Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models

Matt Pritsker
Federal Reserve Board

I study the finite sample distribution of one of Aït-Sahalia's (1996c) nonparametric tests of continuous-time models of the short-term riskless rate. The test rejects true models too often because interest rate data are highly persistent but the asymptotic distribution of the test (and of the kernel density estimator on which the test is based) treats the data as if it were independently and identically distributed. To attain the accuracy of the kernel density estimator implied by its asymptotic distribution with 22 years of data generated from the Vasicek model in fact requires 2755 years of data.

Many continuous-time models in finance specify a stochastic process for the short rate of interest and then use the dynamics of the process to derive the term structure of interest rates and to price derivative securities. Because of the short rate's importance, a literature that estimates these dynamics-empirically has developed. The most frequently-used estimation techniques have been maximum likelihood and generalized method of moments (GMM).[1] However, both of these approaches have limitations. Maximum likelihood estimation is restrictive because few short-rate models have known likelihood functions. GMM tech-

---

The author thanks Yacine Aït-Sahalia, Torben Andersen, Mark Fisher, and Young Ho Eom for useful discussions, thanks Ruth Wu and Amy Ashton for valuable research assistance, and thanks Ellen Dykes for editorial assistance. All errors are the author's. The views expressed in this article are those of the author and not those of the Board of Governors of the Federal Reserve System or other members of its staff. Address correspondence to Matt Pritsker, The Federal Reserve Board, Mail Stop 91, Washington, DC 20551, or e-mail: mpritsker@frb.gov.

[1] For examples of maximum likelihood estimation see Chen and Scott (1993) or Pearson and Sun (1994). For examples of GMM estimation see Chan et al. (1992).

The Review of Financial Studies Fall 1998 Vol. 11, No. 3, pp. 449–487

niques are more general, but generate mixed results when applied to interest rate data because the parameter estimates are highly sensitive to the choice of moment conditions and may be of low quality if these moments are estimated imprecisely [Eom (1998)].

Recently new approaches for estimation and hypothesis testing have been developed. The new approaches use either semiparametric [Gallant and Tauchen (1995); Andersen and Lund (1997)] methods to improve the moment conditions for parameter identification with a technique known as efficient method of moments (EMM) or they identify the parameters using features of the data that were estimated using nonparametric methods.

This article focuses on the recent finance literature that has identified the short-rate process using nonparametric methods [Siddique (1994), Aït-Sahalia (1996a, b, c), Stanton (1997)]. All of these methods estimate the marginal (or unconditional) distribution of the short rate using a kernel density estimator and then use the result either alone or as part of other non-parametric procedures to identify the short-rate process.[2] The most striking result using these methods is contained in Aït-Sahalia (1996c). Aït-Sahalia constructed a test of whether any standard one-factor model of the short rate could generate data that are consistent with a nonparametric estimate of the short rate's unconditional distribution. Using the asymptotic distribution of the test and 22 years of daily U.S. short-rate data, he strongly rejected virtually every well-known one-factor diffusion model of the short-rate.

The strength of the rejections using Aït-Sahalia's test appears to show that nonparametric estimates are very powerful for distinguishing among short-rate models. However, there is also strong reason for doubt because the time series of U.S. interest rates is dependent and highly persistent, but the asymptotic distribution of the test statistic has the unusual property that it treats the observations as if they were independently and identically distributed (i.i.d.).[3] By not accounting for the effects of dependence in finite samples, the asymptotic distribution probably understates the variance of the test statistic. This understatement could cause the test to reject true interest rate models too often.

This property of Aït-Sahalia's test—that it treats observations as if i.i.d.—is not shared by tests based on standard estimation methods such as maximum likelihood, GMM, or EMM, but is shared by some tests that use nonparametric kernel density estimates.[4] For this class of tests, because the

[2] For example, Siddique and Stanton both use a kernel estimate of the unconditional distribution of the short rate to perform nonparametric kernel regressions.

[3] By "highly persistent" I mean that the time series of interest rates is strongly positively correlated through time. Using monthly data from McCulloch and Kwon (1993), Pagan, Hall, and Martin (1995) estimate that 1-month zero coupon rates have a first-order autocorrelation coefficient of 0.98.

[4] Maximum likelihood methods implicitly account for time-series dependence in their specification of the likelihood function. Moment-based estimators such as EMM and GMM account for dependence in their standard error calculations [Newey and West (1987)]. Nevertheless, for very persistent time series such

nonparametric kernel density estimator converges more slowly than time-series dependence decays, the asymptotic distribution of the test ignores time series dependence. The slow rate of convergence of the kernel density estimator is an additional reason to be concerned about the accuracy of tests based on its asymptotic distribution in small samples.[5]

To investigate the finite sample performance of Aït-Sahalia's test, I studied the test's size and power when short-rates are generated by the Vasicek (1977) model (Section 1 of this article). My principal result is that the test rejects true models much too often when asymptotic critical values are used. After the critical values are adjusted for size, the test has low power in distinguishing between the Vasicek and Cox-Ingersoll-Ross models compared with the power of a conditional moment-based specification test (Section 2).

To investigate why Aït-Sahalia's test performs poorly in finite samples, in section 3 I used the Vasicek model to contrast the finite sample distribution of the kernel density estimate with the asymptotic distribution on which the test is based. I found that the asymptotic distribution of the kernel density estimator severely understates its finite sample bias, variance, and covariance. The most dramatic finding is that to achieve the estimator precision implied by the asymptotic distribution with 22 years of daily data actually requires more than 2750 years of data. In addition, section 3 contains interesting results on the way that the quality of the kernel density estimate is affected by the persistence of the interest rate process, the number of years of data used in estimation, and the frequency with which the data is sampled.

The practical implications of my results for the estimation of interest rate models are discussed in section 4. I conclude by suggesting directions for further research.

1. Aït-Sahalia's Test

Suppose the instantaneous short rate of interest follows a stationary Markov diffusion process of form

$$dr = \mu(r, \beta)dt + \sigma(r, \beta)dw, \qquad (1)$$

where the vector $\beta \in \mathcal{B}$ represents the parameters of the short-rate process for some parametric model of the short rate, and $\mathcal{B}$ is the set of admissible parameters for the model.

as U.S. interest rates, persistence has been shown to cause small-sample bias in parametric estimates of short-rate processes [Ball and Torous (1996)] and in regression-based tests of the expectations hypothesis of the term structure of interest rates [Bekaert, Hodrick, and Marshall (1997)].

[5] The nonparametric kernel density estimator converges more slowly than estimators based on methods such as GMM, EMM, or maximum likelihood. Castellana and Leadbetter (1986) and Robinson (1986) provide details on the rate of convergence of kernel density estimators. Aït-Sahalia (1996b) discusses the rates of convergence of functions of kernel density estimators.

The process for r in Equation (1) can equivalently be described by its probability transitions. For times t and s, when $t < s$, the transition density $\pi(r_s, s \mid r_t, t, \beta)$ denotes the probability density that r at time s is equal to r_s given r at time t is equal to r_t. The unconditional probability density that r is equal to u is denoted by the marginal density $\pi(u, \beta)$. The marginal density function and transition density function are denoted by $\pi(., \beta)$, and $\pi(., . \mid ., ., \beta)$, respectively.

The two densities contain different information about the process. The transition density describes the short-run time-series behavior of the process because it shows how r at time s depends on r at time t when the time between the observations is finite. The marginal density describes the long-run behavior of the process because it is the limit of the transition density as the time between observations goes to infinity.[6]

Aït-Sahalia (1996c) tests the null hypothesis that the instantaneous short rate follows a single-factor parametric diffusion with a form as given in Equation (1) against the alternative that r is generated by a more general single-factor diffusion of form

$$dr = \mu(r)dt + \sigma(r)dw. \tag{2}$$

The identifying information in the test is the marginal density of r. The null hypothesis restricts r's marginal density to have functional form $\pi(., \beta)$ where $\beta \in \mathcal{B}$, whereas the marginal density under the alternative has the more general functional form $\pi(.)$. Aït-Sahalia uses this restriction to test the null hypothesis by measuring the distance between two estimates of $\pi(.)$. The first, $\hat{\pi}(.)$, is a nonparametric estimate of $\pi(.)$ that is consistent over a wide class of continuous-time models. The second, $\pi(., \hat{\beta})$, is the closest density to $\hat{\pi}(.)$ that could have been generated under the parametric null.[7] The distance measure used in the test is the integrated squared difference between the two density estimates weighted by $\pi(.)$:

$$M = \int_u [\pi(u, \beta) - \pi(u)]^2 \pi(u)\, du. \tag{3}$$

A normalized estimate of this distance is

$$\hat{M} = N h_N \int_{\underline{x}}^{\overline{x}} [\pi(u, \hat{\beta}) - \hat{\pi}(u)]^2 \hat{\pi}(u)\, du, \tag{4}$$

[6] Formally,

$$\pi(u, \beta) = \lim_{s \to \infty} \pi(u, s \mid r_t, t, \beta).$$

$\pi(u, \beta)$ does not depend on r_t because the dependence between observations goes to zero as the time between them approaches infinity.

[7] Aït-Sahalia also introduced a test that uses identifying information contained in this transitions of the interest rate process. I have not examined the size or the power of this transition-based test.

where N is the number of interest rate observations in the sample, h_N is the smoothing bandwidth used in the nonparametic estimator $\hat{\pi}(.)$, $\underline{r}$ and $\overline{r}$ are the highest and lowest realizations of r in the data, and the parameter vector $\hat{\beta}$ is chosen to minimize the distance measure $\hat{M}$.[8]

Under the null hypothesis, $\hat{\pi}(.)$ and $\pi(., \hat{\beta})$ are consistent estimators of $\pi(.)$, and thus their distance will shrink to zero as the sample size grows. If the null hypothesis is false, the distance will not shrink to zero because the nonparametric density estimate will converge to a density outside the parametric class. The failure of the distance to shrink with sample size will cause the null hypothesis to be rejected in samples that are large enough.

Before presenting the asymptotic distribution of Aït-Sahalia's test, I review the nonparametric kernel density estimator that is used to estimate the marginal distribution of the short rate.

1.1 Nonparametric estimation of the marginal density

Nonparametric kernel estimators of the marginal density that r is equal to u approximate the marginal density of r at u as the average number of r realizations in the sample that are close to u weighted by their distance from u. If the interest rate process is sampled every ΔT years over a span of T years, for a total of $N = \frac{T}{\Delta T}$ observations, then the kernel density estimate has form

$$\hat{\pi}(u) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} K\left(\frac{u - r_i}{h}\right). \tag{5}$$

The weights in the estimation are determined by the kernel function $K(.)$ and the kernel bandwidth h, a parameter which determines the degree of smoothing in the kernel density estimate. The density estimator in Equation (5) can be written equivalently as

$$\hat{\pi}(u) = \frac{1}{T} \sum_{s=\Delta T}^{N\Delta T} \frac{1}{h} K\left(\frac{u - r_{(s)}}{h}\right) \Delta T. \tag{6}$$

Equation (6) expresses the kernel density estimate at u as a weighted average of the amount of time that the interest rate process spends near u in the sample. For a given span of data, the kernel density estimate can be made more precise by sampling more frequently. The best one could do would be to sample the data continuously. Although such sampling is not feasible, positing such a case is useful because it establishes an upper bound on the

[8] Formally, $\hat{\beta}$ is the solution to

$$\min_{\hat{\beta} \in \mathcal{B}} \int_u (\pi(u, \hat{\beta}) - \hat{\pi}(u))^2 \hat{\pi}(u)\, du.$$

performance of the density estimate. Continuous sampling corresponds to allowing $\Delta T \to 0$, which causes the density estimate in Equation (6) to converge to the continuous sampling kernel density estimate:[9]

$$\hat{\pi}(u) = \frac{1}{T}\int_0^T \frac{1}{h} K\left(\frac{u - r_{(s)}}{h}\right) ds. \tag{7}$$

The most important element in kernel density estimation is the choice of the kernel bandwidth h. The rate at which h decreases with sample size is used to derive the asymptotic distribution of the kernel density estimate. In finite samples, the bandwidth is chosen sometimes to accentuate particular features of the distribution, but more often to minimize some statistical loss function.[10] The most common loss function is the mean integrated squared error (MISE) of the kernel density estimate, which measures the expected squared distance between the estimated and the true densities integrated over the range of r:

$$MISE\{\hat{\pi}(.)\} = E\int_u \left[\hat{\pi}(u) - \pi(u)\right]^2 du. \tag{8}$$

The MISE can be further decomposed into the density estimate's mean integrated variance (MIVAR) and its mean integrated squared bias (MISB):

$$\begin{aligned} MISE\{\hat{\pi}(.)\} &= E\int_u \left[\hat{\pi}(u) - E\hat{\pi}(u)\right]^2 du \\ &\quad + E\int_u \left[E\hat{\pi}(u) - \pi(u)\right]^2 du \qquad (9) \\ &= MIVAR\{\hat{\pi}(.)\} + MISB\{\hat{\pi}(.)\}. \qquad (10) \end{aligned}$$

Because MISE is not observable, the bandwidth is usually chosen to minimize some proxy for it. Such minimization is often done by selecting the bandwidth that would be optimal if the data came from a particular distribution or by constructing a proxy for MISE using least-squares cross-validation and then choosing the bandwidth to minimize that proxy.[11]

---

[9] For additional information on kernel density estimation with continuous sampling, see Castellana and Leadbetter (1986).

[10] In the estimation of $\pi(u)$, larger bandwidths involve more averaging of data observations far from u. The averaging reduces the variance of $\hat{\pi}(u)$ and emphasizes more "global" features of the density. However, with the smoothing of r realizations that are relatively far from u, local features of the data are lost and bias is introduced. Smaller bandwidths have the opposite effect. Silverman (1986) and Scott (1992) suggest examining density estimates using several different bandwidths because each bandwidth reveals a different aspect of the data.

[11] For example, in the former procedure the MISE-minimizing bandwidth for N i.i.d. draws from the normal distribution with variance σ^2 is approximately $1.06\sigma N^{-1/5}$. The latter procedure is appealing because Stone (1984) showed that for i.i.d. data the bandwidth chosen by least-squares cross-validation converges to a bandwidth that minimizes the MISE of the density estimate as sample size grows. However, Park and Marron (1990) show that it produces very noisy estimates of optimal bandwidth in small samples. It is also

1.2 Asymptotic distribution of the test statistic

Aït-Sahalia (1996c) used the asymptotic distribution of the kernel density estimate to derive the asymptotic distribution of a test statistic based on $\hat{M}$. He shows that if the kernel bandwidth h_N is chosen to shrink with the number of time-series observations N so that $\lim_{N\to\infty} N h_N = \infty$ and $lim_{N\to\infty} N h_N^{4.5} = 0$, then the distribution of a normalized version of the distance measure converges to the standard normal distribution:

$$\lim_{N\to\infty} h_N^{-1/2} V_M^{-1/2} (\hat{M} - E_M) \xrightarrow{d} \mathcal{N}(0,1), \tag{11}$$

where E_M and V_M, the asymptotic mean and variance of $\hat{M}$, are

$$E_M = \left(\int_{-\infty}^{+\infty} K^2(x)\, dx\right)\left(\int_{\underline{x}}^{\overline{x}} \pi^2(u)\, du\right)$$

$$V_M = 2\left(\int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} K(u)K(u+x)\, du\right]^2 dx\right)\left(\int_{\underline{x}}^{\overline{x}} \pi^4(u)\, du\right).$$

It is important to note that the asymptotic distribution of the test statistic in Equation (11) depends only on the marginal density $\pi(.)$ and not on the time-series properties of the process. This means that the asymptotic distribution is the same as it would be if all the observations of r were i.i.d. draws from the marginal distribution. If the time-series properties of the process affect the distribution of $\hat{\pi}(.)$ in finite samples, then they could bias the size of the test.

2. Test Size and Power

2.1 The data-generating process

To analyze the size of the test in Equation (11), I simulated the way the test performs when interest rates evolve according to the Vasicek (1977) model:

$$dr = \kappa(\theta - r)dt + \sigma dw, \tag{12}$$

where the parameters κ and σ are restricted to be positive, and the value of θ is finite.

known to generate downward-biased bandwidth estimates in finite samples with dependent observations.

Under the process in Equation (12), r has normal marginal and transition densities.[12] The marginal density of r is

$$\pi(u \mid \kappa, \theta, \sigma) = \frac{1}{\sqrt{2\pi V_E}} \exp^{-.5\left(\frac{u-\theta}{\sqrt{V_E}}\right)^2}, \tag{13}$$

where

$$V_E = \frac{\sigma^2}{2\kappa}.$$

Equation (13) shows that the marginal density of r is normal with unconditional mean θ and variance $\frac{\sigma^2}{2\kappa}$. The parameter κ determines the persistence of the process by controlling the rate at which interest rates revert toward their unconditional mean. Lowering the value of κ increases persistence because it slows the rate of mean reversion, which increases the correlation between observations of the interest rate process.

The finite sample distribution of the test is likely to depend both on the marginal distribution of r and on the persistence of the interest rate process. To quantify the effect of persistence separately from the effect of changes in the marginal distribution, it is useful to study processes in which persistence can be varied while the marginal distribution is fixed. Vasicek's model has this property [as does the Cox, Ingersoll, and Ross (1985) model] because its marginal distribution is invariant to proportional changes in σ^2 and κ. Therefore by changing σ^2 and κ in the same proportions, one can vary the persistence of the process while holding the marginal density fixed. In my simulation work, I generated data as in Equation (12) for five different specifications of the κ and σ^2 parameters. To make the parameter choices realistic, my baseline model (labeled model 0 in table 1) uses Vasicek model parameter estimates from Aït-Sahalia (1996b), which are based on real data.[13] The parameters in the baseline model are $\kappa = .85837$, $\theta = .089102$, and $\sigma^2 = .0021854$. I also considered models in which the baseline κ and σ^2 were doubled, quadrupled, halved, or quartered. The corresponding models in Table 1 are labeled (-2), (-1), (1), and (2). Models toward the top of the table have less persistence, but all models have the same marginal distribution.

---

[12] The transition density of r is

$$\pi(r_{(s)}, s \mid r_{(t)}, t, \kappa, \theta, \sigma) = \frac{1}{\sqrt{2\pi V(r_{(s)} \mid r_{(t)})}} \exp^{-.5\left(\frac{r_{(s)}-\mu(r_{(s)}|r_{(t)})}{\sqrt{V(r_{(s)}|r_{(t)})}}\right)^2},$$

where $\mu(r_{(s)} \mid r_{(t)}) = \theta + (r_{(t)} - \theta)e^{-\kappa(s-t)}$ and $V(r_{(s)} \mid r_{(t)}) = V_E(1 - e^{-2\kappa(s-t)})$.

[13] Aït-Sahalia estimated these parameters by GMM.

Table 1
Finite sample properties of marginal density test

A: Asymptotic and finite sample 5% critical values

	Vasicek parameters			5% critical values			
Model	κ	θ	σ^2	Asymptotic	22 Year	Lower bound	Upper bound
−2	3.433480	0.089102	0.008742	1.645	11.52	10.21	13.36
−1	1.716740	0.089102	0.004371	1.645	17.00	13.43	20.25
0	0.858370	0.089102	0.002185	1.645	19.47	16.60	23.90
1	0.429185	0.089102	0.001093	1.645	19.17	15.41	25.43
2	0.214592	0.089102	0.000546	1.645	11.08	8.51	18.23

B: Asymptotic and finite sample 1% critical values

	Vasicek parameters			1% critical values			
Model	κ	θ	σ^2	Asymptotic	22 Year	Lower bound	Upper bound
−2	3.433480	0.089102	0.008742	2.33	17.70	15.45	49.87
−1	1.716740	0.089102	0.004371	2.33	29.73	23.90	52.79
0	0.858370	0.089102	0.002185	2.33	44.09	37.30	127.25
1	0.429185	0.089102	0.001093	2.33	36.44	29.55	166.56
2	0.214592	0.089102	0.000546	2.33	39.94	28.67	130.53

C: Empirical rejection frequencies using asymptotic critical values

	Vasicek parameters			5% level		1% level	
Model	κ	θ	σ^2	Rej. freq.	Std. err.	Rej. freq.	Std. err.
−2	3.433480	0.089102	0.008742	45.60%	2.23%	37.80%	2.17%
−1	1.716740	0.089102	0.004371	57.40%	2.21%	49.40%	2.24%
0	0.858370	0.089102	0.002185	51.60%	2.23%	43.60%	2.22%
1	0.429185	0.089102	0.001093	40.80%	2.20%	34.20%	2.12%
2	0.214592	0.089102	0.000546	21.00%	1.82%	18.80%	1.75%

Panels A and B present asymptotic and finite sample critical values for the $\hat{M}$ test in Equation (11) when the short-term interest rate is sampled once a day for 22 years. Lower and upper bounds of a 95% confidence interval for the finite sample critical values are provided. Panel C presents estimates of the probability of rejecting the null when it is true using asymptotic critical values. All finite sample results are based on 500 Monte Carlo similuations. For all tests, the short rate was generated by the Vasicek model:

$$dr = \kappa(\theta - r)dt + \sigma dW,$$

with the parameters shown. The kernel bandwidths used for models −2 through 2 are 0.0140979, 0.175509, 0.0217661, 0.0268048, and 0.0325055, respectively.

2.2 Test implementation

To estimate the size of the test, I performed 500 Monte Carlo simulations for each parameterization of the Vasicek model. For each simulation I generated 22 years of daily data for a total of 5500 observations, and then estimated $\hat{\pi}(u)$ using kernel density estimation with a Gaussian kernel function.[14] The kernel bandwidth used in each simulation exactly minimizes the MISE of the kernel density estimate when the data are sampled continuously. For the data-generating processes in the simulations, the MISE-minimizing band-

[14] I assumed 250 business days per year.

width using daily sampling is virtually the same as the MISE-minimizing bandwidth using continuous sampling. I could choose the bandwidth to exactly minimize MISE because I knew the data-generating process used in the simulations. In his empirical work, Aït-Sahalia chose the bandwidth to minimize a statistical proxy for MISE with a technique known as least-squares cross-validation (LSCV). For comparison, I also conducted a Monte Carlo analysis for my baseline Vasicek model in which the bandwidth was chosen using LSCV.

To compute the test statistics, I used the kernel density estimates and the formula for the marginal density in Equation (13) to generate the following consistent estimates of M, V_M, and E_M:[15]

$$\hat{M} = \min_{\hat{\theta}\in\Theta} N h_N \int_{\underline{r}}^{\overline{r}} [\pi(\hat{\theta}, u) - \hat{\pi}(u)]^2 \hat{\pi}(u)\, du, \tag{14}$$

$$\hat{E}_M = \left(\int_{-\infty}^{+\infty} K^2(x)\, dx\right)\left(\int_{\underline{x}}^{\overline{x}} \hat{\pi}^2(u)\, du\right), \text{ and} \tag{15}$$

$$\hat{V}_M = 2\left(\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} K(u)K(u+x)\, du\right]^2 dx\right)\left(\int_{\underline{x}}^{\overline{x}} \hat{\pi}(u)^4\, du\right). \tag{16}$$

2.3 Empirical results: size

The asymptotic critical values of the test statistic in Equation (11) at the confidence levels of 5% and 1% are 1.645 and 2.33 respectively. Panels A and B of Table 1 contrast the asymptotic critical values of the test with Monte Carlo estimates of the critical values when only 22 years of data are used. Next to each critical-value estimate are the upper and lower bounds of a 95% confidence interval for the true critical values.[16]

---

[15] I also used the alternative consistent estimators

$$\hat{M} = \min_{\hat{\theta}\in\Theta} N h_N (1/N) \sum_{i=1}^{N} [\pi(\hat{\theta}, x_i) - \hat{\pi}(x_i)]^2,$$

$$\hat{E}_M = \left(\int_{-\infty}^{+\infty} K^2(x)\, dx\right)(1/N)\sum_{i=1}^{N} \hat{\pi}(x_i), \text{ and}$$

$$\hat{V}_M = 2\left(\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} K(u)K(u+x)\, du\right]^2 dx\right)(1/N)\sum_{i=1}^{N} \hat{\pi}^3(x_i).$$

Because my results on the size of the test using both sets of estimators were similar, results using this second set of estimators are not reported.

[16] The confidence intervals are based on the order statistics of the Monte Carlo estimates and are fully nonparametric. Details on how to create these confidence intervals are contained in David (1981).

The results in panels A and B are dramatic. The 22-year critical value estimates and the lower bounds of the confidence intervals for these estimates are roughly 10 times the finite sample critical values at both the 5% and 1% confidence levels. Because the lower-bound estimates are less than the true finite sample critical values with at least 95% probability, the evidence for severe size distortion is very strong.

Panel C shows empirical rejection frequencies when asymptotic critical values are used. The main result is that the test rejects far too often. In the case of my baseline model, it rejects about 50% of the time at the 5% confidence level. Panel C can also be used to examine the effects of persistence on the size of the test statistic. Intuitively, more persistence should increase the variance of marginal density estimates because the observations contain less independent information on the marginal distribution. Because the asymptotic critical values of the test statistic do not adjust for persistence, one would expect the test to reject more often in samples with higher persistence. Consistent with this intuition, rejection rates increase in moving from model -2 to model -1, but then, perversely, rejection rates tend to move down as the interest rate process becomes even more persistent. The intuition on persistence is correct; the density estimates from the more persistent processes are of lower quality. However, in this particular case they also appear more Gaussian.[17] Because Aït-Sahalia's test, when using the Vasicek null hypothesis, is formally a test of the distance from the kernel estimate to the class of Gaussian densities, the more Gaussian shape associated with kernel estimates of more persistent processes causes the test to reject less often. To reiterate, this strange phenomenon is an artifact of using a Gaussian kernel and the Vasicek model; it in no way invalidates my main point that high persistence biases the test toward too much rejection.

To complete my empirical work on size, I computed the size of the test when the data are generated as in model 0 and the bandwidth is selected using LSCV.[18] This procedure produced bandwidth choices that were typically half the size of the MISE-minimizing bandwidth and probably would have been lower had I not constrained the cross-validation procedure. As a consequence of these bandwidth choices, the size of the test was dramati-

[17] The kernel density estimates appear more Gaussian for two reasons. First, larger bandwidths make the density estimates more closely resemble the shape of the kernel, which in this case is Gaussian. Second, increasing the persistence while holding the marginal distribution fixed concentrates the data in small samples. This effect further strengthens the tendency of the density estimate to appear like the kernel. As an extreme example, if all of the data in the sample are concentrated at or very close to the number 5 and the kernel function is Gaussian with positive bandwidth h, then the kernel density estimate will be approximately equal to the Gaussian distribution with a mean of 5 and a standard deviation of h.

[18] I used LSCV to search for bandwidths between .5 times and 1.5 times the MISE minimizing bandwidth for continuous sampling. This bandwidth selector is not fully feasible because the search range is centered at the MISE-minimizing bandwidth, which is usually unknown. Nevertheless, my results give some indication of the performance of searching for an optimal bandwidth using LSCV.

cally worse. The null hypothesis was rejected 91.8% of the time at the 5% confidence level and 88.2% of the time at the 1% confidence level.[19]

These size distortions are so severe that they indicate that the asymptotic distribution of the test in Equation (11) is not useful for testing the Vasicek model. The intuition behind the results also suggests that the asymptotics will not be useful for other interest rate models. This inference is confirmed by Conley, Hansen, and Liu (1997), who found serious size distortions when applying this test to other interest rate models.

Although the test has size distortions, it may still be valuable for statistical inference if the test has significant power after controlling for the size distortions. The power of the test is studied in the next section.

2.4 Empirical results: power

To investigate the power of the test, I conducted 500 Monte Carlo simulations. For each simulation I generated 22 years of daily interest rate data using a baseline Cox-Ingersoll-Ross (CIR) model:[20]

$$dr = \kappa(\theta - r)dt + \sigma\sqrt{r}dw, \tag{17}$$

with parameter estimates $\kappa = .89218$, $\theta = .090495$, and $\sigma = \sqrt{.032742}$. I then tested the null hypothesis that the data were generated by my baseline Vasicek model using two different methods.

The first method implemented Aït-Sahalia's test using the approach in Section 3.2, but with finite sample estimates of the critical values.[21] The critical values used in the power computations were the lower bound finite-sample critical-value estimates for model 0 from panels A and B of Table 2. Because the true finite sample critical values are most likely larger than these lower bound estimates, my power results probably overstate the true power of the test.

The second method is a conditional moment test of whether the likelihood function implied by the Vasicek model is properly specified given the data.[22] To implement the test, I estimated the parameters of the Vasicek model using maximum likelihood. If the likelihood function is properly specified, then

[19] The estimated finite sample 5% critical value using LSCV is 61.33 with a 95% confidence interval of [51.51, 71.83]. The estimated finite sample 1% critical value is 114.86 with a 95% confidence interval of [92.85, 239.18]. The 5% and 1% asymptotic critical values are 1.645 and 2.33, respectively.

[20] Like the parameters for my baseline Vasicek model, these parameter estimates come from Aït-Sahalia (1996b) and are GMM estimates based on real data. I therefore consider this my baseline CIR model.

[21] All kernel density estimates used bandwidth .0217661, which minimizes the MISE of the density estimate under the null hypothesis.

[22] See Newey (1985a, b) for additional details on these tests.

Table 2
MISE minimizing bandwidth

$T = 5$ years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.0224	0.0276	0.0333	0.0389	0.0434
1 day	0.0224	0.0276	0.0333	0.0389	0.0434
5 day	0.0224	0.0276	0.0333	0.0389	0.0434
10 day	0.0225	0.0276	0.0333	0.0389	0.0434
20 day	0.0229	0.0277	0.0333	0.0389	0.0434
30 day	0.0234	0.0278	0.0334	0.0389	0.0434

$T = 10$ years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.0181	0.0224	0.0276	0.0333	0.0389
1 day	0.0181	0.0224	0.0276	0.0333	0.0389
5 day	0.0181	0.0224	0.0276	0.0333	0.0389
10 day	0.0183	0.0224	0.0276	0.0333	0.0389
20 day	0.0187	0.0225	0.0276	0.0333	0.0389
30 day	0.0192	0.0227	0.0276	0.0333	0.0389

T = 20 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.0145	0.0181	0.0224	0.0276	0.0333
1 day	0.0145	0.0181	0.0224	0.0276	0.0333
5 day	0.0146	0.0181	0.0224	0.0276	0.0333
10 day	0.0148	0.0181	0.0224	0.0276	0.0333
20 day	0.0154	0.0183	0.0224	0.0276	0.0333
30 day	0.0160	0.0185	0.0225	0.0276	0.0333

T = 50 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.0108	0.0135	0.0169	0.0209	0.0258
1 day	0.0108	0.0135	0.0169	0.0209	0.0258
5 day	0.0110	0.0136	0.0169	0.0209	0.0258
10 day	0.0113	0.0136	0.0169	0.0209	0.0258
20 day	0.0120	0.0139	0.0169	0.0209	0.0258
30 day	0.0126	0.0141	0.0170	0.0209	0.0258

T = 100 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.0087	0.0108	0.0135	0.0169	0.0209
1 day	0.0087	0.0108	0.0135	0.0169	0.0209
5 day	0.0089	0.0109	0.0135	0.0169	0.0209
10 day	0.0093	0.0110	0.0136	0.0169	0.0209
20 day	0.0101	0.0113	0.0136	0.0169	0.0209
30 day	0.0107	0.0117	0.0137	0.0169	0.0209

for each observation the following moment restriction should be satisfied:

$$E\left(\left.\frac{\partial\mathcal{L}(r_t \mid r_{t-\Delta t}, \sigma, \kappa, \theta)}{\partial\sigma}\right| r_{t-\Delta t}\right) = 0, \tag{18}$$

where $\mathcal{L}(r_t \mid r_{t-\Delta t}, \kappa, \theta, \sigma)$ is the conditional log likelihood of r_t given $r_{t-\Delta t}$ and the parameters that generated the data. To test this moment restriction, I estimated the regression

$$\frac{\partial \mathcal{L}(r_t \mid r_{t-\Delta t})}{\partial \sigma} = \beta_0 + \beta_1 r_{t-\Delta t} + u_t \tag{19}$$

and then performed a two-sided t-test of whether $\beta_1 = 0$.[23] This specification test should be powerful for testing the Vasicek null hypothesis against the CIR alternative because the expected value of the score of the likelihood function for the Vasicek model is approximately linear in r if the data were generated under the CIR alternative. Monte Carlo analysis using 500 simulations under the baseline Vasicek null shows that this test has the appropriate size under the null hypothesis.[24]

When the data were generated by the baseline CIR model, Aït-Sahalia's size-corrected test rejects the Vasicek null hypothesis 39.8% of the time at the 5% level and 13.8% of the time at the 1% level. Thus it displays modest power against the CIR model after critical values are adjusted for size. By contrast, the GMM test rejected the null hypothesis for all 500 Monte Carlo runs at the 1% level or better. This result shows that the likelihood-based GMM test is far more powerful for distinguishing between the Vasicek null hypothesis and the CIR alternative.

The modest power of Aït-Sahalia's test suggests that the marginal density may not be estimated precisely enough for a test based on the marginal density alone to be able to distinguish among various short-rate models. The next section studies how well the marginal density is estimated in finite samples.

3. Kernel Estimation and the Vasicek Model

This section derives finite sample properties of kernel density estimates of the marginal distribution of the short rate when interest rates are generated according to the Vasicek model. I chose to focus on this model not because it is fully realistic but because its tractability allows me to derive finite sample results.

The first part of the analysis presents my finite sample results on the effects that persistence, data span, and sampling frequency have on the quality of the nonparametric density estimates and on the choice of the kernel bandwidth. The most important finding here is that in finite samples the optimal choice of bandwidth depends on the persistence of the process but not on the frequency with which the process is sampled. This result

---

[23] Standard errors were computed using Newey and West's method with a lag length of 15 [see Newey and West (1987)].

[24] Using data generated under the Vasicek null hypothesis, this specification test rejected the null hypothesis 4.8% of the time at the 5% level in 500 Monte Carlo runs, and it rejected the null hypothesis 1.6% of the time at the 1% level.

suggests that there may be scope for improving upon bandwidth selection methods that are based on the number of time-series observations but not on the persistence of the process.

The second part of the analysis contrasts the finite sample distribution of the kernel density estimate to the asymptotic distribution of the kernel density estimate. My primary finding is that the asymptotic distribution understates the finite sample bias, variance, and covariance of the kernel density estimate. This finding explains why inference based on the asymptotic distribution of the kernel density estimate may be misleading.

3.1 Finite sample results

This section studies the finite sample performance of kernel density estimators when the interest rate data are generated by the Vasicek model [Equation (12)] and when the kernel function is Gaussian. Because the marginal and transition densities of the Vasicek process are also Gaussian, I could derive analytic or near-analytic expressions for the estimator's finite sample bias, variance, covariance, and MISE. I used these expressions to study how MISE and the MISE-minimizing bandwidth change with sampling frequency, span of the data, and persistence of the data-generating process.

Proposition 1 presents analytical expressions for the bias and MISE for kernel estimates of the marginal distribution of the Vasicek process when the process is sampled at N discrete intervals each of length ΔT. The expression for MISE was derived earlier by Wand (1992) using a technique that is different from the one used here. Expressions for the finite sample variance and covariance of the kernel density estimate are contained in parts C and D of the Appendix, respectively.

Proposition 1. *If $r_{(0)}$ is drawn from the marginal distribution of the Vasicek process given in Equation (13), and r then evolves according to the diffusion in Equation (12), and $\hat{\pi}(u) = \frac{1}{N}\sum_{s=\Delta T}^{N\Delta T}\frac{1}{h}K(\frac{u-r_{(s)}}{h})$, where K(.) is the standard Gaussian kernel, then*

$$E[\hat{\pi}(u)] = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{h^2+V_E}}\exp^{-.5\left(\frac{u-\theta}{\sqrt{h^2+V_E}}\right)^2}, \tag{20}$$

$$MISB\{\hat{\pi}(.)\} = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{2h^2+2V_E}}+\frac{1}{\sqrt{2V_E}}-\frac{2}{\sqrt{(h^2+2V_E)}}\right), \tag{21}$$

$$MIVAR\{\hat{\pi}(.)\} = \frac{1}{N^2\sqrt{2\pi}}\left[\frac{N}{\sqrt{2h^2}}+2\sum_{l=2}^{N}\sum_{m=1}^{l-1}\frac{1}{\sqrt{2h^2+2V_E(1-e^{-\kappa(l-m)\Delta t})}}\right.$$
$$\left.-\frac{N^2}{\sqrt{2h^2+2V_E}}\right] \tag{22}$$

$$MISE\{\hat{\pi}(.)\} = MISB\{\hat{\pi}(.)\} + MIVAR\{\hat{\pi}(.)\}, \tag{23}$$

where, $V_E = \frac{\sigma^2}{2\kappa}$.

Proof. See part C of the Appendix.

Proposition 2 presents results that are analogous to those in Proposition 1 when the kernel density estimate uses continuous sampling as in Equation (7). Expressions for the variance and covariance are contained in parts C and D of the Appendix, respectively.

Proposition 2. *If* $r_{(0)}$ *is drawn from the marginal distribution of the Vasicek process given in Equation (13), and* r *then evolves according to the diffusion in Equation (12), and* $\hat{\pi}(u) = \frac{1}{Th}\int_0^T K(\frac{u-r_{(s)}}{h})ds$, *where K(.) is a Gaussian kernel, then* $E\hat{\pi}(u)$ *and MISB are as in Equations (20) and (21), and*

$$MIVAR\{\hat{\pi}(.)\} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2h^2+2V_E}}\frac{4}{\kappa T^2} \times \int_0^T \ln\left[\frac{1+\sqrt{1-\psi(t)}}{1+\sqrt{1-\psi(0)}}\right]dt, \tag{24}$$

$$MISE\{\hat{\pi}(.)\} = MISB\{\hat{\pi}(.)\} + MIVAR\{\hat{\pi}(.)\}, \tag{25}$$

where $V_E = \frac{\sigma^2}{2\kappa}$ *and* $\psi(t) = (\frac{V_E}{h^2+V_E})e^{-\kappa t}$.

Proof. See part B of the Appendix.

The expressions in Propositions 1 and 2 are very tractable. Equation (20) shows that the expected kernel density estimate under the Vasicek model is Gaussian, but with a variance that exceeds that of the marginal distribution by h, the kernel bandwidth. Therefore the form of the kernel density estimator's bias for the Vasicek model is that its variance is too large. Because the expected kernel density estimate depends only on h and the marginal distribution, bias does not depend on the persistence of the process. Persistence does, however, affect the mean integrated variance of the kernel density estimate. This effect can be seen by noting that κ enters the expression for MIVAR separately from V_E in Propositions 1 and 2. For the kernel density estimate with discrete sampling, Proposition 3 illustrates the effect of persistence in two extreme cases.

Proposition 3. *Under the conditions of Proposition 1, for fixed N and* V_E,
(i) $\lim_{\kappa\to\infty} MIVAR\{\hat{\pi}(.)\} = N^{-1}(2\pi)^{-.5}[(2h^2)^{-.5} - (2h^2+2V_E)^{-.5}]$.
(ii) $\lim_{\kappa\to 0} MIVAR\{\hat{\pi}(.)\} = (2\pi)^{-.5}[(2h^2)^{-.5} - (2h^2+2V_E)^{-.5}]$.

Proof. Results (i) and (ii) follow from taking limits.

Both of these results are intuitive. The first result corresponds to the extreme case of persistence going to zero. As κ goes to infinity the rate of mean reversion of the interest rate process goes to infinity and the correlation between the interest rate observations goes to zero. In this limiting case, because the interest rate observations are Gaussian and their correlations go to zero, the observations become i.i.d. Gaussian random variables, and the MIVAR converges to the MIVAR for kernel estimation with N i.i.d. Gaussian observations.

Result (ii) corresponds to an extreme in which the data are very persisent. As κ goes to zero, mean reversion goes to zero, and the correlation between discretely spaced observations of the process approaches one. Because the correlation of all N observations approaches one in this limit, the $N-1$ observations beyond the first do not contribute any new information on the marginal density. Therefore MIVAR in this case converges to the MIVAR from one observation of the interest rate process.

The results in Proposition 3 are also partially replicated for the continuous sampling case. When κ goes to infinity while V_E is held fixed, the MIVAR goes to zero. Intuition for this result is that, as κ goes to infinity, the continuous sampling kernel estimate uses a continuum of i.i.d. observations, forcing the MIVAR to zero. I have been unable to generate limiting results as κ goes to zero in the continuous sampling case; however, in the analysis below I have computed the effects of reducing κ while holding V_E fixed and have found that doing so increases MISE.

To further analyze MISE, I used the results of Propositions 1 and 2 to compute the effect that persistence, sampling frequency, and data span have on MISE and on the choice of the MISE-minimizing bandwidth. My analysis used both continuous and discrete sampling for six different spans of data and six different sampling frequencies. Table 2 presents results on optimal bandwidth choice when the bandwidth is chosen to minimize MISE. Table 3 presents results on the MISE when the bandwidth is chosen optimally. The Vasicek parameters and models in these tables are identical to those in Table 1.[25]

Two important stylized facts are immediately apparent from Table 2. For the amounts of mean reversion considered in the table, the choice of optimal bandwidth is highly insensitive to the frequency with which the data are sampled and highly sensitive to the persistence of the data-generating process.[26] If either fact is ignored in bandwidth selection, the resulting bandwidth could be quite poor. For example, the bandwidth selection rule for i.i.d. Gaussian

---

[25] Tables 2 and 3 assume that there are 240 rather than 250 business days per year in order to use the same spans of data for all sampling frequencies. To obtain the same span of data for all sampling frequencies requires that the number of business days per year be divisible by all the sampling frequencies.

[26] The optimal bandwidth increases 50% or more when moving from the least persistent model (-2) to the most persistent model (2).

Table 3
Mean integrated squared error

T = 5 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.5466	0.8544	1.2337	1.6090	1.9007
1 day	0.5467	0.8544	1.2337	1.6090	1.9007
5 day	0.5474	0.8546	1.2338	1.6090	1.9007
10 day	0.5498	0.8553	1.2340	1.6091	1.9007
20 day	0.5587	0.8579	1.2348	1.6094	1.9009
30 day	0.5719	0.8622	1.2361	1.6099	1.9011

T = 10 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.3302	0.5466	0.8544	1.2337	1.6090
1 day	0.3302	0.5466	0.8544	1.2337	1.6090
5 day	0.3310	0.5468	0.8544	1.2337	1.6090
10 day	0.3331	0.5474	0.8546	1.2338	1.6091
20 day	0.3409	0.5498	0.8553	1.2340	1.6091
30 day	0.3520	0.5536	0.8564	1.2343	1.6092

T = 20 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.1910	0.3302	0.5466	0.8544	1.2337
1 day	0.1911	0.3302	0.5466	0.8544	1.2337
5 day	0.1917	0.3304	0.5467	0.8544	1.2337
10 day	0.1937	0.3310	0.5468	0.8544	1.2337
20 day	0.2005	0.3331	0.5474	0.8546	1.2338
30 day	0.2094	0.3365	0.5484	0.8549	1.2338

T = 50 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.0882	0.1590	0.2780	0.4673	0.7455
1 day	0.0882	0.1590	0.2780	0.4673	0.7455
5 day	0.0889	0.1592	0.2781	0.4673	0.7455
10 day	0.0906	0.1597	0.2782	0.4674	0.7455
20 day	0.0957	0.1616	0.2788	0.4675	0.7455
30 day	0.1019	0.1645	0.2797	0.4677	0.7456

T = 100 years	Model				
Sampling frequency	−2	−1	0	1	2
Continuous	0.0478	0.0882	0.1590	0.2780	0.4673
1 day	0.0478	0.0882	0.1590	0.2780	0.4673
5 day	0.0478	0.0882	0.1590	0.2780	0.4673
10 day	0.0478	0.0882	0.1590	0.2780	0.4673
20 day	0.0478	0.0882	0.1590	0.2780	0.4673
30 day	0.0478	0.0882	0.1590	0.2780	0.4673

data chooses the bandwidth as approximately $1.06\sqrt{V_E}N^{(-1/5)}$, where N is the number of time-series observations and V_E is the estimated variance of the marginal distribution. If this rule is applied for time-series data sampled once every 30 days or sampled once every day, the latter bandwidths will be about half the size of the former even though my results indicate that the appropriate bandwidths should be roughly the same for both. Table 2 also shows that, as expected, the optimal bandwidth shrinks with the span of the data.

Table 3 reveals similar information for MISE. MISE is insensitive to sampling frequency: for model 0 and 20 years of data, sampling the data daily instead of monthly reduces the MISE by a meager 0.25%. The table also shows MISE is very sensitive to persistence and decreases with the span of the data.

A final insight from Table 3 concerns choosing the bandwidth using automatic methods such as LSCV. The basic principle behind cross-validation is to estimate the kernel density while omitting some observations, and then to use the quality of the fit for the omitted observations in choosing the kernel bandwidth. In i.i.d. data, this type of procedure is intuitive. If the bandwidth is too small, the data will be overfitted. Conversely, if the bandwidth is too large, the data will be oversmoothed. In both cases, the fit for the omitted data will be poor.

A key element of the i.i.d. reasoning is that the data being omitted in cross-validation add significant new information to the estimation problem. Interest rate data that are sampled daily add almost no more new information than data that are sampled monthly (Table 3). In effect, this means that the observations from persistent data sampled at a high frequency will be clustered very close together and therefore will not provide significant additional information about the marginal distribution. The cross-validation algorithm will interpret the clustering as fine detail in small samples and will fit this spurious fine detail by choosing a bandwidth that is too small. This downward bias phenomenon is also documented in Hart and Vieu (1990) who propose a fix that involves cross-validating on observations that are spread less closely in time.[27]

3.2 Quality of the asymptotic approximation

In this section I contrast the finite sample and asymptotic properties of kernel density estimates of the marginal distribution of interest rates in the Vasicek model.

Robinson (1983) provides conditions on the kernel function, time-series dependence, true marginal density, and the rate of bandwidth shrinkage, such that the distribution of the kernel density estimates at n discretely separated points $u_1, \ldots, u_n$ are asymptotically unbiased, normally distributed, and uncorrelated. Moreover, this asymptotic distribution is the same as that from a kernel density estimate that uses i.i.d. draws from the marginal distribution of the interest rate process. Robinson emphasizes that these asymptotic results should not be taken too seriously because in finite samples the quality of the asymptotic approximation will depend on both serial dependence and the size of the bandwidth.

[27] Silverman (1986) also notes the possibility of severe downward bias in bandwidths chosen by cross-validation when the data have been rounded or when large numbers of observations are nearly identical to other observations in the data.

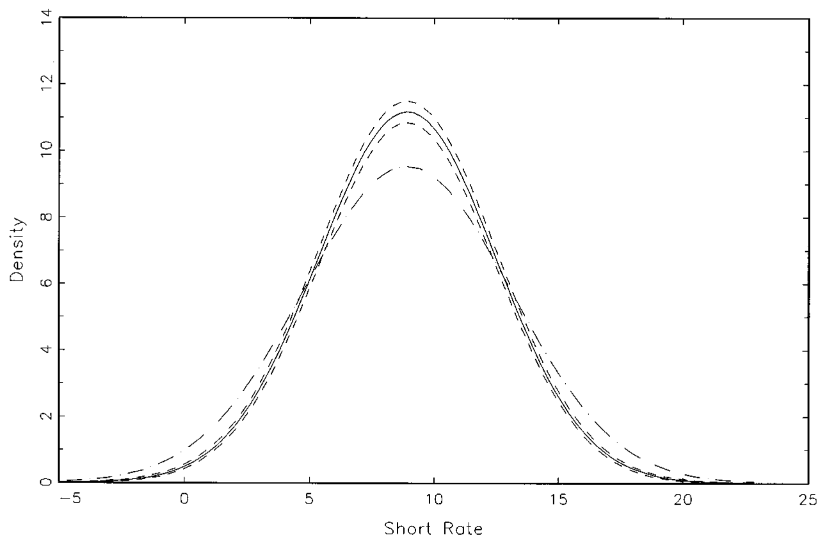

Figure 1
True and expected marginal density estimates
For 22 years of daily data that are generated by the baseline Vasicek model (Table 1, model 0), the figure shows the expected kernel density estimate (dashes and dots), the true marginal density (solid), and a plus or minus 2 standard deviation unnormalized asymptotic confidence band (dashed) centered at the true marginal density.

My results show that the asymptotic approximation of the finite sample properties understate the finite sample magnitudes of the bias, variance, covariance, and correlation of the kernel density estimate. Figures 1–3 illustrate these results using 22 years of daily sampled data for the baseline Vasicek model.

Figure 1 plots the true (solid) and expected (dashes and dots) kernel density estimates for the baseline Vasicek model from Table 1. The difference between the true and expected densities is the bias. Although this bias is zero asymptotically, the finite sample bias appears to be important. One measure of the bias's importance is the extent to which it affects statistical inference in tests that use the asymptotic distribution of the statistic but ignore the bias.

To perform a "visual test," Figure 1 contains an asymptotic plus or minus 2 standard deviation confidence band (dashes) around the true marginal distribution of r.[28] If the density is estimated at a set of discretely spaced points and the estimated distribution function lies outside the confidence

[28] The confidence band in the figure is the unnormalized asymptotic standard deviation. For example, the unnormalized asymptotic standard deviation of $\pi(u)$, the kernel density estimate for interest rate level u is equal to $(Nh)^{-.5}\sqrt{\frac{1}{2\sqrt{\pi}}\pi(u)}$.

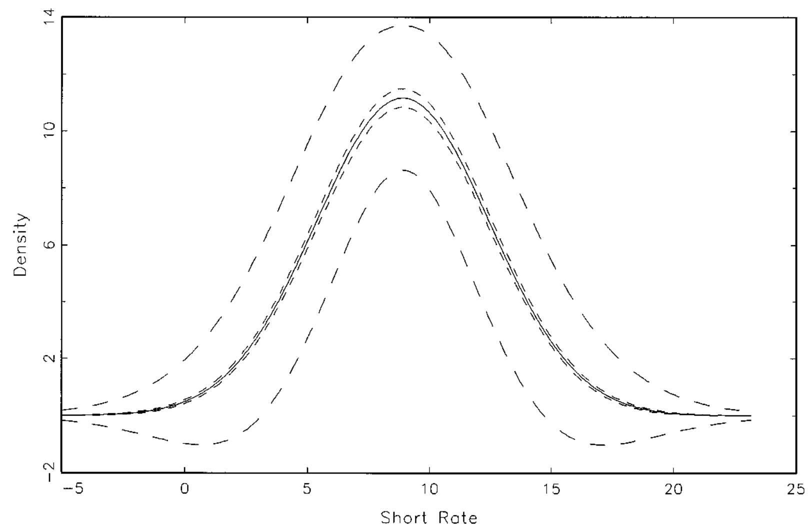

Figure 2
Finite sample and asymptotic confidence bands
For 22 years of daily data that are generated by the baseline Vasicek model (Table 1, model 0), the figure shows the true marginal density (solid), an unnormalized plus or minus 2 standard deviation asymptotic confidence band (short dashes) centered at the true density, and a plus or minus 2 standard deviation finite sample confidence band (long dashes) centered at the true density.

band at more than 5% of those points, a test based on the asymptotic distribution will almost always reject (at the 5% confidence level) the null hypothesis that the marginal distribution of the data is equal to the true marginal distribution. Because the magnitude of the finite sample bias is much larger than the confidence band, the density estimate typically will lie mostly outside of the asymptotic confidence bands, causing the null hypothesis to be rejected too often.

Figure 2 illustrates that the asymptotic distribution not only understates the finite sample bias, it also severely understates the finite sample variance of the kernel density estimate. The curves at the top and bottom of the figure (long dashes) are finite sample plus or minus 2 standard deviation confidence bands around the true density (solid). The curves in the middle of the figure (short dashes) are the asymptotic plus or minus 2 standard deviation confidence bands. The true density estimates are far more variable than implied by the asymptotic distribution. Therefore the null hypothesis that the estimated marginal distribution is equal to the true distribution will be rejected far too often using statistical tests based on the asymptotic variance.

An additional feature of the asymptotic distribution is that kernel estimates at different points of the support of the distribution are asymptotically

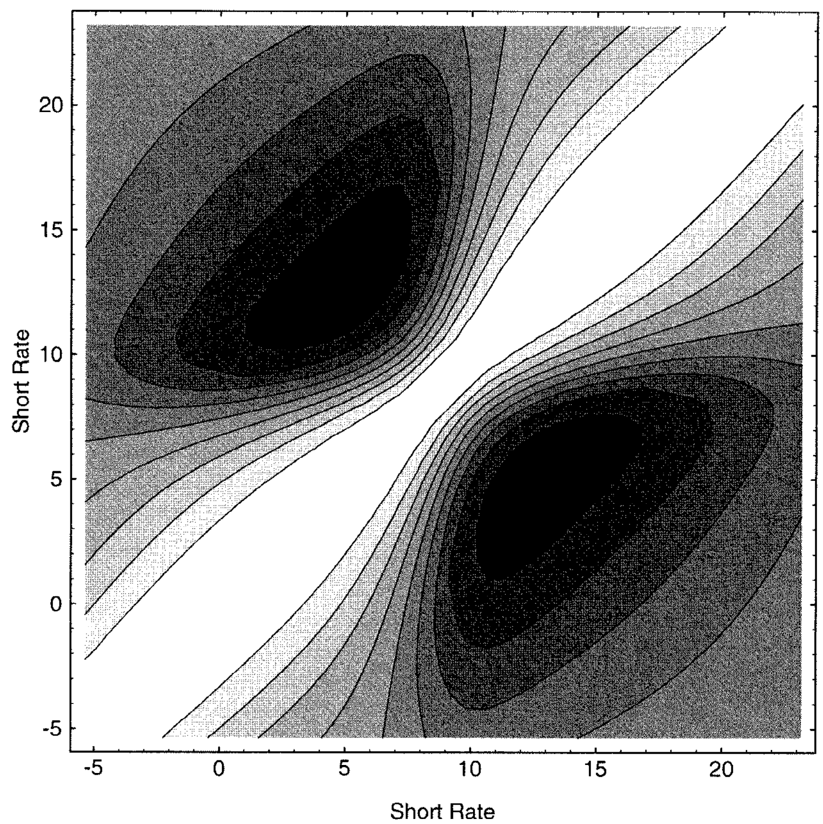

Figure 3
Finite sample correlation of kernel density estimator
For 22 years of daily data that are generated by the baseline Vasicek model (Table 1, model 0), the figure presents a contour plot of the correlation between estimates of the marginal density at different levels of the short rate. For example, the correlation between the estimated density at $r = 8$ and the estimated density at $r = 8$ is 1. The contours represent gradations of 0.2 in correlation. The diagonal white bands represents correlations of 0.8 to 1.0, the adjacent band 0.6 to 0.8, and so on.

uncorrelated. To contrast this asymptotic result with the finite sample correlation, Figure 3 presents a contour plot of the correlation function of the kernel density estimator for the baseline model with 22 years of data, sampled daily. The contours represent gradations of 0.2 in correlation. For example, the white band down the center of each graph represents the points that have correlations of at least 0.8. The band just below this one has correlations from 0.6 to 0.8, and so on. Contrary to the asymptotic distribution, the finite sample correlations in the figure are substantial.

Figure 3 also shows that density estimates above the mean of the distribution of the short rate are usually negatively correlated with density estimates below the mean, and positively correlated with density estimates above the mean. An intuitive explanation for this feature of the correlation function is that the first interest rate observation in the sample has an important influence on subsequent observations. For example, in the Vasicek model if the starting value of the process is above the mean of the distribution, future values are expected to stay above the mean as well. As a result, conditional on the first observation being above the mean, the kernel density estimate will be biased toward assigning too much probability mass above the mean and too little below the mean. This bias worsens with the persistence of the process and remains a feature of the distribution of the estimator even with 100 years of data [Pritsker (1997)].

A final method of examining the adequacy of the asymptotic distribution of the kernel density estimator is to focus directly on the property that its asymptotic distribution is the same as if the data were independently and identically distributed. To do so, I computed the amount of data required to achieve various target levels of MISE if the data were generated by the baseline Vasicek model, or if the data were i.i.d. draws from its marginal distribution. If the "as if i.i.d." property of the asymptotic distribution is reasonable, then the number of years of data required to achieve a particular target level of MISE should be about the same in both cases.

I examined two target MISE levels. The first is 0.008. To achieve an MISE of 0.008 with the i.i.d. draws requires 22 years of daily data (5500 observations). To achieve the same MISE with data from the baseline Vasicek model requires 2755 years of continuously sampled data.[29] Put differently, to achieve the accuracy implied by the asymptotic distribution of the kernel density estimator with 22 years of daily data (Aït-Sahalia's sample size) actually requires thousands of years of data. This indicates that the asymptotic distribution drastically overstates the amount of information that actually is in the data.

To quantify the information that is in the data, I considered an MISE target of 0.5116. To achieve this MISE with data from the baseline Vasicek model requires 22 years of continuously sampled data. To achieve the same MISE with the i.i.d. data requires only 16 i.i.d. draws from the marginal distribution. Put differently, interest rates are so persistent that 22 years of continuously sampled data contains very little information that is useful for identifying the marginal distribution with nonparametric density estimation.

These final two results provide our final intuition on the size and power of Aït-Sahalia's test. Because the asymptotic distribution of the kernel density estimate overstates the information that is in the data, the test rejects

[29] The MISEs were computed using MISE-minimizing bandwidths. The formula for MISE with the continuous sampling kernel density estimate is contained in Proposition 2. The formula for MISE for i.i.d. observations from the marginal distribution is the sum of the expression for MISB from Proposition 1 with the expression for MIVAR from the first result in Proposition 3.

too often. Because the marginal distribution is not well identified using nonparametric density estimation, the test has low power.

4. Conclusions and Directions for Further Research

In this article I examined one of Aït-Sahalia's (1996c) nonparametric tests of models of the short rate of interest, and I examined the properties of the nonparametric kernel density estimator on which the test was based. Our primary results showed that Aït-Sahalia's test rejected true interest rate models far too often. One reason for the high rejection rate appears to be that the asymptotic distribution of the test is the same as if the data were independently and identically distributed even when the data are actually highly dependent. After asymptotic critical values were adjusted for size, the test was found to have low power relative to a conditional moment-based specification test. I also studied the role of persistence and sampling frequency in kernel density estimation and found that the estimator's MISE and optimal bandwidth choice are very sensitive to persistence but highly insensitive to the frequency with which the data is sampled.

The most important implication of these results is that inference based on the asymptotic distribution of kernel density estimators has the potential to be very misleading. In this regard, Aït-Sahalia's test, the nonparametric kernel density estimator, the nonparametric kernel regression estimator, and the localized test function estimator of Conley et al. (1997), are especially problematic because their asymptotic distributions don't depend on persistence, although their finite sample distributions do. Until the adequacy of the asymptotic distributions of nonparametric estimation methods becomes better understood, there appears to be a strong need to compute standard errors and confidence intervals for nonparametric estimators by the use of other methods such as the parametric bootstrapping procedure discussed in Conley, Hansen, and Liu (1997).

The limited examination of power performed here suggests that tests tailored to particular parametric alternatives are more powerful than the nonparametric method considered here. Therefore interest rate models should first be tested against plausible parametric alternatives using tests specifically tailored for those alternatives. If the model passes the first set of tests then nonparametric tests should be applied to search for forms of misspecification that may have been missed by the parametric tests.

My results also show there may be scope for improving the bandwidth selection methods that are used in density estimation by basing the choice of bandwidth on the persistence of the interest rate process. Conley, Hansen, and Liu (1997) studied different interest rate processes than those considered here and also found that persistence influences bandwidth selection.

There is a clear need for more research on the finite sample properties of nonparametric estimation methods and their performance relative to more standard methods for identifying the short rate's dynamics. Among other things, I hope future research in this area compares the bias that occurs from using nonparametric methods with the discretization biases that occur from using other methods, such as quasi-maximum likelihood estimation.

Appendix

This appendix provides detailed results on kernel density estimation when the time series of interest rates is generated under the Vasicek process. The appendix is divided into four main parts. Part A presents notation and provides results which simplify the derivations. For the kernel density estimator which uses continuous sampling [Equation (7)] part B presents results on mean integrated squared bias, mean integrated squared variance, and mean integrated squared error. Part C presents results which are similar to those in part B, but for the kernel density estimator which uses discrete sampling [Equation (5)]. Part C also contains expressions for the variance of the kernel density estimate for both discrete and continuous sampling. Part D contains formulae that are useful for computing the covariance of the density estimates.

A. Notation, Assumptions, and Four Basic Results

A.1 Assumptions

Throughout the appendix, all results are for data generated under the Vasicek (1977) model over the time interval $[0, T]$. Interest rates at time 0 are drawn from the marginal distribution of the Vasicek process [Equation (13)] and then evolve according to the process in Equation (12). The kernel function that is used in all estimation is the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-.5u^2}$.

A.2 Notation

The marginal density under the Vasicek model is $\pi(u)$; its formula is contained in Equation (13). The transition density for the Vasicek model is $\pi(r_s, s, \mid r_t, t)$. For convenience the transition density's dependence on the parameters of the Vasicek model is suppressed. The formula for the transition density is

$$\pi(r_{(s)}, s \mid r_{(t)}, t) = \frac{1}{\sqrt{2\pi V(r_{(s)} \mid r_{(t)})}} \exp^{-.5\left(\frac{r_{(s)} - \mu(r_{(s)} \mid r_{(t)})}{\sqrt{V(r_{(s)} \mid r_{(t)})}}\right)^2},$$

where

$$\begin{aligned} \mu(r_{(s)} \mid r_{(t)}) &= \theta + (r_{(t)} - \theta)e^{-\kappa(s-t)} \text{ and} \\ V(r_{(s)} \mid r_{(t)}) &= V_E(1 - e^{-2\kappa(s-t)}). \\ V_E &= \frac{\sigma^2}{2\kappa}. \end{aligned}$$

A.3 Basic results

The following four results will be stated without proof.

Result 1.

$$\pi(r_s) = \int \pi(r_s, s \mid r_t, t)\pi(r_t)\, dr_t.$$

Result 2.

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-.5(\frac{y-u}{\sigma_1})^2} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-.5(\frac{u-z}{\sigma_2})^2}\, du$$
$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}} e^{-.5\left(\frac{y-z}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)^2}.$$

Result 3.

$$\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}\sigma_1} e^{-.5(\frac{y-u}{\sigma_1})^2}\right)^2 dy = \frac{1}{\sqrt{2}\sqrt{2\pi}\sigma_1}.$$

Result 4.

$$\frac{1}{\sqrt{2\pi}\sigma_1} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-.5(\frac{u-y}{\sigma_1})^2} e^{-.5(\frac{y-\theta}{\sigma_2})^2}$$
$$= \frac{1}{\sqrt{2\pi}S_1} \frac{1}{\sqrt{2\pi}S_2} e^{-.5(\frac{y-(\lambda u+(1-\lambda)\theta)}{S_1})^2} e^{-.5(\frac{u-\theta}{S_2})^2},$$

where $S_1 = \sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}$, $S_2 = \sqrt{\sigma_1^2+\sigma_2^2}$, *and* $\lambda = \frac{\sigma_2^2}{\sigma_2^2+\sigma_1^2}$.

B. Kernel Estimation with Continuous Sampling

In this section, all results are proved for the kernel density estimator [Equation (7)] which uses continuous sampling.

B.1 The expected kernel density estimate

To compute the expected kernel density estimate, I will exploit the linearity of the integral operator and the fact that r follows a Markov process. Therefore, if r starts from r_0 at time 0, then

$$\begin{aligned} E(\hat{\pi}(u) \mid r_0) &= \frac{1}{Th}\int_0^T \left[EK\left(\frac{u-r_s}{h}\right) \mid r_0\right] ds \\ &= \frac{1}{Th}\int_0^T \left[\int_{r_s} K\left(\frac{u-r_s}{h}\right)\pi(r_s, s \mid r_0, 0)\, dr_s\right] ds. \end{aligned}$$

The unconditional expected value of the kernel density estimate is given by the expression

$$E\hat{\pi}(u) = \int_{r_0} \left(\frac{1}{Th} \int_0^T \left[\int_{r_s} K\left(\frac{u - r_s}{h}\right) \pi(r_s, s \mid r_0, 0) dr_s \right] ds \right) \pi(r_0)\, dr_0.$$

By Result 1 and Fubini's theorem this simplifies considerably to become

$$\begin{aligned} E\hat{\pi}(u) &= \int_{r_s} K\left(\frac{u - r_s}{h}\right) \pi(r_s)\, dr_s. \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{h^2 + \frac{\sigma^2}{2\kappa}}} e^{-.5\left(\frac{u-\theta}{\sqrt{h^2+\frac{\sigma^2}{2\kappa}}}\right)^2}. \end{aligned}$$

The second line of the equality follows from Result 2.

B.2 Mean integrated squared bias

The expression for the mean integrated squared bias is

$$MISB = \int_u (E\hat{\pi}(u) - \pi(u))^2\, du.$$

By plugging in the expressions for $\pi(u)$ and $E\hat{\pi}(u)$, and then using Results 2 and 3, the formula for the mean integrated squared bias simplifies to the following analytical form:

$$MISB = \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sqrt{2h^2 + \frac{\sigma^2}{\kappa}}} + \frac{1}{\sqrt{2(\frac{\sigma^2}{2\kappa})}} - \frac{2}{\sqrt{(h^2 + \frac{\sigma^2}{\kappa})}} \right).$$

B.3 Mean integrated variance

The expression for the mean integrated variance of the kernel estimate is

$$MIVAR = E \int_u \left(\hat{\pi}(u) - E\hat{\pi}(u)\right)^2\, du.$$

By Fubini's theorem, and simplification, this can be written as

$$MIVAR = \int_u \left(E\left[\hat{\pi}^2(u)\right] - (E\hat{\pi}(u))^2\right)\, du.$$

By Result 3 this expression simplifies further to become

$$MIVAR = \int_u E\left[\hat{\pi}^2(u)\right]\, du - \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sqrt{2h^2 + \frac{\sigma^2}{\kappa})}} \right).$$

It is very difficult to compute the first term in the expression for mean integrated variance. To do so, it is useful to define some auxilary expressions. Define

$$F(t) = \int_0^t K\left(\frac{u - r_s}{h}\right) ds.$$

Although the function $F(t)$ depends on the random realizations of the process r, over each increment of time dt, the function is locally deterministic with derivative[30]

$$dF(t) = K\left(\frac{u - r_t}{h}\right) dt.$$

Using elementary rules of calculus this implies

$$d(F(t)^2) = 2F(t)dF(t).$$

Therefore an alternative expression for the square of the kernel density estimate is

$$\hat{\pi}(u)^2 = \frac{1}{T^2h^2}\int_0^T 2F(t)dF(t).$$

Substituting for $F(t)$ and $dF(t)$ and rearranging yields

$$\hat{\pi}(u)^2 = \frac{2}{T^2h^2}\int_0^T \left[\int_0^t K\left(\frac{u - r_s}{h}\right) K\left(\frac{u - r_t}{h}\right) ds\right] dt.$$

The above expression is very convenient because it means one can express $E\int_u \hat{\pi}(u)^2\, du$ using the law of iterated conditional expectations, and Result 1, producing

$$E\int_u \hat{\pi}(u)^2\, du = \int_u \left\{\frac{2}{T^2h^2}\int_0^T \left(\int_0^t E\left[K\left(\frac{u - r_s}{h}\right)K\left(\frac{u - r_t}{h}\right)\right] ds\right) dt\right\} du.$$

By the law of iterated conditional expectations, and by Result 1, the

[30] In the next increment of time, F will grow by the amount

$$K\left(\frac{u - r_{(t+dt)}}{h}\right) dt.$$

Applying Itô's lemma to r_{t+dt}, the above expression can be written in Taylor series form as

$$K\left(\frac{u - r_{(t+dt)}}{h}\right) dt = K\left(\frac{u - r_t}{h}\right) dt + K_r(.)dr * dt + .5K_{rr}(.)(dt^2).$$

All of the stochastic terms in the expansion are of smaller order than dt and can therefore be dropped. As a result, F is locally deterministic with the derivative given in the text.

expectation of the term inside the inner braces can be written as

$$E\left[K\left(\frac{u-r_s}{h}\right)K\left(\frac{u-r_t}{h}\right)\right] = \int_{r_s} K\left(\frac{u-r_s}{h}\right)\left\{\int_{r_t} K\left(\frac{u-r_t}{h}\right)\pi(r_t, t \mid r_s, s)\, dr_t\right\}\pi(r_s)dr_s.$$

Substitution of the above expression into the one before it provides an expression for the first term of the MIVAR:

$$E\int_u \hat{\pi}(u)^2\, du = \frac{2}{T^2h^2}\int_u \left\{\int_0^T \left(\int_0^t \left[\int_{r_s} K\left(\frac{u-r_s}{h}\right) \times \left\{\int_{r_t} K\left(\frac{u-r_t}{h}\right)\pi(r_t, t \mid r_s, s)\, dr_t\right\}\pi(r_s)dr_s\right] ds\right) dt\right\} du.$$

Because the kernel function is Gaussian, I can use Fubini's theorem and apply Result 2 three times to produce the much more simplified expression

$$E\int_u \hat{\pi}(u)^2\, du = \frac{2}{T^2}\frac{1}{\sqrt{2\pi}}\int_0^T \left[\int_0^t \frac{1}{\sqrt{2h^2 + \frac{\sigma^2}{\kappa}(1 - e^{-\kappa(t-s)})}}\, ds\right] dt.$$

It is convenient to rewrite this expression as

$$E\int_u \hat{\pi}(u)^2\, du = \frac{2}{T^2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2h^2 + \frac{\sigma^2}{\kappa}}}\int_0^T \left[\int_0^t \frac{ds}{\sqrt{1 - \left(\frac{\frac{\sigma^2}{\kappa}}{2h^2 + \frac{\sigma^2}{\kappa}}\right)e^{-\kappa(t-s)}}}\right] dT,$$

and then make the substitution

$$z = \sqrt{\left(\frac{\frac{\sigma^2}{\kappa}}{2h^2 + \frac{\sigma^2}{\kappa}}\right)e^{-\kappa(t-s)}}.$$

The above integral simplifies further to become

$$E\int_u \hat{\pi}(u)^2\, du = \frac{2}{\kappa}\frac{2}{T^2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2h^2 + \frac{\sigma^2}{\kappa}}}\int_0^T \left[\int_{\underline{z}}^{\overline{z}} \frac{dz}{z\sqrt{1-z^2}}\right] dT,$$

where

$$\underline{z} = \sqrt{\left(\frac{\frac{\sigma^2}{\kappa}}{2h^2 + \frac{\sigma^2}{\kappa}}\right) e^{-\kappa t}} \text{ and } \overline{z} = \sqrt{\left(\frac{\frac{\sigma^2}{\kappa}}{2h^2 + \frac{\sigma^2}{\kappa}}\right)}.$$

To compute the integral in terms of z, I make the additional substitution $z = \cos(\rho)$, and then integrate and simplify producing

$$\begin{aligned} E \int_u \hat{\pi}(u)^2\, du \;=\; & \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2h^2 + \frac{\sigma^2}{\kappa}}} \\ & \times \left\{ 1 + \frac{2}{\kappa} \frac{2}{T^2} \int_0^T ln \left[\frac{1 + \sqrt{1 - \psi(t)}}{1 + \sqrt{1 - \psi(0)}} \right] dt \right\}, \end{aligned}$$

where

$$\psi(t) = \left(\frac{\frac{\sigma^2}{\kappa}}{2h^2 + \frac{\sigma^2}{\kappa}}\right) e^{-\kappa t}.$$

Substitution of the above expression as the first piece of the mean integrated variance for the Vasicek process with a Gaussian kernel produces the final result,

$$MIVAR(Vasicek) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2h^2 + \frac{\sigma^2}{\kappa}}} \frac{2}{\kappa} \frac{2}{T^2} \int_0^T ln \left[\frac{1 + \sqrt{1 - \psi(t)}}{1 + \sqrt{1 - \psi(0)}} \right] dt,$$

where $\psi(t)$ is defined above.

C. Discrete Sampling Kernel Density Estimates

In this section I will derive expressions for the first and second moments of

$$\hat{\pi}(u) = \frac{1}{Nh} \sum_{s=\Delta s}^{N\Delta s} K\left(\frac{u - r_s}{h}\right).$$

C.1 The expected kernel density estimate

To solve for the expected squared density estimate with discrete sampling is straightforward and similar to the method of solving for continuous sampling.

$$\begin{aligned} E\hat{\pi}(u) \;&=\; E \frac{1}{Nh} \sum_{s=\Delta s}^{N\Delta s} K\left(\frac{u - r_s}{h}\right) \\ &=\; \frac{1}{N} \sum_{s=\Delta s}^{N\Delta s} \int_{r_0} \left(\int_{r_s} \frac{1}{h} K\left(\frac{u - r_s}{h}\right) \pi(r_s, s \mid r_0, 0)\, dr_s \right) \pi(r_0)\, dr_0 \end{aligned}$$

$$= \frac{1}{N}\sum_{s=\Delta s}^{N\Delta s}\int_{r_s}\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\pi(r_s)\,dr_s$$

$$= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{h^2+\frac{\sigma^2}{2\kappa}}}e^{-.5\left(\frac{u-\theta}{\sqrt{h^2+\frac{\sigma^2}{2\kappa}}}\right)^2}.$$

The second equality is by the law of iterated conditional expectations. The third equality is by Fubini's theorem and Result 1. The final equality follows by Result 2.

C.2 The expected squared kernel density estimate

The expression for the expected square of the kernel density can be written as

$$\begin{aligned}
E\hat{\pi}(u)^2 &= E\left(\frac{1}{Nh}\sum_{s=\Delta s}^{N\Delta s}K(\frac{u-r_s}{h})\right)^2\\
&= \left(\frac{1}{N}\right)^2\left(\sum_{s=\Delta s}^{N\Delta s}E\left[\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\right]^2\right.\\
&\qquad\left.+\,2\sum_{t=\Delta t}^{N\Delta t}\sum_{s=\Delta s}^{t-\Delta t}E\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\frac{1}{h}K\left(\frac{u-r_t}{h}\right)\right).
\end{aligned}$$

The first piece of this expectation is relatively easy to find. Basic algebra shows the square of a Gaussian kernel is

$$\begin{aligned}
\left[\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\right]^2 &= \frac{1}{\sqrt{2\pi}h}\frac{1}{\sqrt{2\pi}h}e^{-.5\left(\frac{u-r_s}{\frac{h}{\sqrt{2}}}\right)^2}\\
&= \frac{1}{h\sqrt{2}\sqrt{2\pi}}\frac{1}{\frac{h}{\sqrt{2}}}K\left(\frac{u-r_s}{\frac{h}{\sqrt{2}}}\right).
\end{aligned}$$

The above expression simplifies by applying the formula for $E\hat{\pi}(u)$, producing the answer

$$\begin{aligned}
E\left[\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\right]^2 &= \frac{1}{h\sqrt{2}\sqrt{2\pi}}E\frac{1}{\frac{h}{\sqrt{2}}}K\left(\frac{u-r_s}{\frac{h}{\sqrt{2}}}\right)\\
&= \frac{1}{h\sqrt{2}\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}\sqrt{\frac{h^2}{2}+\frac{\sigma^2}{2\kappa}}}e^{-.5\left(\frac{u-\theta}{\sqrt{\frac{h^2}{2}+\frac{\sigma^2}{2\kappa}}}\right)^2}.
\end{aligned}$$

The second piece of the expectation is more difficult because it involves a double sum of terms with general form

$$
\begin{aligned}
&E\left[\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\frac{1}{h}K\left(\frac{u-r_t}{h}\right)\right]\\
&= \int_{r_0}\left(\int_{r_s}\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\right.\\
&\quad\left.\times\left\{\int_{r_t}\frac{1}{h}K\left(\frac{u-r_t}{h}\right)\pi(r_t,t\mid r_s,s)\,dr_t\right)\pi(r_s,s\mid r_0,0)\,dr_s\right)\pi(r_0)\,dr_0.\\
&= \int_{r_s}\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\left\{\int_{r_t}\frac{1}{h}K\left(\frac{u-r_t}{h}\right)\pi(r_t,t\mid r_s,s)\,dr_t\right\}\pi(r_s)\,dr_s.\\
&= \int_{r_s}\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\frac{1}{\sqrt{2\pi}}\\
&\qquad\times\ \frac{1}{\sqrt{h^2+V(r_t,t\mid r_s,s)}}\exp^{-.5\left(\frac{u-\mu(r_t,t\mid r_s,s)}{\sqrt{h^2+V(r_t,t\mid r_s,s)}}\right)^2}\pi(r_s)\,dr_s,
\end{aligned}
$$

where $V(r_t, t \mid r_s, s)$ is the variance of r_t given r_s, and $\mu(r_t, t \mid r_s, s)$ is the mean of r_t given r_s.

The first equality above follows by the law of iterated expectations; the second equality follows from Result 1; and the third equality follows from Result 2. The expression simplifies to become the product of three Gaussian density functions multiplied by some additional factors. This makes it possible to apply additional results to further simplify the integral. Substitution of the expressions for $V(r_t, t \mid r_s, s)$ and $\mu(r_t, t \mid r_s, s)$, and some algebra makes it possible to represent the above integral as the product of three Gaussian density functions, each of which involve r_s minus some mean divided by some variance. One can then apply Result 4 to any two of these density functions. The result expression contains the product of only two Gaussian density functions that involve r_s. Result 2 can be applied to these functions to integrate out r_s. This produces a final complicated expression which, for the sake of notational compactness, will be written as $G(\kappa, t-s, V_E, h)$:

$$
\begin{aligned}
G(\kappa,t-s,V_E,h) &= E\left[\frac{1}{h}K\left(\frac{u-r_s}{h}\right)\frac{1}{h}K\left(\frac{u-r_t}{h}\right)\right]\\
&= C(\kappa,t-s,V_E,h)\frac{1}{\sqrt{2\pi}\,V_1(\kappa,t-s,V_E,h)}\\
&\quad\times e^{-.5\left(\frac{u-\theta}{V_1(\kappa,t-s,V_E,h)}\right)^2}\frac{1}{\sqrt{2\pi}}\frac{1}{V_2(h,V_E)}e^{-.5\left(\frac{u-\theta}{V_2(h,V_E)}\right)^2},
\end{aligned}
$$

where

$$V_E = \frac{\sigma^2}{2\kappa}$$

$$C(\kappa, t-s, V_E, h) = \frac{e^{\kappa(t-s)}}{e^{\kappa(t-s)} - \frac{V_E}{V_E+h^2}}$$

$$V_1(\kappa, t-s, V_E, h) = \sqrt{\frac{(h^2+V_E)e^{2\kappa(t-s)} - V_E + \frac{h^2 V_E}{h^2+V_E}}{\left(e^{\kappa(t-s)} - \frac{V_E}{V_E+h^2}\right)^2}}$$

$$V_2(h, V_E) = \sqrt{h^2+V_E}.$$

Although this expression is onerous, it only involves a leading term multiplied by the product of two Gaussian density functions. More importantly, it is fully analytical, which means it is possible to generate (for a given bandwidth) an analytical expression for the variance of the kernel density estimate at each point u.

C.3 Variance of the discrete sampling kernel density estimate

The variance of $\hat{\pi}(u)$ is given by $E[\hat{\pi}(u)]^2 - [E\hat{\pi}(u)]^2$. This variance has the following form:

$$\begin{aligned}
\operatorname{var}[\hat{\pi}(u)] = \left(\frac{1}{N^2}\right) \Bigg(& \frac{N}{h\sqrt{2}\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}\sqrt{\frac{h^2}{2}+\frac{\sigma^2}{2\kappa}}} \exp^{-.5\left(\frac{u-\theta}{\sqrt{\frac{h^2}{2}+\frac{\sigma^2}{2\kappa}}}\right)^2} \\
& + 2\sum_{t=1}^{N}\sum_{s=1}^{t-1} G(\kappa, t-s, V_E, h) \Bigg) \\
& - \frac{1}{\sqrt{2}\sqrt{2\pi}\sqrt{V_E+h^2}} \frac{1}{\sqrt{2\pi}\sqrt{\frac{V_E+h^2}{2}}} \exp^{-.5\left(\frac{u-\theta}{\sqrt{\frac{V_E+h^2}{2}}}\right)^2},
\end{aligned}$$

where $G(\kappa, t-s, V_E, h)$ is from Section C.2.

C.4 Variance of the continuous sampling kernel density estimate

To compute the variance of the continuous sampling kernel density estimate, it suffices to take limits of the variance for the discrete sampling kernel density estimate as the sampling interval goes to zero. Define sampling interval Δt and Δs so that

$$\frac{1}{N}=\frac{\Delta t}{T}=\frac{\Delta s}{T},$$

and redefine $G(.)$ to remove the discreteness captured by the term ΔT, that is, define $G^*(\kappa, t-s, V_E, h)$ as

$$\begin{aligned} G^*(\kappa, t-s, V_E, h) \;=\; & C^*(\kappa, t-s, V_E, h)\frac{1}{\sqrt{2\pi}V_1^*(\kappa, t-s, V_E, h)} \\ & \times e^{-.5\left(\frac{u-\theta}{V_1^*(\kappa, t-s, V_E, h)}\right)^2}\frac{1}{\sqrt{2\pi}}\frac{1}{V_2(h, V_E)}e^{-.5\left(\frac{u-\theta}{V_2(h, V_E)}\right)^2}, \end{aligned}$$

where

$$\begin{aligned} V_E \;&=\; \frac{\sigma^2}{2\kappa} \\ C^*(\kappa, t-s, V_E, h) \;&=\; \frac{e^{\kappa(t-s)}}{e^{\kappa(t-s)}-\frac{V_E}{V_E+h^2}} \\ V_1^*(\kappa, t-s, V_E, h) \;&=\; \sqrt{\frac{(h^2+V_E)e^{2\kappa(t-s)}-V_E+\frac{h^2V_E}{h^2+V_E}}{\left(e^{\kappa(t-s)}-\frac{V_E}{V_E+h^2}\right)^2}} \\ V_2(h, V_E) \;&=\; \sqrt{h^2+V_E}. \end{aligned}$$

After applying this change in notation, the variance of the continous sampling kernel density estimate can be expressed as

$$\begin{aligned} \lim_{\Delta t\to 0}\lim_{\Delta s\to 0}\left(\frac{1}{T^2}\right) & \left(\frac{T\Delta s\Delta t}{h\sqrt{2}\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}\sqrt{\frac{h^2}{2}+\frac{\sigma^2}{2\kappa}}}\exp^{-.5\left(\frac{u-\theta}{\sqrt{\frac{h^2}{2}+\frac{\sigma^2}{2\kappa}}}\right)^2}\right. \\ & \left. +\;2\sum_{t=\Delta t}^{T}\left\{\sum_{s=\Delta s}^{t-\Delta s}G^*(\kappa, t-s, V_E, h)\Delta s\right\}\Delta t\right) \\ & -\frac{1}{\sqrt{2}\sqrt{2\pi}\sqrt{V_E+h^2}}\frac{1}{\sqrt{2\pi}\sqrt{\frac{V_E+h^2}{2}}}\exp^{-.5\left(\frac{u-\theta}{\sqrt{\frac{V_E+h^2}{2}}}\right)^2}. \end{aligned}$$

The first term inside the large parenthesis goes to zero in the limit; the second term converges to a double integral. The term outside the parenthesis remains unchanged. Therefore the final result for the variance of the continuous sampling kernel density estimate is

$$\begin{aligned}\text{var}[\hat{\pi}(u)] \;=\; & \frac{2}{T^2}\int_{t=0}^{T}\left\{\int_{s=0}^{t} G^*(\kappa, t-s, V_E, h)\, ds\right\} dt \\ & -\frac{1}{\sqrt{2}\sqrt{2\pi}\sqrt{V_E+h^2}}\frac{1}{\sqrt{2\pi}\sqrt{\frac{V_E+h^2}{2}}}\exp^{-.5\left(\frac{u-\theta}{\sqrt{\frac{V_E+h^2}{2}}}\right)^2}.\end{aligned}$$

Unfortunately this integral is too difficult to reduce further.

C.5 MIVAR of discrete sampling kernel density estimate

The mean integrated variance of the kernel density estimate is simply the integral of the variance with respect to u. The first and third terms of the expressions for variance are trivial to integrate. The middle term involves the product of two Gaussian densities and hence can be integrated using Result 2. Simplification produces the result that is contained in Proposition 2.

C.6 Mean integrated squared bias

Because the expected kernel density estimate is the same in both the continuous and discrete sampling cases, the mean integrated squared bias is also the same in both cases and is presented in the results for the kernel density estimate with continuous sampling.

D. Covariance of the Kernel Density Estimates

For kernel density estimates $\hat{\pi}(u)$ and $\hat{\pi}(v)$ of the marginal distribution of r at points u and v, this section derives expressions for $\text{cov}(\hat{\pi}(u), \hat{\pi}(v))$.

D.1 Covariance with discrete sampling kernel estimate

The formula for the covariance is

$$\text{cov}(\hat{\pi}(u), \hat{\pi}(v)) = E[\hat{\pi}(u)\hat{\pi}(v)] - [E\hat{\pi}(u)][E\hat{\pi}(v)].$$

The only piece of this formula that has not been solved for is $E[\hat{\pi}(u)\hat{\pi}(v)]$. It can be expressed as

$$\begin{aligned}E[\hat{\pi}(u)\hat{\pi}(v)] \;=\; & E\left(\frac{1}{N}\sum_{t=1}^{N}\frac{1}{h}K\left(\frac{u-r_t}{h}\right)\frac{1}{N}\sum_{s=1}^{N}\frac{1}{h}K\left(\frac{v-r_s}{h}\right)\right) \\ =\; & \frac{1}{N^2h^2}E\left(\sum_{t=1}^{N}K\left(\frac{u-r_t}{h}\right)K\left(\frac{v-r_t}{h}\right)\right.\end{aligned}$$

$$+\sum_{t=1}^{N} K\left(\frac{u-r_t}{h}\right)\sum_{s=1}^{t-1} K\left(\frac{v-r_s}{h}\right)$$
$$\left.+\sum_{t=1}^{N} K\left(\frac{v-r_t}{h}\right)\sum_{s=1}^{t-1} K\left(\frac{u-r_s}{h}\right)\right).$$

The right-hand side of this expression involves three sets of terms. However, the second and third set are virtually identical since one can be obtained from the other by switching u and v. Therefore it is only necessary to compute

$$E\left(K\left(\frac{u-r_t}{h}\right)K\left(\frac{v-r_t}{h}\right)\right) \text{ and } E\left(K\left(\frac{u-r_t}{h}\right)K\left(\frac{v-r_s}{h}\right)\right).$$

Using Results 1 and 4, I find that

$$\begin{aligned} E\left(K\left(\frac{u-r_t}{h}\right)K\left(\frac{v-r_t}{h}\right)\right) &= J_1(u, v, h, \theta, V_E) \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\frac{h^2}{2}+V_E}} \exp^{-.5\left(\frac{.5(u+v)-\theta}{\sqrt{\frac{h^2}{2}+V_E}}\right)^2} \\ &\quad \times \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2h^2}} \exp^{-.5\left(\frac{u-v}{\sqrt{2h^2}}\right)^2}. \end{aligned}$$

Similarly, using Results 1 and 4, I find that

$$\begin{aligned} &E\left(K\left(\frac{u-r_t}{h}\right)K\left(\frac{v-r_s}{h}\right)\right) \\ &= J_2(u, v, h, \theta, V_E, \kappa, (t-s)) \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{h^2+V_E}} \exp^{-.5\left(\frac{v-\theta}{\sqrt{h^2+V_E}}\right)^2} \\ &\quad \times \frac{1}{\sqrt{2\pi}}\frac{e^{\kappa(t-s)}}{\sqrt{\left(\frac{h^2V_E}{h^2+V_E}\right)+(h^2+V_E)e^{2\kappa(t-s)}-V_E}} \\ &\quad \times \exp^{-.5\left(\frac{(u-\theta)e^{\kappa(t-s)}+(\theta-v)\frac{V_E}{h^2+V_E}}{\sqrt{\left(\frac{h^2V_E}{h^2+V_E}\right)+(h^2+V_E)e^{2\kappa(t-s)}-V_E}}\right)^2}. \end{aligned}$$

Therefore

$$E[\hat{\pi}(u)\hat{\pi}(v)] = \frac{1}{N^2}\left(\sum_{t=1}^{N} J_1(u, v, h, \theta, V_E) + \sum_{t=1}^{N}\sum_{s=1}^{t-1} J_2(u, v, h, \theta, V_E, \kappa, (t-s)) + \sum_{t=1}^{N}\sum_{s=1}^{t-1} J_2(v, u, h, \theta, V_E, \kappa, (t-s))\right).$$

When $[E\hat{\pi}(u)][E\hat{\pi}(v)]$ is subtracted from the above expression, it produces the final result for the covariance function.

D.2 Covariance with continuous sampling

To compute the covariance for continuous sampling I will follow the same procedure that I used to compute the variance for continuous sampling. I begin by taking limits of the covariance for discrete sampling as the time between observations goes to zero. I then change notation to represent continous time sampling. This yields

$$E[\hat{\pi}(u)\hat{\pi}(v)] = \frac{1}{T^2}\left(\int_{t=0}^{T}\int_{s=0}^{t} J_2^*(u, v, h, \theta, V_E, \kappa, (t-s))\, ds\, dt + \int_{t=0}^{T}\int_{s=0}^{t} J_2^*(v, u, h, \theta, V_E, \kappa, (t-s))\, ds\, dt\right).$$

References

Aït-Sahalia, Y., 1996a, "Do Interest Rates Really Follow Continuous-Time Markov Diffusions?" working paper, University of Chicago.

Aït-Sahalia, Y., 1996b, "Nonparametric Pricing of Interest Rate Derivative Securities," *Econometrica*, 64, 527–560.

Aït-Sahalia, Y., 1996c, "Testing Continuous-Time Models of the Spot Interest Rate," *Review of Financial Studies*, 9, 385–426.

Andersen, T. G., and J. Lund, 1997, "Continuous-Time Stochastic Volatility Models of the Short-Term Interest Rate," *Journal of Econometrics*, 77, 343–377.

Ball, C. A., and W. N. Torous, 1996, "Unit Roots and the Estimation of Interest Rate Dynamics," *Journal of Empirical Finance*, 3, 215–238.

Bekaert, G., R. J. Hodrick, and D. A. Marshall, 1997, "On Biases in Tests of the Expectations Hypothesis of the Term Structure of Interest Rates," *Journal Financial Economics*, 44, 309–348.

Castellana, J. V., and M. R. Leadbetter, 1986, "On Smoothed Probability Density Estimation for Stationary Processes," *Stochastic Processes and their Applications*, 21, 179–193.

Chan, K. C., et al., 1992, "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate," *Journal of Finance*, 47, 1209–1227.

Chen, R-R, and L. Scott, 1993, "Maximum Likelihood Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates," *Journal of Fixed Income*, 3, 14–31.

Conley, T. G., L. P. Hansen, and W. F. Liu, 1997, "Bootstrapping the Long Run," *Macroeconomic Dynamics*, 1, 279–311.

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

Cox, J. C., J. E. Ingersoll, and S. A. Ross, 1985, "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–407.

David, H. A., 1981, *Order Statistics*, John Wiley and Sons, New York.

Eom, Y. H., 1998, "On Efficient GMM Estimation of Continuous-Time Asset Dynamics: Implications for the Term Structure of Interest Rates," working paper, Federal Reserve Bank of New York.

Gallant, A. R., and G. Tauchen, 1995, "Estimation of Continuous-Time Models for Stock Returns and Interest Rates," working paper, University of North Carolina at Chapel Hill.

Hardle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Hart, J. D., and P. Vieu, 1990, "Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data," *Annals of Statistics* 18, 873–890.

Jones, M. C., and M. P. Wand, 1995, *Kernel Smoothing*, Chapman and Hall, London.

Karlin, S., and H. M. Taylor, 1981, *A Second Course in Stochastic Processes*, Academic Press, New York.

McCulloch, J. H., and H. L. Kwon, 1993, "U.S. Term Structure Data, 1947–1991," Working Paper 93-6, Ohio State University.

Newey, W. K., 1985a, "Generalized Method of Moments Specification Testing," *Journal of Econometrics*, 29, 229–256.

Newey, W. K., 1985b, "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 53, 1047–1070.

Newey, W. K., and K. D. West, 1987, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Pagan, A. R., A. D. Hall, and V. Martin, 1995, "Modeling the Term Structure," working paper, Australia National University.

Park, B. U., and J. S. Marron, 1990, "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85, 66–72.

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, 1279–1304.

Pritsker, M. G., 1997, "Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models," Working Paper 1997–26, Finance and Economic Discussion Series, Federal Reserve Board.

Robinson, P. M., 1983, "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis*, 4, 185–207.

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 Statistical Mathematics*, 38, 539–549.

Scott, D. W., 1992, *Multivariate Density Estimation, Theory, Practice, and Visualization*, John Wiley and Sons, New York.

Siddique, A. R., 1994, "Nonparametric Estimation of Mean and Variance and Pricing of Securities," working paper, Georgetown University.

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

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

Stone, C. J., 1984, "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *Annals of Statistics*, 12, 1285–1297.

Vasicek, Oldrich, 1977, "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics*, 5, 177–188.

Wand, M. P., 1992, "Finite Sample Performance of Density Estimators under Moving Average Dependence," *Statistics and Probability Letters*, 13, 109–115.