Is the Short Rate Drift Actually Nonlinear?*

David A. Chapman† Neil D. Pearson‡

June 16, 1998

Abstract

Virtually all existing continuous-time, single-factor term structure models are based on a short rate process that has a linear drift function. However, there is no strong a priori argument in favor of linearity, and Stanton (1997) and Aït-Sahalia (1996) employ nonparametric estimation techniques to conclude that the drift function of the short rate contains important nonlinearities. Comparatively little is known about the finite-sample properties of these estimators, particularly when they are applied to frequent sampling of a very persistent process, like short term interest rates. In this paper, we apply these estimators 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 by Stanton (1997) and Aït-Sahalia (1996). These results, along with the results of a simple GMM estimation procedure applied to the Stanton and Aït-Sahalia data sets, imply that nonlinearity of the short rate drift is not a robust stylized fact.

*A previous version of this paper circulated under the title "Nonparametric Estimation of Continuous-Time Markov Processes: A Monte Carlo Analysis." We would like to thank Yacine Aït-Sahalia, Anil Bera, Murray Carlson, John Cochrane, Eric Hughson, Narasihman Jegadeesh, George Pennachi, Matt Pritsker, an anonymous referee, and seminar participants at the University of Illinois at Urbana-Champaign, The University of Texas at Austin, and the University of Utah for helpful comments. We would also like to thank Yacine Aït-Sahalia and Richard Stanton for generously supplying us with their data sets.

†Finance Department, Graduate School of Business, The University of Texas at Austin, Austin, Texas 78712-1179. Phone: (512) 471-6621. e-mail: chapman@eco.utexas.edu. A copy of this paper is available online at http://www.bus.utexas.edu/~chapmand/cp.html

‡Finance Department, University of Illinois at Urbana-Champaign, 340 Commerce West, 1206 South Sixth Street, Champaign, Illinois 61820. Phone: (217) 244-0490. e-mail: pearson2@uiuc.edu.

1 Introduction

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, which 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), in the time-homogenous continuous-time versions of the Black, Derman, and Toy (1990) (hereafter, BDT) and Black and Karisinski (1991) (hereafter, BK) models, and in the empirical models considered by Chan, Karolyi, Longstaff, and Sanders (1992) (hereafter, CKLS). 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 are 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, Stanton (1997) and Aït-Sahalia (1996) have recently proposed nonparametric estimators of the drift and diffusion functions. A key finding in these papers is that the estimated drift function is highly non-linear, especially for large values of the interest rate process. Stanton (1997) finds that the estimated drift drops sharply as the interest rate increases beyond about 14 percent, while Aït-Sahalia (1996) 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" [Aït-Sahalia (1996), page 387].

These results are inconsistent with *all* of the models cited above. In Vasicek, CIR, Pearson and Sun, Brennan and Schwartz, Courtadon, and in the empirical models considered by CKLS, the drift is linear, while in continuous-time versions of the BDT and BK 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 which is consistent with the drift and

diffusion functions estimated by Stanton (1997) and Aït-Sahalia (1996), 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 Stanton (1997) and Aït-Sahalia (1996) by repeatedly simulating the sample paths of the CIR square-root process. We consider three different parameterizations, all of which have the same stationary density but different levels of persistence. The most persistent parameterization is consistent with the time series properties of short-term US Treasury yields. The second parameterization has persistence equal to that of the one week Eurodollar yield used in Aït-Sahalia (1996), and the final parameterization provides a lower bound case with an implied monthly first-order autocorrelation coefficient of 0.867. In addition to variation in persistence of the process, we also consider three different simulation lengths corresponding to $7,500$, $15,000$, and $30,000$ daily observations.

We apply the nonparametric estimators to each of the simulated sample paths, and thereby construct many estimates of the drift and diffusion functions. Applying the exact procedure in Stanton (1997), we find that the typical estimated drift function displays non-linearities at high interest rates of exactly the sort found in Stanton, even though the simulated sample paths were generated by a process with a linear drift. The explanation for the poor performance of the nonparametric estimators is a truncation of the distribution that occurs in finite samples but is eliminated asymptotically. This result 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 within a bandwidth of 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 (at best) only a modest reduction in the spurious nonlinearity. In fact, we are unable to find a variant of the kernel regression estimator and bandwidth choice that enables accurate estimation of both the drift and the diffusion functions.

A consistent result for Stanton's estimator (including the jackknife version of the kernel regression estimator) is that the performance improves with decreases in persistence and increases in sample size. The estimates are accurate over the entire range of the data only for implausibly large sample sizes and for persistence levels that are implausibly low for short-term interest rate data. It should be noted, however, that our findings regard-

ing the finite-sample performance of the kernel regression estimator is not entirely negative. Except at very low levels of the process, the diffusion estimator is generally quite accurate. The accuracy of the diffusion estimator is perhaps not surprising, for it has been widely known at least since Merton (1980) that high frequency data permits very precise estimation of the diffusion coefficient.

The estimator in Aït-Sahalia (1996) also suggests that the drift function contains important nonlinearities. In order to understand its finite-sample properties, we applied a simplified version of this estimator to the simulated square-root sample paths. Specifically, we assumed that the exact form of the diffusion function was known, and we attempted to estimate the parameters of the drift, starting the nonlinear optimization problem from the true parameter values. The results of this exercise demonstrate that there is substantial uncertainty associated with the drift parameter that determines the nonlinearity at high levels of the square-root process. In particular, the standard deviation of the estimated value of this parameter across the one hundred simulations is an order of magnitude larger than the uncertainty associated with any of the other drift parameters. The point estimate of this parameter is also extremely sensitive to the manner in which the nonparametric density estimator is constructed, and it can easily suggest important nonlinearities where none exist in the data.

In a related paper, Pritsker (1997) studies the finite-sample properties of Gaussian kernel estimators of the steady-state density, when the data are generated by an Ornstein-Uhlenbeck process. Since this process is Gaussian, he is able to compute the exact mean integrated squared error, optimal kernel density bandwidth parameter, and other properties of the density estimator. His results focus on the finite-sample biases in density estimators of Markov processes with very weak mean reversion (so-called "near unit root" behavior). His principle conclusion is that, in order to achieve finite-sample results that are consistent with asymptotic theory, more frequent sampling (i.e., daily versus monthly) is less important than a long span of data. He also examines the optimal choice of the bandwidth parameter in this context and finds that it differs substantially from the rules generated for *iid* processes. However, he does not examine kernel regression estimators of the type used in Stanton (1997); nor does he address any issues of the finite-sample bias in the nonparametric drift and diffusion estimators in Aït-Sahalia (1996) or Stanton (1997) and their relation to the biases in the estimation of the steady-state density.

A reasonable conclusion to draw from the Monte Carlo evidence is that it is difficult to use the nonparametric estimators in Stanton (1997) and Aït-

Sahalia (1996) 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 generalized method of moments (GMM) estimator, of the form introduced in CKLS, to the data from both Stanton (1997) and Aït-Sahalia (1996). The evidence from this estimator is not consistent with the results of the nonparametric estimators. When GMM is applied to Stanton's Treasury bill data, there is no statistically significant evidence of nonlinearity, and when GMM is applied to Aït-Sahalia's Eurodollar data, the implied nonlinearity in the drift is marginally significant but of the opposite sign from the nonparametric estimates. An alternative would be to apply more sophisticated moment-based estimators to answer the question. However, there is (as yet) no finite sample evidence, that we are aware of, to support an alternate estimator as clearly superior in practice.

The balance of the paper is organized as follows: Section 2 introduces notation, some basic concepts from the kernel density estimation and kernel regression literature, and the estimator introduced in Stanton (1997). The Monte Carlo simulation of Stanton's estimator – including an explanation for the finite-sample nonlinearity – is contained in Section 3, and Section 4 presents the results for the jackknife kernel regression estimator. Section 5 examines a simplified version of the estimator in Aït-Sahalia (1996), and Section 6 presents a simple discretized GMM estimation applied to the data in Stanton (1997). The conclusions and implications for future work contains in Section 7.

2 Kernel Estimates of the Drift and the Diffusion Functions

2.1 Basic Definitions

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, $\mu : \mathbb{R} \to \mathbb{R}$ is the *drift function*, and $\sigma : \mathbb{R} \to \mathbb{R}_+$ is the *diffusion function*. Equation (1) is a short-hand notation for the stochastic integral equation

$$x_t = x_0 + \int_0^t \mu(x_s)\,ds + \int_0^t \sigma(x_s)\,dB_s, \tag{2}$$

where the first integral in (2) is an ordinary Riemann integral and the second integral is an Itô stochastic integral. In the context of this paper, x_t is interpreted as the "instantaneous" or "short" rate.

Under technical conditions on the transition function, x_t is a *diffusion process*.[1] This implies that

$$E\left[x_{t+\Delta} - x_t \mid x_t\right] = \mu\left(x_t\right)\Delta + o\left(\Delta\right) \tag{3}$$

and

$$E\left[\left(x_{t+\Delta} - x_t\right)^2 \mid x_t\right] = \sigma^2\left(x_t\right)\Delta + o\left(\Delta\right), \tag{4}$$

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

Finally, as Arnold (1974) states, the dynamics of the Markov process x_t are completely described by its associated *infinitesimal generator*

$$\begin{aligned}\mathcal{A}\left[f\left(x_t\right)\right] &\equiv \lim_{\Delta \downarrow 0} \frac{E\left[f\left(x_{t+\Delta}\right) \mid x_t\right] - f\left(x_t\right)}{\Delta} \\ &= \mu\left(x_t\right) f'\left(x_t\right) + \frac{1}{2}\sigma^2\left(x_t\right) f''\left(x_t\right), \end{aligned} \tag{5}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a bounded measurable function, f' denotes the first derivative of f with respect to its single argument, and f'' is the second derivative of f.[2] This operator will turn out to be very useful in characterizing alternate approaches to the estimation of the drift and diffusion functions.

2.2 A *Brief* Review of Alternative Estimators

As (3), (4), and (5) suggest, estimation of the drift and diffusion functions is central to understanding both the long-run properties and the short-run dynamics of x_t. One natural approach to estimating μ and σ is to posit specific functional forms that are consistent with a process whose transition density is known in closed-form. For example, if $\mu\left(x\right) = \kappa\left(\theta - x\right)$ and $\sigma\left(x\right) = \sigma\sqrt{x}$, then Cox, Ingersoll, and Ross (1985) demonstrate that the

[1] See Arnold (1974), Section 2.5.

[2] In defining the infinitesimal generator, (5) imposes the assumption of time-homogeneity implied by the form of (1). For the infinitesimal generator of a more general non time-homogeneous Markov process, see Equations (2.4.1), (2.4.5), and page 42 in Arnold (1974).

stationary density of the unique solution to the SDE defined by these functions is a gamma distribution, and the transition density of the process is a noncentral chi-squared density.

Lo (1988) describes how to estimate the parameters of x_t using the method of maximum likelihood, where the likelihood function is constructed from the known form of the transition density.[3] This approach is elegant and the associated estimator possesses all of the desirable properties of a maximum likelihood estimator. Unfortunately, it is usually only practical with the small set of diffusions whose transition densities are known in closed-form.[4]

When likelihood-based estimation is impractical, a variety of moment-based estimators are available. The simplest approach is to discretize (1) using a simple Euler approximation and apply the generalized method of moments introduced in Hansen (1982). CKLS do this in the case where the drift is of the form $\mu(x) = \alpha + \beta x$ and the (local) variance is of the form $\sigma^2(x) = \sigma^2 x^{2\gamma}$. Of course, some of the moments constructed are only approximately correct.

Hansen and Scheinkman (1995) describe a method for constructing exact method-of-moment estimators based on the infinitesimal generator of the continuous-time process.[5] Duffie and Glynn (1996) develope an alternate exact moment-based estimator by sampling at random, as opposed to deterministic, intervals. Finally, Duffie and Singleton (1993) describe simulated moment estimators.

2.3 The Nonparametric Estimator in Stanton (1997)

A nonparametric density estimator is constructed in a standard way, following Silverman (1986), for example.[6] Let $\left\{x_t^{\Delta}\right\}_{t=1}^{T}$ be a sample of size T from

[3] See Pearson and Sun (1994) for an application of this approach.

[4] Santa-Clara (1995) uses simulation methods to extend the likelihood-based estimation approach to cases where the partial differential equation defining the transition density (the Kolmogorov backward or forward equation) cannot be solved explicitly. The drawback to this approach is that it is computationally intensive.

[5] Specifically, for the case of time-homogeneous scalar diffusions, Hansen and Scheinkman (1995) propose two classes of moment conditions: (i) $E(\mathcal{A}[\phi(x_t)]) = 0$ and (ii) $E(\phi(x_t)\mathcal{A}[\phi(x_{t+1})] - \phi(x_{t+1})\mathcal{A}[\phi(x_t)]) = 0$, where ϕ is any "test" function in a dense subset of the set of (almost-surely) square-integrable functions.

[6] Since – thanks in large part to the work of Aït-Sahalia (1996), Boudoukh, Whitelaw, Richardson, and Stanton (1997), and Stanton (1997) – kernel estimation is familiar to financial economists, the treatment given in this section will be brief, and it is designed primarily to introduce required notation. For more detail, see Silverman (1986) or Härdle (1990).

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.[7] If the stationary density of x_t is denoted $\pi(x)$, a kernel estimator is of the form

$$\widehat{\pi}(z_i) \equiv \frac{1}{T \cdot h} \sum_{t=1}^{T} K\left(\frac{z_i - x_t^{\Delta}}{h}\right); \tag{6}$$

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 (6). Stanton (1997) uses a *Gaussian kernel*:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right); \qquad u \in (-\infty, +\infty). \tag{7}$$

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). Unfortunately, the choice of the bandwidth parameter in a time series context in which the data are highly autocorrelated is problematic. Most of the results on bandwidth choice in the kernel estimation literature apply only to data which are independent draws from a given (although unknown) density, i.e., "*iid* observations." Pritsker (1997) evaluates the impact of persistence in the process on the choice of optimal bandwidth in the context of a Gaussian process.

His results can be summarized as follows: (1) The bandwidth that minimizes the mean integrated squared error relative to the true density is very sensitive to the autocorrelation in the data, and it is much larger than the optimal choice in the case of *iid* data; (2) The optimal bandwidth is very insensitive to the frequency with which the data is sampled, which means that there is only a small change in the optimal bandwidth for monthly versus daily observations; and (3) The optimal bandwidth is decreasing in the

[7] 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.

span of the data. However, it does not decline dramatically with realistic increases in the sample size. For example, for parameter values consistent with the data used in Aït-Sahalia (1996), Pritsker (1997) computes the decrease in the optimal bandwidth in moving from ten years of daily data to twenty years of daily data is only about twenty percent.

In the Monte Carlo simulations examined in the next section, two choices of the bandwidth parameter are used for each parameter combination. The first choice is the optimal bandwidth for the *iid* case: $h = \widehat{\sigma} T^{-1/5}$, where $\widehat{\sigma}$ is the sample standard deviation of the data and T is the sample size. This choice has three advantages in the current analysis. (*i*) These bandwidths are the ones used in Stanton (1997), and this choice is, therefore, helpful in understanding the finite-sample properties of his estimator. (*ii*) It is much smaller than the bandwidths suggested in Pritsker (1997). Finally, (*iii*) the *iid* bandwidth procedure produces bandwidth values that are close to those chosen by other common data-dependent bandwidth selection procedures such as "plug-in" and "solve the equation" methods.[8] The second bandwidth choice is the one implied by the simulation results in Pritsker (1997).[9] Even though they are explicitly optimal only for the Gaussian case, they are the only choices available that explicitly vary with the degree of persistence of the sampled process.

The estimators in Stanton (1997) are based directly on (3) and (4), above.[10] In particular, "inverting" these equations yields:

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

and

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

The essence of Stanton's approach is to apply the Nadaraya-Watson (N-W) kernel regression estimator to construct nonparametric estimates of the

[8] Whether or not they are close to the bandwidths that would be chosen by cross-validation methods is an open question, since cross-validation is too computationally expensive in the current context.

[9] The parameter choices used in examining the square-root diffusion are deliberately chosen to be consistent with the unconditional moments of the parameterizations examined in Pritsker (1997). Two points should be emphasized strongly: (1) The bandwidth choices in Pritsker (1997) were constructed in the context of density estimators and not kernel regression estimators, and (2) they were developed for Gaussian processes while the square-root process has a noncentral chi-squared conditional distribution function.

[10] Stanton (1997) justifies his estimators using Taylor series expansions involving the infinitesimal generator, but for the simple approximations examined here, the estimators are an immediate consequence of the definition of a diffusion process.

conditional expectations in (8) and (9):

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

and

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

for $i = 1, \ldots, N$, where $\left\{x_t^{\Delta}\right\}_{t=1}^{T}$ 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 (7), and h is chosen in the manner described earlier.

An examination of (10) and (11) suggests that a potential problem with the N-W estimators occurs in the tails of the empirical estimate of the density. The term $\sum_{t=1}^{T-1} K\left(\frac{z_i - x_t^{\Delta}}{h}\right)$ is $Th \cdot \widehat{\pi}\left(z_i\right)$ and when the estimate of the density is small, the estimator may become numerically unstable. Intuitively, even though there may be a large number of observations of x_t, there are (by definition) very few observations in the extremes of the tails. Since the kernel regression estimator (in effect) only uses "local" data in estimating the conditional moments, they may be measured with a substantial amount of error.

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

3.1 Simulating a Square-Root Diffusion

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

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

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, and (as noted earlier), its (true) steady-state density is

$$\pi(x) = \frac{\omega^{\nu}}{\Gamma(\nu)} x^{\nu-1} \exp(-\omega x), \tag{13}$$

where $\omega \equiv 2\kappa/\sigma^2$, $\nu \equiv 2\kappa\theta/\sigma^2$, and Γ is the Gamma function.

The transition density for the square-root diffusion between any two dates s and t $(s > t)$ is:

$$p(x_s, s \mid x_t, t) = c \exp(-u - v) \left(\frac{v}{u}\right)^{q/2} I_q\left(2\sqrt{uv}\right), \tag{14}$$

where

$$c \equiv \frac{2\kappa}{\sigma^2 \left[1 - \exp(-\kappa(s-t))\right]},$$

$$u \equiv c x_t \exp(-\kappa(s-t)),$$

$$v \equiv c x_s,$$

$$q \equiv \nu - 1,$$

and I_q is the modified Bessel function of the first kind of order q. This is the noncentral chi-square density with $2q + 2$ degrees of freedom and noncentrality parameter $2u$. The form of (14) is given in CIR and in Feller (1951). The unconditional moments of the square-root process are

$$E[x] = \theta, \tag{15}$$

$$\text{Var}[x] = \frac{\theta\sigma^2}{2\kappa}, \tag{16}$$

and

$$\text{Corr}[x_{t+\Delta}, x_t] = \exp(-\kappa\Delta). \tag{17}$$

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 can be constructed in two steps: (1) Draw an initial value from the Gamma density (13); and (2) Using this initial value, simulate successive observations by drawing from the noncentral chi-squared density (14).[11] The simulated sample paths of (12) are constructed assuming that

[11] In the simulations reported below, this is accomplished using the noncentral chi-squared random number generator (implemented using the **ncx2rnd** function) in Matlab, Version 5.0.

the length of time between observations of the diffusion is $\Delta = 1/250$, corresponding to daily observations. The length of the simulated series, T_{sim}, is chosen from the set $\{7500, 15000, 30000\}$, in order to understand the impact of increasing sample size on the properties of the estimators.

κ is chosen from the set $\{0.21459, 0.85837, 1.71624\}$. The first value implies a (monthly) autocorrelation of the short rate of 0.982, which is consistent with the upper end of the estimates of this parameter based on US 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), and the third choice for κ is a lower bound case of a first-order autocorrelation coefficient equal to 0.867.[12] 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 (1996) data set.[13] Given a (θ, κ) pair, the value of σ is chosen in order to set (16) equal to the sample variance in the Aït-Sahalia (1996) data set. This implies a set of values for σ equal to $\{0.07830, 0.15660, 0.22143\}$, where the ordering is consistent with the ordering of the κ values.

3.2 The Results

The results of applying the nonparametric density estimator in (6) to the simulated sample paths associated with each model parameterization, simulation length, and bandwidth choice are shown in Figures 1 through 3. These figures all have the same structure: The left column of graphs reports the density estimates for the *iid* bandwidth, and the right column reports the same results for the Pritsker bandwidth. Table 1 reports the bandwidth parameter values used for all simulation length and parameter combinations. As required by the theory of kernel estimation, both sets of bandwidths decrease as the sample size increases. The Pritsker bandwidths are substantially larger than the corresponding *iid* bandwidths.[14]

Each row of each figure reports the results for simulations of a given

[12] As noted earlier, these parameter values are also chosen to be consistent with the analysis in Pritsker (1997) of the optimal bandwidth choice in the presence of extremely persistent data.

[13] This choice of θ is not exactly equal to the value of 0.08912 in Aït-Sahalia's data, but it is close. There is no evidence to suggest that a change in the long-run mean of the simulated data will have an important effect on the analysis of the estimators conducted below. The slight difference was necessary to ensure the integer degrees of freedom necessary for the noncentral chi-squared random number generator in Matlab.

[14] We would like to thank Matt Pritsker for generously providing the bandwidths that correspond to the sample sizes used in the simulations reported here.

sample size, with the sample size increasing toward the bottom of the figure. Figure 1 reports the results for the simulation with the greatest level of persistence. Figure 2 is the intermediate persistence parameterization, and Figure 3 reports the results for the least persistent version of the square-root process. Each graph also has a common structure. It consists of three lines. The dashed line is the pointwise average at each gridpoint in the support of the stationary density across the one hundred simulations. The two dotted lines are the 25^{th} and 75^{th} percentile points for the density (again, for each gridpoint). Notice that the true densities are the same in every graph in each figure. This is a direct result of the choices of κ and σ in each parameterization. All of the estimated densities are closer to the true density as T_{sim} increases, exactly as expected. It is also the case that the estimates are more accurate as κ increases; i.e., as persistence in the data decreases. This is also perfectly reasonable, since lower persistence implies – in a heuristic sense – more data.

The results from applying the kernel regression estimators from Stanton (1997) to the simulated square-root data are shown in Figures 4 through 9. Figures 4 through 6 examine the pointwise average and 25-th and 75-th percentile points of the drift estimates for the three parameterizations varying both the bandwidth and the number of simulated observations.

The results for the *iid* bandwidth choice are striking. The kernel regression estimates can exhibit substantial nonlinearity. First, consider the top left graph in Figure 4. This corresponds to the sample size and bandwidth choice used in Stanton (1997). The average drift estimator is only close to the true drift for a small interval in the heart of the stationary density. It diverges rapidly from the true linear drift at both the extreme lower and upper ends of the support of the stationary density. Equally important, the 25^{th} and 75^{th} percentile bands indicate that there is substantial variation in individual estimates of the drift function and that these estimates are particularly inaccurate in the right tail of the density. It would be easy on the basis of a picture like this and the asymptotic theory to conclude that the drift of the underlying process is highly nonlinear. However, this nonlinearity is entirely spurious.

The drift estimates, based on the *iid* bandwidth choice, for other choices of persistence and other sample sizes are consistent with the results for the $\kappa = 0.21459$ and $T_{sim} = 7,500$ case. The remaining graphs in Figure 4 indicate that the estimator becomes more accurate and the dispersion in the estimates across the 100 simulated sample paths becomes smaller for larger sample sizes. For the case of $T_{sim} = 30,000$, the *iid*-based drift estimator is accurate for a wide range of values in the heart of the support of the

stationary density, although it still exhibits some spurious nonlinearity for levels of the process in excess of sixteen percent. Figures 5 and 6 confirm the pattern observed in Figure 4. The nonparametric estimators do become more accurate as the persistence in the data decreases (as κ increases). Intuitively, this effect is similar to an increase in the sample size. In summary, the accuracy of the *iid* drift estimator improves as T_{sim} increases and as κ increases, but the spurious nonlinearity in the drift estimator only becomes negligible for unrealistic parameter values and sample sizes.

The kernel regression estimators of the drift based on the bandwidth recommendations in Pritsker (1997) are generally accurate, although they are slightly nonlinear.[15] This is true for all parameterizations and for all sample sizes examined in Figures 4 through 6. If drift estimation is the sole objective of the analysis, the prediction from these simulations is clear: Choose a large bandwidth that oversmooths the stationary density. Of course, this prescription is only appropriate if the true drift is linear, because what is really going on in these figures (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.

A partial explanation for the poor performance of the *iid* drift estimator at high levels of x is that the kernel regression estimator only uses local data in defining the regression function around a point z_i in the partition of the support of the stationary density. This means that, even though there may be, for example, 7500 observation in total, there are (by construction) very few observations in the long right tail of the density. In effect, there is a "small sample" problem in constructing the nonparametric estimator at high levels of x. This explains why the estimator is imprecise where there are few observations (and it explains why the oversmoothed estimator does not exhibit this problem), but it does not explain the direction of the bias, which is discussed in the next section.

Figures 7 through 9 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 does a good job of providing an accurate estimate of the diffusion function, for all but the most extreme combinations of persistence and sample size, and the estimates based on the larger Pritsker bandwidths are less accurate because they oversmooth the diffusion function. However, the

[15] This is most readily apparent in Figure 6.

Pritsker estimates improve with both increases in sample size and decreases in persistence. 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) shows how to modify the estimator to constrain it to go through zero.

3.3 Why *Is* the Estimated Drift Nonlinear?

From equation (3), it follows 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. Asymptotic theory (for example, Robinson (1983, 1986)) establishes the pointwise consistency of kernel regression estimators in a time series context. Therefore, the spurious non-linearity documented in the previous section must be a finite sample property of the estimators. Figure 10, which is based on a single sample path from the parameterization $\kappa = 0.85837$, $\sigma = 0.15660$, and $\theta = 0.085711$ using 7500 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 $x_t \geq 0.14$. This is approximately the point at which the average estimated drift function in the top left plot of Figure 5 starts to diverge from the true drift.

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

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

where the second equality uses the parameter values and $\Delta = 1/250$ (one day) used to simulate the process. This line is plotted in Figure 10, along with the estimated OLS regression line using only the tail observations. Notably, the northeast 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 non-positive, 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 northeast corner of Figure 10 is empty by necessity, not 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 between $x_{t+\Delta} - x_t$ 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 $x^{\max} = \infty$.

In order to understand the effect of the truncation of the distribution of realized interest rates, we consider how our beliefs about the distribution of the realizations of a square root process, and particularly the conditional mean of our beliefs, are affected by the knowledge that the realizations lie in a range $[x^{\min}, x^{\max}]$. Recall that the transition density of the square-root process is given in (14). Over such a short time interval the density is approximately symmetric, and the conditional mean and (approximate) drift are

$$E[x_{t+\Delta}|x_t] = \theta + e^{-\kappa\Delta}(x_t - \theta)$$

and

$$\begin{aligned} \mu(x_t) &= [E[x_{t+\Delta}|x_t] - x_t]/\Delta + o(\Delta)/\Delta \\ &\approx \left[\theta\left(1 - e^{-\kappa\Delta}\right) + \left(e^{-\kappa\Delta} - 1\right)x_t\right]/\Delta \end{aligned}$$

Using the parameter vector above, when $x_t = 0.16$ the conditional mean and drift are 0.1598 and approximately -0.06, respectively.

Next, suppose that we condition on the knowledge that $x^{\min} \leq x_{t+\Delta} \leq x^{\max}$. Conditional on both this inequality and x_t, it is straightforward to show that our beliefs about $x_{t+\Delta}$ are given by the density

$$g\left(x_{t+\Delta}|x_t \text{ and } x_{t+\Delta} \in \left[x^{\min}, x^{\max}\right]\right) = \begin{cases} \frac{f(x_{t+\Delta}|x_t)}{\int_{x^{\min}}^{x^{\max}} f(u|x_t)du} & \text{if } x_{t+\Delta} \in \left[x^{\min}, x^{\max}\right], \\ 0 & \text{otherwise.} \end{cases}$$

Figure 11 illustrates this density when $x_t = 0.16$, $x^{\min} = 0.0132$, and $x^{\max} = 0.1618$. Since x_t is close to $x^{\max}$, the truncation at $x^{\min}$ has little effect on the conditional density and the truncation at $x^{\max}$ has the effect of skewing the distribution to the left. It is clear that, when x_t is close to $x^{\max}$, the truncation at $x^{\max}$ reduces the conditional mean of $x_{t+\Delta}$. We can calculate the conditional mean and the (approximate) drift from

$$E\left[x_{t+\Delta}|x_t \text{ and } x_{t+\Delta} \in \left[x^{\min}, x^{\max}\right]\right] = \frac{\int_{x^{\min}}^{x^{\max}} u f(u|x_t) du}{\int_{x^{\min}}^{x^{\max}} f(u|x_t) du}$$

and

$$\mu(x_t) \approx \left(E\left[x_{t+\Delta}|x_t \text{ and } x_{t+\Delta} \in \left[x^{\min}, x^{\max}\right]\right] - x_t\right)/\Delta.$$

Using the same parameter vector as above, when $x_t = 0.16$ these are 0.15977 and -0.0567, respectively. As expected, the conditional mean is smaller, and the absolute value of the drift is larger, than when we did not condition on the fact that $x^{\min} \leq x_{t+\Delta} \leq x^{\max}$.

This calculation, while intuitive, actually understates the bias. The expected value of the next observation is not

$$E\left[x_{t+\Delta} \mid x_t \text{ and } x_{t+\Delta} \in \left[x^{\max}, x^{\min}\right]\right], \tag{18}$$

but rather the expected value conditional on all of the other observations being in the interval $\left[x^{\max}, x^{\min}\right]$. For x_t near $x^{\max}$, this is smaller than (18), while for x_t near $x^{\min}$, it is larger. Due to the dimensionality of the integral, we cannot present an explicit calculation of this expectation. However, our Monte Carlo analysis of the kernel regression estimator provides an approximation of this integral.

It is important to note that this bias is distinct from 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, while the bias demonstrated above will occur even when this does not occur. For example, the *iid* bandwidths reported in Table 1 are all approximately 0.005. As long as z is a distance of approximately 0.015 from the boundary, a Gaussian kernel using these bandwidths will assign a trivial weight to the boundary points, so that the boundary value bias will not be material. Comparing the *iid* panels of Figures 1 through 6, one can see that the bias occurs well within the interior of the distribution of the data, more than 0.015 from the boundary. In addition, in the next section, we demonstrate that the (boundary) bias-corrected jackknife estimator of Rice (1984) does not eliminate the problem.

To summarize this discussion, Figures 10 and 11 illustrate that, when x_t is near $x^{\max}$, $x_{t+\Delta}$ is more likely to be less than x_t than the true drift implies. Similarly, when x_t is near $x^{\min}$, $x_{t+\Delta}$ is more likely to be greater than x_t than the true drift implies. This is the source of the bias in the N-W kernel regression estimator in (10) that is used by Stanton (1997).

Intuitively, at each point x in some subset of the support of x_t, it estimates the drift by taking the realizations of the process that are near x, and seeing what typically is the realization of the process at the next time step. When x is near $x^{\min}$ (or $x^{\max}$) the realization at the next time step is more likely to be high (or low) than the true drift implies, biasing the estimation of the drift.

4 The Jackknife Kernel Estimator

4.1 Defining the Estimator

As noted earlier, it is well-known in the kernel regression literature that these estimators can exhibit increased bias within a bandwidth of the upper and lower boundaries of the support of the sample observations.[16] Rice (1984) proposes the *jackknife kernel estimator* as a solution to this "boundary effect."[17] 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 $\left[\underline{b}, \overline{b}\right]$ and assume that the kernel function is defined over the closed interval $[-1, 1]$.[18] Recall that the simple N-W kernel estimator with bandwidth parameter h, described earlier, has the general form

$$m_h(x) = \frac{\sum_{t=1}^{T-1} \left(x_{t+1}^{\Delta} - x_t^{\Delta}\right) K\left(\frac{x - x_t^{\Delta}}{h}\right)}{\sum_{t=1}^{T-1} K\left(\frac{x - x_t^{\Delta}}{h}\right)}. \tag{19}$$

Let any gridpoint $x \in \left[\underline{b}, \overline{b}\right]$ 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

$$\omega_K(0, \rho) \equiv \int_{\underline{b}}^{\rho} K(z)\, dz \tag{20}$$

and

$$\omega_K(1, \rho) \equiv \int_{\underline{b}}^{\rho} zK(z)\, dz. \tag{21}$$

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

[17] This estimator is also discussed in Section 4.4 of Härdle (1990).

[18] Obviously, this analysis precludes the case of the Gaussian kernel used earlier.

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

$$m_h^J(x) = (1+\phi)\, m_h(x) - \phi m_{\xi h}(x), \tag{22}$$

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

$$\phi \equiv \frac{\omega_K(1,\rho)/\omega_K(0,\rho)}{\xi\omega_K(1,\rho/\xi)/\omega_K(0,\rho/\xi) - \omega_K(1,\rho)/\omega_K(0,\rho)}. \tag{23}$$

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

The jackknife kernel estimator will be implemented using a quartic kernel

$$K_q(z) = \frac{15}{16}\left(1 - z^2\right)^2, \tag{24}$$

for $|z| \leq 1$. This kernel is defined on the compact interval $[-1,1]$, and it permits an explicit computation of (20) and (21)

$$\omega_{K_q}(0,\rho) = \frac{3}{16}\rho^5 - \frac{5}{8}\rho^3 + \frac{15}{16}\rho + \frac{1}{2},$$

$$\omega_{K_q}(1,\rho) = \frac{5}{32}\rho^6 - \frac{15}{16}\rho^4 + \frac{15}{16}\rho^2 - \frac{5}{32}.$$

The use of the quartic rather than the Gaussian kernel means that the data-dependent *iid* bandwidth choices used in the previous section need to be modified. We are unaware of any simple heuristic rule for this kernel, but following the analysis in Chapter 5 of Härdle (1990), the data dependent bandwidth was set at three times the corresponding Gaussian *iid* bandwidth. The next section considers whether the boundary correction in the jackknife kernel estimator result in substantially improved inference about the non-linearity of the drift function.

4.2 A Monte Carlo Evaluation of the Jackknife Estimator

We apply the same Monte Carlo analysis to this estimator as was applied to the original Stanton (1997) estimator. Figures 12 through 17 reproduce Figures 4 through 6 using the jackknife kernel. For ease of comparison, the scales on the axes are identical across the two sets of pictures. First, consider the *iid* bandwidth case in Figure 12. The jackknife kernel results in

a noticeable improvement relative to the standard N-W estimator in Figure 4, but correcting the boundary effect does not eliminate the spurious nonlinearity in the kernel regression estimator. This is true for all three sample sizes. Of course, the overall performance of the estimator improves with the length of the simulated series. The results in Figures 13 and 14 are similar to those in Figure 12. As in the case of the standard N-W kernel regression estimator, the estimates improve with decreases in the persistence of the process.

The jackknife kernel results based on the larger bandwidths consistent with Pritsker (1997) are quite interesting. As in the earlier analysis, these bandwidths are not optimal because the process is not Gaussian. In the case of the jackknife kernel, these bandwidths are also incorrect because we are no longer using a Gaussian kernel. Nonetheless, this case demonstrates the impact of a large bandwidth on jackknife estimator. The estimates of the drift function are substantially worse than the (oversmoothed) regular N-W estimates. They exhibit the kind of spurious nonlinearity that is common with the *iid* bandwidth setting. This is true across all levels of autocorrelation and sample size, although the usual effects apply here as well. These results are interesting because they are evidence of one case in which the choice of the kernel seems to be important in determining the performance of the estimator.

The effects of the boundary correction on the diffusion estimators are shown in Figures 15 through 17. As in Figures 7 through 9, the estimators based on the *iid* bandwidth choice are quite accurate. The boundary correction has a strong positive impact on the diffusion estimator based on the (larger) Pritsker bandwidths. In particular, these estimates are now quite accurate and virtually indistinguishable from the *iid* bandwidth results. There appears to be a trade-off in the wider bandwidth choices between an accurate estimation of the drift or the diffusion function. In summary, the overall effect of the boundary correction in the jackknife estimator is to reduce the spurious nonlinearity in the data, but it does not eliminate it. This implies that the spurious nonlinearity documented using the N-W kernel regression algorithm is not simply a result of the boundary effect, but rather it follows from the truncation argument in the last section and the extreme persistence of interest rate data.

5 The Estimator in Aït-Sahalia (1996)

As noted in the introduction, Aït-Sahalia (1996) also concludes that there are statistically and economically significant nonlinearities in the drift of the short-rate. Since his estimation approach is semi-nonparametric, we will now consider some Monte Carlo evidence on its performance.

5.1 The Definition

The general form of the drift and diffusion estimators proposed in Aït-Sahalia (1996) are based on the stationary (or invariant) distribution of x.[19] 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 (6), as described above. The second step in the estimation of the drift and diffusion functions is to use results in Karlin and Taylor (1981) to relate the stationary density $\pi(x)$ to the drift and diffusion functions.[20] 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.[21] 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 representation of the drift and diffusion functions

[19] Aït-Sahalia (1996) also develops drift and diffusion estimators based on the transition densities, using the forward and backward Kolmogorov equations, but we do not evaluate them here.

[20] See, in particular, Karlin and Taylor (1981) Chapter 15, Section 5 (pages 220-221) and Section 6 (pages 241-242) for the connection, through the stationary form of the Kolmogorov forward equation, between the drift and diffusion functions and the stationary density.

[21] The notation in (25) generally follows Aït-Sahalia (1996).

in (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 nesting 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

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

so that $\psi \equiv (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_2, \beta_3)'$. Of course, there are a number of restrictions on the elements of ψ that are necessary to ensure that μ and σ actually imply that (1) has a unique solution and that the implied stationary density exists.[22]

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 (1996) proposes a minimum mean square distance measure. In particular, he chooses

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

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

$$\widehat{\psi}^* \equiv \underset{\psi\in\Psi}{\arg\min}\, \frac{1}{T}\sum_{i=1}^{T}\left(\pi(x_t;\psi) - \widehat{\pi}(x_t)\right)^2, \tag{29}$$

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

5.2 Some Monte Carlo Evidence

The general form of the estimator as described in Section 5.1 cannot be applied directly to the case of the square-root diffusion (12) since, as Aït-Sahalia (1996) notes, (27) is not identified under the null hypothesis. Since

[22] See Aït-Sahalia (1996), Equation (24).

our focus is on the estimate of the drift function, we will make the strong assumption that the true diffusion function is known precisely, including the value of σ. In this case,

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

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

Finding good starting values for the parameter estimation is a nontrivial issue. One approach is to follow Aït-Sahalia (1996) and use a feasible generalized least squares (FGLS) algorithm based on the Euler approximation to the SDE implied by the drift and variance functions (26) and (27), respectively. Instead, in the results reported below, we started the optimization problem at the true parameter vector; i.e., $\alpha_0^0 = \kappa\theta$, $\alpha_1^0 = -\kappa$, and $\alpha_2^0 = \alpha_3^0 = 0.0$. While this is admittedly unrealistic in a practical context, if anything, it will bias the results in favor of the estimator.[23]

Tables 2 through 7 examine the finite-sample performance of the Aït-Sahalia (1996) estimator across the 100 sample paths for each simulation length and parameter combination. Tables 2 and 3 examine the case of $\kappa = 0.21459$, Tables 4 and 5 are for $\kappa = 0.85837$, and Tables 6 and 7 examine $\kappa = 1.71624$. There are a few common results across these three pairs of tables. First, the accuracy of the estimators – as measured by the mean of the distribution across the one hundred simulations – is increasing in the length of the simulated series. Second, the standard deviation of the estimates is also decreasing in the simulation length. Third, the cross-sectional standard deviation of all parameter estimates is increasing in σ. Finally, the standard deviations of the estimators are uniformly lower for the (larger) Pritsker bandwidths, when compared with the results for the corresponding *iid* bandwidths.

There are also striking differences across the tables. First, the results for the *iid* bandwidth cases show that the point estimate of the drift parameters α_0, α_1, and α_3 are reasonably accurate, but the coefficient on the quadratic

[23] The numerical calculations for minimizing the sum of squared errors in (29) for each of the 100 simulated sample paths was conducted using the "*minerr*" function in Mathcad7 (Professional). The sum of squared errors is minimized using a modification of the Levenberg-Marquardt method (LM algorithm). The actual code is based on the MINPACK algorithms developed by the Argonne National Laboratory with an additional random perturbation step added at the end of the first convergence, in order to reduce the chance that the algorithm stops at a local minimum. For details on the LM algorithm, see the MathCad 6.0 Manual or Moré (1977).

term in the drift, α_2, is consistently negative and skewed to the left for the two parameterizations that are consistent with the data used in prior studies. This negative sign is reversed in Table 6 for the case of $\kappa = 1.71624$ with T_{sim} equal to 15,000 or 30,000. In all cases, however, the standard deviation across simulations for α_2 is substantially larger than the cross-sectional standard deviation on any of the other drift parameters. This is consistent with the idea that it is very difficult to measure with accuracy the nonlinearity in the drift for large levels of the square-root process. The concavity of the drift at high levels of the short rate, documented in earlier studies, may be partly a reflection of this bias in this parameter estimate.

The Pritsker bandwidth results – shown in Tables 3, 5, and 7 – also produce reasonably accurate estimates of α_0, α_1, and α_3, but the point estimates of α_2 are now positive and the sample distribution is skewed to the right! The sample standard deviation of the estimates of this parameter are, again, large. The point estimates decrease as T_{sim} increases, but (as noted above) they increase substantially with increases in σ. Apparently, the data dependent bandwidth selection procedure for the Pritsker bandwidths does not decrease at a fast enough rate.

One possible response to comparing the results for the *iid* versus Pritsker bandwidths might be termed the "Goldilocks effect;" i.e., small bandwidth choices result in downward biased estimates and large bandwidth choices result in upward biased estimates, so a choice in the middle will be "just right."[24] It may well be the case that a gridsearch over possible bandwidth choices will result in an estimator that is unbiased in finite-samples for realistic persistence parameters and sample sizes *for the underlying square-root diffusion*, but that choice will undoubtedly be dependent on the choice of the true underlying process. More importantly, this parameter that governs nonlinearity in the fitted drift for large levels of the process is inherently measured with a great deal of uncertainty. The standard deviation across the one hundred simulations is consistently an order of magnitude larger than the standard deviations for any other parameter.

This result is consistent, heuristically, with the arguments made in the previous section to explain the biases in the estimator in Stanton (1997). The term $\alpha_3 x^{-1}$ and has relatively little impact on the drift function for moderate and large values of x, while the term $\alpha_2 x^2$ has relatively little

[24] In fact, the actual bandwidth choice used in Aït-Sahalia (1996) lies roughly in the middle of the range spanned by the *iid* and Pritsker bandwidth choices. Results not reported in the paper show that this choice produces a (slightly) upward biased estimate of the quadratic parameter, with a standard deviation across simulations that is virtually identical to Panel A of Table 4. These results are available upon request.

impact for moderate and small values of x. Thus, the estimates of α_2 and α_3 place a great deal of weight on the realizations near $x^{\min}$ and $x^{\max}$, respectively. We have already seen that when x_t is near $x^{\min}$, $x_{t+\Delta}$ is more likely to be greater than x_t than the true drift implies, and when x_t is near $x^{\max}$, $x_{t+\Delta}$ is more likely to be less than x_t than the true drift implies. The spurious non-linearity produced by the estimator is due to the fact that it accommodates this by letting α_2 and α_3 be non-zero, even though the true drift is linear with $\alpha_2 = \alpha_3 = 0$.

6 So, Is the Drift Actually Nonlinear?

The primary conclusions of the Monte Carlo analyses presented above are as follows. First, there is a truncation 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, and kernel regression estimators are particularly sensitive to this bias since they use local data more intensively than conventional parametric estimators. In fact, if one "corrects" the results in Stanton (1997) for the bias we have documented, the drift appears to be nearly linear. Second, the performance of the kernel regression estimators are sensitive to the choice of the bandwidth parameter, which is difficult to specify a priori in the absence of a specific underlying null hypotheses for the regression functions. Third, both the Stanton (1997) and Aït-Sahalia (1996) 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. Overall, the Monte Carlo results indicate that Stanton (1997) and Aït-Sahalia (1996)do not provide convincing evidence of non-linearity.

To confirm this, we consider an alternate estimation approach to address the question of the linearity or nonlinearity of the short rate drift. The simplest alternative is to consider a GMM approach applied to a discretized version of a generalized short rate process. In particular, using a simple extension of the estimator in CKLS, we estimate

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

where

$$E\left(\varepsilon_{t+1}\right) = 0 \qquad \text{and} \qquad E\left(\varepsilon_{t+1}^2\right) = \sigma^2 x_t^{2\gamma}. \tag{32}$$

Since GMM estimation is now standard in the macroeconomics and finance literature, there is no need to describe the general approach. The

precise application to the problem at hand is explained in detail in CKLS. Two additional moment conditions are added, in this case, in order to exactly identify the model's six parameters. The vector of moment conditions are

$$f_t(\psi) \equiv \begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_{t+1} x_t \\ \varepsilon_{t+1} x_t^2 \\ \varepsilon_{t+1} x_t^{-1} \\ \varepsilon_{t+1}^2 - \sigma^2 x_t^{2\gamma} \\ \left(\varepsilon_{t+1}^2 - \sigma^2 x_t^{2\gamma}\right) x_t \end{bmatrix}, \tag{33}$$

where $\psi \equiv \left(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma^2, \gamma\right)'$. The moment weighting matrix used in the estimation and in calculating the (asymptotic) covariance matrix is calculated using a Bartlett kernel with 60 lags.[25,26]

The data used in the GMM estimation are the daily observations on the three-month Treasury bill from January 1965 to July 1995 used in Stanton (1997) and the daily observations on the seven day Eurodollar rate used in Aït-Sahalia (1996).[27] The results of the simple, discretized GMM estimation are presented in Table 8. There is little evidence against linearity in the drift of the three-month Treasury bill yield, but the point estimate of the quadratic term in the drift using the Eurodollar rate is significantly *greater* than zero. This is in marked contrast to the nonparametric estimators, which suggest a negative coefficient. The point estimates for the Treasury bill data suggests that there is extremely strong persistence in the data.

The parameter estimates in Table 8 are consistent with "volatility induced stationarity" of the interest rate process, as described in Conley, Hansen, Luttmer, and Scheinkman (CHLS).[28] The evidence in favor of lin-

[25] The estimates were also constructed using the pre-whitened covariance matrix estimator with automatic lag truncation selection as described in Newey and West (1994). This resulted in no material differences in the point estimates of the parameters or in the estimates of the parameter standard deviations.

[26] 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.

[27] We would like to thank both Yacine Aït-Sahalia and Richard Stanton for generously providing their data sets.

[28] "One special case of volatility-induced stationarity is when the drift is constant and positive and the variance elasticity exceeds one. Another is when the drift is linear . . . (in x) . . . and the variance elasticity exceeds two." [CHLS, page 534]

earity in the drift of the three-month Treasury bill yield is not entirely inconsistent with the Federal funds rate data analyzed in CHLS. There, for a pre-specified variance elasticity of three (which is close to the $2 \times \gamma = 2 \times 1.623 = 3.246$ point estimate in Table 6), specification tests based on the stationary density produced mixed results.[29]

It is well-known that the simple GMM estimator used to produce Table 8 induces a discretization bias that comes from the first-order approximation used in (31) and (32). Some of the evidence in Stanton (1997) on the effect of higher-order approximations on estimates based on daily data suggests that this bias may not be too large, but it remains an open question, and an application of the techniques introduced in CHLS – once their finite-sample properties are better understood – would provide important additional information on the issues addressed in this section.

7 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 (1997). 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, while undersmoothed estimates are particularly susceptible to the biases documented above. Surprisingly these bandwidth issues also carry over into the semi-nonparametric estimator of Aït-Sahalia (1996). In addition, a major conclusion from a Monte Carlo analysis of this drift estimator is that the standard errors on the quadratic term are so large as to make useful inference problematic at best.

The overall conclusion that we draw from the Monte Carlo evidence is that the nonparametric and semi-nonparametric 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, GMM estimation based on a simple first-order discretization of the data, as in CKLS, is examined. The evidence from this estimator is mixed. The Treasury bill data does not produce any statistically reliable

[29] Specifically, tests using the Fed funds data and six "test functions" failed to reject the linear model at a conventional ten percent significance level, but estimates based on eight test functions did clearly reject the model.

evidence of nonlinearity. The Eurodollar data suggests a positive coefficient on the quadratic term in the drift, which is inconsistent with the evidence presented in Stanton (1997) and Aït-Sahalia (1996). It is possible that this question can be resolved by the application of the more sophisticated moment-based estimation techniques of CHLS or Duffie and Glynn (1996). However, until the finite-sample performance of these estimators is understood, the nonlinearity or linearity of the short rate drift will have to remain an open question.

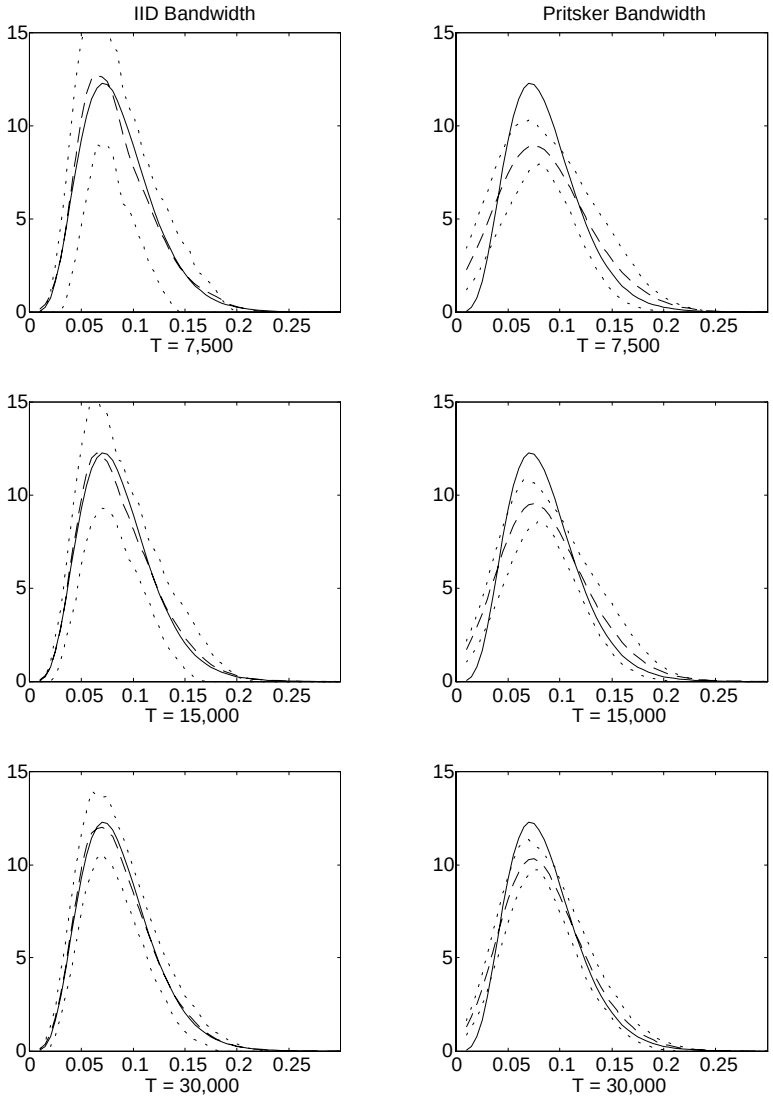

Figure 1: The Stationary Density of the Square-Root Process. $\kappa = 0.21459$, $\theta = 0.085711$, and $\sigma = 0.07830$. The solid line is the true gamma density. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

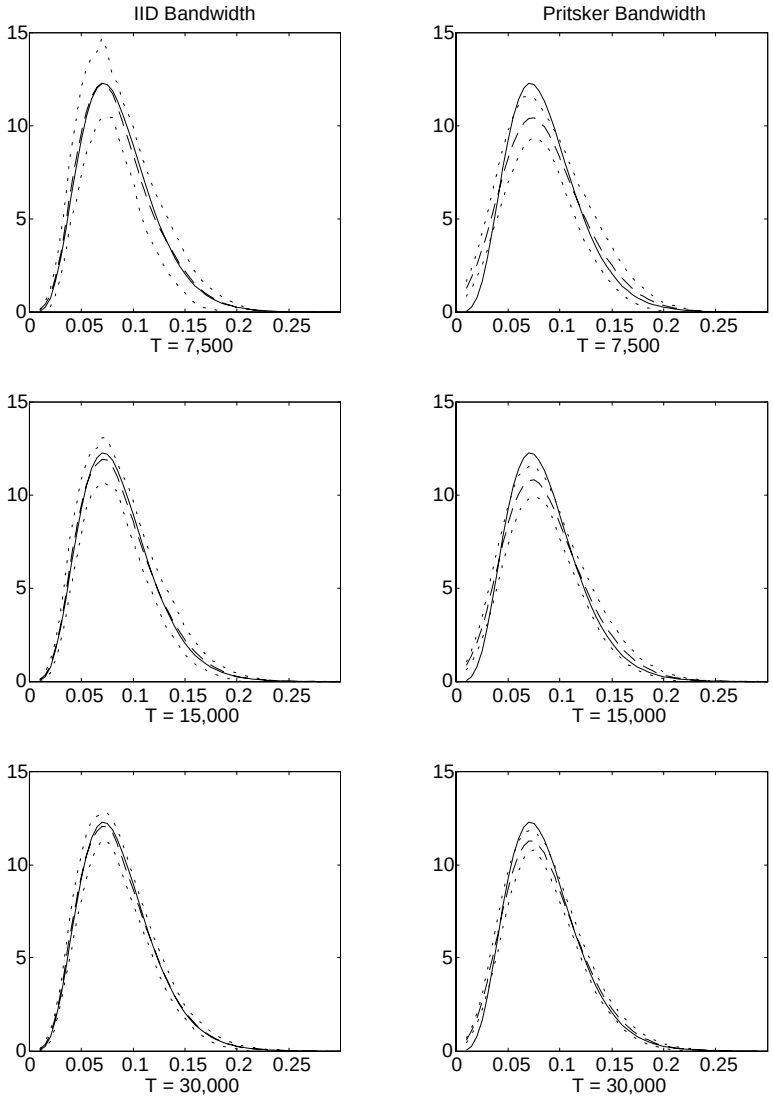

Figure 2: The Stationary Density of the Square-Root Process. $\kappa = 0.85837$, $\theta = 0.085711$, and $\sigma = 0.15660$. The solid line is the true gamma density. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

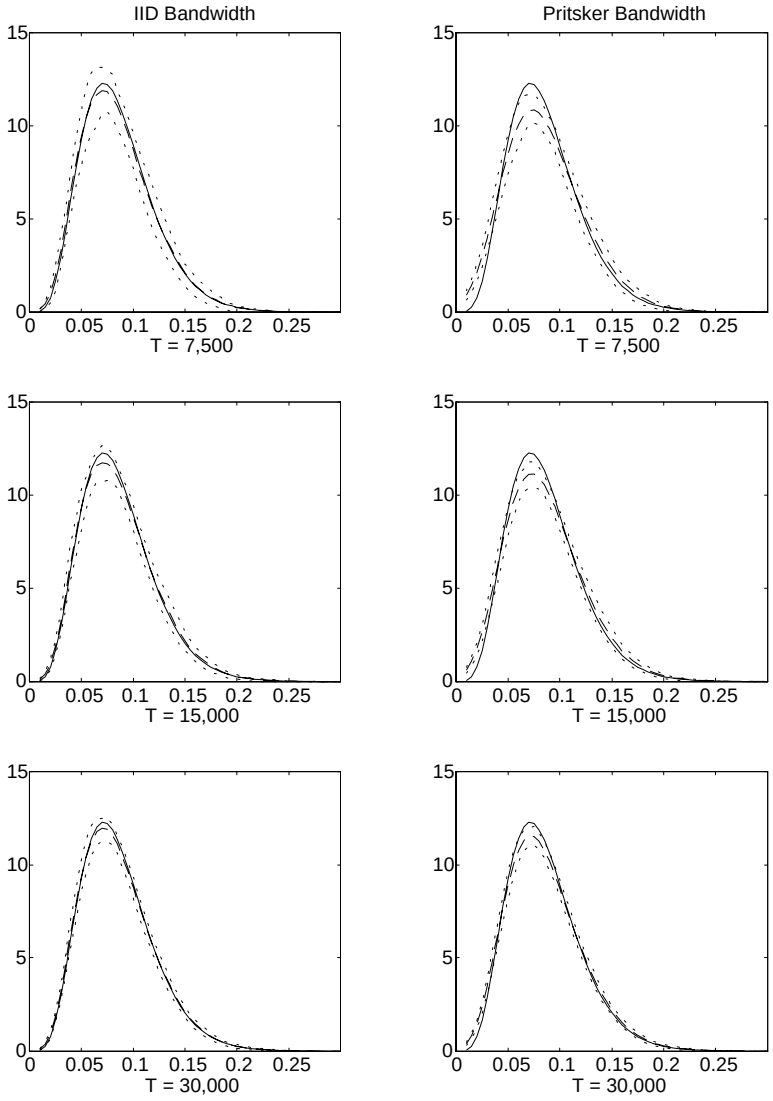

Figure 3: The Stationary Density of the Square-Root Process. $\kappa = 1.71624$, $\theta = 0.085711$, and $\sigma = 0.22143$. The solid line is the true gamma density. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

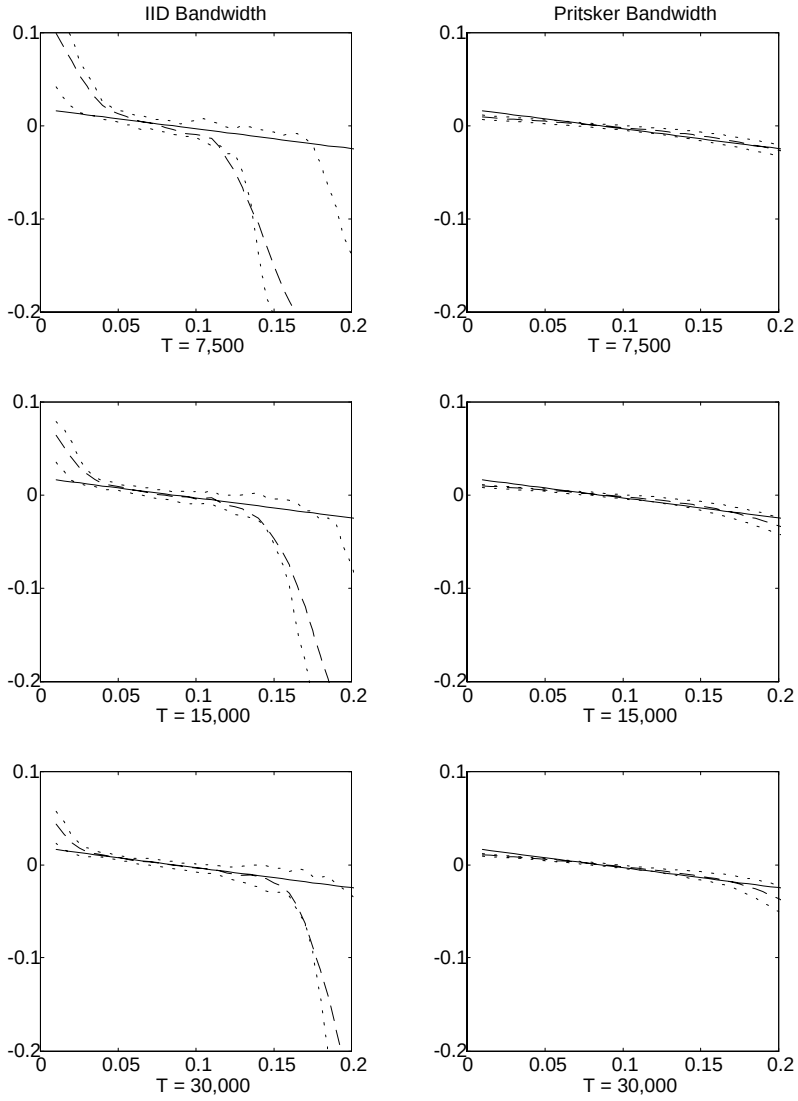

Figure 4: The Drift Function Using the Estimator in Stanton (1997). $\kappa = 0.21459$, $\theta = 0.085711$, and $\sigma = 0.07830$. The solid line is the true drift. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

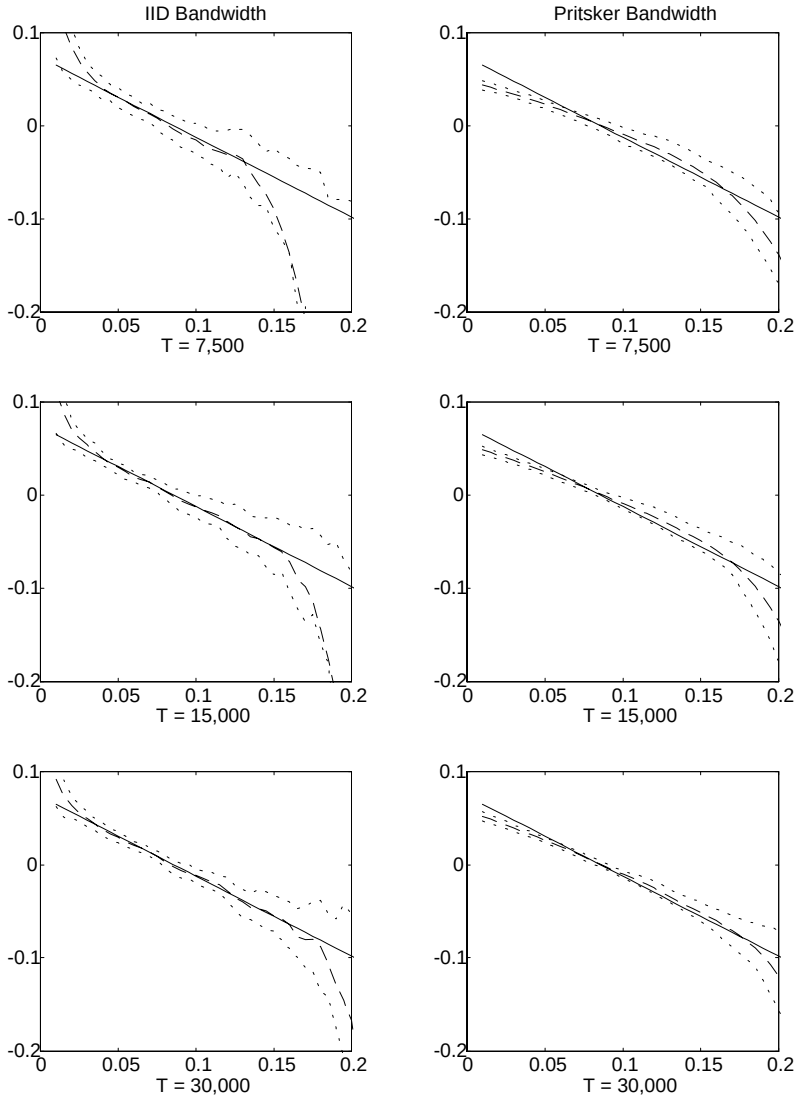

Figure 5: The Drift Function Using the Estimator in Stanton (1997). $\kappa = 0.85837$, $\theta = 0.085711$, and $\sigma = 0.15660$. The solid line is the true drift. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

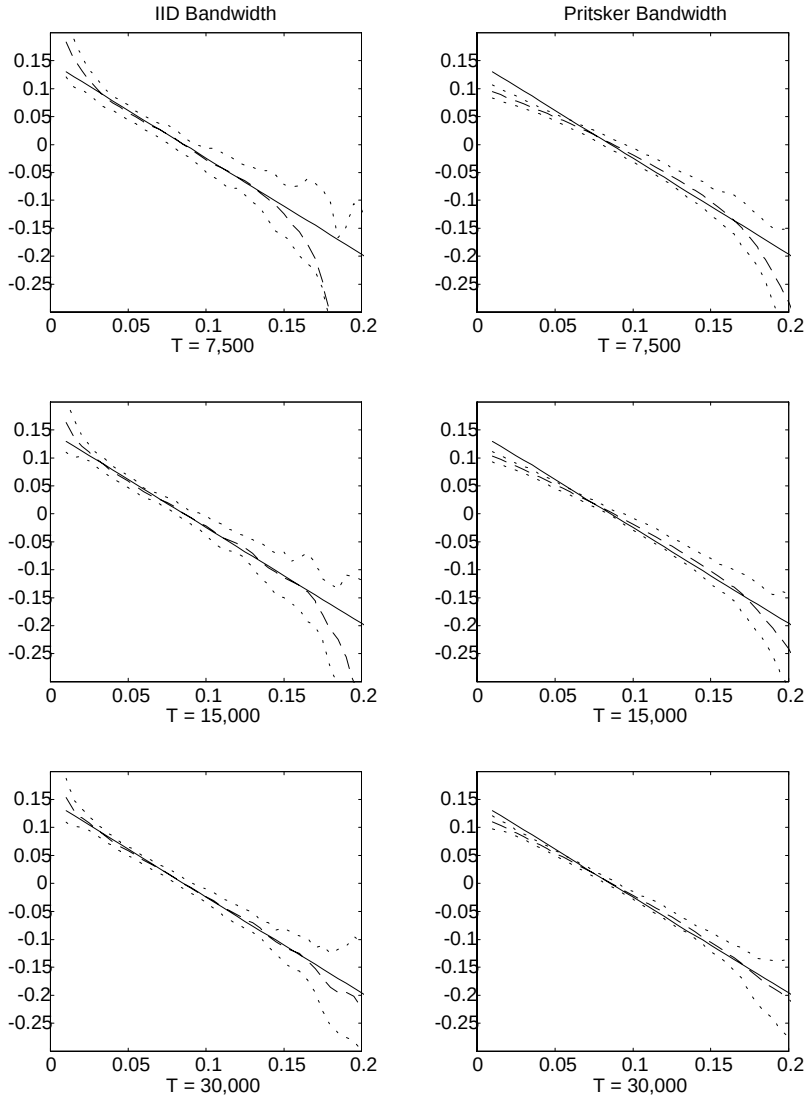

Figure 6: The Drift Function Using the Estimator in Stanton (1997). $\kappa = 1.71624$, $\theta = 0.085711$, and $\sigma = 0.22143$. The solid line is the true drift. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

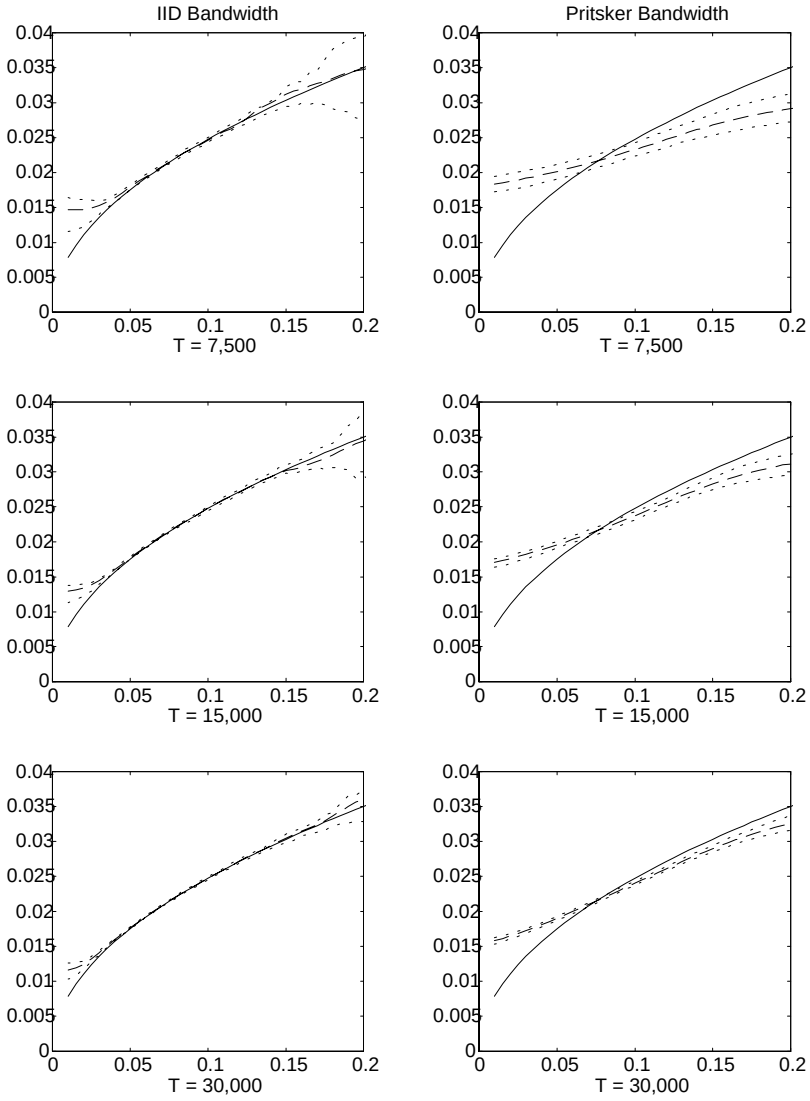

Figure 7: The Diffusion Function Using the Estimator in Stanton (1997). $\kappa = 0.21459$, $\theta = 0.085711$, and $\sigma = 0.07830$. The solid line is the true diffusion. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

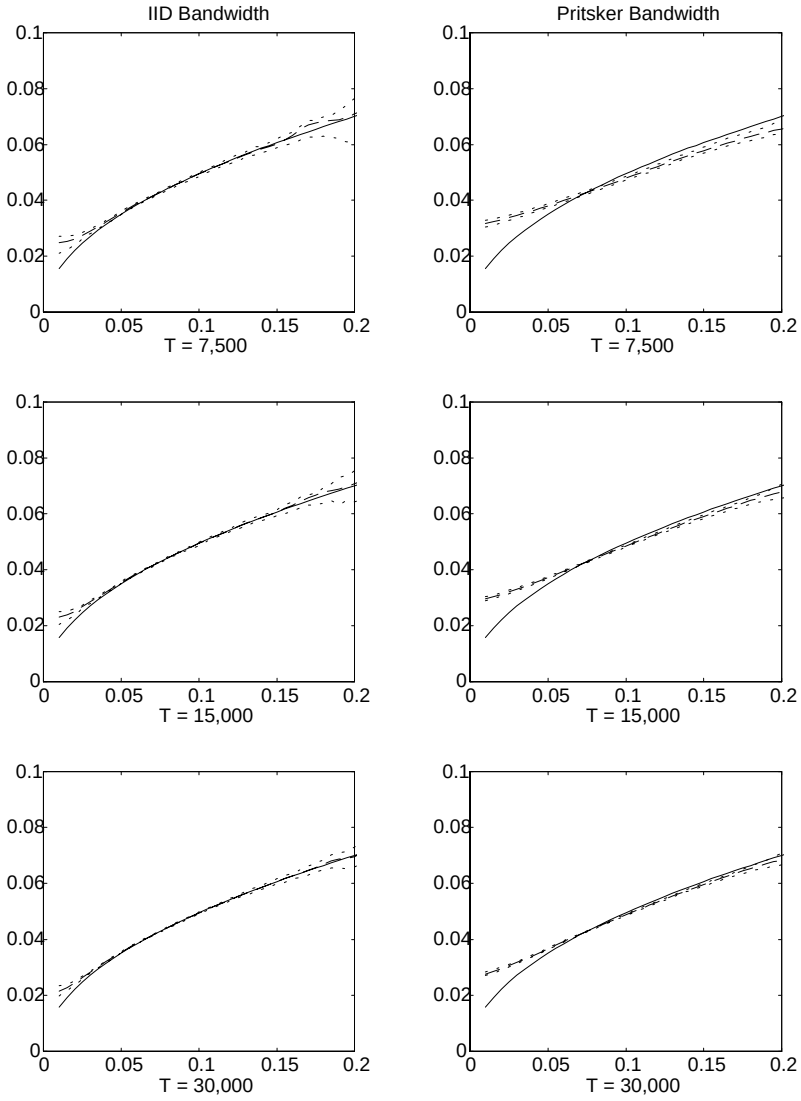

Figure 8: The Diffusion Function Using the Estimator in Stanton (1997). $\kappa = 0.85837$, $\theta = 0.085711$, and $\sigma = 0.15660$. The solid line is the true diffusion. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

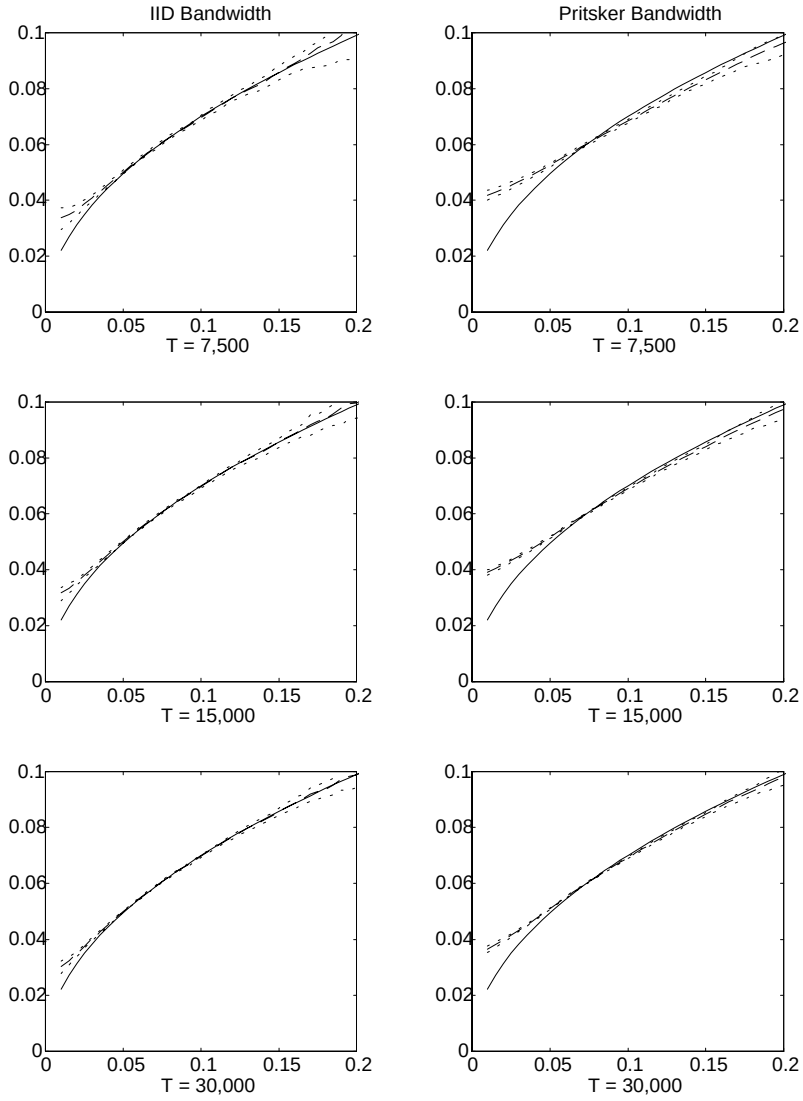

Figure 9: The Diffusion Function Using the Estimator in Stanton (1997). $\kappa = 1.71624$, $\theta = 0.085711$, and $\sigma = 0.22143$. The solid line is the true diffusion. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

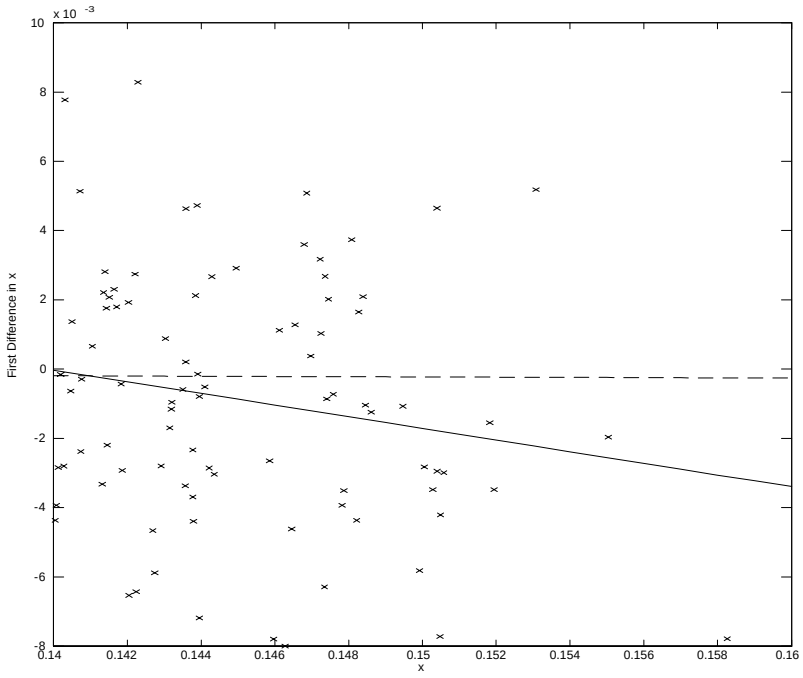

Figure 10: A Plot of $[x_{t+\Delta} - x_t]$ versus x_t in the tail of a sample path. The parameter values are: $\kappa = 0.85837$, $\theta = 0.085711$, $\sigma = 0.15660$, $\Delta = 1/250$, and $T_{sim} = 7,500$. The dashed line is the true regression function and the solid line is the OLS regression through the data points.

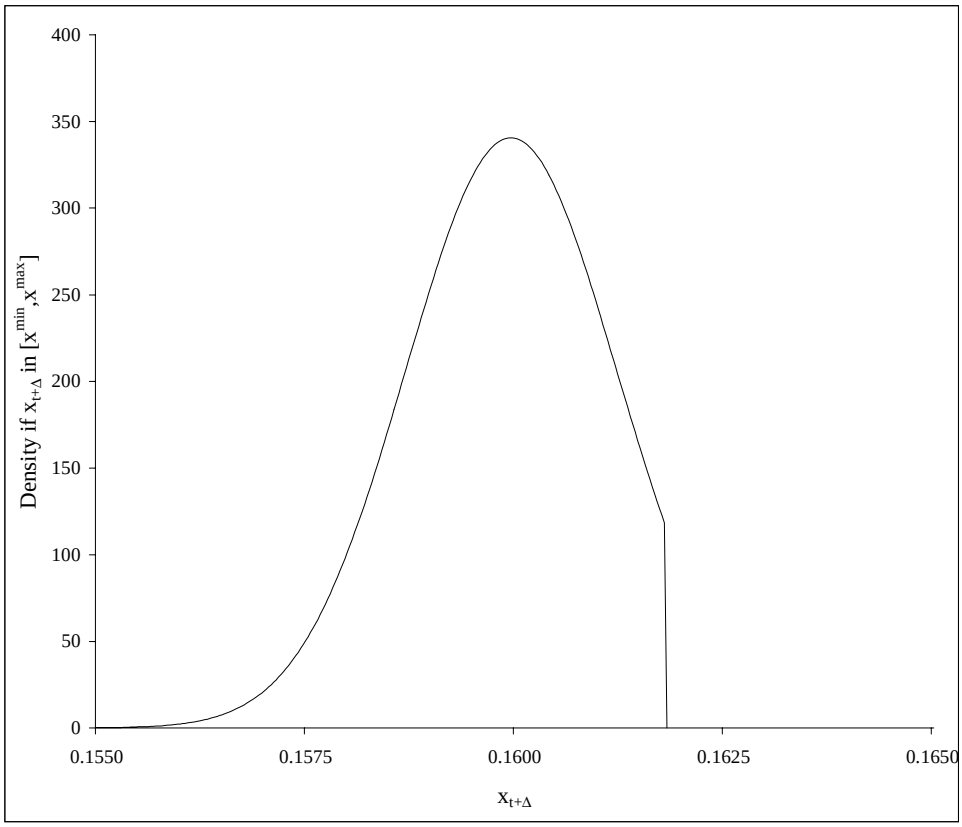

Figure 11: Truncated Conditional Density of the Square-Root Diffusion Process for $x_t = 0.16$, $x^{\min} = 0.013$ and $x^{\max} = 0.1618$.

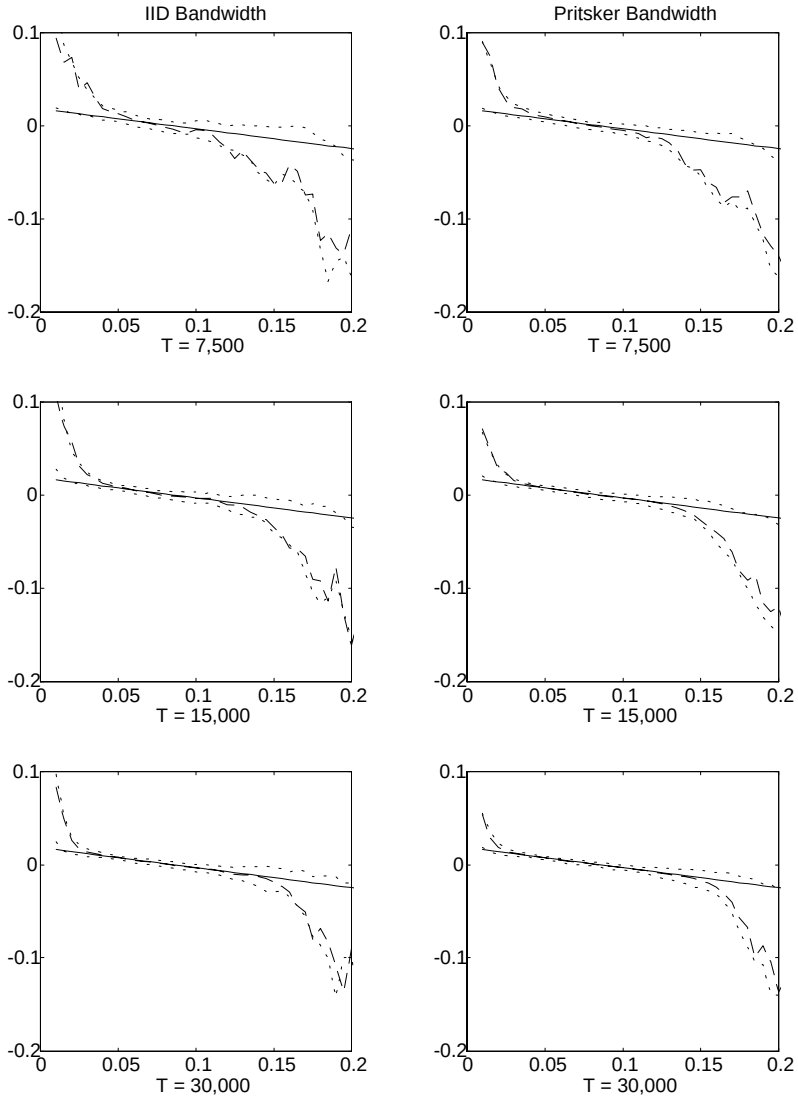

Figure 12: The Drift Function Using the Jackknife Estimator. $\kappa = 0.21459$, $\theta = 0.085711$, and $\sigma = 0.07830$. The solid line is the true drift. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

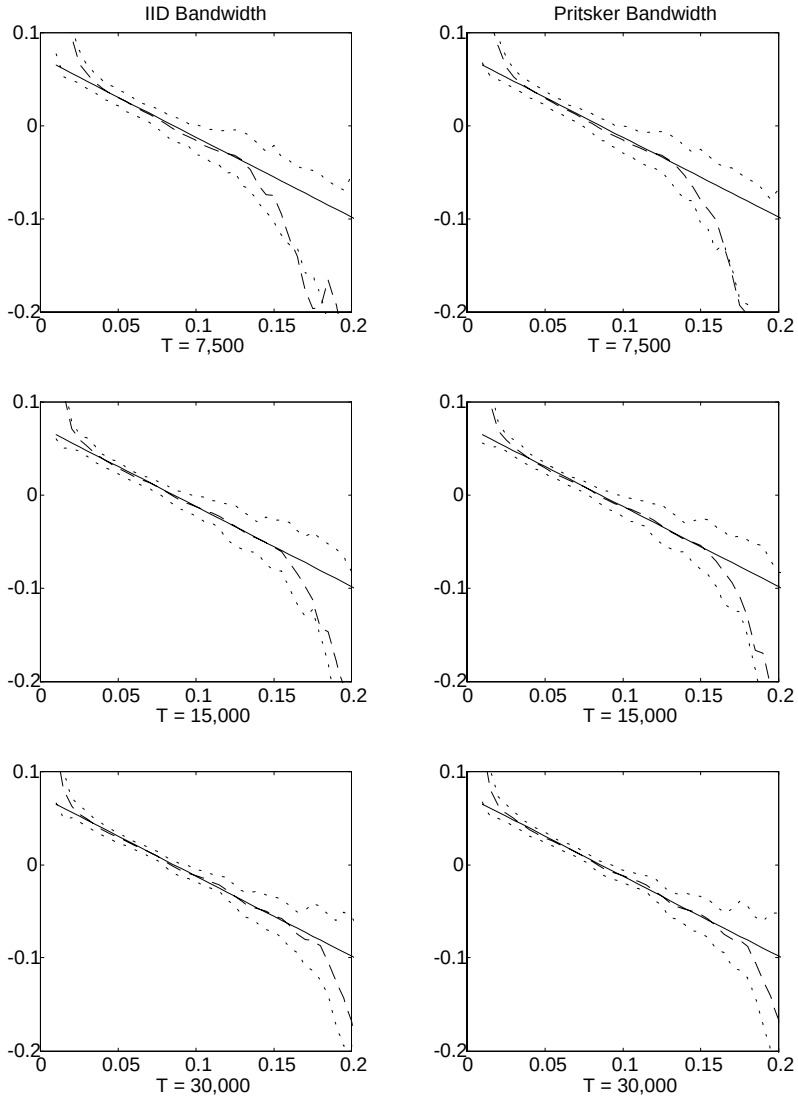

Figure 13: The Drift Function Using the Jackknife Estimator. $\kappa = 0.85837$, $\theta = 0.085711$, and $\sigma = 0.15660$. The solid line is the true drift. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

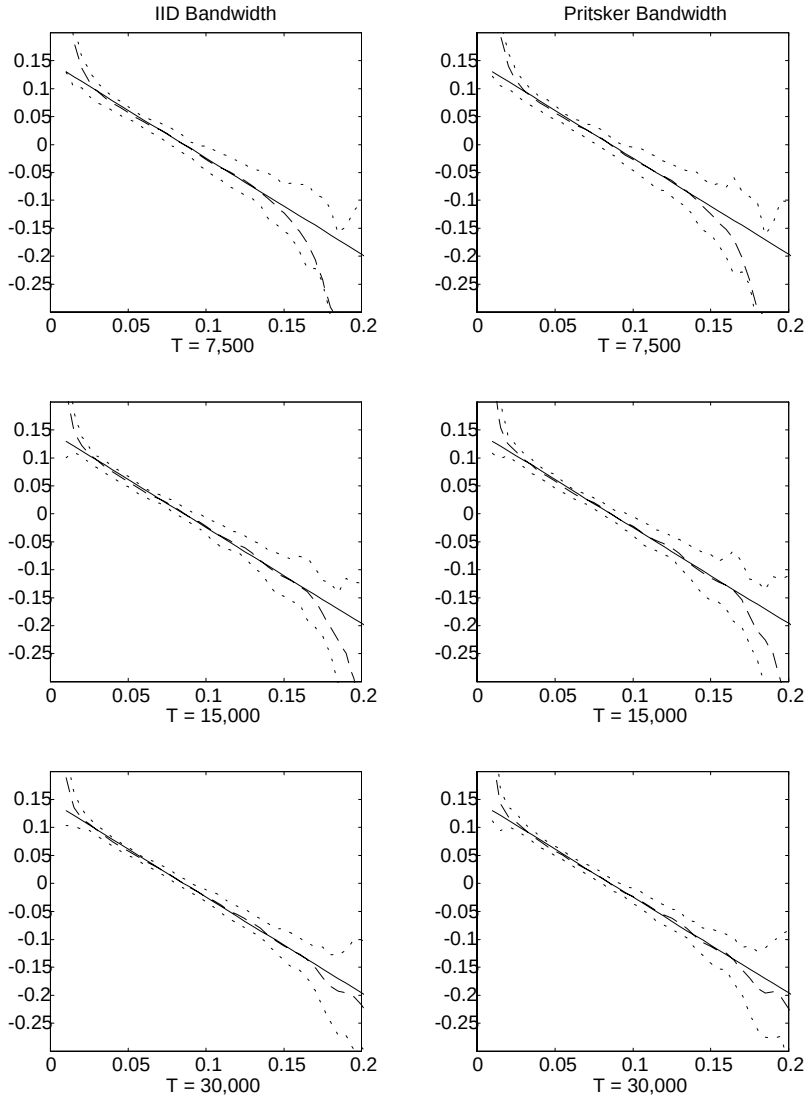

Figure 14: The Drift Function Using the Jackknife Estimator. $\kappa = 1.71624$, $\theta = 0.085711$, and $\sigma = 0.22143$. The solid line is the true drift. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

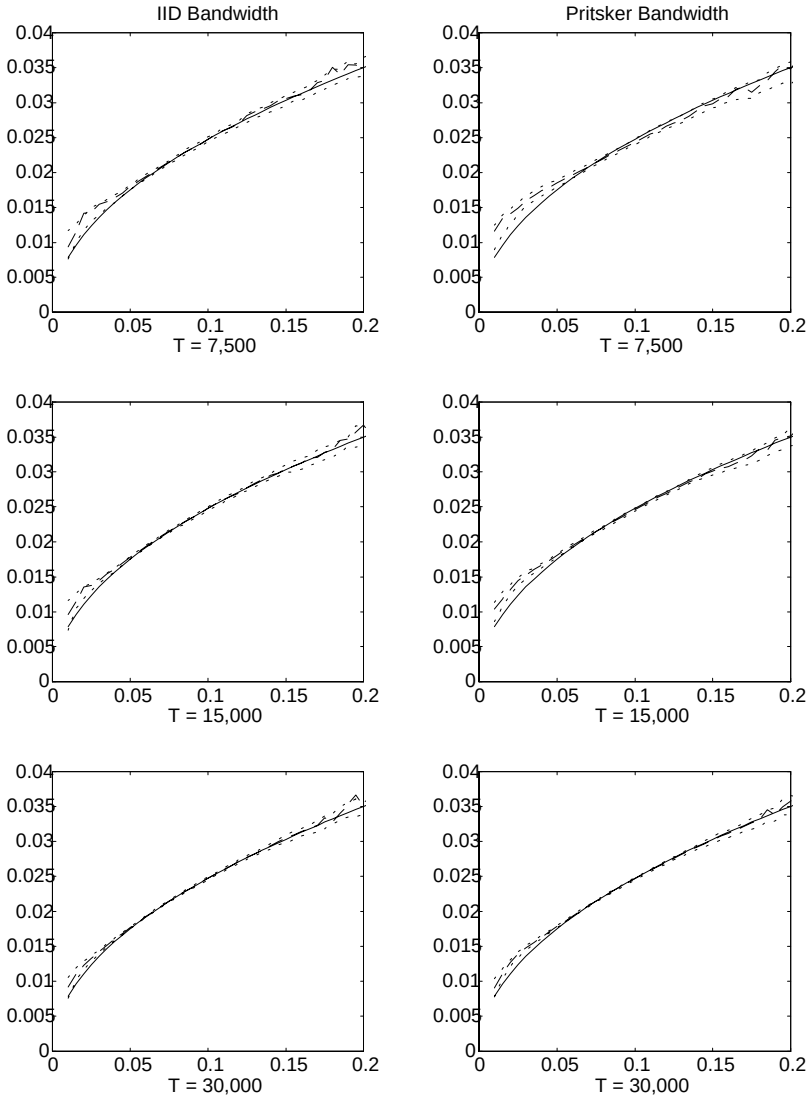

Figure 15: The Diffusion Function Using the Jackknife Estimator. $\kappa = 0.21459$, $\theta = 0.085711$, and $\sigma = 0.07830$. The solid line is the true diffusion. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

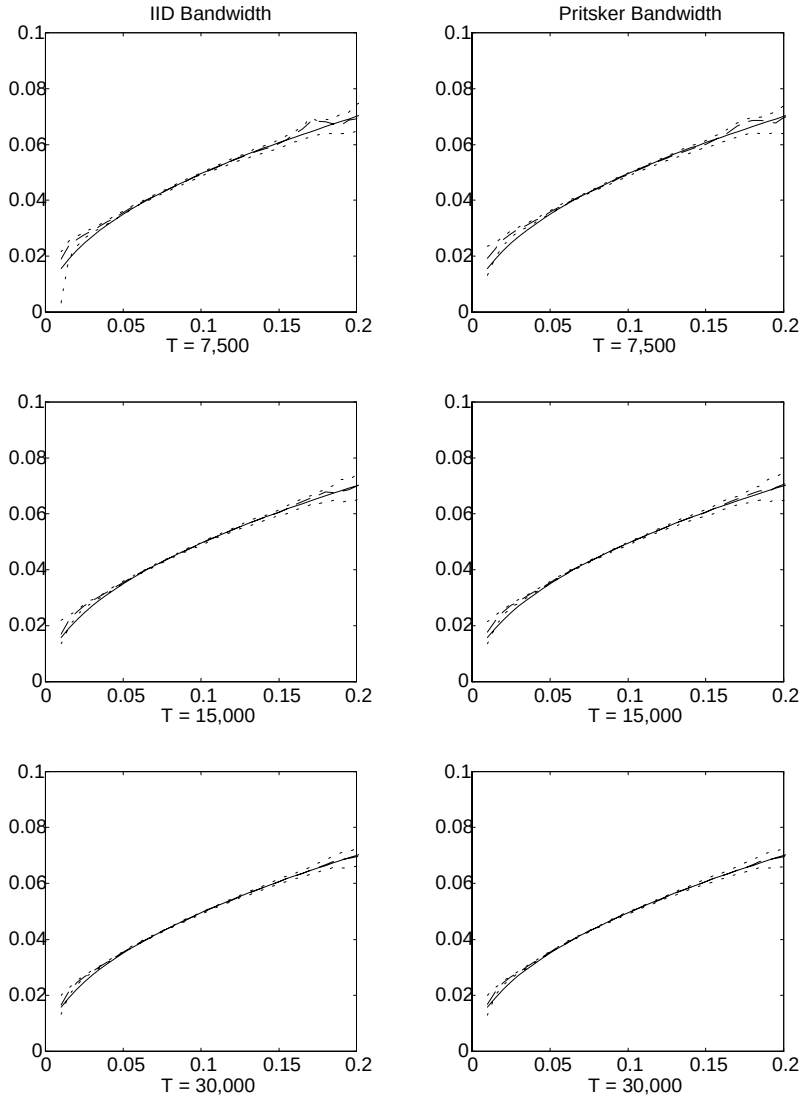

Figure 16: The Diffusion Function Using the Jackknife Estimator. $\kappa = 0.85837$, $\theta = 0.085711$, and $\sigma = 0.15660$. The solid line is the true diffusion. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

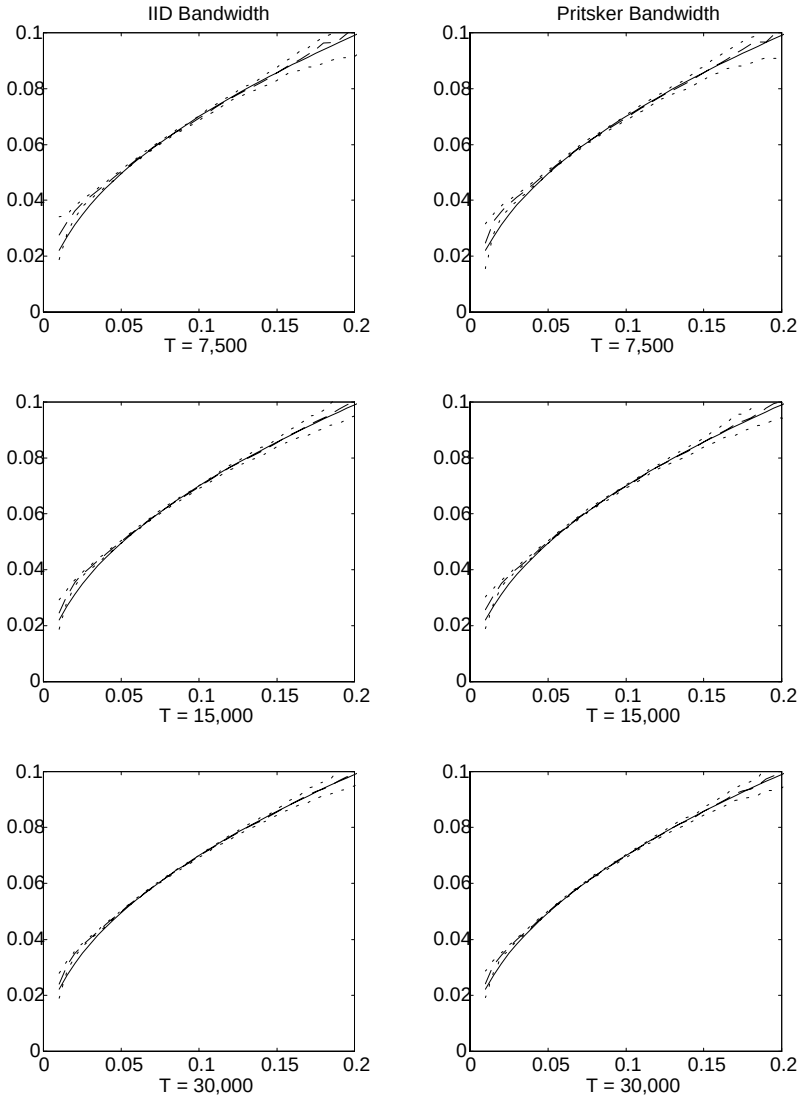

Figure 17: The Diffusion Function Using the Jackknife Estimator. $\kappa = 1.71624$, $\theta = 0.085711$, and $\sigma = 0.22143$. The solid line is the true diffusion. The dashed line is the pointwise average across the 100 simulations, and the dotted lines are the pointwise 25-th and 75-th percentiles.

Table 1: Bandwidth Parameter Choices

Panel A: $\kappa = 0.21459$ and $\sigma = 0.07830$.

Simulation Length	IID Bandwidth	Pritsker Bandwidth
$7,500$	0.0048	0.0299
$15,000$	0.0047	0.0245
$30,000$	0.0043	0.0198

Panel B: $\kappa = 0.85837$ and $\sigma = 0.15660$.

Simulation Length	IID Bandwidth	Pritsker Bandwidth
$7,500$	0.0056	0.0198
$15,000$	0.0051	0.0159
$30,000$	0.0045	0.0128

Panel C: $\kappa = 1.71624$ and $\sigma = 0.22143$.

Simulation Length	IID Bandwidth	Pritsker Bandwidth
$7,500$	0.0057	0.0159
$15,000$	0.0051	0.0128
$30,000$	0.0045	0.0102

For all three panels, $\theta = 0.085711$. The "IID Bandwidth" is the average bandwidth across all 100 simulations, for a given simulation length and combination of parameters, of the data-dependent selection rule $h = \widehat{\sigma}T^{-1/5}$, where $\widehat{\sigma}$ is the sample average standard deviation for the simulated data set. The "Pritsker Bandwidth" are comparable (for the sample sizes reported above) to those Table 2 of Pritsker (1997) as the bandwidth that minimizes the mean integrated squared error of the estimated stationary density (for the case of a Gaussian process) with parameter values and the number of observations. They were generously provided by Matt Pritsker.

Table 2: The Distribution of the Aït-Sahalia (1996) Parameter Estimates: $\kappa = 0.21459$, $\sigma = 0.07830$, and IID Bandwidth.

Panel A: $T_{sim} = 7,500$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0184	0.0029	0.0166	0.0183	0.0192	0.0184
α_1	−0.2402	0.0470	−0.2661	−0.2252	−0.2117	−0.2146
α_2	−0.4410	0.6647	−0.8009	−0.2291	−0.0001	0.0000
α_3	0.0003	0.0004	−0.0000	0.0002	0.0005	0.0000

Panel B: $T_{sim} = 15,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0180	0.0013	0.0175	0.0181	0.0185	0.0184
α_1	−0.2266	0.0251	−0.2386	−0.2196	−0.2109	−0.2146
α_2	−0.1590	0.3410	−0.3886	−0.0954	0.0643	0.0000
α_3	0.0001	0.0002	−0.0000	0.0001	0.0003	0.0000

Panel C: $T_{sim} = 30,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0183	0.0008	0.0181	0.0183	0.0187	0.0184
α_1	−0.2209	0.0172	−0.2273	−0.2159	−0.2117	−0.2146
α_2	−0.1011	0.2474	−0.2399	−0.0597	0.0498	0.0000
α_3	0.0001	0.0002	−0.0000	0.0000	0.0001	0.0000

The form of the estimated drift function is:

$$\widehat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$$

while the true drift function is $\mu(x) = \kappa(\theta - x)$. Estimation for each sample path is started at the true parameter values. All statistics are computed using the 100 simulated sample paths of a square root diffusion.

Table 3: The Distribution of the Aït-Sahalia (1996) Parameter Estimates: $\kappa = 0.21459$, $\sigma = 0.07830$, and Pritsker Bandwidth.

Panel A: $T_{sim} = 7,500$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0189	0.0013	0.0184	0.0188	0.0192	0.0184
α_1	−0.2035	0.0101	−0.2113	−0.2044	−0.1971	−0.2146
α_2	0.1677	0.1260	0.0829	0.1416	0.2412	0.0000
α_3	−0.0001	0.0002	−0.0002	−0.0002	−0.0000	0.0000

Panel B: $T_{sim} = 15,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0188	0.0010	0.0183	0.0187	0.0193	0.0184
α_1	−0.2045	0.0108	−0.2121	−0.2046	−0.1975	−0.2146
α_2	0.1472	0.1140	0.0662	0.1340	0.2164	0.0000
α_3	−0.0001	0.0002	−0.0002	−0.0001	−0.0000	0.0000

Panel C: $T_{sim} = 30,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0188	0.0008	0.0184	0.0187	0.0190	0.0184
α_1	−0.2057	0.0095	−0.2128	−0.2059	−0.2018	−0.2146
α_2	0.0989	0.0966	0.0533	0.1047	0.1611	0.0000
α_3	−0.0001	0.0001	−0.0001	−0.0001	−0.0000	0.0000

The form of the estimated drift function is:

$$\widehat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$$

while the true drift function is $\mu(x) = \kappa(\theta - x)$. Estimation for each sample path is started at the true parameter values. All statistics are computed using the 100 simulated sample paths of a square root diffusion.

Table 4: The Distribution of the Aït-Sahalia (1996) Parameter Estimates: $\kappa = 0.85837$, $\sigma = 0.15660$, and IID Bandwidth.

Panel A: $T_{sim} = 7,500$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0733	0.0034	0.0717	0.0734	0.0749	0.0736
α_1	−0.8782	0.0632	−0.9164	−0.8659	−0.8425	−0.8584
α_2	−0.3749	0.8634	−0.9324	−0.3391	0.3266	0.0000
α_3	0.0003	0.0008	−0.0003	0.0001	0.0007	0.0000

Panel B: $T_{sim} = 15,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0735	0.0026	0.0724	0.0735	0.0749	0.0736
α_1	−0.8605	0.0450	−0.8798	−0.8569	−0.8323	−0.8584
α_2	−0.0909	0.6843	−0.4903	−0.0011	0.4053	0.0000
α_3	0.0000	0.0004	−0.0003	−0.0000	0.0002	0.0000

Panel C: $T_{sim} = 30,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0736	0.0011	0.0732	0.0737	0.0742	0.0736
α_1	−0.8577	0.0268	−0.8718	−0.8576	−0.8431	−0.8584
α_2	−0.0385	0.4179	−0.3281	−0.0155	0.2478	0.0000
α_3	−0.0000	0.0003	−0.0002	−0.0001	0.0001	0.0000

The form of the estimated drift function is:

$$\widehat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$$

while the true drift function is $\mu(x) = \kappa(\theta - x)$. Estimation for each sample path is started at the true parameter values. All statistics are computed using the 100 simulated sample paths of a square root diffusion.

Table 5: The Distribution of the Aït-Sahalia (1996) Parameter Estimates: $\kappa = 0.85837$, $\sigma = 0.15660$, and Pritsker Bandwidth.

Panel A: $T_{sim} = 7,500$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0745	0.0022	0.0733	0.0740	0.0766	0.0736
α_1	−0.8308	0.0300	−0.8555	−0.8342	−0.8096	−0.8584
α_2	0.4023	0.4265	0.1247	0.3644	0.7115	0.0000
α_3	−0.0002	0.0005	−0.0005	−0.0004	−0.0000	0.0000

Panel B: $T_{sim} = 15,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0746	0.0022	0.0735	0.0745	0.0761	0.0736
α_1	−0.8316	0.0287	−0.8556	−0.8327	−0.8082	−0.8584
α_2	0.3348	0.3754	0.1427	0.3784	0.6189	0.0000
α_3	−0.0003	0.0004	−0.0005	−0.0003	−0.0001	0.0000

Panel C: $T_{sim} = 30,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.0742	0.0018	0.0733	0.0740	0.0748	0.0736
α_1	−0.8366	0.0219	−0.8544	−0.8378	−0.8216	−0.8584
α_2	0.2313	0.2809	0.0823	0.2469	0.4169	0.0000
α_3	−0.0002	0.0002	−0.0004	−0.0002	−0.0001	0.0000

The form of the estimated drift function is:

$$\widehat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$$

while the true drift function is $\mu(x) = \kappa(\theta - x)$. Estimation for each sample path is started at the true parameter values. All statistics are computed using the 100 simulated sample paths of a square root diffusion.

Table 6: The Distribution of the Aït-Sahalia (1996) Parameter Estimates: $\kappa = 1.71624$, $\sigma = 0.22143$, and IID Bandwidth.

Panel A: $T_{sim} = 7,500$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.1473	0.0052	0.1455	0.1470	0.1483	0.1471
α_1	−1.7252	0.0855	−1.7763	−1.7160	−1.6843	−1.7162
α_2	−0.2051	1.2058	−0.8552	−0.0081	0.5295	0.0000
α_3	0.0001	0.0010	−0.0005	−0.0000	0.0006	0.0000

Panel B: $T_{sim} = 15,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.1468	0.0031	0.1450	0.1470	0.1482	0.1471
α_1	−1.7100	0.0584	−1.7383	−1.7075	−1.6715	−1.7162
α_2	0.1041	0.8522	−0.4080	0.3000	0.7522	0.0000
α_3	−0.0001	0.0006	−0.0005	−0.0001	0.0003	0.0000

Panel C: $T_{sim} = 30,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.1470	0.0027	0.1463	0.1472	0.1479	0.1471
α_1	−1.7100	0.0408	−1.7275	−1.7130	−1.6858	−1.7162
α_2	0.0977	0.5919	−0.2842	0.1157	0.5170	0.0000
α_3	−0.0001	0.0004	−0.0004	−0.0001	0.0001	0.0000

The form of the estimated drift function is:

$$\widehat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$$

while the true drift function is $\mu(x) = \kappa(\theta - x)$. Estimation for each sample path is started at the true parameter values. All statistics are computed using the 100 simulated sample paths of a square root diffusion.

Table 7: The Distribution of the Aït-Sahalia (1996) Parameter Estimates: $\kappa = 1.71624$, $\sigma = 0.22143$, and Pritsker Bandwidth.

Panel A: $T_{sim} = 7,500$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.1486	0.0049	0.1466	0.1480	0.1509	0.1471
α_1	−1.6624	0.0630	−1.7110	−1.6635	−1.6185	−1.7162
α_2	0.6496	0.7654	0.2405	0.7211	1.1033	0.0000
α_3	−0.0005	0.0008	−0.0010	−0.0006	−0.0001	0.0000

Panel B: $T_{sim} = 15,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.1493	0.0035	0.1470	0.1484	0.1513	0.1471
α_1	−1.6713	0.0469	−1.7063	−1.6650	−1.6330	−1.7162
α_2	0.5209	0.6306	0.1543	0.6056	0.9284	0.0000
α_3	−0.0005	0.0006	−0.0009	−0.0005	−0.0001	0.0000

Panel C: $T_{sim} = 30,000$.

			Percentiles			
	Mean	Std. Dev.	25^{th}	50^{th}	75^{th}	True Value
α_0	0.1475	0.0029	0.1463	0.1473	0.1487	0.1471
α_1	−1.6841	0.0379	−1.7152	−1.6870	−1.6628	−1.7162
α_2	0.3987	0.4341	0.1679	0.4163	0.6847	0.0000
α_3	−0.0003	0.0004	−0.0006	−0.0004	−0.0001	0.0000

The form of the estimated drift function is:

$$\widehat{\mu}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^{-1}$$

while the true drift function is $\mu(x) = \kappa(\theta - x)$. Estimation for each sample path is started at the true parameter values. All statistics are computed using the 100 simulated sample paths of a square root diffusion.

Table 8: GMM Estimates of the Nonlinear Drift Model

Parameters	Stanton (1997) T-Bill Data	Aït-Sahalia (1996) Eurodollar Data
α_0	0.0011	0.0039
	(0.0007)	(0.0025)
α_1	-0.0164	-0.0588
	(0.0097)	(0.0319)
α_2	0.0746	0.2596
	(0.0395)	(0.1094)
α_3	-0.0000	-0.0001
	(0.0000)	(0.0001)
σ^2	0.0054	0.0142
	(0.0024)	(0.0117)
γ	1.6234	1.4550
	(0.1024)	(0.1925)
Sample Size:	Jan 1965 – July 1995 $T = 7975$	June 1973 – Feb 1995 $T = 5505$

The parameters are estimated from

$$x_{t+1} - x_t = \alpha_0 + \alpha_1 x_t + \alpha_2 x_t^2 + \alpha_3 x_t^{-1} + \varepsilon_{t+1}$$

and

$$E\left(\varepsilon_{t+1}\right) = 0 \qquad \text{and} \qquad E\left(\varepsilon_{t+1}^2\right) = \sigma^2 x_t^{2\gamma}$$

using a weighting matrix based on a Bartlett kernel with 60 time lags. Standard errors are reported in parentheses.

References

[1] Aït-Sahalia, Y., 1996, "Testing Continuous-Time Models of the Spot Interest Rate," *The Review of Financial Studies* 9, pages 385-426.

[2] Arnold, L., 1974, *Stochastic Differential Equations: Theory and Applications.* (New York: John Wiley & Sons).

[3] Black, F., E. Derman, and W. Toy, 1990, "A One-Factor Model of Interest Rates and Its Application to Treasury Bond Options," *Financial Analysts Journal*, pages 33-39.

[4] Black, F., and P. Karasinski, 1991, "A One-Factor Model of Interest Rates and Its Application to Treasury Bond Options," *Financial Analysts Journal*, pages 52-59.

[5] Boudoukh, J., R. F. Whitelaw, M. Richardson, and R. Stanton, 1997, "Pricing Mortgage-Backed Securities in a Multifactor Interest Rate Environment: A Multivariate Density Estimation Approach," *The Review of Financial Studies* 10, pages 405-446.

[6] Brennan, M. J., and E. S. Schwartz, 1979, "A Continuous-Time Approach to the Pricing of Bonds," *Journal of Banking and Finance* 3, pages 133-155.

[7] Chan, K. C., G. A. Karolyi, F. A. Longstaff, and A. B. Sanders, 1992, "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate," *Journal of Finance* 47, pages 1209-1227.

[8] Conley, T. G., L. P. Hansen, E. G. J. Luttmer, and J. A. Scheinkman, 1997, "Short-Term Interest Rates as Subordinated Diffusions," *The Review of Financial Studies* 10, pages 525-577.

[9] Courtadon, G., 1982, "The Pricing of Options on Default-Free Bonds," *Journal of Financial and Quantitative Analysis* 17, pages 75-100.

[10] Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross, 1985, "A Theory of the Term Structure of Interest Rates," *Econometrica* 53, pages 321-346.

[11] Duffie, D., and P. Glynn, 1996, "Estimation of Continuous-Time Markov Processes Sampled at Random Time Intervals," Mimeo, Graduate School of Business, Stanford University.

[12] Duffie, D., and K. J. Singleton, 1993, "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica* 61, pages 929-952.

[13] Feller, W., 1951, "Two Singular Diffusion Problems," *Annals of Mathematics* 54, pages 173-182.

[14] Hansen, L. P., 1982, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50, pages 1029-1054.

[15] Hansen, L. P., and J. A. Scheinkman, 1995, "Back to the Future: Generating Moment Implications for Continuous-Time Markov Processes," *Econometrica* 63, pages 767-804.

[16] Härdle, W., 1990, *Applied Nonparametric Regression.* (Cambridge: Cambridge University Press).

[17] Karlin, S., and H. M. Taylor, 1981, *A Second Course in Stochastic Processes.* (San Diego: Academic Press, Inc.)

[18] Lo, A. W., 1988, "Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data," *Econometric Theory* 4, pages 231-247.

[19] Merton, R. C., 1980, "On Estimating the Expected Return on the Market: An Exploratory Investigation," *Journal of Financial Economics* 8, pages 323-361.

[20] Moré, J. J., 1977, "The Levenberg-Marquardt Algorithm: Implementation and Theory," pages 105-116 in *Numerical Analysis*, G. A. Watson (editor), *Lecture Notes in Mathematics 630*, Springer-Verlag.

[21] Newey, W. K., and K. D. West, 1994, "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies* 61, pages 631-653.

[22] Pearson, N. D., and T. Sun, 1994, "Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll, and Ross Model," *Journal of Finance* 49, pages 1279-1304.

[23] Pritsker, M., 1997, "Nonparametric Density Estimation of Tests of Continuous Time Interest Rate Models," forthcoming in *The Review of Financial Studies.*

[24] Rice, J. A., 1984, "Boundary Modification for Kernel Regression," *Communications in Statistics, Theory and Methods* 13(7), pages 893-900.

[25] Robinson, P. M., 1983, "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis* 4, pages 185-207.

[26] Robinson, P. 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, pages 539-549.

[27] Santa-Clara, P., 1995, "Simulated Likelihood Estimation of Diffusions With An Application to the Short Term Interest Rate," Mimeo, INSEAD.

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

[29] Stanton, R., 1997, "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *Journal of Finance* 52, pages 1973-2002.