ISBA 2000, Proceedings, pp. 000–000

Bayesian Model Choice for the Short Term US Interest Rate

MICHEL LUBRANO
GREQAM-CNRS, Marseille, France and
CORE, Louvain la Neuve, Belgique.

SUMMARY

The paper 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 a nonparametric estimate of the data density. The method is applied to empirical models of the US short term interest rate. The constant elasticity of variance model of Chan, Karolyi, Longstaff and Sanders (1992) with different drift functions is confronted to various alternative parametric formulations which draw their justifications from a preliminary nonparametric analysis. The paper concludes by some implications for the term structure.

Keywords: MODEL SELECTION; POSTERIOR PREDICTIVE DENSITY; DIFFUSION EQUATIONS; INTEREST RATES.

1. INTRODUCTION

This paper is concerned with model choice and model selection for empirical models of the short term U.S. interest rate. It builds on the tradition initiated by Box (1980) and Rubin (1984) where the performance of a model is cast in terms its predictive ability to reproduce some characteristics of the observed data. The approach here is extended to measuring the predictive ability of a model to reproduce the sample distribution using a non-parametric distance between the simulated posterior predictive density and the empirical density of the observed sample.

An important and practical question is to know among which models to select. In a bond pricing model, the short term interest rate r is usually modelled by a continuous time diffusion equation having the general form

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

where W is a standard Brownian motion. The usual empirical practice reported for instance in Chan, Karolyi, Longstaff and Sanders (1992) (CKLS for now on) consists in proposing

[0]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.

a parameterisation for the drift and volatility functions $\mu(r)$ and $\sigma(r)$ and then to discretise the model. In order to obtain a tractable model for option or bond pricing, there is an incentive to restrict oneself to the class of linear drift functions while a wider choice is available for the volatility function. In order to fit better the data, CKLS have promoted the volatility function $\sigma(r) = \sigma r^{\gamma}$ with $\gamma > 1$. However, as second order stationarity is obtained only for $\gamma < 1$ when the drift is linear as shown by Broze, Scaillet and Zakoian (1995), there is a clear sign of misspecification for existing models. Ait-Sahalia (1996) claims that the linearity of the drift function is the main source of misspecification. He proposes instead to use a non-linear formulation with

$$\mu(r) = \alpha_0 + \alpha_1 r + \alpha_2 r^2 + \alpha_3/r \tag{2}$$

In this case, a negative α_2 guaranties ergodicity and second order stationarity even when $\gamma > 1$. Does this close the debate? Certainly not as for instance a switching mechanism is a credible alternative which was used e.g. in Pfann, Schotman and Tschernig (1996). We must have a tool to determine what are the potential and credible alternatives for parameterising the drift and volatility functions. In this paper, we consider classical non-parametric tools (such as data density estimation and kernel regression) to explore graphically the potential shape of drift and volatility functions suggested by the data.

The paper is organised as follows. In section 2, we present the model selection procedure. Section 3 introduces the data set and its non-parametric exploration. Section 4 analyses various parameterisations including two regime models. Section 5 examines implications for the term structure and concludes.

2. BAYESIAN MODEL EVALUATION

Suppose that we have various alternative parameterisations for the drift and volatility functions. We have to validate and compare these choices in their ability to reproduce the characteristics of the data.

2.1. *General setting*

We build on the old idea of comparing actual data y_d (a sample of size T) to hypothetical replicated samples y_{rep} (each of size T) obtained by simulating the model. If the model under inspection fits the data in a satisfactory way, the observed sample and the simulated samples should not look too different. To apply the idea, the main stream Bayesian literature (see references below) defines a statistics $g(y_d)$ to compute a tail area probability

$$\Pr\{g(y) \geq g(y_d) \,|\, M\} \tag{3}$$

which quantifies the extremeness of the observed value $g(y_d)$ with reference to the distribution of the transformation $g(.)$ of the random variable y which represents the possible outcomes of the model. A small tail area probability casts doubts on the model. For Box (1980), this probability is computed in reference to the predictive distribution attached to model M

$$p(y \,|\, M) = \int f(y \,|\, \eta, M)\, \varphi(\eta \,|\, M)\, d\eta \tag{4}$$

which represents all the possible outcomes of the data y if model M (data density and prior distribution) were true. In this writing, $f(y \mid \eta, M)$ is the data density, η the vector of parameters and $\varphi(\eta \mid M)$ the prior density. Rubin (1984) develops the posterior predictive approach arguing that if in some situations it may be interesting to operate unconditional frequency checks, in other cases it may be preferable "to regard η as a fixed feature of the replications" and then to average over the posterior distribution of η. This is the posterior predictive approach. Rubin (1984) says "posterior meaning conditional on observed values and predictive meaning the distribution of a future observable quantity". This last approach knew a significant development in the statistical literature with Meng (1994), Gelman and Meng (1996) or Gelman, Meng and Stern (1996). With the posterior predictive approach, interest is focused on the fit of the model to the observed data. Whereas with the pure predictive approach, concern is stressed on reliability of the model for every possible outcome, at the cost of specifying an informative prior for η. See Gelman, Meng and Stern (1996) for more details concerning the comparison of the prior predictive approach and the posterior predictive approach.

The posterior predictive approach is particularly convenient in a MCMC setting. Suppose that we have already obtained N draws from the posterior density. Given a draw η_i, we draw an hypothetical sample y_i^{rep} from the sampling distribution. We compute the $p-$value as the proportion of cases in which the simulated transformation $g(y_i^{rep})$ exceed the realised value $g(y_d)$

$$\frac{1}{N}\sum_{i=1}^{N} ID[g(y_i^{rep}) > g(y_d)] \tag{5}$$

where ID is the indicator function.

2.2. *Measuring distances between densities*

The choice of the transformation $g(.)$ is crucial for the method. For instance Rubin (1984) notes that the first two moments of two densities may be equal while the gap between order statistics may still be different. Consequently if $g(.)$ is the empirical mean or variance, the method will not reject the normal model when the true sampling model is Cauchy.

The transformation $g(.)$ is used to summarise the observed sample into a single number. We develop here the idea that it is better to avoid choosing a particular transformation. Consequently we have to compare directly the density of the observed sample with the density of replicated samples obtained by simulating the posterior predictive density. We must first choose a measure of discrepancy between densities and second propose a nonparametric method to estimate it. The ϕ divergence is popular in the Bayesian literature, especially the Kullback-Leibler divergence and the χ^2 distance. However both are asymmetric. If f and g are two densities to be compared, the distance between f and g is not the same as that between g and f when measured with the above two instruments. On the contrary, the Hellinger distance defined as

$$D_H = 2(1 - \int \sqrt{f(x)\, g(x)}\, dx) = 2\left(1 - \int \frac{\sqrt{f(x)g(x)}}{f(x)} dF(x)\right) \tag{6}$$

is symmetric. The usual non-parametric procedure (see e.g. Pagan and Ullah 1999) consists in approximating the Stieltjes integral $\int f(x)dF(x)$ by its Monte Carlo estimate

$\sum f(x_i)/T$ where the x_i are distributed according to $F(x)$ and using a kernel estimate for $f(x)$, $\sum K_h((x_i - x)/h)/Th$ where h is the window size and K_h a smoothing kernel. Consequently if y_d and y_{rep} are the observed and simulated data, a nonparametric estimate of D_H is given by:

$$\hat{D}_H = 2\left(1 - \frac{1}{T}\sum_i \frac{\sqrt{\sum_j K_h(y_d[i] - y_d[j]) \sum_j K_h(y_d[i] - y_{rep}[j])}}{\sum_j K_h(y_d[i] - y_d[j])}\right) \qquad (7)$$

where $y_d[i]$ indicates the ith element of y_d. The classical practice would go on by finding a normally distributed asymptotic approximation for (7) (see e.g. Li 1996). In a Bayesian framework, we want to avoid these asymptotic approximations. We can first reduce the sampling variance of (7) by replacing in (6) the Monte Carlo evaluation of the Stieltjes integral (which converges at the slow rate of $\sqrt{T}$) by the Simpson rule (for which the error can be more easily controlled)

$$\hat{D}_H = 2\left(1 - \frac{d}{3}\sum_i w_i \sqrt{\hat{f}(y_i)\hat{g}(y_i)}\right). \qquad (8)$$

where d is the size of the increment in y_i, w_i 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 and is evaluated at T equidistant points y_i starting at the minimum of the sample and ended at its maximum. The computational cost is even smaller in (8) than in (7).

Then for each of the N draws of η, we draw a replicated sample y_i^{rep} of the same size T as the observed sample y_d and we compute corresponding Hellinger distance D_{H_i} using (8). We obtain N values D_{H_i} which approximate the small sample distribution of D_H. Note that when $N \to \infty$, we tend to the exact small sample distribution of D_H when the usual classical procedure can derive only an asymptotic approximation (see e.g. Li 1996). With these N draws, we are free to compute several quantities. We can decide for instance to select the model producing the smallest average distance if we compute the mean of the distribution. Or we can compute quantiles of the distribution and select the model with the smallest 0.95% quantile. Or we can compute the posterior probability that $D_H < 0.10$ and select the model with the highest corresponding probability[1].

2.3. *How to simulate the predictive density*

We are typically in the context of a dynamic model which gives the conditional density of the observations, but not its marginal density which is defined as

$$f(y_t \,|\, \eta_i) = \int f(y_t \,|\, y_{t-1}, \eta_i)\, f(y_{t-1} \,|\, \eta_i)\, dy_{t-1} \qquad (9)$$

Suppose that we have a draw y_{t-1}^i from $f(y_{t-1} \,|\, \eta_i)$. We can immediately get a draw y_t^i from $f(y_t \,|\, \eta_i)$ by drawing it in $f(y_t \,|\, y_{t-1}, \eta_i)$, using y_{t-1}^i as the conditioning variable. Of

[1]The level 0.10 is chosen as D_H is at value in [0,2].

course the process has to be initiated. Depending on the starting point, it will converge to the true marginal distribution after a certain number of steps. We can initiate the process by drawing y_0 from the empirical distribution of the observed sample y_d. To get rid of the influence of the initial condition, we have in practice to discard the first few draws.

3. NONPARAMETRIC EVIDENCES FOR THE US TREASURY BILL RATE

We consider monthly observations of the three month U.S. Treasury bill rate between

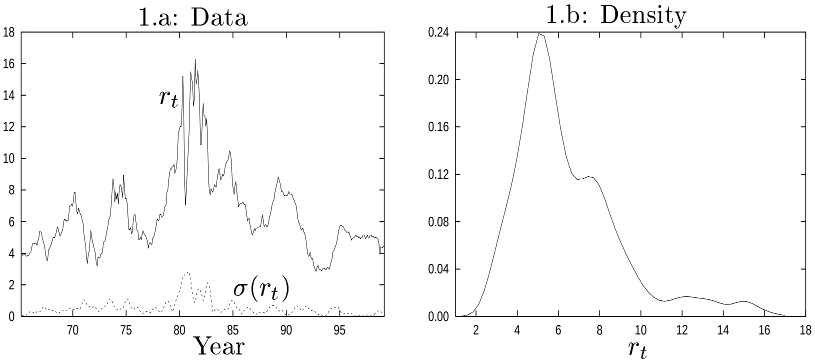

Figure 1. *Graph and density of the US short term interest rate*

1965:1 and 1999:3. Figure 1.a displays the data together with a non-parametric estimate of the volatility computed by local averaging over 13 observations weighted by the Epanechnikov kernel. 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. It led to very high interest rates which explain the very long tail of the empirical density displayed in Figure 1.b (obtained using a Gaussian kernel and Silverman's rule to determine the bandwidth). Note that this density has two modes which may suggest a mixture model. These modes are not the result of the smoothing method as they are still present when we try other values for the bandwidth selected around the optimal value.

Let us discretise the original continuous time model (1) using the Euler discretisation with a unit discretisation step

$$\Delta r_t = m(r_{t-1}) + s(r_{t-1})\epsilon_t \tag{10}$$

where ϵ_t is a normal error with zero mean and unit variance. A non-parametric regression of Δr_t over r_{t-1} gives an estimate for the discretised drift function $m(r)$. Figure 2.a shows that it is 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 very negative linear slope. There is also a slight mean reversion for rates lower than 6 per cent. This is coherent with parameterisation (2) proposed by Ait-Sahalia (1996) for the seven day Eurodollar deposit rate. But a linear drift with a break as in Pfann, Schotman and Tschernig (1996), Gray (1996) and Pai and Pedersen (1998) may be also plausible.

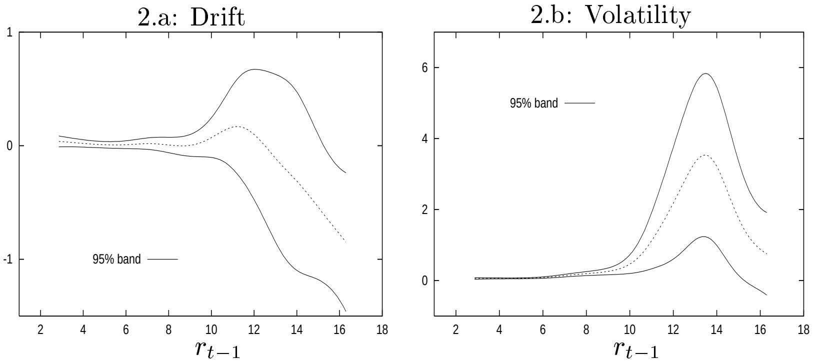

Figure 2. *Non-parametric estimates for the discretised diffusion equation*

A non-parametric regression of $(\Delta r_t)^2$ over r_{t-1} minus $\hat{m}(r)^2$ gives an estimate for the discretised squared volatility function $s^2(r)$. Figure 2.b shows the association between high interest rates and high volatility, but invalidates the CKLS parameterisation $\sigma(r) = \sigma r^\gamma$. Even if the decrease in volatility may seem poorly determined, an everlasting increase seems to be rejected. A parameterisation like $s^2(r) = \sigma/(1 + \gamma(r - c)^2)$ can mimic rather well the bell shape displayed in Figure 2.b. A Vasicek (1977) model with a two regime variance may also prove a good candidate if one does not believe in the decrease in volatility for very high rates.

4. ANOTHER BAYESIAN LOOK AT INTEREST RATE MODELS

4.1. *Single regime models*

Does the failure of usual models comes from the linearity of the drift function as suggested in Ait-Sahalia (1996)? And what is the exact role of the volatility function once the non-linearity of the drift is taken into account? The model we start from is

$$\Delta r_t = \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-1}^2 + \alpha_3/r_{t-1} + \sigma\epsilon_t\sqrt{h_t} \tag{11}$$

with various alternatives for the volatility function

- The Vasicek (1977) model which assumes a constant volatility with $h_t = 1$
- The CEV model of CKLS (1992) which may generate an explosive volatility with $h_t = r_{t-1}^{2\gamma}$
- The Cauchy volatility (CAV) function $h_t = 1/(1+\gamma(r_{t-1}-c_r)^2)$ which assumes that volatility reaches a maximum at $r_{t-1} = c$ and decreases after at a speed determined by γ.

Bayesian inference for this type of partially non-linear discretised model is detailed in Chapter 7 of Bauwens, Lubrano and Richard (1999). It relies on the Griddy Gibbs sampler of Bauwens and Lubrano (1998) for the nonlinear parameters and on conditional direct simulation for the other parameters. The prior is non-informative, except for γ where an exponential prior $\varphi(\gamma) \propto \exp(-\gamma)$ is necessary to solve an empirical identification problem between σ and γ. Table 1 reports various summary statistics on the Hellinger

Table 1. *D_H for single regime models*

Model		$h_t = 1$			$h_t = r^{2\gamma}$			$h_t = \dfrac{1+\gamma}{1+\gamma(r-c_r)^2}$	
	μ	$q_{0.95}$	$p_{0.10}$	μ	$q_{0.95}$	$p_{0.10}$	μ	$q_{0.95}$	$p_{0.10}$
N.L.	0.31	0.52	0.10	0.04	0.08	1.00	0.05	0.09	0.96
L.D.	-	-	-	0.06	0.14	0.83	0.07	0.14	0.78

Computations were based on 250 replications. 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. '-' indicates negative draws of r. μ is the empirical mean of D_H. $q_{0.95}$ is the 95% quantile and $p_{0.10}$ is the probability that $D_H < 0.10$.

distance between the data density and the posterior predictive density. The comparison between the various models indicates that

- Heteroskedasticity has definitely to be taken into account. The Vasicek model, with its constant volatility, is dominated and even fails to produce positive interest rates when the drift is linear.

- A nonlinear drift function is better in every case, giving weight to Ait-Sahalia (1996) assertion.

- When equipped with a nonlinear drift, the CEV and CAV models are comparable. There is a slight advantage for the CEV model, but at the price of an estimated γ greater than one as $\mathrm{E}(\gamma \mid y) = 1.37$. This is in contradiction with the stylised facts depicted by the non-parametric analysis.

4.2. *The impact of a two regime modelling*

The model displaying the smallest Hellinger distance is still not satisfactory as it entails an ever increasing volatility which is in contradiction with the preliminary non-parametric analysis. Threshold models can be an interesting alternative which has been tried by several authors as mentioned in section 3. We start from the general switching model

$$\begin{aligned} \Delta r_t &= \begin{cases} \alpha_{10} + \alpha_{11} r_{t-1} + \alpha_{12}/r_{t-1} + u_t & \text{if } r_{t-1} < c_1 \\ \alpha_{20} + \alpha_{21} r_{t-1} + u_t & \text{if } r_{t-1} \geq c_1 \end{cases} \\ u_t &= \sigma \epsilon_t \sqrt{h_t} \end{aligned} \tag{12}$$

This is not the canonical model of Pfann, Schotman and Tschernig (1996), as there is a non-linear term in the drift function of the first regime, $1/r_{t-1}$, non-linear term which plays the role of a penalty function when the interest rate become too small. For modelling h_t, we keep the previous formulations, but extend the Vasicek model to consider

$$\sigma^2\, h_t = \begin{cases} \sigma_1^2 & \text{if } r_{t-1} \le c_2 \\ \sigma_2^2 & \text{if } r_{t-1} > c_2 \end{cases} \tag{13}$$

where the squared volatility is still constant within a regime, but varies across regimes. There are 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.

Table 2. *D_H for two regime models*

Model	σ_1^2, σ_2^2			$h_t = r^{2\gamma}$			$h_t = \dfrac{1+\gamma}{1+\gamma(r-c_r)^2}$		
	μ	$q_{0.95}$	$p_{0.10}$	μ	$q_{0.95}$	$p_{0.10}$	μ	$q_{0.95}$	$p_{0.10}$
N.L.	0.05	0.08	0.99	0.05	0.09	0.96	0.03	0.05	1.00
L.D.	-	-	-	0.06	0.14	0.89	0.06	0.11	0.92

The introduction of a threshold mechanism changes the ordering among the models as shown in Table 2. The CAV model becomes by far the best choice and the Vasicek model is vastly improved. However, a simple linear switching mechanism is not enough as the best results are all the time obtained when a non-linearity is added in the first regime. Figure 3 compares the empirical distribution of the Hellinger distance for models with and without a threshold mechanism so that we can visualise the improvement.

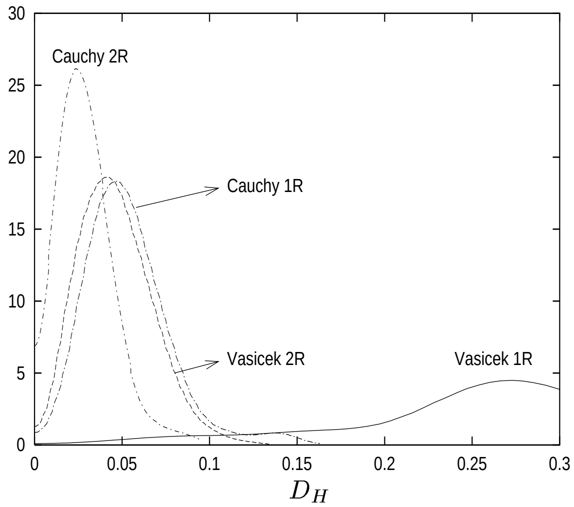

Figure 3. *The impact of a two-regime modelling*

We finally give inference results in Table 3 for the non-linear drift function case. For the switching Vasicek model, the change of regime occurs first for the volatility ($\mathrm{E}(c_2 \mid y) =$

Table 3. *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.

9.94) and then for the drift ($\mathrm{E}(c_1 \mid y) = 12.86$). For the Cauchy volatility model, the change of regime occurs roughly for the same level of r for the drift function ($\mathrm{E}(c_1 \mid y) = 12.97$ and for the volatility function ($\mathrm{E}(c_r \mid y) = 13.06$).

6. CONCLUSION

In this paper, we have proposed a new method for comparing models based on the posterior predictive density. A non-parametric measure of distance is used to compare the density of the observations to the density of artificially replicated samples drawn from the posterior predictive density. We have shown how non-parametric tools can help to select potential candidate models for the short term U.S. interest rate. In Lubrano (2000), we show how these models can be used to predict the yield curve, i.e. the relation between the riskless short term interest rate and a long term interest rate. Analytical solutions to the problem of pricing bonds exist only for very simple stochastic processes of the short term interest rate. For instance Vasicek's (1977) result is based on the mean reverting Ornstein-Uhlenbeck process which has a linear drift. This model gives a linear yield curve. A non-linear drift entails automatically a non-linear yield curve, which can be found only by simulating a transformation of the empirical model (risk neutralisation by introducing the price of risk in the drift function). As shown in Lubrano (2000), the performance of the best fitting empirical model can be seriously damaged by this operation of risk neutralisation so that the best fitting model may no longer be the best model for bond pricing.

REFERENCES

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

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

Bauwens, L. and Lubrano, M. (1998). Bayesian inference in GARCH models using the Gibbs sampler. *The Econometrics Journal*, **1**, C23–C46.

Bauwens, L., Lubrano, M. and Richard, J.F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press.

Broze, L., Scaillet, O. and Zakoian, J.M. (1995). Testing for continuous-time models of the short-term interest rate. *Journal of Empirical Finance* **2**, 199–223.

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

Gelman A. and Meng X.L. (1996) Model checking and model improvement. *Practical Markov Chain Monte Carlo* (W. Gilks, S. Richardson and S. Spiegelhalter, eds.). London: Chapman and Hall, 189–201.

Gelman A., Meng X.L. and Stern H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733-807 (with discussion).

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

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

Lubrano, M. (2000). Bayesian Non-linear Modellings of the Short Term US Interest Rate: the help of non-parametric tools. *Tech. Rep.*, No 2000/38, CORE, Universite catholique de Louvain, Belgium.

Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 1142–1160.

Pagan, A. and Ullah A. (1999). *Nonparametric Econometrics*. Cambridge University Press, Cambridge, UK.

Pai, J. and Pedersen, H. (1998). The Bayesian approach to regime switching models of the short interest rate: theory, calibration and implementation. *Tech. Rep.*, U Manitoba, Canada..

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

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

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