Statistical Inference for Stochastic Processes manuscript No.
(will be inserted by the editor)

Pre-averaged kernel estimators for the drift function of a diffusion process in the presence of microstructure noise

Wooyong Lee · Priscilla E. Greenwood · Nancy Heckman · Wolfgang Wefelmeyer

November 23 2015/ Received: xxxx / Accepted: xxxx

Abstract We consider estimation of the drift function of a stationary diffusion process when we observe high-frequency data with microstructure noise over a long time interval. We propose to estimate the drift function at a point by a Nadaraya–Watson estimator that uses observations that have been pre-averaged to reduce the noise. We give conditions under which our estimator is consistent and asympotically normal. Its rate and asymptotic bias and variance are the same as those without microstructure noise. To use our method in data analysis, we propose a data-based cross-validation method to determine the bandwidth in the Nadaraya–Watson estimator. Via simula-

W. Lee

Department of Economics, University of Chicago, USA

P. Greenwood · N. Heckman

Statistics Department, University of British Columbia, Vancouver BC, Canada

W. Wefelmeyer (Corresponding Author)

Mathematical Institute, University of Cologne, Germany, wefelm@math.uni-koeln.de

tion, we study several methods of bandwidth choices, and compare our estimator to several existing estimators. In terms of mean squared error, our new estimator outperforms existing estimators.

Keywords diffusion process, nonparametric estimation, discrete observations, high-frequency observations, microstructure noise, pre-averaging, drift estimation, Nadaraya–Watson estimator

1 Introduction

Consider a one-dimensional time-homogeneous diffusion process given by the stochastic differential equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,$$

where μ and σ are real-valued functions, the *drift* and the *diffusion* functions, respectively, and $(W_t)_{t\geq 0}$ is a Wiener process. We want to estimate the drift function μ at a point x.

Suppose first that we have no microstructure noise, and we observe the process at times $t = i\Delta$ for $i = 1,\ldots,n$, where $\Delta \to 0$ and $n\Delta \to \infty$. Under our Assumption 3 of stationarity, we have $\mu(x) = \lim_{\Delta\to 0} E(X_{t+\Delta} - X_t | X_t = x)/\Delta$. Hence a natural estimator of $\mu(x)$ is the *Nadaraya–Watson* estimator suggested by Arfi (1995) and Stanton (1997),

$$\overline{\mu}_{NW}(x) = \frac{\sum_{i=1}^{n-1} S_i K_h(X_{i\Delta} - x)}{\sum_{i=1}^{n-1} K_h(X_{i\Delta} - x)},$$

with slopes $S_i = (X_{(i+1)\Delta} - X_{i\Delta})/\Delta$, where $K_h(x) = K(x/h)/h$ for a kernel K and a bandwidth h. Several competitors to this estimator can be found in the literature.

Bandi and Phillips (2003) modify the estimator by *double smoothing*, replacing the slopes S_i by smoothed slopes obtained as moving averages

$$\tilde{S}_i = \frac{1}{|\mathrm{B}_i|} \sum_{k \in \mathrm{B}_i} \frac{X_{(k+1)\Delta} - X_{k\Delta}}{\Delta}$$

where $\mathrm{B}_i = \{k : |X_{k\Delta} - X_{i\Delta}| \leq \ell\}$ with bandwidth ℓ tending to zero at an appropriate rate as specified in Theorem 3 of Bandi and Phillips (2003). They show that the resulting estimator

$$\overline{\mu}_{BP}(x) = \frac{\sum_{i=1}^{n-1} \tilde{S}_i K_h(X_{i\Delta} - x)}{\sum_{i=1}^{n-1} K_h(X_{i\Delta} - x)}$$

is consistent and determine the asymptotic distribution both in ergodic and nonergodic cases. Moloche (2001) and Bandi and Phillips (2009a) suggest extensions with local polynomial smoothers in place of kernel estimators. Bandi and Nguyen (2004) obtain analogous results for diffusions with jumps. Hoffmann (1999) constructs rate-optimal estimators for μ based on wavelet thresholding. Comte et al. (2007) and (2012) construct a rate-optimal penalized least squares estimator and local polynomial smoothers for μ based on piecewise polynomial splines. For the multivariate version of the diffusion process, Schmisser (2013) extends the result of Comte et al. to the multivariate version of the diffusion process. Similar results exist for *continuously* observed diffusions. For one-dimensional diffusions we refer to Dalalyan and Kutoyants (2002) and Dalalyan (2005). For multivariate diffusions see Strauch (2015, 2016).

Suppose now that at times $t_i = i\Delta$ with $i = 1, \ldots, n$ we observe the process up to *microstructure noise* ε_{ni}. This means that the observations are

$$Y_{i\Delta} = X_{i\Delta} + \varepsilon_{ni}, \quad i = 1, \ldots, n,$$

with errors $\varepsilon_{n1}, \dots, \varepsilon_{nn}$ that are independent of the process $(X_t)_{t \geq 0}$. Such measurement errors are in particular observed in financial time series. For example, Zhou (1996) reported the presence of measurement error in foreign exchange rates data, and Jones (2003) argued for the presence of measurement error in the seven-day Eurodollar rates dataset of Aït-Sahalia (1996). When microstructure noise is present, we can use the *pre-averaging* approach introduced in Podolskij and Vetter (2009) in the context of volatility. See also Jacod et al. (2009) and Chapter 16 of Jacod and Protter (2012). Assume that $n = mr$. Decompose the time indices into m disjoint blocks of length r and take averages of the observations over each block,

$$\overline{Y}_j = \frac{1}{r} \sum_{i=1}^{r} Y_{((j-1)r+i)\Delta}.$$

Our version of the Nadaraya–Watson estimator based on these averages is

$$\hat{\mu}_{Ave}(x) = \frac{\sum_{j=1}^{m-2} \frac{\overline{Y}_{j+2} - \overline{Y}_{j+1}}{r\Delta} K_h(\overline{Y}_j - x)}{\sum_{j=1}^{m-2} K_h(\overline{Y}_j - x)}.$$

For technical convenience, we have shifted the slope in the numerator one time block forward. Our main result is analogous to Corollary 2 of Bandi and Phillips (2003), now with microstructure noise. In our notation, their result reads as follows:

Let μ and σ be twice continuously differentiable. Let the kernel $K \in L^2(\mathbb{R})$ be a bounded, symmetric and continuously differentiable density function such that $\int s^2 K(s)\, ds < \infty$ and K' is absolutely integrable. Assume that μ and σ grow locally at most as fast as x, and that σ is positive. Let X be stationary with stationary density f. Let $n \to \infty$, $\Delta \to 0$, $n\Delta \to \infty$ and $(n\Delta/h)(\Delta \log(1/\Delta))^{1/2} \to 0$. Then the assumption of Theorem 1 below holds for $\overline{\mu}_{BP}$.

A different estimator is treated by Schmisser (2011). She constructs a rate-optimal penalized least squares estimator for μ, following the approach of Comte et al. (2007, 2010). The optimal rate of convergence is the same as for our estimator.

The structure of this article is as follows. In Section 2, we introduce our assumptions and state our main result, the consistency and asymptotic normality of our estimator. In Section 3, we discuss the choice of the bandwidth h and the block size r. In Section 4, we describe our simulation study. The proof of the main result is in Section 5.

2 Result

We use the following assumptions.

Assumption 1 As $n = mr \to \infty$, the sequence of positive real numbers $\Delta = \Delta_n$ and the sequence of positive integers $r = r_n$ satisfy $\Delta \to 0$, $n\Delta \to \infty$, $r \to \infty$, $r\Delta \to 0$.

Note that this implies that the integers $m = m_n = n/r$ satisfy $m \to \infty$.

Assumption 2 The functions μ and σ are twice differentiable on $\mathbb{R}$, and the second derivatives are Hölder with exponent ε for some $\varepsilon > 0$. In addition, $\sigma^2(x) > 0$ for all $x \in \mathbb{R}$.

Assumption 3 The solution process $\{X_t\}$ is positive recurrent and strictly stationary with stationary density f. In addition, both $E(\mu^2(X_0))$ and $E(\sigma^2(X_0))$ are finite.

Sufficient conditions for positive recurrence and ergodicity are: Both μ and σ are globally Lipschitz, σ is bounded and bounded away from zero, and $x\mu(x) \lesssim -|x|^\alpha$ for large $|x|$ and some $\alpha \geq 1$.

Assumption 4 The kernel $K \in L^2(\mathbb{R})$ is a bounded, symmetric and continuously differentiable density function such that $\int_{-\infty}^{\infty} s^2 K(s)ds < \infty$. In addition, its derivative, K', is bounded and is in $L^1(\mathbb{R})$.

Assumption 5 The errors $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ are independent, and independent of the process $(X_t)_{t\geq 0}$. Also, $E(\varepsilon_{ni}) = 0$ and $\mathrm{Var}(\varepsilon_{ni})$ is bounded by a finite constant σ_ε^2 for all i and n.

In the literature, the ε_{ni} with $i = 1, \ldots, n$ are usually assumed to be identically distributed. Some authors assume that $\mathrm{Var}(\varepsilon_{ni}) = \sigma_\varepsilon^2$ for all i and n; see e.g. Zhang et al. (2005). In contrast, some papers in the literature, and most of the papers in the rounding error literature according to Jacod et al. (2009), assume that $\mathrm{Var}(\varepsilon_{ni}) = a_n \sigma_\varepsilon^2$ for all i, where $a_n \to 0$ as $n \to \infty$; see e.g. Bandi et al. (2009b). Our Assumption 5 includes both specifications as special cases.

Now we state our main result.

Theorem 1 *Suppose Assumptions 1–5 hold, and let*

$$\Big(\frac{n\Delta}{h}\Big)^2 r\Delta \log\frac{1}{r\Delta} \to 0, \quad n\Delta h \to \infty, \quad \frac{n}{h^3 r^2} \to 0.$$

Then the following results hold for each x such that $f(x) > 0$:

(1) If $n\Delta h^5 \to 0$, then

$$(n\Delta h)^{1/2}\big(\hat{\mu}_{Ave}(x) - \mu(x)\big) \Rightarrow N\Big(0, \frac{\sigma^2(x)}{f(x)} \int K^2(s)\,ds\Big).$$

(2) If $n\Delta h^5$ is bounded, then

$$(n\Delta h)^{1/2}\big(\hat{\mu}_{Ave}(x) - \mu(x) - h^2\Gamma(x)\big) \Rightarrow N\Big(0, \frac{\sigma^2(x)}{f(x)} \int K^2(s)\,ds\Big),$$

where

$$\Gamma(x) = \left(\mu'(x)\frac{f'(x)}{f(x)} + \frac{1}{2}\mu''(x)\right)\int s^2 K(s)\,ds.$$

In order to keep the technical details at a minimum, we have restricted attention to the simplest case. The result can be generalized in several directions.

1. Since we reduce our proof to that of Bandi and Phillips (2003), we must have their assumptions on μ and σ, namely that they are twice continuously differentiable. We expect analogous results assuming more generally that μ and σ are, say, r-times differentiable with r-th derivative fulfilling a certain Hölder condition. The optimal bandwidth must then be chosen accordingly.

2. Bandi and Phillips (2003) also construct an estimator for the diffusion function σ. In the ergodic case, Corollary 3 of their paper, their estimator needs to be standardized differently from the estimator for the drift function μ. In particular, the asymptotic variance of their estimator depends on the number of observations. We do not know whether and how their result carries over to observations with microstructure noise.

3. The errors $\varepsilon_{n1},\dots,\varepsilon_{nn}$ need not be independent. The proof would go through without changes as long as the variance of the averages $\bar{\varepsilon}_j = \sum_{i=1}^{r} \varepsilon_{n((j-1)+i)}$ is of the order $1/r$ uniformly in j, for example if a central limit theorem holds for the errors.

4. For the case without microstructure noise, Bandi and Phillips (2003), Theorem 3, obtain a central limit theorem for their drift estimator that covers also a non-ergodic (null-recurrent) situation. The estimator is then normed with an estimator of the "chronological" local time of the process. We do not know whether this result also carries over to the case of observations with microstructure noise.

5. As noted above, again for the case without microstructure noise, Schmisser (2013) obtains a rate-optimal penalized least-squares estimator for the drift function in the multivariate case. We expect our result on kernel estimators also to carry over to the multivariate case, with a correspondingly changed bias-variance trade-off of course.

6. Again for the case without microstructure noise, Schmisser (2014) obtains bounds for the rate of penalized least-squares estimators for the drift function of certain one-dimensional processes that are the sum of a diffusion process and a centered pure-jump Lévy process. Presumably, in the case with microstructure noise, our result for kernel estimators also carries over to processes with jumps.

7. Like Bandi and Phillips (2003) we restrict attention to equally spaced observations. In finance mathematics, when estimating the (integrated) realized volatility, one often has observations at non-equidistant times. We are not aware of results for estimators of the drift and diffusion functions μ and σ under such observation schemes but expect appropriate versions to hold.

The proof is in Section 5. In particular, we obtain the following result for a bandwidth $h = (n\Delta)^{-1/5}$:

$$(n\Delta)^{2/5}\big(\hat{\mu}_{Ave}(x) - \mu(x)\big) \Rightarrow N\Big(\Gamma(x), \frac{\sigma^2(x)}{f(x)} \int K^2(s)\,ds\Big).$$

The bandwidth rate of $(n\Delta)^{-1/5}$ is optimal in the sense of minimizing asymptotic mean squared error, as discussed in Section 3.

3 Bandwidth Choices

For our simulation study in Section 4 we use three different methods of choosing the bandwidth, based on minimizing the asymptotic mean squared error (AMSE) at x, the asymptotic mean integrated squared error (AMISE), and a cross-validation criterion. For an overview of bandwidth choice methods, see e.g. Jones et al. (1996) and the references therein.

Bandwidth minimizing AMSE. By Theorem 1(2), the AMSE at x of $\hat{\mu}_{AVE}(x)$ is

$$\text{AMSE}(x) = h^4 \Gamma^2(x) + \frac{\sigma^2(x) \int K^2(s)\, ds}{f(x) n\Delta h}.$$

When $\Gamma^2(x) \neq 0$, the minimizer of AMSE(x), the *oracle bandwidth* $h_{opt}(x)$, satisfies

$$n\Delta h_{opt}^5(x) = \frac{\sigma^2(x) \int K^2(s)\, ds}{4\Gamma^2(x) f(x)}. \tag{3.1}$$

We use h_{opt} in our simulation study, where we know $\Gamma(x)$, $\sigma(x)$ and $f(x)$. However, in data analysis, $\Gamma(x)$, $\sigma(x)$ and $f(x)$ are unknown, and so we must use an estimator of $h_{opt}(x)$.

Bandwidth minimizing AMISE. A global bandwidth, that is, a bandwidth not depending on x, is obtained as follows. The AMISE of $\hat{\mu}_{Ave}(x)$ is defined by

$$\text{AMISE} = \int \text{AMSE}(x) f(x)\, dx = h^4 \int \Gamma^2(x) f(x)\, dx + \frac{\int \sigma^2(x)\, dx \int K^2(s)\, ds}{n\Delta h}.$$

When $\int \Gamma^2(x) f(x)\, dx \neq 0$, the minimizer of AMISE, the *oracle global bandwidth* h_{opt}^{integ}, is defined by

$$(n\Delta)(h_{opt}^{integ})^5 = \frac{\int \sigma^2(x)\, dx \int K^2(s)\, ds}{4 \int \Gamma^2(x) f(x)\, dx}. \tag{3.2}$$

Again, in our simulation study, we can use h_{opt}^{integ}, but in data analysis, we must use an estimator of h_{opt}^{integ}.

Bandwidth based on H-block cross-validation. For data analysis, we do not use estimators of $h_{opt}(x)$ or h_{opt}^{inter}. Rather, we use H-block cross-validation, as proposed by Chu and Marron (1991) and further developed by Burman et al. (1994), who coined the name “H-block”. It adapts the well-known leave-one-out cross-validation of Stone (1974) to dependent data. For an integer H we estimate the prediction error given the bandwidth h by

$$\widehat{PE}(h) = \sum_{k=1}^{m-1} \left(\hat{\mu}_k(\overline{Y}_k) - \frac{\overline{Y}_{k+1} - \overline{Y}_k}{r\Delta} \right)^2, \tag{3.3}$$

where $\hat{\mu}_k(\overline{Y}_k)$ is our estimator $\hat{\mu}_{Ave}(x)$ evaluated at $x = \overline{Y}_k$ and calculated without the data $\overline{Y}_{j+1} - \overline{Y}_j$, $j = k-H, \ldots, k+H$, that is,

$$\hat{\mu}_k(\overline{Y}_k) = \frac{\sum_{j \in A_k} \frac{\overline{Y}_{j+1} - \overline{Y}_j}{r\Delta} K_h(\overline{Y}_j - \overline{Y}_k)}{\sum_{j \in A_k} K_h(\overline{Y}_j - \overline{Y}_k)} \tag{3.4}$$

for the set of indices $A_k = \{j = 1, \ldots, m-1 : j \neq k-H, \ldots, k+H\}$. The integer H is chosen so that the dependence between $\overline{Y}_k$ and $\overline{Y}_j$, $j \in A_k$ is “weak enough”. For the implementation, H can be chosen by looking at the empirical autocorrelation function of the $\overline{Y}_j$’s. In $\widehat{PE}(h)$, we use $(\overline{Y}_{k+1} - \overline{Y}_k)/(r\Delta)$ as it will provide a good target value for $\hat{\mu}_k(\overline{Y}_k)$, since the pre-averaged process $\{\overline{Y}_j\}$ is close to the underlying, unobserved process. Then the cross-validation bandwidth h_{cv} is the minimizer of $\widehat{PE}(h)$:

$$h_{cv} = \operatorname{argmin} \widehat{PE}(h). \tag{3.5}$$

4 Simulation Study

We carry out a simulation study to assess the finite sample performance of our estimator. We simulate data with two kinds of underlying models for the drift coefficient μ:

$$dX_t = 0.858 \times (0.086 - X_t)dt + 0.157\sqrt{X_t}dW_t, \tag{4.1}$$

$$dX_t = -(X_t - 1)(X_t + 1)^2 dt + 2dW_t. \tag{4.2}$$

The process defined by Equation (4.1) is called a Cox–Ingersoll–Ross (CIR) process and is used as an underlying model for a short-term interest rate process. The value of a CIR process at time t equals the annual interest rate, and the time is measured in years, with a year being 250 days (counting business days only). Following the parameter choice of Chapman and Pearson (2000), we use the parameter values $(0.858, 0.086, 0.157)$ in Equation (4.1) to match the solution process's monthly (i.e. 21st-order) autocorrelation, unconditional mean and unconditional variance to the corresponding sample quantities of the dataset of Aït-Sahalia (1996). We use the process defined by (4.2) to study the performance of our estimator when the true drift coefficient is nonlinear. Note that (4.2) satisfies the sufficient conditions for Assumption 3 mentioned after Assumption 3, but (4.1) does not. However, Assumption 3 holds also for the CIR process by Bhan and Mandrekar (2010) and Jin et al. (2013).

We generated 1,000 discretely observed independent sample paths for each of models (4.1) and (4.2) at time increments of $\Delta = 1/250$, which represents daily observations, and with the number of observations $n = 5505$, which is the sample size of the dataset of Aït-Sahalia (1996). To generate these sample paths, we used the companion R package to Iacus (2008), sde: Simulation and Inference for Stochastic Differential Equations (Iacus, 2014).

We then added independent and identically normally distributed measurement errors to the generated discretely observed sample paths. For model (4.1), we took 0.002 as the standard deviation of our measurement errors. This value is an estimate of the standard deviation of the measurement error of the dataset of Aït-Sahalia (1996), proposed by Jones (2003). We note that the value 0.002 is 5.7% of the unconditional standard deviation of the solution process of model (4.1). We also set the standard deviation of the measurement error added to model (4.2) to be 5.7% of the unconditional standard deviation of the solution process of model (4.2), that is, to be 0.0661.

Besides our pre-averaged estimator $\hat{\mu}_{Ave}(x)$, we considered four other estimators. The first two are versions of the Nadaraya–Watson estimator $\overline{\mu}_{NW}(x)$ and the double-smoothing estimator $\overline{\mu}_{BP}(x)$ of Bandi and Phillips (2003), but now using data $Y_{i\Delta}$, $i = 1, \ldots, n$, with microstructure noise,

$$\hat{\mu}_{NW}(x) = \frac{\sum_{i=1}^{n-1} S_i K_h(Y_{i\Delta} - x)}{\sum_{i=1}^{n-1} K_h(Y_{i\Delta} - x)}, \qquad \hat{\mu}_{BP}(x) = \frac{\sum_{i=1}^{n-1} \overline{S}_i K_h(Y_{i\Delta} - x)}{\sum_{i=1}^{n-1} K_h(Y_{i\Delta} - x)},$$

with $S_i = (Y_{(i+1)\Delta} - Y_{i\Delta})/\Delta$ and $\overline{S}_i = (1/|B_i|)\sum_{k \in B_i}(Y_{(k+1)\Delta} - Y_{k\Delta})/\Delta$. Here $B_i = \{k : |Y_{k\Delta} - Y_{i\Delta}| \leq \ell\}$. The other two are *subsampled* versions $\hat{\mu}_{NWS}(x)$ and $\hat{\mu}_{BPS}(x)$, using only the subsample $Y_{jr\Delta}$, $j = 1, \ldots, m$, of $Y_{i\Delta}$, $i = 1, \ldots, n$.

For all estimators, the kernel K was equal to the standard normal density. In *Ave*, *NWS* and *BPS* we chose $r = 5$, yielding pre-averaged data equal to weekly averages for *Ave* and subsampled data equal to weekly closing prices (i.e. every fifth value) for *NWS* and *BPS*, assuming 5 business days a week.

For each estimator, we set h equal to the oracle bandwidth based on AMSE(x), defined in (3.1). Note that all estimators have the same oracle bandwidths because they have the same asymptotic biases and variances. For *BPS*, we used $\ell = h_{opt}^{integ}$. This is motivated by the result of Bandi and Phillips (2003) that h and ℓ should be of the same order of magnitude (Remark 5 in Bandi and Phillips, 2003). We choose ℓ independent of x due to the high computational cost when ℓ depends on x.

In addition to these oracle bandwidths, we used cross-validation bandwidths but for only three estimators, *Ave*, *NWS* and *BPS*, due to the high computational cost. For *Ave*, we used $\widehat{PE}$ as in Equation (3.3). For *NWS* and *BPS*, we used

$$\widehat{PE}(h) = \sum_{k=1}^{m-1} \left(\hat{\mu}_k(Y_{kr\Delta}) - \frac{Y_{(k+1)r\Delta} - Y_{kr\Delta}}{r\Delta} \right)^2,$$

where $\hat{\mu}_k$ is either $\hat{\mu}_{NWS}$ or $\hat{\mu}_{BPS}$ calculated without $Y_{jr\Delta}$, $j = k - H, \ldots, k + H$. For *BPS*, we chose to minimize $\widehat{PE}(h)$ with respect to h with the restriction that $\ell = h$, since, as noted above, h and ℓ should be the same order of magnitude. We set $H = 150$ for *Ave*, *NWS* and *BPS* by the observation that, for most sample paths, the empirical autocorrelation functions of the averaged and the subsampled data reached zero before the time lag reached 150.

In order to numerically minimize $\widehat{PE}$ for each sample path, we first calculated $D = \max_i Y_{i\Delta} - \min_i Y_{i\Delta}$, and we found the local minimum of $\widehat{PE}$ by computing $\widehat{PE}(h)$ for $h = D/30, 2D/30, \ldots, D$. If there were multiple bandwidths that attained local minima, we took the largest bandwidth, which is a common practice when using cross-validation. In our simulation result, a grid of 30 values was fine enough to detect the local minima. We obtained an interior minimizer of $\widehat{PE}$ for *Ave* and *NWS* for every sample path we generated. This was not the case for *BPS*. For the *BPS* estimator, for some sample paths, the curve $h \mapsto \widehat{PE}(h)$ evaluated at the grid of the bandwidths was monotonically decreasing in h. In this case, we picked D as the bandwidth. This occurred in 365 of 1000 sample paths for model (4.1) and in 214 of 1000 sample paths for model (4.2). Whenever our chosen bandwidth was D, the *BPS* function estimate was a constant function. These constant function estimates had very little variation in the intercept. For example, the standard deviation of the intercepts for model (4.2) was 0.07, which is small considering that the drift coefficient of model (4.2) ranges from -1 to 1 at our evaluation points.

To assess our estimators, we evaluated them pointwise at each point in the grid which consists of 100 equispaced points ranging from the 20th percentile to the 80th percentile of the stationary density

$$f(x) = \frac{1}{G\sigma^2(x)} \exp\left(2\int_0^x \frac{\mu(y)}{\sigma^2(y)}\,dy\right)$$

with norming constant G defined by $\int f(x)\,dx = 1$.

Table 1 summarizes the estimated expected integrated squared errors (ISE) of the estimators. For each combination of the model, the estimator and the bandwidth choice method, we approximated the ISE for each sample path by a stationary-density-weighted sum of the squared errors over the grid of evaluation points. We provide the mean of the 1,000 ISEs along with the standard error of the mean in Table 1.

According to Table 1, if we use the oracle bandwidth, *Ave* has a smaller expected ISE than any other listed estimators except for *BPS*. If we use the cross-validation bandwidth, *Ave* outperforms both *NWS* and *BPS*. We note that using the oracle bandwidth instead of cross-validation increases the expected ISE, except for the *BPS* es-

Table 1 Means (and standard errors, i.e. standard deviations/$\sqrt{1000}$) of the integrated squared errors (ISE) of candidate estimators over 1,000 sample paths. Labels *NW* and *BP* stand for the Nadaraya–Watson (single smoothing) and the double-smoothing estimators of Bandi and Phillips (2003), respectively. Labels *NWS* and *BPS* represent the subsampled NW and BP. Label *Ave* stands for the pre-averaging estimator.

	ISE, Model (4.1)		ISE, Model (4.2)	
Estimator	Oracle	CV	Oracle	CV
NW	1.475 (0.048)	—	94.8 (2.1)	—
BP	0.628 (0.022)	—	50.9 (1.5)	—
NWS	0.479 (0.016)	0.187 (0.010)	48.9 (1.1)	27.84 (0.8)
BPS	0.194 (0.008)	0.283 (0.011)	27.2 (0.8)	22.67 (0.7)
Ave	0.197 (0.007)	0.118 (0.003)	24.2 (0.6)	15.83 (0.4)

timator in model (4.1). Thus we recommend using the *Ave* estimator with the cross-validation bandwidth.

5 Proof of Theorem 1

We write our estimator as $\hat{\mu}_{Ave}(x) = N(x)/D(x)$ with

$$N(x) = \frac{1}{m-2}\sum_{j=1}^{m-2} \frac{\overline{Y}_{j+2} - \overline{Y}_{j+1}}{r\Delta} K_h(\overline{Y}_j - x),$$

$$D(x) = \frac{1}{m-2}\sum_{j=1}^{m-2} K_h(\overline{Y}_j - x).$$

We compare it with the Nadaraya–Watson estimator based on the (unobserved) subsample $X_{jr\Delta}$, $j = 0, \ldots, m-2$, which we write $\hat{\mu}_X = N_X(x)/D_X(x)$ with

$$N_X(x) = \frac{1}{m-2}\sum_{j=1}^{m-2} \frac{X_{jr\Delta} - X_{(j-1)r\Delta}}{r\Delta} K_h(X_{(j-1)r\Delta} - x),$$

$$D_X(x) = \frac{1}{m-2}\sum_{j=1}^{m-2} K_h(X_{(j-1)r\Delta} - x).$$

We show that $(n\Delta h)^{1/2}\big(\hat{\mu}_{Ave}(x)-\hat{\mu}_X(x)\big)=o_p(1)$. Then the result of Theorem 1 follows from the corresponding result for $\hat{\mu}_X$, which is proved in Theorems 1–3 of Bandi and Phillips (2003) and equation (3.47) and Theorem 7 of Bandi and Phillips (2009a). They also show $D_X(x)\to f(x)$ almost surely.

In order to prove $(n\Delta h)^{1/2}\big(\hat{\mu}_{Ave}(x)-\hat{\mu}_X(x)\big)=o_p(1)$, we show that

$$D(x)-D_X(x)=o_p(1), \tag{5.1}$$

$$(n\Delta h)^{1/2}\big(N(x)-N_X(x)\big)=o_p(1). \tag{5.2}$$

The proof will use the following lemma repeatedly. Similarly as $\overline{Y}_j$ we define

$$\overline{X}_j=\frac{1}{r}\sum_{i=1}^{r}X_{((j-1)r+i)\Delta},\quad \bar{\varepsilon}_j=\frac{1}{r}\sum_{i=1}^{r}\varepsilon_{n((j-1)r+i)}.$$

Lemma 1 *Suppose that Assumptions 1, 2 and 3 hold. Define*

$$\kappa=\kappa_n=\max_{j\le m}\sup_{(j-1)r\Delta\le s\le jr\Delta}|X_s-X_{(j-1)r\Delta}|,$$

$$\gamma=\gamma_n=\max_{j\le m}E\big((\overline{X}_j-X_{(j-1)r\Delta})^2\big).$$

Then the following hold.

(1) $\kappa^2=O(r\Delta\log(1/(r\Delta)))$ *almost surely.*

(2) $\gamma\le\beta r\Delta$ *for some* $\beta<\infty$.

(3) $\max_{j\le m}E\big((X_{jr\Delta}-X_{(j-1)r\Delta})^2\big)\le\phi r\Delta$ *for some* $\phi<\infty$.

(4) $\max_{j\le m}|\overline{X}_j-X_{(j-1)r\Delta}|\le\kappa$.

(5) $\max_{j\le m}E\big|\overline{X}_j-X_{(j-1)r\Delta}\big|\le\sqrt{\gamma}$.

(6) $\max_{j\le m}|\overline{X}_{j+1}-\overline{X}_j|\le 3\kappa$.

(7) $\max_{j\le m}E\big|\overline{X}_{j+1}-\overline{X}_j\big|\le\phi^*\sqrt{r\Delta}$ *for some* $\phi^*<\infty$.

Note that (1) and the assumptions of Theorem 1 imply that

$$\frac{n\Delta}{h}\,\kappa \to 0 \quad \text{almost surely.} \tag{5.3}$$

Proof. The proof of (1) uses Lévy's modulus of continuity of diffusions. See Friz and Victoir (2004), Theorem 22.

The proof of (2) follows from the two bounds

$$E\Big[\Big(\int_a^b \mu(X_s)ds\Big)^2\Big] \leq (b-a)^2 E(\mu^2(X_0)), \tag{5.4}$$

$$E\Big[\Big(\int_a^b \sigma(X_s)dW_s\Big)^2\Big] \leq (b-a)E(\sigma^2(X_0)). \tag{5.5}$$

To see this, write

$$\begin{aligned}\overline{X}_j - X_{(j-1)r\Delta} &= \frac{1}{m-2}\sum_{j=1}^{m-2}\big(X_{((j-1)r+i)\Delta} - X_{(j-1)r\Delta}\big)\\ &= \frac{1}{m-2}\sum_{j=1}^{m-2}\int_{(j-1)r\Delta}^{((j-1)r+i)\Delta}\big(\mu(X_s)ds + \sigma(X_s)dW_s\big)\end{aligned}$$

and use $(a+b)^2 \leq 2(a^2+b^2)$ to apply the above two bounds.

The proof of (3) is similar, also using (5.4) and (5.5).

Conclusion (4) follows directly from the definitions of κ and $\overline{X}_j$. Conclusion (5) follows by the Cauchy–Schwarz inequality. Conclusion (6) follows by rewriting

$$\bar{X}_{j+1} - \bar{X}_j = (\bar{X}_{j+1} - X_{jr\Delta}) + (X_{jr\Delta} - X_{(j-1)r\Delta}) - (\bar{X}_j - X_{(j-1)r\Delta}),$$

then using the definition of κ and (4). Conclusion (7) follows similarly, but using (3) with the Cauchy–Schwarz inequality, plus (5) and the bound on γ in (2). ■

Throughout the proof of Theorem 1, we use the fact that K is bounded and Lipschitz continuous, the latter following by the existence and boundedness of K'. Therefore,

$$|K_h(x+x_0)-K_h(x)| \le \frac{M}{h^2}|x_0|, \quad |K_h(x)| \le \frac{M}{h}, \quad |K_h'(x)| \le \frac{M}{h^2}, \tag{5.6}$$

where M equals the maximum of $\sup_x |K(x)|$ and $\sup_x |K'(x)|$.

Instead of proving (5.1), we prove the stronger statement that

$$D(x) - D_X(x) = o_p((n\Delta h)^{-1/2}).$$

We bound

$$\begin{aligned} |D(x) - D_X(x)| &\le \frac{1}{m-2}\sum_{j=1}^{m-2} \left|K_h(\overline{Y}_j - x) - K_h(X_{(j-1)r\Delta} - x)\right| \\ &\le \frac{M}{h^2}\frac{1}{m-2}\sum_{j=1}^{m-2} |\overline{Y}_j - X_{(j-1)r\Delta}|. \end{aligned} \tag{5.7}$$

Using the definition of $\overline{Y}_j$ and Lemma 1(4), we have

$$|\overline{Y}_j - X_{(j-1)r\Delta}| = |\overline{X}_j - X_{(j-1)r\Delta} + \bar{\varepsilon}_j| \le \kappa + |\bar{\varepsilon}_j|. \tag{5.8}$$

Combining (5.7) and (5.8), we get

$$|D(x) - D_X(x)| \le \frac{M}{h^2}\kappa + \frac{M}{h^2}\frac{1}{m-2}\sum_{j=1}^{m-2} |\bar{\varepsilon}_j|. \tag{5.9}$$

We show the first term is $o((n\Delta h)^{-1/2})$ almost surely by showing the stronger result that the first term is $o((n\Delta h)^{-1})$ almost surely, recalling the assumption of Theorem 1 that $n\Delta h \to \infty$. We write

$$n\Delta h \frac{\kappa}{h^2} = \left(\frac{n\Delta}{h}\kappa\right)$$

which is $o(1)$ almost surely by (5.3). To study the second term in (5.9) we use the inequality $E|\bar{\varepsilon}_j| \le \sigma_\varepsilon/\sqrt{r}$ and bound $(n\Delta h)^{1/2}$ times the expected value of the second term by

$$\sqrt{n\Delta h}\,\frac{M}{h^2}\frac{\sigma_\varepsilon}{\sqrt{r}} = M\sigma_\varepsilon\sqrt{\frac{n}{h^3 r^2}r\Delta}$$

which converges to 0 by Assumption 1 and the assumptions in Theorem 1. It follows from the Markov inequality that the second term in (5.9) is $o_p((n\Delta h)^{-1/2})$.

In order to prove (5.2), we write

$$\begin{aligned}\sqrt{n\Delta h}\big(N(x)-N_X(x)\big) &= \frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\sum_{j=1}^{m-2}\big[(\bar{Y}_{j+2}-\bar{Y}_{j+1})K_h(\bar{Y}_j-x)\\ &\qquad -(X_{jr\Delta}-X_{(j-1)r\Delta})K_h(X_{(j-1)r\Delta}-x)\big]\\ &\equiv \frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\sum_{j=1}^{m-2}\big[A_j+B_j+C_j+D_j+E_j\big], \qquad (5.10)\end{aligned}$$

where

$$\begin{aligned}A_j &= (\bar{Y}_{j+2}-\bar{Y}_{j+1})\big[K_h(\bar{Y}_j-x)-K_h(\bar{X}_j-x)\big],\\ B_j &= \big[\bar{Y}_{j+2}-\bar{Y}_{j+1}-(\bar{X}_{j+2}-\bar{X}_{j+1})\big]K_h(\bar{X}_j-x),\\ C_j &= (\bar{X}_{j+2}-\bar{X}_{j+1})\big[K_h(\bar{X}_j-x)-K_h(X_{(j-1)r\Delta}-x)\big],\\ D_j &= \big[\bar{X}_{j+2}-\bar{X}_{j+1}-(X_{(j+1)r\Delta}-X_{jr\Delta})\big]K_h(X_{(j-1)r\Delta}-x),\\ E_j &= \big[X_{(j+1)r\Delta}-X_{jr\Delta}-(X_{jr\Delta}-X_{(j-1)r\Delta})\big]K_h(X_{(j-1)r\Delta}-x).\end{aligned}$$

To study the contributions of the A_j to (5.10), we bound $E|A_j|$ using the bounds involving K_h in (5.6):

$$\begin{aligned}\mathrm{E}|A_j| &\le \frac{M}{h^2}\mathrm{E}|(\bar{Y}_{j+2}-\bar{Y}_{j+1})\bar{\varepsilon}_j| = \frac{M}{h^2}\mathrm{E}|\bar{Y}_{j+2}-\bar{Y}_{j+1}|\mathrm{E}|\bar{\varepsilon}_j|\\ &\le \frac{M}{h^2}\big(\mathrm{E}|\overline{X}_{j+2}-\overline{X}_{j+1}|+\mathrm{E}|\bar{\varepsilon}_{j+2}-\bar{\varepsilon}_{j+1}|\big)\mathrm{E}|\bar{\varepsilon}_j|\\ &\le \frac{M}{h^2}\Big(\phi^*\sqrt{r\Delta}+\frac{2\sigma_\varepsilon}{\sqrt{r}}\Big)\frac{\sigma_\varepsilon}{\sqrt{r}}\end{aligned}$$

by Lemma 1 (7), and the fact that $\mathrm{E}|\bar{\varepsilon}_j| \le \sigma_\varepsilon/\sqrt{r}$ for all j. Therefore, we bound the contribution of the A_j to (5.10) by

$$\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\sum_{j=1}^{m-2}\mathrm{E}|A_j| \le \frac{\sqrt{n\Delta h}}{r\Delta}\frac{M}{h^2}\Big(\phi^*\sqrt{r\Delta}+\frac{2\sigma_\varepsilon}{\sqrt{r}}\Big)\frac{\sigma_\varepsilon}{\sqrt{r}}$$

which converges to 0, since, by Lemma 1 and the assumptions of Theorem 1,

$$\frac{\sqrt{n\Delta h}}{r\Delta}\frac{1}{h^2}\sqrt{r\Delta}\frac{1}{\sqrt{r}} = \sqrt{\frac{n}{h^3r^2}} \to 0$$

and

$$\frac{\sqrt{n\Delta h}}{r\Delta}\frac{1}{h^2}\frac{1}{r} = \frac{n}{r^2h^3}\frac{h^2}{\sqrt{n\Delta h}} \to 0.$$

We consider the contribution of the B_j to (5.10). Since $B_j = (\bar{\varepsilon}_{j+2} - \bar{\varepsilon}_{j+1})K_h(\bar{X}_j - x)$,

$$\sum_{j=1}^{m-2} B_j = \bar{\varepsilon}_m K_h(\bar{X}_{m-2} - x) - \bar{\varepsilon}_2 K_h(\bar{X}_1 - x) + \sum_{j=1}^{m-3} \bar{\varepsilon}_{j+2}\left[K_h(\bar{X}_j - x) - K_h(\bar{X}_{j+1} - x)\right].$$

Using the boundedness and Lipschitz continuity of K, we obtain

$$E\Big|\sum_{j=1}^{m-2} B_j\Big| \le M\Big[\frac{E|\bar{\varepsilon}_m|}{h} + \frac{E|\bar{\varepsilon}_2|}{h} + \frac{1}{h^2}\sum_{j=1}^{m-3} E|\bar{\varepsilon}_{j+2}|E|\bar{X}_{j+1} - \bar{X}_j|\Big].$$

By Lemma 1 (7), $E|\bar{X}_{j+1} - \bar{X}_j| \le \phi^*\sqrt{r\Delta}$. Therefore, since $E|\bar{\varepsilon}_j| \le \sigma_\varepsilon/\sqrt{r}$, for some constant C,

$$\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}E\Big|\sum_{j=1}^{m-2} B_j\Big| \le C\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\Big(\frac{1}{h\sqrt{r}} + \frac{m-2}{h^2}\sqrt{\Delta}\Big). \tag{5.11}$$

To show that this converges to 0, recall that $m = n/r$ and write

$$\Big(\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\frac{1}{h\sqrt{r}}\Big)^2 = \Big(\frac{m}{m-2}\Big)^2\frac{1}{m^2}\frac{n}{r^3\Delta h} = \Big(\frac{m}{m-2}\Big)^2\frac{1}{rn\Delta h}.$$

This converges to 0, since $r \to \infty$ and $n\Delta h \to \infty$. Now consider the remaining portion of (5.11):

$$\Big(\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\frac{m-2}{h^2}\sqrt{\Delta}\Big)^2 = \frac{n}{h^3r^2}$$

which converges to 0 by the third assumption made on the bandwidth h made in Theorem 1.

Studying the C_j, D_j and E_j is more delicate, requiring a first order Taylor expansion of K and a closer analysis of the behaviour of the diffusion component of the X process. Specifically, when computing expected values, we will use the following.

$$0 = E\Big(\int_a^b \sigma(X_s)\, dW_s \,\Big|\, X_a \Big) = E\Big(\int_a^b \sigma(X_s)\, dW_s \,\Big|\, X_t, t \le a \Big). \tag{5.12}$$

To study the C_j, first write

$$C_j = \big(\overline{X}_{j+2} - \overline{X}_{j+1}\big) K_h'(\xi_j - x) \big(\overline{X}_j - X_{(j-1)r\Delta}\big), \tag{5.13}$$

where ξ_j is a value between $X_{(j-1)r\Delta}$ and $\overline{X}_j$. By the integral expression for X, for $l \ge k$,

$$X_{lr\Delta} - X_{kr\Delta} = \int_{kr\Delta}^{lr\Delta} \mu(X_s)\, ds + \int_{kr\Delta}^{lr\Delta} \sigma(X_s)\, dW_s \equiv \mathscr{M}_{kr}^{lr} + \mathscr{W}_{kr}^{lr}.$$

So we can write

$$\overline{X}_{j+2} - \overline{X}_{j+1} = \frac{1}{r} \sum_{i=1}^{r} \mathscr{M}_{jr+i}^{(j+1)r+i} + \frac{1}{r} \sum_{i=1}^{r} \mathscr{W}_{jr+i}^{(j+1)r+i}$$

and

$$\overline{X}_j - X_{(j-1)r\Delta} = \frac{1}{r} \sum_{k=1}^{r} \mathscr{M}_{(j-1)r}^{(j-1)r+k} + \frac{1}{r} \sum_{k=1}^{r} \mathscr{W}_{(j-1)r}^{(j-1)r+k}.$$

Substituting these in the expression for C_j and expanding yields

$$C_j = C_{1j} + C_{2j}$$

where

$$\begin{aligned} C_{1j} = \frac{1}{r^2} \sum_{i=1}^{r} \sum_{k=1}^{r} \Big[&\mathscr{M}_{jr+i}^{(j+1)r+i} \mathscr{M}_{(j-1)r}^{(j-1)r+k} + \mathscr{M}_{jr+i}^{(j+1)r+i} \mathscr{W}_{(j-1)r}^{(j-1)r+k} \\ &+ \mathscr{W}_{jr+i}^{(j+1)r+i} \mathscr{M}_{(j-1)r}^{(j-1)r+k} \Big] K_h'(\xi_j - x) \end{aligned}$$

and

$$C_{2j} = \frac{1}{r^2} \sum_{i=1}^{r} \sum_{k=1}^{r} \mathscr{W}_{jr+i}^{(j+1)r+i} \mathscr{W}_{(j-1)r}^{(j-1)r+k} K_h'(\xi_j - x).$$

To bound the terms in C_{1j}, we use the fact that K_h' is bounded by M/h^2 and also use the Cauchy–Schwarz inequality and (5.4) and (5.5). For instance,

$$\begin{aligned}
\mathrm{E}\left|\mathscr{M}_{jr+i}^{(j+1)r+i}\mathscr{M}_{(j-1)r}^{(j-1)r+k}\right| &\leq \sqrt{\mathrm{E}\Big(\Big[\mathscr{M}_{jr+i}^{(j+1)r+i}\Big]^2\Big)\mathrm{E}\Big(\Big[\mathscr{M}_{(j-1)r}^{(j-1)r+k}\Big]^2\Big)} \\
&\leq E(\mu^2(X_0))r^2\Delta^2
\end{aligned}$$

by (5.4). By the same reasoning, $\mathrm{E}\big|\mathscr{M}_{jr+i}^{(j+1)r+i}\mathscr{W}_{(j-1)r}^{(j-1)r+k}\big|$ and $\mathrm{E}\big|\mathscr{W}_{jr+i}^{(j+1)r+i}\mathscr{M}_{(j-1)r}^{(j-1)r+k}\big|$ are both bounded by a constant times $(r\Delta)^{3/2}$. Since $r\Delta \to 0$, $(r\Delta)^{3/2}$ is of larger magnitude than $r^2\Delta^2$. Therefore, for some constant $\mathscr{C}$, the contribution of the C_{1j} to (5.10) is bounded by

$$\begin{aligned}
\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\mathrm{E}\Big(\sum_{j=1}^{m-2}|C_{1j}|\Big) &\leq \mathscr{C}M\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}(m-2)(r\Delta)^{3/2}\frac{1}{h^2} \\
&= \mathscr{C}M\Big[\Big(\frac{n\Delta}{h}\Big)^2 r\Delta\frac{1}{n\Delta h}\Big]^{1/2} \to 0
\end{aligned}$$

since $n\Delta h \to \infty$ and

$$\Big(\frac{n\Delta}{h}\Big)^2 r\Delta = \Big(\frac{n\Delta}{h}\Big)^2 r\Delta\log\Big(\frac{1}{r\Delta}\Big)\frac{1}{\log\Big(\frac{1}{r\Delta}\Big)} \to 0, \qquad (5.14)$$

by the conditions of Theorem 1 and Assumption 1, that $r\Delta \to 0$.

To bound $\sum_{j=1}^{m-2}C_{2j}$ we will use the fact that, for fixed i and k and for $j \neq l$,

$$\mathrm{E}\Big(\mathscr{W}_{jr+i}^{(j+1)r+i}\mathscr{W}_{(j-1)r}^{(j-1)r+k}K_h'(\xi_j - x)\mathscr{W}_{lr+i}^{(l+1)r+i}\mathscr{W}_{(l-1)r}^{(l-1)r+k}K_h'(\xi_l - x)\Big) = 0. \qquad (5.15)$$

To see this, suppose that $j > l$. Then

$$\begin{aligned}
&\mathrm{E}\Big(\mathscr{W}_{jr+i}^{(j+1)r+i}\mathscr{W}_{(j-1)r}^{(j-1)r+k}K_h'(\xi_j - x) \\
&\qquad \mathscr{W}_{lr+i}^{(l+1)r+i}\mathscr{W}_{(l-1)r}^{(l-1)r+k}K_h'(\xi_l - x)\Big|X_t, t \leq (jr+i)\Delta\Big) = 0
\end{aligned}$$

by (5.12), since $\mathscr{W}_{(j-1)r}^{(j-1)r+k}$, $K_h'(\xi_j - x)$, $\mathscr{W}_{lr+i}^{(l+1)r+i}$, $\mathscr{W}_{(l-1)r}^{(l-1)r+k}$, and $K_h'(\xi_l - x)$ all depend on X_t with $t \leq (jr+i)\Delta$.

To bound $\sum_{j=1}^{m-2} C_{2j}$, we first interchange the order of summation:

$$\sum_{j=1}^{m-2} C_{2j} = \frac{1}{r^2} \sum_{i=1}^{r} \sum_{k=1}^{r} \sum_{j=1}^{m-2} \mathscr{W}_{jr+i}^{(j+1)r+i} \mathscr{W}_{(j-1)r}^{(j-1)r+k} K_h'(\xi_j - x) \tag{5.16}$$

and bound the expectation of the absolute value of the inner sum, using, in order, the Cauchy–Schwarz inequality, (5.15), (5.6), the independence of $\mathscr{W}_{jr+i}^{(j+1)r+i}$, and $\mathscr{W}_{(j-1)r}^{(j-1)r+k}$, and (5.5):

$$\begin{aligned}
&E\Big|\sum_{j=1}^{m-2} \mathscr{W}_{jr+i}^{(j+1)r+i} \mathscr{W}_{(j-1)r}^{(j-1)r+k} K_h'(\xi_j - x)\Big| \\
&\le \Big(E\Big[\sum_{j=1}^{m-2} \mathscr{W}_{jr+i}^{(j+1)r+i} \mathscr{W}_{(j-1)r}^{(j-1)r+k} K_h'(\xi_j - x)\Big]^2\Big)^{1/2} \\
&= \Big(E\sum_{j=1}^{m-2} \Big[\mathscr{W}_{jr+i}^{(j+1)r+i} \mathscr{W}_{(j-1)r}^{(j-1)r+k} K_h'(\xi_j - x)\Big]^2\Big)^{1/2} \\
&\le \frac{M}{h^2}\Big(\sum_{j=1}^{m-2} E\Big[\mathscr{W}_{jr+i}^{(j+1)r+i} \mathscr{W}_{(j-1)r}^{(j-1)r+k}\Big]^2\Big)^{1/2} \\
&= \frac{M}{h^2}\Big(\sum_{j=1}^{m-2} E\Big[\mathscr{W}_{jr+i}^{(j+1)r+i}\Big]^2 E\Big[\mathscr{W}_{(j-1)r}^{(j-1)r+k}\Big]^2\Big)^{1/2} \le \frac{M}{h^2}(m-2)^{1/2} r\Delta E[\sigma^2(X_0)].
\end{aligned}$$

So the contribution of the C_{2j} to (5.10) is bounded by

$$\begin{aligned}
&\frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\Big|\sum_{j=1}^{m-2} E(C_{2j})\Big| \le \frac{\sqrt{n\Delta h}}{(m-2)r\Delta}\frac{M}{h^2}(m-2)^{1/2} r\Delta E(\sigma^2(X_0)) \\
&= ME(\sigma^2(X_0))\sqrt{\frac{m}{m-2}}\sqrt{\Big(\frac{n\Delta}{h}\Big)^2 r\Delta h \frac{1}{(n\Delta h)^2}}
\end{aligned}$$

which converges to 0 by (5.14) and the assumptions that $h \to 0$ and $n\Delta h \to \infty$.

The analyses of $\sum_{j=1}^{m-2} D_j$ and $\sum_{j=1}^{m-2} E_j$ are similar so we only present the analysis of the contribution of the D_j to (5.10). Write

$$\begin{aligned}\sum_{j=1}^{m-2} D_j &= \sum_{j=1}^{m-2} [\bar{X}_{j+2} - X_{(j+1)r\Delta}] K_h\big(X_{(j-1)r\Delta} - x\big)\\ &\quad - \sum_{j=1}^{m-2} [\bar{X}_{j+1} - X_{jr\Delta}] K_h\big(X_{(j-1)r\Delta} - x\big)\\ &= \sum_{j=2}^{m-2} [\bar{X}_{j+1} - X_{jr\Delta}] \big[K_h(X_{(j-2)r\Delta} - x) - K_h(X_{(j-1)r\Delta} - x)\big]\\ &\quad + [\bar{X}_m - X_{(m-1)r\Delta}] K_h(X_{(m-3)r\Delta} - x)\\ &\quad - [\bar{X}_2 - X_{r\Delta}] K_h(X_0 - x). \end{aligned} \tag{5.17}$$

We can write the first term in (5.17) as

$$-\sum_{j=2}^{m-2} [\bar{X}_{j+1} - X_{jr\Delta}] K_h'\big(\xi_j - x\big)\ [X_{(j-1)r\Delta} - X_{(j-2)r\Delta}]$$

for some ξ_j between $X_{(j-2)r\Delta}$ and $X_{(j-1)r\Delta}$. This sum's contribution to (5.10) converges to 0 in probability, by the same argument that was used in bounding $\sum_{j=1}^{m-2} C_j$, in the calculations following equation (5.13).

To study the contribution of the last two terms of (5.17), write

$$E\big|[\bar{X}_{l+1} - X_{lr\Delta}] K_h(X_{kr\Delta} - x)\big| \le \frac{M}{h} E\big|\bar{X}_{l+1} - X_{lr\Delta}\big| \le \frac{M}{h}\sqrt{\beta r\Delta}$$

by Lemma 1 (2) and (5). Thus, we can bound the contribution to (5.10) of the last two terms of (5.17) by a constant times

$$\frac{\sqrt{n\Delta h}}{mr\Delta}\frac{1}{h}\sqrt{r\Delta} = \sqrt{\frac{n}{m^2hr}} = \sqrt{\frac{r\Delta}{nh\Delta}}$$

since $m = n/r$. This converges to 0 since $r\Delta \to 0$ and $nh\Delta \to \infty$.

Acknowledgements The research of W. Lee and N. Heckman was supported by the Natural Sciences and Engineering Research Council of Canada. We thank the referees for reading the paper carefully and for suggesting a number of improvements.

References

1. Aït-Sahalia, Y. (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica* **64**, 527–560.

2. Arfi, M. (1995). Nonparametric drift estimation from ergodic samples. *J. Nonparametr. Statist.* **5**, 381–389.

3. Bandi, F. M. and Nguyen, T. H. (2003). On the functional estimation of jump-diffusion models. *J. Econometrics* **116**, 293–328.

4. Bandi, F. M. and Phillips, P. C. B. (2003). Fully nonparametric estimation of scalar diffusion models. *Econometrica* **71**, 241–283.

5. Bandi, F. M. and Phillips, P. C. B. (2009a). Nonstationary continuous-time processes. In: *Handbook of Financial Econometrics, Vol. 1* (Y. Aït-Sahalia and L. P. Hansen, eds.), 139–201, North-Holland/Elsevier, Amsterdam.

6. Bandi, F. M., Corradi, V. and Moloche, G. (2009b). Bandwidth selection for continuous time Markov processes. Preprint, Department of Economics, University of Warwick.

7. Bhan, C. and Mandrekar, V. (2010). Recurrence properties of term structure models. Int. J. Contemp. Math. Sci. 5 (2010), no. 33-36, 16451652.

8. Burman, P., Chow, E. and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika* **81**, 351–358.

9. Chapman, D. A. and Pearson, N. D. (2000). Is the short rate drift actually nonlinear? *J. Finance* **55**, 355–388.

10. Chu, C.-K. and Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.* **19**, 1906–1918.

11. Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**, 514–543.

12. Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2012). Non-parametric estimation of the coefficients of ergodic diffusion processes based on high-frequency data. In: *Statistical Methods for Stochastic Differential Equations*, (M. Kessler, A. Lindner and M. Sørensen, eds.), 341–381, Monogr. Statist. Appl. Probab. **124**, CRC Press, Boca Raton.
13. Dalalyan, A. (2005). Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Statist.* **33**, 2507–2528.
14. Dalalyan, A. S. and Kutoyants, Y. A. (2002). Asymptotically efficient trend coefficient estimation for ergodic diffusion. *Math. Methods Statist.* **1**, 402–427.
15. Friz, P. and Victoir, N. (2005). Approximations of the Brownian rough path with applications to stochastic analysis. *Ann. Inst. H. Poincaré Probab. Statist.* **41**, 703–724.
16. Hoffmann, M. (1999). Adaptive estimation in diffusion processes. *Stochastic Process. Appl.* **79**, 135–163.
17. Iacus, S. M. (2008). *Simulation and Inference for Stochastic Differential Equations. With R examples.* Springer Series in Statistics, Springer, New York.
18. Iacus, S. M. (2014). sde: Simulation and Inference for Stochastic Differential Equations. R package version 2.0.13, url = http://CRAN.R-project.org/package=sde.
19. Jin, P., Mandrekar, V., Rüdiger, B. and Trabelsi, C. (2013). Positive Harris recurrence of the CIR process and its applications. *Commun. Stoch. Anal.* **7**, 409–424.
20. Jacod, J., Li, Y., Mykland, P. A., Podolskij, M. and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Process. Appl.* **119**, 2249–2276.
21. Jacod, J. and Protter, P. (2012). *Discretization of Processes.* Stochastic Modelling and Applied Probability **67**, Springer, Heidelberg.
22. Jones, C. S. (2003). Nonlinear mean reversion in the short-term interest rate. *Rev. Financial Stud.* **16**, 793–843.
23. Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91**, 401–407.
24. Moloche, G. (2001). Local nonparametric estimation of scalar diffusions. Preprint, Munich Personal RePEc Archive.
25. Podolskij, M. and Vetter, M. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli* **15**, 634–658.

26. Schmisser, E. (2011). Non-parametric drift estimation for diffusions from noisy data. *Statist. Decisions* **28**, 119–150.
27. Schmisser, E. (2013). Penalized nonparametric drift estimation for a multidimensional diffusion process. *Statistics* **47**, 61–84.
28. Schmisser, E. (2014). Non-parametric adaptive estimation of the drift for a jump diffusion process. *Stochastic Process. Appl.* **124**, 883–914.
29. Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *J. Finance* **52**, 1973–2002.
30. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, with discussion. *J. Roy. Statist. Soc. Ser. B* **36**, 111–147.
31. Strauch, C. (2015). Sharp adaptive drift estimation for ergodic diffusions: the multivariate case. *Stochastic Process. Appl.* **125**, 2562–2602.
32. Strauch, C. (2016). Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions. *Probab. Theory Related Fields* **164**, 361–400.
33. Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *J. Amer. Statist. Assoc.* **100**, 1394–1411.
34. Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *J. Bus. Econom. Statist.* **14**, 45–52.