Is the Short Rate Drift Actually Nonlinear?
Author(s): David A. Chapman and Neil D. Pearson
Source: *The Journal of Finance*, Vol. 55, No. 1 (Feb., 2000), pp. 355-388
Published by: Blackwell Publishing for the American Finance Association
Stable URL: http://www.jstor.org/stable/222558
Accessed: 24/11/2009 12:22

Is the Short Rate Drift Actually Nonlinear?

DAVID A. CHAPMAN and NEIL D. PEARSON*

ABSTRACT

Aït-Sahalia (1996) and Stanton (1997) use nonparametric estimators applied to short-term interest rate data to conclude that the drift function contains important nonlinearities. We study the finite-sample properties of their estimators by applying them to simulated sample paths of a square-root diffusion. Although the drift function is linear, both estimators suggest nonlinearities of the type and magnitude reported in Aït-Sahalia (1996) and Stanton (1997). Combined with the results of a weighted least squares estimator, this evidence implies that nonlinearity of the short rate drift is not a robust stylized fact.

A COMMON APPROACH IN MODELING THE TERM STRUCTURE of interest rates and pricing interest rate derivatives is to express interest rates in terms of one or more state variables that follow continuous-time Markov processes. In time homogeneous one-factor models there is only one state variable, which is usually taken to be the short or instantaneous rate of interest. This is the case, for example, in Vasicek (1977), Cox, Ingersoll, and Ross (1985) (hereafter, CIR), the translated CIR model discussed in Pearson and Sun (1994), Brennan and Schwartz (1979), Courtadon (1982), the time homogenous continuous-time versions of the Black, Derman, and Toy (1990) and Black and Karasinski (1991) models, and in the empirical models considered by Chan et al. (1992). In these and similar models, the properties of the interest rate process are determined entirely by the drift and diffusion functions (defined below). Thus, the problem of selecting among the models above, or determining that none of them is appropriate and that an alternative model is needed, comes down to choosing or estimating the drift and diffusion functions.

Unfortunately, theory provides little guidance about these choices. The appropriate specification of the drift and diffusion remains, for the most part, an unanswered question. At least partly for these reasons, Aït-Sahalia (1996) and Stanton (1997) have recently proposed nonparametric estimators of the drift and diffusion functions. A key finding in these papers is that the

* Chapman is from the University of Texas at Austin and Pearson is from the University of Illinois at Urbana-Champaign. We thank Yacine Aït-Sahalia, Jonathan Berk, Murray Carlson, John Cochrane, Andy Filardo, Eric Hughson, Narasimhan Jegadeesh, Chris Jones, George Pennachi, René Stulz (the editor), Robert Whitelaw, an anonymous referee, seminar participants at the University of Illinois at Urbana-Champaign, the University of Texas at Austin, the University of Utah, and the "Asset Pricing" session of the 1998 NBER Summer Institute, and in particular, Matt Pritsker for helpful comments and discussions.

estimated drift function is highly nonlinear, especially for large values of the interest rate process. Stanton finds that the estimated drift drops sharply as the interest rate increases beyond about 14 percent, while Aït-Sahalia rejects all of the parametric models he considers, and finds that "[t]he linearity of the drift imposed in the literature appears to be the main source of misspecification."

These results are inconsistent with *all* of the models cited above. In Vasicek (1977), CIR, Pearson and Sun (1994), Brennan and Schwartz (1979), Courtadon (1982), and in the empirical models considered by Chan et al. (1992), the drift is linear, and in the continuous-time versions of the Black et al. (1990) and Black and Karasinski (1991) models, the drift of the natural log of the interest rate is linear (in the natural log). Other than the flexible parametric specification introduced by Aït-Sahalia (1996), we are not aware of any parametric model that is consistent with the drift and diffusion functions estimated by Aït-Sahalia (1996) and Stanton (1997), and it is tempting to conclude that the existing set of interest rate models is inadequate. On the other hand, these new results may, in part, be artifacts of the estimation procedure, rather than fundamental features of the data.

We perform a Monte Carlo study of the finite sample properties of the nonparametric estimators of Aït-Sahalia (1996) and Stanton (1997). Applying the exact form of the kernel regression estimator proposed in Stanton to repeated simulated sample paths of a CIR square root process, we find that the typical estimated drift function displays nonlinearities at high interest rates of exactly the sort found in Stanton, even though the true drift is linear. A consistent result for Stanton's estimator is that its performance improves with decreases in the persistence of the underlying process and increases in the sample size. However, the spurious nonlinearity remains for sample sizes that are twice as large as any currently available interest rate data set. The explanation for the poor performance of this nonparametric estimator is a truncation of the distribution that occurs in finite samples but is eliminated asymptotically. This truncation of the data implies that the kernel regression residuals are correlated with the regressor. This bias is particularly severe for data generated by a very persistent underlying process.

As we explain below, this issue is distinct from the well-known increase in the bias of the kernel regression estimators near the boundary of the support of the data. In particular, in order to account for this "boundary effect," we repeat the Monte Carlo experiment using the jackknife kernel proposed in Rice (1984). This estimator offers no reduction in the spurious nonlinearity of the estimated drift function. In fact, for some bandwidth choices, reducing the boundary bias increases the overall bias because the boundary bias partially offsets the truncation bias. However, using the jackknife kernel to correct the boundary bias has a uniformly positive effect on the kernel regression estimates of the diffusion function. In fact, estimating the diffusion function is the notable success of the kernel regression approach.

The estimator in Aït-Sahalia (1996) also suggests that the drift function contains important nonlinearities. In order to understand its finite sample properties, we apply a simplified version of this estimator to the simulated square root sample paths. The results of this exercise demonstrate that there is substantial uncertainty associated with the drift parameters due to multicollinearity induced by the functional form chosen in Aït-Sahalia. In particular, the magnitudes and signs of the parameter estimates are implausibly large and their standard deviations are also large. Considering the point estimates alone from this estimator can easily suggest important nonlinearities where none exist in the data.

A reasonable conclusion to draw from the Monte Carlo evidence is that it is difficult to use the nonparametric estimators in Aït-Sahalia (1996) and Stanton (1997) to produce reliable inferences about the question asked in the title of this paper. In order to provide additional evidence on the nature of the short rate drift, we apply a weighted least squares (WLS) estimator, similar to the one used in Chan et al. (1992), to the data from both Aït-Sahalia and Stanton. A separate Monte Carlo study of this estimator suggests that it does not display spurious nonlinearity. The evidence from the WLS estimator is not consistent with the results of the nonparametric estimators, in that it is difficult to find statistically reliable evidence against a constant drift in both data sets. However, the point estimates of the drift coefficients are quite different, and the WLS results are not conclusive.

Overall, our results provide a clear warning about the application of nonparametric estimators to time series data and, in particular, to the need for cautious interpretation of the results. These techniques are inherently ill equipped to provide reliable information about the properties of any time series in the extreme tails of the distribution. Unfortunately, this is precisely where the nonlinearity appears in estimates of the short rate drift. This problem is exacerbated when the data consist of high-frequency observations of a very persistent process. When considered in conjunction with the WLS results, the general conclusion for interest rate modeling is that time series methods applied to the short rate alone are incapable of resolving the issue of the nonlinearity of the drift.

If these methods imply little about the tail behavior of interest rates, then we are left with their implications for the shape of the drift in the heart of the distribution. Here, both Aït-Sahalia's and Stanton estimators agree: there is virtually no evidence of mean reversion in short-term interest rates. Nonetheless, there are strong theoretical reasons to believe that short rates cannot exhibit the asymptotically explosive behavior implied by a random walk model. The answer to resolving this very fundamental question lies in considering the identifying restrictions imposed simultaneously on the properties of short and long rates by an economic model of the term structure of interest rates.

Jones (1998) addresses the question of the linearity or nonlinearity of the drift using a Bayesian approach. The results of his analysis suggest that the finding of nonlinearity is highly dependent on whether or not stationarity is imposed in the prior distribution. Consistent with the findings reported in this paper, Jones concludes by stating: "Despite the fact that over five thousand observations of daily data are available, the data sample is effectively small. Precise statements about the shape of the drift function . . . are impossible to make."

The balance of the paper is organized as follows: Section I introduces notation, some basic concepts from the kernel density estimation and kernel regression literature, and the estimator developed in Stanton (1997). The Monte Carlo simulation of Stanton's estimator is contained in Section II, and Section III presents the results for the jackknife kernel regression estimator. Section IV provides an explanation for the finite sample nonlinearity of the kernel regression estimators, and Section V examines a simplified version of the estimator in Aït-Sahalia (1996). Section VI presents the results of a Monte Carlo simulation of two variants of a simple discretized least squares estimator, and also contains an application of the preferred estimator to the data in Aït-Sahalia and Stanton. Section VII briefly concludes.

I. The Estimator in Stanton (1997)

Let $\{x_t; t \geq 0\}$ be defined as the unique, time homogeneous Markov process that solves a stochastic differential equation (SDE) of the form

$$dx_t = \mu(x_t)dt + \sigma(x_t)dB_t, \tag{1}$$

where $\{B_t; t \geq 0\}$ is a scalar Brownian motion, $\mu : \mathbb{R} \rightarrow \mathbb{R}$ is the *drift function*, and $\sigma : \mathbb{R} \rightarrow \mathbb{R}_+$ is the *diffusion function*. In this paper, x_t is interpreted as the "instantaneous" or "short" rate of interest.

Under technical conditions on the transition function, x_t is a *diffusion process*. This implies that

$$E[x_{t+\Delta} - x_t | x_t] = \mu(x_t)\Delta + o(\Delta) \tag{2}$$

and

$$E[(x_{t+\Delta} - x_t)^2 | x_t] = \sigma^2(x_t)\Delta + o(\Delta), \tag{3}$$

where Δ is a discrete (but arbitrarily small) time step in a sequence of observations of the process x_t and $o(\Delta)$ is the asymptotic order symbol used to denote a function ζ such that $\lim_{\Delta \downarrow 0} \zeta(\Delta)/\Delta = 0$.

Let $\{x_t^\Delta\}_{t=1}^T$ be a sample of size T from the continuous time process x_t, observed at the discrete interval Δ. Furthermore, let $\{z_i\}_{i=1}^N$ be a set of N points defining an equally spaced partition of a subset of the support of the stationary density.[1] If the stationary density of x_t is denoted $\pi(x)$, a kernel estimator is of the form

$$\hat{\pi}(z_i) \equiv \frac{1}{T \cdot h} \sum_{t=1}^{T} K\left(\frac{z_i - x_t^\Delta}{h}\right), \tag{4}$$

[1]In many applications, the support of the stationary density is $(0,\infty)$. In practice, the partition points $\{z_i\}_{i=1}^N$ are chosen to capture virtually all of the probability mass of the stationary density.

for $i = 1,2,\ldots,N$, where K is a *kernel function* satisfying the condition

$$\int_{-\infty}^{\infty} K(y)\,dy = 1.$$

The kernel function provides a method of weighting "nearby" observations in order to construct a smoothed histogram, which is the density estimator in equation (4). Stanton (1997) uses a *Gaussian kernel*:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right); \qquad u \in (-\infty,+\infty). \tag{5}$$

The parameter h is called the *smoothing parameter* or the *bandwidth* of the density estimator, and h determines the width of the kernel function around any partition point z_i. It specifies how (and how many) "neighboring" points of x_t^Δ are to be considered in constructing the density estimator at z_i.

The nonparametric density estimator is completely defined by the pair (K,h). The choice of the bandwidth h is inherently subjective, and there is no single accepted approach to its selection.[2] A common selection method is to use a version of a data-dependent rule called "least squares cross validation" (hereafter, LSCV). The first step in this procedure is to estimate the regression function for a set of bandwidths $\{h_1,\ldots,h_k\}$. The LSCV bandwidth is chosen as the solution to the following problem:

$$\min_{\{h_i\}} T^{-1} \sum_{t=1}^{T} [y_t - \hat{m}_{h_i}(x_t)]^2 \omega(x_t) \Xi(T^{-1}h_i), \tag{6}$$

where $i = 1,\ldots,k$, $\hat{m}_{h_i}(x_t)$ is the fitted value of the kernel regression estimated at x_t, $\omega(x_t)$ is a function that weights the observation at x_t, and $\Xi(T^{-1}h_i)$ is a function that "penalizes" small bandwidths.

The role of ω is to reduce boundary biases in the bandwidth selection procedure by reducing the weight in the sum of squares in problem (6) associated with extreme levels of x_t. In the analysis that follows, ω is 1 if x_t is in the 5th to 95th percentile range and zero otherwise.[3] The penalty function is the Shibata model selector

$$\Xi(T^{-1}h_i) = 1 + 2T^{-1}h_i^{-1}K(0), \tag{7}$$

where $K(0)$ is the height of the Gaussian kernel at zero. Intuitively, in the absence of a function explicitly penalizing small bandwidths, the LSCV criterion function in problem (6) would always select the smallest available bandwidth since this would imply a fitted regression curve that tries to go through as many of the actual data points as possible.

[2] See Härdle (1990, Chap. 5) for a detailed description of alternative selection approaches.

[3] Figure 5.9 in Härdle (1990) suggests that the LSCV bandwidth choice is not sensitive to the percentage of observations ignored in the tails, as long as that percentage is in the range of two percent to eight percent in each tail.

In the Monte Carlo simulations that follow, three bandwidth choices are used for each parameter combination: The LSCV bandwidth, the "Stanton" bandwidth, and the independent and identically distributed (IID) bandwidth. The first choice is the solution to the least squares cross validation problem (6). The IID bandwidth is optimal for IID data, and it is defined as $h^{iid} = \hat{\sigma}T^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation of the data and T is the sample size. The Stanton bandwidth is the one actually used in Stanton (1997): $h^s = 4 \cdot \hat{\sigma}T^{-1/5}$.[4]

The estimators in Stanton (1997) are based directly on equations (2) and (3), above. In particular, "inverting" these equations yields:

$$\mu(x_t) = \frac{1}{\Delta} E[x_{t+\Delta} - x_t | x_t] + \frac{o(\Delta)}{\Delta} \tag{8}$$

and

$$\sigma(x_t) = \sqrt{E[(x_{t+\Delta} - x_t)^2 | x_t] \frac{1}{\Delta} + \frac{o(\Delta)}{\Delta}}. \tag{9}$$

The essence of Stanton's approach is to apply the Nadaraya–Watson (N-W) kernel regression estimator to construct nonparametric estimates of the conditional expectations in equations (8) and (9):

$$\hat{\mu}(z_i) = \frac{1}{\Delta} \frac{\sum_{t=1}^{T-1} (x_{t+1}^{\Delta} - x_t^{\Delta}) K\left(\frac{z_i - x_t^{\Delta}}{h}\right)}{\sum_{t=1}^{T-1} K\left(\frac{z_i - x_t^{\Delta}}{h}\right)} \tag{10}$$

and

$$\hat{\sigma}(z_i) = \sqrt{\frac{\sum_{t=1}^{T-1} (x_{t+1}^{\Delta} - x_t^{\Delta})^2 K\left(\frac{z_i - x_t^{\Delta}}{h}\right)}{\sum_{t=1}^{T-1} K\left(\frac{z_i - x_t^{\Delta}}{h}\right)} \frac{1}{\Delta}}, \tag{11}$$

for $i = 1, \ldots, N$, where $\{x_t^{\Delta}\}_{t=1}^{T}$ is a sample of size T from the continuous time process x_t, $\{z_i\}_{i=1}^{N}$ is a set of N points defining an equally spaced partition of a subset of the support of the stationary density, K is the Gaussian kernel defined by equation (5), and h is chosen in the manner described earlier.

[4] Though this is not the bandwidth parameter reported in Stanton (1997), Richard Stanton indicated in a private communication that he used h^s. He chose it using a heuristic smoothing approach.

I. A Monte Carlo Analysis of the Estimator in Stanton (1997)

A. *Simulating a Square Root Diffusion*

In order to evaluate the finite sample performance of Stanton's estimator, the SDE in equation (1) is assumed to have the form of a square root diffusion process, introduced into the term structure literature in CIR:

$$dx_t = \kappa(\theta - x_t)\,dt + \sigma\sqrt{x_t}\,dB_t, \tag{12}$$

where θ defines the long-run mean of x_t, κ determines the speed at which the process returns to the long-run mean, and σ helps to define the instantaneous variance of the process. By construction, the drift of this process is linear.

The unconditional moments of the square root process are

$$E[x] = \theta, \tag{13}$$

$$\mathrm{Var}[x] = \frac{\theta\sigma^2}{2\kappa}, \tag{14}$$

and

$$\mathrm{Corr}[x_{t+\Delta}, x_t] = \exp(-\kappa\Delta). \tag{15}$$

The choice of κ, the parameter that determines the persistence of the process, is particularly important.

For any given set of parameter values, a simulated sample path for the square root diffusion is constructed in two steps. First, an initial value is drawn from the steady state (Gamma) density. Then using this initial value, successive observations are simulated by drawing from the transition (noncentral chi squared) density. The simulated sample paths of equation (12) are constructed assuming that the length of time between observations of the diffusion is $\Delta = 1/250$, corresponding to daily observations. We consider simulation lengths of 7,500 and 15,000 daily observations. The first value is roughly the size of the data set in Stanton (1997), and the second value allows us to evaluate the effect on the estimator of doubling the sample size.

κ is set to either 0.21459 or 0.85837. The first value implies a (monthly) autocorrelation of the short rate of 0.982, which is consistent with the upper end of the range of estimates of this parameter based on U.S. interest rate data. The second choice of κ implies a first-order (monthly) autocorrelation coefficient that is equal to that of the Eurodollar data used in Aït-Sahalia (1996). The choice of θ is not particularly important for the Monte Carlo study, and it is set at 0.085711 in order to be consistent with the Aït-Sahalia data set. Given a (θ,κ) pair, the value of σ is chosen in order to set equation

(14) equal to the sample variance in the Aït-Sahalia data set. This implies choices of σ equal to 0.07830 and 0.15660, where the ordering is consistent with the ordering of the κ values.

B. The Results

The results from applying the kernel regression estimators from Stanton (1997) to the simulated square root data are shown in Figures 1 through 4. These figures all have the same structure. The left column of graphs reports the drift estimates for a sample of size 7,500, and the right column reports the same results for a sample of size 15,000. The first row of each figure reports the results for the LSCV bandwidth. The second row shows the results for the Stanton bandwidth, and the final row is for the IID bandwidth. Each graph within each figure also has a common structure. The solid line is the true drift function. The dashed line is the pointwise mean at each gridpoint in the support of the stationary density across the 5,000 simulations. The two dotted lines are the 25th and 75th percentile points at each gridpoint. Notice that the true drift functions are the same in every graph in each figure. This is a direct result of the choices of κ and θ in each parameterization.

The results for the drift estimates for different bandwidths are striking. In *all* cases, the kernel regression estimates exhibit some spurious nonlinearity. First, consider the left column in Figure 1. The LSCV bandwidth delivers the best performance, but the average drift function diverges from the linear drift for values of x in excess of 0.12. Equally important, the 25th and 75th percentile bands indicate that there is substantial variation in individual estimates of the drift function and that these estimates are particularly inaccurate for extremely high levels of x.

The Stanton bandwidth choice does less smoothing than the LSCV bandwidth.[5] It exhibits more extreme nonlinearity at high levels of x, and it also has an interquartile range that is increasing in the level of x. It would be easy to conclude from the Stanton bandwidth picture (in the absence of the plot of the true drift), that the drift function is nonlinear. The IID bandwidth picture presents the most spectacular evidence of spurious nonlinearity in the drift. Clearly, h^{iid} is a poor choice. However, the reason for including h^{iid} will be clear in Figures 3 and 4 when the diffusion functions are examined.

Figures 1 and 2 also report estimates of drift functions for larger sample sizes and lower persistence levels. These estimates are consistent with the results for the $\kappa = 0.21459$ and $T_{sim} = 7{,}500$ case. The right columns in both Figures 1 and 2 indicate that the estimator becomes more accurate and the dispersion in the estimates across the 5,000 simulated sample paths becomes smaller for larger sample sizes, although spurious nonlinearity is still evident in all panels. It is worth noting that for $T_{sim} = 15{,}000$, the LSCV and

[5] For example, the average value (standard deviation) of the LSCV bandwidth is 0.0230 (0.0049) across the 5,000 simulations used in producing the left column of Figure 1. The average value (standard deviation) of the Stanton bandwidth is 0.0193 (0.0057).

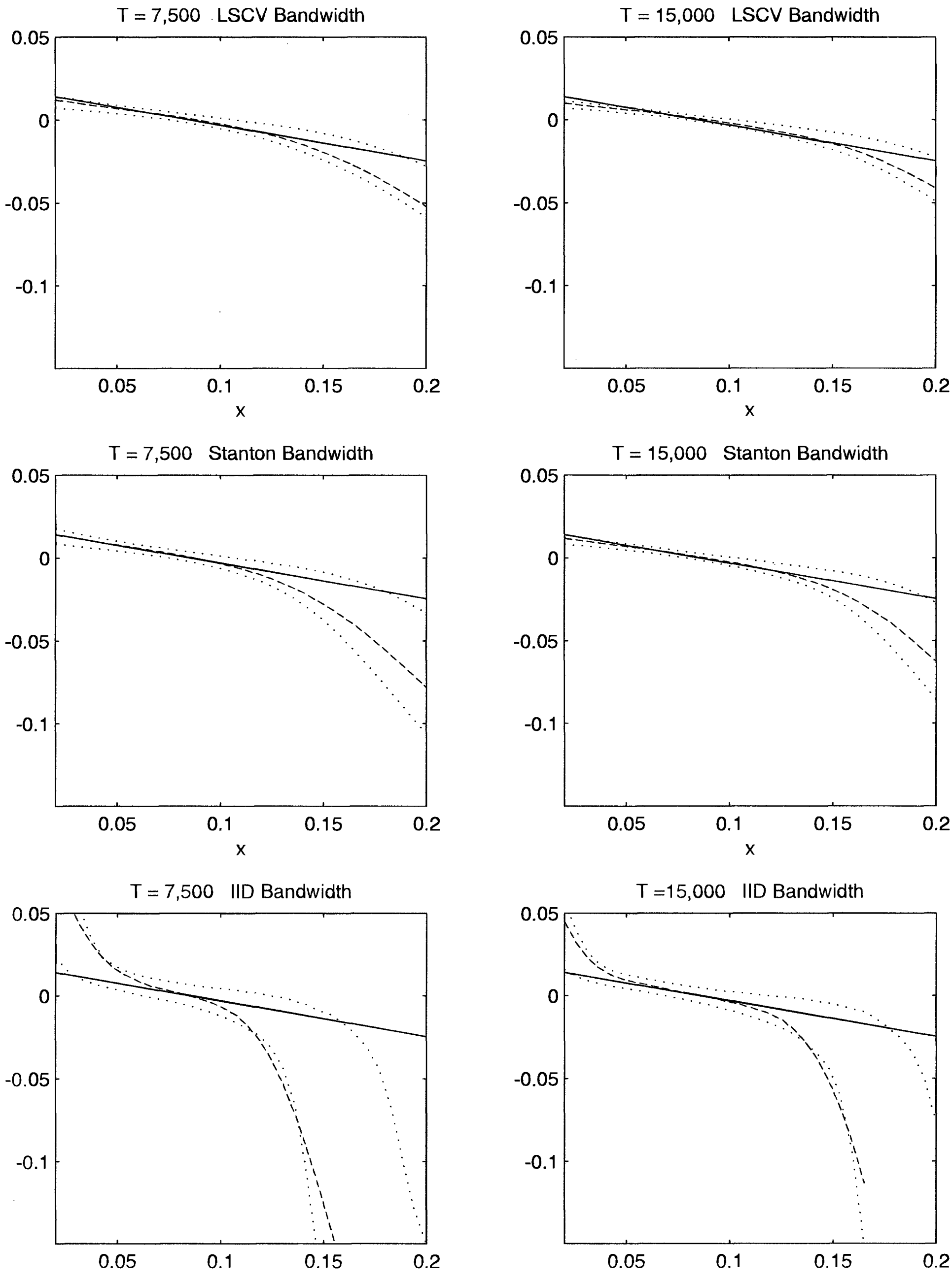

Figure 1. The drift function using the estimator in Stanton (1997) for true drift defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.21459$, $\theta = 0.08571$. In these simulations, the diffusion function is $\sigma(x) = \sigma\sqrt{x}$, where $\sigma = 0.07830$. The solid line is the true drift. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles across the 5,000 simulations.

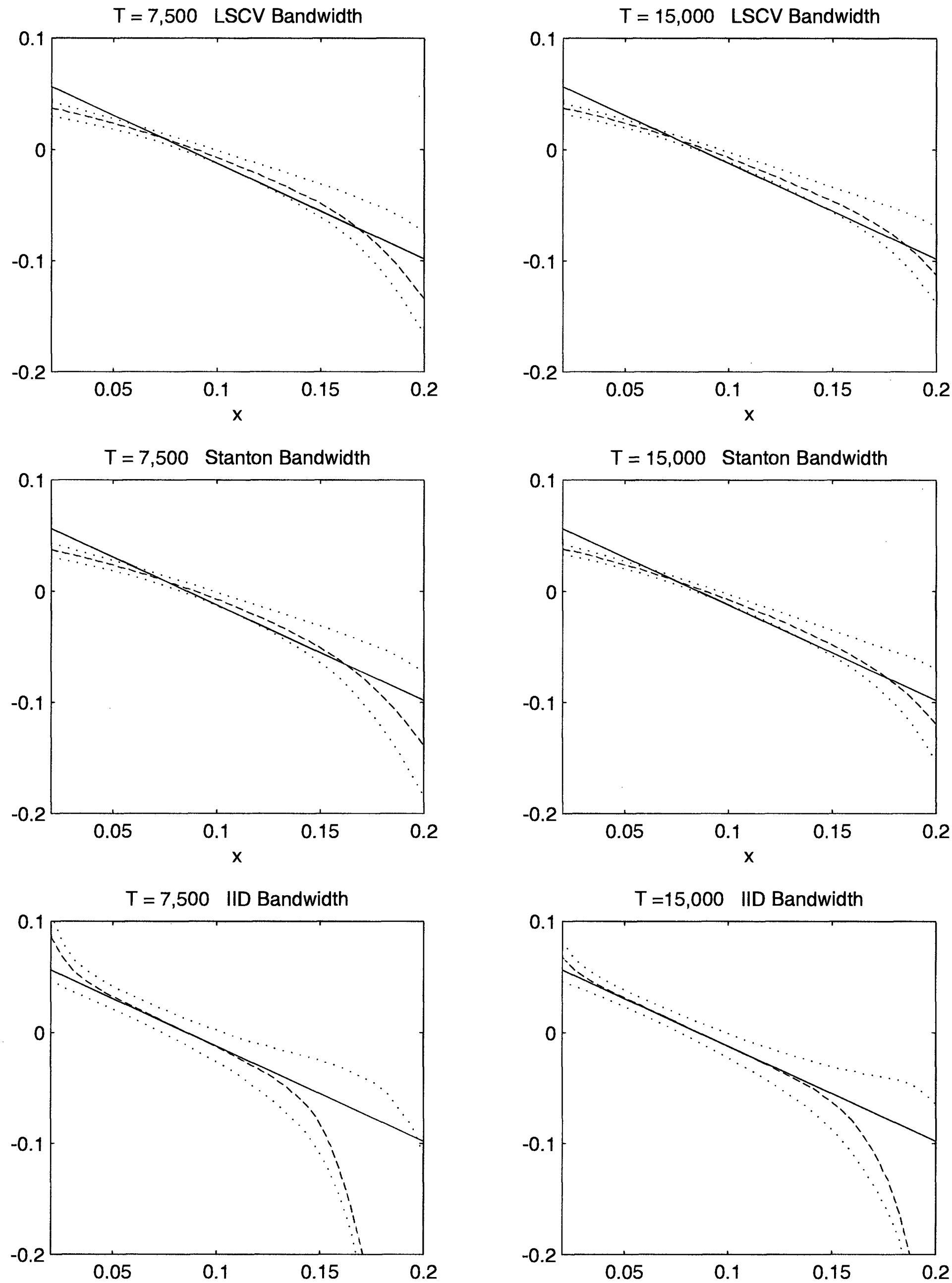

Figure 2. The drift function using the estimator in Stanton (1997) for true drift defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.85837$, $\theta = 0.08571$. In these simulations, the diffusion function is $\sigma(x) = \sigma\sqrt{x}$, where $\sigma = 0.15660$. The solid line is the true drift. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles across the 5,000 simulations.

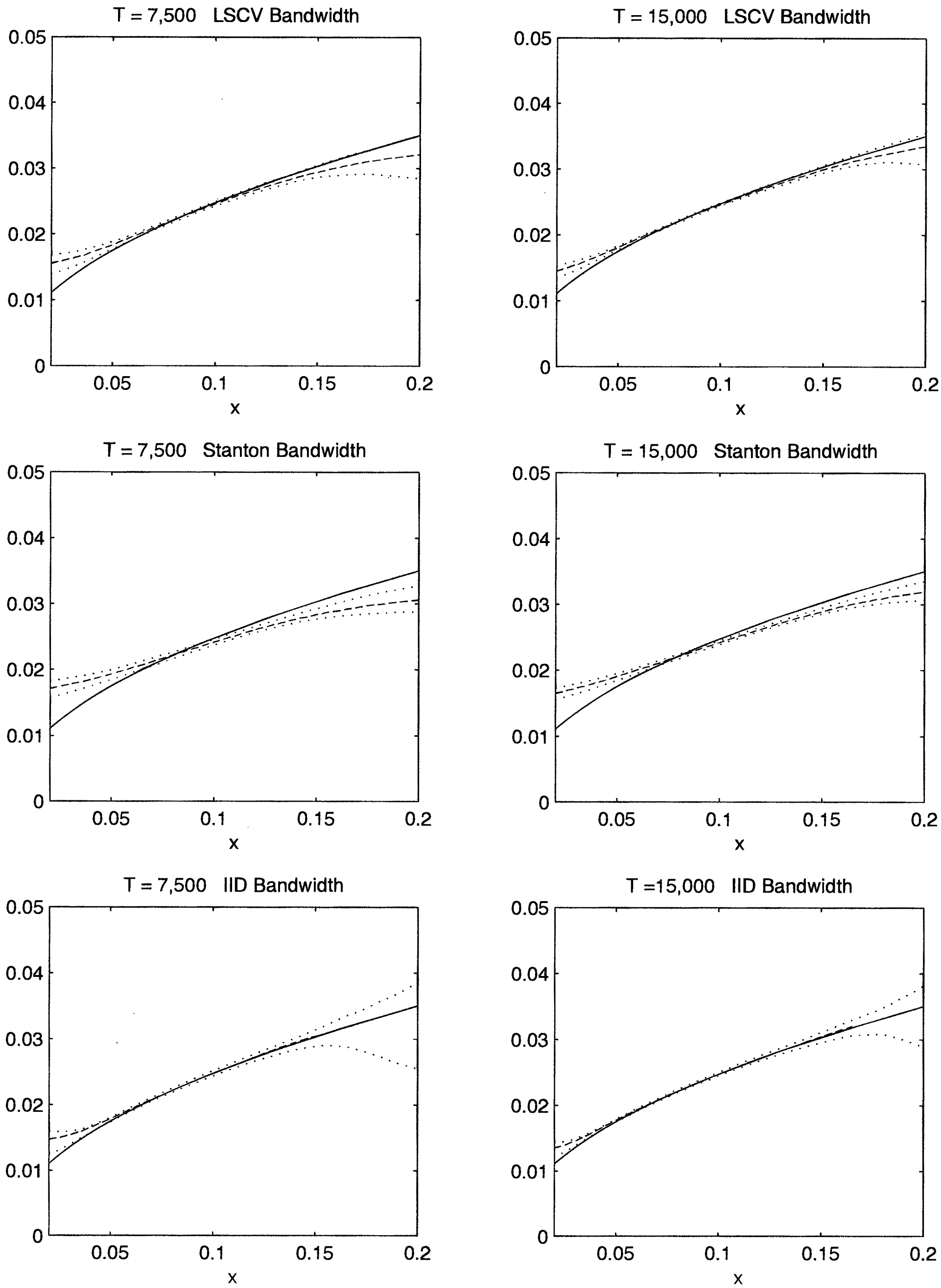

Figure 3. The diffusion function using the estimator in Stanton (1997) for the true diffusion function defined as $\sigma(x) = \sigma\sqrt{x}$, where $\sigma = 0.07830$. In these simulations, the true drift is defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.21459$ and $\theta = 0.08571$. The solid line is the true diffusion. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles.

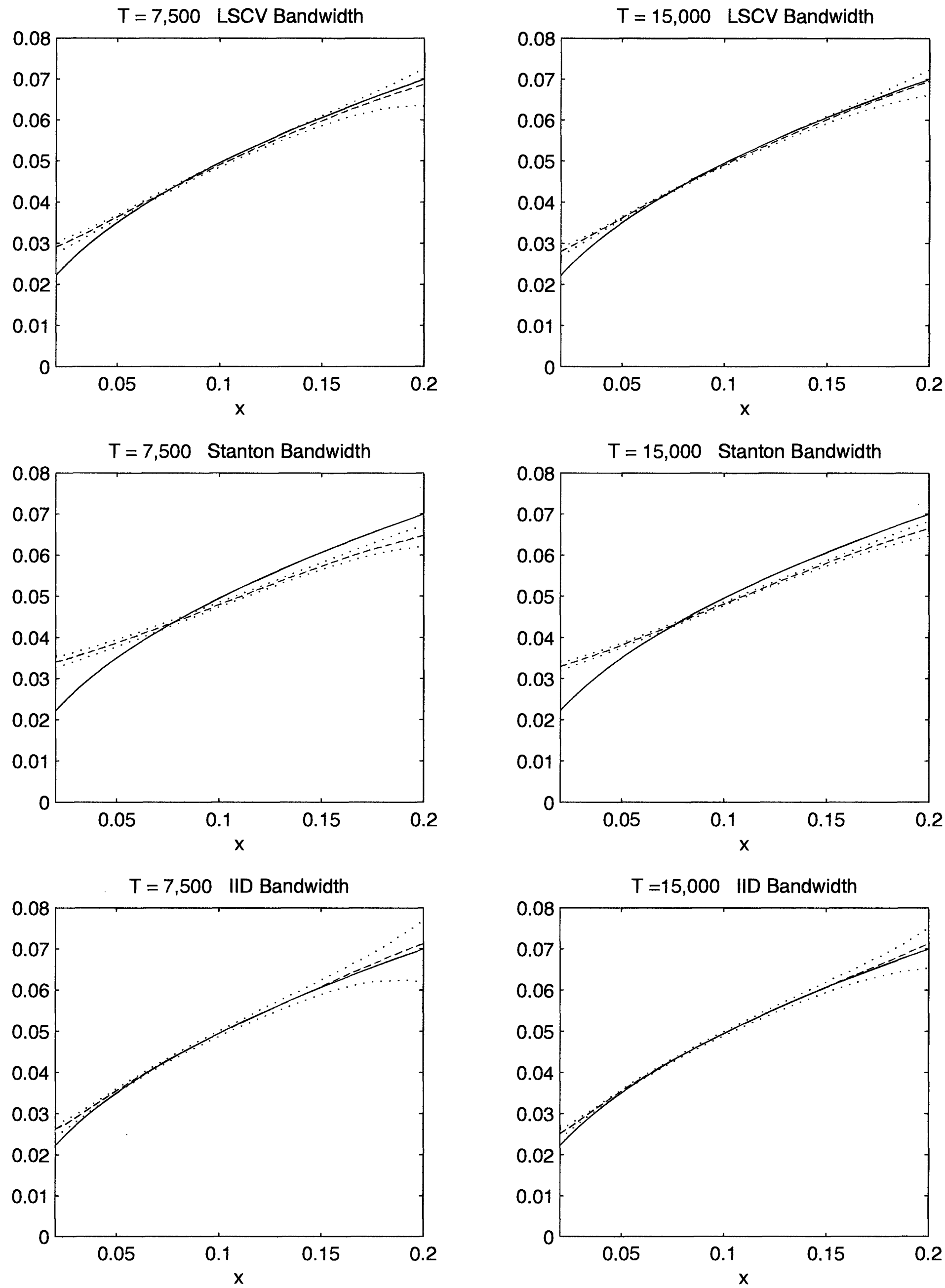

Figure 4. The diffusion function using the estimator in Stanton (1997) for the true diffusion function defined as $\sigma(x) = \sigma\sqrt{x}$, where $\sigma = 0.15660$. In these simulations, the true drift is defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.85837$ and $\theta = 0.08571$. The solid line is the true diffusion. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles across the 5,000 simulations.

Stanton bandwidth averages are very similar, indicating that $h^{LSCV} \approx h^S$ on average for larger sample sizes. The drift estimators based on the IID bandwidth also become more accurate as the persistence in the data decreases (as κ increases). Intuitively, this effect is similar to an increase in the (effective) sample size.

If drift estimation is the sole objective of the analysis, the prediction from these simulations is clear: Choose a large bandwidth that oversmooths the kernel regression function. Of course, this prescription is only appropriate if the true drift is linear, because what is really going on in Figures 1 and 2 (as is apparent from the diffusion results presented below) is that the oversmoothed nonparametric estimator has simply obliterated all of the detail in the estimated function. Even if the true drift had been nonlinear, the oversmoothed estimator would have suggested linearity.

Figures 3 and 4 show the results of applying the kernel regression estimators of the diffusion function to the simulated data sets. Generally speaking, they are the exact opposite of the drift estimation case. The IID estimator provides an accurate estimate of the diffusion function for the sample sizes and persistence parameters considered in these figures. The estimates based on the larger Stanton bandwidths are less accurate because they oversmooth the diffusion function. It is interesting that the data-dependent LSCV bandwidth also selects too much smoothing for the diffusion.[6] The primary failing of the diffusion estimator is its inability to capture the nonlinearity of the true diffusion at low levels of x_t. However, Stanton (1997) does provide an estimator constrained to go through the origin. Presumably, this would do a better job at matching the diffusion function for low levels of the short rate, although it would still underestimate the diffusion at high levels of x.

One possible explanation for the spurious nonlinearity is the well-known boundary value bias problem associated with kernel regression estimators. The boundary value bias occurs at points for which the kernel overlaps the boundary of the data.[7] The average value of the IID bandwidth for both choices of κ and for both sample sizes is roughly 0.005. As long as z is a distance of approximately 0.015 from the boundary, a Gaussian kernel using these bandwidths assigns a trivial weight to the boundary points, so that the boundary value bias should not be the sole explanation for the results in Figures 1 and 2. Also, when the true drift is linear, the direction of the boundary value bias near the upper limit of the data is opposite of the spurious nonlinearity apparent in Figures 1 and 2. However, in order to assess this issue more completely, the following section examines the finite sample properties of a kernel regression estimator that explicitly corrects for boundary bias.

[6] The oversmoothing of the LSCV bandwidth procedure for the diffusion may be due to the fact that boundary data are not weighted in the LSCV minimization problem, in order to decrease bias in the bandwidth choice. However, most of the curvature in $\sigma(x)$ occurs for $x \leq 0.05$. So, the LSCV procedure wants to fit a linear (oversmoothed) function to the data.

[7] Formally, the bias is $O(h^2)$ in the middle of the density, but only $O(h)$ near the boundaries.

III. The Jackknife Kernel Estimator

Rice (1984) proposes the *jackknife kernel estimator* as a solution to this "boundary effect." The essence of this correction is to use Richardson extrapolation to modify the shape of the kernel when the grid point being evaluated is within one bandwidth of either the upper or lower boundary of the approximate support of the density. Let the support of the density be the closed interval $[\underline{b},\bar{b}]$ and assume that the kernel function is defined over the closed interval $[-1,1]$. Recall that the simple N-W kernel estimator with bandwidth parameter h, described earlier, has the general form

$$m_h(x) = \frac{\sum_{t=1}^{T-1} (x_{t+1}^{\Delta} - x_t^{\Delta}) K\left(\frac{x - x_t^{\Delta}}{h}\right)}{\sum_{t=1}^{T-1} K\left(\frac{x - x_t^{\Delta}}{h}\right)}. \tag{16}$$

Let any gridpoint $x \in [\underline{b},\bar{b}]$ be written in the form $x = \rho h$, where ρ is a new parameter whose role is to express the position of x relative to the boundary and the bandwidth. Define the following terms:

$$\omega_K(0,\rho) \equiv \int_{\underline{b}}^{\rho} K(z)\,dz \tag{17}$$

and

$$\omega_K(1,\rho) \equiv \int_{\underline{b}}^{\rho} zK(z)\,dz. \tag{18}$$

Note, by the definition of the kernel function, if $\rho \geq 1$, then the basic properties of K imply that $\omega_K(0,\rho) = 1$ and $\omega_K(1,\rho) = 0$.

For $\rho < 1$, the jackknife kernel estimator is defined as

$$m_h^J(x) = (1 + \phi)m_h(x) - \phi m_{\xi h}(x), \tag{19}$$

which is a weighted average of two kernel estimators with bandwidths h and ξh. ϕ is the weighting parameter which is optimally (from the criterion of bias elimination) set equal to

$$\phi \equiv \frac{\omega_K(1,\rho)/\omega_K(0,\rho)}{\xi\omega_K(1,\rho/\xi)/\omega_K(0,\rho/\xi) - \omega_K(1,\rho)/\omega_K(0,\rho)}. \tag{20}$$

Rice (1984) advocates the use of $\xi = 2 - \rho$.

The jackknife kernel estimator will be implemented using a quartic kernel

$$K_q(z) = \tfrac{15}{16}(1 - z^2)^2, \tag{21}$$

defined on the compact interval $[-1,1]$, which permits an explicit computation of equations (17) and (18). The use of the quartic rather than the Gaussian kernel means that the bandwidth choices used in the previous section need to be modified. We use the method of a "canonical kernel" to map from the Gaussian bandwidths to the corresponding quartic bandwidths.[8] The next subsection considers the finite sample performance of the boundary correction in the jackknife kernel estimator.

We apply the same Monte Carlo analysis to the jackknife estimator as was applied to the original Stanton (1997) estimator. Figures 5 through 8 reproduce Figures 1 through 4 using the jackknife kernel. For ease of comparison, the scales on the axes are identical across the two sets of pictures. Correcting the boundary bias results in a slight improvement in the kernel regression estimators based on the IID bandwidth, but the estimators based on the LSCV and Stanton bandwidths are actually worse! For example, the average kernel regression estimate in the graph in the middle of the left column of Figure 5 (corresponding to Stanton's sample size and persistence level) is much more nonlinear than the corresponding picture in Figure 1. The next section explains this apparently anomalous result.

The effects of the boundary correction on the diffusion estimators are shown in Figures 7 and 8. As in Figures 3 and 4, the estimators based on the smaller bandwidth choice are quite accurate. The boundary correction has an important impact on the diffusion estimator for the larger bandwidth estimators. They are now substantially more accurate, although the Stanton bandwidth choice still exhibits the worst performance of the three choices examined. In summary, the overall effect of the boundary correction in the jackknife estimator is to reduce the spurious nonlinearity in the data, but not to eliminate it. This verifies that the spurious nonlinearity documented using the N-W kernel regression algorithm is not simply a result of the boundary effect.

IV. Why *Is* the Estimated Drift Nonlinear?

The definition of a diffusion process establishes that estimation of the drift function is equivalent (up to $o(\Delta)$) to estimation of the conditional mean function. Thus, understanding the biases in the estimation of the drift amounts to understanding the biases in the estimation of the conditional mean.[9] Asymptotic theory (see, e.g., Robinson (1983, 1986)) establishes the pointwise

[8] See Section 5.4 in Härdle (1990) for a description of this technique.

[9] We thank (without implicating) Matt Pritsker for helping to clarify the following discussion of the biases in the kernel regression estimators.

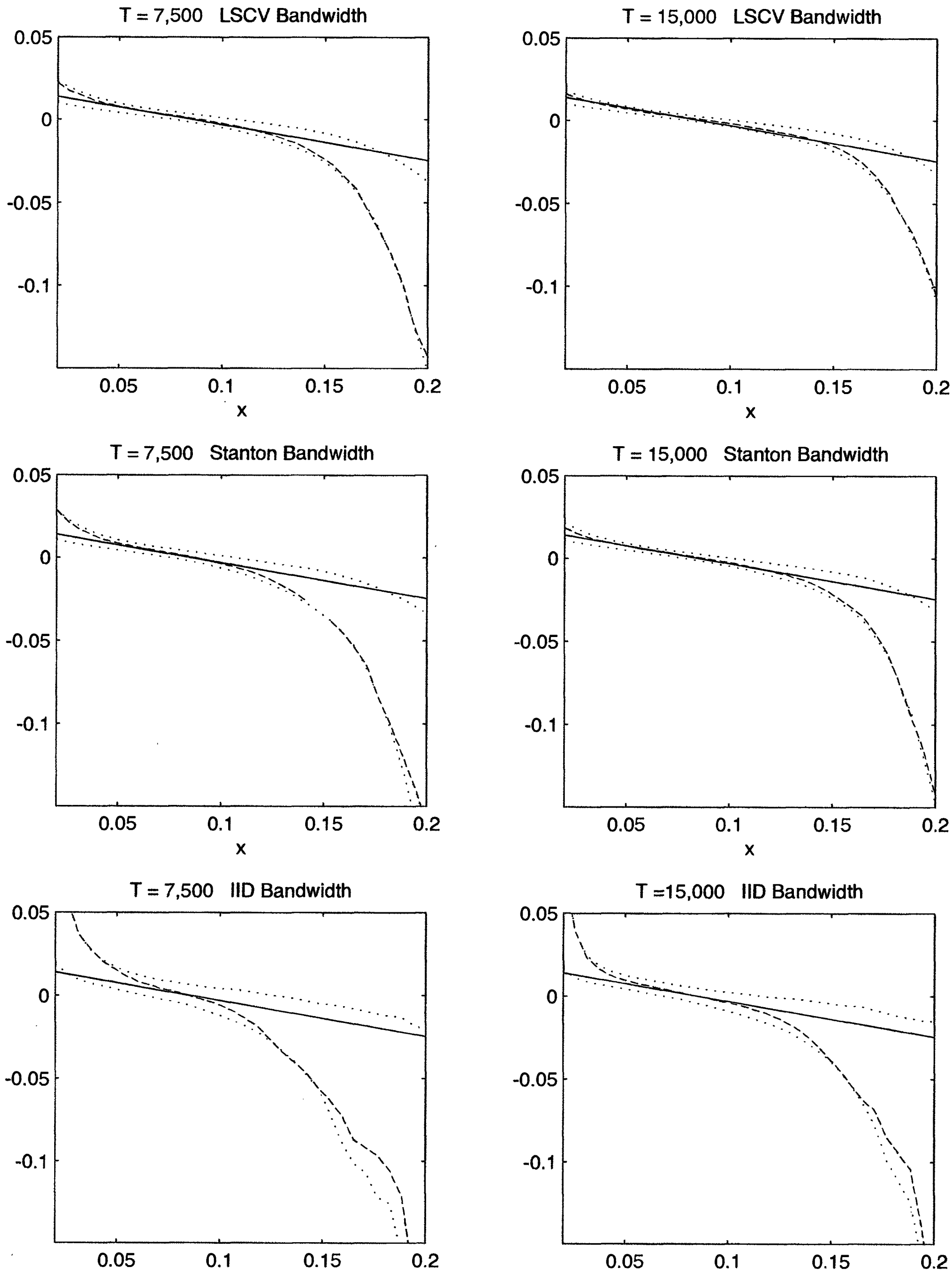

Figure 5. The drift function using the jackknife estimator for true drift defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.21459$ and $\theta = 0.08571$. In these simulations, the diffusion function is $\sigma(x) = \sigma\sqrt{x}$, where $\sigma = 0.07830$. The solid line is the true drift. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles across the 5,000 simulations.

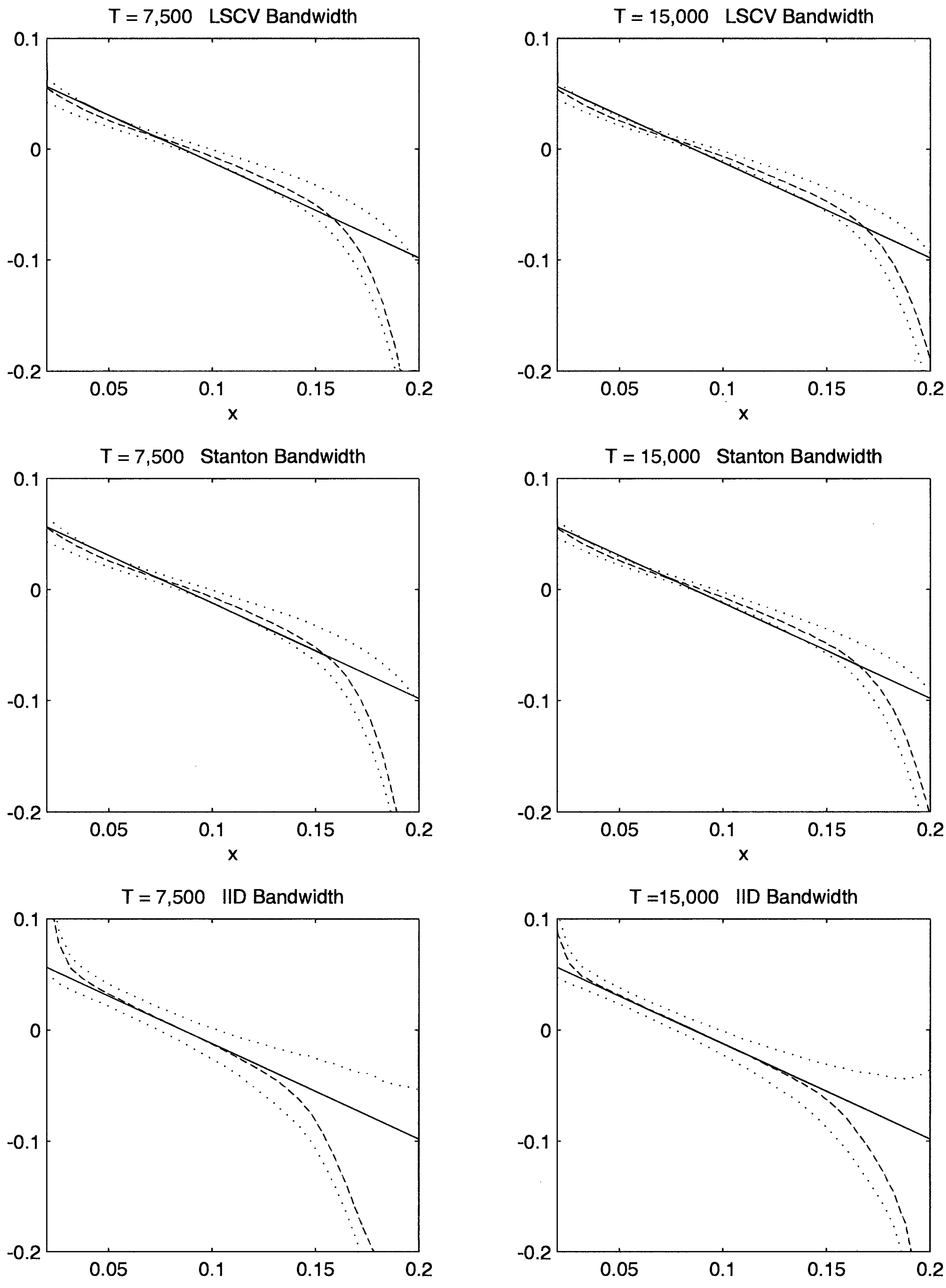

Figure 6. The drift function using the jackknife estimator for true drift defined as $\mu(x) = \kappa(\theta - x)$, where κ = 0.85837 and θ = 0.08571. In these simulations, the diffusion function is $\sigma(x) = \sigma\sqrt{x}$, where σ = 0.15660. The solid line is the true drift. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles across the 5,000 simulations.

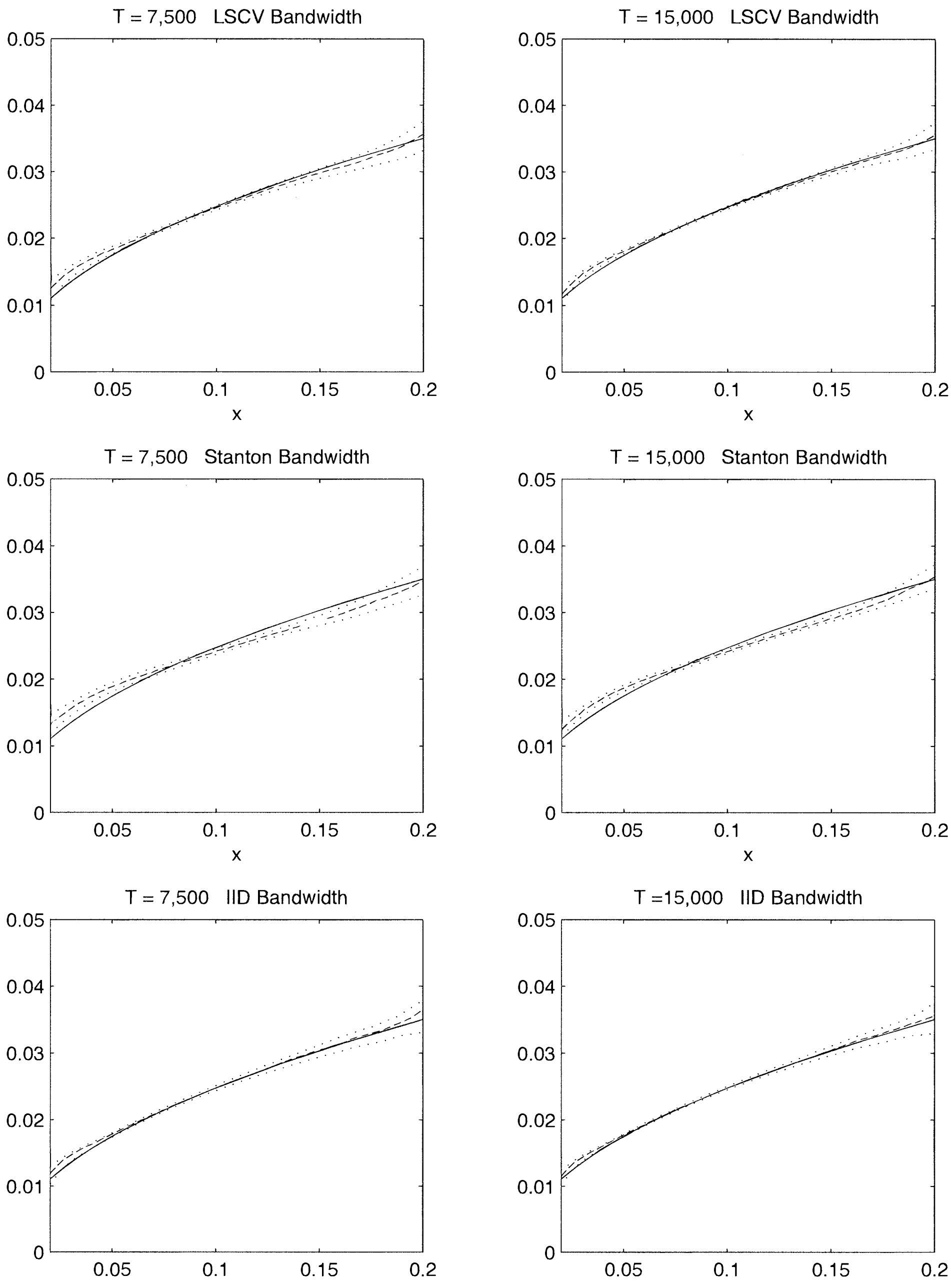

Figure 7. The diffusion function using the jackknife estimator for the true diffusion function defined as $\sigma(x) = \sigma\sqrt{x}$, where σ = 0.07830. In these simulations, the true drift is defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.21459$ and $\theta = 0.08571$. The solid line is the true diffusion. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles across the 5,000 simulations.

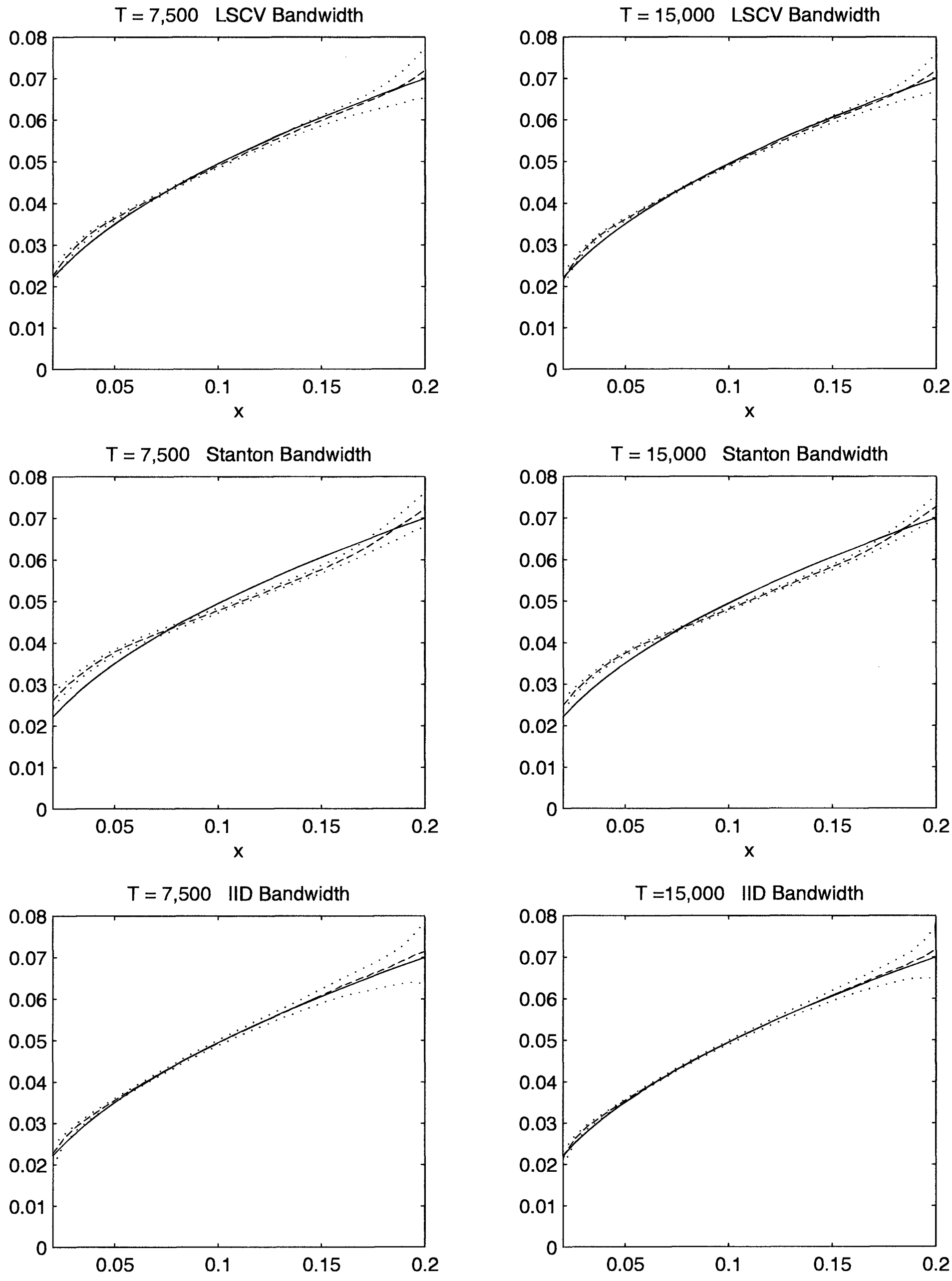

Figure 8. The diffusion function using the jackknife estimator for the true diffusion function defined as $\sigma(x) = \sigma\sqrt{x}$, where $\sigma = 0.15660$. In these simulations, the true drift is defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.85837$ and $\theta = 0.08571$. The solid line is the true diffusion. The dashed line is the pointwise mean across the 5,000 simulations, and the dotted lines are the pointwise 25th and 75th percentiles across the 5,000 simulations.

consistency of kernel regression estimators in a time series context. Therefore, the spurious nonlinearity documented in the previous section must be a finite sample property of the estimators.

There are two distinct sources of these biases. The first is the boundary bias associated with the weighting that occurs through the choice of kernel function and bandwidth parameter. In particular, the drift at any point z_i is estimated as an average of the responses of changes in x at points near z_i. If the weight function is symmetric and the true drift is linear, this bias is small. This occurs for values of z_i in the center of the support of the density of x. However, if z_i is close to the boundary of the support, where "close" is defined relative to the bandwidth value, then the weights are skewed toward the center of the distribution. This implies a downward bias in the drift for values of z_i close to the upper boundary and an upward bias for z_i close to the upper boundary. This type of bias also affects estimates of the diffusion function.

The second source of finite sample bias in the kernel regression estimator is from a truncation of the upper limit of any finite sample. Figure 9, which is based on a single sample path from the parameterization $\kappa = 0.85837$, $\sigma = 0.15660$, and $\theta = 0.08571$ using 7,500 observations, illustrates the source of the problem. The figure shows the interest rate, and the subsequent change in the interest rate, in the right-hand tail of the distribution. Specifically, it shows the ordered pairs $(x_t, x_{t+\Delta} - x_t)$ for which $x_t \geq 0.14$.

The relation between the expected change in the interest rate and its level implied by the true square root process is

$$\begin{aligned} E[x_{t+\Delta} - x_t | x_t] &= \theta(1 - e^{-\kappa\Delta}) + (e^{-\kappa\Delta} - 1)x_t \\ &= 0.00029 - 0.00343x_t, \end{aligned} \tag{22}$$

where the second equality uses the parameter values and $\Delta = 1/250$ (one day) to simulate the process. This line is plotted in Figure 9, along with the estimated OLS regression line using only the tail observations. Notably, the upper right-hand corner of the figure is empty, and the estimated regression has a slope of -0.17, which is clearly more steeply sloped than the true relation between rate levels and subsequent changes.

The relation between $x_{t+\Delta} - x_t$ and x_t is negatively sloped, not by chance but because of the truncation of the realized distribution of interest rates at 0.1583, the largest realization in the sample. The interest rate change that follows the realization of $x_t = 0.1583$ must be nonpositive simply because 0.1583 is the largest interest rate in the sample. More generally, in any sample it must be the case that $x_{t+\Delta} - x_t \leq x^{\max} - x_t$, where $x^{\max}$ is the largest realization. Thus, the upper right corner of Figure 9 is empty by necessity, not by chance, and the negative relation between $x_{t+\Delta} - x_t$ and x_t is not just an artifact of this sample path. As discussed above, the N-W estimator used by Stanton uses only local information in estimating the regression function, and therefore accommodates the negative relation be-

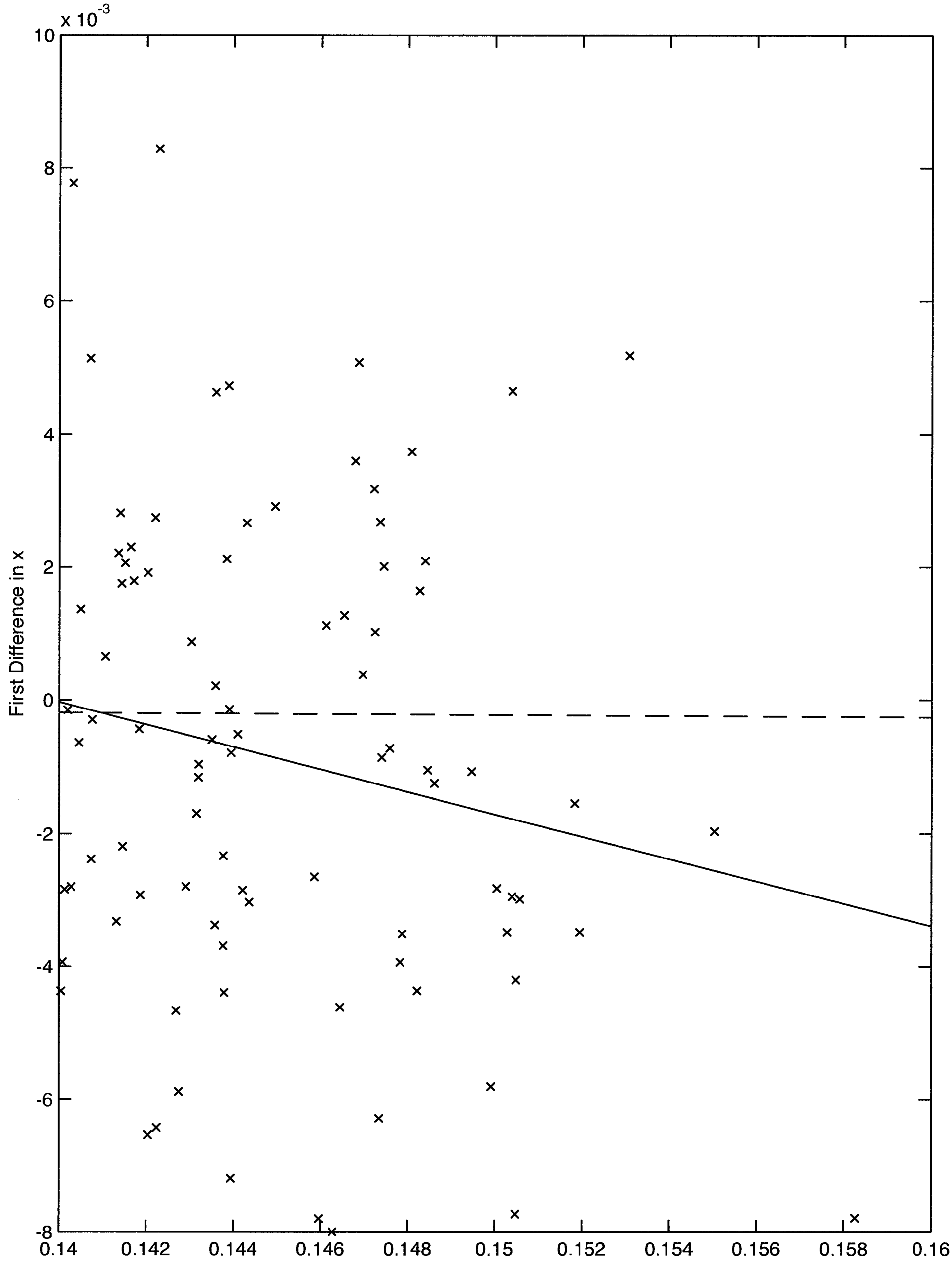

Figure 9. A plot of $[x_{t+\Delta} - x_t]$ versus x_t in the tail of a sample path. The true drift is defined as $\mu(x) = \kappa(\theta - x)$, where $\kappa = 0.85837$, $\theta = 0.08571$. In this simulation, the true diffusion function is $\sigma(x) = \sigma\sqrt{x}$, where $\sigma = 0.15660$. Additionally, $\Delta = 1/250$ and $T_{sim} = 7{,}500$. The dashed line is the true regression function and the solid line is the OLS regression through the data points.

tween $x_{t+\Delta} - x_t$ and x_t in the tail of the distribution. The truncation that is the source of the problem fails to occur only in the limiting case of an infinite sample, because then $x^{\max} = \infty$.

An alternative interpretation is that the truncation bias is due to correlation of the residuals with the regressors for extreme values of the regressors. Specifically, consider the nonparametric regression model

$$x_{t+\Delta t} - x_t = \mu(x_t) + \varepsilon_{t+\Delta t}, \tag{23}$$

along with the standard condition

$$E(\varepsilon_{t+\Delta t}|x_t) = 0. \tag{24}$$

In our case the condition (24) is not satisfied. Rather, $E(\varepsilon_{t+\Delta t}|x_t)$ is much less than zero for x_t near the upper limit of the data, and $E(\varepsilon_{t+\Delta t}|x_t)$ is much greater than zero for x_t near the lower limit of the data. The kernel regression estimator estimates the expectation of $x_{t+\Delta t} - x_t$ conditional on the current realization of x_t and the fact that all of the realizations are in the interval $[x^{\min}, x^{\max}]$. This differs considerably from $E[x_{t+\Delta t} - x_t|x_t]$ when x_t is near either the upper or lower limit of the range of observed interest rates. Given the correlation between the residuals and the regressors, the results in Figures 1 and 2 and 5 and 6 are not surprising.

The interaction of the boundary bias and the truncation or "correlated residuals" bias explains the patterns of bias in Figures 1 through 8. In estimating the drift, large bandwidths result in a cancellation between the boundary bias and the truncation or "correlated residuals" bias. The boundary bias is less severe, and thus there is less cancellation, for smaller bandwidths. This explains why in Figures 1 and 2 the bias is greatest with the IID bandwidth and smallest with the LSCV bandwidth. Comparing the corresponding panels of Figures 1 and 2, one can see that for the Stanton and IID bandwidths the bias is more severe for the parameter choice that implies less persistence (Figure 1). This occurs because increases in persistence reduce the effective sample size, which increases the "correlated residuals" bias. The LSCV bandwidth typically increases with persistence, which enhances the cancellation effect over some ranges of κ. This causes the overall bias to decrease with an increase in persistence.

The striking feature of the drift estimates in Figures 5 and 6 is that the jackknife correction makes the overall bias more severe for the Stanton and LSCV bandwidths. By construction, the jackknife correction reduces the boundary bias, which lessens the bias cancellation discussed above. The result is the striking increase in the overall bias for the large bandwidths.

The story is somewhat different for estimates of the diffusion function. Here, large bandwidths induce a severe boundary bias and residual bias is not present, due to the form of the regression function. This results in a larger bias for larger bandwidths. Bandwidth choices that increase with persistence, like the LSCV bandwidth, worsen the boundary bias and exhibit increases in bias for larger persistence. Finally, the jackknife correction removes the boundary bias. This reduces the bias of all of the larger bandwidth choices, which removes the apparent relation of bias to persistence.

V. The Estimator in Aït-Sahalia (1996)

The general form of the drift and diffusion estimators proposed in Aït-Sahalia (1996) are based on the stationary (or invariant) distribution of x. Given a sample of T (discrete time) observations of x, they are constructed in four steps: (i) estimate the stationary density using a fully nonparametric estimator, (ii) develop an explicit connection between the drift and diffusion functions and the stationary density using the stationary form of the Kolmogorov forward equation, (iii) choose flexible parametric forms for the drift and diffusion functions, and (iv) choose the parameters of the drift and diffusion to make the stationary density implied by the drift and diffusion function in (iii) as "close" as possible (in a specific sense) to the nonparametric estimate in (i). These steps are now described in more detail.

The first step is implemented using a standard Gaussian kernel and equation (4), as described above. The second step in the estimation of the drift and diffusion functions is to relate the stationary density $\pi(x)$ to the drift and diffusion functions.[10] In particular, if $\mu(x;\psi)$ and $\sigma(x;\psi)$ are specific functional forms for the drift and diffusion functions, using parameter vector ψ, then

$$\pi(x;\psi) = \frac{\xi(\psi)}{\sigma^2(x;\psi)} \exp\left\{ \int^x \frac{2\mu(u;\psi)}{\sigma^2(u;\psi)}\, du \right\}, \tag{25}$$

where the lower limit of integration is arbitrary and $\xi(\psi)$ is a constant that ensures that $\pi(x;\psi)$ integrates to one. The heart of the estimation procedure developed in Aït-Sahalia (1996) is the observation that, if $\mu(x;\psi)$ and $\sigma(x;\psi)$ are adequate representations of the drift and diffusion functions in equation (1), then—for some parameter choice ψ^*—the parameterized density $\pi(x;\psi^*)$ should be close to the nonparametric density estimated from the data.

The third step in the estimation of the drift and diffusion functions is to choose flexible functional forms that are capable of taking on a variety of possible shapes. Aït-Sahalia (1996) selects

$$\mu(x;\psi) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1} \tag{26}$$

and

$$\sigma^2(x;\psi) = \beta_0 + \beta_1 x + \beta_2 x^{\beta_3}, \tag{27}$$

[10] See, in particular, Karlin and Taylor (1981, Chap. 15, Sec. 5, pp. 220–221, and Sec. 6, pp. 241–242) for the connection, through the stationary form of the Kolmogorov forward equation, between the drift and diffusion functions and the stationary density.

so that $\psi \equiv (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_2, \beta_3)'$. Of course, a number of restrictions on the elements of ψ are necessary to ensure that μ and σ actually imply that equation (1) has a unique solution and that the implied stationary density exists.[11]

The fourth—and final—step in the estimation process is to state the precise sense in which $\pi(x;\psi)$ and $\pi(x)$ should be "close," and how this idea can be used to select the parameter vector ψ^* from the parameter space Ψ. Aït-Sahalia proposes a minimum mean square distance measure. In particular, he chooses

$$\psi^* \equiv \arg\min_{\psi \in \Psi} E[(\pi(x;\psi) - \pi(x))^2], \tag{28}$$

where the expectation is taken with respect to the true stationary density $\pi(x)$. Estimation is actually performed using

$$\hat{\psi}^* \equiv \arg\min_{\psi \in \Psi} \frac{1}{T} \sum_{t=1}^{T} (\pi(x_t;\psi) - \hat{\pi}(x_t))^2, \tag{29}$$

which is effectively a "nonlinear regression" problem in which the dependent variable is the nonparametric density evaluated at a data point ($\hat{\pi}(x_t)$) and the fitted nonlinear function is given by equation (25).

The general form of the estimator as described above cannot be applied directly to the case of the square root diffusion in equation (12) because, as Aït-Sahalia (1996) notes, equation (27) is not identified under the null hypothesis. Since our focus is on the estimate of the drift function, we make the strong assumption that the true diffusion function is known precisely, including the value of σ. In this case,

$$\int_{\underline{x}}^{x} \frac{2\mu(u;\alpha)}{\sigma^2(u;\beta)}\, du = \frac{2\alpha_0}{\sigma^2}(\ln x - \ln \underline{x}) + \sum_{j=1}^{2} \frac{2\alpha_j}{j\sigma^2}(x^j - \underline{x}^j) - \frac{2\alpha_3}{\sigma^2}\left(\frac{1}{x} - \frac{1}{\underline{x}}\right), \tag{30}$$

where again the lower bound of integration, $\underline{x}$, is arbitrary.

Under these assumptions, given an estimate of the stationary density, Aït-Sahalia's estimator reduces to a simple ordinary least squares problem in a transformation of $\hat{\pi}(x_t)$.[12] The stationary density is estimated using a Gaussian kernel with two bandwidth choices. The first bandwidth is the LSCV bandwidth for the density estimator (as described in Chapter 3 of Silverman (1986)). The second bandwidth is the IID bandwidth, which is oversmoothed relative to the LSCV choice. Using the algorithm described in Section II.A,

[11] See Aït-Sahalia (1996, eq. (24)).

[12] Specifically, the dependent variable is $y_t = \frac{1}{2}\sigma^2 \log(\hat{\pi}_t \sigma^2 x_t)$, which is regressed on a constant (to capture the arbitrary $\underline{x}$), $\log x$, x, x^2, and x^{-1}. The coefficients on the last four terms are α_0 through α_3, respectively.

Table I

Aït-Sahalia (1996) Parameter Estimates: LSCV Bandwidth

The estimated drift function is $\hat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$, and the true drift function is $\mu(x) = \kappa(\theta - x)$ with $\kappa = 0.85837$, $\sigma = 0.15660$, and LSCV bandwidth. All statistics are computed using ordinary least squares applied to the 5,000 simulated sample paths of a square root diffusion.

			Percentiles			
	Mean	Std. Dev.	25th	50th	75th	True Value
			Panel A: T_{sim} = 7,500			
α_0	0.0086	0.2370	−0.0981	0.0324	0.1421	**0.0736**
α_1	0.0351	3.1805	−1.8836	−0.2537	1.5957	**−0.8584**
α_2	−2.0191	6.7816	−5.3202	−1.1770	2.1855	**0.0000**
α_3	−0.0016	0.0057	−0.0037	−0.0007	0.0013	**0.0000**
			Panel B: T_{sim} = 15,000			
α_0	0.0339	0.2053	−0.0606	0.0495	0.1463	**0.0736**
α_1	−0.2361	2.8113	−1.9057	−0.4619	1.1884	**−0.8584**
α_2	−1.5385	6.0796	−4.5215	−0.8528	2.1912	**0.0000**
α_3	−0.0009	0.0048	−0.0027	−0.0002	0.0015	**0.0000**

5,000 sample paths of length 7,500 and length 15,000 for $\kappa = 0.85837$, $\theta = 0.08571$, and $\sigma = 0.15660$ are simulated and the Aït-Sahalia estimator is applied to each simulated sample path.

Tables I and II contain the results of this experiment. The accuracy of the estimators, as measured by the mean of the distribution across the 5,000 simulations, is increasing in the length of the simulated series. It is also true that the standard deviation of the estimates is decreasing in the simulation length. Finally, the standard deviations of the estimators are uniformly lower for the (larger) IID bandwidths, when compared with the results for the corresponding LSCV bandwidths.

However, the most striking feature of these tables is the consistently poor performance of the estimator. The parameter on the x^{-1} term in the regression, α_3, is measured accurately, but it is the notable exception. In both panels of both tables, the average estimates of α_0 and α_1 are too small (in absolute value). For the LSCV bandwidth and T_{sim} = 7,500, the average point estimate of α_1 has the wrong sign. The average estimates of α_2 are large and negative for the LSCV bandwidth and smaller but still negative for the IID bandwidth. If the average parameter estimate is compared with the standard deviation of the parameter estimates across the 5,000 simulations, none of the estimates appear to be significantly different from zero. This fact highlights the imprecision inherent in this estimation procedure. Based on these simulations, it is very difficult to have any confidence in the point estimates constructed in this manner.

Table II

Aït-Sahalia (1996) Parameter Estimates: IID Bandwidth

The estimated drift function is $\hat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$, while the true drift function is $\mu(x) = \kappa(\theta - x)$ with $\kappa = 0.85837$, $\sigma = 0.15660$, and IID bandwidth. All statistics are computed using ordinary least squares applied to the 5,000 simulated sample paths of a square root diffusion.

			Percentiles			
	Mean	Std. Dev.	25th	50th	75th	True Value
			Panel A: $T_{sim} = 7{,}500$			
α_0	0.0506	0.1278	−0.0162	0.0624	0.1280	**0.0736**
α_1	−0.5343	1.7764	−1.6716	−0.6819	0.4557	**−0.8584**
α_2	−0.7322	3.7838	−2.7990	−0.2579	1.7942	**0.0000**
α_3	−0.0005	0.0028	−0.0017	−0.0001	0.0010	**0.0000**
			Panel B: $T_{sim} = 15{,}000$			
α_0	0.0633	0.1141	0.0040	0.0705	0.1309	**0.0736**
α_1	−0.6607	1.6712	−1.6905	−0.7674	0.2631	**−0.8584**
α_2	−0.5208	3.4877	−2.4537	−0.1254	1.8147	**0.0000**
α_3	−0.0002	0.0024	−0.0012	0.0001	0.0012	**0.0000**

The direction of the bias in α_2 is consistent, heuristically, with the "correlated residuals" bias documented in connection with the kernel regression estimator, but the explanation is different. Here the bias appears because the estimated marginal density is truncated at the boundaries, compared with the true density. As a result, for a given (nonvanishing) diffusion, the estimated drift must stop the interest rate from "escaping" too far from the observed range of the sample. With the wider bandwidth, the truncation of the marginal density and the bias are less severe.[13]

Additionally, the estimates are very imprecise due to multicollinearity induced by the choice of functional form in equation (26).[14] Recall that the classic symptoms of multicollinearity include coefficient estimates of implausible sign or magnitude and large standard errors on estimated parameters combined with jointly significant explanatory variables. Although only a simplified version of the Aït-Sahalia estimator is examined here, it is implausible that a more complicated nonlinear version of the model does not suffer from these same problems.

[13] We also obtain results (not reported) using the Pritsker (1998) bandwidth, which is considerably wider than the IID and LSCV bandwidths. Using this bandwidth, the mean of α_2 changes sign; that is, the bias is reversed. This occurs because the Pritsker bandwidth results in an estimate of the marginal density with much fatter tails.

[14] This same multicollinearity problem is apparent in the simple weighted least squares estimator examined in the next section.

VI. Is the Drift Actually Nonlinear?

The primary conclusions of the Monte Carlo analyses presented above are as follows. First, there is a truncation or correlated residuals bias in any finite sample which complicates inference about the conditional mean of the short rate away from the center of the support of the density. Kernel regression estimators are particularly sensitive to this bias because they use local data more intensively than conventional parametric estimators. In fact, if one "corrects" the results in Stanton (1997) for the bias we document, the drift appears to be nearly linear. Second, the performance of the kernel regression estimators is sensitive to the choice of the bandwidth parameter, which is difficult to specify a priori in the absence of a specific null hypothesis for the regression functions. This point is illustrated in Figure 10, which shows three estimates of the drift function using the Stanton data and the three bandwidth rules examined in the Monte Carlo simulation. Third, both the Aït-Sahalia (1996) and Stanton (1997) estimators are not very efficient in the tails of the density, which makes precise inference about linearity or nonlinearity of the true drift function difficult. Taken as a whole, the Monte Carlo results indicate that Aït-Sahalia (1996) and Stanton (1997) do not provide convincing evidence of nonlinearity.

Since these problems appear to be inherent in nonparametric estimators, in order to answer the question raised in the title of the paper, consider a weighted least squares estimator applied to a discretized version of a generalized short rate process:

$$x_{t+\Delta} - x_t = [\alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + \alpha_3 x_t^{-1}]\Delta + \varepsilon_{t+\Delta}, \tag{31}$$

where $\Delta x_{t+\Delta} \equiv x_{t+\Delta} - x_t$, Δ is the (constant) interval between observations and

$$E(\varepsilon_{t+\Delta}) = 0 \quad \text{and} \quad E(\varepsilon_{t+\Delta}^2) = (\sigma^2\Delta)x_t^{2\gamma}. \tag{32}$$

The estimator is constructed in three steps. First, the drift parameters are estimated using ordinary least squares. Given estimates of the drift parameters, equation (32) implies that the diffusion parameters can be estimated using a log linear regression of the (scaled) squared residuals from equation (31).[15] Finally, the diffusion parameters are used (in conjunction with x_t) to construct the weights applied to both the dependent and independent variables in equation (31) to estimate the weighted least squares drift parameter estimates. The standard errors of the coefficients estimated from these

[15] This estimation procedure is very close to the exactly identified case of the generalized method of moments (GMM) estimator used in Chan et al. (1992). They differ only in the moment conditions associated with the diffusion parameters. It is well known (e.g., see Hamilton (1994)) that the choice of the weighting matrix is irrelevant in the exactly identified case, since the GMM objective function is minimized by choosing the parameters that make each moment condition as close to zero as possible.

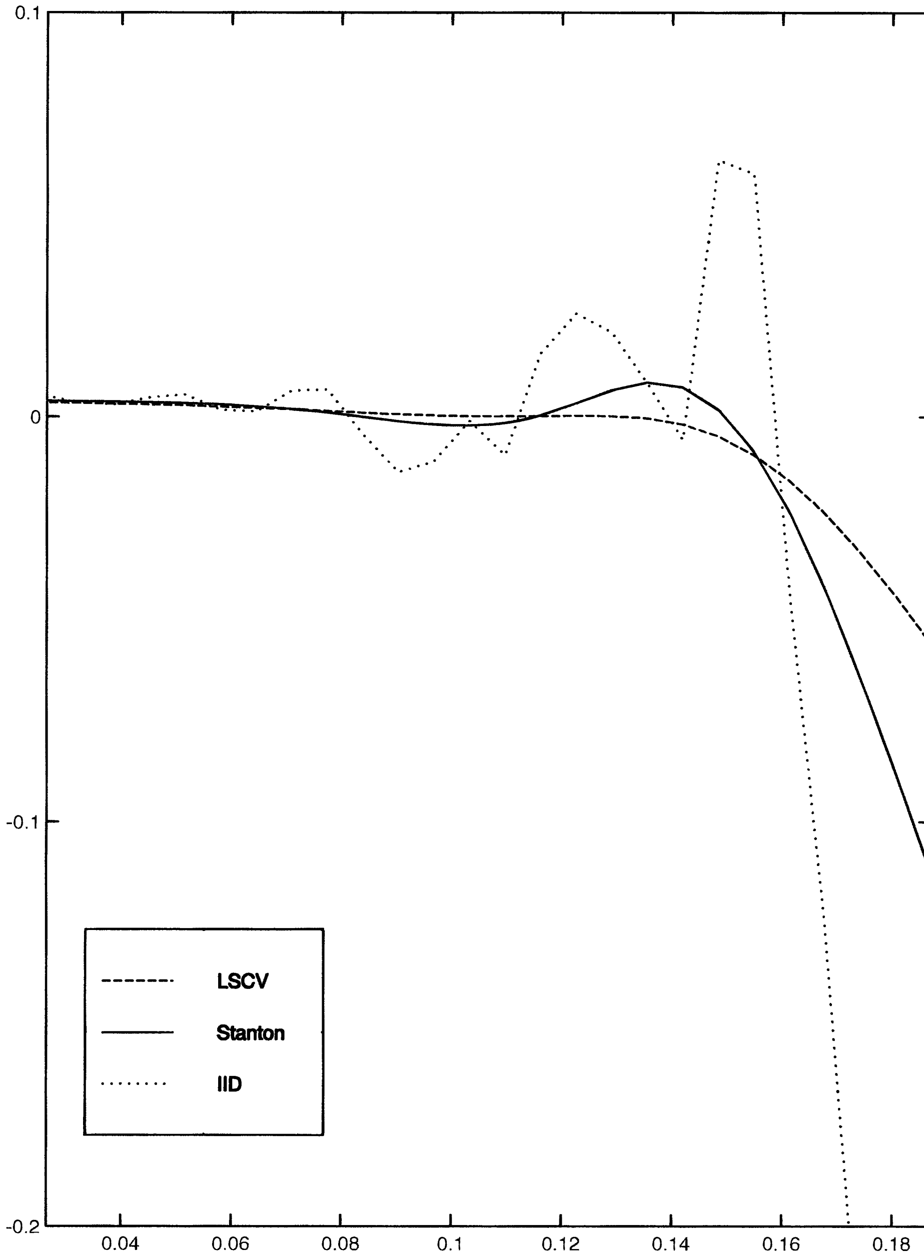

Figure 10. A plot of the estimated drift function of the three-month Treasury bill using three different bandwidth parameter choices. "LSCV" is the true drift function $\mu(x)$ estimated using the least squares cross validation bandwidth. "Stanton" is the same drift function estimated using the Stanton bandwidth, and "IID" is the drift function estimated using the bandwidth for independent and identically distributed data. The data are from Stanton (1997), and they consist of daily observations of the yield on a three-month Treasury bill from January 1965 to July 1995, for a total of 7,975 observations.

regressions can be calculated in a standard manner for weighted least squares under the assumption that the null hypothesis is correctly specified, or they can be calculated from the residuals using a prewhitened autocorrelation and heteroskedasticity consistent covariance matrix estimator and an automatic lag truncation selection procedure as described in Andrews and Monohan (1992) or Newey and West (1994).[16]

It is well known that a simple moment-based estimator from equations (31) and (32) induces a discretization bias through the use of a first-order approximation to the true discretely sampled moments. The magnitude of this bias should not be large when the estimator is based on daily data. However, a simulation of its finite sample properties permits a direct assessment of the estimator's performance. Panel A of Table III reports the summary statistics from the empirical distribution of the least squares parameter estimates based on 5,000 simulated sample paths of length $T_{sim} = 7{,}500$ with $\kappa = 0.21459$, $\theta = 0.08571$, and $\sigma = 0.07830$, where the data are simulated in exactly the same manner as in the analysis of the nonparametric estimators.

The first row in Table III, Panel A reports the true value of the parameters. The estimate of γ, which is one-half the elasticity of variance of the diffusion process, is quite accurate and most of the mass of the empirical density is within 0.05 of the true value. α_1 is an estimate of $-\kappa$, and the sampling distribution of this estimator is biased downward and skewed to the left. This is a manifestation of the well-known problems associated with the estimation of the autocorrelation parameter of a process that almost has a "unit root" (in the autoregressive polynomial representation of the series). This downward bias in the estimate of κ results in an upward bias in the estimate of α_0 because $-\alpha_0/\alpha_1$ is an estimate of the unconditional mean of x. In fact, the mean of the sample estimates of α_0 and α_1 imply an unconditional mean of 0.08407, which is close to the true value of $\theta = 0.08571$. The discretization bias becomes apparent in the estimate of σ. All else equal, the downward bias in the estimate of κ implies, from the analogy to the discrete-time first-order autoregressive process, a *larger* estimate of σ. However, Panel A shows that the least squares estimate of σ is approximately half of the true parameter value.

Table III, Panel B, shows the summary statistics of the empirical distribution of the parameter estimates of equations (31) and (32) across the 5,000 simulated sample paths. This is a misspecified model since it includes the quadratic and inverse terms which are not a part of the true data generating process. Assuming that the additional variables are not correlated with the true regressors (or the error term), standard regression theory predicts that the sampling distributions of both α_2 and α_3 should be centered at zero and estimates of α_0 and α_1 should be unaffected by the inclusion of irrelevant regressors. Panel B shows that there are dramatic effects on the point estimates of the drift pa-

[16] The lag truncation parameter is also referred to in the literature as the bandwidth parameter. It defines how many autocovariances are used in constructing the heteroskedasticity and autocorrelation consistent covariance matrix estimator. It is analogous to the bandwidth parameter in the kernel regression literature in the sense that it is both difficult to specify a priori and its choice has a significant impact on the performance of the covariance estimator.

Table III

Distribution of the Weighted Least Squares Parameters

The results reported in the table are based on 5,000 simulated sample paths of length T_{sim} = 7,500 observations. The true drift function is $\mu(x) = \kappa(\theta - x)$ and the true diffusion function is $\sigma(x) = \sigma\sqrt{x}$. The linear model estimates the drift function $\Delta x_{t+\Delta} = \alpha_0 + \alpha_1 x_t + \varepsilon_{t+1}$, and the nonlinear model estimates the drift function $\Delta x_{t+\Delta} = \alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + \alpha_3 x_t^{-1} + \varepsilon_{t+\Delta}$. In both specifications, the error term satisfies $E_t(\varepsilon_{t+\Delta}) = 0$ and $E_t(\varepsilon_{t+\Delta}^2) = \sigma^2 x_t^{2\gamma}$.

Panel A: Linear Drift Model

	α_0	α_1			σ	γ
True value	**0.0184**	**−0.2146**			**0.0783**	**0.5000**
Mean	0.0305	−0.3628			0.0416	0.4993
Std. Dev.	0.0155	0.1838			0.0042	0.0395
5th Per.	0.0119	−0.7181			0.0352	0.4354
10th Per.	0.0143	−0.5984			0.0366	0.4505
25th Per.	0.0194	−0.4602			0.0388	0.4745
50th Per.	0.0272	−0.3289			0.0414	0.4991
75th Per.	0.0382	−0.2287			0.0441	0.5238
90th Per.	0.0508	−0.1653			0.0467	0.5478
95th Per.	0.0602	−0.1354			0.0486	0.5636

Panel B: Nonlinear Drift Model

	α_0	α_1	α_2	α_3	σ	γ
True value	**0.0184**	**−0.2146**	**0.0000**	**0.0000**	**0.0783**	**0.5000**
Mean	−0.2853	3.6558	−16.5896	0.0081	0.0416	0.4995
Std. Dev.	0.3699	4.2823	18.3498	0.0117	0.0042	0.0394
5th Per.	−0.9698	−0.6465	−50.9428	−0.0002	0.0351	0.4350
10th Per.	−0.7021	0.0250	−38.5272	0.0004	0.0366	0.4510
25th Per.	−0.3863	1.0658	−22.7839	0.0016	0.0389	0.4750
50th Per.	−0.1816	2.5491	−11.7452	0.0045	0.0414	0.4993
75th Per.	−0.0680	5.0773	−5.4086	0.0102	0.0440	0.5233
90th Per.	−0.0007	8.8580	−0.9149	0.0202	0.0468	0.5472
95th Per.	0.0451	11.7002	1.5533	0.0291	0.0487	0.5632

rameters and virtually no impact on the parameters estimated from the residuals of the drift function regression. In particular, the drift parameters have implausible magnitudes and implausible signs and very large standard deviations of the empirical distribution. These are all classic symptoms of multicollinearity induced (in this case) by the choice of the nonlinear functional form.

Table IV reports the simulation results from a regression that addresses the multicollinearity problem by using orthogonalized and scaled monomials in x to capture any possible nonlinearity in the drift. These new regressors span the same space as the original regressors, but allow for easier interpretation; that is, they are the incremental effect of adding the nonlinear terms. The regression is

$$\Delta x_{t+\Delta} = [\alpha_0 + \alpha_1 x_t + \alpha_2 p_2(x_t) + \alpha_3 p_3(x_t)]\Delta + \varepsilon_{t+\Delta}. \quad (33)$$

Table IV

Weighted Least Squares Estimates Using Orthogonalized Polynomials

The results reported in the table are based on 5,000 simulated sample paths of length T_{sim} = 7,500 observations. The true drift function is $\mu(x) = \kappa(\theta - x)$ and the true diffusion function is $\sigma(x) = \sigma\sqrt{x}$. The nonlinear model estimates the drift function $\Delta x_{t+\Delta} = \alpha_0 + \alpha_1 x_t + \alpha_2 p_2(x_t) + \alpha_3 p_3(x_t) + \varepsilon_{t+\Delta}$, where $p_2(x_t)$ and $p_3(x_t)$ are orthogonalized and normalized second- and third-order monomials in x. The precise construction of these regressors is described in the text. The error term satisfies $E_t(\varepsilon_{t+\Delta}) = 0$ and $E_t(\varepsilon_{t+\Delta}^2) = \sigma^2 x_t^{2\gamma}$.

	α_0	α_1	α_2	α_3	σ	γ
True value	**0.0184**	**-0.2146**	**0.0000**	**0.0000**	**0.0783**	**0.5000**
Mean	0.0314	−0.3726	−0.0879	−0.1701	0.0416	0.4994
Std. Dev.	0.0159	0.1882	0.1333	0.1256	0.0042	0.0395
5th Per.	0.0116	−0.7204	−0.3151	−0.3881	0.0351	0.4354
10th Per.	0.0145	−0.6147	−0.2578	−0.3337	0.0366	0.4505
25th Per.	0.0200	−0.4729	−0.1677	−0.2432	0.0389	0.4749
50th Per.	0.0283	−0.3406	−0.0825	−0.1633	0.0414	0.4993
75th Per.	0.0394	−0.2350	−0.0066	−0.0908	0.0440	0.5240
90th Per.	0.0518	−0.1703	0.0681	−0.0206	0.0468	0.5475
95th Per.	0.0611	−0.1319	0.1204	0.0243	0.0487	0.5634

$p_2(x_t)$ is formed as the regression residual of x_t^2 onto a constant and x_t, which is then scaled to have a standard deviation equal to the standard deviation of r_t. In a similar manner, $p_3(x_t)$ is defined by regressing x_t^{-1} onto a constant, x_t, and $p_2(x_t)$. The parameter estimates in Table IV are consistent with Panel A of Table III and with the true underlying process. The distributions for α_0, α_1, σ, and γ in Table IV are virtually identical to Panel A of Table III.[17] The estimates of α_1 and α_2 are not significantly different from zero, using the true cross-sectional standard deviation. The negative sign of the mean of the empirical distribution is consistent with the residual correlation bias discussed in conjunction with the nonparametric estimators, but the mean is closer to zero.

Having verified that the bias in the regression estimator is not large, we turn to the actual data. Two different actual data series are used as proxies for the unobservable short rate in the least squares estimation procedure. The first is daily observations on the three-month Treasury bill from January 1965 to July 1995, the series used in Stanton (1997), and the second is daily observations on the seven-day Eurodollar rate, the series used in Aït-Sahalia (1996).[18] The results of the weighted least squares estimation of equations (33) and (32) are presented in Table V.

There is little evidence against a *nonzero* drift in the three-month Treasury bill yield; that is, α_0 through α_3 are (individually) not significantly different from zero using the robust standard errors calculated from the

[17] The estimates for α_0 and α_1 are not exactly identical because the weighting procedure makes the regressors only approximately orthogonal.

[18] We thank both Yacine Aït-Sahalia and Richard Stanton for providing us with their data sets.

Table V

Weighted Least Squares Estimates of a Nonlinear Model

The parameters are estimated for the nonlinear drift function

$$\frac{\Delta x_{t+\Delta}}{\Delta} = \alpha_0 + \alpha_1 x_t + \alpha_2 p_2(x_t) + \alpha_3 p_3(x_t) + \varepsilon_{t+\Delta},$$

where $\Delta x_{t+\Delta} = x_{t+\Delta} - x_t$, Δ is the time step (in units of a fraction of a year), and $p_2(x_t)$ and $p_3(x_t)$ are orthogonalized and normalized second- and third-order monomials in x. Given fitted values for the drift, the diffusion parameters are estimated from a log linear regression using the (scaled) squared residuals from the initial regression

$$\ln(\tilde{\varepsilon}^2_{t+\Delta}) = \ln(\sigma^2) + 2\gamma \ln(x_t) + \xi_{t+\Delta},$$

where $\tilde{\varepsilon}_t = \sqrt{\Delta}\hat{\varepsilon}_t$ and $\hat{\varepsilon}_t$ is the residual from the nonlinear drift function regression. Given the fitted values for the residuals, the drift parameters are reestimated by weighted least squares. Ordinary least squares standard errors are reported in parentheses. Heteroskedasticity and autocorrelation consistent standard errors, based on the prewhitened Bartlett kernel with automatic lag truncation selection as in Andrews and Monohan (1992) and Newey and West (1994), are reported in square brackets.

Parameters	Stanton (1997) T-Bill Data	Aït-Sahalia (1996) Eurodollar Data
α_0	0.0180	0.1307
	(0.0008)	(0.0010)
	[0.0164]	[0.0371]
α_1	−0.2526	−1.5735
	(0.0132)	(0.0139)
	[0.2873]	[0.5423]
α_2	−0.1989	−1.4778
	(0.0129)	(0.0143)
	[0.2021]	[0.8034]
α_3	−0.0944	−0.5416
	(0.0076)	(0.0094)
	[0.2794]	[1.3068]
$\ln(\sigma^2)$	−3.2599	−3.6478
	(0.2437)	(0.3114)
	[0.6010]	[1.1137]
2γ	2.7847	2.4569
	(0.0878)	(0.1194)
	[0.2167]	[0.4392]
Sample size	Jan 1965–July 1995	June 1973–Feb 1995
	$T = 7975$	$T = 5505$

consistent covariance matrix estimator. Although the parameter estimates are significantly different from zero using the conventional WLS standard errors, these standard errors are correct only if the null hypothesis is correctly specified. It is important to note that these results do not imply that the drift function is linear. They do demonstrate a lack of evidence of nonlinearity in the short rate drift, but this may be an artifact of the poor finite sample performance of the robust covariance matrix estimator from Andrews and Monohan (1992) and Newey and West (1994) (also, the specific functional form of equation (33) rules out certain forms of nonlinearity). The point estimates of the drift terms in the Eurodollar rate regression are larger in magnitude than the Treasury bill data, but they are also not significantly different from zero using robust standard errors.

VII. Conclusions

There is no definitive answer to the question posed in the title of this paper. The Monte Carlo evidence developed and presented above demonstrates that there are quantitatively significant biases in kernel regression estimators of the drift advocated in Stanton (1997). These biases produce precisely the kind of nonlinearity at high levels of the process reported in Stanton. Furthermore, the finite sample performance of these estimators is dependent on the choice of the bandwidth parameter. Oversmoothed estimates always tend to suggest linearity in the drift, and undersmoothed estimates are particularly susceptible to the biases documented above.

The Aït-Sahalia (1996) estimator is also incapable of producing convincing evidence of a nonlinear drift. Although the sample truncation still introduces a bias, the primary issue here is the multicollinearity induced by the flexible functional form for the drift function. The parameter estimates on the quadratic term are large and measured imprecisely for all sample size and bandwidth combinations.

The conclusion that we draw from the Monte Carlo evidence is that the nonparametric and seminonparametric estimators examined in the paper simply cannot produce reliable evidence of nonlinearity in the drift when applied to short-term interest rate data. Therefore, in order to assess this question, least squares estimation based on a simple first-order discretization of the data, as in Chan et al. (1992), is examined. These estimators do not produce statistically reliable evidence of nonlinearity. Overall, the evidence suggests that time series methods alone are not capable of producing evidence of nonlinearity in the drift.

REFERENCES

Aït-Sahalia, Yacine, 1996, Testing continuous-time models of the spot interest rate, *Review of Financial Studies* 9, 385–426.

Andrews, Donald W. K., and J. Christopher Monohan, 1992, An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator, *Econometrica* 60, 953–966.

Black, Fischer, Emmanuel Derman, and W. Toy, 1990, A one-factor model of interest rates and its application to Treasury bond options, *Financial Analysts Journal* 46, 33–39.

Black, Fischer, and Piotr Karasinski, 1991, Bond and option pricing when short rates are lognormal, *Financial Analysts Journal* 47, 52–59.

Brennan, Michael J., and Eduardo S. Schwartz, 1979, A continuous-time approach to the pricing of bonds, *Journal of Banking and Finance* 3, 133–155.

Chan, K. C., G. Andrew Karolyi, Francis A. Longstaff, and Anthony B. Sanders, 1992, An empirical comparison of alternative models of the short-term interest rate, *Journal of Finance* 47, 1209–1227.

Courtadon, Georges, 1982, The pricing of options on default-free bonds, *Journal of Financial and Quantitative Analysis* 17, 75–100.

Cox, John C., Jonathan E. Ingersoll, Jr., and Stephen A. Ross, 1985, A theory of the term structure of interest rates, *Econometrica* 53, 321–346.

Hamilton, James D., 1994, *Time Series Analysis* (Princeton University Press, Princeton, N.J.).

Härdle, Wolfgang, 1990, *Applied Nonparametric Regression* (Cambridge University Press, Cambridge, U.K.).

Jones, Christopher S., 1998, Nonlinear mean reversion in the short-term interest rate, Unpublished paper, The Wharton School, University of Pennsylvania.

Karlin, Samuel, and Howard M. Taylor, 1981, *A Second Course in Stochastic Processes* (Academic Press, San Diego.).

Merton, Robert C., 1980, On estimating the expected return on the market: An exploratory investigation, *Journal of Financial Economics* 8, 323–361.

Newey, Whitney K., and Kenneth D. West, 1994, Automatic lag selection in covariance matrix estimation, *Review of Economic Studies* 61, 631–653.

Pearson, Neil D., and Tong-Shen Sun, 1994, Exploiting the conditional density in estimating the term structure: An application to the Cox, Ingersoll, and Ross model, *Journal of Finance* 49, 1279–1304.

Pritsker, Matthew, 1998, Nonparametric density estimation of tests of continuous time interest rate models, *Review of Financial Studies* 11, 449–487.

Rice, J. A., 1984, Boundary modification for kernel regression, *Communications in Statistics, Theory and Methods* 13, 893–900.

Robinson, Peter M., 1983, Nonparametric estimators for time series, *Journal of Time Series Analysis* 4, 185–207.

Robinson, Peter M., 1986, On the consistency and finite-sample properties of nonparametric kernel time series regression, autoregression and density estimators, *Annals of the Institute of Statistics and Mathematics* 38, 539–549.

Silverman, B. W., 1986, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London).

Stanton, Richard, 1997, A nonparametric model of term structure dynamics and the market price of interest rate risk, *Journal of Finance* 52, 1973–2002.