$, we have

$$\begin{aligned} I_{n,\xi} &= \frac{2T_n}{T_n(\mathcal{C})}\frac{1}{T_n^2}\sum_{k=1}^{T_n}\sum_{k'=1}^{T_n}\left(\frac{1}{\xi}\sum_{j_k=\tau_{k-1}+1}^{\tau_k}\sum_{i_{k'}=\tau_{k'-1}+1, i'_k\neq j_k}^{\tau_{k'}}\frac{1}{p_s(X_{j_k-1})}K\left(\frac{X_{i_{k'}-1}-X_{j_k-1}}{\xi}\right)\right. \\ &\left.\sigma^2(X_{j_k})(u_{j_k}^2-1)\left(\widehat{\mu}_{\widehat{h}_n}(X_{i_{k'}-1})-\mu(X_{i_{k'}-1})\right)^2 1\left\{X_{j_k-1}\in\mathcal{C}\right\}\right)=o_p(II_{n,\xi}) \end{aligned}$$

because of Lemma 7. The terms in Eq. (19) and in Eq. (23) are $o_p\left(d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)\right)$ by a similar argument.

Q.E.D.

Proof of Theorem 2.

$$\begin{aligned} \overline{CV}(\xi) &= \widetilde{CV}(\xi)+\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})\left(u_j^2-1\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\ &= CV(\xi)+\widehat{\text{Error}}(\xi)+\widehat{\text{Error}}+\frac{1}{T_n(\mathcal{C})}\sum_{j=2}^{n}\left(\sigma^2(X_{j-1})\left(u_j^2-1\right)\right)^2 1\left\{X_{j-1}\in\mathcal{C}\right\} \\ &= \overline{\overline{CV}}(\xi)+\widehat{\text{Error}}(\xi), \end{aligned}$$

where $\widehat{\text{Error}}$ denotes the first five terms in Eq. (16). We can write

$$\bar{\xi}=\arg\min_{\xi}\overline{\overline{CV}}(\xi)\overset{a.s.}{=}\arg\min_{\xi}CV(\xi).$$

For $\widetilde{\text{Cross}}(\xi)$ defined as in Eq. (55), we have

$$\begin{aligned} \left|\frac{\overline{\overline{CV}}(\bar{\xi})}{\inf_{\xi\in\Xi_n}d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}-1\right| &= \sup_{\xi\in\Xi_n}\left|\frac{\bar{d}_{A,\bar{\xi}}\left(\widetilde{\sigma}^2,\sigma^2\right)+\widehat{\text{Error}}(\bar{\xi})-\widetilde{\text{Cross}}(\bar{\xi})}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}-1\right| \\ &\leq \sup_{\xi\in\Xi_n}\left|\frac{\bar{d}_{A,\bar{\xi}}\left(\widetilde{\sigma}^2,\sigma^2\right)}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}-1\right|+\sup_{\xi\in\Xi_n}\left|\frac{\widetilde{\text{Cross}}(\bar{\xi})}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\right|+\sup_{\xi\in\Xi_n}\left|\frac{\widehat{\text{Error}}(\bar{\xi})}{d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}\right| \\ &= o_p(1) \end{aligned} \tag{57}$$

by the results in the proofs of Lemma 6 and Lemma 8. Now, write

$$\begin{aligned} &\left|\frac{d_{A,\bar{\xi}}\left(\widetilde{\sigma}^2,\sigma^2\right)}{\inf_{\xi\in\Xi_n}d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}-1\right| \\ &=\left|\frac{\overline{CV}(\bar{\xi})+\widetilde{\text{Cross}}(\bar{\xi})-\left(\bar{d}_{A,\bar{\xi}}\left(\widetilde{\sigma}^2,\sigma^2\right)-d_{A,\bar{\xi}}\left(\widetilde{\sigma}^2,\sigma^2\right)\right)}{\inf_{\xi\in\Xi_n}d_{A,\xi}\left(\widetilde{\sigma}^2,\sigma^2\right)}-1\right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \frac{\overline{\overline{CV}}(\bar{\xi}) - \widehat{\text{Error}}(\bar{\xi}) + \widetilde{\text{Cross}}(\bar{\xi}) - \left(\bar{d}_{A,\bar{\xi}}\left(\widetilde{\sigma}^2, \sigma^2\right) - d_{A,\bar{\xi}}\left(\widetilde{\sigma}^2, \sigma^2\right) \right)}{\inf_{\xi \in \Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2\right)} - 1 \right| \\
&\leq \left| \frac{\overline{\overline{CV}}(\bar{\xi})}{\inf_{\xi \in \Xi_n} d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2\right)} - 1 \right| + \sup_{\xi \in \Xi_n} \left| \frac{\widehat{\text{Error}}(\bar{\xi})}{d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2\right)} \right| \\
&\quad + \sup_{\xi \in \Xi_n} \left| \frac{\widetilde{\text{Cross}}(\bar{\xi})}{d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2\right)} \right| + \sup_{\xi \in \Xi_n} \left| \frac{\bar{d}_{A,\bar{\xi}}\left(\widetilde{\sigma}^2, \sigma^2\right) - d_{A,\bar{\xi}}\left(\widetilde{\sigma}^2, \sigma^2\right)}{d_{A,\xi}\left(\widetilde{\sigma}^2, \sigma^2\right)} \right| \\
&= o_p(1),
\end{aligned}
$$

by Eq. (57). Q.E.D.

References

F. M. Bandi. Short-term interest rate dynamics: a spatial approach. *Journal of Financial Economics*, 65(1):73–110, 2002.

M. W. Brandt. Portfolio choice problems. In Y. Aït-Sahalia and L. P. Hansen, editors, *Handbook of Financial Econometrics: Tools and Techniques*, volume 1, chapter 5, pages 269–336. North-Holland, San Diego, 2010.

J. Campbell and R. Shiller. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1(3):195–228, 1988.

C. L. Cavanagh, G. Elliott, and J. H. Stock. Inference in models with nearly integrated regressors. *Econometric Theory*, 11:1131–1147, 10 1995.

P. D. Feigin and R. L. Tweedie. Random coefficient autoregressive processes: a markov chain analysis of stationarity and finiteness of moments. *Journal of Time Series Analysis*, 6(1):1–14, 1985.

J. Fleming, C. Kirby, and B. Ostdiek. The economic value of volatility timing. *The Journal of Finance*, 56(1):329–352, 2001.

J. Gao, S. Kanaya, D. Li, and D. Tjøstheim. Uniform consistency for nonparametric estimators in null recurrent time series. *Econometric Theory*, forthcoming, 2014.

E. Guerre. Design-adaptive pointwise nonparametric regression estimation for recurrent markov time series. Technical report, 2004.

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

W. Härdle. Approximations to the mean integrated squared error with applications to optimal bandwidth selection for nonparametric regression function estimators. *Journal of Multivariate Analysis*, 18(1):150–168, 2 1986.

W. Härdle and J. S. Marron. Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, 13(4):1465–1481, 1985.

W. Härdle and P. Vieu. Kernel regression smoothing of time series. *Journal of Time Series Analysis*, 13(3):209–232, 1992.

H. A. Karlsen and D. Tjøstheim. Nonparametric estimation in null recurrent time series. Technical report, 1998.

H. A. Karlsen and D. Tjøstheim. Nonparametric estimation in null recurrent time series. *The Annals of Statistics*, 29(2):372–416, 2001.

T. Y. Kim and D. D. Cox. Bandwidth selection in kernel smoothing of time series. *Journal of Time Series Analysis*, 17(1):49–63, 1996.

D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1):159–178, 1992.

J. Lewellen. Predicting returns with financial ratios. *Journal of Financial Economics*, 74 (2):209–235, 2004.

G. Moloche. Kernel regression for nonstationary harris-recurrent processes. Technical report, 2001.

D. Nolan and D. Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15(2):780–799, 1987.

E. Nummelin. *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, 1984.

D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

P. M. Robinson. Nonparametric estimators of times series. *Journal of Time Series Analysis*, 4(3):185–207, 1983.

M. Schienle. Nonparametric nonstationary regression with many covariates. Technical report, 2010.

R. F. Stambaugh. Bias in regressions with lagged stochastic regressors. Working paper, University of Chicago, 1986.

R. F. Stambaugh. Predictive regressions. *Journal of Financial Economics*, 54(3):375–421, 1999.

W. Torous and R. Valkanov. Boundaries of predictability: Noisy predictive regressions. Working paper, UCLA, 2000.

R. Valkanov. Long-horizon regressions: theoretical results and applications. *Journal of Financial Economics*, 68(2):201–232, 2003.

Q. Wang and P. C. B. Phillips. Structural nonparametric cointegrating regression. *Econometrica*, 77(6):1901–1948, 2009.

P. Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability & Its Applications*, 5(3):302–305, 1960.

Y. Xia and W. Li. Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of Multivariate Analysis*, 83(2):265 – 287, 2002.

S. Yakowitz. Nonparametric density and regression estimation for markov sequences without mixing assumptions. *Journal of Multivariate Analysis*, (30):124–136, 1989.

| | | linear autoregression | | | | | |
| | | nonparametric | | | OLS | | |
T	ρ	bias	stdev	RMSE	bias	stdev	RMSE
200	0.1	0.161	0.249	0.203	0.014	0.174	0.121
	0.5	0.215	0.323	0.264	0.003	0.153	0.109
	0.9	0.157	0.297	0.199	0.009	0.088	0.054
	0.99	0.104	0.822	0.267	0.012	0.050	0.021
	1	0.150	1.185	0.372	0.010	0.045	0.015
500	0.1	0.126	0.186	0.151	0.001	0.113	0.078
	0.5	0.150	0.227	0.186	0.001	0.096	0.068
	0.9	0.095	0.159	0.118	0.005	0.051	0.033
	0.99	0.048	0.387	0.156	0.004	0.022	0.011
	1	0.072	0.980	0.252	0.004	0.016	0.006
$1,000$	0.1	0.095	0.142	0.118	0.004	0.081	0.057
	0.5	0.108	0.174	0.137	0.001	0.070	0.047
	0.9	0.065	0.114	0.086	0.004	0.035	0.023
	0.99	0.025	0.167	0.105	0.002	0.013	0.008
	1	0.046	0.842	0.200	0.002	0.008	0.003

Table 1: Linear autoregression: bias, standard deviation ("stdev"), and root mean-square error ("RMSE") of estimates of $\mu(x)$, averaged over a grid x-values. "CV" denotes the nonparametric estimator using the cross-validated bandwidth and "OLS" the standard linear least-squares estimator.

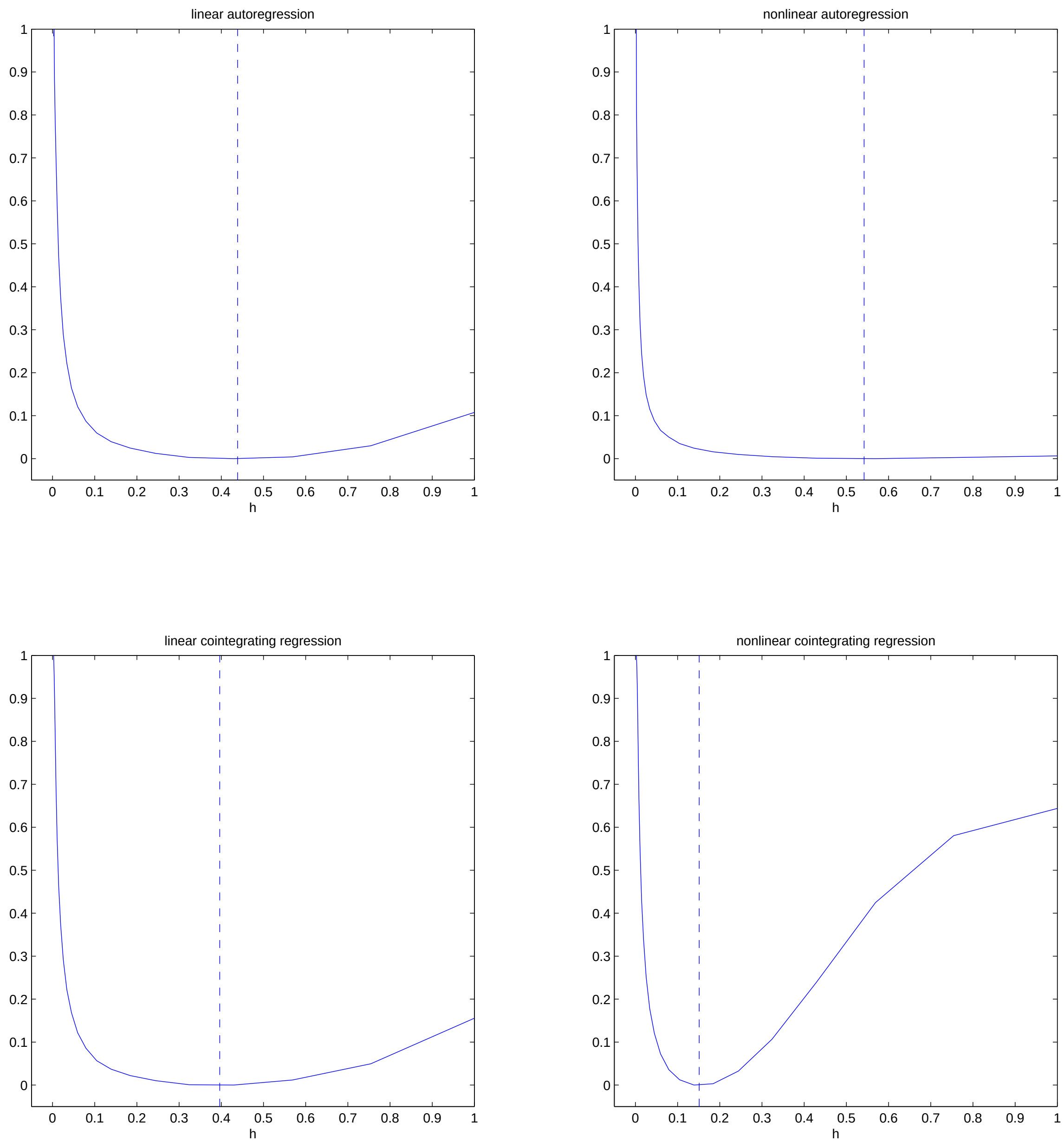

Figure 1: Cross-validation objective function (solid line) together with the selected bandwidth (dashed line), both averaged over all Monte Carlo samples, for $T = 500$ and $\rho = 0.9$.

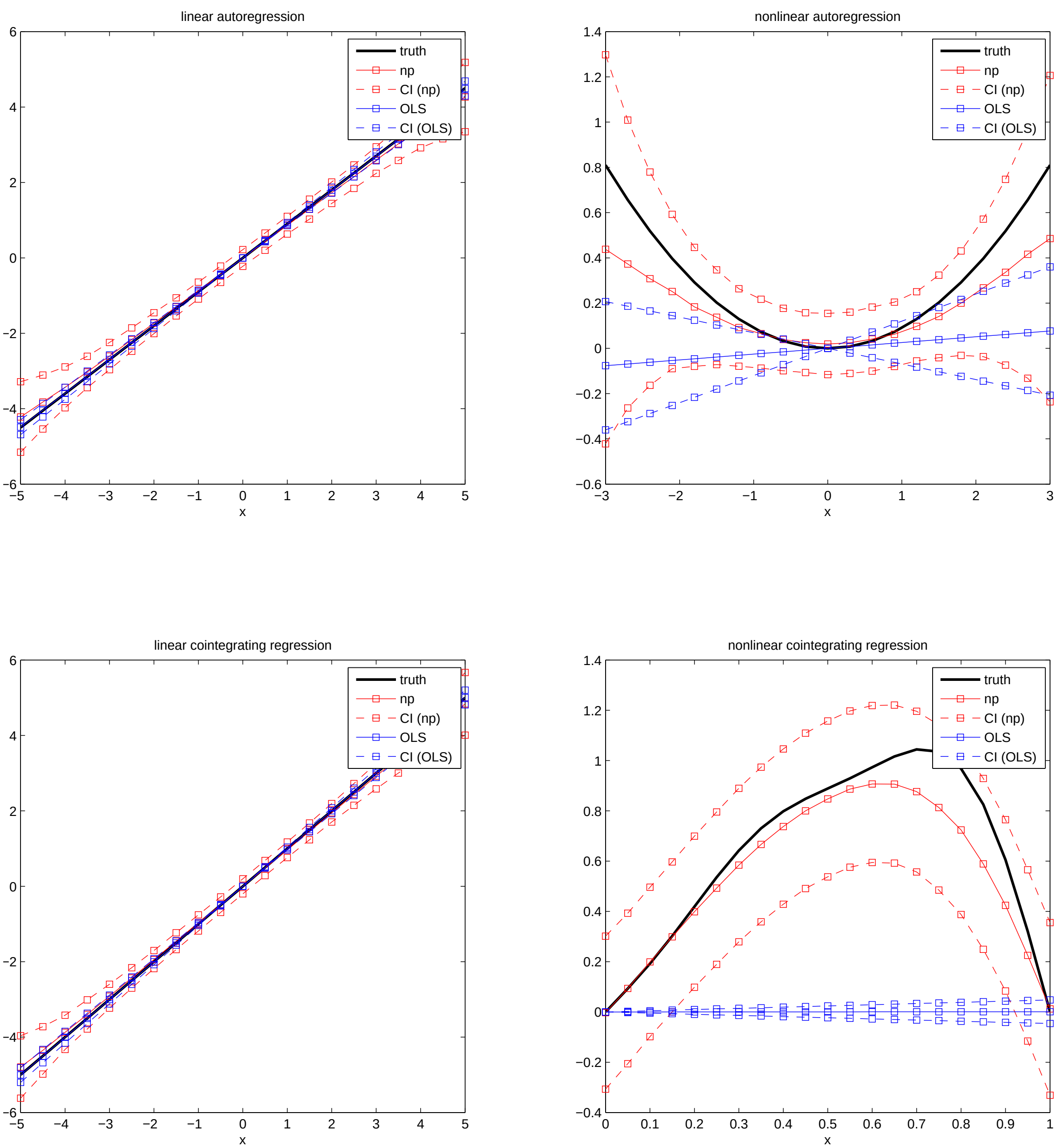

Figure 2: Estimated regression functions together with pointwise empirical confidence intervals. "np" refers to the nonparametric estimator with the cross-validated bandwidth and "OLS" to the linear least-squares estimator, each averaged over all Monte Carlo samples, for $T = 500$ and $\rho = 0.9$.

		nonlinear autoregression					
		nonparametric			OLS		
T	ρ	bias	stdev	RMSE	bias	stdev	RMSE
200	0.1	0.016	0.114	0.066	0.022	0.105	0.075
	0.5	0.059	0.127	0.087	0.129	0.106	0.130
	0.9	0.095	0.139	0.116	0.243	0.111	0.243
	0.99	0.101	0.142	0.123	0.263	0.113	0.263
	1	0.102	0.143	0.123	0.265	0.113	0.265
500	0.1	0.012	0.074	0.044	0.023	0.068	0.050
	0.5	0.048	0.088	0.067	0.123	0.069	0.123
	0.9	0.065	0.107	0.088	0.241	0.072	0.241
	0.99	0.066	0.111	0.088	0.267	0.073	0.267
	1	0.066	0.111	0.088	0.269	0.073	0.269
$1,000$	0.1	0.014	0.051	0.030	0.024	0.049	0.038
	0.5	0.044	0.066	0.057	0.124	0.049	0.124
	0.9	0.056	0.089	0.069	0.241	0.052	0.241
	0.99	0.057	0.086	0.070	0.266	0.052	0.266
	1	0.057	0.086	0.070	0.268	0.052	0.268

Table 2: Nonlinear autoregression: bias, standard deviation ("stdev"), and root mean-square error ("RMSE") of estimates of $\mu(x)$, averaged over a grid x-values. "CV" denotes the nonparametric estimator using the cross-validated bandwidth and "OLS" the standard linear least-squares estimator.

		linear cointegrating regression					
		nonparametric			OLS		
T	ρ	bias	stdev	RMSE	bias	stdev	RMSE
200	0.1	0.269	0.534	0.397	0.005	0.180	0.128
	0.5	0.219	0.424	0.311	0.002	0.155	0.110
	0.9	0.107	0.242	0.165	0.001	0.084	0.054
	0.99	0.058	0.603	0.197	0.000	0.038	0.023
	1	0.104	0.823	0.262	0.000	0.030	0.016
500	0.1	0.200	0.347	0.281	0.001	0.114	0.079
	0.5	0.143	0.271	0.205	0.000	0.101	0.070
	0.9	0.070	0.147	0.104	0.003	0.050	0.032
	0.99	0.031	0.272	0.123	0.001	0.020	0.012
	1	0.059	0.703	0.180	0.000	0.011	0.006
1,000	0.1	0.147	0.271	0.209	0.005	0.080	0.055
	0.5	0.127	0.197	0.166	0.003	0.067	0.046
	0.9	0.058	0.103	0.082	0.002	0.034	0.024
	0.99	0.017	0.128	0.083	0.002	0.012	0.007
	1	0.033	0.610	0.149	0.000	0.006	0.003

Table 3: Linear cointegrating regression: bias, standard deviation ("stdev"), and root mean-square error ("RMSE") of estimates of $\mu(x)$, averaged over a grid x-values. "CV" denotes the nonparametric estimator using the cross-validated bandwidth and "OLS" the standard linear least-squares estimator.

		nonlinear cointegrating regression					
		nonparametric			OLS		
T	ρ	bias	stdev	RMSE	bias	stdev	RMSE
200	0.1	0.069	0.178	0.132	0.724	0.044	0.724
	0.5	0.070	0.190	0.133	0.728	0.039	0.728
	0.9	0.114	0.234	0.174	0.731	0.021	0.731
	0.99	0.154	0.416	0.263	0.731	0.010	0.731
	1	0.193	0.521	0.324	0.731	0.007	0.731
500	0.1	0.046	0.118	0.084	0.722	0.029	0.722
	0.5	0.051	0.122	0.093	0.729	0.025	0.729
	0.9	0.058	0.158	0.112	0.730	0.012	0.730
	0.99	0.099	0.262	0.181	0.731	0.005	0.731
	1	0.141	0.461	0.257	0.731	0.003	0.731
$1,000$	0.1	0.029	0.091	0.065	0.724	0.019	0.724
	0.5	0.036	0.095	0.070	0.730	0.017	0.730
	0.9	0.046	0.121	0.091	0.731	0.008	0.731
	0.99	0.064	0.194	0.138	0.731	0.003	0.731
	1	0.108	0.418	0.213	0.731	0.001	0.731

Table 4: Nonlinear cointegrating regression: bias, standard deviation ("stdev"), and root mean-square error ("RMSE") of estimates of $\mu(x)$, averaged over a grid x-values. "CV" denotes the nonparametric estimator using the cross-validated bandwidth and "OLS" the standard linear least-squares estimator.

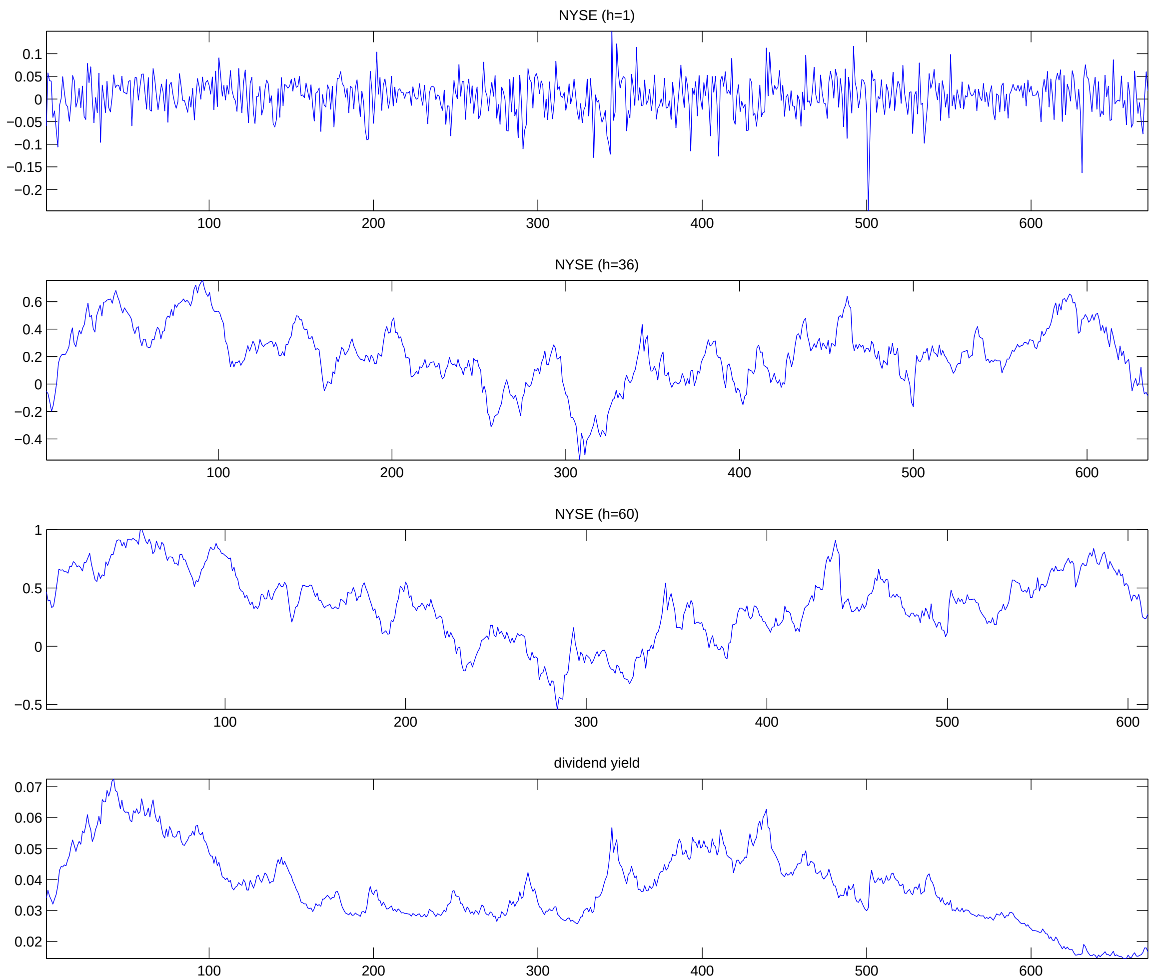

Figure 3: NYSE and dividend yield sample.

	$\gamma = 1$	$\gamma = 5$	$\gamma = 10$
$\tau = 1$	0.0021	0.0012	0.0011
$\tau = 36$	0.0619	0.0343	0.0309
$\tau = 60$	0.0422	0.0179	0.0148

Table 5: Annualized fee Δ.

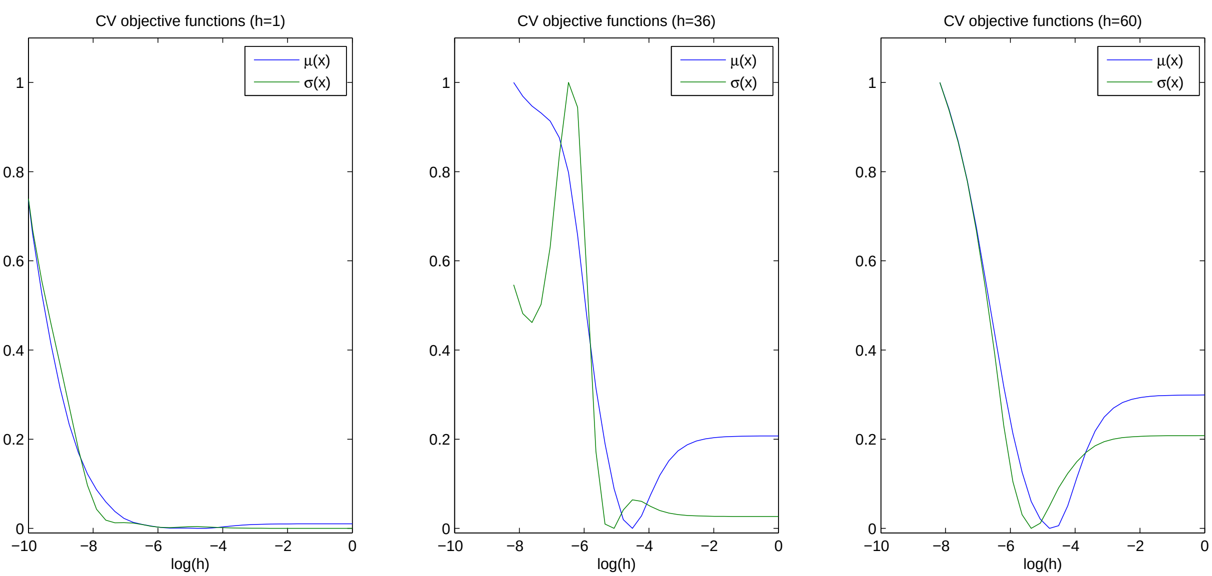

Figure 4: CV criterion functions.

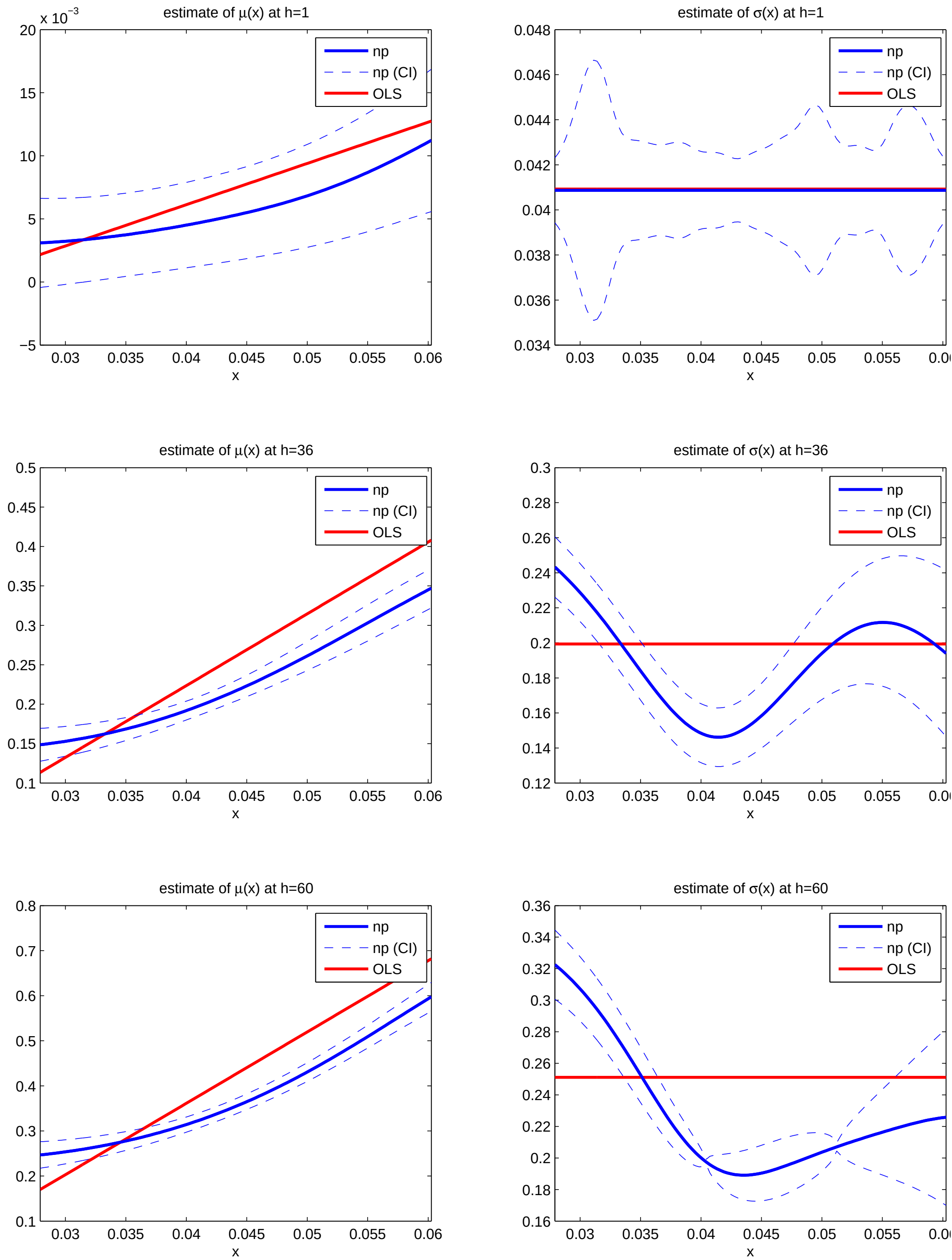

Figure 5: Nonparametric ("np") estimates $\hat{\mu}_{\hat{h}_{1,n}}(x)$ and $\hat{\sigma}_{\hat{h}_{2,n}}(x)$ of $\mu(x)$ and $\sigma(x)$ compared to linear least-squares ("OLS") estimates.

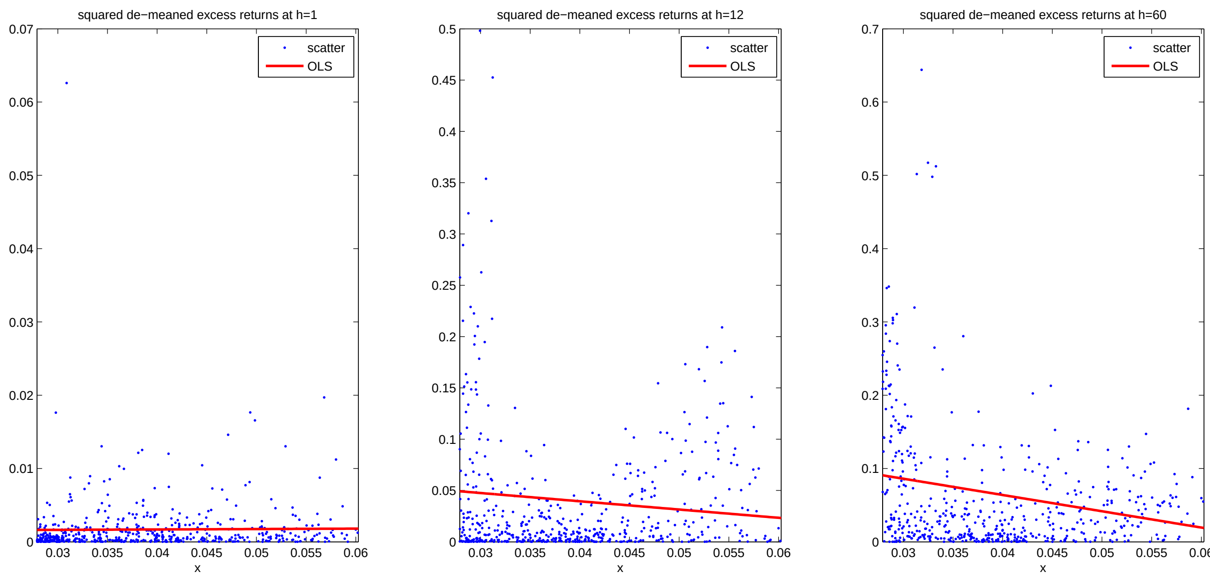

Figure 6: Scatter plots of squared de-meaned excess returns against the dividend-price ratio X_t, together with OLS regression fit.

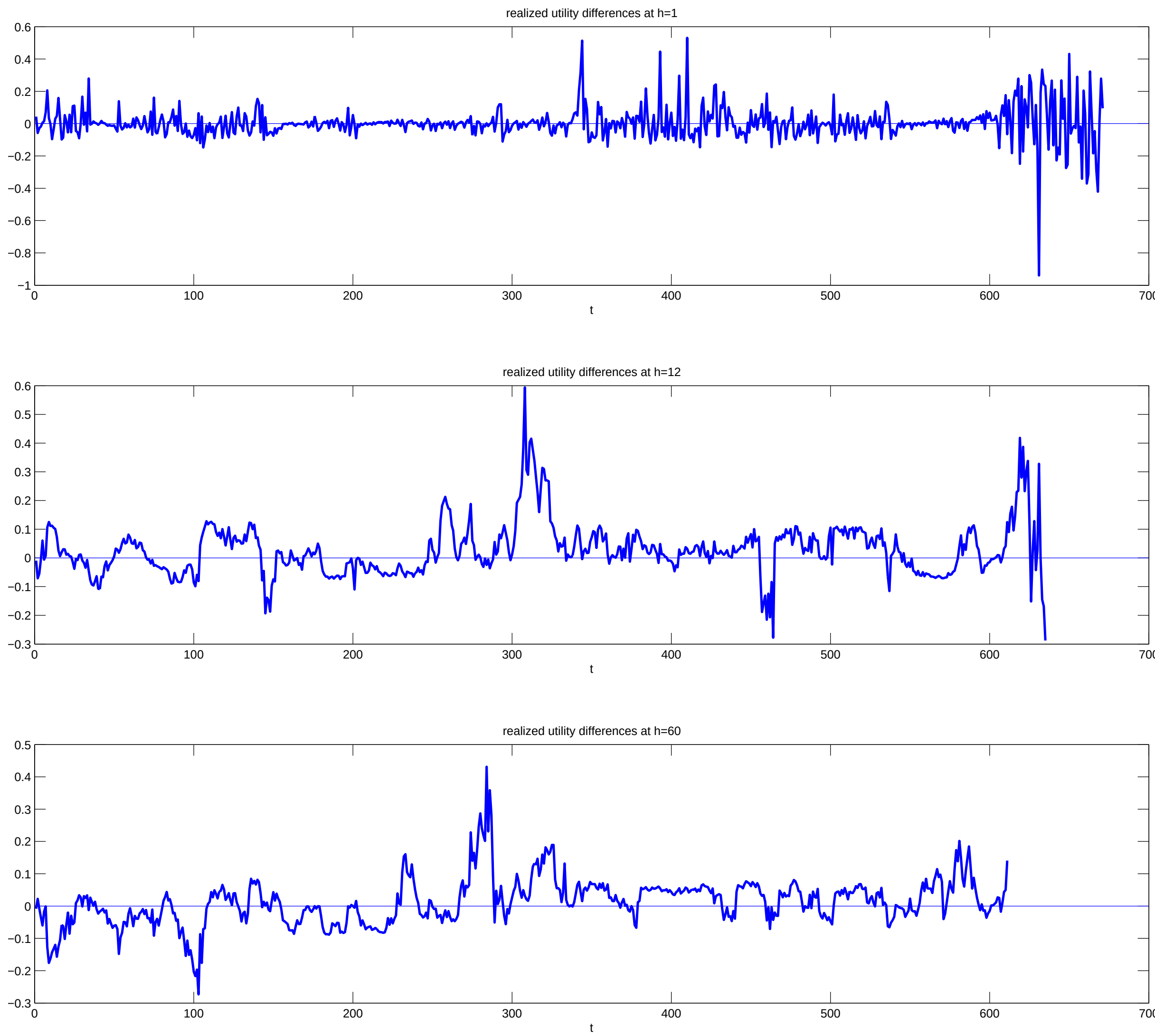

Figure 7: Difference of nonlinear and linear investors' realized utilities over time for $\gamma = 10$.