Nonparametric Estimation of Multifactor Continuous Time Interest Rate Models

Chris Downing
Board of Governors of the Federal Reserve System *

*I thank Matt Pritsker, Torben Andersen, Jesper Lund, and the participants at the 1998 Conference of the Society for Computational Economics for helpful discussions. I remain responsible for any errors. The views expressed in this paper are those of the author and are not necessarily those of the Board of Governors or members of its staff. Address correspondence to Chris Downing, Federal Reserve Board, Mail Stop 89, Washington, DC 20551. The author may also be reached by phone: (202) 452-2378, fax: (202) 452-5296, or e-mail: cdowning@frb.gov.

Abstract

This paper studies the finite sample properties of the kernel regression method of Boudoukh, Richardson, Stanton and Whitelaw (1998) for estimating multifactor continuous–time term structure models. Monte Carlo simulations are employed, with a grid-search technique to find the optimal kernel bandwidth. The performance of the estimator is also studied under model misspecification. Irrelevant regressors reduce efficiency and induce additional biases in the estimates. Using Treasury bill data, I test whether the estimates produced by the nonparametric estimator are statistically distinguishable from estimates obtained under a parametric model. The kernel regressions pick up nonlinearities that the parametric model cannot capture.

In a series of recent papers, researchers in finance have developed nonparametric methods for estimating the drift and diffusion functions of continuous time stochastic processes. Stanton (1997) pioneered a method based on the theory of weak approximations of the expectations of functions of stochastic processes. His methodological innovation was to estimate the expectations using kernel regression methods, and then invert them in order to recover the drift and diffusion functions of the underlying processes. The method has been applied to the problem of estimating univariate continuous time models of the term structure. More recently, Boudoukh et al. (1998) extended the estimator to the problem of estimating multivariate term structure models. Although different in important respects, the method developed by Ait–Sahalia (1996) is related to the Stanton and BRSW estimators in that it also relies on nonparametric techniques and is also applied to the problem of pricing interest rate derivative securities.[1]

One of the more provocative conclusions reached by Ait–Sahalia (1996), Stanton (1997), and Boudoukh et al. (1998) is that the short rate drift appears to be nonlinear. This conclusion is at odds with the rest of the term structure literature, because in virtually all previous work, the short rate is modeled with a linear drift. In part to investigate the robustness of this result, Pritsker (1998) and Chapman and Pearson (1999) look at the properties of the Stanton and Ait–Sahalia estimators in finite samples. In both of these papers, the authors concluded that the nonlinearity result is not robust, and *could* be an artifact of the finite sample properties of the estimator. However, the authors do not formally test this hypothesis.

In this paper, I study the finite sample properties of the BRSW estimator for multifactor models.[2] Monte Carlo simulations of data from the stochastic volatility model of Andersen and Lund (1997a) are used to examine how closely the estimator fits the known drift and diffusion functions. The Andersen and Lund (1997a) model is used because it provides a reasonably good fit to Treasury data, although in their final analysis the authors reject the model using a chi–squared test.

I first focus on the problem of kernel bandwidth selection. Because the asymptotically optimal bandwidths are functions of the derivatives of the unknown joint density of the data generating process, I use a grid–search technique to find the bandwidths that minimize a sum of squared errors criterion. I find that, even with the optimal bandwidths, the estimator exhibits a high degree of bias with forty years of data simulated at weekly frequency. However, the sampling variance of the estimator is high, so that from a hy-

pothesis testing point of view, the biases are likely to be irrelevant.

The performance of the BRSW estimator is also analyzed under model misspecification. The results show that if one uses the BRSW estimator to fit a misspecified model in which irrelevant arguments of the drift and diffusion functions are included, the efficiency of the estimator decreases markedly. Somewhat more surprising is the result that including irrelevant conditioning variables introduces additional bias in the estimates. The additional biases are a result of adding dimensions along which biases from truncation and correlated residuals can affect the estimator.

These biases and inefficiencies highlight that, while nonparametric estimators might free one from the need to specify the particular functional forms for the various estimands, one still must correctly specify the arguments of the functions (and thus the correct set of conditioning variables in the kernel regressions). In other words, nonparametric estimators do not obviate issues of specification; rather, such issues are removed to a higher level of generality.

My main conclusion is that the BRSW estimator, and related kernel regression methods, are primarily useful as diagnostic tools when used in the context of term structure modeling. Given the problems associated with bandwidth selection when the data are autocorrelated, and given the problems of calculating reliable standard errors for kernel regression estimators, it is more productive to use the kernel regression methods to test if a given parametric specification is an adequate description of the data. In other words, the more general kernel regression estimator can be used to try and "pick up" nonlinearities in the data that a parametric model might miss. An important advantage of this approach is that the finite sample distributions of test statistics based on the BRSW estimator can be bootstrapped under the null hypothesis that the parametric model is the "true" data generating process. Thus, one can produce quantiles for the hypothesis test statistics that are robust against finite sample biases in the BRSW estimator. I demonstrate this by applying the BRSW estimator to test the Andersen and Lund (1997a) model of the term structure. The results of the hypothesis tests show that the biases of the BRSW estimator do not fully explain the differences between the estimates obtained under the BRSW estimator and the parametric estimator. There appear to be significant nonlinearities in the evolution of the short rate that the parametric model cannot capture.

This paper is organized as follows. In the next section, I examine the dynamic behavior of the Andersen and Lund (1997a) stochastic volatility model, which is used in the Monte Carlo simulations in the following sec-

tions. Section II discusses the BRSW estimator and kernel regression, and contains the main results on fitting the Andersen and Lund (1997a) model. Section III discusses the performance of the estimator in the context of model misspecification. Section IV presents the results of hypothesis tests on the Andersen and Lund (1997a) model, and the final section concludes.

I Dynamic Behavior of the Stochastic Volatility Model

In this section, I discuss the calculation of weak solutions of the Andersen and Lund (1997a) model (henceforth, the "AL model"). An interesting feature of the AL model is that it fails to satisfy the conditions sufficient to guarantee the existence of a unique solution, raising questions about the stability of the system, as well as questions about the existence of a stationary density. Maintaining the assumption that the system has a solution, I use a weak numeric solution algorithm and an extension of the Kolmogorov-Smirnov test to determine whether or not the transition densities of the system converge at long trajectories. From the results, we can conclude that the system has a stationary density at the parameters considered.

The specification of the AL model is given as:

$$dr_t = \kappa_1(\mu - r_t)dt + \sigma_t\sqrt{r_t}dW_{1,t} \quad (1)$$

$$d\log\sigma_t^2 = \kappa_2(\theta - \log\sigma_t^2)dt + \xi dW_{2,t}, \quad (2)$$

where W_1 and W_2 are independent standard Wiener processes.

The set of sufficient conditions for the existence of a solution to this system includes the conditions that the drift and diffusion functions satisfy *Lipschitz* and *growth* conditions (see Karatzas and Shreve (1991) and Ait–Sahalia (1996) for different formulations of the conditions). The specification of the diffusion function of the interest rate process (1) causes the system to violate the growth condition. The relevant condition is given by:

$$\sigma^2 r + \xi^2 \leq k(1 + r^2 + (\log\sigma^2)^2). \quad (3)$$

This condition must apply uniformly in t, meaning that the constant k must apply for all t simultaneously. It is easy to show that there is no k that satisfies condition (3). For any k, let $\log\sigma^2 = r$, so that $\sigma^2 = e^r$. Substituting,

we have:

$$\begin{aligned} e^r r + \xi^2 &\leq k(1 + 2r^2), \text{ or} \quad (4) \\ \frac{e^r r + \xi^2}{(1 + 2r^2)} &\leq k. \quad (5) \end{aligned}$$

The left-hand side of (5) clearly diverges as $r \to \infty$, showing that the growth condition is violated by the model. In essence, the model fails to satisfy the growth condition because the diffusion function in the interest rate process involves an exponential transformation of the volatility state variable.

To make the exponential transform in the interest rate diffusion explicit, rewrite the AL model in the following equivalent form [3]:

$$\begin{aligned} dr_t &= \kappa_1(\mu - r_t)dt + \sqrt{e^{\sigma_t} r_t} dW_{1,t} \quad (6) \\ d\sigma_t &= \kappa_2(\theta - \sigma_t)dt + \xi dW_{2,t}, \quad (7) \end{aligned}$$

Because it fails to satisfy the growth condition, there might not be a unique Ito process in $\Re^2$ that satisfies (6) – (7). In practice, it's difficult to use numeric methods to verify the existence of a unique solution. I assume that a solution exists, and instead focus on the dynamic stability of the system. For certain parameterizations of the drift and diffusion functions, the model will exhibit explosive behavior, and thus fail to have a stationary density. Determining whether or not the model is explosive is a problem to which we can apply a numeric solution algorithm.

Kloeden and Platen (1995) derive a number of algorithms for computing weak solutions of systems of SDEs like the AL model. The solution algorithms operate on a finite time interval $[0, T]$. A key feature of the algorithms is the discretization of the time interval into M smaller time steps of length Δ, where $\Delta = \frac{T}{M}$. The simplest method is the Euler scheme, which has a degree of accuracy that is inversely proportional to the length of the time step Δ. The following set of recursive formulae show how to generate values of r and σ:

$$\begin{aligned} r_t &= r_{t-1} + \kappa_1(\mu - r_{t-1})\Delta + \sqrt{e^{\sigma_{t-1}} r_{t-1} \Delta} \eta_{1,t} \quad (8) \\ \sigma_t &= \sigma_{t-1} + \kappa_2(\theta - \sigma_{t-1})\Delta + \xi\sqrt{\Delta}\eta_{2,t}, \quad (9) \end{aligned}$$

where $\eta_{1,t}$ and $\eta_{2,t}$ are independent standard normal deviates, and r_0 and σ_0^2 are given. Where necessary, I'll use $\tilde{r}$ and $\tilde{\sigma}$ to indicate values of r and σ computed from the discrete system in (8) and (9).

Understanding the dynamic behavior of the AL model, as well as evaluating the nonparametric estimator in the next section, both boil down to computing the expectations of different functions of the state variables r and σ:

$$E\left[f(r_T, \sigma_T)\right], \tag{10}$$

where $f(\cdot)$ is a smooth function. Kloeden and Platen (1995) prove that the expectation of $f(\cdot)$, calculated at $(\tilde{r}_T,\tilde{\sigma}_T)$, converges to the true expectation as $\Delta \to 0$:

$$\lim_{\Delta \to 0} \left| E\left[f(r_T, \sigma_T)\right] - E\left[f(\tilde{r}_T, \tilde{\sigma}_T)\right] \right| = 0. \tag{11}$$

By choosing

$$f(r, \sigma) = (r, \sigma), \tag{12}$$

we can use the Euler scheme to compute the moments of transition densities of the AL model.

It is useful to first consider whether or not the transition densities appear to be converging in location and scale. To do so, I use Monte Carlo simulations to generate moments of the transition densities of the model. From each of 25 different starting points, equally dispersed on the square of values:

$$\left\{(r, \sigma) : 0.02 \leq r \leq 0.20, -7.00 \leq \log \sigma^2 \leq -5.00\right\} \tag{13}$$

I simulate 1,000 batches of 100 trajectories. The last point of each trajectory is saved, forming a batch of 100 draws from the transition density defined by the starting point and the length of the trajectories. I compute the mean and variance of each batch of saved points. Thus, at the end of a run, we have 1,000 independent draws of the first two moments of each of the 25 transition densities. Eight such runs are completed, the first with trajectories one year in length, the second with five year trajectories, and so on for ten, twenty, thirty, forty, fifty, and finally sixty year trajectories. The parameters employed are shown in table I, and $\Delta = \frac{1}{52}$.[5]

Table II displays univariate statistics for the pooled data ($N = 25,000$), with which we can perform some unscientific "eyeball tests" for convergence. If the null hypothesis of convergence is correct, the moments of the transition densities should converge to the moments of the stationary density. The means should converge as follows:

$$\lim_{T \to \infty} E[r_T] = \mu = 0.0596, \tag{14}$$

$$\lim_{T \to \infty} E[\sigma_T] = \theta = -6.3599. \tag{15}$$

Examining the values in the second column (labeled 'Mean') of table II, it's clear that the first moments ($E[\cdot]$ values) of the transition densities are converging to these values. The interest rate mean hits the value in (14) at around thirty years, and then bounces around within a narrow confidence interval. The volatility mean converges quite rapidly and very precisely to the value in (15), reflecting the higher degree of mean reversion in the volatility drift function. [6]

The second moments should converge approximately as follows:

$$\lim_{T\to\infty} \mathrm{Var}[r_T] \approx 0.00032 \tag{16}$$

$$\lim_{T\to\infty} \mathrm{Var}[\sigma_T] \approx 0.7780, \tag{17}$$

The approximate value for the second moment of r is calculated as the variance of the stationary density of a square–root process:

$$dr_t = \kappa_1(\mu - r_t)dt + \sigma\sqrt{r_t}dW_t \tag{18}$$

with σ fixed at e^θ. The variance is given by $\frac{e^\theta \mu}{2\kappa_1}$. The approximate value for the second moment of σ is calculated as the stationary variance of a constant diffusion process:

$$d\sigma_t = \kappa_2(\theta - \sigma_t)dt + \xi dW_t. \tag{19}$$

The variance of this process is given by $\frac{\xi^2}{2\kappa_2}$. From column seven of table II, it appears that the variances ($\mathrm{Var}[\cdot]$ values) are converging to neighborhoods of the values in (16) and (17). In the case of the interest rate process, we would probably reject the null hypothesis that the variance is equal to the value in (16), even for the sixty year trajectories. Of course, this is because the process is not really the square–root process that we used to compute the variance. For the volatility process, we would probably accept the null hypothesis that the variance is equal to the value in (17). The means of $\mathrm{Var}[\sigma.]$ are close to the value in (17), and the standard deviation around the means is relatively large. The variance of the volatility process converges to the value in (17), while the variance of the interest rate process does not converge to (16), because the dependence between the interest rate and volatility processes is expressed in the diffusion function of the interest rate process. The volatility process does in fact evolve like the Vasicek process that we used to compute the variance in (17).

While the transition densities appear to be converging in the first two moments, they still might have different distribution functions. Moreover,

it's hard to assess joint significance using table II. Assuming that a solution to the system exists, we would like to show that the system is stationary, defined to mean that the transition densities converge to a common density with finite moments, as the length of the time interval increases:

$$\lim_{T\to\infty} \pi(r_T, \sigma_T | r_0, \sigma_0) \xrightarrow{d} \pi(r, \sigma), \tag{20}$$

for $r_0 \in \Re^{++}$ and $\sigma_0 \in \Re$, and where $\pi(r_T, \sigma_T | r_0, \sigma_0)$ is the transition density between times 0 and T, and $\pi(r, \sigma)$ is the stationary density. If we use the discrete system in (8)-(9) to make draws from the transition densities defined by different starting points (r_0, σ_0) and time intervals $[0, T]$, and these densities exhibit convergence as T increases, then we can interpret this as evidence supporting our hypothesis that the system has a stationary density at the parameter values in table I.[4]

To rigorously test for convergence in distribution when the true distribution is unknown, we can use an adaptation of the Kolmogorov- Smirnov (KS) test for bivariate densities, due to Fasano and Franceschini (1987). The one dimensional KS test statistic is based on the maximum value of the absolute difference between two cumulative distribution functions. A direct generalization of this statistic to higher dimensions is not possible because cumulative probability is not well defined in more than one dimension. However, an analogous statistic can be based on the integrated probabilities in each the four natural quadrants at a given point (r_i, σ_i). The analog to the KS statistic is the maximum difference over the data points and over the quadrants of the integrated probabilities. In essence, the algorithm for computing the statistic searches through the data for the point at which the difference in the proportions of data in one of the four natural quadrants formed by the point is maximized. Fasano and Franceschini (1987) work out an approximation to the probability of realizing the observed maximum difference in proportions, under the null hypothesis that the two densities are identical.[7]

The transition densities of the discrete system in (8)-(9) converge to the transition densities of the continuous–time system at rate $\sqrt{\Delta}$ (see Kloeden and Platen (1995) or Brandt and Santa–Clara (1999) for proofs). Thus, the discrete system can be used to draw random samples from transition densities that closely approximate the densities of the continuous–time system. To carry out the convergence test, I use two starting values that are widely

apart on the (r, σ) plane. The points that I use are:

$$\{(\mu + 2\hat{\sigma}_r, \theta + 2\hat{\sigma}_\sigma), (\mu - 2\hat{\sigma}_r, \theta - 2\hat{\sigma}_\sigma)\}. \tag{21}$$

The points are two standard deviations away from the long-run means of the processes, and four standard deviations from one another.[8] The standard deviations $\hat{\sigma}_r$ and $\hat{\sigma}_\sigma$ are approximated using the square roots of the values for $\text{Var}[r_{60}]$ and $\text{Var}[\sigma_{60}]$ from table II, respectively. From each of these points, I use the discrete system in (8)-(9) to simulate 20,000 trajectories, saving the last point on each trajectory. The two sets of points form large samples of the two transition densities. The bivariate KS test is applied to the two samples to test whether or not they are drawn from identical distributions. I repeat this exercise for trajectories of lengths between one and forty years. The parameterization of the system and the length of the time step are the same as before.[9]

Table III displays the results. The first column gives the trajectory lengths in years. The second and third columns display the bivariate KS test statistic and the approximate p–value, respectively. From the results, we can conclude that the two distribution functions become indistinguishable after forty years. The approximation to the p–value becomes imprecise for values above 0.2. However, given the large sample sizes, and the results from table II, we can conclude with a high degree of confidence that the system does in fact have a stationary density.

The length of time at which the transition densities appear to converge is consistent with the behavior of the system reported in Andersen and Lund (1997a). In order to simulate draws from the stationary density, Andersen and Lund (1997a) ran the Euler simulator for approximately thirty-eight years. The authors found that using longer trajectories had no significant effects on their results. Their results are consistent with the finding here that the distributions converge at trajectories of around forty years in length.[10]

To sum up, it is reasonable to conclude that, at the parameter values considered here, the AL model is stable and has a stationary density. Both of these features are prerequisites for the consistency of the BRSW estimator, and we will make use of some of the results in table II in what follows. In the next section, we turn to considering the behavior of the BRSW estimator in finite samples.

II Nonparametric Estimation

Assume that the term structure is determined by two state variables, the short rate r and the volatility of the short rate σ:

$$dr_t = \alpha_r(r_t, \sigma_t)dt + \beta_r(r_t, \sigma_t)dW_{r,t}, \quad (22)$$
$$d\sigma_t = \alpha_\sigma(r_t, \sigma_t)dt + \beta_\sigma(r_t, \sigma_t)dW_{\sigma,t}, \quad (23)$$

where $W_{r,t}$ and $W_{\sigma,t}$ are independent Wiener processes, and suppose that we observe data generated from these processes at discrete time intervals of length Δ. The Euler method of the previous section is one way to relate our discrete observations to the drift and diffusion functions of the continuous–time processes. The Euler discretization for this system is given by:[11]

$$r_{t+1} - r_t = \alpha_r\Delta + \beta_r\sqrt{\Delta}\eta_{r,t+1}, \quad (24)$$
$$\sigma_{t+1} - \sigma_t = \alpha_\sigma\Delta + \beta_\sigma\sqrt{\Delta}\eta_{\sigma,t+1}, \quad (25)$$

where, as before, η_r and η_σ are independent standard normal deviates. It's easy to see that the observations in equations (24) and (25) satisfy the following relationships:

$$\frac{1}{\Delta}E\left[r_{t+1} - r_t|F_t\right] = \alpha_r + O(\Delta), \quad (26)$$
$$\frac{1}{\Delta}E\left[\sigma_{t+1} - \sigma_t|F_t\right] = \alpha_\sigma + O(\Delta), \quad (27)$$
$$\frac{1}{\Delta}E\left[(r_{t+1} - r_t)^2|F_t\right] = \beta_r^2 + O(\Delta), \quad (28)$$
$$\frac{1}{\Delta}E\left[(\sigma_{t+1} - \sigma_t)^2|F_t\right] = \beta_\sigma^2 + O(\Delta), \quad (29)$$

where $O(\Delta)$ means terms for which it is true that $\lim_{\Delta\to 0}\frac{O(\Delta)}{\Delta} < \infty$, and F_t denotes the information set at time t. The methodological innovation of Boudoukh et al. (1998) is to note that, if we compute estimates of the first and second conditional moments on the left hand sides of equations (26) - (29), we will have estimates of the drift and diffusion functions accurate to $O(\Delta)$.

In order to estimate the conditional moments in equations (26)-(29) with minimal *a priori* structure on the drift and diffusion functions, a kernel regression method is used. First, we define a grid of interest rate and volatility values at which to estimate the conditional moments. Then, at each grid

value (r_i, σ_j), the estimates of the conditional moments are computed as follows:

$$E\left[r_{i,t+1} - r_{i,t} | (r_i, \sigma_j)\right] = \sum_{t=1}^{T-1} W(t)(r_{t+1} - r_t) \tag{30}$$

$$E\left[\sigma_{i,t+1} - \sigma_{i,t} | (r_i, \sigma_j)\right] = \sum_{t=1}^{T-1} W(t)(\sigma_{t+1} - \sigma_t) \tag{31}$$

$$E\left[(r_{i,t+1} - r_{i,t})^2 | (r_i, \sigma_j)\right] = \sum_{t=1}^{T-1} W(t)(r_{t+1} - r_t)^2, \text{and} \tag{32}$$

$$E\left[(r_{i,t+1} - r_{i,t})^2 | (r_i, \sigma_j)\right] = \sum_{t=1}^{T-1} W(t)(\sigma_{t+1} - \sigma_t)^2, \tag{33}$$

where $W(t)$ is the Nadaraya–Watson product weight function:

$$W(t) = \frac{K_{h_{i,j}}(r_i - r_t) K_{h_{i,j}}(\sigma_j - \sigma_t)}{\sum_{t=1}^{T} K_{h_{i,j}}(r_i - r_t) K_{h_{i,j}}(\sigma_j - \sigma_t)}, \tag{34}$$

and

$$K_{h_{i,j}}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{h_{i,j}}\right)^2} \tag{35}$$

is the Gaussian kernel, and $i, j = 1, 2, \ldots, N$. The smoothing parameters $h_{i,j}$, or "bandwidths," are the way one trades off bias against variance in the fit. Large bandwidths reduce local variation, but increase bias. Small bandwidths fit local phenomena, at the cost of increased variance.

Theoretic results for kernel regression estimators show that the optimal bandwidths will be proportional to $T^{-\frac{1}{6}}$. However, the constant of proportionality is a complicated function of the joint density and its derivatives, the function to be estimated and its derivatives, the bandwidths, and the properties of the kernel function. Since under the AL model the joint density function is not known, it is not possible to derive a closed–form expression for the optimal bandwidth. Instead, one must rely on numerical procedures. I conduct a search over a grid of bandwidth values in order to arrive at an optimal bandwidth for data generated by the AL model using the Euler approximations.[12] For the interest rate drift function, I search over scaling factors $\phi_r = 1, 2, 4, 6, 8, 10, 12$ for the bandwidth $\phi_r \hat{\sigma}_r T^{-\frac{1}{6}}$ that minimizes the sum of squared errors (SSE), computed as the sum over the estimation grid

of the squared deviations of the estimated surface from the true surface. For the interest rate diffusion, I search over a 7×7 grid of integer scaling factors ϕ_r and ϕ_σ to find the bandwidth vector $(\phi_r \hat{\sigma}_r, \phi_\sigma \hat{\sigma}_\sigma) T^{-\frac{1}{6}}$ that minimizes the SSE. For the volatility process functions, I search for scaling factors in the same way as for the interest rate functions. Table IV displays the optimal scaling factors and the associated SSEs.

The results in table IV show that, for the drift functions, the more highly autocorrelated interest rate data require relatively more smoothing. This is because a wider bandwidth leads to more cancellation of biases, and the biases tend to be more serious with more highly autocorrelated data, as will be discussed shortly. The large SSE on α_σ reflects a high degree of bias at extreme values of σ. If along the volatility dimension the solution grid were restricted to values in the range $(-4.9, -6.8)$, for example, the SSE on α_σ would be two orders of magnitude smaller.[13]

In the following discussion, I report pointwise averages for fits of the drift and diffusion functions over a 25×25 grid of equally–spaced values on the square defined by[14]:

$$\{(r, \sigma) : 0.02651 < r < 0.16731, -7.0 < \sigma < -4.6\}. \tag{36}$$

The pointwise averages are computed over 1000 simulations from the AL model. The "true" functions are parameterized using the values shown in table I in the previous section. The simulated data are drawn at a weekly frequency, with twenty–five inter–week draws.[15] Each trajectory is forty years in length. I run off fifty years of data before drawing simulated values, in view of the results from the previous section.[16]

Figures 1 and 2 display the fitted and true surfaces, as well as 95% pointwise confidence surfaces, for the fits obtained using the bandwidth scaling factors in table IV. In general, the fitted surfaces exhibit significant biases near the boundaries of the data, but the sampling variances are so high that the biases are likely to be irrelevant from the point of view of hypothesis testing. Only in a few small regions do the true surfaces "break through" the 95% confidence region. The quality of the fits is in general better for the volatility process, reflecting the higher degree of mean reversion for this process.

As discussed in Chapman and Pearson (1999), two effects induce bias in the estimated surfaces. Near the boundaries of the data, the kernel function is truncated, and since it is symmetric, this skews the weights toward the center

of the data. This can have predictable effects on the estimates. Taking the interest rate drift as an example, near the lower boundary of r, the weights will be biased toward higher values of r where the observed drifts tend to be less positive, or even negative. This biases the estimates near the lower boundary downward. The opposite is true for high values of r. Similar reasoning follows along the volatility dimension, because the volatility process is also mean-reverting.

The second form of bias results from the correlation of the residuals with the regressors near the edges of the data. The nonparametric regression model for the drifts is given by:

$$r_{t+\Delta} - r_t = \alpha_r + \epsilon_{r,t+\Delta} \tag{37}$$

$$\sigma_{t+\Delta} - \sigma_t = \alpha_\sigma + \epsilon_{\sigma,t+\Delta} \tag{38}$$

where the $\epsilon_{\cdot,t+\Delta}$ are disturbances. Unbiased estimation requires that:

$$E\left[\epsilon_{r,t+\Delta}|r_t, \sigma_t\right] = 0, \text{ and} \tag{39}$$

$$E\left[\epsilon_{\sigma,t+\Delta}|r_t, \sigma_t\right] = 0. \tag{40}$$

Bias arises because, in fact, the nonparametric estimator works with a finite data set for which (39) and (40) don't necessarily hold at the boundaries of the data. For example, at the data point where:

$$(r_t, \sigma_t) = (r_{\max}, \sigma), \tag{41}$$

it must be the case that:

$$r_{t+\Delta} - r_t \leq r_{\max} - r_t. \tag{42}$$

In other words, at the upper boundary of the observations on r, the residual in equation (37) must be negative, and *ceterus paribus* this causes downward bias in the point estimate of the drift function of the interest rate process. Moreover, to the extent that the residuals ϵ_r and ϵ_σ are correlated, bias will also be induced in the drift of the volatility process. This form of bias does not affect the diffusion estimates, because the sign of $(r_{t+\Delta} - r_t)^2$ is always positive.

Returning to figures 1 and 2, we see that, for high interest rates, the interest rate drift function estimate is biased upward, indicating that the effect of truncation bias is dominant. The opposite pattern holds for the estimates

of the volatility drift function. The estimates of the diffusion function of the interest rate exhibit complicated patterns of bias, as illustrated in the lower panel of figure 1. This is because the interest rate diffusion is a function of both state variables, and in addition, the interest rate data are highly persistent. The function is well estimated at the center of the data, but toward the corners of the surface, significant biases are in evidence. Looking at the lower panel of figure 2, we find that the surface is estimated with much less bias.

It is useful to compute numerical measures of error, both for diagnostic purposes and as a prelude to the test statistics used below. I compute three error measures, based on the L_1, L_2 and L_∞ norms. To "estimate" the L_1 norm, I use the simple formula

$$\hat{L}_1 = \sum_i \sum_j |\hat{f}_{i,j} - f_{i,j}|, \tag{43}$$

where i and j run over the solution grid, $\hat{f}_{i,j}$ denotes the estimated function value at (r_i, σ_j), and $f_{i,j}$ denotes the true value.[17] The L_2 norm is similar, except that we "integrate" the squared errors over the solution surface:

$$\hat{L}_2 = \sum_i \sum_j (\hat{f}_{i,j} - f_{i,j})^2. \tag{44}$$

Finally, inspired by the Kolmogorov–Smirnov test of the previous section, I compute an estimate of the L_∞ norm:

$$\hat{L}_\infty = \max_{i,j} |\hat{f}_{i,j} - f_{i,j}|. \tag{45}$$

Table V displays these error measures for the surfaces shown in figures 1 and 2.

Examining the results in table V, we see that the measures $\hat{L}_1$ and $\hat{L}_2$ are driven by extreme errors. This can be deduced from the fact that the $\hat{L}_\infty$ measure tends to be large relative to the $\hat{L}_1$ measure. The error measures for β_r and β_σ underscore the success of the kernel method for estimating diffusion functions. In both cases, the $\hat{L}_1$ and $\hat{L}_2$ measures are at least an order of magnitude smaller than the corresponding measures on the drift functions. The relatively large values of the error measures on α_σ highlights the influence of the choice of the solution set on the estimator, noted earlier.

The inefficiency of the estimator can be measured by integrating the region between the upper and lower 95% confidence surfaces. The measure

that I compute is given by:

$$EFF = \sum_i \sum_j (\hat{f}^{(+)}_{i,j} - f^{(-)}_{i,j}), \tag{46}$$

where $\hat{f}^{(+)}$ denotes a point on the upper surface, and $\hat{f}^{(-)}$ on the lower surface. Thus, a larger value for EFF indicates greater inefficiency, the confidence surfaces being farther apart. Table VI displays the calculations for the surfaces in figures 1 and 2. The inefficiency measures in table VI are primarily useful for comparisons between estimators. I defer a discussion of these results until the next section, where I consider the performance of the BRSW estimator when the model is misspecified.

III Misspecification

The estimates in the previous section were computed for the unrealistic case where we assumed *a priori* knowledge of the arguments to the drift and diffusion functions, and could thus use the correct conditioning variables in the kernel regressions. In other words, we estimated the following system:

$$dr_t = \alpha_r(r_t)dt + \beta_r(r_t, \sigma_t)dW_{r,t} \tag{47}$$

$$d\sigma_t = \alpha_\sigma(\sigma_t)dt + \beta_\sigma dW_{\sigma,t}, \tag{48}$$

in which all the arguments coincide with the arguments of the corresponding functions in the AL model.

Suppose we were to estimate the more general system in (22)-(23). In this case, the drift functions and the diffusion function of the volatility process are misspecified. The drift function for the interest rate process depends only on the level of the interest rate, as shown in (47), but under the more general model we will condition on the levels of the interest rate and volatility. Similarly, for the volatility drift, we'll condition on both state variables when in fact the drift only depends on the level of volatility, as shown in (48). The volatility diffusion will be highly misspecified. For this function, we condition on both state variables when in fact the diffusion is constant. It is interesting to look at how these forms of misspecification affect the estimator.

Figures 3 and 4 display the various estimated surfaces. Introducing irrelevant conditioning variables introduces additional biases in the estimates due to the correlations in the residuals at the data boundaries, as discussed

above. Starting with the top panel of figure 3, the surface has a distinct slope along the volatility dimension for high values of r. For low values of r, the surface also has a non–zero slope along the σ axis, although it is less pronounced.

Comparing the top panel of figure 4 to the top panel of figure 2, we see that for the volatility drift, the irrelevant conditioning information leads mainly to a loss of efficiency. There is only slight evidence of increased bias. The results for the volatility diffusion function are similar.

Table VII shows the error measures for the correct and misspecified fits. In general, the errors increase, although there are some important exceptions. For α_r, both the $\hat{L}_1$ and $\hat{L}_2$ measures *improve* under the misspecified model, showing that the introduction of the irrelevant conditioning variable facilitated additional bias cancellations. The diffusion function β_r is correctly specified under both models and thus the error measures don't change. For the volatility process, the irrelevant conditioning information significantly worsens the fit for both the drift and diffusion functions. In sum, the results here and above show that irrelevant conditioning information has an ambiguous effect on the magnitude and sign of bias.

As we would expect, the inclusion of irrelevant conditioning variables results in greater inefficiency. Table VIII displays the inefficiency measure given by equation (46) for the misspecified model. Comparing these values to the values in table VI, we see that the value of EFF is in general greater under the misspecified model. The efficiency loss is greatest for the volatility diffusion, where we have introduced two irrelevant variables. The value of EFF jumps from 54.4 to 90.0. The value of EFF for β_r doesn't change because in both cases we've estimated the function with both conditioning variables.

The main points to take away from the results of this section and the previous section are that the kernel regression estimator has significant finite sample biases, but that the variance of the estimator is high enough that there is reason to doubt that the biases are relevant for hypothesis testing. In a real–data situation, of course, one can't know the sampling variance, or the degree of bias in the estimator. In light of these facts, a question that plays to the strengths of the kernel estimator is to ask if the estimator produces estimates of the drift and diffusion functions that are statistically distinguishable from a *known parametric estimator*. In other words, does the more general kernel estimator "pick up" anything in the data that the parametric estimator might be missing? In this context, Monte Carlo methods can be used to bootstrap the finite–sample distributions of statistics based

on the nonparametric estimator.

IV Hypothesis Tests

In this section, I use Treasury bill data to test the hypothesis that the BRSW estimator produces estimated surfaces that are statistically indistinguishable from the surfaces implied by the estimates in table I. The test proceeds in two stages. First, the quantiles of three different test statistics for the BRSW estimator are bootstrapped under the null hypothesis that the AL model is the true data generating process. Second, the BRSW estimator is applied to the Treasury data, and the values of the test statistics are computed. Finally, the values of the test statistics computed for the Treasury data are compared to the bootstrapped quantiles.

The Treasury data used to proxy the riskless short rate are the same data that are used by Andersen and Lund (1997a). I use the three-month Treasury–bill yield, at weekly (Wednesday) frequency from 1962-1999.[18] The data are obtained from the H.15 release of the Federal Reserve System. I convert the series from a bank discount basis to an investment basis prior to analysis, and Tuesday values are substituted for Wednesday values when the Wednesday value is missing.

I also make use of data on the slope of the term structure. The data used to form the slope of the term structure are the same data used in Boudoukh et al. (1998). I use the yields on Treasury securities at constant, ten–year maturities, again from the H.15 release. The slope of the term structure is computed as the difference between the ten–year rate and the three–month rate.

The slope of the term structure is used in the estimation procedure because the volatility process is not directly observable. Estimates of the volatilities are obtained by first fitting the level and slope data using the BRSW estimator, and using the estimates of the interest rate diffusion function from this first stage to compute the implied volatilities. The three–month rates and implied volatilities are used in the estimation of the functions of the "true" processes.[19] In the first stage, I estimate $\beta_r(r_t, S_t)$ from the following system:

$$dr_t = \alpha_r(r_t, S_t)dt + \beta_r(r_t, S_t)dW_{r,t} \quad (49)$$

$$dS_t = \alpha_S(r_t, S_t)dt + \beta_S(r_t, S_t)dW_{S,t}, \quad (50)$$

where S_t is the slope of the term structure at time t. The estimate $\hat{\beta}_r(r, S)$ is then used to infer the volatility process observations. An observation s_t is obtained by plugging in (interpolating where necessary) the observed values (r_t, S_t) to obtain $\hat{\beta}_r(r_t, S_t) = s_t$.[20] Finally, to make the volatility process consistent with the AL model, I make the transformation $\sigma_t = \ln(s_t^2/r_t)$.

The series of implied σ values has the same unconditional mean as the volatility process estimated by Andersen and Lund (1997a). The estimated unconditional mean for the AL model is reported in table I as -6.3599. The unconditional mean of the volatility values inferred using the BRSW estimator and observations on the level and slope of the term structure is -6.3557. It is reassuring that two estimators agree on this parameter.

Using the Monte Carlo methods of the previous section, I bootstrap the distribution of three different statistics, under the null hypothesis that the AL model is the "true" data generating process. The test statistics are the mean squared error (MSE), mean absolute error (MAE) and maximum absolute deviation (MAD), defined as follows:

$$MSE = \frac{\delta(r,\sigma)}{N^2}\hat{L}_2 \tag{51}$$

$$MAE = \frac{\delta(r,\sigma)}{N^2}\hat{L}_1 \tag{52}$$

$$MAD = \delta(r,\sigma)\hat{L}_\infty, \tag{53}$$

where $\hat{L}_1$, $\hat{L}_2$ and $\hat{L}_\infty$ are defined in equations (43)-(45) in the previous section, and $\delta(r, \sigma)$ is a "trimming function" used to reduce the effect of boundary biases on the statistics. I used $\delta(\cdot)$ to trim the solution grid to a 21×21 square, thus removing the outer two rings of data. The quantiles of the statistics are found by compiling the values of the statistics for 1,000 simulated draws from the AL model using the Euler method of the previous section. Table IX displays the 90% and 95% quantiles for the three statistics.

If the null hypothesis is true, when we apply the BRSW estimator to the Treasury data and compute the statistics on the resulting estimated surfaces, we should obtain values for the statistics that fall into the middle of the bootstrapped distributions. If the null hypothesis is false, the statistics will fall into the upper tails of the distributions, and we can conclude that the kernel estimator is "picking up" something in the data that is missed by the parametric estimator.

The distributions of the test statistics are computed under the misspecified model; this allows the for the best chance of picking up something in

the data that the parametric model might miss. For each function and each statistic, I search over an 18×18 grid for the pair of integer scaling values (ϕ_r, ϕ_σ) that produce bandwidths $(\phi_r \hat{\sigma}_r, \phi_\sigma \hat{\sigma}_\sigma) T^{-\frac{1}{6}}$ that minimize the statistic in question. This approach finds the bandwidth values that minimize the statistics for the model that maximizes the likelihood of finding significant differences between the nonparametric and parametric estimates. The statistic–minimizing bandwidths are shown in table X. Figures 5 and 6 display the fitted surfaces for the bandwidth values that minimize the MSE criterion, as well as the surfaces under the AL model.

The observed statistics are displayed in table XI. Except for the interest rate drift function, the null hypothesis is rejected at the 95% level for each function and statistic. For the interest rate drift function (α_r), the 90^{th} quantiles are 0.000048, 0.0057 and 0.013 for the MSE, MAE, and MAD statistics, respectively. From table XI, we see that the observed statistics are 0.000033, 0.0049, and 0.0089, respectively – all less than the associated quantile values and thus within the 90% acceptance region. For the interest rate diffusion (β_r), we see that the observed statistic values are greater than the 95% quantiles for each statistic, indicating rejection of the hypothesis that the Treasury data are drawn from the distribution implied by the AL model. Similarly, the observed statistics for the volatility process functions (α_σ and β_σ) indicate rejection of the null hypothesis.

In sum, the results support the conclusion that the Treasury data are not generated by the AL model. However, the results do not support the conclusion of nonlinearities in the interest rate drift function. The results indicate that the interest rate diffusion and the volatility process drift and diffusion functions exhibit nonlinearities that are not captured by the AL model.

It is important to emphasize that these hypothesis test results are robust against any residual kernel biases that may be present in the estimates, because we have bootstrapped the finite sample distributions of the statistics under the null hypothesis that the parametric model is the true data generating process. The quantiles that are reported in table IX are thus "corrected" for kernel bias by the bootstrap.

V Conclusion

In this essay, I used Monte Carlo simulations from the Andersen and Lund (1997a) stochastic volatility model of interest rates to study the finite sample properties of the BRSW estimator. The estimator exhibited complicated patterns of bias and a high sampling variance. The introduction of irrelevant conditioning information resulted in increased inefficiency in all cases, and increased bias in most cases. I tested whether the BRSW estimates were statistically distinguishable from the parametric estimates, and found that the BRSW estimator indeed appeared to be picking up dynamics in the data that the parametric estimator missed.

As part of this research, I worked out a method to test whether or not a system of stochastic differential equations is stationary. The algorithm that I used for performing the test involved the first–order Euler discretization scheme for simulating trajectories from the model, and an extension of the Kolmogorov–Smirnov test. As mentioned earlier, it would be useful to extend the bivariate Kolmogorov–Smirnov test to the case of k–samples. It is possible that the k–sample generalization can be derived much the same way that the univariate k–sample KS test is derived from its two sample analogue. While the full k–sample bivariate statistic would be computationally burdensome to calculate, the wide range of applications for which it would be useful would seem to justify its development.

In the econometrics literature, and in the research pipeline, there are many different estimators for the drift and diffusion functions of continuous time stochastic processes. For example, one can turn to the efficient method of moments estimator of Gallant and Tauchen (1996) or the simulated likelihood method of Brandt and Santa–Clara (1999). It would be useful to compare the finite sample properties of these estimators against a common benchmark, such as the maximum likelihood estimator for a model in which the transition densities are known in closed form. To date, little work has been done to understand the relative performance of the different estimators.

Appendix

Kernel regression, particularly in multiple dimensions, is necessarily a computationally intensive procedure. However, a parallel computer can make short work of even fairly large problems, because kernel regression lends itself easily to parallelization. In this appendix, I discuss a very simple algorithm that I've developed for doing kernel regression on a parallel computer.

In two dimensions, kernel regression using the Nadaraya-Watson estimator essentially boils down to computing the following formula repeatedly over a grid of solution points:

$$\hat{f}(x_i, y_j) = \sum_{t=1}^{T} W(t) g(x_t, y_t; x_i, y_j), \tag{54}$$

where $W(t)$ is the weighting function from equation (34) in the body of the paper, and $g(\cdot)$ is a known function of the data and the solution point. We compute this equation for $\{x_i, y_j\}_{i,j=1}^{N}$.

A naive parallel algorithm for this problem is to simply break up the solution grid into chunks, and to assign the chunks to the available processors. This algorithm is in general inefficient unless one also works out an algorithm for balancing the load across the processors, which is a difficult problem, particularly on a shared machine. A more efficient approach is to rely on the operating system for load balancing, and to assign small bits of the task (single grid points) to lightweight processes for execution. The bit of pseudo–code below shows how I implemented such an algorithm using the pthreads library on a Sun workstation running the Sun Solaris 2.6 operating system.

The outer while loop checks the completion condition, where the size of the problem is given by the parameter $n = N$. The if–statement inside the while loop ensures that a limited number of threads are running at one time, where the maximum number of threads is given by nt. This mechanism prevents the program from loading the machine with so many lightweight processes that they begin to compete with one another for resources, degrading performance. When the limit nt is reached, the algorithm waits for threads to join (terminate), and then fires off more threads as needed. The routine `Kernel_Thread` is the routine in which the actual computations are done.

```
i = 0;
count = 0;
while ( i < n ) {
  if ( count < nt ) {
    if ( pthread_create((pthread_t *) &thread_id,
                        (pthread_attr_t *) &thread_attributes,
                        Kernel_Thread,
                        (void *) (thread_data + i)) ) {
      perror("pthread_create");
      return;
    }
    count++;
    i++;
  } else {
    thr_join((thread_t)   0,
             (thread_t *) &thread_id,
             (void **)    NULL);
    count--;
  }
}
```

The algorithm is efficient, driving a Sun Ultrasparc with three processors to around 80% of maximum efficiency in terms of cpu utilization. Over a solution grid with 144 points, using 2,080 data points, the algorithm computed 4,000 iterations of the BRSW estimator for the AL model in approximately eleven minutes. When the number of data points was increased to 208,000, the program drove the machine to nearly maximum efficiency, and ran in one hour, forty minutes.

References

Ait–Sahalia, Y.: 1996, Testing continuous–time models of the spot interest rate, *The Review of Financial Studies* **9**, 385–426.

Andersen, T. G. and Lund, J.: 1997a, Estimating continuous-time stochastic volatility models of the short term interest rate, *Journal of Econometrics* **77**(2), 343–377.

Andersen, T. G. and Lund, J.: 1997b, Stochastic volatility and mean drift in the short rate diffusion: Sources of steepness, level and curvature in the yield curve. Working paper.

Boudoukh, J., Richardson, M., Stanton, R. and Whitelaw, R. F.: 1998, The stochastic behavior of interest rates: Implications from a multifactor, nonlinear continuous-time model. Working paper.

Brandt, M. W. and Santa–Clara, P.: 1999, Simulated likelihood estimation of multivariate diffusions with an application to interest rates and exchange rates with stochastic volatility. Working paper.

Chapman, D. A. and Pearson, N. D.: 1999, Is the short rate drift actually nonlinear? Forthcoming in *Journal of Finance*.

Duffie, D. and Kan, R.: 1996, A yield–factor model of interest rates, *Mathematical Finance* **6**, 379–406.

Fasano, G. and Franceschini, A.: 1987, A multidimensional version of the kolmogorov–smirnov test, *Monthly Notices of the Royal Astronomical Society* **225**, 155–170.

Gallant, A. and Tauchen, G.: 1996, Which moments to match?, *Econometric Theory* **12**, 657–681.

Härdle, W.: 1990, *Applied Nonparametric Regression*, Cambridge University Press.

Karatzas, I. and Shreve, S. E.: 1991, *Brownian Motion and Stochastic Calculus*, Springer–Verlag, New York, NY.

Kloeden, P. E. and Platen, E.: 1995, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P.: 1994, *Numerical Recipes in C*, second edn, Cambridge University Press, Cambridge.

Pritsker, M.: 1998, Nonparametric density estimation and tests of continuous time interest rate models, *Review of Financial Studies* **11**(3), 449–487.

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

Footnotes

1. In what follows, I refer to the the Boudoukh et al. (1998) estimator for multifactor models as the "BRSW estimator."

2. The Ait–Sahalia (1996) estimator is difficult to adapt to multivariate models, so I do not consider it here.

3. One can verify that (6)-(7) are equivalent to (1)-(2) using Ito's Lemma and the transformation $\hat{\sigma}_t = \log \sigma_t^2$. In equations (6)-(7), I have omitted the '^' symbol on σ_t for notational brevity.

4. It is important to keep in mind our maintained hypothesis that the system has a unique solution. We might conclude that the system is stationary, but if our maintained hypothesis is in error, the transition densities could be converging to the stationary density of a *different* system! This is similar to the problems that can arise when solving a partial differential equation with a finite difference algorithm that is inconsistent. However, as we'll see below, the transition densities appear to converge, and there is no evidence of convergence to the "wrong" density.

5. In private communications, the authors indicated that the parameters reported in Andersen and Lund (1997a) reflect rescalings of the diffusion function. The parameter values in table I are from Andersen and Lund (1997b), in which the authors correct the values for the rescaling. In tests similar to those reported here, I found that the system was borderline stationary, perhaps even nonstationary, at the values actually published in Andersen and Lund (1997a).

6. The standard deviations are reported at zero due to rounding. In reality they are on the order of 10^{-14}. The tight standard deviations reflect the use of the antithetic variance reduction technique.

7. Unlike the standard one dimensional KS test statistic, the bivariate statistic is slightly distribution–dependent. In future work, I plan to study the test statistic a little more closely. For more information on the test statistic, see the paper cited in the text and Press, Teukolsky, Vetterling and Flannery (1994).

8. Picking points farther out in the tails of the distribution will bias the test toward finding convergence at longer trajectories. On the other hand, from the results in table II, we can make some assessment of the probability of observing the points that are chosen for the test. One should pick points far enough out in the tails so that the probability of observing points that could generate different results is very low, but not so far out that the test becomes computationally infeasible.

9. It would be useful to have a k–sample bivariate Kolmogorov–Smirnov test, with which one could simultaneously test the convergence of bivariate transition densities defined by a surface of k starting points. To my knowledge, no such test has been developed.

10. It's unclear how the efficient method of moments estimator used in Andersen and Lund (1997a), or other simulation estimators, are affected when the first draws of simulated trajectories are not drawn from the stationary density of the process. To my knowledge, a formal study of the issue has not been completed. In related work, Brandt and Santa–Clara (1999) report that *fixing* the first observation has little effect on the simulated maximum likelihood estimator that they develop, but the extent to which this finding generalizes to other estimators is unknown. Of course, the effects must be limited in a large sample, simply because the effect of any single observation on the likelihood function will be limited. In the main, it is a small sample issue.

11. When no confusion will arise, in what follows I omit the arguments to the drift and diffusion functions. They are to be understood.

12. The cross-validation approach to bandwidth selection is not useful for highly autocorrelated data. See Härdle (1990) for a short discussion, and Pritsker (1998) for a more in–depth discussion of the problems.

13. The solution grid was chosen to be consistent with the hypothesis tests in section four.

14. The set of valid (r, σ) values is defined by the observed Treasury bill data in section four. The set contains all of the observed data points.

15. The inter-week draws ensure that, during the simulations, the discretized process for the interest rate never takes on negative values.

In addition, with the inter–week draws, the data are simulated at a degree of accuracy that is greater than the accuracy of the nonparametric estimator. Thus, the accuracy of the weak solution does not bound the accuracy of the estimator.

16. A parallel kernel estimator is used in order to manage the computational load. The parallel kernel estimator is discussed in the appendix.

17. To be precise, one ought to compute the L_1 norm using a quadrature integration method or the like, especially if the function surfaces exhibit radical gradients. Because our surfaces are very well–behaved, the simple formulas used here suffice for our purposes.

18. Andersen and Lund (1997a) use data for the period from 1954-1995; in all other respects the series are the same.

19. See Duffie and Kan (1996) for a discussion.

20. To estimate the interest rate diffusion, I used bandwidths $(\hat{\sigma}_r, \hat{\sigma}_\sigma)T^{-\frac{1}{6}}$ $= (7.384985e-03, 3.579158e-03)$, where the $\hat{\sigma}$ symbols denote sample standard deviations.

Table I: Parameter Values

This table lists the parameter values used in the Monte Carlo simulations throughout the paper. The parameters are taken from Andersen and Lund (1997b). The stochastic system is given by:

$$\begin{aligned} dr_t &= \kappa_1(\mu - r_t)dt + \sigma_t\sqrt{r_t}dW_{1,t} \\ d\log\sigma_t^2 &= \kappa_2(\theta - \log\sigma_t^2)dt + \xi dW_{2,t} \end{aligned}$$

Parameter	Value
κ_1	0.1633
μ	0.0595
κ_2	1.0397
θ	-6.3599
ξ	1.2719

Table II: Simulation Results

This table reports the results of Monte Carlo simulations to generate moments of the transition densities of the AL model. From each of 25 different starting points, 1,000 batches of 100 trajectories are simulated. The last point of each trajectory is saved, forming a batch of 100 draws from the transition density defined by the starting point and the length of the trajectory. The mean and variance of each batch of saved points is then computed. At the end of a run, the procedure produces 1,000 independent draws of the first two moments of each of the 25 transition densities. Eight such runs are completed, the first with trajectories one year in length, the second with five year trajectories, and so on for ten, twenty, thirty, forty, fifty, and finally sixty year trajectories. In the table, the 'Mean' columns show the average over the 25 densities of the moment in question, and the 'Std Dev' columns show the dispersion of this moment over the 25 densities. The 'Min' and 'Max' columns show the minimums and maximums of each moment over the 25 densities, respectively.

Moment	Mean	Std Dev	Min	Max	Moment	Mean	Std Dev	Min	Max
$E[r_1]$	0.0769	0.0420	0.0163	0.1401	$\text{Var}[r_1]$	0.0002	0.0001	0.0000	0.0011
$E[r_5]$	0.0685	0.0218	0.0353	0.1050	$\text{Var}[r_5]$	0.0004	0.0002	0.0000	0.0020
$E[r_{10}]$	0.0634	0.0097	0.0462	0.0830	$\text{Var}[r_{10}]$	0.0005	0.0001	0.0001	0.0018
$E[r_{20}]$	0.0602	0.0022	0.0529	0.0682	$\text{Var}[r_{20}]$	0.0004	0.0001	0.0001	0.0019
$E[r_{30}]$	0.0596	0.0013	0.0547	0.0658	$\text{Var}[r_{30}]$	0.0004	0.0001	0.0002	0.0016
$E[r_{40}]$	0.0595	0.0012	0.0547	0.0647	$\text{Var}[r_{40}]$	0.0004	0.0001	0.0001	0.0016
$E[r_{50}]$	0.0594	0.0012	0.0549	0.0651	$\text{Var}[r_{50}]$	0.0004	0.0001	0.0001	0.0015
$E[r_{60}]$	0.0594	0.0012	0.0547	0.0651	$\text{Var}[r_{60}]$	0.0004	0.0001	0.0001	0.0016
$E[\sigma_1]$	-6.2339	0.2473	-6.5838	-5.8841	$\text{Var}[\sigma_1]$	0.6971	0.1399	0.2701	1.4404
$E[\sigma_5]$	-6.3580	0.0037	-6.3632	-6.3527	$\text{Var}[\sigma_5]$	0.7932	0.1590	0.2567	1.6098
$E[\sigma_{10}]$	-6.3598	0.0000	-6.3599	-6.3598	$\text{Var}[\sigma_{10}]$	0.7936	0.1596	0.3470	1.5934
$E[\sigma_{20}]$	-6.3599	0	-6.3599	-6.3599	$\text{Var}[\sigma_{20}]$	0.7945	0.1591	0.2875	1.5405
$E[\sigma_{30}]$	-6.3599	0	-6.3599	-6.3599	$\text{Var}[\sigma_{30}]$	0.7950	0.1589	0.2838	1.6439
$E[\sigma_{40}]$	-6.3599	0	-6.3599	-6.3599	$\text{Var}[\sigma_{40}]$	0.7942	0.1589	0.2723	1.6174
$E[\sigma_{50}]$	-6.3599	0	-6.3599	-6.3599	$\text{Var}[\sigma_{50}]$	0.7950	0.1597	0.3147	1.5621
$E[\sigma_{60}]$	-6.3599	0	-6.3599	-6.3599	$\text{Var}[\sigma_{60}]$	0.7934	0.1593	0.3054	1.5795

Table III: Bivariate KS Test Results

This table displays the results of the bivariate Kolmogorov-Smirnov test for convergence in distribution of the transition densities of the AL system. Two transition densities are tested for convergence. The densities are defined by starting points that are two standard deviations away from the long–run means of each process, and about four standard deviations away from one another, and by the length of the trajectories. The first column displays the length of the trajectories, the second column shows the test statistic, and the final column shows the p–Value.

Years	KS	p–Value
1	0.9991	0.0000
5	0.8735	0.0000
10	0.4928	0.0000
20	0.1061	0.0000
30	0.0240	0.0012
40	0.0100	0.5327

Table IV: Scaling Factors and Sum of Squared Errors

This table reports the results of a grid search for the optimal scaling factors on the bandwidths of the kernel estimator for each function of the system. The first column lists the function, and the second and third columns display the relevant scaling factors that minimized the sum–of–squared error criterion. The final column displays the resulting SSE value.

Function	ϕ_r	ϕ_σ	SSE
α_r	6.0	–	0.00273
β_r	2.0	1.0	0.00277
α_σ	–	1.0	0.14943
β_σ	–	2.0	0.00829

Table V: Error Measures

This table reports measures of error in fit for the kernel estimates displayed in figures 1 and 2. The error measures are defined as:

$$\begin{aligned}\hat{L}_1 &= \sum_i \sum_j |\hat{f}_{i,j} - f_{i,j}|, \\ \hat{L}_2 &= \sum_i \sum_j (\hat{f}_{i,j} - f_{i,j})^2, \\ \hat{L}_\infty &= \max_{i,j} |\hat{f}_{i,j} - f_{i,j}|,\end{aligned}$$

where $\hat{f}_{i,j}$ denotes the kernel estimate at point i, j on the solution grid, and f denotes the true value.

Function	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$
α_r	1.744913e-02	2.799722e+00	9.110867e-03
β_r	4.321830e-03	9.906933e-01	1.201149e-02
α_σ	4.221677e-01	1.005452e+01	1.003910e-01
β_σ	9.791103e-03	2.473750e+00	3.958000e-03

Table VI: Inefficiency Measure

This table reports the value of an inefficiency measure for the estimates displayed in figures 1 and 2. The inefficiency measure is defined as:

$$EFF = \sum_i \sum_j (\hat{f}_{i,j}^{(+)} - \hat{f}_{i,j}^{(-)}),$$

where $\hat{f}_{i,j}^{(+)}$ denotes the upper 95% confidence value at point (r_i, σ_j) on the solution surface, and $\hat{f}_{i,j}^{(-)}$ denotes the lower 95% value.

Function	EFF
α_r	6.56
β_r	9.06
α_σ	1068.59
β_σ	54.44

Table VII: Error Measures under Misspecification

This table reports measures of error in fit for the kernel estimates displayed in figures 3 and 4. The error measures are defined as:

$$\begin{aligned}\hat{L}_1 &= \sum_i \sum_j |\hat{f}_{i,j} - f_{i,j}|, \\ \hat{L}_2 &= \sum_i \sum_j (\hat{f}_{i,j} - f_{i,j})^2, \\ \hat{L}_\infty &= \max_{i,j} |\hat{f}_{i,j} - f_{i,j}|,\end{aligned}$$

where $\hat{f}_{i,j}$ denotes the kernel estimate at point i, j on the solution grid, and f denotes the true value.

Function	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$
α_r	1.718060e-02	2.773228e+00	9.949927e-03
β_r	4.321830e-03	9.906933e-01	1.201149e-02
α_σ	5.777287e-01	1.257005e+01	1.707380e-01
β_σ	3.256718e-02	4.120876e+00	1.014200e-02

Table VIII: Inefficiency Measure under Misspecification

This table reports the value of an inefficiency measure for the estimates displayed in figures 3 and 4. The inefficiency measure is defined as:

$$EFF = \sum_i \sum_j (\hat{f}_{i,j}^{(+)} - \hat{f}_{i,j}^{(-)}),$$

where $\hat{f}_{i,j}^{(+)}$ denotes the upper 95% confidence value at point (r_i, σ_j) on the solution surface, and $\hat{f}_{i,j}^{(-)}$ denotes the lower 95% value.

Function	EFF
α_r	6.67
β_r	9.06
α_σ	1201.86
β_σ	90.09

Table IX: Bootstrapped Quantiles

This table reports the bootstrapped quantiles of three different statistics, computed under the null hypothesis that the AL model is the "true" data generating process. The test statistics are the mean absolute error (MAE), mean squared error (MSE), and maximum absolute deviation (MAD), defined as follows:

$$MAE = \frac{\delta(r,\sigma)}{N^2}\hat{L}_1 \quad (55)$$

$$MSE = \frac{\delta(r,\sigma)}{N^2}\hat{L}_2 \quad (56)$$

$$MAD = \delta(r,\sigma)\hat{L}_\infty, \quad (57)$$

where $\hat{L}_1$, $\hat{L}_2$ and $\hat{L}_\infty$ are defined as:

$$\begin{aligned}\hat{L}_1 &= \sum_i \sum_j |\hat{f}_{i,j} - f_{i,j}|,\\ \hat{L}_2 &= \sum_i \sum_j (\hat{f}_{i,j} - f_{i,j})^2,\\ \hat{L}_\infty &= \max_{i,j} |\hat{f}_{i,j} - f_{i,j}|,\end{aligned}$$

where $\hat{f}_{i,j}$ denotes the kernel estimate at point i, j on the solution grid, and f denotes the true value. The function $\delta(r,\sigma)$ is a "trimming function" used to reduce the effect of boundary biases on the statistics. I used $\delta(\cdot)$ to trim the solution grid to a 21×21 square, thus removing the outer two rings of data. The quantiles of the statistics are found by compiling the values of the statistics for 1,000 simulated draws from the AL model using the Euler method.

	95^{th} Quantiles				90^{th} Quantiles		
Function	MSE	MAE	MAD	Function	MSE	MAE	MAD
α_r	0.000054	0.006009	0.013931	α_r	0.000048	0.005736	0.013033
β_r	0.000034	0.003841	0.020105	β_r	0.000027	0.003423	0.018463
α_σ	1.305474	0.406894	1.453917	α_σ	1.128532	0.355319	1.193143
β_σ	0.001622	0.040272	0.042565	β_σ	0.001057	0.032504	0.035641

Table X: Bandwidth Scalings for Observed Statistic Values

This table reports the bandwidth scalings for the kernel estimates based on Treasury data. For each function and each statistic, I search over an 18×18 grid for the pair of integer scaling values (ϕ_r, ϕ_σ) that produce bandwidths $(\phi_r \hat{\sigma}_r, \phi_\sigma \hat{\sigma}_\sigma) T^{-\frac{1}{6}}$ that minimize the statistic in question. This approach finds the bandwidth values that minimize the statistic that maximizes the probability of finding significant differences between the BRSW and EMM estimators.

	MSE		MAE		MAD	
Function	ϕ_r	ϕ_σ	ϕ_r	ϕ_σ	ϕ_r	ϕ_σ
α_r	4.0	12.0	4.0	12.0	4.0	12.0
β_r	1.0	1.0	2.0	1.0	1.0	12.0
α_σ	1.0	10.0	1.0	12.0	12.0	1.0
β_σ	6.0	8.0	6.0	6.0	12.0	12.0

Table XI: Observed Statistic Values

This table reports statistic values computed on Treasury data. The statistics are defined as:

$$MAE = \frac{\delta(r,\sigma)}{N^2}\hat{L}_1 \quad (58)$$

$$MSE = \frac{\delta(r,\sigma)}{N^2}\hat{L}_2 \quad (59)$$

$$MAD = \delta(r,\sigma)\hat{L}_\infty, \quad (60)$$

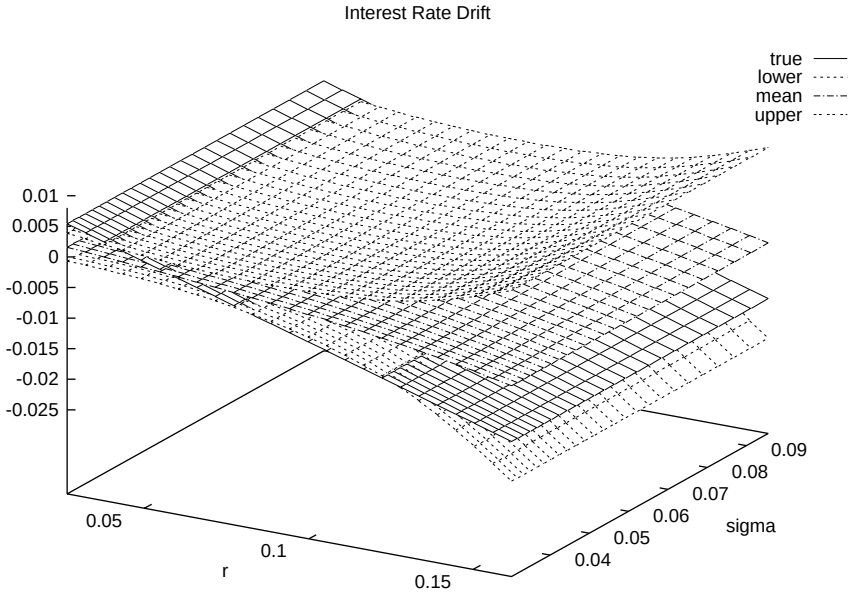
where $\hat{L}_1$, $\hat{L}_2$ and $\hat{L}_\infty$ are defined as:

$$\begin{aligned}\hat{L}_1 &= \sum_i \sum_j |\hat{f}_{i,j} - f_{i,j}|,\\ \hat{L}_2 &= \sum_i \sum_j (\hat{f}_{i,j} - f_{i,j})^2,\\ \hat{L}_\infty &= \max_{i,j} |\hat{f}_{i,j} - f_{i,j}|,\end{aligned}$$

where $\hat{f}_{i,j}$ denotes the kernel estimate at point i, j on the solution grid, and f denotes the value implied by the AL model. The function $\delta(r,\sigma)$ is a "trimming function" used to reduce the effect of boundary biases on the statistics. I used $\delta(\cdot)$ to trim the solution grid to a 21×21 square, thus removing the outer two rings of data. The bootstrapped quantiles of the statistics are displayed in table IX.

Function	MSE	MAE	MAD
α_r	0.000033	0.004941	0.008971
β_r	0.004405	0.041261	0.204983
α_σ	10.724300	2.713032	5.731216
β_σ	0.031650	0.150124	0.318428

Figure 1: Estimates for Interest Rate Process



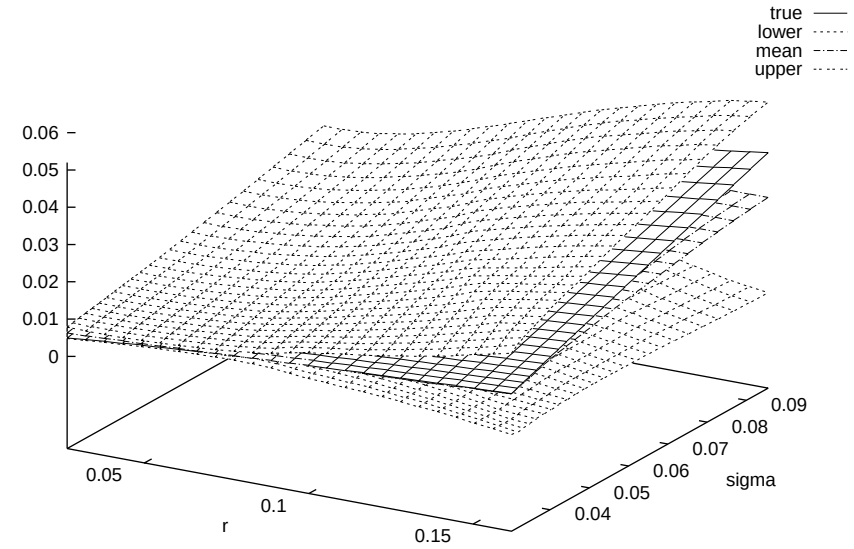


Figure 2: Estimates for Volatility Process

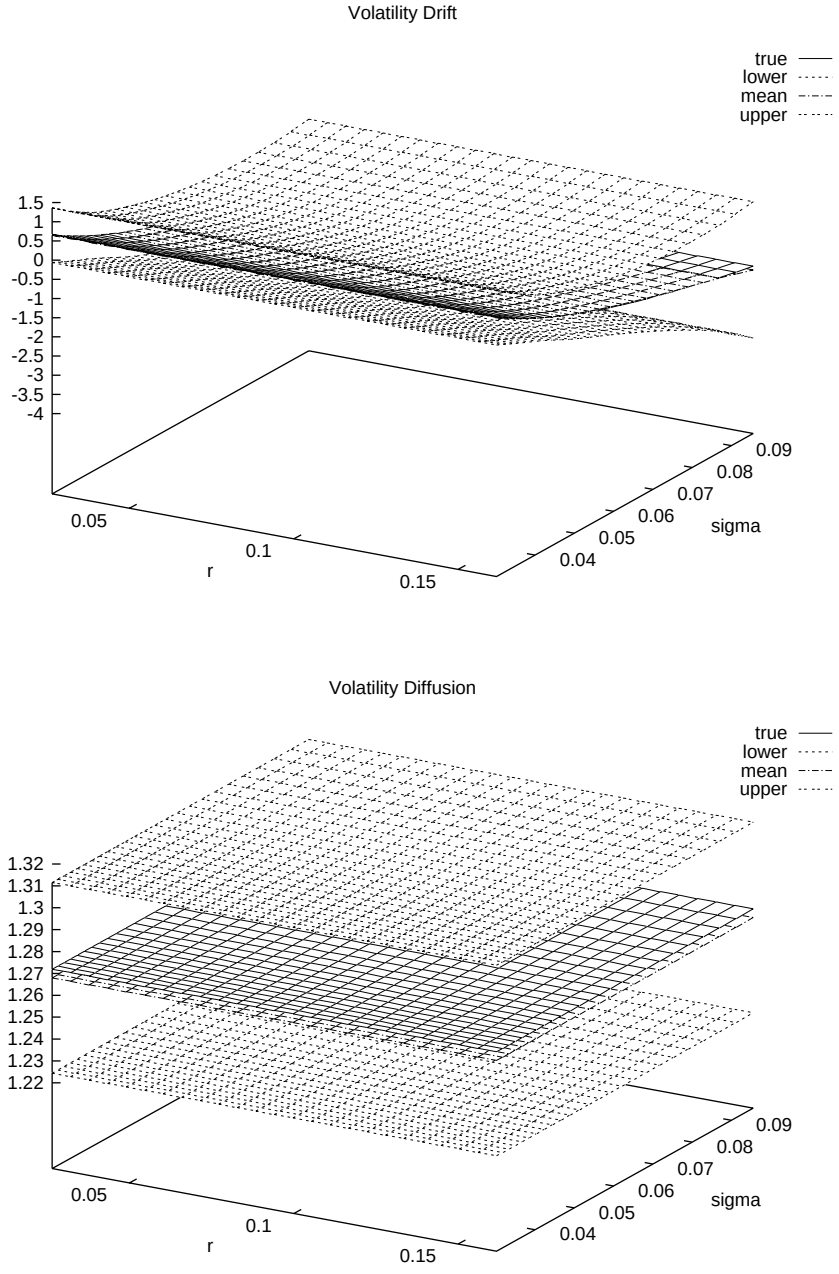
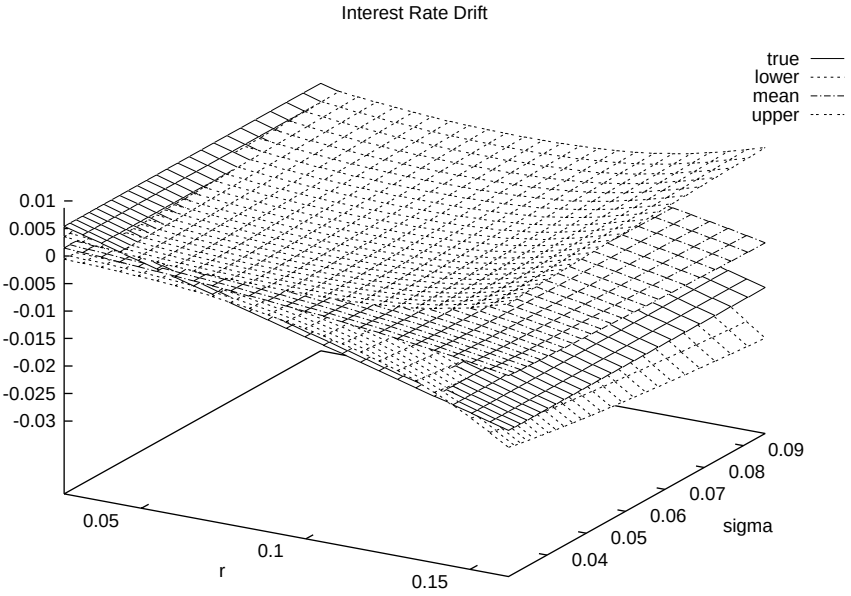


Figure 3: Estimates for Misspecified Interest Rate Process



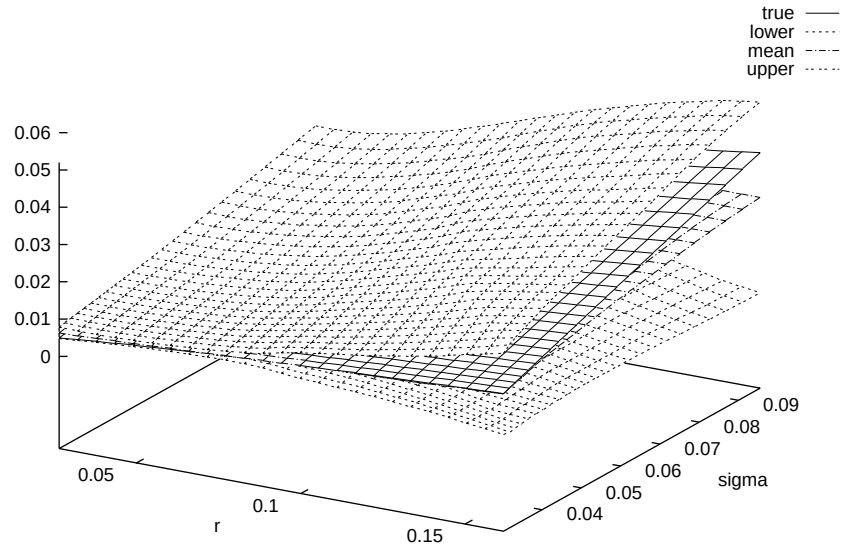


Figure 4: Estimates for Misspecified Volatility Process

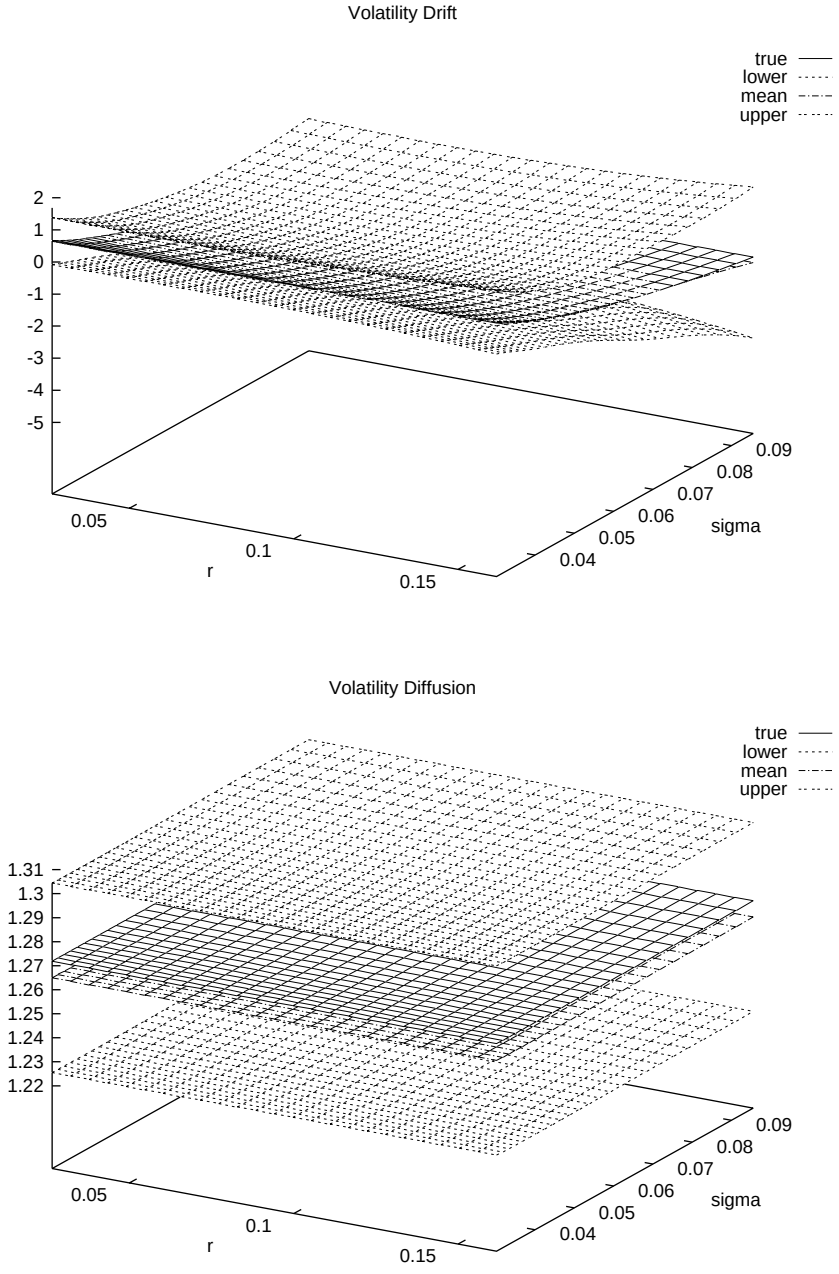
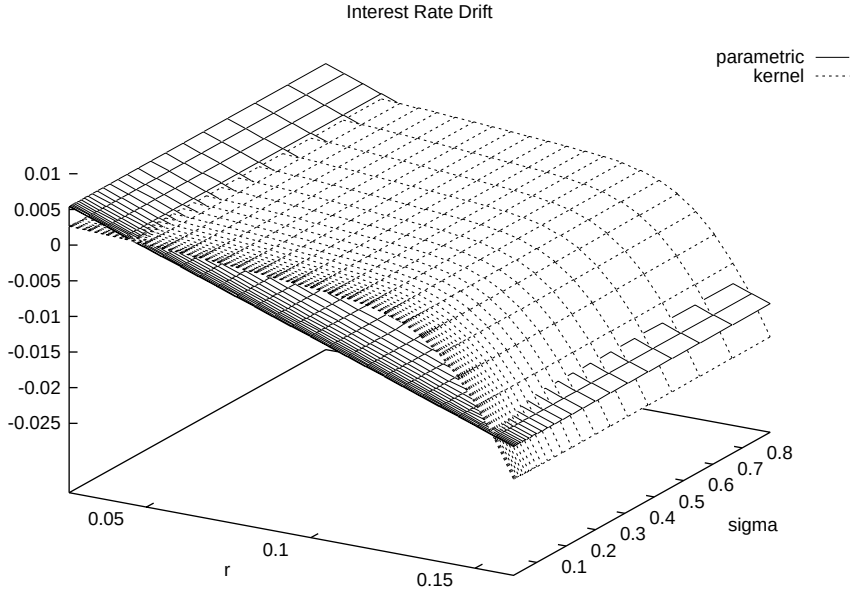


Figure 5: Interest Rate Process



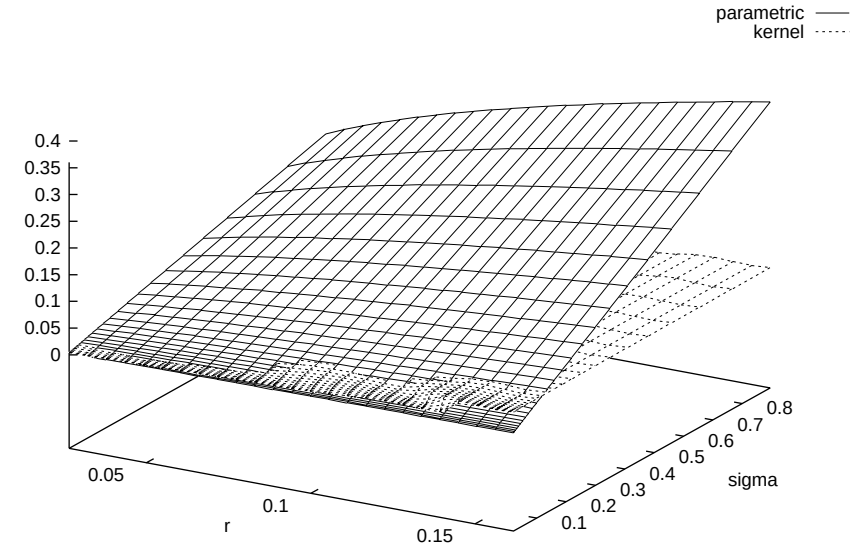


Figure 6: Volatility Process