Nonparametric estimation of diffusion process: a closer look

Orazio Di Miscia*

20 December 2004

*This research was carried out when the author was visiting the School of Economics and Management at the University of Aarhus. This research has been supported by a Marie Curie Fellowship of the European Community Programme Improving the Human Research Potential and the Socio-Economic Knowledge Base under contract number HPMT-CT-2000-00139. I would like to thank Bent Jesper Christensen and participants at the workshop "Topics in Applied Economics and Finance", University of Aarhus. Only the author is responsible for any omission and mistakes. E-mail *orazio.dimiscia1@bancaintesa.it*

Abstract

A Monte Carlo simulation is performed to investigate the finite sample properties of a nonparametric estimator, based on discretely sampled observations of continuous-time Ito diffusion process. Chapman and Pearson (2000) studies finite-sample properties of the nonparametric estimator of Aït-Sahalia (1996) and Stanton (1997) and they find that nonlinearity of the short rate drift is not a robust stylized fact but it's an artifacts of the estimation procedure. This paper examine the finite sample properties of a different nonparametric estimator within the Stanton (1997)'s framework.

1 Introduction

The purpose of this paper is to design and perform a small Monte Carlo simulation experiment to investigate the finite sample properties of a particular nonparametric estimator, based on discretely sampled observations, of a continuous-time Ito diffusion process represented by the following stochastic differential equation (SDE):

$$dX_t = \mu(X, t; \theta)dt + \sigma(X, t; \theta)dW_t \tag{1}$$

where $\{W_t, t \geq 0\}$ is a standard Brownian motion. The functions $\mu(\cdot)$ and $\sigma^2(\cdot)$ are respectively the drift (or instantaneous mean) and the diffusion (or instantaneous variance) functions of the process.

In particular, using the same philosophy of Stanton (1997) the simulation study aims to investigate the performance of a slight different nonparametric estimator for diffusion and drift functions.

The estimator is developed when only discretely sampled data of the continuous-time diffusion process are available. The continuous record of observations of the process between the sampling points is unobservable. Therefore the data generating process (DGP) in this simulation study requires that the explicit transition density functions of the diffusion processes is known, in order that the realizations of the process can be observed at discrete time along the exact continuous sampling path.

For this reason, the stochastic process is the Cox-Ingersoll-Ross squared-root process for which the DGP is well-known.

Section 2 is devoted to summarize some theoretical aspects about the estimation of the diffusion process from discretely sampled data, pointing the attention to the difficulties to estimate the drift coefficient. Section 3 explains the nonparametric estimator and the strategies to select the optimal bandwidth element. Section 4 show the Monte Carlo simulation results.

2 Estimation of the Diffusion Process from Discretely Sampled Data

Estimation of the Ito diffusion process or stochastic differential equation (SDE) has been considered in the literature for many years, with most of the papers being concerned with estimating the drift and diffusion functions from continuously sampled data. Unfortunately, in practice, more often than not we can only obtain data in discrete time since the dynamics of the process can be much faster than the sampling rate. With discretely sampled observations from the continuous sampling path, identification and estimation of the continuous-time Ito diffusion process proves to be much more complicated and difficult.

Table 1— Some specifications of the interest rate process

$\mu(r)$	$\sigma(r)$	Reference
$\beta(\alpha - r)$	σ	Vasicek (1977)
$\beta(\alpha - r)$	$\sigma r^{1/2}$	Cox, Ingersoll, and Ross (1985)
$\beta(\alpha - r)$	σr	Courtadon (1982)
$\beta(\alpha - r)$	σr^{λ}	Chan, Karolyi, Longstaff, and Sanders (1992)
$\beta(\alpha - r)$	$\sqrt{\sigma + \lambda r}$	Duffie and Kan (1996)

The first parametric estimator of the coefficients of a stationary diffusion process from discretely sampled observations is the maximum likelihood estimator proposed by Dacunha-castelle and Florence-Zmirou (1986). Other parametric estimators include the maximum likelihood estimators derived by Lo (1988) for more general jump-diffusion processes, the method of moments based on simulated sampling paths from given parameter values proposed by Duffie and Singleton (1993) and many others.

The importance of the problem is also related to the wide usage of diffusion processes in the finance literature, to model the dynamics of certain financial variables, e.g., the stock prices, the exchange rates, and the term structure of interest rates. Due to the estimation problem, however, all the diffusion models in the finance literature have to rely on parametric or semiparametric specifications for the drift and diffusion functions in order to implement available estimation methods based on discretely observed data (see table (1) for some parametric specifications related to term structure interest rate estimation).

Such specifications allow estimation of the parameters via the use of common parametric estimators, such as MLE, NLS (or OLS), or GMM. However, the discussions and empirical results show that both parametric and semiparametric specifications impose very strong and unrealistic assumptions on the underlying process of the model.

Even if the researcher takes into account these difficulties he should keep in mind that the identification and estimation of the drift function requires stronger conditions than the diffusion function. This is the so-called "*aliasing problem*" for a system of linear stochastic differential equations (SDE), as discussed in Phillips (1973) and Hansen and Sargent (1983). Phillips (1973) points out that, unless there are sufficient a priori restrictions on the parameters of a system of linear stochastic differential equations, we cannot distinguish between structures generating cycles whose frequencies differ by integer multiples of the reciprocal of the observation period. Similarly, it is impossible to identify a nonlinear diffusion process without imposing any structural restrictions on the model. Especially, the drift term of the diffusion process (univariate or multivariate) cannot be directly identified on a fixed time interval, no matter how frequently the observations are sampled, as the Cameron-Martin-Girsanov transformation (see Oksendal (1995)) can always be applied to give an otherwise unnoticeable change in the drift.

Moreover to highlight again the difficulties about drift estimation, as some authors have already observed (see Merton (1980)), even though the diffusion term of a stochastic process can be estimated very precisely when the sampling interval is small, the estimates of the drift term tend to have low precision. The parametric estimates of the drift function specification can perform very poorly even with large samples of data, no matter how frequently the observations are sampled over a short sampling period. The following example can help to illustrate the problem. Suppose that the short term interest rate follows a Brownian motion with drift process:

$$dX_t = \mu dt + \sigma dW_t$$

where μ and σ are constants. The maximum likelihood estimator of μ from equispaced discretely sampled observations $\{X_{t_{1=0}}, X_{t_2}, \ldots, X_{t_{n=T}}\}$ is the average of the log difference of short rate:

$$\hat{\mu} = (1/T) \sum_{i=1}^{n} \log \left(\frac{X_{t_i}}{X_{t_{i-1}}} \right) \qquad \hat{\mu} = (\log X_T - \log X_0)/T$$

It's easy to verify that $\hat{\mu}$ is a consistent estimator of μ as $\mu|X \sim N(\mu, \sigma^2/T)$. However, it is also very easy to see that, for any finite sample of observations, $\hat{\mu}$ is very sensitive to the first and last observations of the sample and is actually determined only by these two values. Thus, if we have a sample of, say, 5,000 observations, it is only the first and last observations that matter for the estimate of μ. Moreover, $\hat{\mu}$ has no efficiency gains even if we increase the sample size by reducing the sampling interval over fixed T.

Thus, it is not hard to see that the estimate of μ will not be robust in that it will be very sensitive to the sampling path and/or the discrete observations along the sampling path.

However, these difficulties has no worried the researcher since that greatest attention was devoted to the diffusion function estimation.

One reason is that the diffusion function, as the second moment and the measurement of instantaneous volatility of the stochastic process, is of more interest in modelling the movements of interest rates, asset prices, or exchange rates. For instance, the volatility of the riskless interest rate is one of the key determinants of the value of contingent claims and one of the key factors determining optimal portfolio hedging strategies for risk-averse investors. Therefore, to predict the movements of derivative security prices, to hedge an investment portfolio, or to create a certain leverage within a portfolio, the volatility of the prices of underlying assets is the major factor to be considered. Another and maybe more direct reason is that, in the famous Black-Scholes option pricing formula, the prices of derivative securities are affected by the price of underlying assets only through its instantaneous volatility, i.e. the diffusion function. The drift function does not appear in the option pricing formula at all due to an assumption that, in the economy, there exists a risk-free asset with non-stochastic rate of return.However as Lo and Wang (1995) point out, predictability of an asset's return is typically induced by the drift and will affect the prices of options on that asset, even though the drift term does not enter the option pricing formula. Moreover, in models with stochastic spot interest rates, both the diffusion function and drift function will enter the derivative security pricing formulation. Therefore the prices of derivative securities in these cases are explicitly affected by the price of underlying assets through not only the diffusion function but also drift function. From this point of view, the drift function estimation is as important as the diffusion function estimation.

However, to avoid some of the problems due to the parametric specification of the diffusion process, a nonparametric approach has been recently used.

The first nonparametric diffusion function estimator is proposed by Florens-Zmirou (1993) which imposes no restriction on either the drift term or diffusion term, but the procedure leaves the drift term unidentified and the diffusion function estimator can not be used for the construction of the drift function estimator. Aït-Sahalia (1996) proposes a nonparametric diffusion function estimator based on the linear mean-reverting drift function for the strictly stationary diffusion processes. Stanton (1997) develops approximations to the true drift and diffusion functions and estimates these approximations nonparametrically.

3 The nonparametric estimator

Let $\{r_t; t \geq 0\}$ be defined as the unique, time homogeneous Markov process that solves a stochastic differential equation (SDE) of the form:

$$dr_t = \mu(r_t) + \sigma(r_t) dB_t$$

where r_t is the "instantaneous" or "short" rate, $\{B_t; t \geq 0\}$ is a scalar Brownian motion, $\mu : \mathbb{R} \rightarrow \mathbb{R}$ is the drift function and $\sigma : \mathbb{R} \rightarrow \mathbb{R}_+$. Following Stanton (1997), drift and diffusion function are approximate using the following equation (first order approximation):

$$\begin{aligned} \mu(r_t) &= \frac{1}{\Delta}\mathrm{E}_t[r_{t+\Delta} - r_t] + O(\Delta) \\ \sigma(r_t) &= \sqrt{\frac{1}{\Delta}\mathrm{Var}_t(r_{t+\Delta} + O(\Delta)} \end{aligned}$$

where $\mathrm{E}_t[\cdot]$ and $\mathrm{Var}_t(\cdot)$ are the first and second conditional moments.

The essence of Stanton's approach is to apply the Nadaraya-Watson (see Nadaraya (1964) and Watson (1964)) kernel regression estimator to construct nonparametric estimates of the conditional moments above.

The nonparametric estimator for the conditional moments used in this paper is the local linear regression (see Loader (1999) and Härdle (1990) for an extensive treatment). In this framework the analysis start with the general form of the regression function:

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the response variable related to the sample predictor variables $\mathbf{X} = (\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_n)$ with $\mathbf{x}_i \in \mathbb{R}^d$ and ϵ is an error term. The goal is to estimate $f(\mathbf{X})$ since that:

$$f(\mathbf{x}) = \mathrm{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$$

the conditional moment used to approximate drift and diffusion function of diffusion process.

Taylor's theorem says that under mild condition is possible to approximate $f(\cdot)$ "locally" around a point, $\mathbf{x}_i$ by a polynomial:

$$f(\mathbf{u}) \approx a_0 + a_1(\mathbf{u} - \mathbf{x}_i) + \frac{1}{2}a_2(\mathbf{u} - \mathbf{x}_i)^2 + O(\Delta^n)$$

with $n = 2, \ldots$, $\mathbf{u} \in [(\mathbf{x}_i - \mathbf{h}), (\mathbf{x}_i + \mathbf{h})]$ and $\mathbf{h}$ the *width* of the interval of $\mathbf{x}_i$.

A compact vector notation for polynomials is:

$$a_0 + a_1(\mathbf{u} - \mathbf{x}_i) + \frac{1}{2}a_2(\mathbf{u} - \mathbf{x}_i)^2 = \langle \mathbf{a}, \mathbf{A}(\mathbf{u} - \mathbf{x}_i) \rangle$$

where $\langle \cdot \rangle$ is the inner product in the euclidean space and

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \qquad \mathbf{A}(v) \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix}$$

The coefficient vector, $\mathbf{a}$, must be estimated for each point by minimizing the locally weighted sum of squares. To implement the estimator is necessary to fix a set of N grid points, $\{\mathbf{z}_i\}_{i=1}^N$ defining an equally spaced partition of a subset of the domain of $f(\cdot)$. The procedure minimize the sum of square distance between each grid point, $\{\mathbf{z}_i\}$ and the whole sample of predictor variables. The sum of squares is weighted in order to impose less weight to the sample points far from the $\{\mathbf{z}_i\}$.

It's possible to restate the problem in matrix form. Rewrite the matrix $\mathbf{X}$ in order to consider the polynomial approximation of $f(\mathbf{X})$:

$$\mathbf{X} = \begin{pmatrix} \mathbf{A}(\mathbf{x}_1 - \mathbf{z}_i)' \\ \mathbf{A}(\mathbf{x}_2 - \mathbf{z}_i)' \\ \cdots \\ \mathbf{A}(\mathbf{x}_n - \mathbf{z}_i)' \end{pmatrix}$$

Let consider the diagonal matrix $\mathbf{W}$ with the j-th entries $w_j(\mathbf{z}_i) = W(\frac{\mathbf{x}_j - \mathbf{z}_i}{\mathbf{h}})$. The weighted sum of squares in order to estimate the coefficient vector $\mathbf{a}$ can be written in matrix form:

$$\arg\min_{\mathbf{a}} (\mathbf{Y} - \mathbf{X}\mathbf{a})' \mathbf{W} (\mathbf{Y} - \mathbf{X}\mathbf{a})$$

If $\mathbf{WX}$ has full column rank, least squares theory give the explicit expression:

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

Considering $\mathbf{z}_i = \mathbf{x}_i$ then the local regression estimation is:

$$\hat{\mathbf{a}} = e_1^{'}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

When matrix $\mathbf{A}$ has only one column it's possible to show that $\hat{\mathbf{a}}$ is the Nadaraya-Watson estimator. This paper consider also the second column of matrix $\mathbf{A}$ that permit to estimate the local linear regression.

The choice of the local linear regression is due to the well-known "boundary bias" problem of the Nadaraya-Watson estimator and the superior performance of local linear regression in this case. Chapman and Pearson (2000) consider a boundary correction of the Nadaraya-Watson estimator but the local linear regressor is simpler to estimate, due to the closed form and give the same solution to the boundary-bias mentioned above.

3.1 The optimal choice of the bandwidth parameter

In a kernel regression framework, the "parameters" to estimate are the kernel function and the bandwidth (or smoothing) element. Empirical researches confirm that the optimal kernel function does not improve significantly the "fitting" but the bandwidth estimation is crucial (see Härdle (1990) and Loader (1999) for details).

There is no single accepted approach to its selection and sometimes the better choice depend on a visual inspection of the results obtained.

In this paper a data-dependent bandwidth selectors are used according to different aspects to take into account[1]. For example, one element is the selection of a bandwidth parameter when the kernel regression residuals are correlated with the regressors. This is particular relevant when the researcher analyze a time-series with short/long-rate dependency, a well-known empirical regularity showed by almost all high frequency financial time series.

Opsomer, Wang, and Yang (2001) show that the presence of correlation, if ignored, causes the break down of the commonly used automatic tuning bandwidth selection. The bandwidth tend to be too small compared with the theoretical optimal bandwidth. Roughly speaking, the nonparametric regression "interpolate" instead of "fit" the data.

A common selection method is to use a version of the rule called cross-validation, the "leave-k-out cross-validation" (hereafter LMO-CV where M is for "many"). The first step in this procedure is to estimate the regression function for a set of bandwidths $\{h_1, \ldots, h_2\}$. The LMO-CV bandwidth is the solution of the following problem:

$$\arg\min_{h_i} n^{-1} \sum_{t=1}^{n} (y_t - \hat{f}_i(x_t; h_i))^2 \tag{3}$$

where $\hat{f}_i(x_t; h_i)$ is the "leave-k-out" estimator[2]. In particular, the conditional moments at each data point is estimated using the entire sample, except for the actual data point and its nearest neighbors[3].

[1]The analysis of the statistical properties of different bandwidth selectors is not pursued in this paper

[2]To be precise, it's possible to rewrite the bandwidth element as: $h_i = c * \hat{\sigma} n^{-5}$ and the algorithmic find the optimal $\hat{c}$ (see Wand and Jones (1995))

[3]Due to serial dependence of the simulated data, the cross-validation is performed omitting 100 observations, i.e., four months is either direction of the particular data point in question. Even if the number of omitted observations should be calculated or "calibrated" in some way, this is what other authors do with interest rate data (see Boudoukh, Richardson, Stanton, and WhiteLaw (1999) among others)

An other selector method is the Partial Cross-Validation (PCV). It is performed using eq. (3), without "leave-k-out" estimator. The strategy is to partition the sample into N sub-samples and for each of j-th sub sample find the minimum $h_{i,j}$. The optimal bandwidth is the average of the single $h_{i,j}$ (for detatils about LMO-CV and PCV see Loader (1999) and Plutowski (2000)) .

In the paper other bandwidth choices have been used. The IID bandwidth is optimal for IID data and it is defined as $h^{IID} = \hat{\sigma} n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation of the data and n is the sample size. Following what is said above, it seems a poor choice but is useful to compare the other bandwidths. The "Stanton bandwidth" is also used: it is the same bandwidth used in Stanton (1997). In private communications with Chapman and Pearson he reported that the formula is the result of a heuristic approach and it is equal to $h^{Stanton} = 4\hat{\sigma} n^{-1/5}$ (see Chapman and Pearson (2000)).

The Monte Carlo simulations terminate analyzing the behaviour of other three bandwidth parameter values: $h^{(6)} = 6\hat{\sigma} n^{-1/5}$, $h^{(8)} = 8\hat{\sigma} n^{-1/5}$, $h^{(10)} = 10\hat{\sigma} n^{-1/5}$, $h^{(14)} = 14\hat{\sigma} n^{-1/5}$. According to what has been said above, increasing levels of the bandwidth give the possibility to take into account the serial dependence that will be introduced in the simulation of the square-root process with an ad-hoc choice of parameter set. Obviously this is an heuristic way to evaluate the effect of the serial dependence on the bandwidth choice, but on the other hand there is no single and accepted method to do this[4].

4 Monte Carlo Study

4.1 Simulating a square-root process

In order to evaluate the performance of the Stanton (1997)'s estimator, this paper consider the square-root process introduced in the term structure literature by Cox, Ingersoll, and Ross (1985) (CIR model):

$$dX_t = \kappa(\theta - x_t)dt + \sigma\sqrt{X_t}dB_t$$

where θ defines the long-run mean of X_t, κ determines the speed at which the process returns to the long-run mean, and σ is the instantaneous variance of the process. By construction, the drift of this process is linear. The aim of the Monte Carlo study is to examine the finite sample properties of

[4]The last bandwidth use a scaling factor equal to 14. It is higher than the level used by other authors with data-driven bandwidth selectors applied to interest rate data (see Boudoukh, Richardson, Stanton, and WhiteLaw (1999)).

the nonparametric diffusion function and drift function estimators. For a set of parameter values, a simulated sample path is generated and based on this sample the nonparametric estimators are applied. The simulated sample paths are constructed assuming that the length of time between observations of diffusion is $\Delta = 1/250$ corresponding to daily observations. Each path simulate a daily time series of 7500 observations that correspond to an hypothetical data set starting in 1973. The choice about the length is due to the consideration that there are no longer and reliable data sets for proxies of the short rate.

Practical implementation of the stochastic process simulation require the specification of the data generating process (DGP) that is given by its transition probability density function and based on its eventual Markovian property. In the case of CIR process the transition density is a non-central chi squared). In all simulation, the first 1000 observations are discharged to eliminate any start-up effects.

The unconditional moments of the square-root process are:

$$\mathrm{E}(X) = \theta$$

$$\mathrm{Var}(X) = \frac{\theta\sigma^2}{2\kappa}$$

and

$$\mathrm{Corr}(X_{t+\Delta}, X_t) = e^{-\kappa\Delta}$$

The choice of κ, the parameter that determines the persistence of the process, is particularly important. It is fixed at the value of 0.21459 that implies a (monthly) autocorrelation of the short rate of 0.982, which is consistent with U.S interest rate (see also Chapman and Pearson (2000)). θ is not particular important and it is fixed to 0.085711 that correspond to the level estimated in other papers (see Chapman and Pearson (2000)). Given the value of θ and κ, σ is fixed in order to set the unconditional variance equal to the sample variance in the Stanton (1997) data set ($\sigma = 0.138213$)

4.2 The Results

The results from applying the Stanton's approximation to estimate the drift and diffusion coefficients, using the local linear regression to estimate the conditional moment, are shown in figure 1 through 4.

The figures have the following structure: figure 1 and 2 report the drift coefficient estimates using 8 different bandwidths, while figure 3 and 4 are relative to diffusion coefficient estimates.

In each pane, the solid line is the theoretical drift/diffusion coefficient. The dash-dotted and two dotted lines are respectively the pointwise average and the 2.5 th and 97.5 th percentile points at each grid point[5].

Figure 1 and 2 show that, in all cases, the estimates exhibit some spurious nonlinearities.

The PCV bandwidth delivers the best performance, but the average drift function diverges from the linear drift for value of the short rate in excess of 0.15 (15%). The IID bandwidth presents the most spectacular evidence of spurious nonlinearity in the drift, Clearly h^{IID} is a poor choice. This is the proof that serial dependency can create serious problem for bandwidth selection.

If drift estimation is the sole objective of the analysis, the prediction from these simulations is clear: choose a large bandwidth that oversmooths the kernel regression function (see figure 2). Of course, this is valid only if the true drift is linear, because a large bandwidth obliterate all of the detail in the estimated function. Even if the true drift had been nonlinear, the oversmoothed estimator would have suggested linearity.

Figure 3 and 4 show the results for the diffusion function. Generally speaking, they are the opposite of the drift case.

Using a PCV bandwidth is possible to obtain an estimate that mimic the behaviour of theoretical diffusion coefficient. Also the IID estimator provides an accurate estimate of the diffusion function. The nonparametric approach seems particular indicated to estimate the instantaneous volatility of the square-root process.

To conclude, the spurious nonlinearities exhibited by the kernel regression at high and low level of the data range, are not related to the boundary-bias problem that is eliminated using the local linear regression. Chapman and Pearson (2000) give a possible explanation of these nonlinearities: the truncation of the upper limit of any finite sample. Roughly speaking, if we do not consider a theoretical infinite sample, the realizations of the process cannot have a value greater than the highest value of the sample and this generate the finite sample bias.

5 Conclusion

Robinson (1983) and Robinson (1986) establishes the pointwise consistency of kernel regression estimators in a time series context. Therefore, a possible explanation to spurious nonlinearities showed by drift coefficient estimates could be the finite sample performance of the estimator.

[5] In the analysis the grid points are 32

However, it is not possible to reject nonlinearities in the data: these technique seems ill equipped to provide reliable information about the properties of any time series in the extreme tails of the distribution (see Chapman and Pearson (2000)). This means that the statistical tools are inadequate to investigate the nonlinearity of the drift but this is not a proof for its linearity.

On the other hand, the local linear regression estimate the diffusion coefficients in a good way and this is one of the advantage to use nonparametric methods to estimate diffusion processes.

The conclusion is that, with the statistical tools adopted in this paper, is not possible to obtain reliable conclusion about drift coefficient that seems hard to estimate (see Merton (1980)). Even if recent researches have used nonparametric techniques to reduce the number of arbitrary parametric restrictions imposed on the underlying process, there are some drawbacks, in finite sample, that must be considered and eliminated.

Next step will be to consider a bandwidth selectors with a superior performance when the data show short/long-rate dependency: in my opinion it could be a great improvement.

References

AÏT-SAHALIA, Y. (1996): "Nonparametric Pricing of Interest Rate Derivatives Securities," *Econometrica*, 64, 527–560.

BOUDOUKH, J., M. RICHARDSON, R. STANTON, AND R. WHITELAW (1999): "A Multifactor, Nonlinear, Continuous-Time model of Interest Rate Volatily," Working paper, Stern School of Business.

CHAN, K., F. KAROLYI, F. LONGSTAFF, AND A. SANDERS (1992): "An Empirical Comparison of Alternative Models of the Short–Term Interest Rate," *Journal of Finance*, 47, 1209–1227.

CHAPMAN, D., AND N. PEARSON (2000): "Is the Short Rate Drift Actually Nonlinear?," *Journal of Finance*, 55, 355–388.

COURTADON, G. (1982): "The Pricing of Options on Default-Free Bonds," *Journal of Financial and Quantitative Analysis*, 17, 75–100.

COX, J., J. INGERSOLL, AND S. ROSS (1985): "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–408.

DACUNHA-CASTELLE, D., AND D. FLORENCE-ZMIROU (1986): "Estimation of the coefficient of a diffusion from discrete observations," *Stochastics*, 19, 263–284.

Duffie, D., and R. Kan (1996): "A Yield Factor Model of Interest Rates," *Mathematical Finance*, 6, 379–406.

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

Florens-Zmirou, D. (1993): "On Estimating the diffusion coefficient from Discrete Observations," *Journal of Applied Probability*, 30, 790–804.

Hansen, L., and T. Sargent (1983): "The dimensionality of the aliasing problem," *Econometrica*, 51, 377–388.

Härdle, W. (1990): *Applied nonparametric regression.* Cambridge University Press, Cambridge.

Lo, A. W. (1988): "Maximum likelihood estimation of Generalized Ito Processes with discretely-sampled data," *Econometric Theory*, 4, 231–247.

Lo, A. W., and J. Wang (1995): "Implementing option pricing models when asset returns are predictable," *Journal of Finance*, L,1, 87–129.

Loader, C. (1999): *Local Regression and likelihood.* Springer, New York.

Merton, R. (1980): "On estimating the expected return on the market: an exploratory investigation," *Journal of Financial Economics*, 8, 323–361.

Nadaraya, E. (1964): "On estimating regression," *Theory of Probability and its Applications*, 9, 141–142.

Oksendal, B. (1995): *Stochastic Differential Equations.* Springer, Berlin.

Opsomer, J., Y. Wang, and Y. Yang (2001): "Nonparametric Regression with Correlated Errors," *Statistical Science*, 16, 134–153.

Phillips, P. (1973): "The problem of identification in finite parameter continuous-time model," *Journal of Econometrics*, 1, 351–362.

Plutowski, M. (2000): "Survey: Cross-validation in theory and practice," Working paper, Department of Computational Science Research, David Sarnoff Research Center, Princeton.

Robinson, P. M. (1983): "Nonparametric estimators for time series," *Journal of Time Series Analysis*, 4, 185–207.

——— (1986): "On the consistency and finite-sample properties of non-parametric kernel time series regression, autoregression and density estimators," *Annals of the Institute of Statistics and Mathematics*, 38, 539–549.

STANTON, R. (1997): "A Non parametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *Journal of Finance*, 52, 1973–2002.

VASICEK, O. (1977): "An Equilibrium Characterization of the Term Structure," *J. of Financial Economics*, 5, 177–188.

WAND, M., AND M. JONES (1995): *Kernel Smoothing*. Chapman and Hall.

WATSON, G. (1964): "Smooth regression analysis," *Sankhya Series*, 26, 359–372.

Figure 1— Drift function using local linear regression

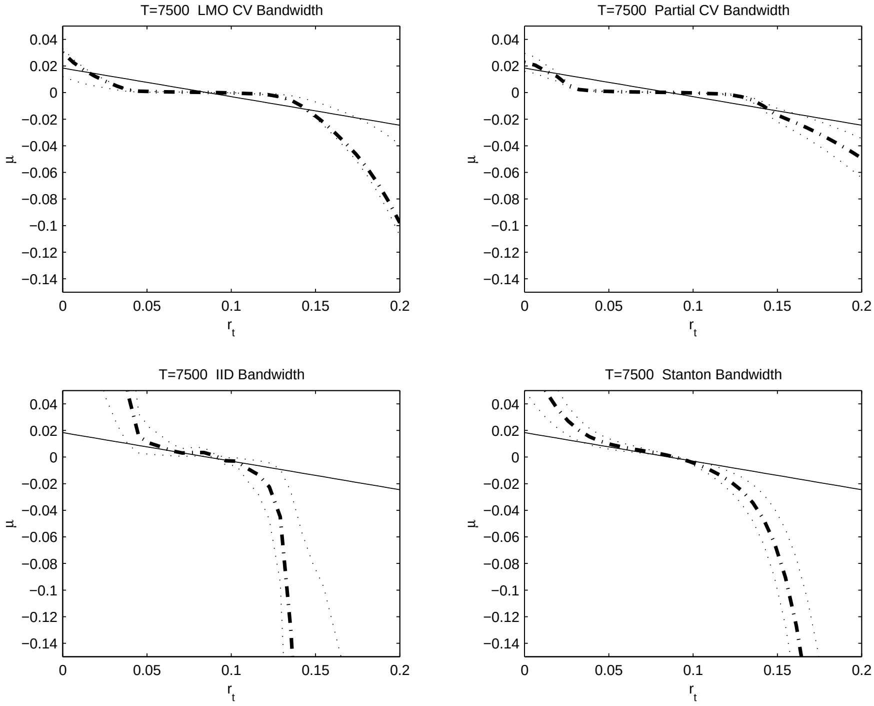

Figure 2— Drift function using local linear regression

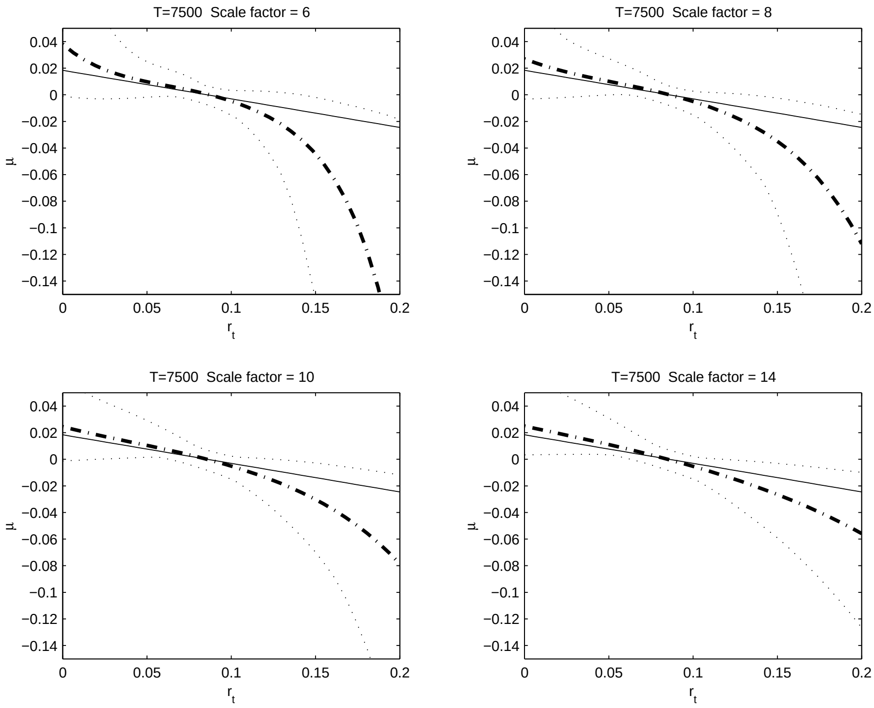

Figure 3— Diffusion function using local linear regression

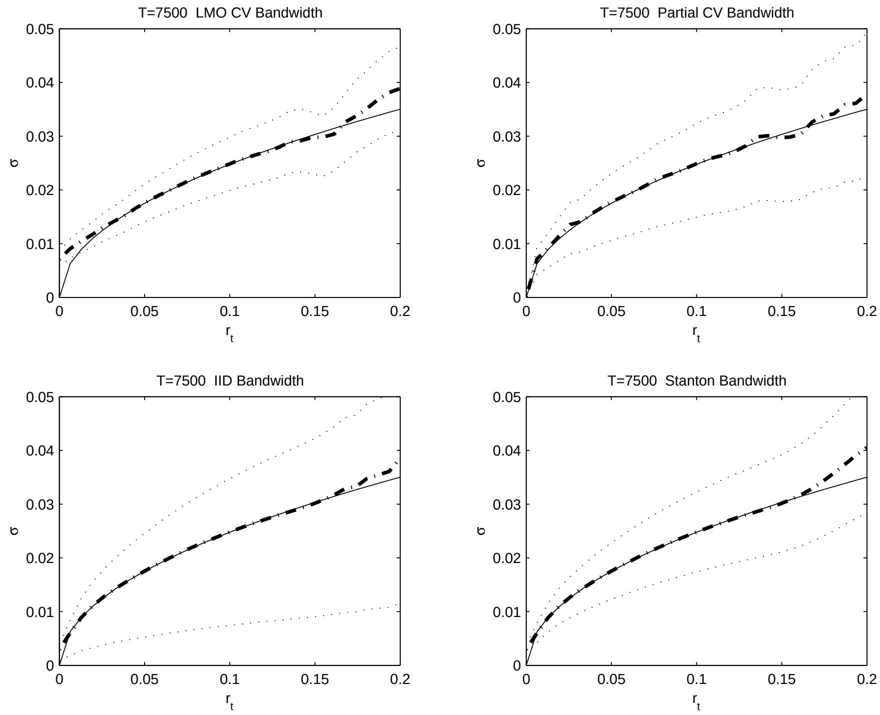

Figure 4— Diffusion function using local linear regression