Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes

Garland B. Durham
Department of Economics, University of Iowa, Iowa City, IA 52242-1000 (*garland-durham@uiowa.edu*)

A. Ronald Gallant
Department of Economics, University of North Carolina, Chapel Hill, NC 27599-3305 (*ron_gallant@unc.edu*)

Stochastic differential equations often provide a convenient way to describe the dynamics of economic and financial data, and a great deal of effort has been expended searching for efficient ways to estimate models based on them. Maximum likelihood is typically the estimator of choice; however, since the transition density is generally unknown, one is forced to approximate it. The simulation-based approach suggested by Pedersen (1995) has great theoretical appeal, but previously available implementations have been computationally costly. We examine a variety of numerical techniques designed to improve the performance of this approach. Synthetic data generated by a Cox-Ingersoll-Ross model with parameters calibrated to match monthly observations of the U.S. short-term interest rate are used as a test case. Since the likelihood function of this process is known, the quality of the approximations can be easily evaluated. On datasets with 1,000 observations, we are able to approximate the maximum likelihood estimator with negligible error in well under 1 min. This represents something on the order of a 10,000-fold reduction in computational effort as compared to implementations without these enhancements. With other parameter settings designed to stress the methodology, performance remains strong. These ideas are easily generalized to multivariate settings and (with some additional work) to latent variable models. To illustrate, we estimate a simple stochastic volatility model of the U.S. short-term interest rate.

Stochastic differential equations (SDE's) often provide a convenient way to model economic and financial data, and their use has become increasingly common in recent years. Although the process specified by a stochastic differential equation is defined in continuous time, the data which are available are typically sampled at discrete time intervals. The resulting estimation problem turns out to be nontrivial, and considerable energy has been expended in developing computationally (and statistically) efficient estimation schemes.

In this article, we focus primarily on scalar, time-homogeneous processes. In particular, we consider the diffusion process generated by an SDE of the form

$$
\begin{aligned}
dX &= \mu(X;\theta)\,dt + \sigma(X;\theta)\,dW \\
X(t_0) &= X_0
\end{aligned}
\tag{1}
$$

with parameter vector θ. Suppose that the sample $\{X_i = X(t_i), i = 0, \ldots, n\}$ is available for analysis. The observations need not be equally spaced.

Ideally, one would like to know the transition density, which would allow one to compute the maximum likelihood estimator with its usual optimality properties. Although exact transition densities are known in only a few isolated cases, several approaches toward approximating the transition density have been proposed.

Lo (1988) suggests numerically solving the Fokker–Planck partial differential equation for each observation. Pedersen (1995b) suggests a simulation-based approach which involves integrating out unobserved states of the process at intermediate points between each pair of observations (see also Santa-Clara 1995; Brandt and Santa-Clara 2002). While this approach, commonly known as simulated maximum likelihood estimation (SMLE), is able to come arbitrarily close to the true transition density, previously available implementations have been computationally burdensome.

Other approaches have been proposed which are much less computationally costly. For example, the process described by (1) has a first-order approximation given by the discrete-time process

$$
\begin{aligned}
\widetilde{X}_{i+1} &= \widetilde{X}_i + \mu(\widetilde{X}_i;\theta)\Delta_i + \sigma(\widetilde{X}_i;\theta)\Delta_i^{1/2}\epsilon_i \\
\Delta_i &= t_{i+1} - t_i, \qquad \epsilon_i \sim N(0,1).
\end{aligned}
\tag{2}
$$

Under mild regularity conditions, the maximum likelihood estimator based on this approximation is known to converge to the true maximum likelihood estimator as the sampling interval goes to zero (Florens-Zmirou 1989). While this approach is very appealing from a computational viewpoint, the approximation may not be sufficiently accurate for the sampling frequencies at which reliable data are available.

There are various ways in which one might improve upon this idea. Elerian (1998) suggests replacing the Gaussian density in (2) by a noncentral chi-squared density which is derived from the Milstein scheme, an order 2.0 weak approximation to the true process. Shoji and Ozaki (1998) linearize the SDE, obtaining an approximating Ornstein–Uhlenbeck process (the exact transition density of an Ornstein–Uhlenbeck process is known). Kessler (1997) approximates the transition function by a Gaussian density with first and second moments obtained from higher order Ito–Taylor expansions. Aït-Sahalia (2001) approximates the transition density using a Hermite function with coefficients obtained using higher order Ito–Taylor expansions. Except for Aït-Sahalia (2001), these methods still require the sampling interval to go to zero to obtain convergence to the true transition density. While this requirement also holds for Aït-Sahalia's approach with a Hermite function and Ito–Taylor expansion of fixed order, Aït-Sahalia's approximation may be made arbitrarily accurate with fixed sampling frequency by using a Hermite function and Ito–Taylor expansion of sufficiently high order (given some regularity conditions).

Various method-of-moments approaches have also been proposed. Chan, Karolyi, Longstaff, and Sanders (1992) use moments based on Equation (2). Duffie and Singleton (1993), Gallant and Tauchen (1997), Bibby and Sørensen (1995), and Gouriéroux, Monfort, and Renault (1993) compute expectations using simulation-based methods. Hansen and Scheinkman (1995) and Duffie and Glynn (1996) use moment conditions obtained from the infinitesimal generator.

The simulation-based methods can be computationally costly, but have the advantage of being easily adapted to diffusions with unobserved state variables. Stochastic volatility models and term structure models are important applications where these techniques have been found useful. The efficient method of moments proposed by Gallant and Tauchen (1996) approaches the efficiency of maximum likelihood asymptotically, and provides a convenient set of diagnostic measures for model specification.

Markov chain Monte Carlo (MCMC) methods have been proposed by Eraker (2001), Jones (1999a), and Elerian, Chib, and Shephard (2001). There is a close relationship between MCMC methods and SMLE. For example, Elerian et al. point out that their importance sampler can also be used with the simulation-based approach of Pedersen (1995b) to substantially reduce the computational effort required to obtain reasonably accurate likelihood approximations.

In this article, we focus on the SMLE approach. The basic idea is quite simple. Suppose that one wishes to obtain the transition density $p(x_t, t; x_s, s)$. The first-order approximation $p^{(1)}(x_t, t; x_s, s)$ defined by (2) will be accurate if the interval $[s, t]$ is sufficiently short. Otherwise, one may partition the interval $s = \tau_1 < \tau_2 < \cdots < \tau_M = t$ such that the first-order approximation is sufficiently accurate on each subinterval. The random variables $X(\tau_1), \ldots, X(\tau_{M-1})$ are, of course, unobserved, and must be integrated out. Because the process is Markovian, one obtains

$$p(x_t, t; x_s, s) \approx p^{(M)}(x_t, t; x_s, s) \tag{3}$$

$$\equiv \int \prod_{m=0}^{M-1} p^{(1)}(u_{m+1}, \tau_{m+1}; u_m, \tau_m) \times d\lambda(u_1, \ldots, u_{M-1}) \tag{4}$$

where λ denotes the Lebesgue measure, and we use the convention $u_0 = x_s$ and $u_M = x_t$ to conserve notation. Monte Carlo integration is generally the only feasible way to evaluate this integral.

The theoretical issues involved with this approach are already reasonably well understood. Sufficient conditions for the approximation in (3) to converge are known. While it is certainly of value to extend these conditions, we do not undertake this task here. The theories of Monte Carlo integration and maximum likelihood estimation have also been extensively studied. Nonetheless, although the simulation-based approach is attractive from a theoretical point of view, the computational burden associated with previous implementations has hindered its widespread use. We have found that it can be quite costly to attain even the degree of accuracy provided by the simple first-order approximation (2). It is this shortcoming which we seek to address.

We attack the problem of computational efficiency from two directions. We first seek to improve the approximation in Equation (3). This allows one to attain a given level of accuracy with fewer intermediate points. We consider extrapolation techniques and the use of alternatives to the first-order (Euler) approximation of the subtransition densities. Secondly, we examine techniques to accelerate the convergence of the Monte Carlo integration. We consider several importance samplers and random schemes. Finally, we consider transforming the model in such a way as to make the volatility function constant. Working with the transformed rather than the original model turns out to provide a useful improvement in both the accuracy of the approximation (3) as well as the performance of the Monte Carlo integration used to compute (4).

As a test case, we use the square-root specification proposed by Cox, Ingersoll, and Ross (1985) as a model for the short-term interest rate. Parameter settings are calibrated to match monthly observations of the U.S. short-term interest rate. This model has the advantage that the transition density is available in closed form, which allows us to easily evaluate the accuracy of our approximations. We also tested our techniques using other parameter settings and models with similar results.

On simulated datasets of 1,000 observations, we are able to obtain estimates in well under 1 min (running FORTRAN code on a 750 MHz PC) which differ negligibly from those obtained by maximizing the exact log-likelihood function. Achieving comparable accuracy without our acceleration techniques would require something on the order of a 10,000-fold increase in computational effort.

Much of the discussion in this article may be readily adapted to the multivariate setting. With some additional work, the ideas can also be extended to latent variable models. We outline an approach to approximating the transition density of a continuous-time stochastic volatility model, and illustrate by estimating a simple model over weekly observations of the U.S. treasury bill rate. We speculate that much carries over to the time-inhomogeneous case as well; however, we have not examined such extensions carefully. Although it should be possible to apply techniques similar to those considered here to jump diffusions, this is also beyond the scope of this article.

A more extensive application illustrating the techniques discussed in this article may be found in Durham (2000). Further exploration of these and related techniques in multivariate and latent variable settings is underway.

The structure of this article is as follows. Section 1 introduces the notation, and provides some theoretical results, Section 2 describes the benchmarks which we will use for evaluation of our techniques, Section 3 examines the performance of the simulation-based method without any of our acceleration techniques, Section 4 considers the issue of bias reduction, Section 5 considers the issue of variance reduction, Section 6 discusses the results of our numerical experiments, Section 7 extends these ideas to the stochastic volatility model, Section 8 provides an application, and Section 9 concludes.

1. BACKGROUND

To begin, we define some notation, and provide a brief discussion of the theoretical framework. Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let W be a Brownian motion defined on it. Let $\{\mathcal{F}_t, t \geq 0\}$ be the filtration generated by W and augmented by the P-null sets of $\mathcal{F}$. Let Θ be a compact subset of $\mathbb{R}^d$. We are interested in the parameterized family of scalar diffusion processes $\{X(t; \theta), \theta \in \Theta\}$ generated by the time-homogeneous SDE

$$dX = \mu(X; \theta)\, dt + \sigma(X; \theta)\, dW$$
$$X(t_0; \theta) = X_0.$$

Assumption 1. For each $\theta \in \Theta$, (5) has a nonexploding, unique weak solution.

By nonexploding, we mean that there is zero probability that the process diverges to infinity over any fixed time interval. Sufficient conditions ensuring Assumption 1 are well known (e.g., Karatzas and Shreve 1991, sec. 5.5). For example, it suffices that μ and σ satisfy global Lipschitz and linear growth conditions. A variety of extensions is also available. Explosiveness would preclude the existence of a transition density, and is thus disallowed. Note that stationarity is not required.

For $s < t$, suppose that $X(t; \theta) | X(s; \theta)$ has a transition density $p(x_t, t; x_s, s, \theta)$, and let

$$p^{(1)}(x_t, t; x_s, s, \theta) \\ = \phi\big(x_t; x_s + \mu(x_s)(t-s), \sigma^2(x_s)(t-s)\big), \quad (5)$$

where $\phi(x; \mu, \sigma^2)$ is the Gaussian density, be its first-order approximation. Let $s = \tau_0 < \cdots < \tau_M = t$ be a partition of the interval $[s, t]$, and let

$$p^{(M)}(x_t, t; x_s, s, \theta) \\ = \int \prod_{m=1}^{M} p^{(1)}(u_m, \tau_m; u_{m-1}, \tau_{m-1}, \theta)\, d\lambda(u_1, \ldots, u_{M-1}) \quad (6)$$

where $u_0 = x_s$, $u_M = x_t$, and λ denotes the Lebesgue measure. This will serve as our approximating density. For clarity, we will often refer to $p^{(1)}(\cdot)$ as a *subtransition density* (or occasionally simply *subdensity*) when used in this context.

Suppose that one has a set of observations $\{X_i = X(t_i; \theta^o), i = 0, \ldots, n\}$ of the process generated by (5) with unknown parameter vector θ^o, and let $P_{\theta^o, n}$ denote the probability measure induced by $\{X_0, \ldots, X_n\}$. Let $l_n(\theta) = \sum_{i=1}^{n} \log p(X_i, t_i; X_{i-1}, t_{i-1}, \theta)$ and $l_n^{(M)}(\theta) = \sum_{i=1}^{n} \log p^{(M)}(X_i, t_i; X_{i-1}, t_{i-1}, \theta)$ denote the log-likelihood functions associated with the exact and approximate densities, respectively.

Assumption 2. For all $s < t$, x_s in the support of $X(s; \theta^o)$, $\theta \in \Theta$. And $M \geq 1$, the densities $p(\cdot, t; x_s, s, \theta)$ and $p^{(M)}(\cdot, t; x_s, s, \theta)$ exist.

Pedersen (1995a) provides sufficient conditions for Assumption (2) to hold, as well as regularity conditions ensuring that

$$\lim_{M \to \infty} p^{(M)}(\cdot, t; x_s, s, \theta) = p(\cdot, t; x_s, s, \theta) \qquad \text{in } L^1(\lambda). \quad (7)$$

We note that Pedersen's results are obtained for multivariate processes. Pedersen's Theorem 2 allows for time-inhomogeneous processes. While this theorem requires a constant diffusion function, we will see in Section 2 that, for scalar processes at least, this does not impose a material constraint. Pedersen's Theorem 3 allows for a variable diffusion function, but imposes other conditions.

Although Pedersen's results assume Lipschitz and linear growth conditions on $\mu(\cdot)$ and $\sigma(\cdot)$ that are not satisfied for many applications of economic interest (including notably the CIR square root process), we speculate that suitable extensions should be possible using localization arguments along the lines of, for example, Karatzas and Shreve (1991, thm. 5.2.5). Similarly, we will examine subtransition densities other than the simple first-order approximation shown in (5) (see Section 4) and alternative random number schemes (see Section 5) which are not covered by Pedersen's results. Again, these extensions seem plausible, but formal justification is left for future work. The goal of this article is practical rather than theoretical, and Pedersen's results will serve as a convenient starting point. In particular, we assume the following, which Pedersen's Theorem 4 shows to be an immediate consequence of (7).

Assumption 3. For each $\theta \in \Theta$,

$$\lim_{M \to \infty} l_n^{(M)}(\theta) = l_n(\theta) \qquad \text{in probability under } P_{\theta^o, n}.$$

The difficulty is how to efficiently evaluate the integral in Equation (6). Monte Carlo integration is generally the only feasible approach. To perform Monte Carlo integration, one requires an importance sampler. Fix $s < t$, x_s, x_t, θ, and M, and let $q(u_1, \dots, u_{M-1})$ denote a probability density on $\mathbb{R}^{M-1}$. This will be our importance sampler. Some techniques for constructing efficient importance samplers are discussed in Section 5.

Let $\{\mathbf{u}_k = (u_{k,1}, \dots, u_{k,M-1}), k = 1, \dots, K\}$ be independent draws from q, and let

$$p^{(M,K)}(x_t, t; x_s, s, \theta) = \frac{1}{K} \sum_{k=1}^{K} \frac{\prod_{m=1}^{M} p^{(1)}(u_{k,m}, \tau_m; u_{k,m-1}, \tau_{m-1}, \theta)}{q(u_{k,1}, \dots, u_{k,M-1})} \quad (8)$$

where $u_{k,0} = x_s$ and $u_{k,M} = x_t$ for all k. Then, given Assumption 4 below, the strong law of large numbers implies that

$$\lim_{K \to \infty} \left| p^{(M,K)}(x_t, t; x_s, s, \theta) - p^{(M)}(x_t, t; x_s, s, \theta) \right| = 0 \text{ a.s.} \quad (9)$$

A somewhat stronger condition provides $\sqrt{n}$ convergence [see Geweke (1989)].

Assumption 4. Let $U_0 = x_s$, $U_M = x_t$, $\theta \in \Theta$, and q be fixed, and let $(U_1, \dots, U_{M-1})$ be a random vector with density q. Then

$$\mathrm{E}\left[\frac{\prod_{m=1}^{M} p^{(1)}(U_m, \tau_m; U_{m-1}, \tau_{m-1}, \theta)}{q(U_1, \dots, U_{M-1})} \right] < \infty.$$

Our goal is to approximate $\log l_n(\theta)$ for a given realization of the process. For this, it will suffice to be able to approximate $p(x_t, t; x_s, s, \theta)$ for arbitrary $s < t$, x_s, x_t, and θ. If we can do this with arbitrary precision, and if the log-likelihood function is continuous and Θ is compact, then we can evaluate the maximum likelihood estimator at this realization with any desired level of accuracy. We do not treat the estimator obtained by optimizing the approximate log-likelihood with a fixed setting of the tuning parameters as an object of independent interest.

2. BENCHMARKS

The specification

$$dX = \theta_2(\theta_1 - X)\,dt + \theta_3\sqrt{X}\,dW \quad (10)$$

with θ_1, θ_2, and θ_3 positive was proposed by Cox et al. (1985) to model short-term interest rates. Since this model has a known transition density and is frequently used in applications, it provides a convenient means of evaluating the effectiveness of our numerical methods. If we let $\Delta = t - s > 0$, $c = 2\theta_2/[\theta_3^2(1 - e^{-\theta_2\Delta})]$, and $Y = 2cX$, then $Y_t|Y_s$ is distributed as noncentral chi-squared with $4\theta_2\theta_1/\theta_3^2$ degrees of freedom and noncentrality parameter $Y_s e^{-\theta_2\Delta}$ or, equivalently,

$$p(x_t, t; x_s, s) = ce^{-u-v}(v/u)^{q/2} I_q(2\sqrt{uv}) \quad (11)$$

where $u = cx_s e^{-\theta_2\Delta}$, $v = cx_t$, $q = 2\theta_2\theta_1/\theta_3^2 - 1$, and $I_q(\cdot)$ is the modified Bessel function of the first kind of order q.

For any experiments where synthetic data from the CIR model are required, we generate them directly using draws from the noncentral chi-squared.

Our base case uses the parameter settings $\theta^o = (.06, .5, .15)$ and $\Delta = 1/12$. These settings are identical to those used in Aït-Sahalia (2001) for ease of comparison, and are said to be calibrated to match monthly observations of the U.S. treasury bill rate. We also test the methods discussed in this article with other models and parameter settings with similar results.

We have found that better results are often obtained if the SDE is first transformed to make the diffusion term of constant magnitude. With the CIR model, for example, setting $Y = \sqrt{X}$ and applying Ito's lemma gives

$$dY = \left[\frac{\theta_2}{2Y}(\theta_1 - Y^2) - \frac{\theta_3^2}{8Y} \right] dt + \frac{\theta_3}{2}\,dW.$$

If $p_Y(y_t, t; y_s, s)$ denotes the transition density of the transformed process, then the density of the original process is obtained in the usual manner by

$$p(x_t, t; x_s, s) = p_Y(y_t, t; y_s, s)\left| \frac{dy}{dx} \right| = \frac{p_Y(\sqrt{x_t}, t; \sqrt{x_s}, s)}{2\sqrt{x_t}}.$$

In general, the appropriate transformation is given by $Y = G(X)$, where G satisfies $G'(x) = 1/\sigma(x)$. The constant of integration is irrelevant. Ito's lemma then implies

$$\begin{aligned} dY &= G'(X)\,dX + \frac{1}{2}G''(X)\sigma^2(X)\,dt \\ &= \left[\frac{\mu(X)}{\sigma(X)} - \frac{1}{2}\sigma'(X) \right] dt + dW \\ &= \left[\frac{\mu[G^{-1}(Y)]}{\sigma[G^{-1}(Y)]} - \frac{1}{2}\sigma'[G^{-1}(Y)] \right] dt + dW. \end{aligned}$$

In many cases, G can be obtained analytically; otherwise, it may require a numerical integration. This does not pose any serious difficulties. If the parameter vector enters into σ nonlinearly, the transformation will have to be recomputed for each candidate parameter, which may be inconvenient.

This transformation goes back to at least Doss (1977), and is also used by Shoji and Ozaki (1998) and Aït-Sahalia (2001). (In contrast to those papers, our methodology does not *require* that the model be transformed.) The reason underlying its effectiveness appears to be that it makes the process closer to Gaussian. This improves the performance of the approximation $p^{(M)}$, as well as that of the importance samplers.

We compare the effectiveness of the various approximation techniques using several different measures. First, we look at some density plots. We fix a value for x_s, and consider a range of values for x_t. For each value of x_t, we approximate the density $p(x_t, t; x_s, s)$ a number of times using different seeds for the random number generator. We then compute the difference between the true and approximate log densities, and plot the median and interquartile range of the approximation errors for each x_t.

The figures in this article are obtained using the process defined by (10), with $\Delta = t - s = 1/12$, $x_s = .10$, $x_t \in [.05, .15]$, and 1,024 repetitions. For reference, the exact transition density with these settings is shown in Figure 1.

Since the object of ultimate interest is the log-likelihood, it seems appropriate to examine the error in the log rather

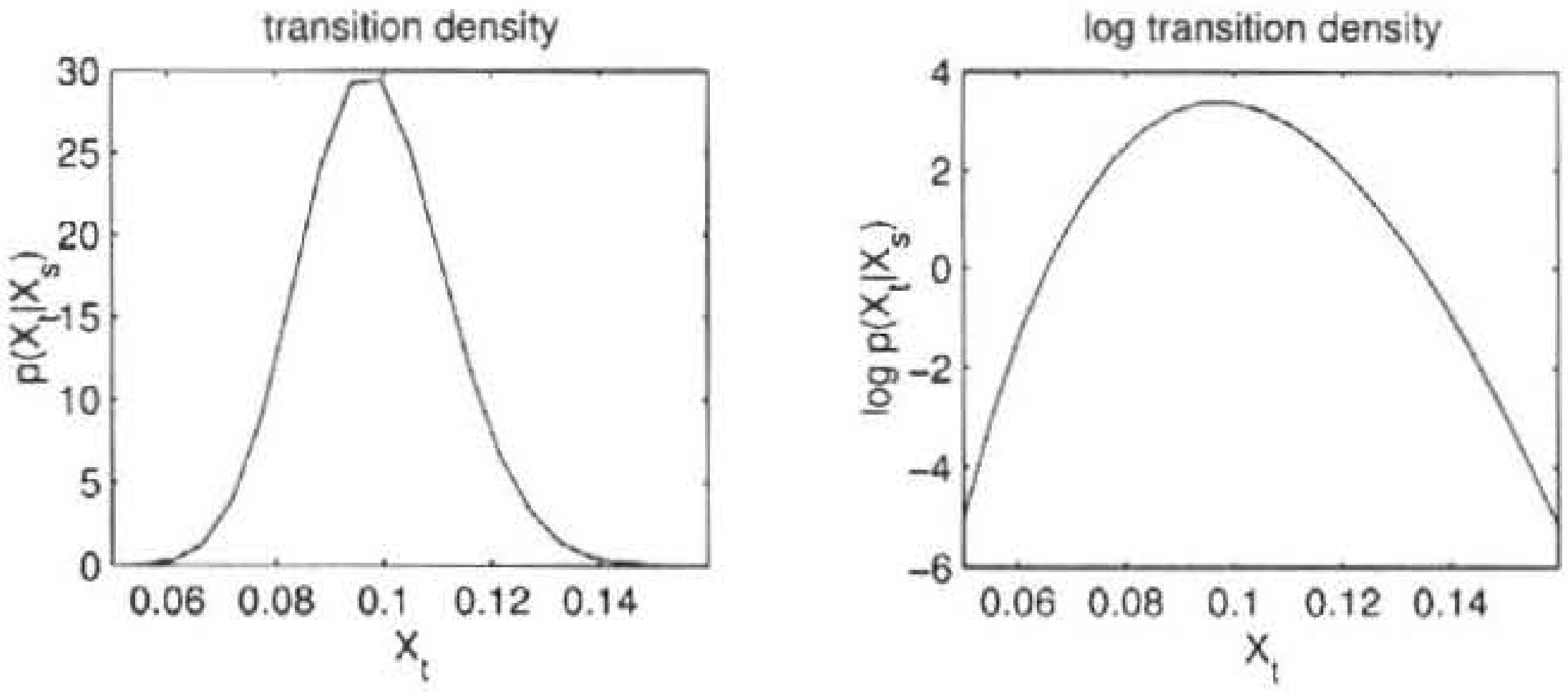


Figure 1. True Transition Density and Log Transition Density for the CIR Model Given $\Delta = t - s = 1/12$, $X_s = .10$, and $\theta = (.5, .06, .15)$.

than the level of the density. Suppose, for example, that $p(x_t, t; x_s, s) = 10$, and we are able to compute it with an error of $\pm .1$. This term would contribute an error of $\pm .01$ to the log-likelihood. On the other hand, if $p(x_t, t; x_s, s) = .2$, the contribution to the log-likelihood of the same approximation error would be in the range of $[-.7, .4]$. If the approximation error is greater than the level of the density, the computed density can be negative. This is clearly catastrophic for the log-likelihood. These distinctions are obscured if one examines the approximation error for the level of the density.

The second measure which we examine is the root mean squared error (RMSE) of the log-density approximation. We approximate this by generating $n = 100{,}000$ simulated observations from the model, and computing a sample analog, that is,

$$\mathrm{RMSE} = \left\{ \int \left(\log \hat{p}(y|x) - \log p(y|x) \right)^2 p(y, x)\, dy\, dx \right\}^{1/2} \quad (12)$$

$$\approx \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\log \hat{p}(x_{i+1}|x_i) - \log p(x_{i+1}|x_i) \right)^2 \right\}^{1/2} \quad (13)$$

where we have denoted the approximate transition density by $\hat{p}$. It is convenient to assume that the integral in (12) exists. At any rate, the sum in (13) certainly exists for a fixed realization $\{x_1, \ldots, x_n\}$, which is all we really need in order to compare across approximation techniques.

Finally, we are interested in the accuracy of the parameter estimates obtained by maximizing the approximate rather than the exact log-likelihood. To measure this, we generate $J = 512$ data sets of length $n = 1{,}000$, and compute $\hat{\theta}$ for each repetition using the exact log-likelihood and the various approximations. We compute the RMSE of the exact maximum likelihood estimates with respect to the parameter vector used to actually generate the data, and the RMSE of the simulated maximum likelihood estimates with respect to the exact maximum likelihood estimates, that is,

$$\mathrm{RMSE}_{\text{TRUE-MLE}} = \left\{ \frac{1}{J} \sum_{j=1}^{J} \left(\hat{\theta}_{\text{MLE}} - \theta^{o} \right)^2 \right\}^{1/2}$$

$$\mathrm{RMSE}_{\text{MLE-SMLE}} = \left\{ \frac{1}{J} \sum_{j=1}^{J} \left(\hat{\theta}_{\text{MLE}} - \hat{\theta}_{\text{SMLE}} \right)^2 \right\}^{1/2}.$$

A reasonable goal might be to obtain an approximation error on the order of 1% of the error inherent in the MLE itself. We are able to easily obtain this goal for our test case.

Virtually any method which one might reasonably consider should be able to approximate the log-likelihood function with arbitrary precision given sufficient time. The key issue is how quickly one is able to obtain sufficiently accurate results. Thus, we also report computational costs.

As a matter of implementation, variance in the Monte Carlo integral can result in a great deal of jaggedness in the likelihood surface, which will severely degrade the performance of the optimizer. However, this issue is easily addressed if, for each evaluation of the likelihood function, one uses the same seed for the random number generator used to draw samples for the Monte Carlo integration. This is especially critical if one is computing numerical derivatives. At any rate, for many of the methods which we examine, it is relatively straightforward to obtain analytical derivatives.

3. SIMULATION METHOD WITHOUT ACCELERATION TECHNIQUES

To establish a baseline, we begin by examining the simulation method as implemented by Pedersen (1995b), that is, without any of our acceleration techniques. The importance sampler used by Pedersen is constructed by simulating paths on each subdivided interval using the Euler scheme. Suppose that $s < t$, $x_s = X(s)$, and $x_t = X(t)$ are given. The importance sampler is defined by the mapping $T^{(M)}: (W_1, \ldots, W_{M-1}; \theta) \mapsto (u_1, \ldots, u_{M-1})$ given by the recursion

$$u_{m+1} = u_m + \mu(u_m; \theta)\delta + \sigma(u_m; \theta)\delta^{1/2} W_{m+1}, \qquad m = 0, \ldots, M-2 \quad (14)$$

where $u_0 = x_s$, $\delta = (t - s)/M$, and $W = (W_1, \ldots, W_{M-1})$ is a multivariate standard normal.

In this case, Equation (8) simplifies considerably. Since the density of the importance sampler q is identical to the first $M - 1$ factors of the numerator, they cancel, and one is left with

$$p^{(M,K)}(x_t, t; x_s, s, \theta) = \frac{1}{K} \sum_{k=1}^{K} p^{(1)}(x_t, t; u_{k,M-1}, \tau_{M-1}, \theta) \quad (15)$$

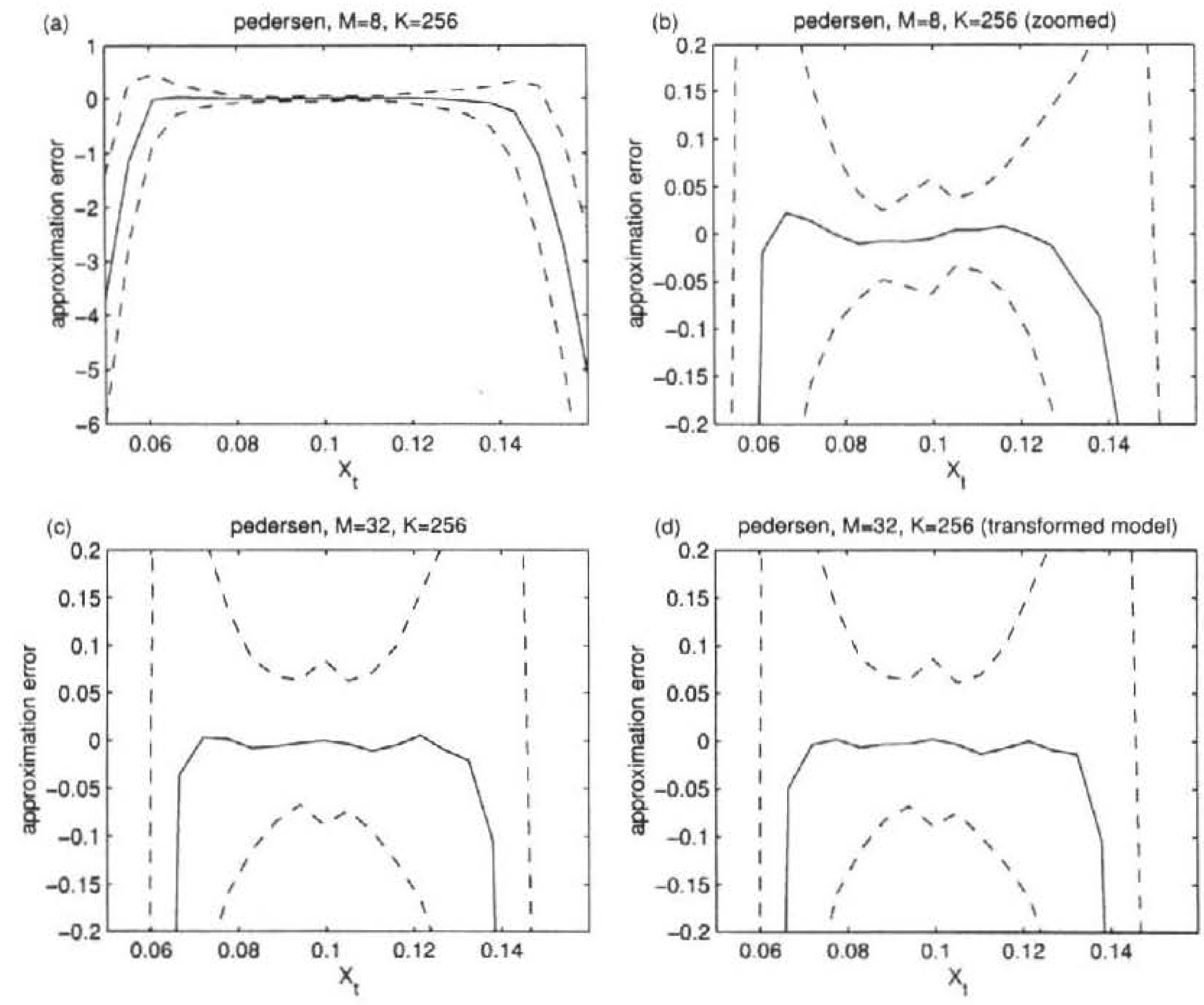


Figure 2. Approximation Error, $\log \hat{p}(X_t,t;X_s,s) - \log p(X_t,t;X_s,s)$, Using Pedersen's Method Given $\Delta = t - s = 1/12$, $X_s = .10$, and $\theta = (.5,.06,.15)$. The median and interquartile range over 1,024 repetitions are plotted. The untransformed model is used in panels (a)–(c), and the transformed model is used in panel (d).

where the $\{u_{k,M-1}, k = 1,\ldots,K\}$ are drawn from the $(M-1)$st component of q. An alternate interpretation of Equation (15) is to consider the right-hand side as the sample analog of $\mathrm{E}[p^{(1)}(x_t, t; u_{M-1}, \tau_{M-1}, \theta)]$, where the expectation is over u and with respect to the distribution induced by $X(\tau_{M-1})|X(s) = x_s$.

Throughout this article, we use the method of antithetic variates when drawing random numbers. This is a commonly used variance-reduction technique in simulation-based methods. To implement antithetic variates, one draws only $K/2$ samples from the multivariate normal, and simulates two paths from each: $T_+^{(M)}(W) = T^{(M)}(W)$ is as described above, and $T_-^{(M)}(W) = T^{(M)}(-W)$ is its "mirror image." While we have found antithetic variates to provide only marginal benefit, the cost is also negligible.

Figure 2 illustrates the approximation error which results from computing the log density using this approach. The settings $K = 256$ and $M = 8$ or $M = 32$ are used (recall that K is the number of sample paths and M is the number of subintervals). Panels (a)–(c) use the untransformed model. Panel (d) uses the transformed model, which appears to provide little benefit in this case. Increasing M reduces bias, but at the cost of greater variance. Reducing the variance is costly since it is $\mathcal{O}(K^{-1/2})$.

Upon comparison with the tables and figures in Section 6, the reason why this approach has not seen widespread use is readily apparent. It would take a great deal of effort even to match the accuracy of the simple first-order approximation, at least for our test model.

4. BIAS-REDUCTION TECHNIQUES

There are two sources of approximation error which we wish to address: bias due to the first-order approximation used in the construction of $p^{(M)}$, and variance resulting from the Monte Carlo integration.

We begin with the bias. While it is possible to drive the bias to zero by partitioning the intervals between observations sufficiently finely, this can be computationally costly. We examine two approaches toward reducing the number of subintervals required to obtain a given level of accuracy. The first is to replace the first-order approximation used in Equation (5) by a higher order method. There are several possibilities which one might try.

Elerian (1998) suggests using a transition density derived from a scheme due to Milstein (1978). If the volatility function $\sigma(\cdot)$ is constant, the density is identical to that of the first-order approximation; otherwise, it is given by

$$p_{\text{Elerian}}(x_t, t; x_s, s) = \frac{z_t^{-1/2}}{|A|\sqrt{2\pi}} \exp\left(-\frac{C + z_t}{2}\right) \cosh(\sqrt{C z_t})$$

where

$$z_t = \frac{x_t - B}{A}$$
$$\Delta = t - s$$
$$A = \frac{\sigma(x_s)\sigma'(x_s)\Delta}{2}$$
$$B = -\frac{\sigma(x_s)}{2\sigma'(x_s)} + x_s + \mu(x_s)\Delta - \frac{\sigma(x_s)\sigma'(x_s)\Delta}{2}$$
$$C = \frac{1}{(\sigma'(x_s))^2\Delta}.$$

Note that $\sigma'(\cdot)$ denotes the derivative of σ.

Kessler (1997) suggests using a Gaussian transition density, but rather than using the first-order approximations for the mean and variance, he proposes using higher order Ito–Taylor approximations. We try a second-order implementation, that is,

$$p_{\text{Kessler}}(x_t, t; x_s, s) = \phi(x_t; \tilde{\mu}, \tilde{\sigma}^2)$$

where

$$\begin{aligned}\tilde{\mu} &= x_s + \mu(x_s)\Delta + \left[\mu(x_s)\mu'(x_s) + \frac{\sigma^2(x_s)\mu''(x_s)}{2}\right]\frac{\Delta^2}{2}\\ \tilde{\sigma}^2 &= x_s^2 + \{2\mu(x_s)x_s + \sigma^2(x_s)\}\Delta\\ &\quad + \{2\mu(x_s)[\mu'(x_s)x_s + \mu(x_s) + \sigma(x_s)\sigma'(x_s)]\\ &\qquad + \sigma^2(x_s)[\mu''(x_s)x_s + 2\mu'(x_s) + \sigma'(x_s)\sigma'(x_s)\\ &\qquad\quad + \sigma(x_s)\sigma''(x_s)]\}\frac{\Delta^2}{2} - \tilde{\mu}^2.\end{aligned}$$

Notice that, for some models and parameter settings, it is possible to obtain $\tilde{\sigma}^2 < 0$. The code should include a check to watch out for this.

Shoji and Ozaki (1998) suggest a method which they refer to as *local linearization*. Their approach requires a model with constant volatility; however, as shown in Section 3, this results in little loss of generality. Given

$$dX = \mu(X)\,dt + \sigma\,dW$$

(σ is constant) and fixed x_s, one begins with an application of Ito's lemma:

$$d\mu(X) = \frac{1}{2}\sigma^2\mu''(X)\,dt + \mu'(X)\,dX.$$

Using the first-order Taylor expansion, we define

$$\tilde{\mu}(x_t) = \mu(x_s) + \mu'(x_s)(x_t - x_s) + \frac{1}{2}\sigma^2\mu''(x_s)(t - s).$$

The approximate density will be obtained from

$$d\widetilde{X} = \tilde{\mu}(\widetilde{X})\,dt + \sigma\,dW,$$

which is an Ornstein–Uhlenbeck process. One obtains

$$p_{\text{Sh\&Oz}}(x_t, t; x_s, s) = \phi(x_t; \tilde{\mu}, \tilde{\sigma}^2)$$

where

$$\tilde{\mu} = x_s + \frac{\mu(x_s)}{\mu'(x_s)}K + \frac{\sigma^2\mu''(x_s)}{2[\mu'(x_s)]^2}[K - \mu'(x_s)\Delta]$$
$$\tilde{\sigma}^2 = \frac{\sigma^2}{2\mu'(x_s)}(e^{2\mu'(x_s)\Delta} - 1)$$
$$K = e^{\mu'(x_s)\Delta} - 1.$$

Nowman (1997) suggests a similar approach, but simply treating the volatility as if it were constant on each sample interval rather than transforming the model so that it actually is constant. While he examines only the special case where the drift function is linear, a plausible extension would be to use a first-order Taylor expansion for the drift function as described above. The resulting approximation is analogous to that of Shoji and Ozaki, but replacing σ by $\sigma(x_s)$.

While Elerian (1998) uses the Milstein density in the context of a simulation-based approach, Kessler (1997), Shoji and Ozaki (1998), and Nowman (1997) approximate the transition density between observations directly (i.e., without using intermediate points).

Another approach to obtaining higher order methods is extrapolation. Given, for example, a first-order method, one may construct a second-order method as follows:

$$\begin{aligned}p^{(M)} &= p + K\Delta + \mathcal{O}(\Delta^2)\\ p^{(2M)} &= p + K\Delta/2 + \mathcal{O}(\Delta^2)\\ p_E^{(2M)} &= 2p^{(2M)} - p^{(M)}\\ &= p + \mathcal{O}(\Delta^2)\end{aligned}$$

where K is some unknown constant. If the approximate likelihoods are stochastic (i.e., computed by simulation), extrapolation reduces bias, but at the cost of greater variance. Since it is possible to obtain a negative value for the extrapolated density, any implementation of this technique should check for positivity, and fall back to the nonextrapolated value in case of trouble.

Extrapolation is a well-known bias-reduction technique for computing expectations of diffusion processes (see Kloeden and Platen 1992, sec. 15.3). That we are able to apply the technique in the present context is because our approach to approximating the transition density is essentially an expectation. For Pedersen's method, it is easy to see from Equation (15) that

$$p^{(M)}(x_t, t; x_s, s) = \int p^{(1)}(x_t, t; u, \tau_{M-1}, \theta)\,dP_{M-1}^{(M)}(u) \quad (16)$$

where $P_{M-1}^{(M)}$ is the measure induced by the Euler scheme at τ_{M-1}. In general, one obtains

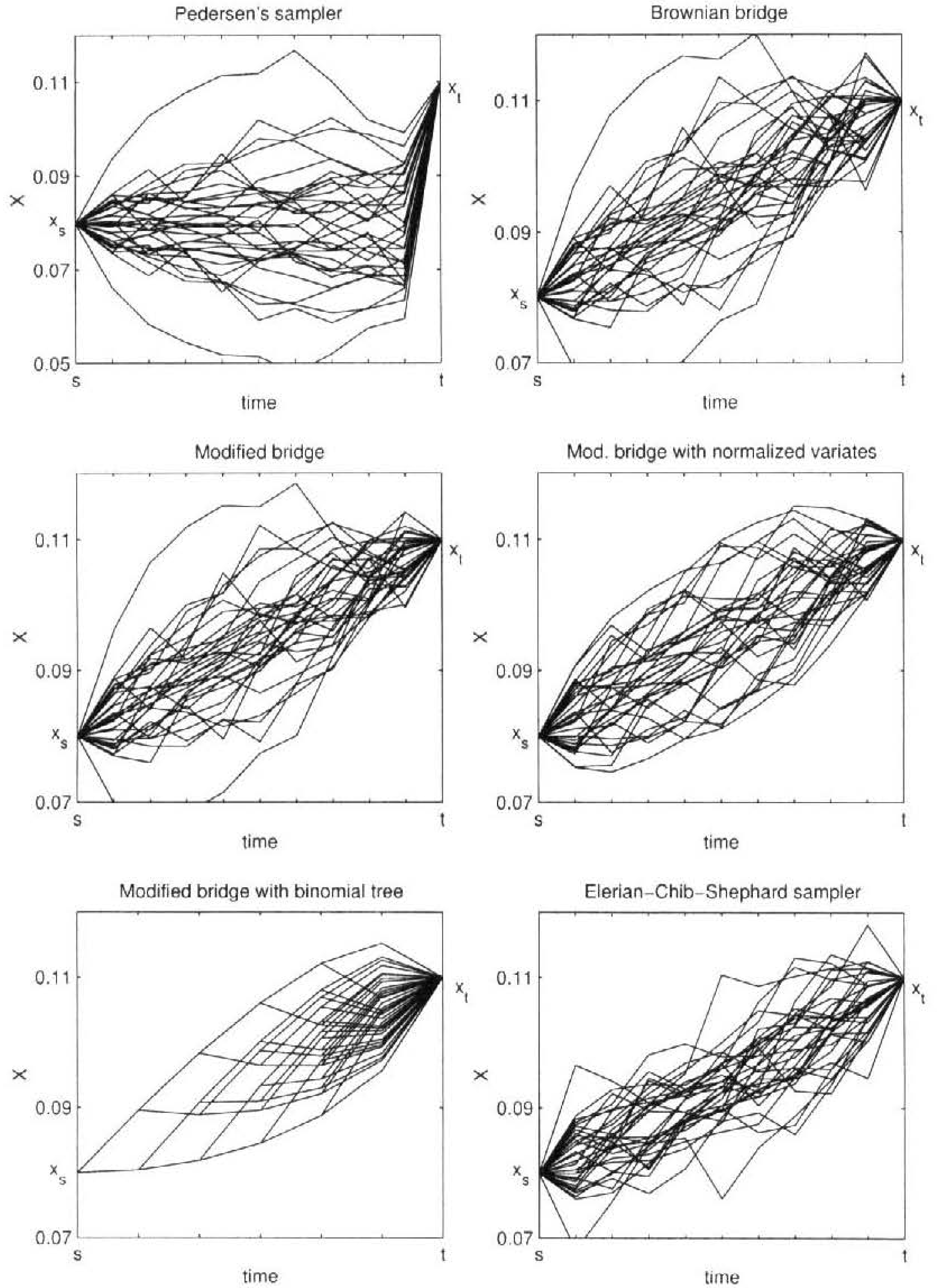
$$p^{(M)}(x_t, t; x_s, s) = \int p^{(1)}(x_t, t; u, \tau_{M-1}, \theta)\,\rho_{M-1}^{(M)}(u)\,dQ_{M-1}^{(M)}(u) \quad (17)$$

where $Q_{M-1}^{(M)}$ is the measure induced by the $(M-1)$st component of the importance sampler and $\rho_{M-1}^{(M)}$ is the Radon–Nikodym derivative of $P_{M-1}^{(M)}$ with respect to $Q_{M-1}^{(M)}$. These expressions may also be derived directly from (6).

5. VARIANCE-REDUCTION TECHNIQUES

We examine two approaches to reducing the variance of the Monte Carlo integration shown in Equation (8): importance sampling and random number schemes. Some of the techniques are illustrated in Figure 3.

A basic principle of Monte Carlo integration is that one should draw points with higher probability in regions where the integrand is larger. Figure 4 illustrates why Pedersen's method performs so poorly. The paths in the figure are sampled using the Euler scheme with $x_s = .08$, $\Delta = t - s = 1/12$, and the SDE given in Section 2. The terminal point of each path represents a draw from $P_{M-1}^{(M)}$. The curve represents the integrand of the right-hand side of (16) as a function of u with $x_t = .11$. It is clear that most of the samples are drawn from regions where the integrand has little mass. The importance samplers discussed in this section are designed to address this shortcoming. Elerian et al. (2001) appear to have been the first to consider the idea of using efficient importance sampling in this context.


Pedersen's sampler

Brownian bridge

Modified bridge

Mod. bridge with normalized variates

Modified bridge with binomial tree

Elerian–Chib–Shephard sampler

Figure 3. Simulated Paths Drawn Using Various Importance Samplers and Random Schemes.

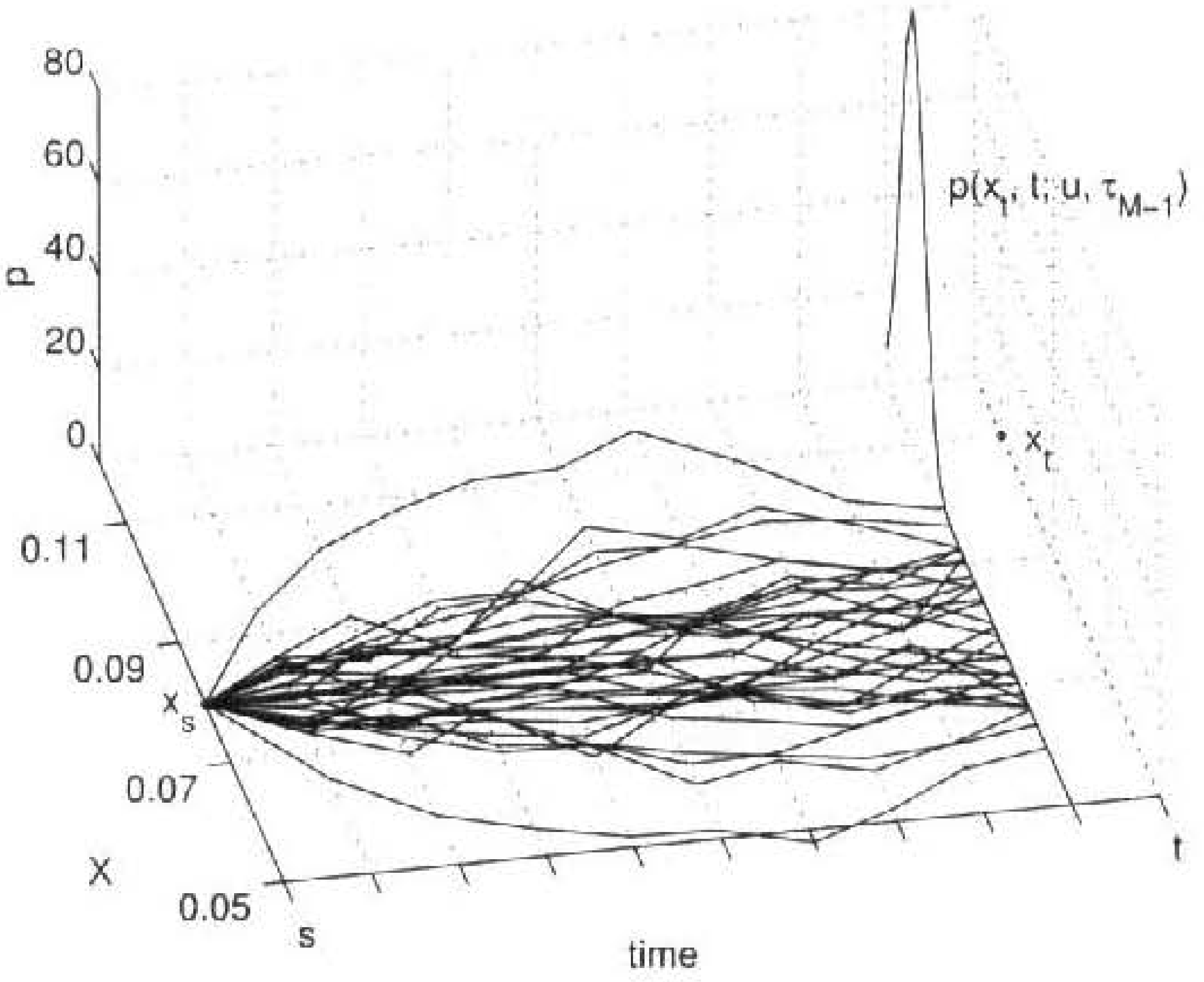


Figure 4. Illustration of Equation (16). The terminal points of the sample paths represent draws from $P^{(M)}_{M-1}$; the curve represents the integrand.

The first importance sampler we consider is based on the Brownian bridge. A Brownian bridge is a Brownian motion started at x_s at time s and conditioned to terminate at x_t at time t. The sampler is constructed in a manner similar to the Euler scheme. In this case, the mapping $T^{(M)}: (W_1, \ldots, W_{M-1}; \theta) \mapsto (u_1, \ldots, u_{M-1})$ is defined by the recursion

$$u_{m+1} = u_m + \tilde{\mu}(u_m, \tau_m)\delta + \sigma(u_m; \theta)\delta^{1/2} W_{m+1}$$

where the drift is given by

$$\tilde{\mu}(x, \tau) = \frac{x_t - x}{t - \tau}.$$

This is a true Brownian bridge if and only if σ is constant (which will be the case if we first transform the model as discussed in Section 2).

Although it is certainly possible to compute the approximate density directly from (8), there is an interesting interpretation of this sampler based on Girsanov's theorem. Consider the processes $dX = \mu(X)\,dt + \sigma(X)\,dW$ and $d\widetilde{X} = \tilde{\mu}(\widetilde{X})\,dt + \sigma(\widetilde{X})\,d\widetilde{W}$ with initial condition $X(s) = \widetilde{X}(s) = x_s$. Girsanov's theorem tells us that the Radon–Nikodym derivative of the probability measure generated by X with respect to that generated by $\widetilde{X}$ is given by

$$d\rho = \rho k(\widetilde{X})\,d\widetilde{W} \tag{18}$$

with initial condition $\rho(s) = 1$ and

$$k(x) = \frac{\mu(x) - \tilde{\mu}(x)}{\sigma(x)}.$$

We can thus obtain the continuous-time expression

$$p(x_t, t; x_s, s) = \int p(x_t, t; u, \tau_{M-1})\,\rho_{M-1}(u)\,dQ_{M-1}(u)$$

where Q_{M-1} is the probability measure induced by $\widetilde{X}(\tau_{M-1})$.

The integral is computed by generating samples $\{(u_{k,M-1}, r_{k,M-1}),\ k = 1, \ldots, K\}$ from the joint process $(\widetilde{X}^{(M)}_{M-1}, \rho^{(M)}_{M-1})$ using the Euler scheme, and then computing

$$p^{(M,K)}(x_t, t; x_s, s, \theta) = \frac{1}{K}\sum_{k=1}^{K} p^{(1)}(x_t, t; u_{k,M-1}, \tau_{M-1})\,r_{k,M-1}.$$

It is easy to show by direct calculation of $\rho^{(M)}_{M-1}$ that this expression is equivalent to (8). We have found it to be more stable to base the Euler scheme for ρ on

$$d(\log\rho) = -\frac{k^2}{2}\,dt + k\,d\widetilde{W}$$

with initial condition $\log\rho(s) = 0$ rather than (18).

The second importance sampler which we consider draws u_{m+1} from a Gaussian density based on the first-order approximation conditional on u_m and x_t. That is, treating u_m and $u_M = x_t$ as fixed, one draws u_{m+1} from the density

$$\begin{aligned}
p(u_{m+1}|u_m, u_M) &= p(u_{m+1}|u_m)p(u_M|u_{m+1})/p(u_M|u_m) \\
&\approx \phi(u_{m+1}; u_m + \bar{\mu}\delta, \bar{\sigma}^2\delta) \\
&\qquad \times \phi(u_M; u_{m+1} + \bar{\mu}\delta^*, \bar{\sigma}^2\delta^*) \\
&\qquad \div \phi(u_M; u_m + \bar{\mu}\delta^+, \bar{\sigma}^2\delta^+) \\
&= \phi(u_{m+1}; u_m + \tilde{\mu}_m\delta, \tilde{\sigma}^2_m\delta)
\end{aligned}$$

where $\delta = (t - s)/M$, $\delta^* = t - \tau_{m+1}$, $\delta^+ = t - \tau_m$, $\bar{\mu} = \mu(u_m)$, $\bar{\sigma} = \sigma(u_m)$, and

$$\tilde{\mu}_m = \left(\frac{u_M - u_m}{t - \tau_m}\right), \qquad \tilde{\sigma}^2_m = \left(\frac{M - m - 1}{M - m}\right)\bar{\sigma}^2.$$

Notice that this importance sampler turns out to be identical to the Brownian bridge sampler, except for the factor $(M - m - 1)/(M - m)$ in the variance. While it is not entirely obvious that this should be the case, we will see that this modification results in much better performance. We will refer to this sampler as the modified Brownian bridge.

The third importance sampler which we consider was proposed by Elerian et al. (2001). The idea is to approximate the target density by a multivariate normal with mean and variance based on a second-order Taylor expansion of the log target density about the mode.

The log target density is given by

$$\log p^{(M)}(u_1, u_2, \ldots, u_{M-1}|x_s, x_t) = \sum_{m=0}^{M-1} \log p^{(1)}(u_{m+1}, \tau_{m+1}; u_m, \tau_m).$$

One samples $u = (u_1, u_2, \ldots, u_{M-1})$ from $N(\mu^*, \Sigma^*)$, where

$$\mu^* = \arg\max_u \log p(u|x_s, x_t)$$

$$\Sigma^* = -\left[\frac{\partial^2}{\partial u'\partial u}\log p(\mu^*|x_s, x_t)\right]^{-1}.$$

In practice, one obtains μ^* by starting with $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_{M-1})$, where $\hat{u}_m = x_s + m(x_t - x_s)/M$, and taking a single Newton step toward the maximum. The derivatives of $\log p(u|x_s, x_t)$ are straightforward, but tedious to compute.

The key feature of this sampler is that it draws paths in one shot rather than recursively. Implementing this sampler requires solving a system of $M-1$ linear equations, and computing a Cholesky decomposition. To obtain reasonable performance, it is essential that one take advantage of the tridiagonal nature of the relevant matrices.

As always with importance sampling, one should ensure that the tails of the sampling density are not too thin; otherwise, it will not be possible to drive down the variance of the Monte Carlo integral despite using a large number of sample paths. One way to address this problem is by using Student t rather than normal increments in the construction of the sample paths. One might also try the approach suggested by Geweke (1989).

The second category of variance-reduction techniques which we examine is random number schemes. The method of antithetic variates, as discussed in Section 4, is one such scheme, although our results suggest that it provides only marginal benefits in this context.

Recall from Equations (16) and (17) that the density approximation may be thought of as an expectation. Kloeden and Platen (1992, sec. 14.1) suggest that, for computing expectations, the Gaussian increments $(W_1, \dots, W_{M-1})$ in Equation (14) (and similar expressions for the other importance samplers) may be replaced by other random variables satisfying appropriate moment conditions.

One possibility is the random variable which takes on the values 1 and -1, each with probability $\frac{1}{2}$. In addition to reducing variance, this scheme gives a speed increase, since generating normal deviates can be a significant fraction of the computational effort. Furthermore, if M is sufficiently small, it is possible to compute the Monte Carlo integral by summing over all possible branches of the binomial tree of length $M-1$. For example, setting $M=8$ would require the computation of $2^7 = 128$ sample paths, which is entirely feasible. Using the techniques discussed in Section 4, it is possible to achieve low bias with small M. In particular, since this random number scheme produces a method with no variance, it is ideally suited for use together with extrapolation.

While expectations computed using the binomial tree scheme are known to converge under appropriate conditions, the properties of this scheme in the context of this article are uncertain. Therefore, we also consider a related scheme which provides much of the same benefit by less drastic means. The idea is to control the "jaggedness" of the sample paths by forcing each vector of increments $(W_1, \dots, W_{M-1})$ to have sample variance one. This may be accomplished simply by using the vector

$$(\widetilde{W}_1, \dots, \widetilde{W}_{M-1}) = \left\{ \frac{1}{M-1} \sum_{m=1}^{M-1} W_m^2 \right\}^{1/2} (W_1, \dots, W_{M-1}),$$

and may be thought of as a weakening of the two-point idea, which forces the sample variance of each individual increment to be equal to 1.

6. NUMERICAL EXPERIMENTS

We first test our various techniques by approximating the transition density as described in Section 2. The settings used for the various approximations will be identified by "sampler-subdensity-M-K," for example, the Brownian bridge sampler used together with the first-order (Euler) subtransition density, $M=8$, and $K=256$ will be identified as "bridge-euler-8-256." The RMSE is also computed for these approximation schemes (as described in Section 2). The results are summarized in Table 1.

Figure 5 illustrates the performance of the various subdensity methods when used to compute the transition density directly (i.e., $M=1$, no simulation). While the error associated with the simple first-order approximation is moderately severe, a factor of 10 improvement is obtained if the model is transformed before applying the first-order approximation. The scheme proposed by Elerian (1998) comes close to obtaining this improvement without needing the transformation step. When applied to the transformed models, the techniques proposed by Shoji and Ozaki (1998) and Kessler (1997) provide an additional order of magnitude improvement over the Euler scheme. Although not shown in the figures, we have found these schemes to be of little benefit when used on the untransformed model. The technique proposed by Nowman (1997) provides nearly no benefit whatsoever.

Figure 6 illustrates the Brownian bridge and modified bridge samplers. The first-order approximation is used for the subtransition densities. The transformation step is not used. Using the Brownian bridge largely solves the main problem associated with Pedersen's method. The modified bridge provides a further dramatic reduction in variance. Notice that panels (c) and (d) of Figure 6 use only $K=8$ sample paths, as compared to $K=256$ for panels (a) and (b) and Figure 2(a)–(d) (Pedersen's method). We see that increasing the number of subintervals brings the expected reduction in bias.

Figure 7(a) and (b) illustrates the use of extrapolation and Elerian's subtransition density scheme, respectively, to reduce bias. Panel (c) shows the variance reduction due to normalized variates. Panel (d) demonstrates that one still obtains the expected reductions in bias and variance from increasing M and K, respectively. All panels in this figure use the untransformed model.

Figure 8 illustrates the Elerian–Chib–Shephard (ECS) sampler. This sampler seems to work well for M relatively small, but the variance goes up dramatically as the number of intermediate points increases. It was not possible to compute the RMSE of the log density approximation with the ECS sampler and $M=32$ because the sampler often chose points below zero (i.e., outside the range of the model). We follow Elerian et al. (2001) by using the transformation $Y = \log X$ in this case.

While it is possible to obtain reasonably accurate results using the untransformed model, Figure 9 shows that first transforming the model provides significant benefits, especially when used with the subdensity scheme of Shoji and Ozaki (1998) and normalized variates. With these settings, we are able to obtain RMSE $\approx .0006$ with only $M=8$ and $K=8$. The computational cost is about 16 s to approximate a likelihood with $n = 100{,}000$ observations using FORTRAN code

Table 1. RMSE of Log Density Approximations

Sampler	Subdensity	M	K	NV[a]	Extrap.[b]	RMSE	Time[c]
			Untransformed model				
None	Euler	1	1			.13678	.2
None	Elerian	1	1			.03550	.2
None	Nowman	1	1			.14467	.2
Pedersen	Euler	8	256			.19353	539.5
Pedersen	Euler	8	256			.19353	538.3
Pedersen	Euler	32	256			.66310	2,169.8
Bridge	Euler	8	256			.05355	550.4
Bridge	Euler	32	256			.05570	2,211.2
Mod bridge	Euler	8	8			.03068	17.5
Mod bridge	Euler	8	8		x	.06103	26.3
Mod bridge	Euler	8	8	x		.02404	19.1
Mod bridge	Euler	32	32	x		.01180	299.3
Mod bridge	Euler	32	128	x		.00713	1,227.1
Mod bridge	Elerian	8	8			.03562	24.6
Mod bridge	Elerian	8	8	x		.01856	25.0
Mod bridge	Elerian	32	32	x		.01206	401.2
Mod bridge	Elerian	32	128	x		.00656	1,603.8
ECS	Elerian	8	8			.07207	30.4
ECS	Elerian	8	8	x		.05164	30.3
ECS[d]	Elerian	32	32	x		.46920	403.6
ECS[d]	Elerian	32	128	x		.25973	1,582.2
			Transformed model				
None	Euler	1	1			.03790	.2
None	Shoji	1	1			.01412	.3
None	Kessler	1	1			.00864	.2
Pedersen	Euler	8	256			.16500	338.2
Pedersen	Euler	32	256			.57507	1,254.5
Mod bridge	Euler	8	8			.00812	10.8
Mod bridge	Euler	64	8			.00255	80.7
Mod bridge	Euler	8	8		x	.01795	16.6
Mod bridge	Shoji	8	8			.00070	16.0
Mod bridge	Shoji	8	8	x		.00057	16.4
Mod bridge	Shoji	32	32	x		.00030	248.7
ECS	Shoji	8	8	x		.00072	25.3
ECS	Shoji	32	32	x		.00020	315.5

[a] An "x" in this column indicates normalized variates were used; otherwise, antithetic variates.

[b] An "x" in this column indicates extrapolation was used.

[c] Computing time (in seconds) required to obtain likelihood for $n = 100{,}000$ observations using FORTRAN code on a 750 MHz PC.

[d] With these settings, the sampler frequently chose points $x_t < 0$; thus, following Elerian et al., we used the transformation $Y = \log X$.

on a PC running at 750 MHz. It would be virtually impossible to obtain anywhere near this level of accuracy using Pedersen's method without our acceleration techniques.

Tables 2 and 3 show the errors which result from estimating parameters by maximizing the approximate rather than exact log-likelihood. Results are shown for several different settings of the model parameters. The errors are estimated by Monte Carlo simulation with 512 repetitions over synthetic datasets of $n = 1{,}000$ observations. The SMLE estimates are obtained using the modified bridge sampler, Shoji and Ozaki's subdensity, $M = 16$ and $K = 16$. For comparison, we also compute parameter estimates using the first-order Euler scheme approximation. The transformed model is used for all of the experiments shown in these tables. It should be noted that the Euler scheme approximations thus obtained can be expected to be significantly better than those typically obtained by practitioners without implementing the transformation (see Fig. 5).

Table 2(a) uses the baseline model settings, $\theta^o = (.06, .5, .15)$ and $\Delta = 1/12$. Recall that these are calibrated to monthly observations of the U.S. short-term interest rate. For these model settings, we also compute parameter estimates using Pedersen's method with $M = 8$ and $K = 256$. Pedersen's method is unable to match even the Euler scheme. On the other hand, the approximation errors of the SMLE estimates obtained using our techniques are negligible (compare with the sample distribution of $\hat{\theta}_{\mathrm{MLE}} - \theta^o$).

Panel (b) increases the volatility of the model. Note that computing the exact likelihood requires the evaluation of a Bessel function, which in turn requires $2\theta_2\theta_1/\theta_3^2 \geq 1$. Setting $\theta^o = (.06, .5, .22)$ comes quite close to this boundary. For some samples, the constraint appears to bind when maximizing the likelihood. These samples are discarded. This model causes our methodology some difficulty, apparently because the data often venture close to the singularity at zero. The estimates are nonetheless quite good.

Panel (c) reduces the model's volatility to $\theta_3 = .03$. Panel (d) sets the mean reversion parameter to $\theta_2 = .4$. Again, the constraint $2\theta_2\theta_1/\theta_3^2 \geq 1$ comes into play. Neither of these tests presents any difficulty to our methodology.

Panel (e) increases the mean reversion parameter by a factor of 10 to $\theta_2 = 5.0$. Panel (f) uses the baseline setting for θ^o, but stretches the sampling interval to two years. Both of these

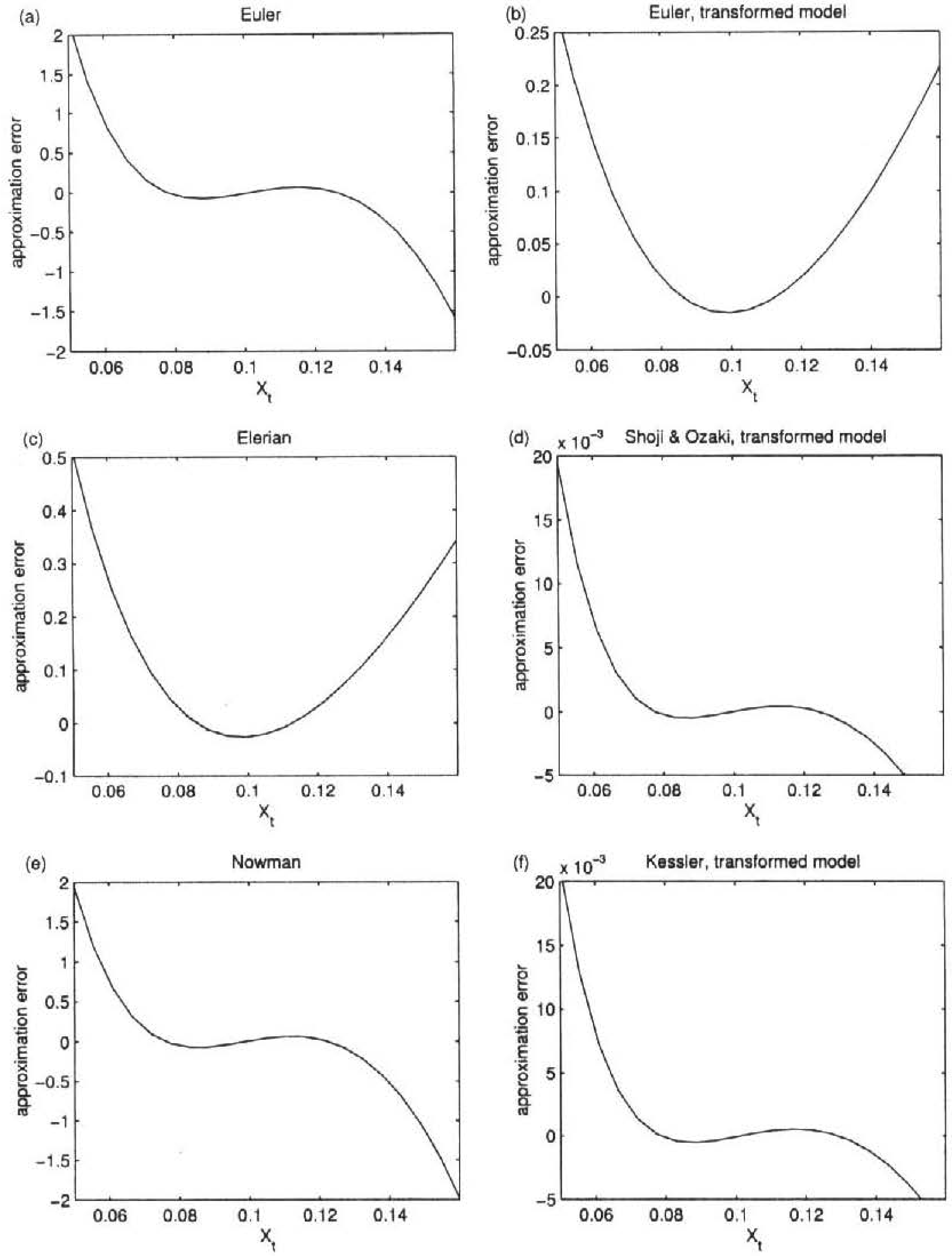


Figure 5. Approximation Error, $\log \hat{p}(X_t, t; X_s, s) - \log p(X_t, t; X_s, s)$, for Various Schemes Without Using Simulation (i.e., $M = 1$, $K = 1$) Given $\Delta = t - s = 1/12$, $X_s = .10$, and $\theta = (.5, .06, .15)$.

settings result in large biases for the first-order approximation, but pose little difficulty for the SMLE technique.

7. STOCHASTIC VOLATILITY

While the previous sections have focused on techniques designed to efficiently approximate the likelihood function for scalar models, most of these ideas are easily generalized to the multivariate setting. With some work, they may also be applied to latent variable models. The short-term interest rate and many other financial time series are well known to exhibit properties such as fat-tails and persistent volatility patterns which are inconsistent with time-homogeneous scalar models (e.g., Ghysels, Harvey, and Renault 1996). A variety of alternative models has been proposed. To illustrate our methodology, we will examine stochastic volatility (SV) models of the form

$$dX = \mu_X(X)\,dt + \sigma_X(X)\exp(H)\,dW_1$$
$$dH = \mu_H(X, H)\,dt + \sigma_H(X, H)\,dW_2.$$

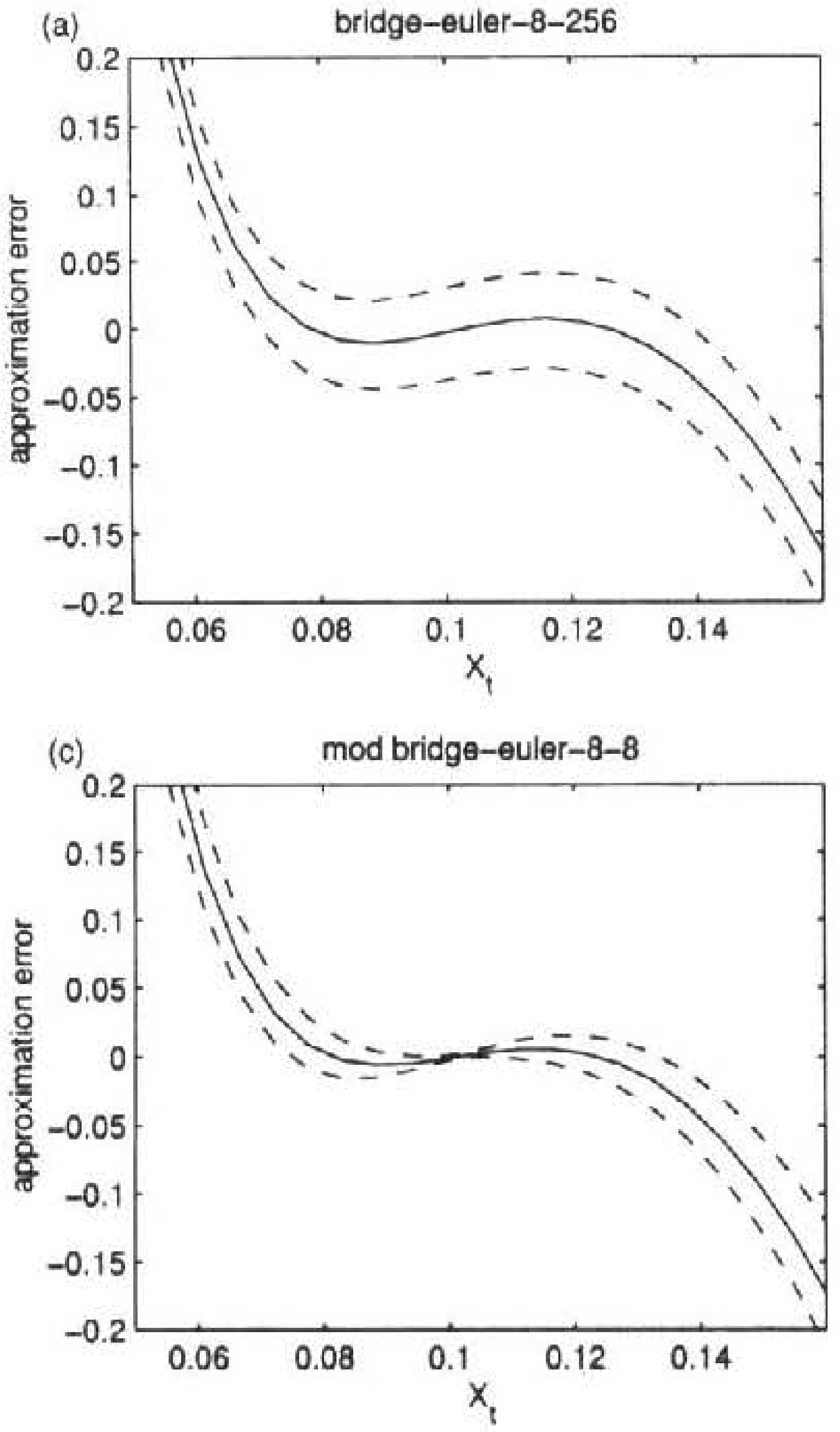


Figure 6. Approximation Error, $\log \hat{p}(X_t, t; X_s, s) - \log p(X_t, t; X_s, s)$, for Various Simulated Likelihood Schemes Given $\Delta = t - s = 1/12$, $X_s = .10$, and $\theta = (.5, .06, .15)$. The median and interquartile range over 128 repetitions are plotted. The untransformed model is used.

Such models have been examined by Gallant and Tauchen (1998), Andersen and Lund (1997), and Eraker (2001) among others. The second component, H, corresponds to an unobserved volatility factor.

In order to obtain a likelihood, the unobserved factor must be integrated out. Several ways of going about this have been proposed for discrete-time models, for example, Danielsson and Richard (1993), Jacquier, Polson, and Rossi (1994), Richard and Zhang (2000), Sandmann and Koopman (1998), Kim, Shephard, and Chib (1998), and Pitt and Shephard (1999). In the continuous-time context, it is less straightforward to integrate out the unobserved factor, and alternative approaches are used. The efficient method of moments approach has been used by Gallant and Tauchen (1998) and others. Methods based on the empirical characteristic function have been proposed by Chacko and Viceira (1998) and Singleton (1997). Markov chain Monte Carlo approaches have been proposed by Eraker (2001), Jones (1999b), and Elerian (1999).

We now outline a technique to approximate the likelihood of continuous time SV models. For simplicity, we will assume that W_1 and W_2 are independent; this is not an essential part of the methodology. We provide a small Monte Carlo study which demonstrates that the procedure is effective and reasonably fast. The methodology is used to estimate an SV model of the U.S. short-term interest rate in Section 8. Further refinements are undoubtedly possible; a more detailed study is currently underway.

The basic idea is relatively straightforward. We are interested in the process $(X(t), H(t))$, where X is observed at times $t_0, t_1, \ldots, t_n$ and H is latent. Let $X_i = X(t_i)$, $H_i = H(t_i)$, and $\mathcal{F}_i = \sigma(H_0, X_0, X_1, \ldots, X_i)$ for $i = 1, \ldots, n$. The goal is to obtain $p(X_{i+1}|\mathcal{F}_i)$. If we knew the distribution of $H_i|\mathcal{F}_i$, we could use

$$p(X_{i+1}|\mathcal{F}_i) = \int p(X_{i+1}|X_i, h_i)\, dP_{H_i|\mathcal{F}_i}(h_i).$$

We will approximate $H_i|\mathcal{F}_i$. Given the distribution of $H_{i-1}|\mathcal{F}_{i-1}$, it can be propagated forward using

$$p(H_{i-1}|\mathcal{F}_i) = \frac{p(X_i|\mathcal{F}_{i-1}, H_{i-1})\, p(H_{i-1}|\mathcal{F}_{i-1})}{p(X_i|\mathcal{F}_{i-1})}$$

$$p(H_i|\mathcal{F}_i) = \int p(H_i|X_i, h_{i-1})\, dP_{H_{i-1}|\mathcal{F}_i}(h_{i-1}).$$

It remains only to find H_0. The approach we take is to estimate H_0 as an unknown parameter, although one could equally well integrate it out using an appropriate importance sampler. This is the basic idea of a particle filter (see, e.g., Pitt and Shephard 1999 and the references therein).

To implement this idea, we use the following procedure. Consider the model

$$dZ = \mu(Z)\, dt + \sigma(Z)\, dW$$

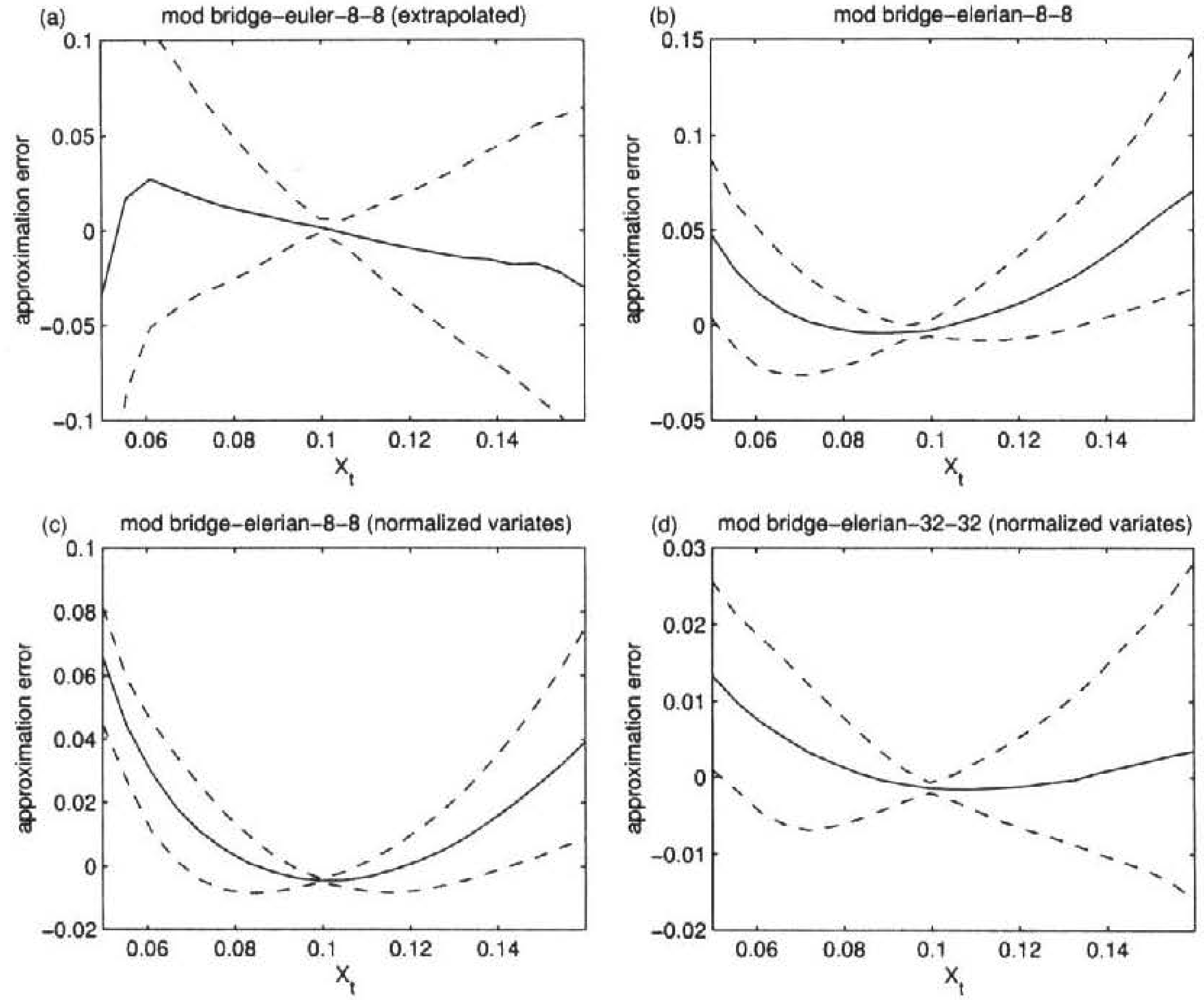


Figure 7. Approximation Error, $\log \hat{p}(X_t, t; X_s, s) - \log p(X_t, t; X_s, s)$, for Various Simulated Likelihood Schemes Given $\Delta = t - s = 1/12$, $X_s = .10$, and $\theta = (.5, .06, .15)$. The median and interquartile range over 128 repetitions are plotted. The untransformed model is used.

where

$$Z = \begin{pmatrix} X \\ H \end{pmatrix}, \mu = \begin{pmatrix} \mu_X(Z) \\ \mu_H(Z) \end{pmatrix}, \sigma = \begin{pmatrix} \sigma_X(Z) & 0 \\ 0 & \sigma_H(Z) \end{pmatrix}, W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}.$$

The problem is to determine the density $p(X_{i+1}|\mathcal{F}_i, \theta)$, where $\mathcal{F}_i = \sigma(H_0, X_0, X_1, \ldots, X_i)$.

The first-order approximation is given by

$$p^{(1)}(z_t, t; z_s, s, \theta) = \phi(z_t; z_s + \mu(z_s)(t - s), \Sigma(z_s)(t - s))$$

where ϕ is the multivariate Gaussian density and $\Sigma = \sigma\sigma^T$. Also, let $s = \tau_0 < \tau_1 < \cdots < \tau_M = t$ be a partition of the interval $[s, t]$, and let

$$p^{(M)}(z_t, t; z_s, s, \theta) = \int \prod_{m=1}^{M} p^{(1)}(w_m, \tau_m; w_{m-1}, \tau_{m-1}, \theta)\, d\lambda(w_1, \ldots, w_{M-1})$$

where $w_m \in \mathbb{R}^2$ for $m = 0, \ldots, M$, λ is the Lebesgue measure in $\mathbb{R}^{2(M-1)}$, and we use the convention $w_0 = z_s$ and $w_M = z_t$. Since H is unobserved, it must be integrated out as well. One obtains

$$p^{(M)}(X_{i+1}; \mathcal{F}_i, \theta) = \int p^{(M)}\left(\begin{pmatrix} X_{i+1} \\ h_{i+1} \end{pmatrix}, t_{i+1}; \begin{pmatrix} X_i \\ h_i \end{pmatrix}, t_i, \theta\right) dP_{H_i, H_{i+1}|\mathcal{F}_i}(h_i, h_{i+1}).$$

As in the scalar case, the integrals are evaluated using Monte Carlo integration. Let $q\left(v_0, \binom{u_1}{v_1}, \binom{u_2}{v_2}, \ldots, \binom{u_{M-1}}{v_{M-1}}, v_M\right)$ be an importance sampler on $\mathbb{R}^{2M}$, and let

$$\mathbf{w}_k = \left(v_{k,0}, \begin{pmatrix} u_{k,1} \\ v_{k,1} \end{pmatrix}, \left(u_{k,2} v_{k,2}\right), \ldots, \begin{pmatrix} u_{k,M-1} \\ v_{k,M-1} \end{pmatrix}, v_{k,M}\right), \quad k = 1, 2, \ldots, K$$

be draws from q. Then one defines

$$p^{(M,K)}(X_{i+1}|\mathcal{F}_i, \theta) = \frac{1}{K} \sum_{k=1}^{K} \frac{\prod_{m=1}^{M} p^{(1)}\left(\binom{u_{k,m}}{v_{k,m}}, \tau_m; \binom{u_{k,m-1}}{v_{k,m-1}}, \tau_{m-1}, \theta\right)}{q\left(v_{k,0}, \binom{u_{k,1}}{v_{k,1}}, \binom{u_{k,2}}{v_{k,2}}, \ldots, \binom{u_{k,M-1}}{v_{k,M-1}}, v_{k,M}\right)}$$

with the convention $u_{k,0} = X_i$ and $u_{k,M} = X_{i+1}$ for all k.

The theoretical framework is essentially unchanged from the scalar case. In particular, sufficient conditions for

$$\lim_{M \to \infty} p^{(M)}(\cdot, t; z_s, s, \theta) = p(\cdot, t; z_s, s, \theta) \qquad \text{in } L^1(\lambda)$$

may be found in Pedersen (1995b). For a fixed realization $(x_0, x_1, \ldots, x_{i+1})$, one obtains (with standard regularity conditions)

$$\lim_{K \to \infty} p^{(M,K)}(x_{i+1}; \mathcal{F}_i, \theta) \to p^{(M)}(x_{i+1}; \mathcal{F}_i, \theta)$$

from the strong law of large numbers in the usual way.

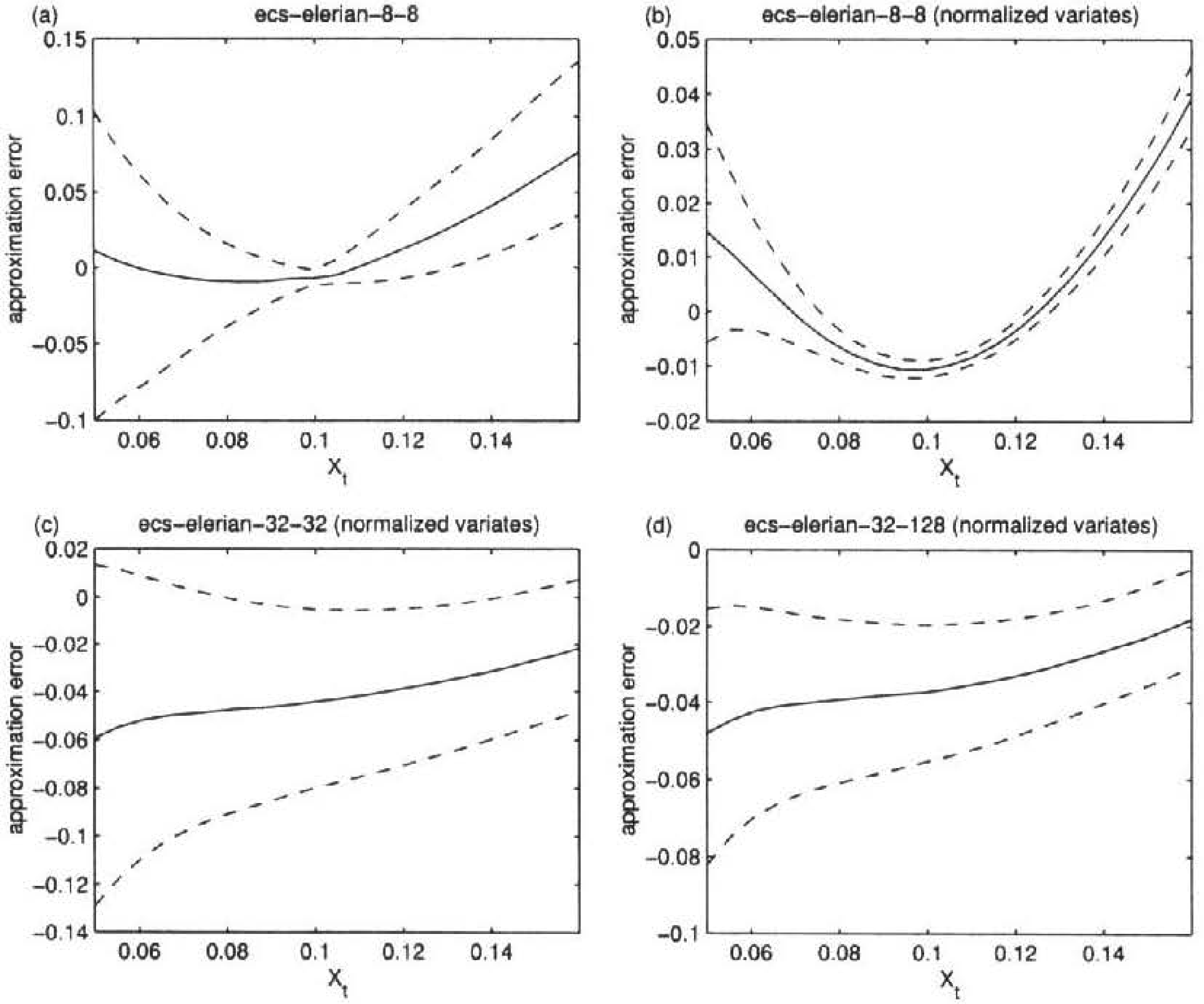


Figure 8. Approximation Error, $\log \hat{p}(X_t, t; X_s, s) - \log p(X_t, t; X_s, s)$, for Various Simulated Likelihood Schemes Given $\Delta = t - s = 1/12$, $X_s = .10$, and $\theta = (.5, .06, .15)$. The median and interquartile range over 128 repetitions are plotted. The untransformed model is used.

The issue of how to construct the importance sampler remains to be addressed. The approach taken in this article is as follows. Suppose that the distribution of $H_i|\mathcal{F}_i$ is known. We need a way to sample

$$\mathbf{w} = \left(v_0, \binom{u_1}{v_1}, \binom{u_2}{v_2}, \ldots, \binom{u_{M-1}}{v_{M-1}}, v_M\right).$$

First, draw v_0 from $H_i|\mathcal{F}_i$, and let $u_0 = X_i$. Now, proceed recursively for $m = 0, \ldots, M-2$: given (u_m, v_m), draw u_{m+1} using the modified Brownian bridge sampler described in Section 5:

$$u_{m+1} \sim N(u_m + \tilde{\mu}_{X,m}\delta, \tilde{\sigma}^2_{X,m}\delta)$$
$$\tilde{\mu}_{X,m} = \left(\frac{X_{i+1} - u_m}{t_{i+1} - \tau_m}\right)$$
$$\tilde{\sigma}^2_{X,m} = \sigma^2_X(u_m, v_m)\left(\frac{M - m - 1}{M - m}\right)$$

and draw v_{m+1} "blindly,"

$$v_{m+1} \sim N(v_m + \tilde{\mu}_{H,m}\delta, \tilde{\sigma}^2_{H,m}\delta)$$
$$\tilde{\mu}_{H,m} = \mu_H(u_m, v_m)$$
$$\tilde{\sigma}^2_{H,m} = \sigma^2_H(u_m, v_m)$$

where $\delta = (t_{i+1} - t_i)/M$. And finally, draw v_M using the same procedure as for $m = 1, \ldots, M-1$.

Now, we need to propagate the distribution of $H_i|\mathcal{F}_i$ forward. Consider the K draws $\{v_{k,M}, k = 1, \ldots, K\}$ from the sampler, together with the corresponding weights

$$\lambda_k = \frac{\prod_{m=1}^{M} p^{(1)}\left(\binom{u_{k,m}}{v_{k,m}}, \tau_m; \binom{u_{k,m-1}}{v_{k,m-1}}, \tau_{m-1}, \theta\right)}{q\left(v_{k,0}, \binom{u_{k,1}}{v_{k,1}}, \binom{u_{k,2}}{v_{k,2}}, \ldots, \binom{u_{k,M-1}}{v_{k,M-1}}, v_{k,M}\right)}.$$

Let $\{\bar{\lambda}_k, k = 1, \ldots, K\}$ denote the weights after normalizing so that they sum to 1. The collection $\{(v_{k,M}, \bar{\lambda}_k), k = 1, \ldots, K\}$ of points and weights may be thought of as representing a discrete approximation to the density $H_{i+1}|\mathcal{F}_{i+1}$.

Various techniques are available to form a distribution function from this approximation and draw points from it. While one could simply use the approximation

$$P_{H_{i+1}|\mathcal{F}_{i+1}}(h|\mathcal{F}_{i+1}) \approx \sum_{k: v_{k,M} \le h} \bar{\lambda}_k,$$

the resulting likelihood function is discontinuous in θ, and will cause difficulties for the optimizer. Therefore, we use a Hermite function to approximate the density, and draw points from it instead.

We repeat the Monte Carlo experiment of Eraker (2001) using the methodology described above. Synthetic datasets of length $n = 500$ are generated from the model

$$dX = (\theta_1 + \theta_2 X)\,dt + \exp(H/2)X^{1/2}\,dW_1$$
$$dH = (\theta_3 + \theta_4 H)\,dt + \theta_5 dW_2 \qquad (19)$$

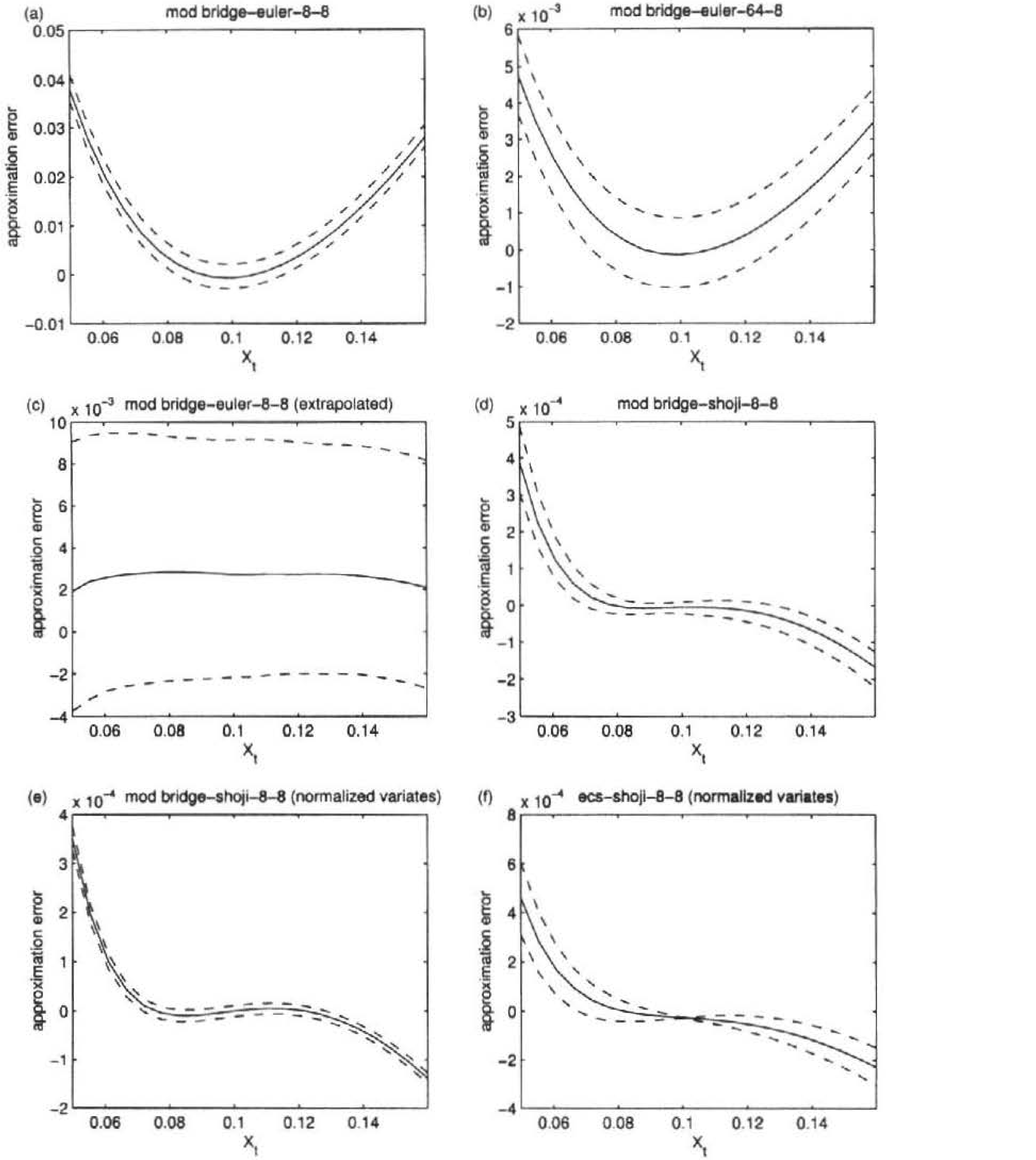


Figure 9. Approximation Error, $\log \hat{p}(X_t,t;X_s,s) - \log p(X_t,t;X_s,s)$, for Various Simulated Likelihood Schemes Given $\Delta = t - s = 1/12$, $X_s = .10$, and $\theta = (.5, .06, .15)$. The median and interquartile range over 128 repetitions are plotted. The transformed model is used.

with parameter vector $\theta^o = (.0002, -.002, -.3, -.03, .3)$. This model is said to be calibrated to the U.S. short-term interest rate with time measured in days. Data are generated using the Euler scheme with 50 subintervals. Estimates are obtained using $K = 128$ sample paths and $M = 8$ subintervals. Little is lost by reducing the number of sample paths to $K = 64$. We do not recommend using less than this. Increasing the number of subintervals does not improve the precision of the estimates significantly. The mean, standard error, and RMSE of parameter estimates over 512 Monte Carlo repetitions are shown in Table 4. For comparison, we reproduce Eraker's results as well. Our results are very similar. The computational cost is about 1.2 s per evaluation of the likelihood function on a 750 MHz PC.

Note that the estimates of θ_1 and θ_3 appear to be quite imprecise. This is due to the parametric form of the model. In contrast, the estimates of θ_1/θ_2 and θ_3/θ_4 are quite good. While we have maintained the model in the form used by Eraker, it might be better to instead use

$$dX = \theta_2(X - \mu)\,dt + \sigma \exp(H) X^{1/2}\,dW_1$$

$$dH = \theta_4 H\,dt + \theta_5 dW_2$$

where $\mu = -\theta_1/\theta_2$ and $\sigma = \exp(-\theta_3/\theta_4)$.

Table 2. Approximation Errors for Parameter Estimates

			θ_1	θ_2	θ_3	log L
(a)	$\theta^o = (.06, .5, .15),\ \Delta = 1/12,\ df^a = 5.33$					
	MLE-TRUE:	Mean	.00061	.04890	.00020	1.54137
		Std. err.	.00796	.12532	.00340	1.24511
		RMSE	.00797	.13442	.00341	1.98074
	MLE-EULER:	Mean	.00026	.01781	.00374	.19031
		Std. err.	.00003	.01103	.00076	.54499
		RMSE	.00026	.02095	.00382	.57700
	MLE-SMLE[b]: (Pedersen)	Mean	−.00062	−.00801	−.00223	8.07971
		Std. err.	.00256	.03681	.00175	4.76355
		RMSE	.00264	.03765	.00284	9.37802
	MLE-SMLE[c]:	Mean	.00000	−.00005	−.00000	−.00469
		Std. err.	.00000	.00073	.00001	.01340
		RMSE	.00000	.00073	.00001	.01419
(b)	$\theta^o = (.06, .5, .22),\ \Delta = 1/12,\ df^a = 2.48$					
	MLE-TRUE:	Mean	−.00031	.05973	.00012	1.55119
		Std. err.	.01136	.12808	.00535	1.25344
		RMSE	.01136	.14126	.00535	1.99386
	MLE-EULER:	Mean	.00069	.04941	.00701	3.01730
		Std. err.	.00014	.03111	.00136	1.90071
		RMSE	.00070	.05838	.00714	3.56548
	MLE-SMLE[c]:	Mean	.00001	−.00357	.00018	.03479
		Std. err.	.00014	.01678	.00119	.93214
		RMSE	.00014	.01714	.00120	.93187
(c)	$\theta^o = (.06, .5, .03),\ \Delta = 1/12,\ df^a = 133.33$					
	MLE-TRUE:	Mean	.00005	.04378	.00001	1.57035
		Std. err.	.00162	.12239	.00070	1.29601
		RMSE	.00162	.12992	.00070	2.03567
	MLE-EULER:	Mean	.00001	.01287	.00067	.00428
		Std. err.	.00000	.00595	.00015	.05930
		RMSE	.00001	.01418	.00069	.05942
	MLE-SMLE[c]:	Mean	.00000	−.00011	−.00000	−.00227
		Std. err.	.00000	.00005	.00000	.00107
		RMSE	.00000	.00013	.00000	.00251

NOTE: Results of Monte Carlo study assessing the quality of parameter estimates obtained using various techniques. Each experiment uses 512 replications, each over synthetic datasets of $n = 1{,}000$ observations. The goal is for the distance from the approximations to the exact MLE to be a small fraction of the distance from the exact MLE to the data-generating parameter.

[a] Degrees of freedom of the exact noncentral chi-square transition density.

[b] Sampler = Pedersen, subdensity = Euler, $M = 8$, $K = 256$, transformed model.

[c] Sampler = modified bridge, subdensity = Shoji and Ozaki, $M = 16$, $K = 16$, normalized variates, transformed model.

Note that our methodology can be easily used to estimate discrete time SV models by setting $M = 1$. We have tested the methodology using some of the models examined by Jacquier et al. (1994) and others with similar results.

8. APPLICATION

To illustrate the methodologies proposed in this article, we estimate some simple models of the short-term interest rate. We use the dataset previously examined by Gallant and Tauchen (1998), which consists of 1809 weekly observations of the three-month treasury bill rate (January 5, 1962–August 30, 1996). Rates are annualized and quoted on a discount basis. The data are plotted in Figure 10.

We first fit a simple scalar model, $dX = (\theta_1 + \theta_2 X)\,dt + \theta_3 X^{\theta_4}\,dW$, commonly referred to as the constant elasticity of volatility (CEV) model. It has been studied previously (using other estimators and data) by Chan et al. (1992), Tauchen (1995), Aït-Sahalia (1996), Conley, Hansen, Luttmer, and Scheinkman (1997), and others.

We also fit the stochastic volatility model given by

$$dX = (\theta_1 + \theta_2 X)\,dt + \theta_3 X^{\theta_4} e^{H}\,dW_1 \tag{20}$$

$$dH = \theta_5 H\,dt + \theta_6\,dW_2 \tag{21}$$

with W_1 and W_2 independent. Similar models have been examined by Gallant and Tauchen (1998), Andersen and Lund (1997), and Eraker (2001). Time is measured in years.

Maximum likelihood estimates and log likelihoods are given in Table 5. For the scalar model, we have used $M = 16$ and $K = 16$ with the modified bridge sampler, transformed model, normalized variance random scheme, and Shoji and Ozaki's subdensity. For the SV model, we use $M = 8$ and $K = 256$ with the techniques described in the preceding section. Since the scalar model is nested within the SV model, the restriction

Table 3. Approximation Errors for Parameter Estimates, Continued

			θ_1	θ_2	θ_3	log L
(d)			$\theta^o = (.06, .4, .15)$, $\Delta = 1/12$, $df^a = 4.27$			
	MLE-TRUE:	Mean	.00077	.04589	.00008	1.55436
		Std. err.	.01019	.11201	.00347	1.28351
		RMSE	.01022	.12099	.00347	2.01539
	MLE-EULER:	Mean	.00027	.01459	.00320	.27005
		Std. err.	.00005	.01066	.00072	.60164
		RMSE	.00027	.01807	.00328	.65920
	MLE-SMLE[c]:	Mean	−.00000	−.00011	−.00001	−.01275
		Std. err.	.00002	.00109	.00004	.03611
		RMSE	.00002	.00109	.00004	.03827
(e)			$\theta^o = (.06, .5, .15)$, $\Delta = 1/12$, $df^a = 53.33$			
	MLE-TRUE:	Mean	.00007	.01506	.00000	1.59353
		Std. err.	.00082	.46302	.00427	1.39283
		RMSE	.00082	.46300	.00427	2.11592
	MLE-EULER:	Mean	.00019	.93515	.00266	.31509
		Std. err.	.00001	.16274	.00257	.73929
		RMSE	.00019	.94919	.02679	.80325
	MLE-SMLE[c]:	Mean	.00000	−.00734	−.00005	−.14917
		Std. err.	.00000	.00423	.00002	.04223
		RMSE	.00000	.00847	.00005	.15502
(f)			$\theta^o = (.06, .5, .15)$, $\Delta = 2$, $df^a = 5.33$			
	MLE-TRUE:	Mean	.00036	.00126	.00038	1.52724
		Std. err.	.00173	.04305	.00564	1.24582
		RMSE	.00176	.04302	.00564	1.97013
	MLE-EULER:	Mean	.00367	.21218	.05275	21.19312
		Std. err.	.00020	.02967	.00458	5.66304
		RMSE	.00367	.21424	.05295	21.93523
	MLE-SMLE[c]:	Mean	.00006	.00712	.00041	−2.63215
		Std. err.	.00004	.00778	.00091	.95334
		RMSE	.00007	.01054	.00100	2.79916

[a] Degrees of freedom of the exact noncentral chi-square transition density.

[b] Sampler − Pedersen, subdensity = Euler, $M = 8$, $K = 256$, transformed model.

[c] Sampler = modified bridge, subdensity = Shoji and Ozaki, $M = 16$, $K = 16$, normalized variates, transformed model.

can be tested using, for example, the likelihood ratio statistic. Although setting $\theta_6 = 0$ causes θ_5 to become unidentified, this issue may be addressed along the lines of, for example, Gallant (1997) or Andrews and Ploberger (1994). In any event, the SV model results in a huge improvement in the log likelihood. The scalar model does not appear to be plausible.

Notice that the estimates for θ_1 and θ_2 are insignificantly different from zero in both models. In a more exhaustive study,

Table 4. Monte Carlo Study for SV Model

	θ_1	θ_2	θ_1/θ_2	θ_3	θ_4	θ_3/θ_4	θ_5
True	.00020	−.00200	−.10000	−.30000	−.03000	.10000	.30000
				SMLE			
Mean	.00089	−.00886	−.10319	−.43425	−.04350	.10040	.31734
RMSE	.00102	.00986	.05523	.27931	.02802	.00493	.06888
				Eraker (2001)			
Mean	.00127	−.01271		−.38174	−.03873		.24297
RMSE	.00154	.01468		.21844	.02209		.07173

NOTE: Mean and root mean squared error of parameters estimated on synthetic data ($n = 500$) generated from (19). The sampling frequency is $\Delta t = 1$. The parameters are calibrated to match U.S. short-term interest rates with time measured in days. The Monte Carlo experiment was run for 512 repetitions. Monte Carlo results from Eraker (2001) are included for comparison.

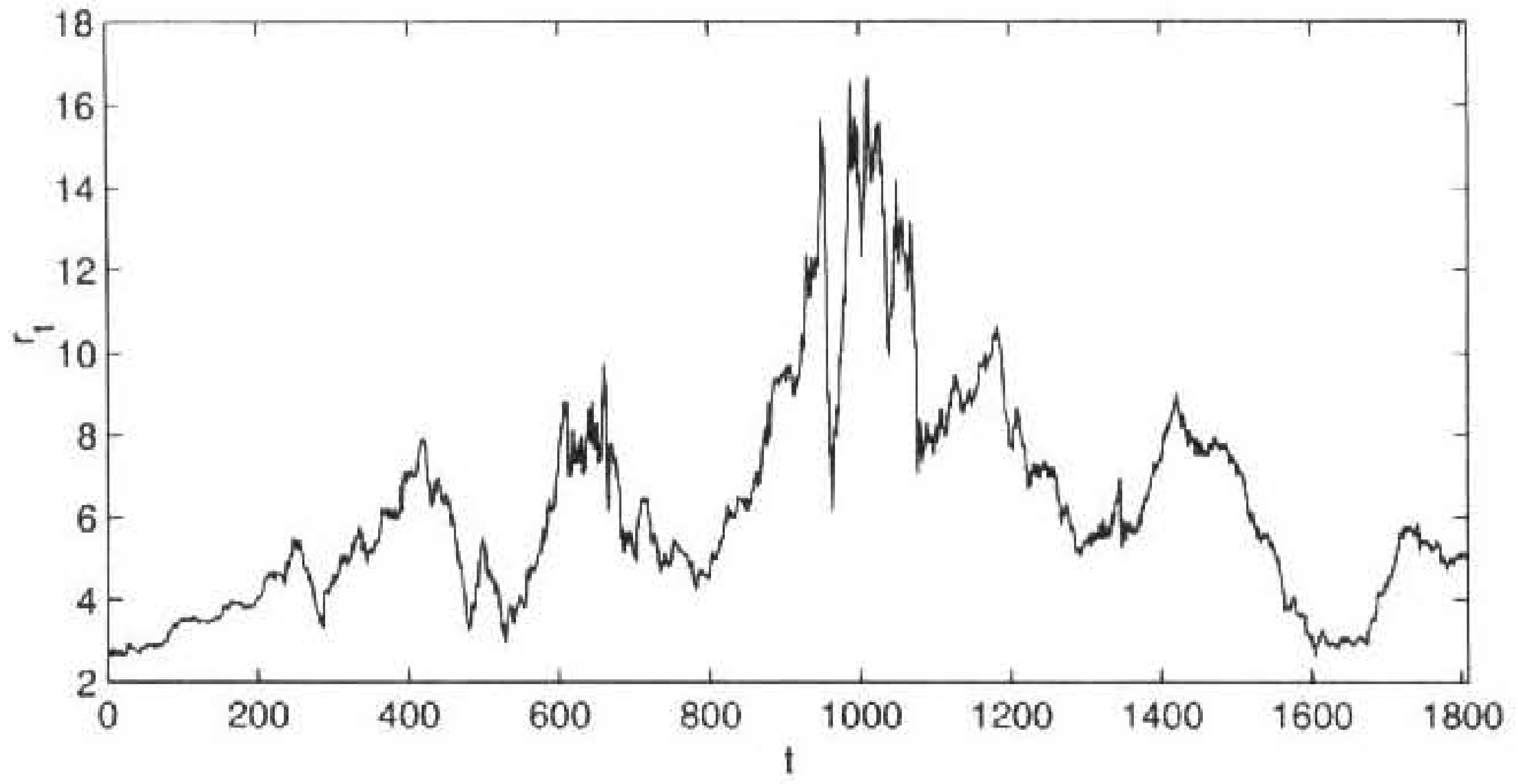


Figure 10. Weekly Observations of the Three-Month Treasury Bill Rate, 1/5/1962–8/30/1996, n = 1809.

Durham (2000) finds that the constant term in the drift is needed to avoid having an attracting boundary at zero, but that the benefits of including additional terms are negligible. This is true for both the scalar and stochastic volatility models.

Notice also that the estimates for θ_3 and θ_4 are similar for both models (the unconditional mean of H is 0, and thus $\exp(H)$ vacillates around 1). The estimates are slightly lower in the SV model since this model generates conditional densities with thicker tails, which helps to catch the extreme events. These parameters are estimated quite precisely in the scalar model. Including the unobserved component reduces the precision of these estimates, but not to an unacceptable degree. There is a great deal of covariance between $\hat{\theta}_4$ and $\hat{\theta}_5$ in the SV model. Volatility tends to be high when interests rates are high; the estimator has difficulty distinguishing whether this is due to persistence or a level effect. Much of the lack of precision in these estimates appears to be due to this.

The volatility component is an Ornstein–Uhlenbeck process. Its unconditional distribution is Gaussian with mean zero and variance $\theta_6^2/(-2\theta_5) \approx .34$. The mean reversion parameter is about -4, which corresponds to a half-life of about two months.

9. CONCLUSIONS

Despite the theoretical advantages of maximum likelihood estimation, it has been seldom used in estimating continuous-time diffusion models. The transition densities are not known for most models of interest, and previously available approximation techniques have either been of questionable accuracy or computationally intensive.

The simulation-based approach suggested by Pedersen (1995b) and Santa-Clara (1995) is appealing from a theoretical and intuitive viewpoint; however, we find that it can be prohibitively costly to attain even the accuracy of the simple first-order approximation. Eraker (2001), Jones (1999a), and Elerian et al. (2001) propose interesting MCMC approaches to estimating diffusion models. Elerian et al. suggest the idea of improved importance sampling to accelerate the convergence of the Monte Carlo integration at the heart of the simulated likelihood approach. However, computational cost remains high.

We build upon this work, examining other importance samplers, alternate random schemes, higher order subtransition densities, extrapolation, and a variance-stabilizing transformation of the model. Combining these ideas results in highly efficient approximations. When applied to synthetic data ($n = 1{,}000$) generated by a CIR model with parameters calibrated to match monthly observations of the U.S. short-term interest rate, we are able to obtain maximum likelihood estimates with a negligible approximation error in well under 1 min. The log-likelihood function itself can be approximated with great accuracy in about .1 s.

Our results suggest that the best performance is obtained using the modified bridge sampler with the subtransition density of Shoji and Ozaki applied to the transformed model. The number of subintervals M and sample paths K must be determined by experimentation.

Future work will undoubtedly uncover further refinements in these techniques, as well as point to situations where one

Table 5. CEV and SV Models Fitted to Weekly Observations of the Three-Month Treasury Bill Rate, Jan. 5, 1962–Aug. 30, 1996

Model	log L	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
CEV	512.32	.8277 (.5004)	−.1049 (.1167)	.1032 (.0038)	1.4502 (.0210)		
SV	913.95	.176 (.291)	.019 (.066)	.105 (.028)	1.247 (.154)	−4.104 (.747)	1.682 (.144)

NOTE: Parameter estimates for CEV model and the SV model in Equation (20). Standard errors are in parentheses below the parameter estimates. Time is measured in years.

or the other particular variant may be preferred. While the main focus of this article is on scalar models, an approach to applying some of the ideas to estimate a two-factor, latent variable model is also proposed. Extending these techniques to models with jump components would also be of interest.

ACKNOWLEDGMENTS

The authors are grateful for the helpful comments and suggestions of Yacine Aït-Sahalia, Evan Anderson, Tim Bollerslev, Michael Brandt, Ai-Ru Cheng, Frank Diebold, Eric Ghysels, Adrian Pagan, Neil Shephard, George Tauchen, and seminar participants at the NBER Summer Institute 2000 and the University of Pennsylvania.

[Received November 2001. Revised November 2001.]

REFERENCES

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

——— (2001), "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approach," *Econometrica*, 70, 223–262.

Andersen, T. G., and Lund, J. (1977), "Estimating Continuous-Time Stochastic Volatility Models of the Short-Term Interest Rate," *Journal of Econometrics*, 77, 343–377.

Andrews, D. W. K., and Ploberger, W. (1994), "Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative," *Econometrica*, 62, 1383–1414.

Bibby, B. M., and Sørensen, M. (1995), "Martingale Estimating Functions for Discretely Observed Diffusion Processes," *Bernoulli*, 1, 17–39.

Brandt, M. W., and Santa-Clara, P. (2002), "Simulated Likelihood Estimation of Diffusions With an Application to Exchange Rate Dynamics in Incomplete Markets," *Journal of Financial Economics*, 63, 161–210.

Chacko, G., and Viceira, L. M. (1998), "Spectral GMM Estimation of Continuous-Time Processes," Working Paper, Harvard Business School.

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992), "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate," *Journal of Finance*, 47, 1209–1228.

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

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

Danielsson, J., and Richard, J.-F. (1993), "Accelerated Gaussian Importance Sampler With Application to Dynamic Latent Variable Models," *Journal of Applied Econometrics*, 8, 153–173.

Doss, H. (1977), "Liens Entre Équations Différentielles Stochastiques et Ordinaires," *Ann. Inst. H. Poincaré*, 13, 99–125.

Duffie, D., and Glynn, P. (1996), "Estimation of Continuous-Time Markov Processes Sampled at Random Time Intervals," Working Paper, Graduate School of Business, Stanford University.

Duffie, D., and Singleton, K. J. (1993), "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61, 929–952.

Durham, G. B. (2000), "Specification Analysis of the Short-Term Interest Rate Using Likelihood-Based Techniques," Working Paper, Department of Economics, University of North Carolina.

Elerian, O. (1998), "A Note on the Existence of a Closed Form Conditional Transition Density for the Milstein Scheme," Working Paper, Nuffield College, Oxford University.

——— (1999), "Simulation Estimation of Continuous-Time Models With Applications to Finance," Ph.D. dissertation, Nuffield College, University of Oxford.

Elerian, O., Chib, S., and Shephard, N. (2001), "Likelihood Inference for Discretely Observed Non-Linear Diffusions," *Econometrica*, 69, 959–993.

Eraker, B. (2001), "MCMC Analysis of Diffusion Models With Application to Finance," *Journal of Business and Economic Statistics*, 19, 177–191.

Florens-Zmirou, D. (1989), "Approximate Discrete-Time Schemes for Statistics of Diffusion Processes," *Statistics*, 20, 547–557.

Gallant, A. R. (1977), "Testing a Nonlinear Regression Specification: A Nonregular Case," *Journal of the American Statistical Association*, 72, 523–530.

Gallant, A. R., and Tauchen, G. (1997), "Estimation of Continuous-Time Models for Stock Returns and Interest Rates," *Macroeconomic Dynamics*, 1, 135–168.

——— (1998), "Reprojecting Partially Observed Systems With Application to Interest Rate Diffusions," *Journal of the American Statistical Association*, 93, 10–24.

——— (1996), "Which Moments to Match?" *Econometric Theory*, 12, 657–681.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1339.

Ghysels, E., Harvey, A., and Renault, E. (1996), "Stochastic Volatility," in *Handbook of Statistics 14, Statistical Methods in Finance*, eds. G. S. Maddala and C. R. Rao, Amsterdam: North-Holland.

Gouriéroux, C., Monfort, A., and Renault, E. (1993), "Indirect Inference," *Journal of Applied Econometrics*, 8, S85–S118.

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

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994), "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics*, 12, 371–389.

Jones, C. S. (1999a), "Bayesian Estimation of Continuous-Time Finance Models," Working Paper, Simon School of Business, University of Rochester.

——— (1999b), "The Dynamics of Stochastic Volatility," Working Paper (Nov. 1999 version), Simon School of Business, University of Rochester.

Karatzas, I., and Shreve, S. E. (1991), *Brownian Motion and Stochastic Calculus*, 2nd ed. New York: Springer.

Kessler, M. (1997), "Estimation of an Ergodic Diffusion From Discrete Observations," *Scandinavian Journal of Statistics*, 24, 211–229.

Kim, S., Shephard, N., and Chib, S. (1998), "Stochastic Volatility: Likelihood Inference and Comparison With ARCH Models," *Review of Economic Studies*, 65, 361–393.

Kloeden, P., and Platen, E. (1992), *Numerical Solution of Stochastic Differential Equations*, Berlin: Springer-Verlag.

Lo, A. W., (1988), "Maximum Likelihood Estimation of Generalized Ito Processes With Discretely Sampled Data," *Econometric Theory*, 4, 231–247.

Milstein, G. (1978), "A Method of Second Order Accuracy Integration of Stochastic Differential Equations," *Theory of Probability and Its Applications*, 23, 396–401.

Nowman, K. (1997), "Gaussian Estimation of Single-Factor Continuous Time Models of the Term Structure of Interest Rates," *Journal of Finance*, 52, 1695–1706.

Pedersen, A. R. (1995a), "Consistency and Asymptotic Normality of an Approximate Maximum Likelihood Estimator for Discretely Observed Diffusion Processes," *Bernoulli*, 1, 257–279.

——— (1995b), "A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations," *Scandinavian Journal of Statistics*, 22, 55–71.

Pitt, M. K., and Shephard, N. (1999), "Filtering via Simulation: Auxiliary Particle Filter," *Journal of the American Statistical Association*, 446, 590–599.

Richard, J.-F., and Zhang, W. (2000), "Accelerated Monte Carlo Integration: An Application to Dynamic Latent Variables Models," in *Simulation-Based Inference in Economics*, eds. R. S. Mariano, T. Schuermann, and M. Weeks, Cambridge, U.K.: Cambridge University Press, chap. 2, pp. 47–70.

Sandmann, G., and Koopman, S. J. (1998), "Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood," *Journal of Econometrics*, 87, 271–301.

Santa-Clara, P. (1995), "Simulated Likelihood Estimation of Diffusions With an Application to the Short-Term Interest Rate," Working Paper, Anderson Graduate School of Management, UCLA.

Shoji, I., and Ozaki, T. (1998), "Estimation for Nonlinear Stochastic Differential Equations by a Local Linearization Method," *Stochastic Analysis and Applications*, 16, 733–752.

Singleton, K. J. (1997), "Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function," Working Paper (Aug. 2000 version), Stanford University.

Tauchen, G. E. (1995), "New Minimum Chi-Square Methods in Empirical Finance," in *Advances in Econometrics*, eds. K. Wallace and D. Kreps, Cambridge, U.K.: Cambridge University Press, chap. 9, pp. 279–317.