Nonparametric Estimation of Multifactor Continuous Time Interest Rate Models

Chris Downing
Board of Governors of the Federal Reserve System [*]

May 1, 1999

Abstract

In this paper we study the finite sample properties of the nonparametric method developed in Stanton (1997), and later extended by Boudoukh, Richardson, Stanton and Whitelaw (1998), for the estimation of the drifts and diffusions of multifactor continuous–time term structure models. Monte Carlo simulations from a known parametric model are employed in order to calculate the performance of the estimator. The paper focuses on the issue of bandwidth selection. The results suggest that, for persistent data generating processes exhibiting stochastic volatility, such as interest rate data, a bandwidth function that varies over the surface of the data is optimal. The paper also presents some results on the performance of the estimator when the model is misspecified. A computationally intensive bandwidth selection procedure is developed in which dynamic graphics and a parallel kernel estimation routine are combined, allowing the researcher to interactively adapt the bandwidth surface to the data.

[*] The views expressed in this paper are those of the author and are not those of the Board of Governors or other 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`.

1 Introduction

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) pioneers a method based on the theory of weak approximations of the expectations of functionals of stochastic processes. His methodological innovation is 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 some important respects, the method developed by Ait–Sahalia (1996) is related to the Stanton and BRSW estimators in that it also relies on nonparametric regression 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 of the previously received works, 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.

In this paper, I try to develop an understanding of the finite sample properties of the BRSW estimator for multifactor models.[2] As noted above, the previous work in this area has convincingly established that the Stanton estimator is not reliable for drawing inferences about the linearity of drift functions of stochastic processes. The BRSW estimator is no different in this regard, and for this reason, I only consider the linearity issue in passing.

[1] In what follows, hopefully to avoid confusion, I refer to the the Ait–Sahalia (1996) estimator as the "Ait–Sahalia estimator," the Stanton (1997) estimator for single factor models as the "Stanton estimator," 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 don't consider it here.

In this research, I focus on the the problem of kernel bandwidth selection, and on the performance of the BRSW estimator when the model is misspecified.

The investigation into the properties of the BRSW estimator follows along lines similar to those followed in Pritsker (1998) and Chapman and Pearson (1999). I use Monte Carlo simulations of data from the stochastic volatility model of Andersen and Lund (1997a) to examine how closely the estimator fits the known drift and diffusion functions. The quality of the fits are assessed graphically, as well as with more formal measures of global and pointwise error.

An important feature of the Stanton and BRSW estimators is their reliance on kernel regression techniques. In kernel regression, one trades off variance against bias in the fit by adjusting a set of smoothing parameters, or "bandwidths," so as to minimize some sort of symmetric, bowl–shaped loss function, or so as to produce different views of the shape of the estimand (depending on one's philosophical stance on the bandwidth selection issue). Pritsker (1998) and Chapman and Pearson (1999) spend a good deal of time on the issue of bandwidth selection in the context of the Stanton estimator. What Pritsker (1998) shows, and Chapman and Pearson (1999) confirms, is that standard techniques of bandwidth selection, be they the "plug–in" methods that use bandwidths which are asymptotically optimal for independently and identically distributed data, or the data–driven methods such as cross–validation, do not work well when confronted with highly persistent data such as interest rates. In the univariate case, they find that oversmoothing the estimates tends to produce the best fits. It is also worth noting that Stanton (1997) uses a heuristic method to choose a bandwidth parameter that oversmooths the data.

In this paper, I build on the works just mentioned by considering the quality of fits produced using a bandwidth surface. Motivation for this approach comes in part from the simple observation that bandwidths which are optimal according to global error measures are not necessarily optimal for estimation at any particular point.[3] It stands to reason that one might achieve better fits by adapting a bandwidth surface to the data. Further

[3] As noted in Härdle (1990):

$$\inf_h \int E\left[\hat{f}_h - f\right]^2 \geq \int \inf_h E\left[\hat{f}_h - f\right]^2, \tag{1}$$

for bandwidth h, true function f, and estimate $\hat{f}_h$.

motivation for this approach comes from the interest in applying the estimator to processes that are stochastically volatile. By oversmoothing the data where the volatility is high, and undersmoothing where the volatility is low, a better overall fit should be produced.

I consider two bandwidth surfaces. The first is a continuous surface, in which a unique bandwidth is assigned to each point on the solution surface. This surface produces very good fits for the interest rate drift function, but results in a deterioration in the quality of the fit of the interest rate diffusion and the volatility drift relative to the benchmarks I describe below. The second surface is a simplified version of the first: Only five unique bandwidths are used to characterize the surface. This scheme produces fits of the interest rate drift that are not as accurate as the fits produced with a continuous surface. On the other hand, the fits of the other functions improve.

These preliminary results suggest that further investigation into the use of bandwidth surfaces is warranted. To further explore the use of bandwidth surfaces, I have developed a graphics interface that allows one to dynamically adjust the complexity of the bandwidth surface in order to adapt the surface to the data. This tool is described in further detail below and in the appendix.

I also examine the performance of the BRSW estimator for misspecified models. When thinking about specification error in the context of continuous time models and nonparametric estimation, it's useful to think about how one can put structure on the model by making *weak* assumptions on the forms of the drift and diffusion functions. What I mean by this is the following. Pretend we know *a priori* that the term structure is determined by two state variables: the short rate, r, and volatility, σ. To capture the dynamics of the short rate, we might write down the following model:

$$\begin{aligned} dr_t &= \alpha_r(r_t, \sigma_t)dt + \beta_r(r_t, \sigma_t)dW_{r,t} && (2) \\ d\sigma_t &= \alpha_\sigma(r_t, \sigma_t)dt + \beta_\sigma(r_t, \sigma_t)dW_{\sigma,t}, && (3) \end{aligned}$$

where the $W_{\cdot,t}$ are independent Wiener processes. This model says that the drifts, $\alpha_\cdot$, and the diffusions, $\beta_\cdot$, are functions of *both* state variables. Alternatively, we might write:

$$\begin{aligned} dr_t &= \alpha_r(r_t)dt + \beta_r(r_t, \sigma_t)dW_{r,t} && (4) \\ d\sigma_t &= \alpha_\sigma(\sigma_t)dt + \beta_\sigma(r_t, \sigma_t)dW_{\sigma,t}, && (5) \end{aligned}$$

where now the drift functions take as arguments particular state variables. These two systems are very different in terms of what they imply about how the state variables evolve. And yet, both are general, in the sense that we haven't said anything about the drift and diffusion functions beyond the dimensionality of their respective domains. [4] Nonetheless, the BRSW estimator will behave very differently when used to estimate the drift and diffusion functions of the two models. The behavior of the estimator will in each case be governed by how the specification of the model stands in relation to the true data generating process. In this study, I try to establish some useful facts about the consistency and efficiency of the BRSW estimator when faced with model misspecification of the type above.

Not surprisingly, 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 rapidly. Somewhat more surprising is the result that, in finite samples, including irrelevant conditioning variables introduces additional bias in the estimates. Work remains to be done in the area of understanding how the estimator behaves when arguments to the functions are omitted.

The biases and inefficiencies caused by misspecifying the drift and diffusion functions highlight the fact 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 to the functions (and thus the correct set of conditioning variables in the kernel regressions). In other words, nonparametric estimators don't obviate issues of specification, they are just removed to a higher level of generality.

This paper is organized as follows. In the next section, I discuss weak numeric solutions of stochastic differential equations, which lie at the foundation of all of this work. I examine the dynamic behavior of the Andersen and Lund (1997a) stochastic volatility model, which is used in the Monte Carlo simulations in section three. The third section discusses the BRSW estimator and kernel regression, and contains the main results on fitting the Andersen and Lund (1997a) model using a variety of bandwidth selection strategies and under different forms of misspecification. The final section concludes.

[4] To guarantee that unique solutions of the systems exist, we also require that the functions belong to fairly restrictive smoothness classes, and that they obey spatial and temporal growth conditions. Later in this essay we'll see that these assumptions, which are usually taken for granted, are *not* always totally innocuous.

2 Weak Numeric Solutions of Stochastic Differential Equations

A weak numeric solution of a system of stochastic differential equations (SDEs) is an algorithm for computing the expected value of functionals of the system's state variables. In this section, I discuss the calculation of weak solutions of the Andersen and Lund (1997a) model. [5] 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 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:

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

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.[6] The diffusion function of the interest rate process (6) fails to satisfy the growth condition. The relevant condition is given by:

$$\sigma^2 r + \xi^2 \leq k(1 + r^2 + (\log\sigma^2)^2). \qquad (8)$$

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 (8). For any k, let $\log\sigma^2 = r$, so that $\sigma^2 = e^r$. Substituting, we have:

$$\begin{aligned} e^r r + \xi^2 &\leq k(1 + 2r^2), \text{ or} && (9) \\ \frac{re^r + \xi^2}{(1 + 2r^2)} &\leq k. && (10) \end{aligned}$$

[5] In what follows, I refer to this as the "AL model."

[6] For more detail on these conditions, see Karatzas and Shreve (1991). Ait–Sahalia (1996) also discusses different formulations of the conditions.

The left-hand side of (10) 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 [7]:

$$\begin{aligned} dr_t &= \kappa_1(\mu - r_t)dt + \sqrt{e^{\sigma_t} r_t} dW_{1,t} && (11) \\ d\sigma_t &= \kappa_2(\theta - \sigma_t)dt + \xi dW_{2,t}, && (12) \end{aligned}$$

Because it fails to satisfy the growth condition, there might not be a unique Ito process in $\Re^2$ that satisfies (11) – (12). 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:

$$\begin{aligned} r_t &= r_{t-1} + \kappa_1(\mu - r_{t-1})\Delta + \sqrt{e^{\sigma_{t-1}} r_{t-1} \Delta}\,\eta_{1,t} && (13) \\ \sigma_t &= \sigma_{t-1} + \kappa_2(\theta - \sigma_{t-1})\Delta + \xi\sqrt{\Delta}\,\eta_{2,t}, && (14) \end{aligned}$$

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 discretized system in (13) and (14).

Understanding the dynamic behavior of the AL model, as well as evaluating the nonparametric estimator in the next section, both boil down to

[7] One can verify that (11)-(12) are equivalent to (6)-(7) using Ito's Lemma and the transformation $\hat{\sigma}_t = \log \sigma_t^2$. In equations (11)-(12), I have omitted the '$\hat{}$' symbol on σ_t for notational brevity.

computing the expectations of different functions of the state variables r and σ:

$$E\left[f(r_T, \sigma_T)\right], \tag{15}$$

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 $\Delta \to 0$:

$$\lim_{\Delta\to 0} \left| E\left[f(r_T, \sigma_T)\right] - E\left[f(\tilde{r}_T, \tilde{\sigma}_T)\right] \right| = 0. \tag{16}$$

By letting $f(r,\sigma)$=(r,σ), we can use the Euler scheme to compute the moments of transition densities of the AL model.

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 as the length of the time interval increases:

$$\lim_{T\to\infty} \pi(r_T, \sigma_T | r_0, \sigma_0) \xrightarrow{d} \pi(r, \sigma), \tag{17}$$

for $r_0 \varepsilon \Re^{++}$ and $v_0 \varepsilon \Re$, 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 simulator to make draws from the transition densities defined by different starting points (r_0, σ_0) and by different 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.[8]

Convergence in distribution is a broad concept. It is perhaps easier 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 $[0.02 \leq r \leq 0.2] \times [-7 \leq \log \sigma^2 \leq -5]$, 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

[8] 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.

Table 1: Parameter Values*

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

* Annualized values.

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 1, and $\Delta = \frac{1}{52}$. [9]

Table 2 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. In table 2, we can look for evidence of this condition in the first two moments of the transition densities. The means should converge as follows:

$$\lim_{T\to\infty} E[r_T] = 0.0596, \tag{18}$$

$$\lim_{T\to\infty} E[\sigma_T] = -6.3599, \tag{19}$$

which are the long–run means that the processes should revert to if they are stationary. Examining the values in the second column (labeled 'Mean') of table 2, 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 (18) 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 (19), reflecting the higher degree of mean reversion

[9] 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 1 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).

in the volatility drift function. [10]

The second moments should converge approximately as follows:

$$\lim_{T\to\infty} \mathrm{Var}[r_T] \approx 0.00032 \tag{20}$$

$$\lim_{T\to\infty} \mathrm{Var}[\sigma_T] \approx 0.7780, \tag{21}$$

The approximate value for the second moment of r is calculated as the stationary variance for a square–root process, holding σ fixed at θ, and 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 Vasicek process, given by $\frac{\xi^2}{2\kappa_2}$. Returning to table 2, it appears that the variances (Var[$\cdot$] values) are converging to neighborhoods of the values in (20) and (21). In the case of the interest rate process, we would probably reject the null hypothesis that the variance is equal to the value in (20), 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 (21). This is because the dependence between the interest rate and volatility processes is expressed in the diffusion function of the interest rate process; the volatility process in fact does evolve like the process that we used to compute the variance.

Of course, the transition densities could appear to be converging in the first two moments, and still have very different distributions. Moreover, it's hard to assess joint significance using table 2. To more rigorously test for convergence in distribution when the true distribution is unknown, we can make use of an adaptation of the Kolmogorov- Smirnov (KS) test to bivariate densities, due to Fasano and Franceschini (1987).

The one dimensional KS test is based on the maximum value of the absolute difference between two cumulative distribution functions. A direct generalization of this test to higher dimensions is not possible because cumulative probability is not well defined in more than one dimension. However, an analogous measure can be based on the integrated probabilities in each of four quadrants around 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

[10] 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.

Table 2: Simulation Results

Moment	Mean	Std Dev	Min	Max
$E[r_1]$	0.0769055	0.0420346	0.0163075	0.1401690
$E[r_5]$	0.0685385	0.0218760	0.0353890	0.1050336
$E[r_{10}]$	0.0634786	0.0097218	0.0462903	0.0830092
$E[r_{20}]$	0.0602762	0.0022962	0.0529907	0.0682432
$E[r_{30}]$	0.0596413	0.0013505	0.0547802	0.0658052
$E[r_{40}]$	0.0595153	0.0012983	0.0547687	0.0647093
$E[r_{50}]$	0.0594871	0.0012930	0.0549952	0.0651942
$E[r_{60}]$	0.0594904	0.0012985	0.0547620	0.0651492
$\text{Var}[r_1]$	0.0002027	0.0001629	0.0000089	0.0011600
$\text{Var}[r_5]$	0.0004750	0.0002337	0.0000743	0.0020648
$\text{Var}[r_{10}]$	0.0005068	0.0001673	0.0001403	0.0018659
$\text{Var}[r_{20}]$	0.0004821	0.0001101	0.0001767	0.0019874
$\text{Var}[r_{30}]$	0.0004725	0.0001025	0.0002012	0.0016414
$\text{Var}[r_{40}]$	0.0004707	0.0001029	0.0001164	0.0016819
$\text{Var}[r_{50}]$	0.0004700	0.0001030	0.0001849	0.0015028
$\text{Var}[r_{60}]$	0.0004704	0.0001028	0.0001762	0.0016165
$E[\sigma_1]$	-6.2339870	0.2473902	-6.5838426	-5.8841313
$E[\sigma_5]$	-6.3580136	0.0037063	-6.3632550	-6.3527723
$E[\sigma_{10}]$	-6.3598901	0.0000194	-6.3599176	-6.3598626
$E[\sigma_{20}]$	-6.3599000	0	-6.3599000	-6.3599000
$E[\sigma_{30}]$	-6.3599000	0	-6.3599000	-6.3599000
$E[\sigma_{40}]$	-6.3599000	0	-6.3599000	-6.3599000
$E[\sigma_{50}]$	-6.3599000	0	-6.3599000	-6.3599000
$E[\sigma_{60}]$	-6.3599000	0	-6.3599000	-6.3599000
$\text{Var}[\sigma_1]$	0.6971374	0.1399842	0.2701711	1.4404008
$\text{Var}[\sigma_5]$	0.7932272	0.1590059	0.2567814	1.6098411
$\text{Var}[\sigma_{10}]$	0.7936240	0.1596618	0.3470085	1.5934727
$\text{Var}[\sigma_{20}]$	0.7945158	0.1591197	0.2875762	1.5405627
$\text{Var}[\sigma_{30}]$	0.7950285	0.1589929	0.2838963	1.6439167
$\text{Var}[\sigma_{40}]$	0.7942603	0.1589761	0.2723940	1.6174558
$\text{Var}[\sigma_{50}]$	0.7950513	0.1597367	0.3147770	1.5621902
$\text{Var}[\sigma_{60}]$	0.7934561	0.1593903	0.3054477	1.5795967

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. [11]

To carry out the test, I use two starting values that are widely apart on the (r, σ) plane. The points that I use are $(\mu+2\hat{\sigma}_r,\theta+2\hat{\sigma}_\sigma)$ and $(\mu-2\hat{\sigma}_r,\theta-2\hat{\sigma}_\sigma)$, points roughly two standard deviations away from the long-run means of the processes, and about four standard deviations from one another.[12] 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 2, respectively. From each of these points, I 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 the same as before.[13]

Table 3 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 transition densities 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 2, we can conclude with a high degree of confidence that the system does in fact have a stationary density.

[11] 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 papers cited above and Press, Teukolsky, Vetterling and Flannery (1994).

[12] Picking points farther out in the tails of the distribution will of course bias the test toward finding convergence at longer trajectories. On the other hand, from the results in table 2, 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.

[13] 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.

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) run the Euler simulator for approximately thirty-eight years. The authors find that using longer trajectories had no discernable effects on their results, which is consistent with our finding here that the transition densities converge at around forty years.[14]

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 2 in what follows. In the next section, we turn to considering the behavior of the BRSW estimator in finite samples.

[14] 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.

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 3: Bivariate KS Test Results

3 Nonparametric Estimation

In this section, I analyze the finite sample performance of the BRSW estimator. Using Monte Carlo simulations of data from the AL model, I measure the accuracy with which the estimator reproduces the drift and diffusion functions of the model. I begin by briefly reviewing the BRSW estimator and kernel estimation, and then discuss some different strategies for selecting bandwidths for the kernel estimator. I present some results on the performance of the estimator under different bandwidth selection strategies, and I present some results on how well the estimator performs when the model is misspecified.

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

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

and suppose that we observe data generated from the true processes in (22) and (23) at discrete intervals of time Δ. The Euler method of the previous section is one way to relate our discrete observations to the drift and diffusion functions of the true processes. The Euler discretization for this system is given by:

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

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:

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

where $O(\Delta)$ means terms for which it is true that $\lim_{\Delta\to 0}\frac{O(\Delta)}{\Delta} < \infty$, and F_t is 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:

$$\begin{aligned} E\left[r_{i,t+1} - r_{i,t} | (r_i, \sigma_j)\right] &= \sum_{t=1}^{T} W(t)(r_i - r_t) & (30) \\ E\left[\sigma_{i,t+1} - \sigma_{i,t} | (r_i, \sigma_j)\right] &= \sum_{t=1}^{T} W(t)(\sigma_i - \sigma_t) & (31) \\ E\left[(r_{i,t+1} - r_{i,t})^2 | (r_i, \sigma_j)\right] &= \sum_{t=1}^{T} W(t)(r_t - r_{t-1})^2, \text{and} & (32) \\ E\left[(r_{i,t+1} - r_{i,t})^2 | (r_i, \sigma_j)\right] &= \sum_{t=1}^{T} W(t)(\sigma_t - \sigma_{t-1})^2, & (33) \end{aligned}$$

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. Wide bandwidths reduce local variation, but increase bias. Narrow bandwidths fit local phenomena, at the cost of increased variance.

In all of the fits reported below, I report pointwise averages for fits of the drift and diffusion functions on a 12 × 12 grid of values, where the averages are taken over 1000 simulations from the AL model. The "true" functions are parameterized using the values shown in table 1 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.

First I consider fits using bandwidths that are asymptotically optimal for data that are independently and identically distributed, even though the data are highly persistent, because the fits serve as useful benchmarks for later comparison.[16] Two bandwidths are used to compute the benchmark fits - one for each dimension. The bandwidth for the interest rate dimension is set to the value that is asymptotically optimal for *iid* data drawn from a distribution with variance 0.0004704, the value of $\text{Var}[r_{60}]$ from table 2 of the previous section. The bandwidth for the volatility dimension is similar, except that the variance of the distribution is taken to be 0.7934561, also taken from table 2.

Figures 1 and 2 display the benchmark fits. The upper panel of each figure displays the estimated drift function and the true drift function, and the lower panel displays the estimated and true diffusion functions. In general, the diffusions are more precisely estimated than the drifts. This follows from the fact that the precision of the drift estimates depends on the span of the data $[0, T]$, while the precision of the diffusion estimates depends on the span of the data *and* the length of the time step Δ.

Two effects cause the estimated surfaces to deviate substantially from the true surfaces at the boundaries of the data. Near the boundaries of the data, since the kernel function is symmetric, the weights are skewed 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 is truncation bias, or bias resulting from the correlation of the residuals with the regressors near the edges of the data.

[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.

[16] Scott (1992) contains a discussion of bandwidth selection strategies for *iid* data.

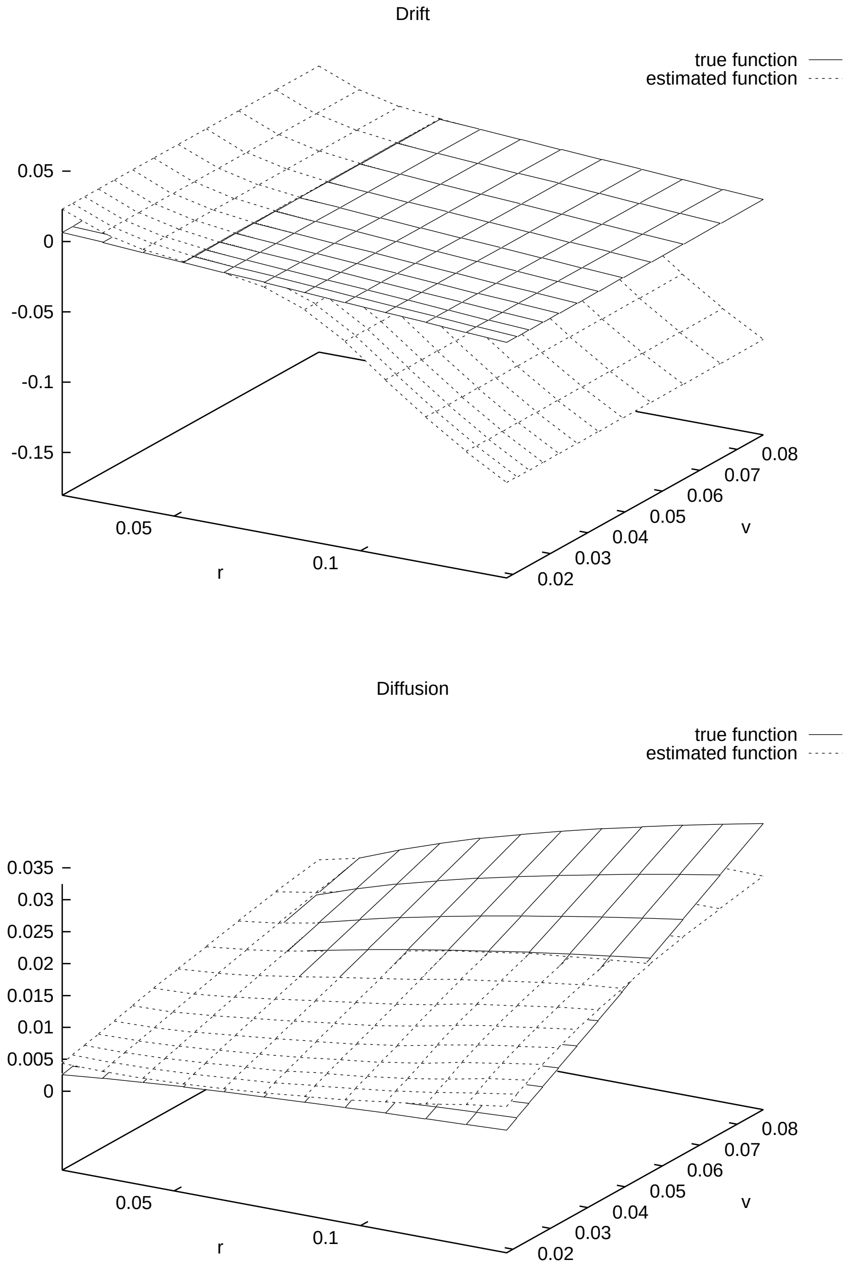


Figure 1: Interest Rate Process Benchmark

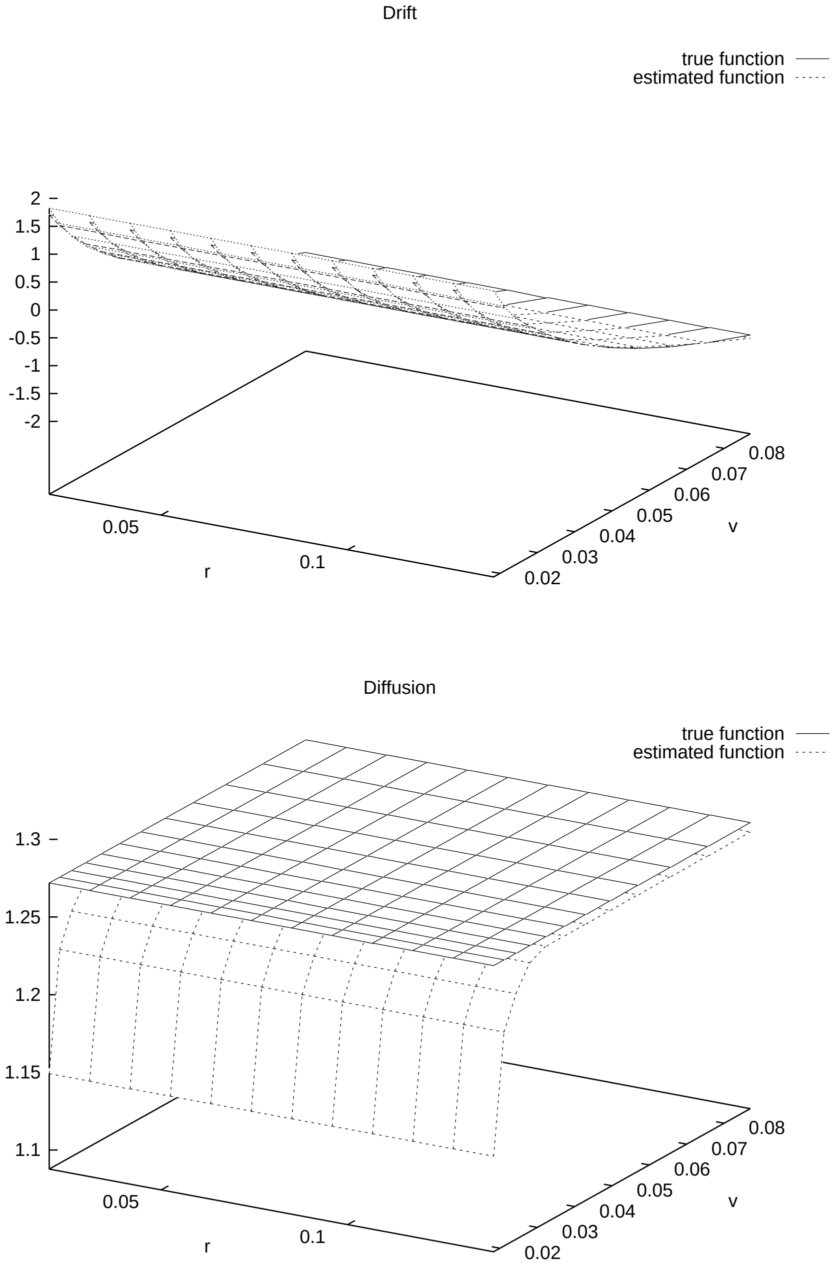


Figure 2: Volatility Process Benchmark Fit

The nonparametric regression model for the drifts is given by:

$$\begin{aligned} r_{t+\Delta} - r_t &= \alpha_r + \epsilon_{r,t+\Delta} && (36) \\ \sigma_{t+\Delta} - \sigma_t &= \alpha_\sigma + \epsilon_{\sigma,t+\Delta} && (37) \end{aligned}$$

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

$$\begin{aligned} E\left[\epsilon_{r,t+\Delta}|r_t, \sigma_t\right] &= 0, \text{ and} && (38) \\ E\left[\epsilon_{\sigma,t+\Delta}|r_t, \sigma_t\right] &= 0. && (39) \end{aligned}$$

Truncation bias arises because, in fact, the nonparametric estimator works with a finite data set for which (38) and (39) don't necessarily hold at the boundaries of the data. For example, at the data point where:

$$(r_t, \sigma_t) = (r\text{max}, \sigma), \qquad (40)$$

it must be the case that:

$$r_{t+\Delta} - r_t \leq r\text{max} - r_t. \qquad (41)$$

In other words, at the upper boundary of the observations on r, the residual in equation (36) 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.

In the top panels of figures 1 and 2, the biases follow patterns similar to those found by Pritsker (1998) and Chapman and Pearson (1999). At high interest rates, the interest rate drift function estimate is biased downward, and vice–versa, indicating that the effect of truncation bias is dominant. A similar pattern holds for the estimates of the volatility drift function, although in general the volatility drift is estimated much more precisely. This is to be expected, because along the volatility dimension, the data are much less persistent (the degree of mean reversion is roughly an order of magnitude higher).

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 precisely, except at the lowest levels of σ, where the data are thin. In this region, the limitations of the "one size fits all" bandwidth strategy is apparent. The estimates are biased strongly downward because the bandwidth is too narrow given the dispersion of the data. If the surface were extended in the direction of higher values of volatility, we would see a similar effect in the other direction.

It is useful to compute some measures of global and pointwise error. I compute three error measures, each capturing a slightly different aspect of the distance between the estimated function and the true function. The first is the mean of the pointwise squared errors, defined by:

$$MSE = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\hat{f}_{i,j} - f_{i,j})^2. \tag{42}$$

As a complement to MSE that doesn't emphasize extreme errors so much, I also compute an estimate of the integrated absolute deviation, given by:

$$IAD = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |\hat{f}_{i,j} - f_{i,j}|. \tag{43}$$

And finally, as a measure of the most egregious error on the surface, I compute the maximal absolute deviation:

$$MAD = \max |\hat{f}_{i,j} - f_{i,j}|, \tag{44}$$

where the maximum is taken over all i and j. The top set of figures in table 4 displays these error measures for the surfaces shown in figures 1 and 2.

Figures 3 and 4 display surfaces formed by the 95% confidence intervals around the point estimates shown in figures 1 and 2. The variance of the estimator increases near the boundaries of the data. Toward the boundaries of the data, the upper and lower confidence surfaces are very far apart, indicating that one could fit any of a variety of nonlinear surfaces in the space between them, none of which are statistically distinguishable. The results show that, with a single draw from the data generating process (as would be the case in reality), it is in fact quite likely the case that one would estimate a surface that exhibits spurious nonlinearities.

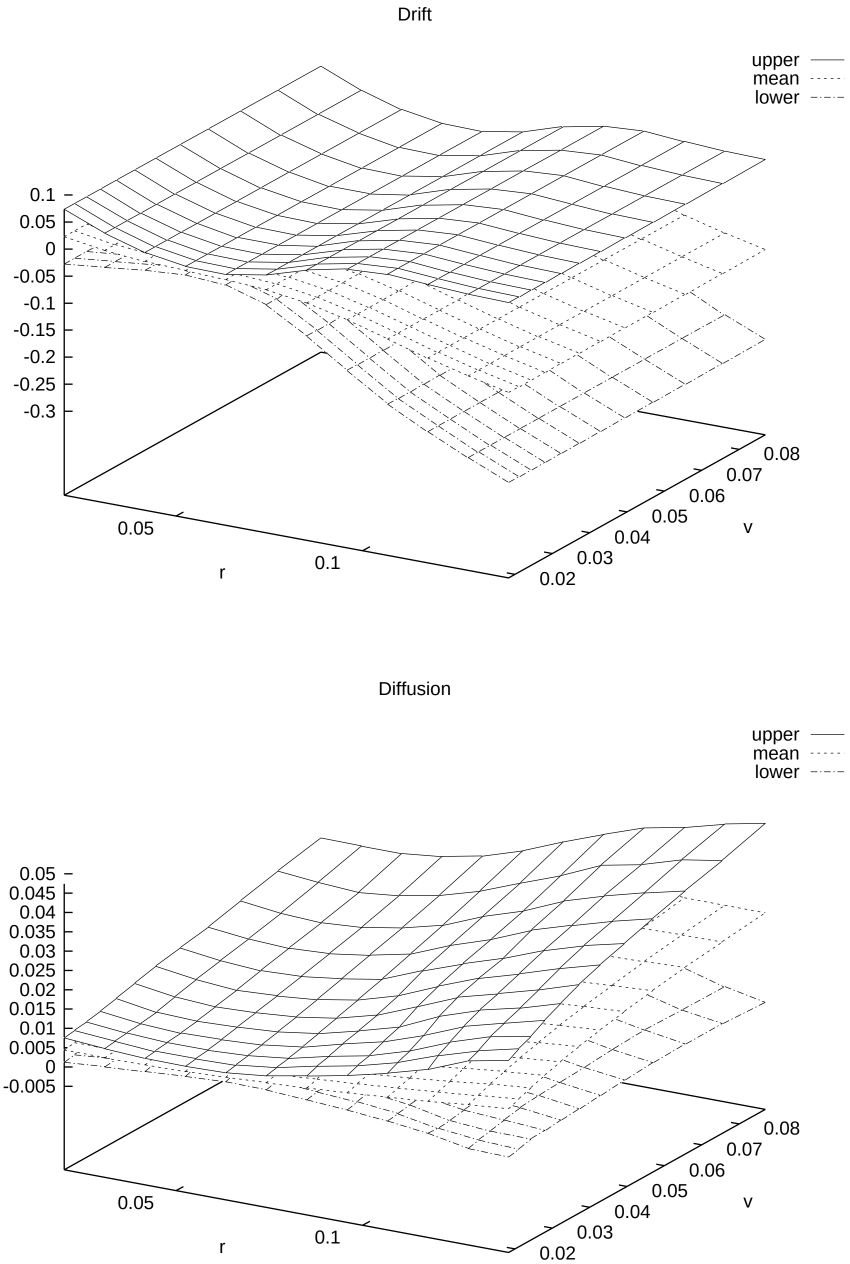


Figure 3: Pointwise Confidence Intervals, Interest Rate Process Benchmark Fit

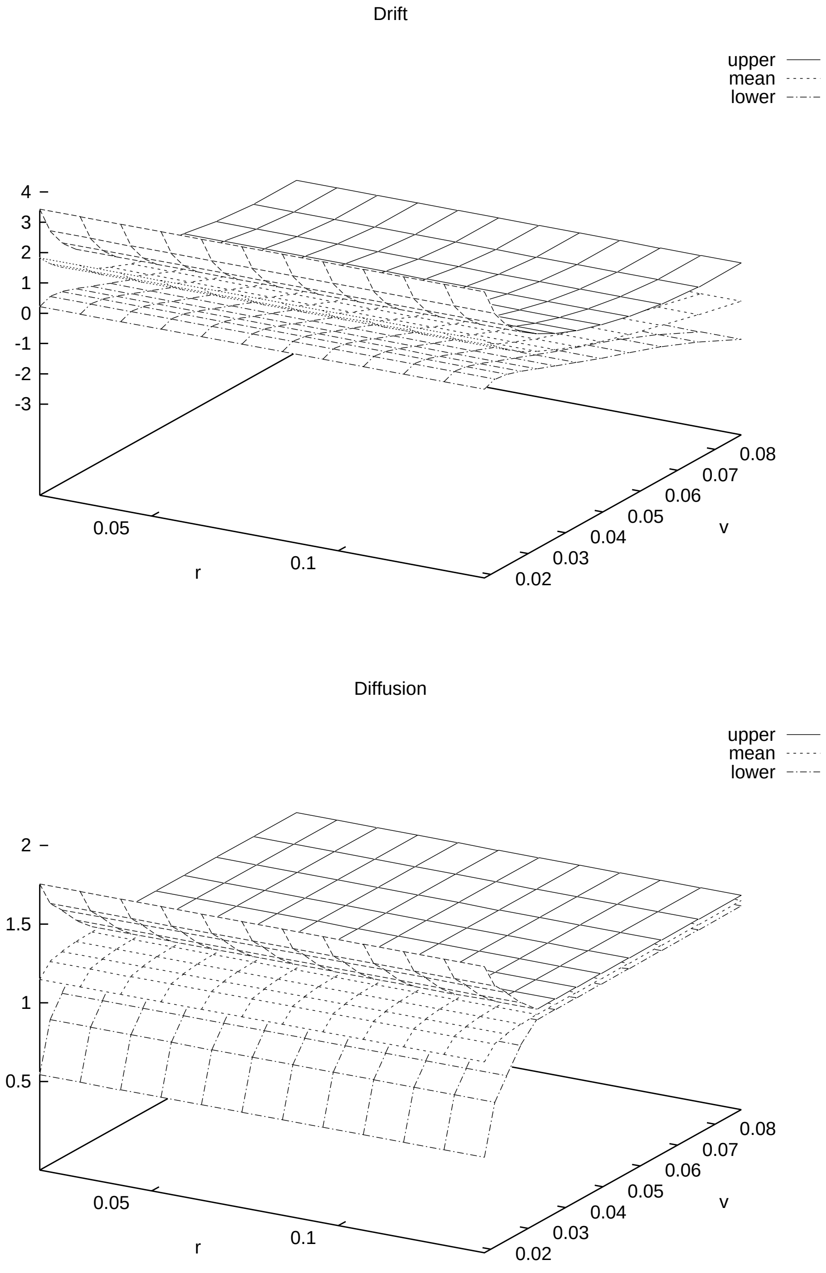


Figure 4: Pointwise Confidence Intervals, Volatility Process Benchmark Fit

3.1 Bandwidth Surfaces

Pritsker (1998) shows that by oversmoothing, one can induce greater cancellation of the truncation and edge biases in the estimators of the drift functions. However, the patterns of bias in the multidimensional case can be complicated, because certain points on the solution grid are in the middle of the data along some dimensions, but at the boundary of the data along other dimensions. Depending on the nature of the estimand, oversmoothing would indeed reduce bias in some regions, but in other regions, oversmoothing could make the fits much more biased. To take advantage of the bias reduction in oversmoothing, we need to locally adapt the bandwidths to the data.

Bias reduction is not the only motivation for working with a bandwidth surface. Because the AL model is stochastically volatile, in certain regions of the state space the noise in the observations will be greater than in other regions. In the noisier regions, we would like to smooth the data more than in other regions. A bandwidth surface allows one to do so.

To begin the investigation into the use of bandwidth surfaces, I first re–estimate the drift and diffusion functions with a bandwidth surface that allows one to assign a separate bandwidth vector to each point on the estimation surface. I use the following formula to compute the bandwidth at each point on the solution grid:

$$h_{i,j} = \hat{\sigma}_{i,j} T^{-\frac{1}{6}}, \tag{45}$$

where $\hat{\sigma}_{i,j}$ measures the dispersion of the data around point $(r_{i,j}, \sigma_{i,j})$. The first coordinate of $\hat{\sigma}_{i,j}$ is computed as:

$$\hat{\sigma}_{i,j}^{(1)} = \frac{1}{T}\sum_{t=1}^{T}(r_t - r_i)^2, \tag{46}$$

with the second coordinate computed analogously.[17] Equation 45 is very simple. It just says that the bandwidth is wider when the data are more dispersed around the solution point, and vice–versa.

Figures 5 and 6 display the fitted surfaces for the simple variable bandwidth selector above. The fits improve dramatically over those shown in figures 1 and 2. Of particular interest is the interest rate drift function.

[17] It might be interesting to try working in covariance information, but I haven't figured out how to do so. Even more interesting would be to use bandwidths that adapt the amount of smoothing to the degree of autocorrelation in the data.

The estimated surface is now almost linear. Looking at the lower panel of figure 5, we see that the bandwidth surface worsens the estimates of the diffusion function. This is to be expected, because the bias cancellations that occur in the estimates of the drift function do not occur here.

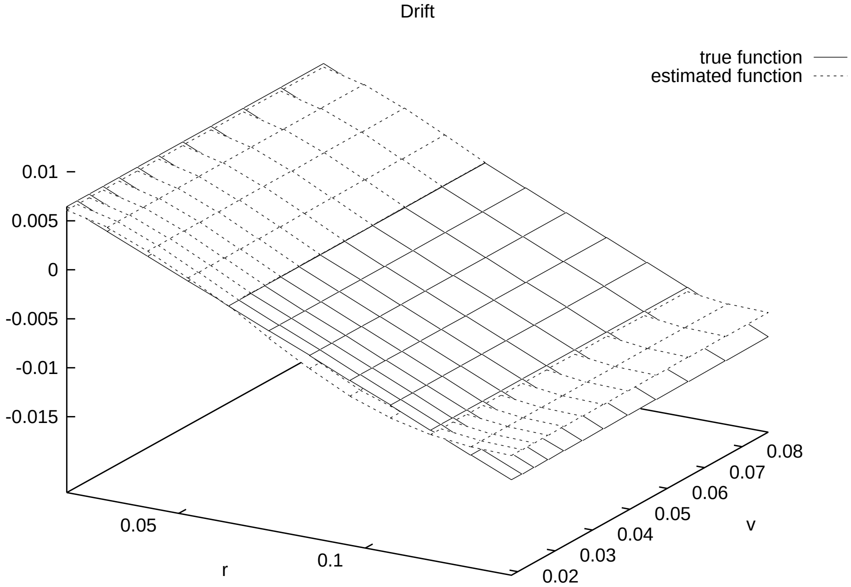
In figure 6, it is apparent that the variable bandwidth surface doesn't improve the fit of the volatility drift function. This is because the volatility process is less persistent relative to the interest rate process. Moreover, in the volatility dimension, the dispersion of the data is homoscedastic, while in the interest rate dimension, it is heteroscedastic. The volatility drift is estimated with a fair degree of precision in both the static and variable bandwidth cases, showing that for well behaved data, the estimates are less sensitive to the way one handles bandwidth selection.

The bandwidth surface improves the estimates of the volatility diffusion function. Comparing the lower panel of figure 6 to the lower panel of figure 2, the bias at low levels of σ that is exhibited in figure 2 has completely disappeared. This is because the bandwidth naturally adapts to the dispersion of the data, oversmoothing the regions where the data are sparse, and smoothing relatively less the regions where observations are more abundant.

The bandwidth surface that we were just working with is defined by a large number of points with unique values. It is interesting to consider whether or not we can achieve similar results with a bandwidth surface defined by fewer unique values. Next I consider a bandwidth surface in which there are only five unique values. The bandwidths in the center of the surface take one value. The bandwidths in the four natural quadrants around the center region also take distinct values. The bandwidth values in each of the five regions are computed using the dispersion of the data around the point in the center of the region. Figure 7 shows the bandwidth values. The top panel shows the values for the r–coordinate, and the lower panel shows the values for the σ–coordinate. In the center of the data, the bandwidths are relatively small. Around the edges, the bandwidths are increasing in value in the dispersion of the data in the respective dimensions.[18]

Figures 8 and 9 show the estimated functions for the simplified bandwidth surface. The results show that the simplified surface does a better job of fitting the interest rate drift function than the benchmark, but not as good

[18] Bear in mind that certain functions use certain dimensions of the bandwidth surface. For example, the interest rate drift only depends on r, and so will only use the bandwidth values in the top panel of figure 7.



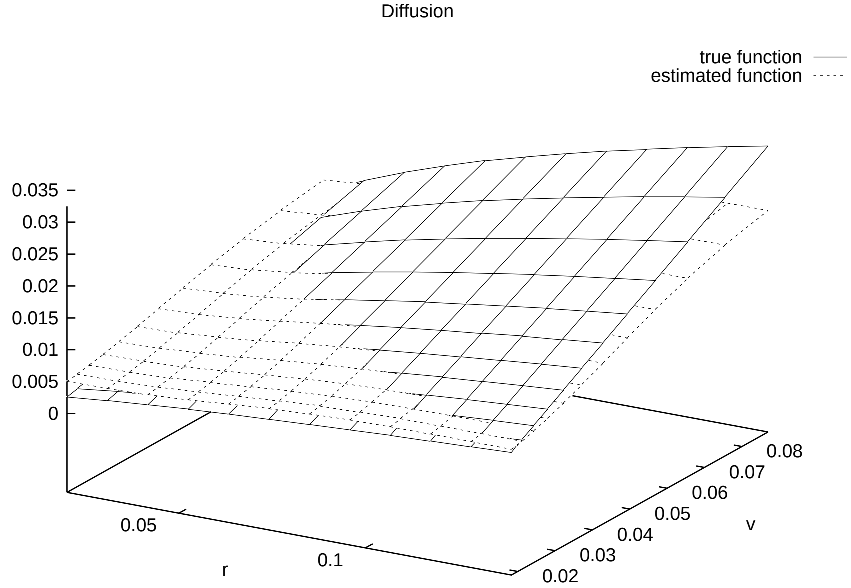


Figure 5: Interest Rate Process with Bandwidth Surface

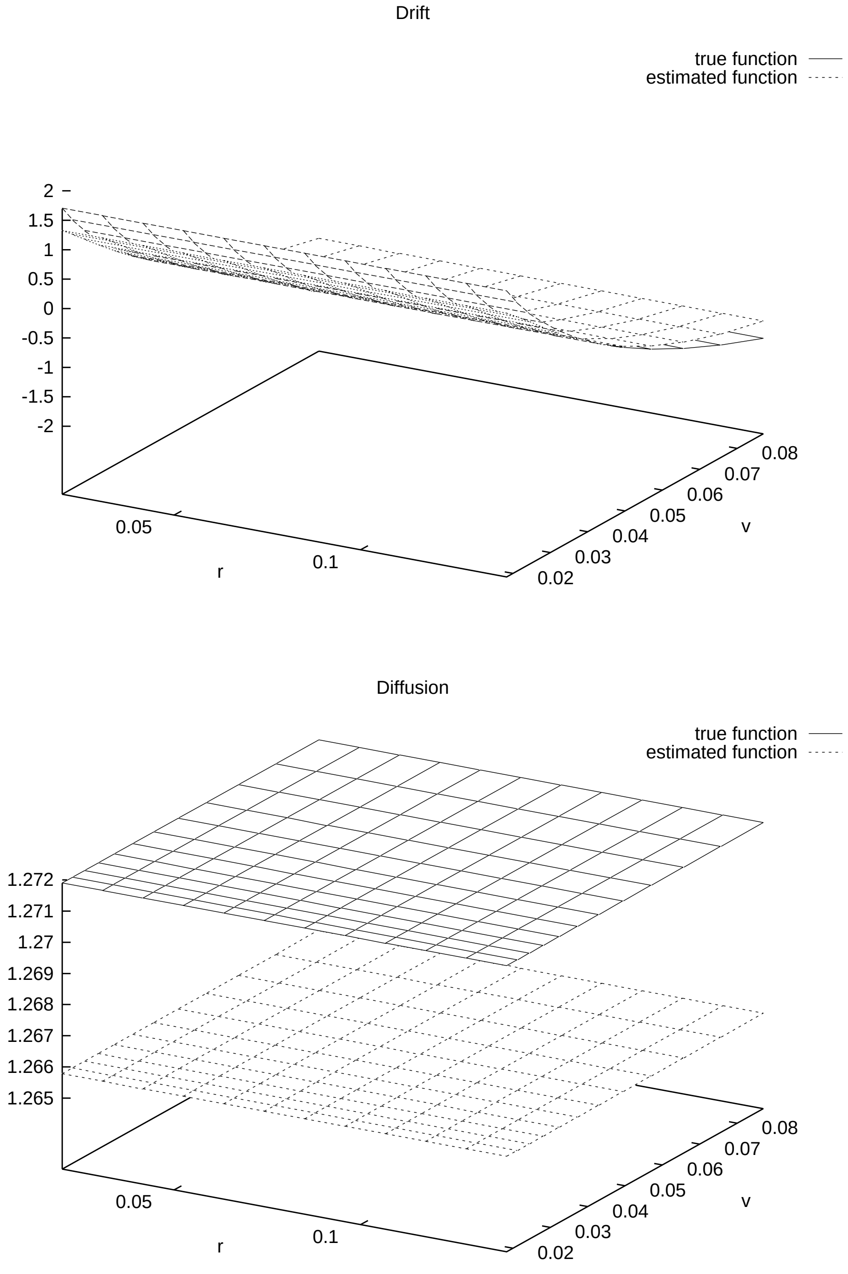


Figure 6: Volatility Process with Bandwidth Surface

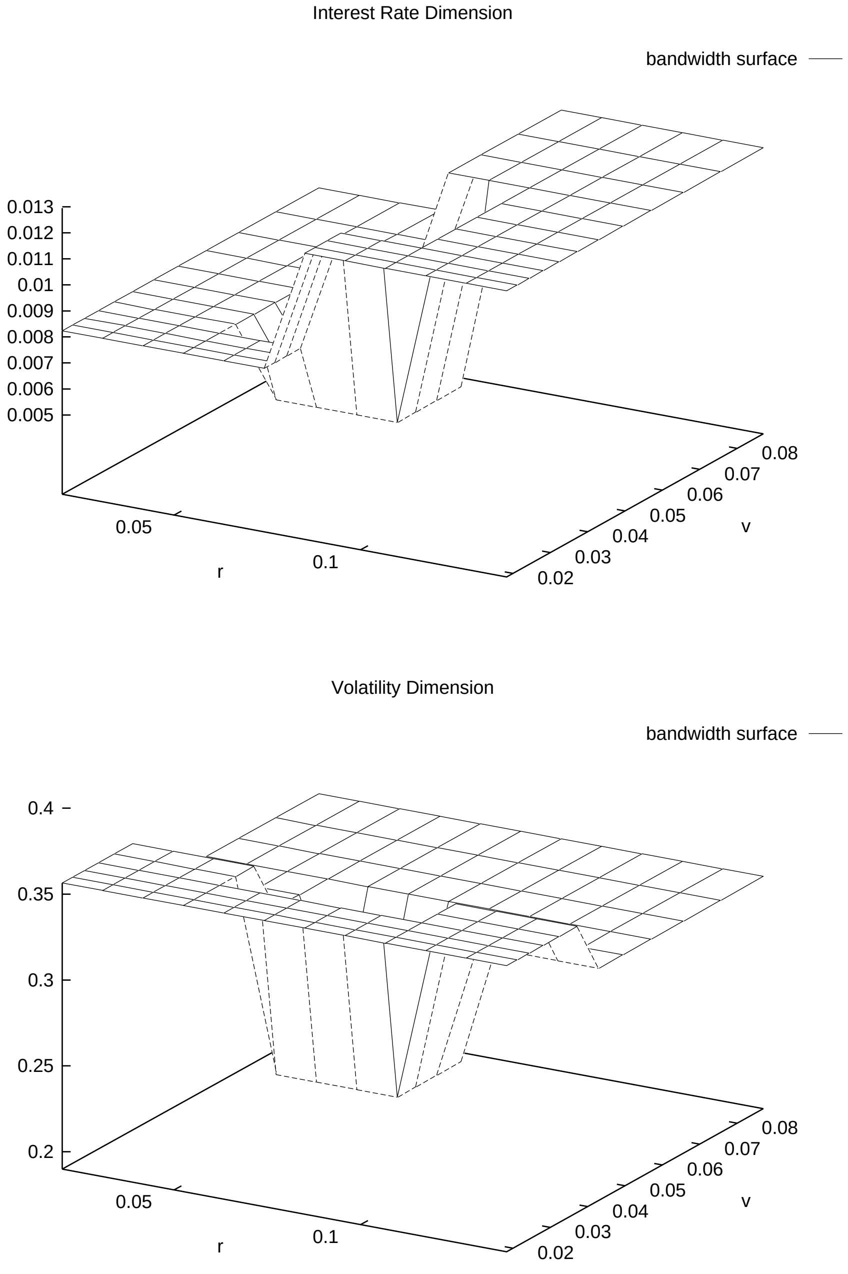


Figure 7: Simplified Bandwidth Surface

a job as the continuous surface. Comparing the top panels of figures 8 and 5, we see that there are some kinks in the surface in figure 8, introduced by the discontinuities in the bandwidth surface, and these biases are reflected in the error measures shown in table 4. For the other functions, the simplified surface does at least as well, if not better, than the continuous surface. In particular, the fit of the volatility drift function is much improved over the fit using a continuous surface.

	Benchmark		
Function	MSE	IAD	MAD
α_r	0.0020932	0.030712	0.09961
β_r	0.0000035	0.001322	0.00823
α_σ	0.0018271	0.029273	0.11936
β_σ	0.0015293	0.021298	0.12280
	Continuous Surface		
α_r	0.0000012	0.000867	0.00248
β_r	0.0000061	0.001725	0.01012
α_σ	0.0311559	0.130926	0.37952
β_σ	0.0000374	0.006117	0.00612
	Simplified Surface		
α_r	0.0000807	0.005354	0.02282
β_r	0.0000037	0.001292	0.00812
α_σ	0.0072974	0.075745	0.13021
β_σ	0.0000374	0.006117	0.00612

Table 4: Error Measures for Bandwidth Surfaces

These preliminary results suggest that there is much to be gained from further exploration of bandwidth surfaces. In particular, one would like to be able to dynamically adjust the complexity of the bandwidth surface so as to adapt to the data the amount of smoothing done in different regions of the solution space. A dynamic graphics interface is well suited to this purpose.

Figure 10 shows screen shots of the main parts of the dynamic graphics interface that I've developed for purposes of doing kernel regressions using bandwidth surfaces.[19] The idea is simple: The interface allows one to use a mouse to select points on the solution grid, and then input a bandwidth

[19] All of the source code for the algorithms discussed in this paper is available upon request.

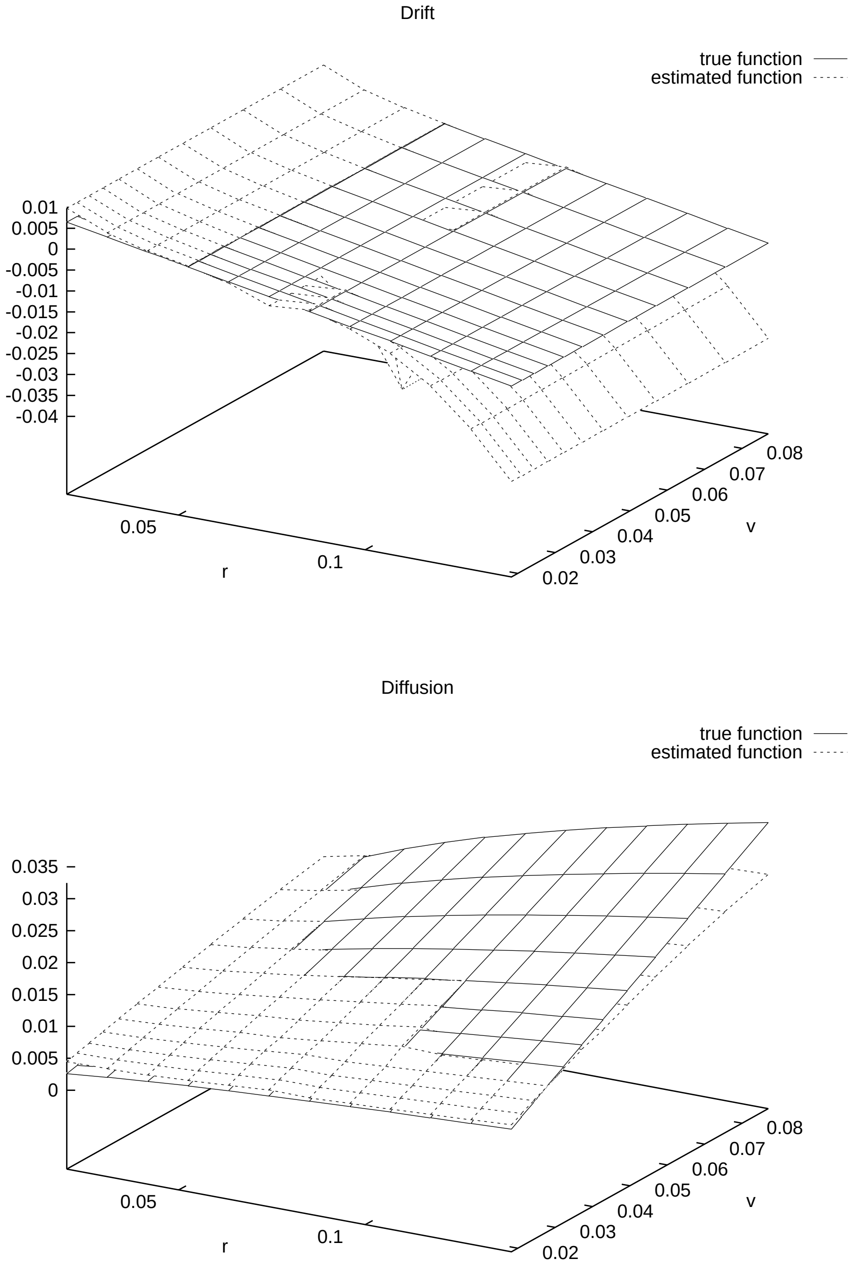
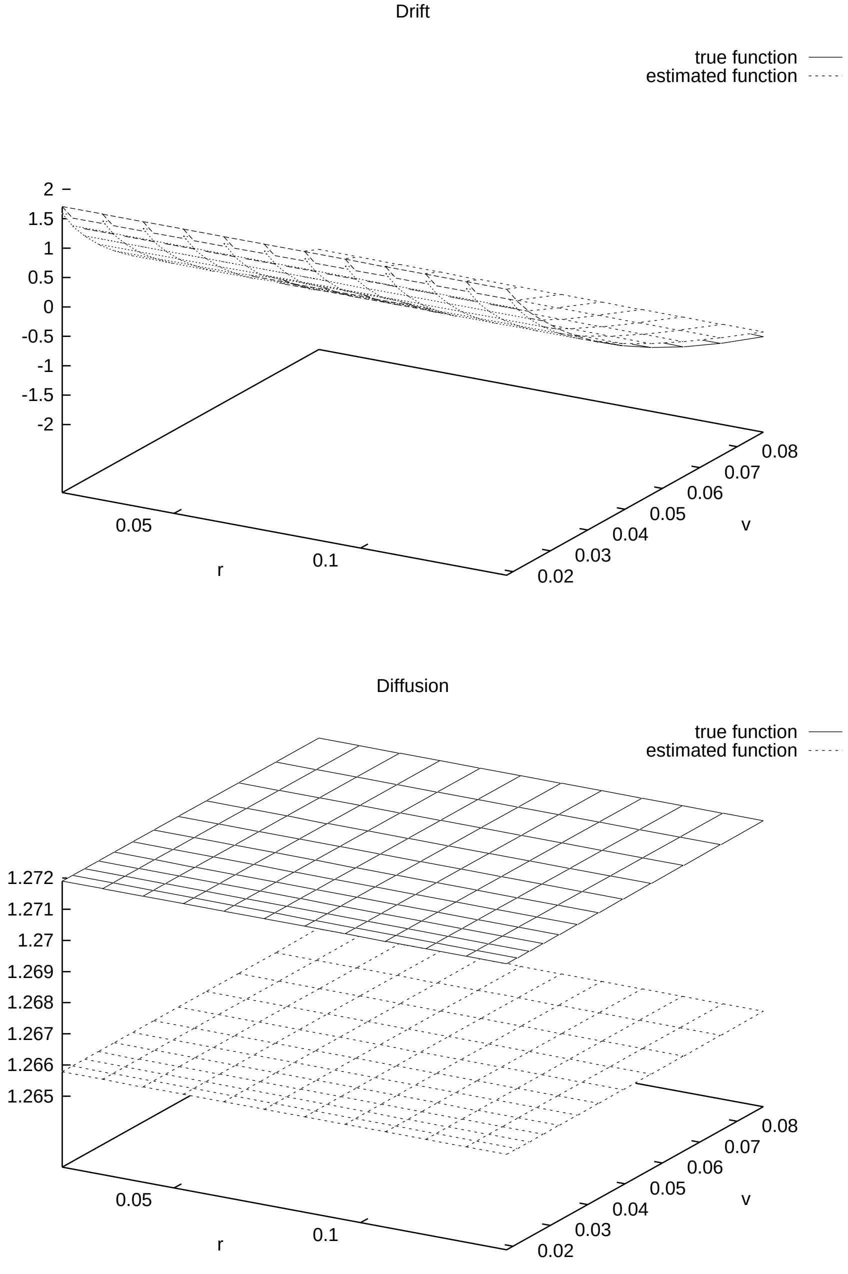


Figure 8: Interest Rate Process with Simplified Bandwidth Surface



Drift

Diffusion

Figure 9: Volatility Process with Simplified Bandwidth Surface

to apply at the selected points. The fit updates in real time, allowing one to interactively explore how the fit responds to different bandwidth surfaces. The top panel of figure 10 shows an example function (which can be rotated), and the lower panel shows the interface to the solution grid, from which points can be selected.

In order for this tool to be useful, the kernel regression algorithm must be fast. In the appendix, I discuss a parallel kernel regression algorithm that is used in conjunction with the dynamic graphics interface.

3.2 Misspecification

The estimates shown in figures 1–4 were computed for the "best case" where it was assumed that we knew *a priori* 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 true data generating process. What would happen if, as is the case in reality, we didn't know what the arguments should be, and we misspecified them? Suppose we estimated the more general system in (22)-(23), for which the drift functions are misspecified, and the diffusion function of the volatility process is misspecified. It is interesting to look at how this form of misspecification affects the bias and efficiency of the estimator.

Figures 11 and 12 display the fitted surfaces and the true function surfaces. The results are surprising because it appears to be the case that introducing irrelevant conditioning variables biases the estimates. Starting with the top panel of figure 11, the surface has a distinct curvature along the volatility dimension for high values of r. For low values of r, the surface also has a non–zero gradient along σ, although it is less pronounced. The irrelevant conditioning variable introduces additional sources of edge and truncation bias in a finite sample.

Comparing the top panel of figure 12 to the top panel of figure 2, we see a similar effect for the estimates of the volatility drift. Note the bowl–shaped pattern of bias along the interest rate dimension in figure 12, which is especially pronounced at low levels of volatility. None of this is seen in figure 2.

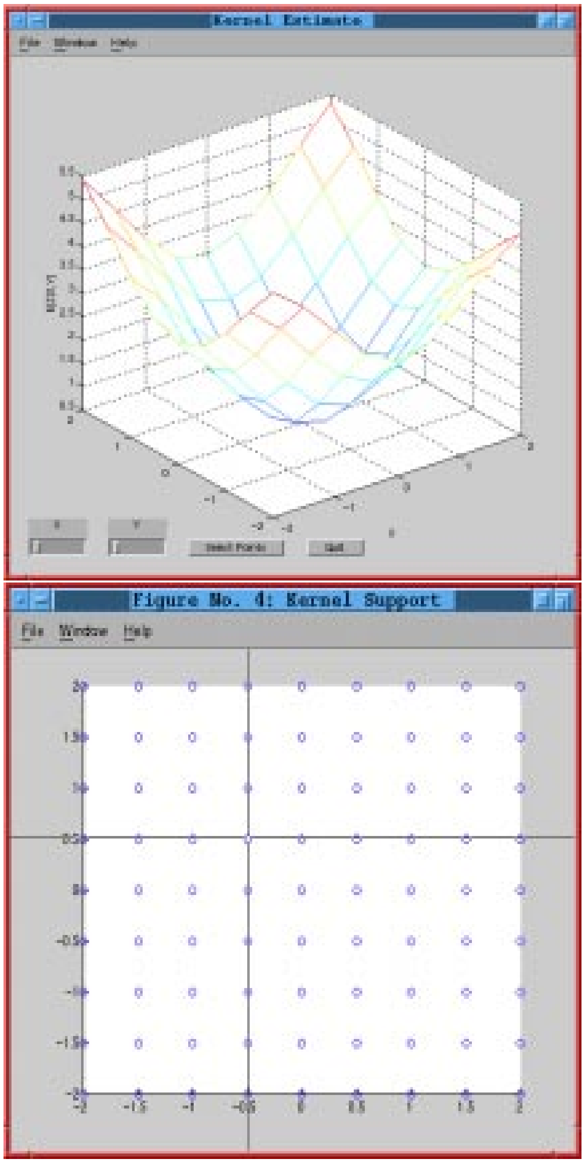


Figure 10: Dynamic Graphics Interface

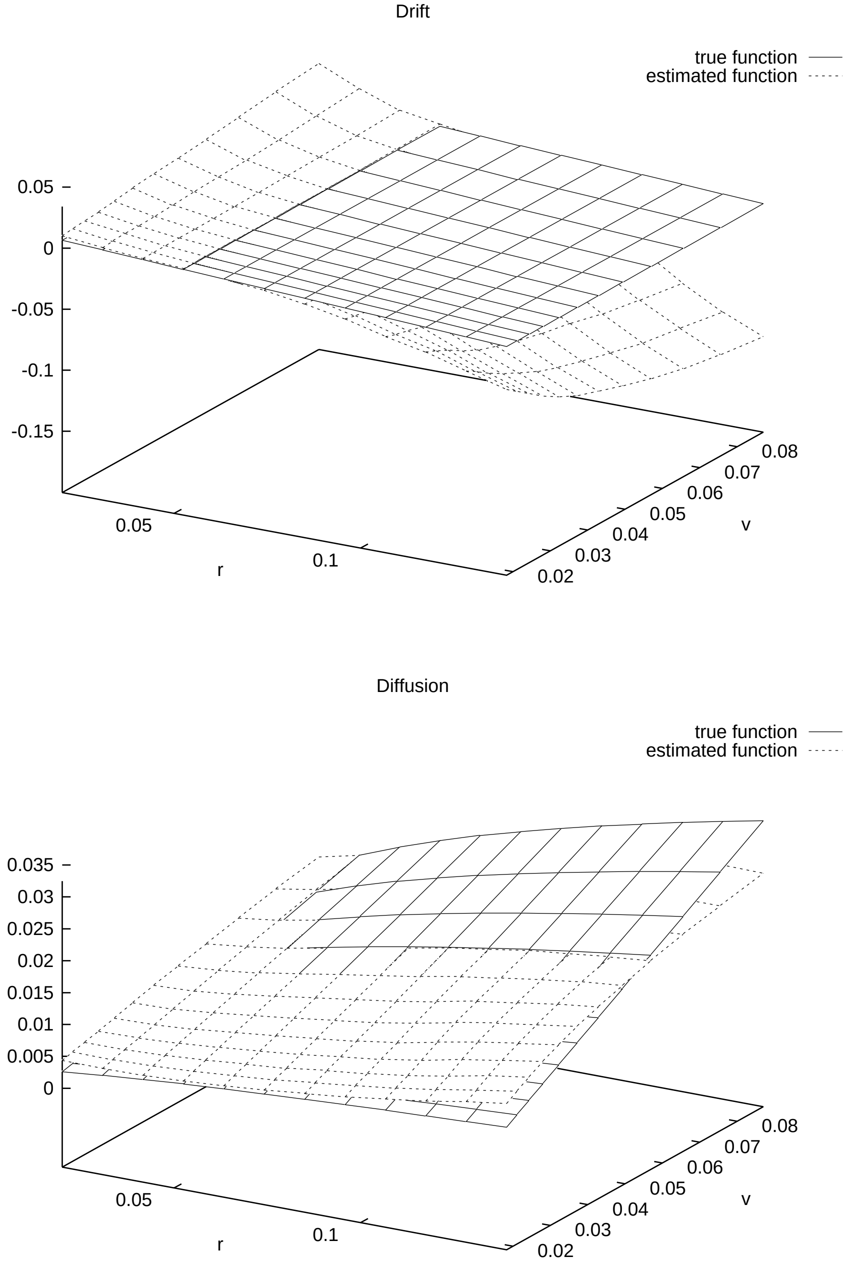


Figure 11: Estimates for Interest Rate Process, IID Bandwidths

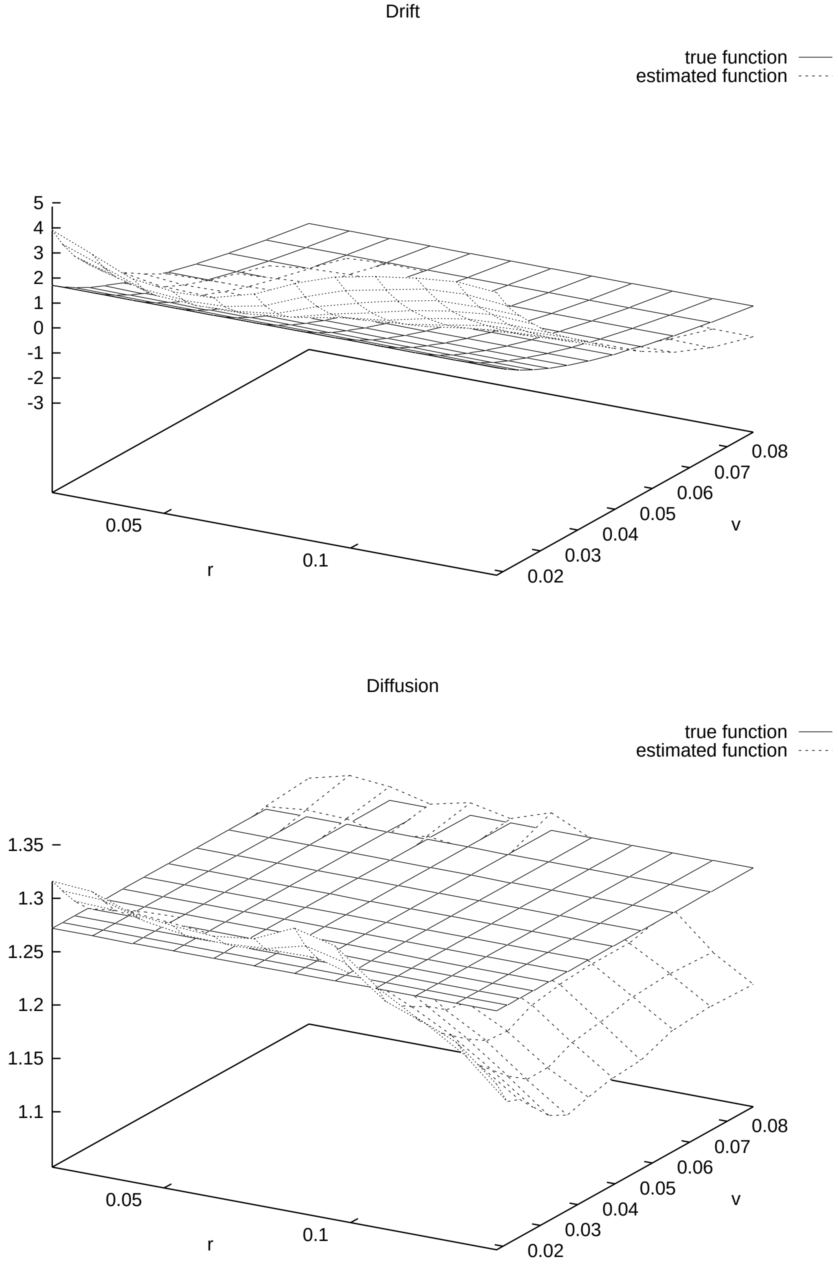


Figure 12: Estimates for Volatility Process, IID Bandwidths

The results for the volatility diffusion function are even more striking. Comparing the lower panels of figures 12 and 2, we see that the fit displayed in figure 12 is radically different from that in figure 2. From the fit in figure 12, one would be led to believe that the diffusion function exhibits very complicated nonlinearities, when in reality it is constant.

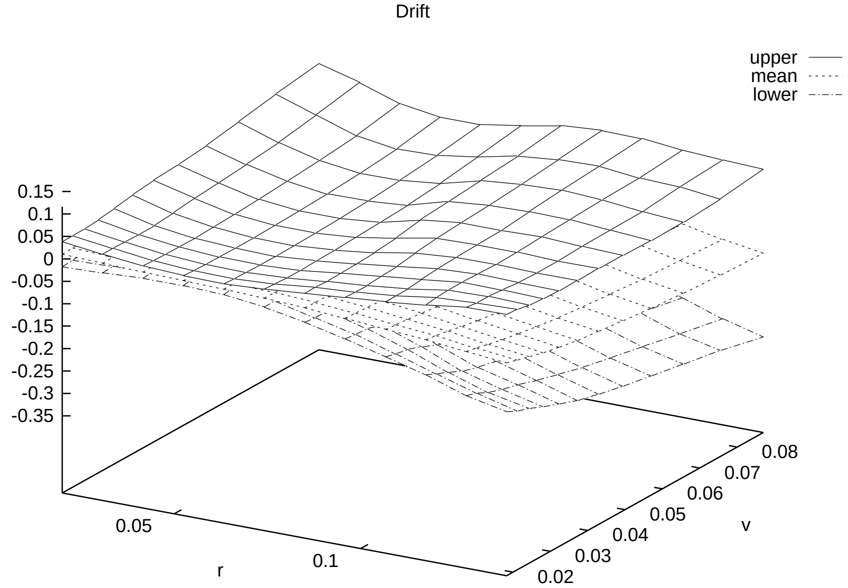
Table 5 shows the error measures for the benchmark and misspecified fits. For the interest rate drift α_r, we see that, despite the fact that the shape of the surface is poorly estimated, the MSE and IAD error measures actually decrease for the misspecified fit. The MAD measure increases. On the other hand, for α_σ and β_σ, the fit worsens according to all three error measures.

	Benchmark		
Function	MSE	IAD	MAD
α_r	0.0020932	0.030712	0.09961
β_r	0.0000035	0.001322	0.00823
α_σ	0.0018271	0.029273	0.11936
β_σ	0.0015293	0.021298	0.12280
	Misspecified		
α_r	0.0013258	0.023513	0.10912
β_r	0.0000035	0.001322	0.00823
α_σ	1.1212730	0.734976	3.14616
β_σ	0.0022579	0.032305	0.13440

Table 5: Error Measures under Misspecification

As we would expect, the inclusion of irrelevant conditioning variables also results in inefficiency. Figures 13 and 14 show the pointwise confidence intervals for the estimates of the misspecified model. In general, the confidence intervals around the estimates are much wider (note the changes of scale on the vertical axes of the plots). In particular, the estimate of the diffusion function of the volatility process is much less precise, which follows from the fact that there are two irrelevant variables.

The forgoing highlights the fact that it's a mistake to think that nonparametric estimation frees one from having to make decisions about the nature of the drift and diffusion functions of the model. One must still correctly specify the arguments to the drift and diffusion functions. The cost of incorrectly specifying the arguments to the functions is greater variance in the estimator and complicated patterns of bias. Unfortunately, theory provides



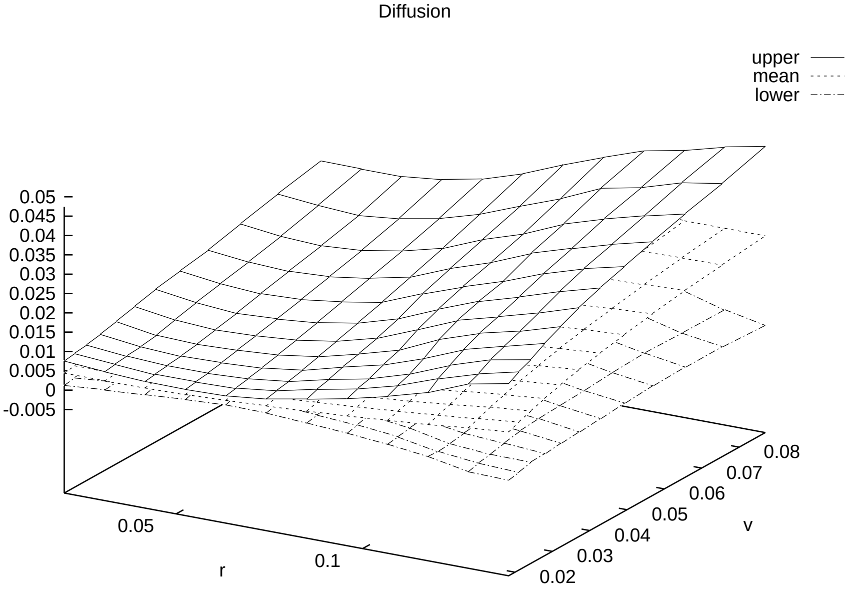


Figure 13: Confidence Intervals, Interest Rate Process, IID Bandwidths

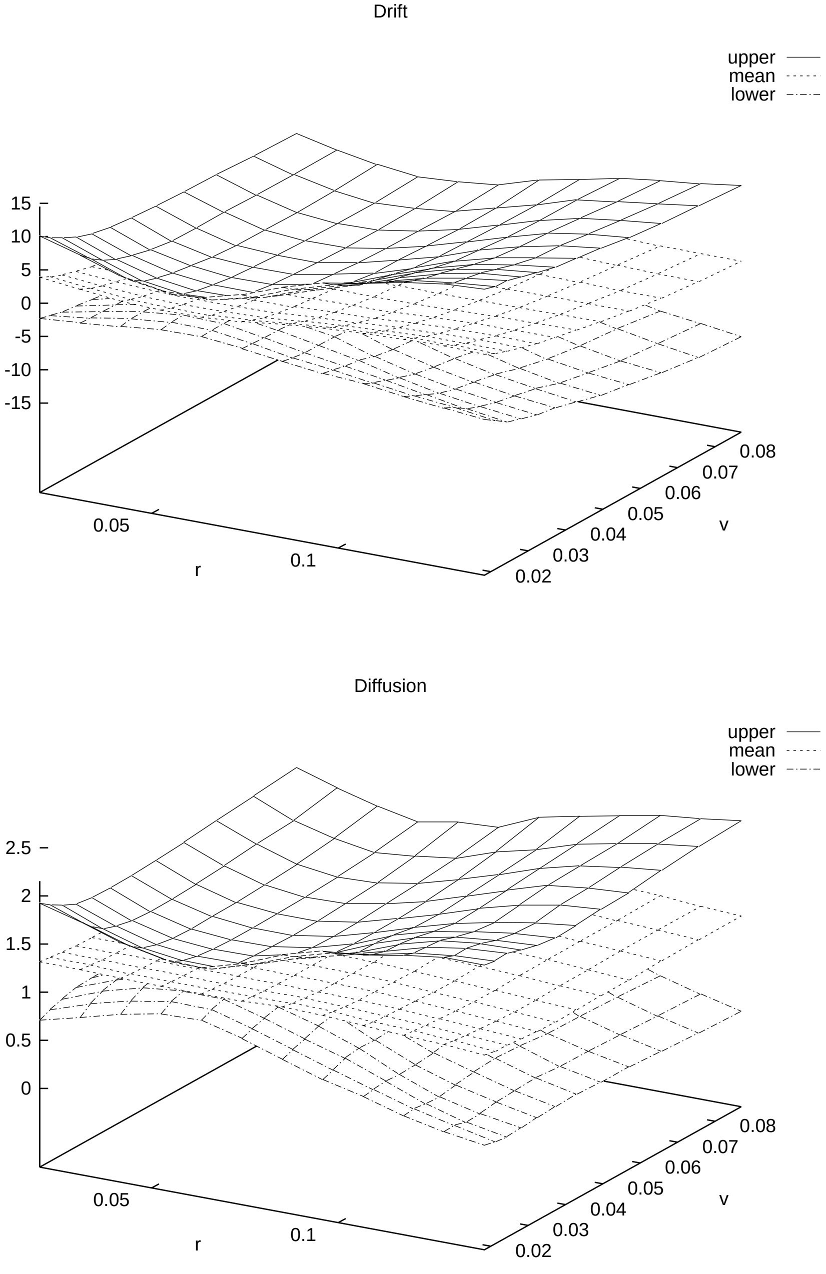


Figure 14: Confidence Intervals, Volatility Process, IID Bandwidths

no more guidance on which arguments to include in the drift and diffusion functions than it does on the forms of the functions.

4 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. Some of the preliminary results are that a locally adapted bandwidth surface appears to be an effective way to handle the problems of kernel estimation in a finite sample. I also reported some results showing how biases and inefficiencies due to model misspecification are expressed in the fit of the BRSW estimator. I found that irrelevant variables introduce bias in the estimates, in addition to reducing efficiency. Work remains to be done to understand how omitted variables affect the estimator.

As part of this research, I worked out a method for testing 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 it's 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 it's development.

Locally adapting a bandwidth surface to the data is a problem that is difficult to solve with standard techniques such as cross–validation, because it requires multi–dimensional function minimization, with an objective function that is fiercely expensive to compute, to boot. I briefly discussed a dynamic graphics interface and a parallel kernel regression algorithm that allow one to adjust the complexity of the bandwidth surface by hand. Further work is needed to explore the usefulness of this tool for applied work. In particular, work is needed to refine the error measures that can serve as a guide in the fitting process.

The analysis in this paper was conducted totally within the context of the BRSW estimator. However, 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.

A Parallel Kernel Regression

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:

$$\hat{f}(x_i, y_j) = \sum_{t=1}^{T} W(t) g(x_t, y_t; x_i, y_j), \tag{49}$$

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("ERROR: Kernel: thr_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.*

Fasano, G. and Franceschini, A.: 1987, *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.

Scott, D. W.: 1992, *Multivariate Density Estimation*, John Wiley & Sons, Inc., New York.

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.