Bayesian Non-linear Modellings of the Short Term US Interest Rate: *the help of non-parametric tools* *

Michel Lubrano[1]

July 1999, revised May 2000

Abstract

This paper is concerned with the empirical investigation of models of the US short term interest rate, using a mixture of classical non-parametric methods and of Bayesian parametric methods. The shape of the drift and volatility functions of the usual diffusion equation are first investigated using a preliminary non-parametric analysis. The paper then develops a Bayesian method for comparing models which is based on the ability of a model to minimise the Hellinger distance between the posterior predictive density and the density of the observed sample. A discretisation of the usual diffusion equation is estimated with different parameterisations which range from variants of the constant elasticity of variance model to various switching models which draw their justifications from the preliminary non-parametric analysis. The paper concludes by some implications for the term structure. It appears that a model good at reproducing the data density is not necessarily the best for simulating the yield curve.

JEL Classification: C11, C14, C22, C52, E43
Keywords: Bayesian econometrics, time series, non- parametric analysis, model evaluation, non-linear modelling, interest rates, term structure.

*Many discussions and critics with Luc Bauwens, Laurence Broze, Christian Hafner, Frédéric Jouneau, Jean-François Richard and Peter Schotman helped a lot in the writing of the paper. Usual disclaimers apply.
[1]GREQAM-CNRS, 2 rue de la Charité, 13002 Marseille, France and CORE, 34 voie du Roman Pays, B-1348 Louvain la Neuve, Belgique
email: lubrano@ehess.cnrs-mrs.fr

1 Introduction

In a bond pricing model, the short term interest rate r is usually modelled by a diffusion equation having the general form

$$dr = \mu(r)dt + \sigma(r)\, dW, \tag{1}$$

where W is the standard Brownian motion. The empirical models reviewed in Chan, Karolyi, Longstaff and Sanders (1992) (CKLS) are parameterised discretisations of this continuous time equation. They share the common characteristics that their drift functions $\mu(r)$ are linear functions of r. The different expressions reviewed for the volatility function $\sigma(r)$ can be constant [Vasicek (1977)], can depend on the square root of r [Cox, Ingersoll and Ross (1985)] or finally can be a constant elasticity function with $\sigma(r) = \sigma r^{\gamma}$. CKLS found that models allowing $\gamma > 1$ capture the short term dynamics of the interest rate better than those which require $\gamma < 1$. However, Broze, Scaillet and Zakoian (1995) have shown that this type of parameterised discretised model is ergodic and second order stationary if and only if $\gamma \leq 1$. Consequently, there is an indication of bad specification for the estimated models where γ appears to be greater than one.

In the presence of fiduciary money, a model for the behaviour of the short term interest rate should be stationary and produce positive and finite interest rates. The model should have a property of mean reversion and thus should produce interest rates that return to their equilibrium level following positive or negative shocks. If short term interest rates are modelled for instance as a random walk, as would be favoured by statistical tests, this leads to incoherencies for the term structure. See for instance Ingersoll (1987, chap 18, p. 396).

In finite samples, interest rates usually do behave as a random walk and the unit root hypothesis is hard to reject from a statistical point of view. The property of mean reversion depends on very few observations and is thus difficult to establish empirically. Ait-Sahalia (1996a) claims that the linearity of the drift function is the main source of misspecification for existing models. He found that the drift is essentially zero as long as the rate has a mid range value and that a non-linear mechanism pulls it back to this region whenever it escapes. Instead of a linear drift, Ait-Sahalia (1996a) suggests a non-linear function of the form $\mu(r) = \alpha_0 + \alpha_1 r + \alpha_2 r^2 + \alpha_3/r$. A negative α_2 drives back high rates to the middle region while the same task is done by a positive α_3 for small rates which come too near to the zero boudary. This non-linear drift with properly signed coefficients guaranties ergodicity and second order stationarity of the parameterised continuous time model even when $\gamma > 1$.

The problem of finding a correct parametric formulation for the drift and volatility functions in a discretised diffusion equation is endless when there is no appropriate tool for guiding the search. A parametric model in a discrete time framework is certainly very attractive for its parsimony, ease of estimation, capability of prediction and the hope to obtain a not too complicated pricing formula for the derivative security. In this paper, we have decided to guide the choice for a particular parametric formulation by using a classical non-parametric preliminary analysis of the data which will be useful to judge between competing models. When a particular parametric formulation has been chosen, one has to check how the selected model performs, its ability to reproduce the characteristics of the data set. A particular approach is developed in this paper for comparing models. The procedure for comparing models is based on comparing the predictive posterior density of the model to the non-parametric estimate of the data density.

Apart from the preliminary non-parametric analysis, the statistical approach of this paper is Bayesian. This framework has been successfully applied in Lubrano (1998) for stock indexes, by Pfann, Schotman and Tschernig (1996) and Pai and Pedersen (1998) for interest rates for instance. Whenever a non-linear model is considered, especially with a threshold, the Bayesian approach is more secure because it is better to average that to try to maximise a likelihood function with possibly many maxima. See Lubrano (2000) for a review. A better understanding of the likelihood is also provided by the Bayesian approach.

The paper is organised as follows. In section 2, we review the main difficulties encoutered for non-parametric inference in diffusion equations and explain how bayesian inference can be conducted for the discretised model. In section 3, we develop a Bayesian strategy to compare parametric models. Section 4 presents the data set, and give a non-parametric estimation for the drift and the volatility function. Section 5 is devoted to the Bayesian analysis of various non-linear models of the interest rate while section 6 analyses the impact of a two-regime modellisation. Section 7 presents some implications for the term structure while section 8 concludes.

2 Parametric and non-parametric inference for diffusion equations

2.1 Introduction

Inference for diffusion equations is not a simple matter, even when $\mu(r)$ and $\sigma(r)$ are fully parameterised. The main difficulty comes from the unavailability of a continuous record of observations. When the observations are discrete, the maximum likelihood estimator can be implemented when there exists an analytic expression for the transition probability or the marginal density of the observations. This is the case for very few processes such as the geometric Brownian motion, the Ornstein-Uhlenbeck process, the Cox-Ingersoll-Ross model. For these models the sampling path can be reconstructed and there exists an exact discretisation. The likelihood function can be written explicitly, even if its evaluation may be difficult as for instance in the Cox-Ingersoll-Ross model. When an exact discretisation does not exist, the Euler discretisation can be used instead as an approximation. For (1) we have:

$$r_n - r_{n-1} = \mu(r_{n-1})\Delta_n + \sigma(r_{n-1})\sqrt{\Delta_n}\epsilon_n, \tag{2}$$

where ϵ_n is a Gaussian white noise with unit variance, Δ_n is the time unit between two observations (discretisation step) and n the index (not the time index) of the observations. If the process is observed in the time interval $[0, T]$, then $\Delta_n = T/n$. When $\Delta_n \to 0$, the discretised process tends to the continuous time process. When Δ_n is strictly positive, there is in general a discretisation bias because the dynamics of the process can be much faster than the sampling rate. CKLS have used GMM in order to try to reduce the discretisation bias, but Brenner, Harjes and Kroner (1996) noticed that GMM are not well behaved when $\gamma > 1$ as then the discretised process is not stationary in this case. The discretisation bias can be reduced if we simulate the unobserved part of the process and then apply indirect inference as in Broze, Scaillet and Zakoian (1995).

In this paper, every time we consider a parametric model, we shall not try to correct for the discretisation bias following in this way a number of authors. For instance, all the authors considering the possibility of GARCH errors neglect this type of correction: Brenner, Harjes and Kroner (1996) as already mentioned, Longstaff and Schwarz (1992), Gray (1996), Koedijk , Nissen, Schotman and Wolf (1997). Papers with a Bayesian approach also neglect the discretisation bias, see the references given above. We shall consequently assume that the discretised model with $\Delta_n = 1$ is the correct model. Bayesian inference for continuous time processes is left for future research. Non-parametric inference for continuous time diffusion processes is detailed in the appendix.

2.2 Non-parametric inference for the discretised model

Once the model is discretised, a non-parametric approach for making inference in (2) is in a way straightforward if we follow the lines recalled in the survey of Härdle, Lütkepohl and Chen (1997) which reviews non-parametric inference for time series models. Let us consider $\Delta_n = 1$ and note the discretised model as

$$\Delta r_t = m(r_{t-1}) + s(r_{t-1})\epsilon_t \tag{3}$$

which means that we consider a regression model where $\Delta r_t = r_t - r_{t-1}$ is the endogenous variable and r_{t-1} the predetermined variable. The conditional expectation function is $m(r_{t-1})$ and the skedastic function $s(r_{t-1})$. As shown in Auestadt and Tjostheim (1990), Tjostheim and Auestadt (1994a,b), the conditional expectation function $m(r_{t-1})$ can be estimated using the Nadaraya-Watson estimator or kernel smoother

$$\hat{m}_h(r) = \frac{\sum_t K[(r_{t-1} - r)/h]\Delta r_t}{\sum_t K[(r_{t-1} - r)/h]} \tag{4}$$

where K is a smoothing kernel (the Normal kernel will be chosen here) and h the smoothing parameter or window size. The skedastic function is estimated similarly by

$$\hat{s}_h(r)^2 = \frac{\sum K[(r_{t-1} - r)/h]\Delta r_t^2}{\sum K[(r_{t-1} - r)/h]} - \hat{m}_h^2(r) \tag{5}$$

The function $\hat{s}_h(r)^2$ can be also considered as the variance of the non-parametric estimate of the conditional expectation $\hat{m}_h(r)$. It is then possible to use it to construct a uniform confidence band for the regression function $\hat{m}_h(r)$. In the IID case, the quantity $\sqrt{Th}[\hat{m}_h(r) - m(r)]$ can be approximated by a Gaussian process $G(r)$. The asymptotic distribution of the supremum of $|G(r)|$ was derived by Bickel and Rosenblatt (1973). Härdle (1990, chap. 4) shows how this result can be used to derive an approximate uniform confidence bound for a given value of r. For the Normal kernel, we can define d_T as

$$d_T = (2\delta \log T)^{1/2} + (2\delta \log T)^{-1/2} \log \left(\frac{0.25}{2\pi^2}\right)^{1/2} \tag{6}$$

where $1/5 < \delta < 1/3$. A $(1 - \alpha)$ confidence band is defined by

$$\hat{m}_h(r) \pm \left(\frac{c_\alpha}{(2\delta \log T)^{1/2}} + d_T\right) \left(\frac{1}{2Th\sqrt{\pi}} \frac{\hat{s}_h^2(r)}{\hat{f}_h(r)}\right)^{1/2} \tag{7}$$

where $\hat{f}_h(r)$ is a non-parametric estimate of the marginal density of r; c_α is defined such that $\exp(\exp(-c_\alpha)) = 1 - \alpha$. For a 95% confidence interval, $c_{0.05} = 3.66$.

The choice of the bandwidth h has to be made so as to minimise the integrated squared error where the error is measured as the difference between the theoretical and the empirical regression curves. This can be approximated by minimising the estimated prediction error

$$\frac{1}{T}\sum_t [\Delta r_t - \hat{m}_h(r_t)]^2, \tag{8}$$

where $\hat{m}_h(r_t)$ is computed as the "leave-one-out" estimator deleting the tth observation in the sums. This automatic method is asymptotically optimal, but computer intensive and it converges very slowly. Convergence can be improved by multiplying (8) by a penalty function to penalise too small a h. The method is described in Rice (1984) which proposes for instance $(1-2T^{-1}h^{-1}K(0))^{-1}$ as a penalty function. A quicker method consists in selecting graphically the optimal h so to obtain a "nice" regression curve.

2.3 Bayesian inference for the discretised model

A non-parametric analysis is useful to suggest particular parameterisations for $m(r)$ and $s(r)$ in the discretised model (3). Most of the time, the proposed parameterised drift function $m(r_{t-1}, \alpha)$ will be linear in α (but nonlinear in r) while the proposed parameterised skedastic function $\sigma\, s(r_{t-1}, \theta)$ will be a highly nonlinear function in θ. Consequently, the discretised model under scrutiny is a dynamic linear regression model with heteroskedastic errors. Conditionally on the nonlinear parameters θ of the skedastic function, this model becomes linear so that Bayesian inference remains simple provided we use a natural conjugate prior for the linear regression parameters α and σ^2. To make this clear, let us call y_t the endogenous variable Δr_t and x_t the set of predetermined variables and define the following transformations

$$\begin{aligned} y_t(\theta) &= y_t/s_t(\theta) \\ x_t(\theta) &= x_t/s_t(\theta) \end{aligned} \tag{9}$$

so that our dynamic linear regression model with heteroskedastic errors can be written as

$$y_t(\theta) = x_t'(\theta)\alpha + v_t \tag{10}$$

where v_t is $\mathrm{N}(0,1)$. The likelihood function of this model is

$$L(\alpha, \sigma^2, \theta; y) \propto \sigma^{-T}[\prod s_t(\theta)]^{-1/2} \exp -\frac{1}{2\,\sigma^2}\sum[y_t(\theta) - x_t'(\theta)\alpha]^2 \tag{11}$$

We shall consider a flat prior on α and $\log(\sigma^2)$. The prior on θ is left unspecified for the moment and is noted $\varphi(\theta)$. The posterior density of this model can be analysed using a partial Gibbs sampler. More precisely, the parameter σ^2 can first be integrated analytically. Second, the conditional posterior density of α given θ is Student with:

$$\varphi(\alpha|\theta, y) \propto f_t(\alpha|\alpha_*(\theta), M_*^{-1}(\theta), s_*(\theta), T) \tag{12}$$

where

$$\left\{ \begin{aligned} M_*(\theta) &= \sum x_t(\theta)x_t'(\theta) \\ \alpha_*(\theta) &= M_*^{-1}(\theta)[\sum x_t'(\theta)y_t(\theta)] \\ s_*(\theta) &= \sum y_t'(\theta)y_t(\theta) - \alpha_*'(\theta)M_*(\theta)\alpha_*(\theta) \end{aligned} \right. \tag{13}$$

Finally, the marginal posterior density of θ has the following form

$$\varphi(\theta|y) \propto [\prod s_t(\theta)]^{-1}[s_*(\theta)]^{-T/2}|M_*(\theta)|^{-1/2}\,\varphi(\theta) \tag{14}$$

The griddy Gibbs sampler[1] of Bauwens and Lubrano (1998) is used to get draws of θ. Conditionally on each draw of θ, a draw of α is obtained using (12). A variance reduction technique to obtain the marginal moments of α and σ^2 consists in averaging the analytical conditional moments along the principles explained in Bauwens, Lubrano and Richard (1999: chap 3).

3 Bayesian model evaluation

A preliminary non-parametric analysis gives visual indications on the choice of a possible parametrisation for the model. In order to verify if these choices are valid, we can measure the ability of the model to reproduce the sample characteristics, as measured by the preliminary non-parametric analysis. In this section, we explore, in a Bayesian framework, how this comparison can be done.

3.1 General setting

Model evaluation is concerned with checking if a particular model, given an observed sample, manages to reproduce correctly some of the characteristics of this sample. This problem has been addressed in the Bayesian literature in many papers. The two seminal papers of Box (1980) and of Rubin (1984) consider the **sampling** properties of data simulated with the predictive density of the model. The predictive distribution attached to model M

$$p(y|M) = \int f(y|\eta, M)\,\varphi(\eta|M)\,d\eta \tag{15}$$

represents the distribution of all possible samples y that would occur if model M were true. When a sample yd is observed, it is used to make inference (revise the prior with Bayes theorem), or also to check model adequacy by computing the probability that it has actually been generated by M. If $g(yd)$ is a function of the observed sample, Box (1980) computes the predictive P-value of this transformation by reference to the predictive distribution $p(g(y)|M)$. A small tail area probability casts doubts on the model. But of course, this approach can hold only with an informative prior $\varphi(\eta|M)$. Rubin (1984) considers hypothetical replications of the observed sample yd and checks if they are close enough to the original sample yd. He considers similar transformations $g(yd)$ as Box (1980) but prefer to gauge them in reference to $f(g(y)|\eta, M)$ averaged not with the prior, but with the posterior density of η. This is what he calls the posterior predictive density. If inference has been done using a Monte Carlo Markov Chain (MCMC) method, this is an easy task because draws from the posterior density of η are available and consequently the posterior predictive density is obtained as a by-product of inference. Suppose that we have observed a sample yd of size T and that a MCMC method generated N draws for η. For each draw η_i, we can generate a new sample y_i of size T from $f(y|\eta_i)$ which represents the data density of our model under inspection and compute "statistics" with $g(y_i)$. With the N draws, we can then estimate the posterior predictive density of $g(.)$ and see in which quantile of it the equivalent statistics $g(yd)$ obtained from the observed sample yd falls. For recent references see Meng (1994).

[1] Note that a direct evaluation of the posterior density of θ is possible with a Simpson rule when the dimension of θ does not exceed 2. But this approach is not convenient for model comparison.

3.2 Measuring distances between densities

The above analysis is made difficult by the choice of the "right" transformation $g(.)$ to make. For instance, in the usual problem of Normal sampling with unknown mean and variance, Rubin notes that the first two moments may be the same between actual and simulated samples, but that gaps between order statistics will be different if the true data generating process is Cauchy. The idea we would like to develop here is that instead of finding adequate statistics (such as transformation of standardised residuals as in the seminal paper of Box (1980)), we prefer to consider the complete posterior predictive distribution and compare it to the empirical distribution of the sample. We have thus to compare the density of the observed sample to the density of a simulated sample of the predictive density. We are thus in the context of comparing two unknown distribution functions. This problem has been addressed in the nonparametric literature, see references in for instance Pagan and Ullah (1999), section 2.9.

Basically, we have first to define a distance between two densities f and g. Most of the nonparametric tests consider the integrated squared difference (or error)

$$ISE = \int [f(x) - g(x)]^2 dx = \int f(x)\, dF(x) + \int g(x)\, dG(x) - 2\int g(x)\, dF(x)$$

where we have developed the squares and used the definition of the Stieltjes integral. The Monte Carlo estimate of $\int f(x)dF(x)$ is $\sum f(x_i)/n$ when the x_i are distributed according to $F(x)$ while the kernel estimate of $f(x)$ is $\sum K((x_i - x)/h)/nh$. Consequently if yd and ys are the observed and simulated data with respective distributions F and G, the integrated squared difference can be estimated by the double sum

$$\hat{I}SE_T = \frac{1}{T^2 h}\sum_i \sum_j \left\{K(\frac{yd_i - yd_j}{h}) + K(\frac{ys_i - ys_j}{h}) - 2K(\frac{yd_i - ys_j}{h})\right\}$$

Using an asymptotic argument, Li (1996) constructs an unbiased classical test based on

$$\tilde{I}SE_T = \frac{1}{T^2 h}\sum_i \sum_{j\neq i} \left\{K(\frac{yd_i - yd_j}{h}) + K(\frac{ys_i - ys_j}{h}) - K(\frac{yd_i - ys_j}{h}) - K(\frac{ys_i - yd_j}{h})\right\}$$

When $h \to 0$ and $Th \to \infty$,

$$T\sqrt{h}\tilde{I}SE_T \to \mathrm{N}(0, \sigma_I^2)$$

where $\sigma_{ISE}^2 = 2\{\int[f(x)+g(x)]^2 dx\}\{\int K^2(u)du\}$. This test, which is said to perform well, has the drawback for us of relying on two asymptotic arguments: one for its evaluation with a Monte Carlo estimator for the integral, one for its distribution. We also note that other tests exists which are based on different measures of discrepancy. For instance Ait-Sahalia (1996a) uses the mean integrated squared error. Other examples are given in Pagan and Ullah (1999). In this paper, we follow of course a different route. We consider first another type of measure of discrepancy which is the ϕ divergence defined by:

$$D_\phi = \int \phi\left(\frac{f(x)}{g(x)}\right) g(x)\, dx \tag{16}$$

For different choices of ϕ, we recover usual measure of distances such as the χ^2 ($\phi(z) = (z-1)^2$), the negative entropy or Kullback-Leibler divergence ($\phi(z) = z\log(z)$), and the

Hellinger distance to which we shall devote our attention. It is obtained for $\phi(z) = (\sqrt{z}-1)^2$. This distance has two advantages. It is symmetric and it is at value between 0 and 2. We have

$$D_H = 2\left(1 - \int \sqrt{f(x)\,g(x)}\,dx\right) = 2\left(1 - \int \frac{\sqrt{f(x)g(x)}}{f(x)} dF(x)\right) \tag{17}$$

We have two ways of evaluating this distance in our case. We can apply the usual classical procedure to get the following estimate:

$$\tilde{D}_H = 2(1 - \frac{1}{T}\sum_i \frac{\sqrt{\sum_j K(\frac{yd_i - yd_j}{h}) \sum_j K(\frac{yd_i - yds_j}{h})}}{\sum_j K(\frac{yd_i - yd_j}{h})}) \tag{18}$$

But we cummulate two types of approximations as already mentioned. We can first reduce the variance of this estimator replacing the Monte Carlo evaluation of the integral by a simple Simpson rule in formula (17)

$$\hat{D}_H = 2\left(1 - \frac{d}{3}\sum_i \sqrt{\hat{f}(y_i)\,\hat{g}(y_i)}\,w_i\right) \tag{19}$$

where d is the size of the increment in y_i, w_i is a weight being 1 for end points, 2 for even points and 4 for odd points. The Kernel estimate is kept for f and g:

$$\hat{f}(y) = \frac{1}{Th}\sum K(\frac{yd_i - y}{h}) \qquad \hat{g}(y) = \frac{1}{Th}\sum K(\frac{ys_i - y}{h}) \tag{20}$$

We can then evaluate the empirical distribution of (19) by simulation as we are in a Bayesian framework.

3.3 Testing using the estimated Hellinger distance

We shall use the estimated Hellinger distance to compare models. The model producing the smallest distance to the empirical distribution of the observations will be the preferred model. A model producing a "small" estimated distance (sufficiently near from zero) will be qualified of well specified. As the Hellinger distance is at value between 0 and 2, it is relatively easy to say if it is small or not. We have just to compute the probability that $\hat{D}_H$ is lower than a given size a, chosen relatively near from zero, typically 0.01, 0.05 or 0.1. This probability corresponds to a posterior predictive P-value for a. We select the model which produces the highest probability, which means selecting the model which has the highest probability of producing a D_H close to zero. In order to compute this probability, we need to simulate the empirical distribution of $\hat{D}_H$, i.e. to devise a simulation skeme for the predictive density.

3.4 How to simulate the predictive density

We are typically in the context of a dynamic model with initial conditions. The marginal density of the first observation is obtained as

$$f(r_1) = \int f(r_1|r_0,\eta)\,\hat{f}(r_0)\,\varphi(\eta|yd)\,dr_0\,d\eta \tag{21}$$

The initial condition r_0 can be drawn from the empirical distribution of the observed sample yd and η is drawn from the posterior density. Conditionally to these two draws, a draw from $f(r_1|r_{0i}, \eta_i)$ can be obtained. If we repeat T times this operation (a draw from $\hat{f}(r_0)$, a draw from $f(\eta|yd)$ and the corresponding draw from $f(r_1|r_{0i}, \eta_i)$), we get a sample of T draws from the posterior predictive distribution of r_1. We compute the Hellinger distance between this first simulated sample and the observed sample yd. We call D_{H1} this value. Once we have obtained this first sample, we can pursue the process with

$$f(r_2) = \int f(r_2|r_1, \eta)\, \hat{f}(r_1)\, \varphi(\eta|yd)\, dr_1\, d\eta \tag{22}$$

which means that this time we consider drawings of r_2, given the previous draws of r_1. We compute a second Hellinger distance. If we continue the process, we get draws from the empirical distribution of $\hat{D}_H$.

4 Non-parametric evidences for the US Treasury bill rate

For the US bond market, the three month Treasury bill rate is often taken as a proxy for the spot rate. See e.g. Pfann, Schotman and Tschernig (1996) or Jiang (1998) among others. The data are taken from the on line data bank of the Federal Reserve bank of St Louis. The monthly observations cover the huge period 1934:1-1999:3. These data are highly heterogeneous and it is certainly the reason why most authors start their sample at the beginning of the sixties. Consequently, we started our sample in 1965:1. Interest rates before this date are very low and not very volatile. Figure 1.a displays the data together with a non-parametric

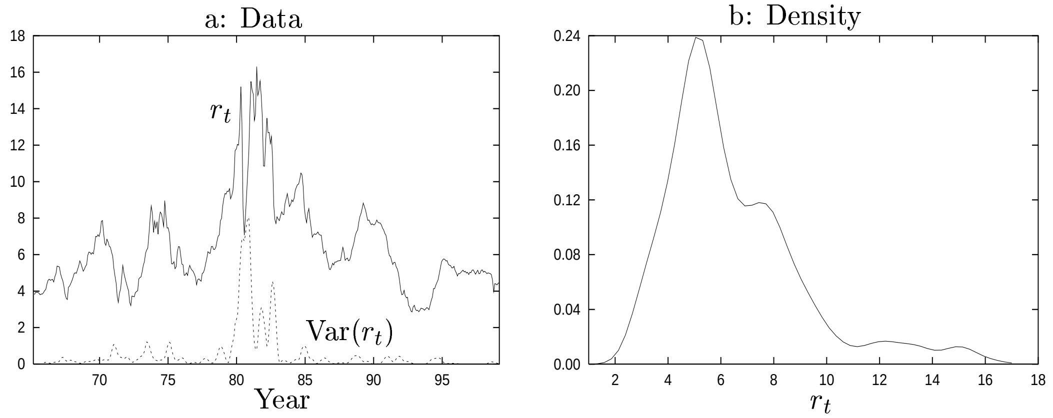


Figure 1: Data, volatility and density of the short term interest rate

estimate of the volatility computed by local averaging over 13 observations weighted by the Epanechnikov kernel. This graph shows that most of the volatility is located in the years 1979-1982. During this period, the U.S. Federal Reserve deviated from its usual practice of targeting nominal interest rates and chose non-borrowed reserves as a new instrument for monetary policy. The same type of switch was operated in most European countries during that period and also led to previously unknown levels and volatility of interest rates.

Figure 1.b displays the non-parametric estimate of the marginal density of r using Silverman's rule to determine the bandwidth. The complexity of the data appears from this

graph. First of all, this density is very skewed. This skewness can be produced only with an heteroskedastic model where probably volatility is a function of past levels. Second, this density has at least two bumps, denoting the mixture of well individualised distributions. Finally, the long right tail with interest rates above 10 per cent is produced by the 1979-1982 influential period.

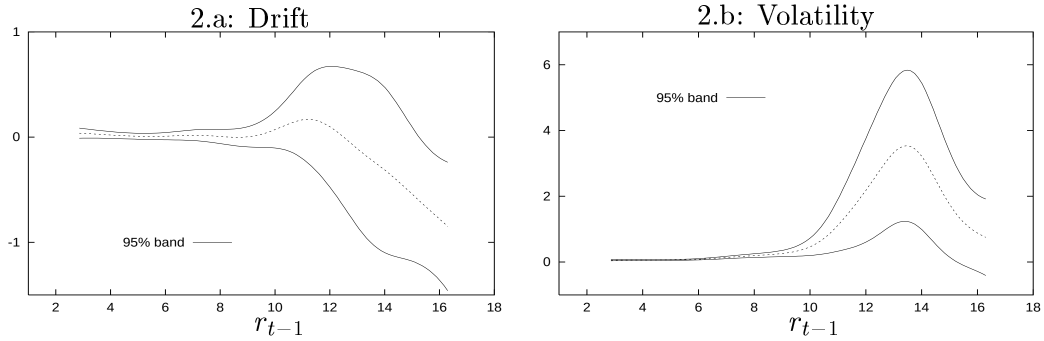


Figure 2: Non-parametric estimation of the discretised diffusion

Let us now estimate of the drift function $m(r)$ of the dicretised model (3) using a non-parametric regression of Δr_t over r_{t-1}. The smoothing parameter h was chosen equal to 2 after visual inspection of the graph of $\hat{m}(r)$. The estimated drift function, displayed in Figure 2.a together with a 95% uniform confidence band, appears to be roughly constant and equal to zero for rates up to 10%. This is in favour of a random walk. For higher rates, the estimated drift function adopts a negative linear slope. However, a closer look reveal a very slight mean reversion for rates lower than 6 per cent. Ait-Sahalia (1996a), for the seven day Eurodollar deposit rate, considers $\mu(r) = \alpha_0 + \alpha_1 r + \alpha_2 r^2 + \alpha_3/r$ in the continuous time case. In the present case, the drift seems to be linear in r_{t-1} with a break. This suggests a two regime model in the drift as in Pfann, Schotman and Tschernig (1996), Gray (1996) and Pai and Pedersen (1998). Note however that the mean reversion effect after $r = 10\%$ is very badly determined as shown by the 95% uniform confidence band. This is because there are comparatively very few observations in the high rate regime (85 observations after 8%, 34 after 10% and 22 after 12% out of 411).

The non-parametric estimate of the volatility function $s(r)$ of the discretised model (3) confirms, as shown in Figure 2.b, the apparent relation between high volatility and high interest rates. However, the shape of the curve really questions the validity of the simple and usual drift function $\sigma(r) = \sigma r^\gamma$. Volatility seems to be constant for low rates and to increase for higher rates. However, volatility decreases after apparently the same level as the one after which the drift function starts to be negative. For compatibility with existing bond pricing models, we can be tempted by maintaining the constant elasticity of variance model. In this case, a two regime model with $\gamma = 0$ in the first regime and $\gamma > 0$ in the second regime is plausible. The parametric formulation suggested by Ait-Sahalia (1996a) for the seven day Eurodollar deposit rate $\sigma^2(r) = \sigma_0 + \sigma_1 r + \sigma_2 r^{2\gamma}$ may constitute an alternative to this two regime model. However, if we really want to take into account the decrease in volatility, the parametric form $s^2(r) = \sigma/(1 + \gamma(r - c)^2)$ may prove rather parsimonious as with two parameters, γ and c, it can mimic rather well the bell shape displayed in Figure 2.b.

5 Another Bayesian look at the CEV model

We shall review and compare various versions of the usual constant elasticity of variance model (CEV) promoted by Chan, Karolyi, Longstaff and Sanders (1992), considering a linear and a non-linear drift. We want in particular to check the assertion of Ait-Sahalia (1996a) according to which the main source of misspecification comes from the linear drift. But we also want to check the role played by the parameterisation of the volatility function. We start from the discretised parametric model

$$\Delta r_t = \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-1}^2 + \alpha_3 / r_{t-1} + \sigma s_t \epsilon_t, \tag{23}$$

where $\epsilon_t \sim \mathrm{N}(0,1)$. We shall consider the following volatility functions

- The simple Vasicek (1977) formulation with $s_t = 1$.
- The CEV model[2] with $s_t^2 = r_{t-1}^{2\gamma}$.
- Finally, we introduce the Cauchy volatility (CAV) function

$$s_t^2 = \frac{1+\gamma}{1+\gamma(r_{t-1}-c_r)^2} \tag{24}$$

  where volatility is an increasing function of r, reaches a maximum at $r_{t-1} = c_r$ and decreases after. The parameter γ monitors the transition between periods of low interest rates and periods of high interest rates[3]. The parameter c_r is in a way a threshold as it gives an indication on the level of r after which volatility decreases.

We want to compare three models (Vasicek, CEV, CAV) with a linear and a non-linear drift. This makes 6 different variants. Let us first examine the posterior probability that the Hellinger distance is lower than a given level in Table 1.

The first conclusion we can draw from this table is that heteroskedasticity has to be taken into account ($s(r) \neq 1$). The Vasicek model is dominated and even fails to produce positive interest rates with a linear drift. The second conclusion is that a non-linear drift function seems better in every case. However, once heteroskedasticity is taken into account, it is difficult to discriminate among the different models.

Let us now give inference results for the CEV and CAV models with a non-linear drift. From Table 2, we see that the parameters of the drift function are more precisely estimated in the CAV model. Its volatility function manages to capture exactly the peak of volatility which appeared around $r = 13$ in the non-parametric analysis. For the CEV model, the posterior expectation of γ is greater than one, leading in theory to a non-stationary model when the drift is linear. Note however that the posterior expectation of α_2 is negative, restoring thus stationarity in the case of a non-linear drift.

The graphs of the implied drift and volatility functions, computed at the posterior expectation of α and θ and displayed in Figure 3, give more details about the ability of the different

[2]Let us recall that a linear drift function ($\alpha_2 = \alpha_3 = 0$) together with $\gamma > 1$ entails non-stationarity, but that stationarity may be recovered with a non-linear drift where $\alpha_2 < 0$, at least in the continuous time case.

[3]Note that γ is a scale parameter which may be slightly exchangeable with σ^2. An exponential prior with $\varphi(\gamma) \propto \exp(-\gamma)$ solves this question.

Table 1: Posterior predictive P-values
for Vasicek, CEV and CAV models

Model	$s_t^2 = 1$			$s_t^2 = r^{2\gamma}$			$s_t^2 = \dfrac{1+\gamma}{1+\gamma(r-c_r)^2}$		
a	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
N.L.	0.00	0.00	0.00	0.00	0.62	1.00	0.00	0.50	0.96
L.D.	-	-	-	0.00	0.56	0.82	0.00	0.28	0.76

Computations were based on 250 periods ahead. The volatility function is detailed on the horizontal axis. The drift function varies with the vertical axis. L.D. means linear drift and N.L. non-linear drift. '-' indicate negative draws. Number corresponds to $\Pr(\hat{D}_H < a)$.

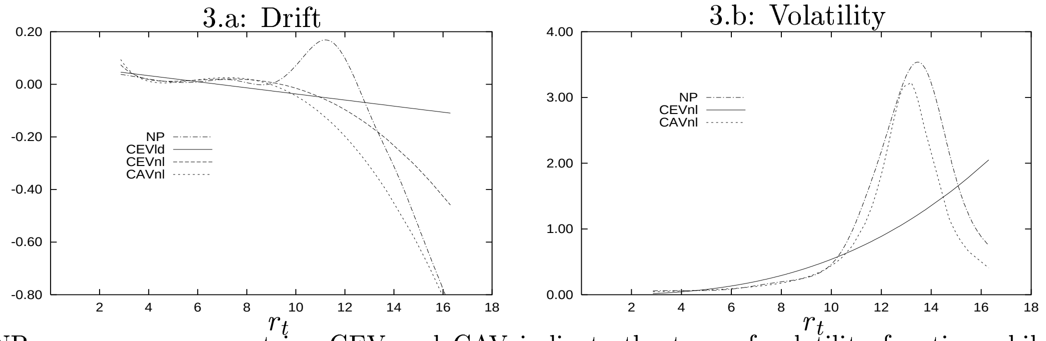
Table 2: Inference results for CEV and CAV models
with a nonlinear drift

Model	α_0	α_1	α_2	α_3	γ	c_r	σ^2
CEV	-0.32	0.14	-0.020	0.24	1.37	-	0.013
	[0.29]	[0.13]	[0.017]	[0.20]	[0.08]	-	
CAV	-0.55	0.25	-0.035	0.40	4.42	13.08	0.60
	[0.28]	[0.12]	[0.015]	[0.21]	[1.34]	[0.30]	

Posterior standard deviations are indicated between squared brackets.

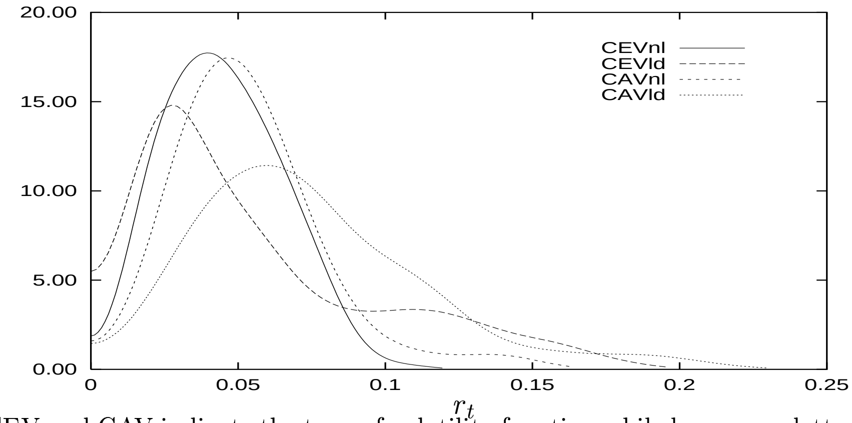
specifications to reproduce some of the characteristics of the data. If there is not much differences between the different nonlinear drift functions, marked differences are present for volatility functions. The CAV model gives a far better account of the volatility present in the data than the CEV model.

In Figure 4, we display the sampling distribution of the Hellinger distance for the four interesting processes. The Vasicek model is discarded as it was clearly rejected by the data from Table 1. It is apparent at first sight that, on average, the Hellinger distance is very close from zero for all models. But there is a large difference in tail behaviour for the models with a linear drift. In particular, the large right tail generated by the CEV model with a linear drift may be due to a problem of non-stationarity, but the CAV model with a linear drift produces the same behaviour. We can confirm the statement made in Ait-Sahalia (1996a) concerning linear drifts, but our conclusions are less strong than his, obtained with a classical procedure, and moreover with a different data set.



NP means non-parametric. CEV and CAV indicate the type of volatility function while lower case letters ld and nl indicate respectively linear drift and non-linear drift.

Figure 3: Implied drift and volatility functions



CEV and CAV indicate the type of volatility function while lower case letters ld and nl indicate respectively linear drift and non-linear drift.

Figure 4: Sampling distribution of the Hellinger distance

6 The impact of a two regime modelling

As a possible alternative to the CEV model, many authors have proposed to consider switching regime models. Pfann, Schotman and Tschernig (1996), Gray (1996), Pai and Pedersen (1998) consider a two regime model with a linear drift, but with different mean reversion coefficients in each regime. Gray (1996) and Pai and Pedersen (1998) consider a Markov switching model à la Hamilton (1989) which renders difficult the modelling of dependence in volatility. Pfann, Schotman and Tschernig (1996) consider a TAR (threshold auto regressive) model à la Tong (1983) which was shown to be more robust than the Markov switching model by Boldin (1996). With this model, they allow for a different linear drift function in each regime, but mainly impose a common volatility function.

Let us consider a transition function F at value in [0,1] defined as

$$F(c_1) = \begin{cases} 1 & \text{if } r_{t-1} > c_1 \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

The model we consider as a starting point is

$$\Delta r_t = [1 - F(c_1)]\, m_1(r) + F(c_1)\, m_2(r) + \sigma s_t \epsilon_t \tag{26}$$

Various drift functions are interesting to investigate. The first question is to know if the switching mechanism manages to exhaust all the non-linearity contained in the data. As most of the time negative interest rates are **the** problem in the simulation step with the discretised model, we consider as the most general formulation

$$\begin{aligned} m_1(r) &= \alpha_{10} + \alpha_{11} r_{t-1} + \alpha_{12}/r_{t-1} \\ m_2(r) &= \alpha_{20} + \alpha_{21} r_{t-1} \end{aligned} \tag{27}$$

In the first regime the term in $1/r_{t-1}$ penalises too small interest rates. Linearity in both regimes is an option that will be tested.

Let us now specify three options for the skedastic function s_t^2. We first consider a generalisation of the Vasicek model which allows for a different variance per regime which remains constant inside each regime

$$\begin{aligned} \sigma^2\, s_t^2 &= [1 - F(c_2)]\sigma_1^2 + F(c_2)\sigma_2^2 \\ &= \sigma^2[1 - F(c_2) + F(c_2)\phi] \end{aligned} \tag{28}$$

where $\phi = \sigma_2^2/\sigma_1^2$. We consider two thresholds parameters c_1 and c_2 to allow the drift and volatility functions to react differently to the level of the interest rate. We then consider the same options as before: the CEV and the CAV volatility functions.

Bayesian inference can be made on these eight variants using the same framework as the one presented in section 2, provided we consider the following transformation of the data

$$\begin{aligned} y_t(\theta) &= y_t/s_t(\theta) \\ x_t(\theta) &= [(1 - F(c_1)x_t', F(c_1)x_t']/s_t(\theta) \end{aligned} \tag{29}$$

so as to rewrite the model as in (10). c_1 is now the first element of θ. We report in Table 3 the posterior predictive P-values of a attached to each variant.

Let us first examine the impact of the modelling choices made for the drift functions on the predictive performance. Clearly, all models prefer to have a non-linear drift in the first regime. The switching mechanism together with a linear drift provides a satisfactory modelling for the drift function in the high interest rate regime, but need the help of an additional non-linear term for the low interest rate regime. If we now turn to the evaluation of the modelling of the volatility function, it appears that the CAV model is the best choice, by far. A switching mechanism inside the volatility function of the Vasicek model performs a very good job and provides a very interesting model when the drift function is non-linear.

Let us now give inference results in Table 4 for the different three volatility functions and a nonlinear switching drift function. When the volatility function is of the Cauchy type, the data indicate that the drift and volatility functions have inflexions for roughly the same level of the interest rate. For the two regime Vasicek model, c_1 and c_2 are on the contrary rather different, just because the two regime skedastic function in this case could be seen as just an approximation to the CAV function. We have now to judge if considering a switching regime

Table 3: Posterior predictive P-values
for various versions of the two regime model

Model	σ_1^2, σ_2^2			$r^{2\gamma}$			$\dfrac{1+\gamma}{1+\gamma(r-c_r)^2}$		
a	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
N.L.	0.00	0.62	0.99	0.01	0.66	0.96	0.04	0.90	1.00
L.D.	-	-	-	0.01	0.60	0.89	0.00	0.51	0.92

Computations were based on 250 periods ahead. The volatility function is detailed on the horizontal axis. The drift function varies with the vertical axis. L.D means a linear drift in each regime, while N.L. corresponds to the full specification. '-' indicate negative draws. Numbers correspond to $\Pr(D_H < a)$.

Table 4: Inference results for the switching Vasicek, CEV and CAV models
with a non-linear drift

Model	α_{10}	α_{11}	α_{12}	α_{20}	α_{21}	c_1	ϕ or γ	c_2 or c_r
Vasicek	-0.07	0.01	0.09	1.36	-0.28	12.86	4.66	9.94
	[0.14]	[0.03]	[0.14]	[2.59]	[0.45]	[0.41]	[0.67]	[0.37]
CEV	-0.10	0.02	0.12	0.47	-0.12	12.88	1.37	
	[0.11]	[0.03]	[0.10]	[2.27]	[0.40]	[0.36]	[0.08]	
CAV	-0.09	0.02	0.11	3.10	-0.57	12.97	4.33	13.06
	[0.12]	[0.03]	[0.11]	[2.80]	[0.47]	[1.32]	[1.32]	[0.29]

Posterior standard deviations are indicated between squared brackets.

model was worth the extra cost of computation. We compare in Figure 6 the best models coming from Table 1 and 3. We have a group of the three models which seem to be equivalent: the CEV, CAV and 2rCEV. One of them has a swithcing drift. But a fourth model dominates this group: the CAV with a switching drift. Consequently, a switching mechanism is clearly a major improvment for certain models. This is particularly true for the Vasicek model which is not displayed in Figure 6.

7 Implications for the term structure

Empirical models of the short term interest rate are used to price bonds of different maturities and bond options. Bond pricing is closely related to the relationships existing between rates corresponding to different maturities, the term structure. In this section, we consider the expectation hypothesis for explaining the term structure. We restrain our attention to a very simple case, that of one maturity. More precisely, we investigate the ability of the different

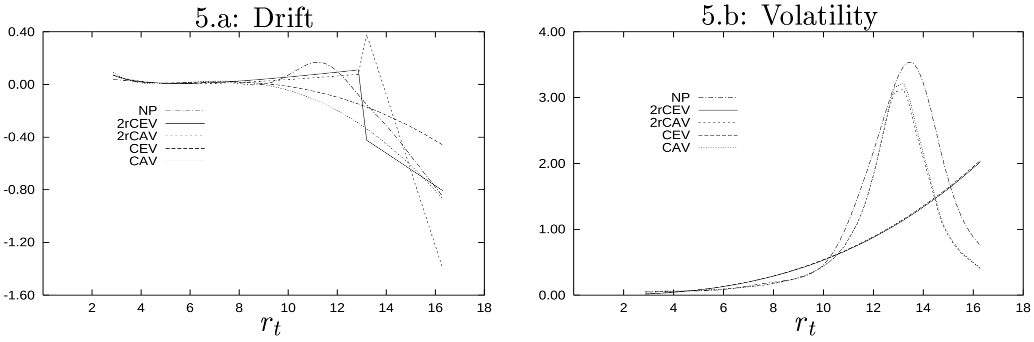


Figure 5: Implied drift and volatility functions

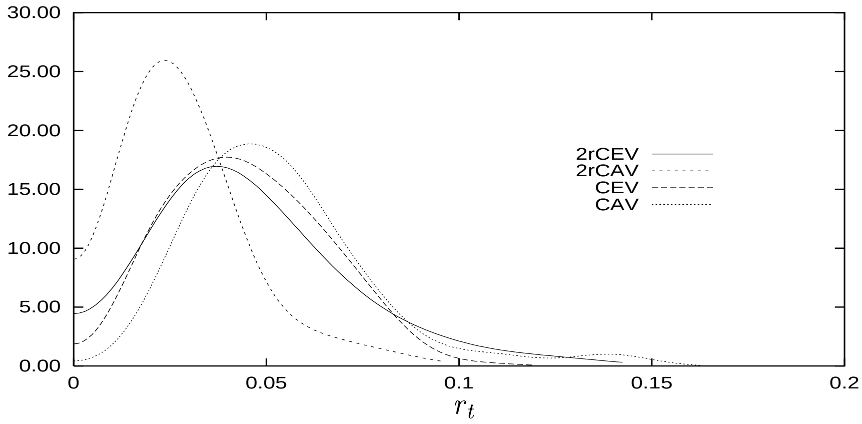


Figure 6: Comparing the best models with D_H

empirical models we have examined so far to predict the level of the ten year US Government bonds. However some details are needed beforehand to explain the expectation hypothesis.

7.1 Term Structure in Continuous Time

Let us consider a zero-coupon bond (or pure discount bond) issued at time t and offering a lump-sum payment of one unit at maturity $t + s$. We are looking for its price P at time t when the riskless rate is r. In the absence of arbitrage, there exists a probability measure such that

$$P_{t,s}(r) = \mathrm{E}[\exp(-\int_t^{t+s} r_u \, du)] \tag{30}$$

the expectation being taken in reference to this probability measure (see Duffie (1996: chap. 7)). The function $P_{t,s}(r)$ is usually called the term structure. It is also common to define the yield to maturity R as

$$R_{t,s}(r) = -\frac{1}{s} \log[P_{t,s}(r)] \tag{31}$$

which gives a prediction of the long term interest rate for maturity s when the spot rate is r.

Let us suppose that the riskless interest rate follows a diffusion process as defined in (1). The absence of arbitrage between bonds of different maturities implies that the price P must satisfy a partial differential equation:

$$\frac{\partial P}{\partial r}(\mu + \lambda\sigma) + \frac{1}{2}\frac{\partial^2 P}{\partial r^2}\sigma^2 + \frac{\partial P}{\partial t} - rP = 0 \tag{32}$$

where λ is the price of risk (see Dumas and Allaz (1996) for a complete derivation). A general solution to this partial differential equation can be found by computing the expectation defined in (30), using not the process of the observed interest rate, but the risk neutral process associated with the observed rate. The two processes are related by

$$dr = [\mu(r) + \lambda\sigma(r)]\, dt + \sigma(r)\, dW \tag{33}$$

where W is a Brownian motion under the historical probability. The parameter λ of the price of risk is not identified in the absence of extra information on the term structure.

The partial differential equation (32) can be solved analytically as a function of the parameters of the diffusion process (1) and of the market price of risk in certain particular cases. For instance, Vasicek (1977) start from an Orstein-Uhlenbeck process which means $\gamma = 0$ in (1) while Cox, Ingersoll and Ross (1985) consider a slightly more general process which corresponds to $\gamma = 0.5$ in (1). But in both cases, the drift function is linear. In these models, the yield of a discount bond to maturity s is an affine function of the riskless rate and of the market price of risk. Linearity comes from the linearity of the drift function. As we have promoted non-linear drift function, we have to use numerical methods to evaluate directly (30). We should get consequently a nonlinear yield curve.

7.2 The yield curve in discrete time

Let us now turn to the discrete time case which is of particular interest here as we have developed our different models for the riskless rate in discrete time. In discrete time, the following formulation of the expectation theory is usually adopted (see e.g. Pfann, Schotman and Tschernig (1996))

$$R_{t,s}(r) = \frac{1}{s}\sum_{i=0}^{s-1} \mathrm{E}_t[r_{t+i}] \tag{34}$$

Actual bonds with maturity longer than one year almost always carry coupons which are paid every year. For long maturity, these coupons account for most of the value of the bonds. So we have to introduce a discount rate $\bar{R}$ so that the yield to maturity becomes a weighted average of future expected short rate. This is the linearised model of Shiller (1979)

$$R_{t,s}(r) = \frac{1-\delta}{1-\delta^s}\sum_{i=0}^{s-1} \delta^s \mathrm{E}_t[r_{t+i}] \tag{35}$$

where $\delta = (1+\bar{R})^{-1}$. Following Campbell and Shiller (1987), this relation implies that if the long and short term rates are I(1) processes, they are cointegrated. Considering the the yield of ten year US Government bonds, s is fixed and will be dropped from the indices. As the short term interest rate we analysed has a maturity of three months, the maturity s of long term bonds is $s = 120/3 = 40$ so that (35) becomes:

$$R_t(r) = \frac{1-\delta}{1-\delta^s}\sum_{i=0}^{39} \delta^s \mathrm{E}_t[r_{t+3i}] \tag{36}$$

We take $\bar{R}$ equal to the empirical mean of R_t and adjust δ for the three month maturity of r_t so that $\delta = (1+\bar{R})^{-3/12}$.

7.3 The simulation procedure

We want to derive the yield curve for a range of different values of the short rate and a constant maturity of 40 quarters using (36). In (36), E_t is a conditional predictive expectation, modified to take into account the price of risk.

The general formula for conditional predictive expectation resulting from the estimated models of the interest rate is

$$\mathrm{E}(r_{T+s}|r_T) = \int r_{T+s}\, f(r_{T+s}|r_{T+s-1},\eta) \cdots f(r_{T+1}|r_T,\eta)\, \varphi(\eta|r_T)\, d\eta\, dr_{T+s} \cdots dr_{T+1} \tag{37}$$

This integral can be evaluated by simulation, using the draws of η and drawing recursively $r_{T+j}, j = 1, s$ according to

$$r_{T+j} = m(r_{T+j-1},\eta_i) + \sigma_i s(r_{T+j-1},\eta_i)\epsilon_i \tag{38}$$

where the starting value r_T is taken from a predefined grid. As we want in fact to take expectations according to a risk neutralised probability, we have to translate $m(r_{T+j-1},\eta_i)$ with $\lambda\, s(r_{T+j-1})$. The drift function (23) of one-regime models is modified as follows

$$\tilde{m}(r) = \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-1}^2 + \alpha_3/r_{t-1} + \lambda\, s(r_{t-1}) \tag{39}$$

while the drift function of two-regime models becomes

$$\begin{aligned} \tilde{m}_1(r) &= \alpha_{10} + \alpha_{11} r_{t-1} + \alpha_{12}/r_{t-1} + \lambda_1\, s(\theta, r_{t-1}) \\ \tilde{m}_2(r) &= \alpha_{20} + \alpha_{21} r_{t-1} + \lambda_2\, s(\theta, r_{t-1}) \end{aligned} \tag{40}$$

allowing for a different impact of the price of risk in each regime. Since $\lambda > 0$, the adjustment has the effect of increasing the mean of the interest rate. When volatility is a function also of r it has the effect of decreasing the amount of mean reversion. The inclusion of the price of risk in the drift function thus modifies the dynamic properties of the model. Consequently stationnarity and positivity are no longer guarantied. Every time a negative draw or an explosive draw is produced, we have chosen to redraw the whole sequence of 120 predictions.

7.4 Empirical results

We have to calibrate λ for the simulations. Confronted to the same type of problem, Ait-Sahalia (1996b) chose the λ which minimised the RMSE between the simulated yield curve and a non parametric estimate. We can estimate the simplified term structure with a non-parametric regression of R_t over r_t. Given this curve displayed in Figure 7, we minimise the RMSE with a grid search over λ. When short term rates are between 6 and 12 per cent on an annual basis, the relation is most likely linear and long rates are proportional to short rates, giving weight to the cointegration model of Campbell and Shiller (1987). In this range, short rates do behave like a random walk. But below 6 per cent and above 12 per cent, the linearity property disappears. Long rates under-react to short rates. This property is more apparent for spot rates greater then 12 per cent than for spot rate lower than 6 per cent.

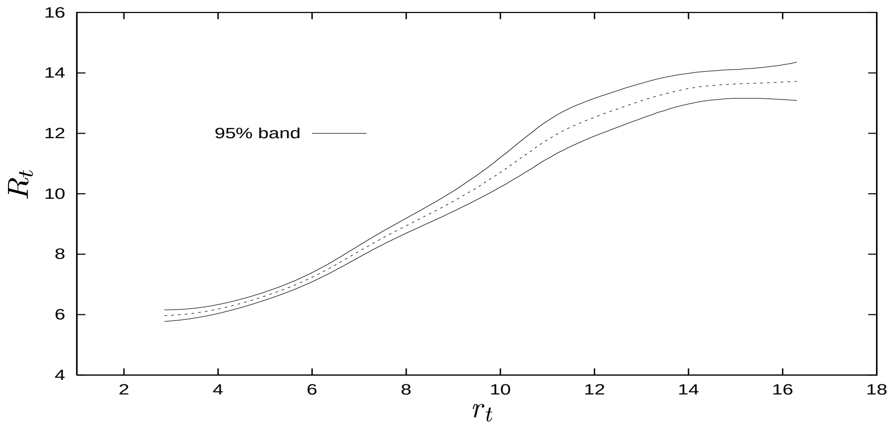


Figure 7: Non-parametric estimate of the yield curve

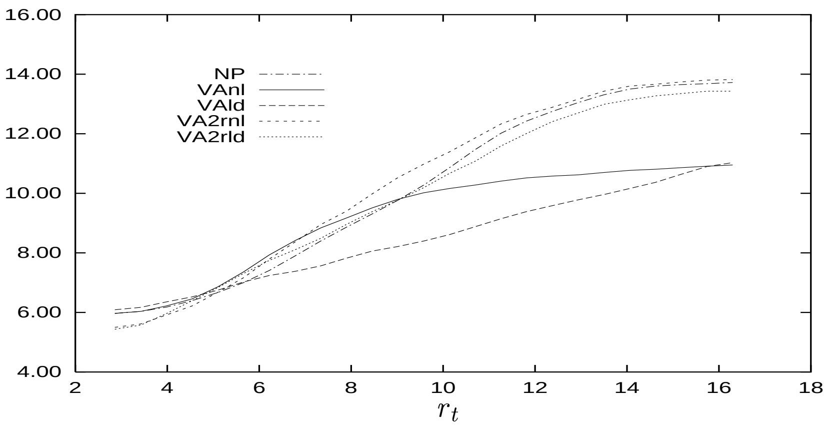


Figure 8: Implied yield curve for the Vasicek volatility

7.4.1 The Vasicek model

As expected, the model with a linear drift (VAld in Figure 8) produces a linear term structure. With $\lambda = 0.04$, the curve is valid for very small r, but then diverges from the non-parametric estimate (RMSE = 2.148). Considering a non-linear drift ((VAnl in Figure 8) with $\lambda = 0.01$ improves greatly the result for r up to 10 per cent. But for higher values, the yield curve is not realistic (RMSE = 1.611). The estimated mean reversion is too large.

A two-regime model can improve greatly this upsetting picture. We have allowed for different volatility and drift functions in the two regimes with two different threshold parameters. This is model 1 in Table 3. Allowing for a different price of risk in each regime with $\lambda_1 = 0.05$ and $\lambda_2 = 0.6$ produces a yield curve which mimics the non-parametric estimate in a much nicer way. The switching mechanism, even with a linear drift in each regime (VA2rld in Figure 8) manages to mimic the curvature at both ends (RMSE = 0.297). A non-linear drift in the first regime does not seem to be useful (RMSE = 0.383).

7.4.2 The constant elasticity model

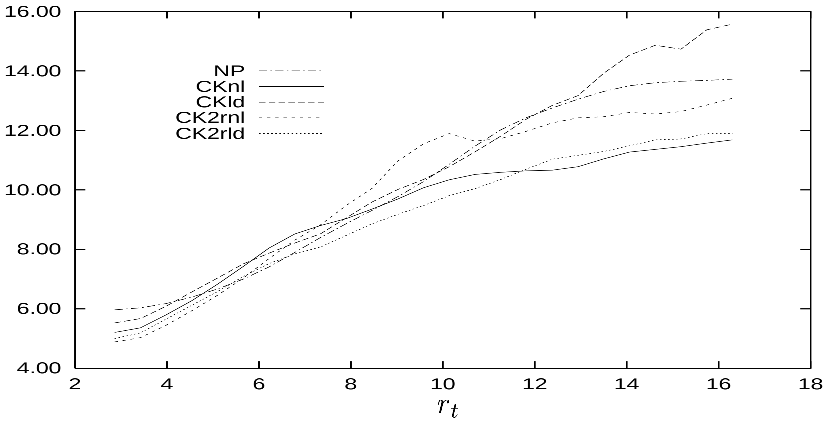


Figure 9: Implied yield curve for the Constant elasticity of volatility model

The constant elasticity of volatility model was estimated with a posterior expectation for γ which was greater than one. As a consequence, the model has a tendency to produce explosive long run prediction when equipped with a linear drift. Moreover, when 500 draws of η were enough with the other models to produce a smooth yield curve, this number was clearly not sufficient here.

Using $\lambda = 0.07$, the implied yield curve with a linear drift is very convincing for $r < 12$, but explodes after that value (CKld in Figure 9, RMSE $= 0.689$). Compared to the Vasicek model, a volatility function which depends on the level of interest rate changes completely the behaviour of the model in term of mean reversion. A non-linear drift function solves the problem of stationarity, but introduces a too strong mean reversion (CKnl in Figure 9, RMSE $= 1.370$) which ends in a result which is worse than before.

The two-regime model is not very much at ease with this type of volatility function as it was not possible to adjust correctly λ_2. We ended with $\lambda_1 = 0.03$ and $\lambda_2 = 0.30$. A higher λ_2 put the model in real trouble. The within regime linear drift performs worse than in the single regime case with RMSE $= 1.283$. A non-linear drift in the first regime is slightly better (RMSE $= 0.752$.)

7.4.3 The Cauchy volatility model

Here again, a price of risk which implies a time-varying volatility ($\lambda\, s(r)$) induces a behaviour which is far different from the single regime Vasicek model. However, we do not experience here the problem of nonstationarity. With $\lambda = 0.09$, the implied yield curve is of course linear, but very close to the non-parametric estimate (RMSE $= 0.461$). A non-linear drift should be avoided (RMSE $= 1.370$) as the implied mean reversion is too strong for high interest rates.

A two-regime drift function with $\lambda_1 = 0.05$ and $\lambda_2 = 0.7$ does not manage to improve the performance of the model as the within regime linear drift leads to RMSE $= 0.585$ while the within regime non-linear drift produces RMSE $= 0.600$.

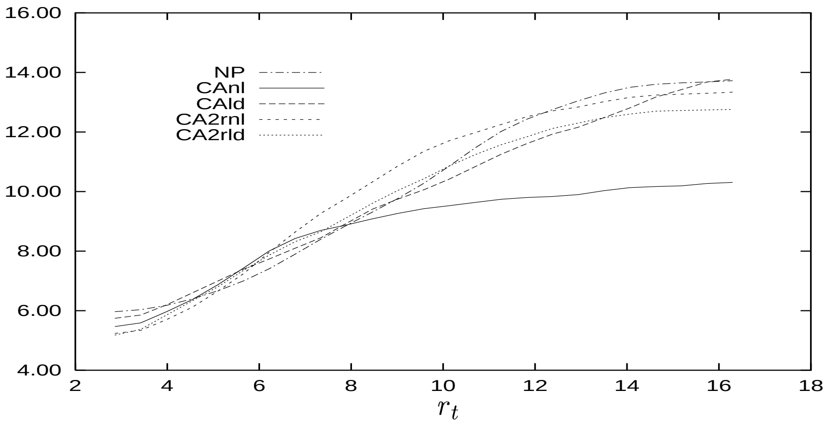


Figure 10: Implied yield curve for the Cauchy volatility model

Table 5: Implicit yield curves

	Vasicek			CEV			CAV		
Drift	λ_1	λ_2	RMSE	λ_1	λ_2	RMSE	λ_1	λ_2	RMSE
Linear	0.04	-	2.148	0.07	-	0.689	0.09	-	0.461
Non-linear	0.01	-	1.611	0.07	-	1.370	0.09	-	2.072
2 reg. lin.	0.05	0.6	0.297	0.03	0.3	1.283	0.05	0.7	0.585
2 reg. N.L.	0.05	0.6	0.383	0.03	0.3	0.752	0.05	0.7	0.601

7.4.4 Comparisons

We can now compare the different implied yield curves in Table 5 where we report the corresponding RMSE measuring the discrepancy between the simulated yield curves and the non-parametric estimate. Ait-Sahalia (1996a) does not favour regime shift models and prefer continuous but non-linear drift functions. When a model is estimated in discrete time, Table 5 is not at all in favour of this assertion as a regime shift in both the drift and the volatility function transform the worse model into the best one: the Vasicek model. We note also that a linear drift function can produce a linear yield curve which approximate quite well the slope of the non-parametric yield curve, provided the volatility function adopt a Cauchy representation.

In Figure 11 we have regroup the best simulations. We have finally the choice between the CAV model with a linear drift that implies a linear yield curve which approximates very well the non-linear and non-parametric yield curve. And the two regime Vasicek model which manages to take into account the non-linearity present in the yield curve for high values of the spot rates, but not for low values.

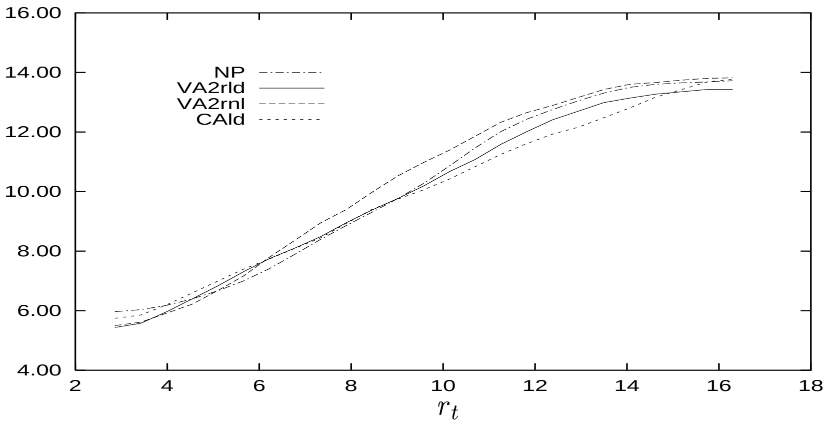


Figure 11: Comparison of models for the volatility

8 Conclusion

In this paper, we have combined the use of classical non-parametric estimation and of Bayesian inference. This mix was, to our point, a success. Exploration of the properties of the likelihood function is a common interest for the two approaches. The non-parametric approach gives intuitions and indications on the shape of potential relations. The Bayesian approach details the posterior density of parameters of interest and detect the possible lack of information given by the sample in certain directions. Moreover, the Bayesian approach provides simulation for the posterior predictive density which can be confronted to the actual sample. This confrontation needs density estimation which is done using again a non-parametric approach.

We have developed a method to evaluate and compare models. This method is concerned with the properties of the dynamic simulation of the estimated model and is thus rather severe. Ait-Sahalia (1996a) used a similar approach in that he compares the estimated density of the sample with the marginal density generated by parametrised continuous time process which is given by (42). The approach of Ait-Sahalia (1996a,b) is clearly different as it considers directly the continuous time process. It appears that it leads to rejecting models much more often than the one we adopted.

When a model is rejected by the predictive approach, there is no indication where the model should be improved. We did not provide or used misspecification tests. However, the preliminary non-parametric estimates of the drift and volatility function provide both a guide first for a priori selecting a particular parametric formulation and a posteriori for confronting the implied drift and volatility functions computed at their posterior expectation to their non-parametric equivalent.

A model may present very good predictive properties and at the same time not be very reliable for applications in finance. There is a gap between an empirical model of the short term interest rate and a modelling of the risk neutral probability. Some of our models could very easily reproduce the empirical distribution of the riskless interest rate. But it was very difficult to introduce the price of risk in a clever way when the drift function is non-linear. A model with a pure non-linear drift is, for this respect, very upsetting when a switching mechanism is much more reliable.

References

Ait-Sahalia Yacine (1996a) Testing continuous-time models of the spot interest rate. *Review of Financial Studies* 9(2), 385-426.

Ait-Sahalia Yacine (1996b) Non-parametric pricing of interest rate derivative securities. *Econometrica* 64(3), 527-560.

Auestadt B. and D. Tjøstheim (1990) Identification of non-linear time series: First order characterisation and order estimation. *Biometrika* 77, 669-687.

Bauwens Luc and Michel Lubrano (1998) Bayesian inference in GARCH models using the Gibbs sampler. *The Econometrics Journal*, Vol. 1, pp. C23-C46. http://www.blackwellpublishers.co.uk/ectj.

Bauwens Luc, Michel Lubrano and Jean-François Richard (1999) *Bayesian Inference in Dynamic Econometric Models.* forthcoming Oxford University Press.

Bickel P.J. and M. Rosenblatt (1973) One some global measures of the deviation of density function estimates. *Annals of Statistics* 1, 1071-1091.

Boldin M.D. (1996) A Check on the robustness of Hamilton's Markov switching model approach to the economic analysis of the business cycle. *Studies in Non-linear Dynamics and Econometrics* 1(1), 35-46.

Box George E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society* A, 143(4), 383-430.

Brenner Robin J., Richard H. Harjes and Kenneth F. Kroner (1996). Another look at models of the short-term interest rate. *Journal of Financial and Quantitative Analysis* 31(1), 85-107.

Broze Laurence, Olivier Scaillet and Jean-Michel Zakoian (1995) Testing for continuous-time models of the short-term interest rate. *Journal of Empirical Finance* 2(3), 199-223.

Campbell John Y. and Robert J. Shiller (1987) Cointegration and tests of present value models. Journal of Political Economy 95(5), 1062-1088.

Chan K.C., G.A. Karolyi, F.A. Longstaff and A.B. Sanders (1992) An empirical comparison of alternative models of the short-term interest rates. *Journal of Finance* 47, 1209-1227.

Cox J.C., J.E. Ingersoll and S.A. Ross (1985) A theory of the term structure of interest rates. *Econometrica* 53, 385-407.

Dacunha-Castelle D. and D. Florens-Zmirou (1986) Estimation of the coefficient of a diffusion from discrete observations. *Stochastics* 19, 263-284.

Duffie Darrell (1996) *Dynamic Asset Pricing Theory.* Princeton University Press, Princeton.

Dumas Bernard and Blaise Allaz (1996) *Financial Securities: Market Equilibrium and Pricing Methods.* Chapman and Hall, London.

Gray Stephen F. (1996) Modeling the conditional distribution of interest rates as a regime switching process. *Journal of Financial Economics* 42, 27-62.

Hamilton J.D. (1989) A new approach to the economic analysis of non- stationary time series and the business cycle. *Econometrica* 57, 357-384.

Härdle Wolfgang (1990) *Applied Non-parametric Regression.* Econometric Society Monographs, Cambridge; New York and Melbourne: Cambridge University Press.

Härdle Wolfgang, Helmut Lütkepohl and Rong Chen (1997) A review of non- parametric time series analysis. *International Statistical Review* 65(1), 49-72.

Ingersoll Jonathan E. (1987) Theory of Financial Decision Making. Rowman and Littlefield Publishers, Savage USA.

Jiang George J. (1998) Non-parametric modeling of U.S. interest rate term structure dynamics and implications on the price of derivative securities. *Journal of Financial and Quantitative Analysis* 33(4), 465-497.

Jiang George and John Knight (1997) A non-parametric approach to the estimation of diffusion processes, with an application to short term interest rate model. *Econometric Theory* 13, 615-645.

Koedijk Kees G., François G.J.A. Nissen, Peter C. Schotman and Christian C.P. Wolf (1997) The Dynamics of short-term interest rates volatility reconsidered. *European Finance Review* 1, 105-130.

Li Qi (1996) Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews* 15(3), 261-274.

Longstaff Francis A. and Eduardo S. Schwarz (1992) Interest rate volatility and the term structure: a two-factor general equilibrium model. The Journal of Finance 47(4), 1259-1282.

Lubrano Michel (1998) Smooth transition GARCH models: A Bayesian perspective. CORE DP No 9866 (http://pgiot.core.ucl.ac.be).

Lubrano, Michel (2000). Bayesian analysis of non-linear time series models with a threshold. In: William A. Barnett, David F. Hendry, Svend Hylleberg, Timo Teräsvirta, Dag Tjøstheim, and Allan Wűrtz (eds), 2000, *Non-linear Econometric Modelling*, Cambridge University Press.

Meng Xiao-Li (1994) Posterior predictive p-values. The Annals of Statistics 22(3), 1142-1160.

Nadaraya E.A. (1964). On estimating regression. *Theory of Probability and its Applications* 10, 186-190.

Parzen E. (1962) On estimation of a probability density and mode. *Annals of Mathematical Statistics* 35, 1065-1076.

Pfann Gerard A., Peter C. Schotman and Rolf Tschernig (1996) Non-linear interest rate dynamics and implications for the term structure. *Journal of Econometrics* 74, 149-176.

Pai Jeffreys and Hal Pedersen (1998) The Bayesian approach to regime switching models of the short interest rate: theory, calibration and implementation. Mimeo.

Rice J. (1984) Bandwidth choice for non-parametric regression. *Annals of Statistics*, 12, 1215-1230.

Rosenblatt M. (1956) Remarks on some non-parametric estimates of a density function. Annals of Mathematical Statistics 27, 642-669.

Rubin Donald B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* 12(4), 1151-1172.

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

Shiller Robert J. (1979) The volatility of long-term interest rates and expectations models of the term structure. Journal of Political economy 87(6), 1190-1219.

Tjøstheim D. and B. Auestadt (1994a) Non-parametric identification of non-linear time series: projection. *Journal of the American Statistical Association* 89, 1398-1409.

Tjøstheim D. and B. Auestadt (1994b) Non-parametric identification of non-linear time series: selecting significant lags. *Journal of the American Statistical Association* 89, 1410-1419.

Tong H. (1983) *Threshold Models in Non-linear Time Series Analysis*, New-York, Springer Verlag.

Vasicek O. (1977) An equilibrium characterisation of the term structure. *Journal of Financial Economics* 5, 177-188.

Watson G.S. (1964) Smooth regression analysis. *Sankhya* Series A 26, 359-372.

Appendix
Non-parametric inference for diffusion processes

It is well known from empirical experience (see e.g. Merton 1980) that in an Ito diffusion equation the volatility is easier to estimate than the drift function. This is easily seen in the following Brownian motion $d \log X_t = \mu dt + \sigma dW_t$ with μ and σ fixed. The MLE of σ is the empirical mean of the log returns $\hat{\mu} = \sum \Delta \log X_t = (\log X_T - \log X_0)/T$. Increasing the frequency of the observations leaves unchanged this estimator. In a non-parametric context, it is not possible to identify the drift function in the absence of extra information. Florens-Zmirou (1993) was the first to study non-parametric inference for Ito diffusion processes with discrete observations. She leaves the drift function unspecified and proposes an estimator for the volatility function which is equivalent to

$$\hat{\sigma}^2(r) = \frac{\sum_{i=1} nK[(r_i - r)/h](r_{i+1} - r_i)^2}{\sum_{i=1} TK[(r_i - r)/h]}, \tag{41}$$

where n is the number of equally spaced observations and T the time period of observation. We note that when $n = T$, this estimator is equivalent to (5), except for the drift correcting term.

If we suppose that the diffusion process is strictly stationary, we can use a property of stochastic differential equations in order to identify the drift and volatility functions, using the marginal and transition probabilities of r. This approach was developed both by Ait-Sahalia

(1996a,b) and by Jiang and Knight (1997). To give some intuition on how this works, let us note that a Gaussian process is entirely characterised by its mean and variance. Because its increments are Gaussian, the distributions (marginal and transition probabilities) of an Ito diffusion process are entirely characterised by their drift and volatility functions under certain regularity conditions. This can be shown (see e.g. Jiang and Knight 1997) by solving the backward and forward Kolmogorov equations which gives an exact relationship between the expression of the marginal density of the observations $f(r_t)$ and the drift and volatility functions:

$$f(r_t) = \frac{A}{\sigma^2(r)} \exp\left\{ 2 \int_{r_0}^{r_t} \frac{\mu(u)}{\sigma^2(u)} \, du \right\}, \tag{42}$$

where A is a normalising constant. Ait-Sahalia (1996b) decides to identify $\mu(r)$ by providing a parametric expression for $\mu(r)$, $\mu(r, \theta)$ and estimating non-parametricaly $f(r)$. After some manipulations, he gets the following estimator for $\sigma^2(r)$

$$\hat{\sigma}^2(r) = \frac{2}{\hat{f}(r)} \int_0^r \mu(u, \hat{\theta}) \hat{f}(u) \, du, \tag{43}$$

which is conditional on a previous estimate of θ. If the drift function is parameterised so as to be mean reverting, e.g. $\mu(r, \theta) = \beta(\alpha - r)$ in the linear case, the properties of the transitional distribution imply that $\mathrm{E}(r_t | r_{t-1}) = \alpha + e^{-\beta}(r_{t-1} - \alpha)$. Then $\hat{\theta}$ can be obtained using an ordinary regression[4]. This setting implies that $\Delta_n = 1$ and that new information is obtained via future observations together with a fixed sampling rate.
Ait-Sahalia (1996b) obviously sticks to a parametric formulation of the drift because he is interested in the pricing of interest rate derivatives securities. He has to introduce the price of risk as a parameter and for that he needs a parametric drift function. Jiang and Knight (1997) have different interests an thus object that the inference results of Ait-Sahalia (1996b) are very sensitive to the chosen parametric formulation of the drift function. They prefer to use the non-parametric estimate of $\sigma(r)$ inspired from Florens-Zmirou (1993) and given in (41). A new estimator for the drift function can then be constructed. Let us solve (42) for $\mu(r)$

$$\mu(r) = \frac{1}{2} \left[\frac{d\sigma^2(r)}{dr} + \sigma^2(r) \frac{f'(r)}{f(r)} \right]. \tag{44}$$

We can now plug $\hat{\sigma}^2(r)$, which is differentiable provided the Kernel is differentiable, into (44) and approximate $f'(r)/f(r)$ by the traditional non-parametric estimator

$$\frac{\hat{f}'(r)}{\hat{f}(r)} = \frac{\sum \frac{1}{h} K' \left(\frac{r_n - r}{h} \right)}{\sum K \left(\frac{r_n - r}{h} \right)}. \tag{45}$$

Jiang and Knight (1997) show that $\hat{\mu}(r)$ and $\hat{\sigma}^2(r)$ are consistent for a fixed T, but a decreasing sampling rate Δ_n.

[4] A non-linear drift function such as $\mu(r, \theta) = \alpha_0 + \alpha_1 r + \alpha_2 r^2 + \alpha_3 / r$ is also possible. It has simply to be estimated by non-linear least squares.