Predicting Short-Term Interest Rates: Does Bayesian Model Averaging Provide Forecast Improvement?

Chew Lian Chua (University of Melbourne)
Sandy Suardi (La Trobe University)
Sarantis Tsiaplias (KPMG Melbourne)

NTU International Conference on Finance (Dec 10, 2010)
Research questions

- Examines the forecasting qualities of Bayesian Moving Average (BMA) over a large set of single factor models of short-term interest rates.

- Considers five BMA methods by way of different posterior model probability constructions: BMA1 and BMA2 based on marginal likelihood over estimation period, BMA3 and BMA4 based on predictive likelihoods computed over forecasting period and a simple MA.

- Is the choice of model, one which exhibits the highest predictive likelihood, dependent on the data frequency used?
Why Model Average?

- No one model may forecast interest rate movements well in all periods.
- Parameter uncertainty. The constant variance elasticity parameter: 0.5 (Cox et al. 1985) or 1.5 (Chan et al., 1992)?
- Combining forecasts on all models and weighting them using model posterior probabilities.
- Vast use of its application: portfolio management & VaR (Pesaran and Zaffaroni, 2004), stock returns predictability (Avramov, 2002), exchange rate forecasts (Wright, 2008)
Key findings

- Pooling forecasts from different short rate models using BMA yields forecast improvements, particularly for BMA forecasts based on recent predictive likelihoods.

- Overwhelming evidence to suggest the BMA based on recent predictive likelihoods give rise to better forecasts than BMA which uses in-sample data to determine posterior model probability.

- The improvement of the predictive likelihood depends on the specification of the diffusion process (whether it allows for noise arrival process or the variance is levels dependent), which in turn relies on the frequency of short rate data examined.
Single Factor Short Rate Models

\[ dr_t = \mu(r_t)dt + \sigma(r_t)dW_t \]

- **Drift term** \( \mu(r_t) \): \( \alpha_0 + \alpha_1 r_t, \alpha_0 + \alpha_1 r_t + \alpha_2 r_t^2 + \frac{\alpha_3}{r_t} \) and others

- **Diffusion process** \( \sigma(r_t) \): \( \sigma r_t^\gamma \) (Chan et al, 1992), \( \sqrt{\beta_0 + \beta_1 r_t + \beta_2 r_t^3} \) (Ait Sahalia, 1996) or GARCH(1,1) with level term (Brenner et al, 1996)

- **Discrete form**

\[ \Delta r_t = X_{it} A_i \Delta t + \varepsilon_t \sqrt{f(Z_{it}, B_i)} \Delta t, \]

\[ E(\varepsilon_t^2 | \Omega_{t-1}) = g(W_{it}, G_i) \]
## Nested Models

<table>
<thead>
<tr>
<th>Model</th>
<th>( X_{it} A_i )</th>
<th>( f(Z_{it}, B_i) )</th>
<th>( g(W_{it}, G_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1: Ait-Sahalia (1996)</td>
<td>( \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r^2_{t-1} + \frac{\alpha_3}{r_{t-1}} )</td>
<td>( \beta_0 + \beta_1 r_{t-1} + \beta_2 r^{\beta_3}_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M2: CKLS (1992)</td>
<td>( \alpha_0 + \alpha_1 r_{t-1} )</td>
<td>( \beta_2 r^{\beta_3}_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M3: Vasicek (1977)</td>
<td>( \alpha_0 + \alpha_1 r_{t-1} )</td>
<td>( \beta_0 )</td>
<td>1</td>
</tr>
<tr>
<td>M4: BS (1980)</td>
<td>( \alpha_0 + \alpha_1 r_{t-1} )</td>
<td>( \beta_2 r^2_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M5: CIR (1985)</td>
<td>( \alpha_0 + \alpha_1 r_{t-1} )</td>
<td>( \beta_1 r_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M6: CEV Cox (1975)</td>
<td>( \alpha_0 )</td>
<td>( \beta_2 r^{\beta_3}_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M7: Merton (1973)</td>
<td>( \alpha_0 )</td>
<td>( \beta_0 )</td>
<td>1</td>
</tr>
<tr>
<td>M8: GBM</td>
<td>( \alpha_1 r_{t-1} )</td>
<td>( \beta_2 r^2_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M9: Dothan (1978)</td>
<td>0</td>
<td>( \beta_2 r^2_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M10: CIR VR (1980)</td>
<td>0</td>
<td>( \beta_2 r^{\beta_3}_{t-1} )</td>
<td>1</td>
</tr>
<tr>
<td>M11: BHK (1996)</td>
<td>( \alpha_0 + \alpha_1 r_{t-1} )</td>
<td>( r^{\beta_3}_{t-1} )</td>
<td>( \text{AGARCH}(1, 1) )</td>
</tr>
<tr>
<td>M12: BHK (1996)</td>
<td>( \alpha_0 + \alpha_1 r_{t-1} )</td>
<td>1</td>
<td>( \text{AGARCHL}(1, 1) )</td>
</tr>
</tbody>
</table>
Two Datasets

- Weekly data on annualised one-month Eurodollar deposit rates obtained from Datastream for period Feb 1975 to Dec 2008.
- Why weekly? Negligible discretization error with daily and weekly data compared to monthly (Stanton, 1997; Jone, 2003)
- General observations and summary statistics of both data: (a) levels dependence particularly during 1979-1983 Volcker monetary regime; (b) volatility clustering; (c) positive skewness (asymmetric GARCH); (d) non-normal distribution; (e) nonstationary 15-minute Eurodeposit rates but stationary for weekly data.
Bayesian Inference

For a given model $M_i$ the posterior density of $\theta_i$ is given by

$$p(\theta_i|R_T, M_i) = \frac{p(R_T|\theta_i, M_i)p(\theta_i|M_i)}{\int p(R_T|\theta_i, M_i)p(\theta_i|M_i)d\theta_i} \quad (1)$$
where the likelihood

\[ p(R_T|\theta_i, M_i) = \prod_{t=2}^{T} p(r_t|R_{T-1}, \theta_i, M_i) = \prod_{t=2}^{T} f(\mu_i, \sigma_{it}^2, v_i) \] for \( v > 2 \) assuming Student’s t distribution for \( \varepsilon_t \).

- For tractability, adopt the normal mixture representation of the Student’s t distribution in which it may be represented as a heteroskedastic normal model with

\[ p(r_t|R_{T-1}, \theta_i, M_i) = N(\mu_i, \sigma_{it}^2 \lambda_{it}) \text{ and } \lambda_{it}|v_i \sim Gamma\left(\frac{v_i}{2}, \frac{v_i}{2}\right) \]
Bayesian Inference

Prior density is decomposed as

\[ p(\theta_i|M_i) = p(A_i|M_i)p(B_i|M_i)p(G_i|M_i)p(v_i|M_i) \]

Prior for \( A_i \) follows a multivariate normal distribution, \( p(B_i) \) is given by the product of the inverse gamma density for each element of \( B_i \), \( p(G_i) \) follows a uniform distribution and \( v_i \) follows the exponential distribution of Geweke (1993).
Sampling Scheme

- Posterior density cannot be sampled from using a known distribution, thus use a combination of Gibbs Sampler and Metropolis Hastings algorithm. Sampling scheme involves iterating through the following five steps (contingent on the model):
  - Draw $A_i$ from $p(A_i|\theta_i \neq A_i, R_T, M_i)$ which is a multivariate normal distribution.
  - Draw $B_i$ from $p(B_i|\theta_i \neq B_i, R_T, M_i)$. This is an inverse gamma distribution for most of the models we consider.
  - Draw $\lambda_t$ from $p(\lambda_t|\theta_i \neq \lambda_t, r_t, M_i)$ for $t = 1, \ldots, T$. The conditional distribution of $\lambda_t$ is a gamma distribution.
  - Draw $G_i$ from $p(G_i|\theta_i \neq \gamma, R_T, M_i)$ using a MH algorithm.
  - Draw $v_i$ from $p(v_i|\theta_i \neq v_i, R_T, M_i)$ using a MH algorithm.
The posterior probability of a model $M_i$ is computed as

$$p(M_i|R_T) = \frac{p(R_T|M_i)p(M_i)}{\sum_{k=1}^{K} p(R_T|M_k)p(M_k)}$$

where $p(R_T|M_i) = \frac{1}{N-N_0} \sum_{j=N_0+1}^{N} p(R_T|\theta_{i}^{(j)}, M_i)p(\theta_{i}^{(j)})$.

The predictive likelihood for $R_{T+S} = \{r_{T+1}, ..., r_{T+S}\}$ is

$$p(R_{T+S}|R_T, M_i) = \frac{1}{N-N_0} \sum_{j=N_0+1}^{N} p(R_{T+S}|\theta_{i}^{(j)}, R_T, M_i)$$

where $\theta_{i}^{(j)}$ is obtained from posterior draws $\{\theta_{i}^{(N_0+1)}, ..., \theta_{i}^{(N)}\}$.
Model Averaging

- A model-free estimate of the predictive likelihood is given by

\[ p(R_{T+S}|R_T) = \sum_{k=1}^{K} p(R_{T+S}|R_T, M_k)p(M_k|R_T) \]

- Augment \( p(M_k|R_T) \) to allow updating at each time step so that

\[ p(M_i|R_{T+1}) = p(M_k|r_{T+1}, R_T) = \frac{p(r_{T+1}|R_T, M_i)p(M_i|R_T)}{\sum_{k=1}^{K} p(r_{T+1}|R_T, M_k)p(M_k|R_{T-1})} \tag{2} \]

- This is the recursive BMA approach adopted by Liu and Maheu (2009) to forecast realised volatility.
Model Averaging

Possible drawback of the construction of (1) and (2) is that posterior model probabilities place little emphasis on recent model performance. Model may have performed poorly in most recent $h$ time periods and yet its posterior probability is very large.

Restrict data available for computation of posterior probability to a subset of recent predictive likelihood such that

$$p(M_i|R_{T-h:T}) = \frac{p(R_{T-h:T}|M_i)p(M_i)}{\sum_{k=1}^{K} p(R_{T-h:T}|M_k)p(M_k)}$$

(3)

where $p(R_{T-h:T}|M_i) = \prod_{t=T-h}^{T} p(r_t|R_{t-1}, M_i)$. 
Empirical Application

- Divide sample into 3 periods: $T_T$ (to construct prior for the estimation period), $T_E$ (to obtain posterior density of parameter vector and in-sample comparison), $T_F$ (to conduct a real time out-of-sample forecasting)

- Weekly data: $T_T = 9/1/75-31/1/85$, $T_E = 7/2/85-29/12/05$, $T_F = 5/1/06-18/12/08$

- High frequency data: $T_T = 19/5/09$-mid day 24/7/09, $T_E = $mid day 24/7/09 - mid day 22/9/09, $T_F = $mid day 22/9/09 - mid day 29/9/09

- $N = 20,000$ posterior draws for $\theta_i$ using data spanning $T_T$ period, burn-in of $N_0 = 5,000$ draws.
Empirical Application

- Prior hyperparameters for $T_T$ period are largely uninformative with (1) individual elements of $A_i \sim N(0, 100)$, (2) $\beta_{ji} \sim \text{Inverse Gamma}(2, 1)$ (3) $v_i \sim \text{Exponential}(0.025)$.

- Generate up to $q = 8$ step ahead forecasts with the associated predictive likelihoods. At each time step, add an additional period of data to $T_E$ before re-estimating the posterior $p(\theta_i|T_E, M_i)$. Generate another $q$-step ahead forecasts and the predictive likelihoods. This recursive procedure continues till 23/10/08.

- BMA1 is obtained from equations (1) and (2) with prior model probabilities computed from the first twenty 1-step ahead predictive likelihoods for the 12 models considered. The marginal likelihoods used to construct (1) is based on the $T_E$ period data and is estimated using the Modified Harmonic Mean method of Gelfand and Dey (1994).
Empirical Application

- BMA2 is calculated the same way as BMA1 but adopts equal prior model probabilities.
- BMA3 is based on equation (3) with equal prior model probabilities for each of the 12 models. The elements of $p(R_{T-h:T} | M_i)$ covers only the predictive likelihoods over $T_F$ period. Model choices are made by reference to the cumulative predictive capacity of the models over the forecast period.
- BMA4 is akin to BMA3 but uses a rolling window of the last 20 predictive likelihoods (i.e. fix $h = 19$). Only recent forecasting performance is considered in determining each model’s weight.
- Simple MA assumes that each model is always given an equal weight irrespective of its predictive likelihood values.
M11 and M12 perform best and the performance is consistent across 8-step ahead periods.

Comparing M7 and M3 (or M6 and M2), a model with linear mean-reverting drift is preferred to a nonstationary one.

Keeping linear drift constant and allowing the elasticity of variance parameter to vary across models M2 to M5, we find CIR (1985) model performs the worst while CKLS model performs best.

Allowing for nonlinearity in the drift (M1) reduces the predictive likelihood value (when compared with M2).

Relative underperformance of M1 to M10 is associated with inability to account for periodic presence of persistently high volatility in short rate data.

News arrival process seems to be pivotal in generating higher predictive likelihoods at weekly frequency (consistent with BHK 1996).
High Frequency Data

- M5 (CIR, 1985) which is associated with the lowest predictive likelihood in weekly data is now the best performing model.

- M12 which is the second best model using low frequency data performs worse off than some of the non-GARCH models (like M2, M4, M5, M6 and M8).

- Greater predictive likelihood is determined by the specification of the elasticity of variance and not due to modelling news arrival process.

- The consensus that CEV models are inadequate in characterising the short rate may only apply to lower frequency modelling of short rate data.
BMA Results

- Evidence in favour of predictive benefits of model averaging using BMA1 and BMA2 is fairly weak. Cumulated predictive likelihoods are clearly lower than better performing models.

- This problem arise because BMA based on marginal likelihoods over the entire estimation period effectively collapses to model selection procedures; both BMA1 and BMA2 have probabilities close to unity for M12 (M1) in the low (high) frequency case. See Amisano and Geweke (2010).

- Cumulated predictive likelihoods of BMA1 and BMA2 are slightly greater than those of simple MA approach, implying construction of predictions using equal model weights is the least effective model averaging approach for short-rate data (particularly with lower frequency data).
But, in the case of BMA3 and BMA4, they do not collapse to unity for any single model.

For BMA3, probability of M11 dominates all models for weekly data (see Fig 2). But for high frequency data, using BMA3, the probability of M5 exceeds that of M11 at the start, from the middle and thereafter of the forecasting period.
BMA Results

- Using a rolling window of 20 periods, results for BMA4 are qualitatively unchanged. However, the posterior model probability plots of M6 and M8 begin to assume some importance in the BMA4 forecast, particularly with weekly data in the last one-third of the forecasting period.

- The difference in the cumulated predictive likelihoods of BMA3 and BMA4 is not significant, implying that primary contribution to forecast improvement for short rate data does not arise from using a rolling window to determine model weights.

- However, there is greater predictive usefulness of restricting the data available to posterior model probabilities to the period covering the forecast period and limiting the weight attached to the performance of the models over the estimation period.
This paper investigates the usefulness of BMA for predicting U.S. short rates observed at weekly and 15-minute frequencies. Pooling forecasts from different short rate models using BMA yields forecast improvements, particularly for BMA forecasts based on recent predictive likelihoods. Overwhelming evidence to suggest the BMA based on recent predictive likelihoods give rise to better forecasts than BMA which uses in-sample data to determine posterior model probability. Results are robust to the choice of prior (i.e. uninformative or diffuse priors) and forecast horizon considered. The improvement of the predictive likelihood depends on the specification of the diffusion process (whether it allows for noise arrival process or the variance is levels dependent), which in turn relies on the frequency of short rate data examined.