Journal of Modern Mathematics and Statistics 4 (2): 73-77, 2010
ISSN: 1994-5388

A Sequential Monte Carlo Approach for Online Stock Market Prediction Using Hidden Markov Model

[1]E. Ahani and [2]O. Abass
[1]Department of Mathematics, [2]Department of Computer Science, University of Lagos, Nigeria

Abstract: This study attempts a development of a Sequential Monte Carlo (SMC) algorithm approach of prediction based on joint probability distribution in Hidden Markov Model (HMM). SMC methods, a general class of monte carlo methods are mostly used for sampling from sequences of distributions. Simple examples of these algorithms are extensively used in the tracking and signal processing literature. Recent developments indicate that these techniques have much more general applicability and can be applied very effectively to statistical inference problems. Firstly, due to the problem involved in estimating the parameter of HMM, the HMM is now represented in a state space model and the Sequential Monte Carlo (SMC) method is used. Secondly, the researchers make the prediction using SMC method in HMM and then develop the corresponding on-line algorithm. At last, the data of daily stock prices in the banking sector of the Nigerian Stock Exchange (NSE) (price index between the years 1st January 2005 to 31st December 2008) are analyzed and experimental results reveal that the method proposed in this manner is effective.

Key words: Sequential monte carlo, hidden markov model, state-space model, stock market, techniques, algorithm

INTRODUCTION

State space or hidden Markov models are convenient means to statistically model a process that varies in time. The state space model (Doucet and Johansen, 2009) of a hidden Markov model is shown by the following two equations:

$$\text{State equation: } X_t | (X_{t-1} = x_{t-1}) \sim f(x_t | x_{t-1}) \quad (1)$$

$$\text{Observation equation: } Y_t | (X_t = x_t) \sim g(y_t | x_t) \quad (2)$$

The state variables x_t and observations y_t may be continuous-valued, discrete-valued or a combination of the two $f(x_t|x_{t-1})$ which indicates the probability density, associated with moving from x_{t-1}- x_t and $g(y_t|x_t)$ are the state (transition) and observation densities. Practically, the x's are the unseen true signals in signal processing (Liu and Chen, 1995), the actual words in speech recognition (Rabiner, 1989), the target features in a multitarget tracking problem (Avitzour, 1995; Gordon *et al*., 1993, 1995), the image characteristics in computer vision (Isard and Blake, 1996), the gene indicator in a DNA sequence analysis (Churchill, 1989) or the underlying volatility in an economical time series (Pitt and Shephard, 1997). Hidden Markov models shown the applications of dynamic state space model in DNA and protein sequence analysis (Krogh *et al*., 1994; Liu *et al*., 1997).

While, using the functions provided by C++ to expand an on-line algorithm of predicting hidden Markov model, this study takes impetus from Johansen (2009) SMCTC: Sequential Monte Carlo in C++. Further supports were derived from some results on predicted and actual data of monthly national air passengers in America. Cheng applied SMC methodology to tackle the problems of optimal filtering and smoothing in hidden Markov models. SMC have also stirred great interest in the engineering and statistical literature (Doucet *et al*., 2000) for a summary of the state of the art). Lately, by Johansen *et al*. (2008), SMC methods have been applied for resolving a marginal Maximum Likelihood problem. In Gordon *et al*. (1993), the application of SMC to optimal filtering was first offered. Here, SMC method is developed for prediction of state by estimating the probability $p(x_t | y_{1:t-1})$.

Hidden Markov model: Although, initially introduced and studied as far back as 1957 and early 1970's, the recent popularity of statistical methods of HMM is not in question. A HMM is a bivariate discrete-time process $\{X_k, Y_k\}_{k\geq 0}$ where $\{X_k\}_{k\geq 0}$ is an homogeneous Markov chain which is not directly observed but can only be observed

Corresponding Author: E. Ahani, Department of Mathematics, University of Lagos, Nigeria

through $\{Y_k\}_{k\geq 0}$ that produce the sequence of observation. $\{Y_k\}_{k\geq 0}$ is a sequence of independent random variables such that the conditional distribution of Y_k only depends on X_k The underlying Markov chain $\{X_k\}_{k\geq 0}$ is called the state. In general, the random variables X_k; Y_k can be of any dimension and of any domain such as discrete, real or complex.

The researchers collect K elements of X_k and Y_k for k = 1, 2,....,K to construct the vectors X_k and Y_k, respectively. Because of the Markov assumption, the probability of the current true state given the immediately previous one is conditionally independent of the other earlier states:

$$p(x_k | x_{k-1}, x_{k-2}, \cdots, x_0) = p(x_k | x_{k-1})$$

Similarly, the measurement at the kth time step is dependent only upon the current state, so is conditionally independent of all other states given the current state:

$$p(y_k | x_k, x_{k-1}, \cdots, x_0) = p(y_k | x_k)$$

Using these assumptions, the probability distribution over all states of the HMM can be written simply as:

$$p(x_0, \cdots, x_k, y_1, \cdots, y_k) = p(x_1)$$

$$p(y_1 | x_1) \prod_{k=2}^{K} p(x_k | x_{k-1})\, p(y_k | x_k)$$

Which is reflected graphically in Fig. 1. Given, $p(x_{k-1} | y_{k-1})$ we can find $p(x_k | y_k)$ using the following prediction and update steps:

$$\text{Prediction: } p(X_k | Y_{1:k-1}) = \int p(X_k | X_{k-1}) p(X_{k-1} | Y_{1:k-1})\, dx_{-1}$$

$$\text{Updating: } p(X_k | Y_{1:k}) = \frac{p(Y_k | X_k) p(X_k | Y_{1:k-1})}{\int p(Y_k | X_k) p(X_k | Y_{1:k-1})\, dx_k}$$

In this case, we use numerical integration which becomes computationally complex when the number of states of x_k are large. One particular Monte Carlo based approach to solve this for the HMM is the SMC.

Sequential monte carlo methods: Since their pioneering contribution in 1993 (Gordon *et al*., 1993), SMC have become a well known class of numerical methods for the solution of optimal estimation problems in non-linear non-Gaussian scenarios. The key idea of SMC method is to represent the posterior density function $p(x_{0k-1} | y_{0k-1})$ at time k-1 by samples and associated weights $\{x^{(i)}_{0:k-1}, w^{(i)}_{0:k-1} i = 1,...., N\}$ and to compute estimates based on

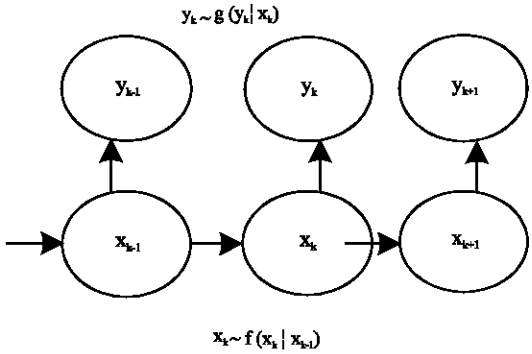

Fig. 1: Probability distribution over all state of HMM

these samples and weights. As the number of samples becomes very large, this Monte Carlo characterization develops into an equivalent representation to the functional description of the posterior probability density function (Sanjeev *et al*., 2002). If we let :

$$\{x^{(i)}_{0:k-1}, w^{(i)}_{0:k-1} i = 1,...., N\}$$

be samples and associated weights approximating the density function:

$$p(x_{0:k-1} | y_{0:k-1}) \{x^{(i)}_{0:k-1}\}_{i=1}^{N}$$

is a set of particles with associated weights;

$$\{w^{(i)}_{0:k-1}\}_{i=1}^{N}$$

with,

$$\sum_{i=1:N} w^{(i)}_{k-1} = 1$$

then the density function are approximated by:

$$p(x_{0:k-1} | y_{0:k-1}) \approx \sum_{i=1}^{N} w^{(i)}_{k-1} \delta(x_{k-1} - x^{(i)}_{k-1})$$

Where, $\delta(x)$ signifies the Dirac delta role. Y_k becomes available when a new observation arrives and the density function $p(x_k | y_k)$ are obtained recursively in two stages:

- Drawing samples $x^i_k \sim p(x_k x_{k-1})$
- Updating weight with the principle of importance sampling (Doucet *et al*., 2000; Sanjeev *et al*., 2002)

The particles are proliferated over time by Monte carlo simulation to get new particles and weights (usually as new information are received), hence forming a series of PDF approximations over time. The reason that it works can be understood from the theory of (recursive) importance sampling.

Procedural functions: This is how it works. We consider a particular algorithm for the SMC, known also as the Sampling Importance Resampling (SIR) (Gordon *et al*., 1993; Carpenter *et al*., 1999; Johansen, 2009). The algorithm can be summarized as follows: The algorithm is initiated by setting k = 1 for which we define $p(x_k|x_{k-1}) = p(x_k)$.

Prediction (for step k): Draw N samples from the distribution $p(x_k|x_{k-1}=s_{k-1}^{(i)})\forall_i$ to form the particles $\{\hat{s}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1:N}$. The weights is:

$$\tilde{w}_k^{(i)} = \frac{\hat{w}_k^{(i)}}{\sum_i \hat{w}_k^{(i)}}$$

Where, $\hat{w}_k^{(i)}$ is calculated from the conditional PDF $p(y_k|x_k = \hat{s}_k^{(i)})$, given observation Y_k:

Resample (for step k): Resample the random measure $\{\hat{s}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1:N}$ obtained in the prediction procedure to get:

$$\left\{s_k^{(i)}, \frac{1}{N}\right\}_{i=1:N}$$

which has uniform weights. The importance of the prediction step is clear by establishing the following results. Using a importance function $q(x_k|y_k)$ satisfying the property:

$$q(x_k|x_{k-1}, y_k) = q(x_k|x_{k-1}, Y_i)$$

$\{\hat{s}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1:N}$ is the random measure for estimating $p(x_k|y_k)$ where $\hat{s}_i = \lfloor \hat{s}_1^{(i)}, \cdots, \hat{s}_k \rfloor$ is the trajectory for particle i and where $\tilde{w}_k^{(i)} = \hat{w}_k(\hat{s}_k^{(i)})$ is the normalized weights of particle I at time k which can be calculated recursively. Let;

$$\hat{W}_k^{(i)} = \hat{w}_k(\hat{s}_k^{(i)})$$

According to the argument, at the kth step, the density function estimate for $p(x_k|y_k)$ is:

$$p(\hat{x}_k|y_k) = \sum_{i=1}^{N} \tilde{w}_k^{(i)}\, \delta(x_k - \hat{s}_k^{(i)})$$

After the density function $\hat{p}(x_k|y_k)$ has been estimated, the observation prediction with some samples with associated weights can be made. Accordingly, $p(\hat{y}_k|y_{k-1})$ are approximated by a new set of samples $\{\hat{y}_k^1, w_{k-1}^{(i)}\}_{i=1:N}$ and the observation prediction equation is:

$$\hat{p}(\hat{y}_k|y_k) = \sum_{i=1}^{N} \tilde{w}_k^{(i)}\, \delta(y_k - y_k^{(i)})$$

Data description: The earlier method is applied to the data sets of daily stock prices in the banking sector of the Nigerian Stock Exchange (price index between the years 1st January 2005 to 31st December 2008). Three hidden states are studied: bull, bear and even. These hidden states along with the observable sequences of large rise, small rise, no change, large drop and small drop were used to develop the hidden Markov model (Fig. 2).

The sequence of observation is obtained by subtracting the prior price from the current price and then with the percentage change gives the classification of the sequence of observation.

Let P_t be the price of an asset at time t, the daily price relative/log return is calculated $r_t = \log p_t / p_{t-1}$. Regularly, stock prices alter in stock markets as seen in the price index on Tuesday, February 5 th 2006; it fell by >100% (Fig. 3). There is no infallible system that indicates the precise movement of stock price. Instead, stock price is subjective to the influence of various factors such as company fundamentals, external factors and market behaviour. These decide the state of the

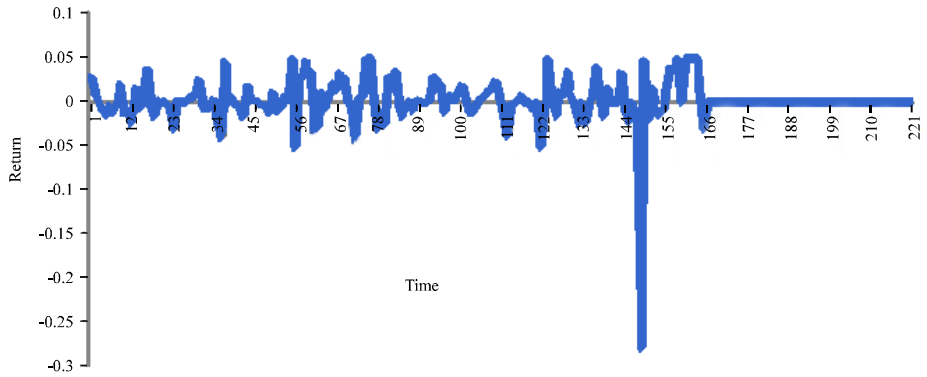

Fig. 2: Daily stock prices in the banking sector of the Nigerian Stock Exchange (price index between the years 1st January 2005 to 31st December 2008)

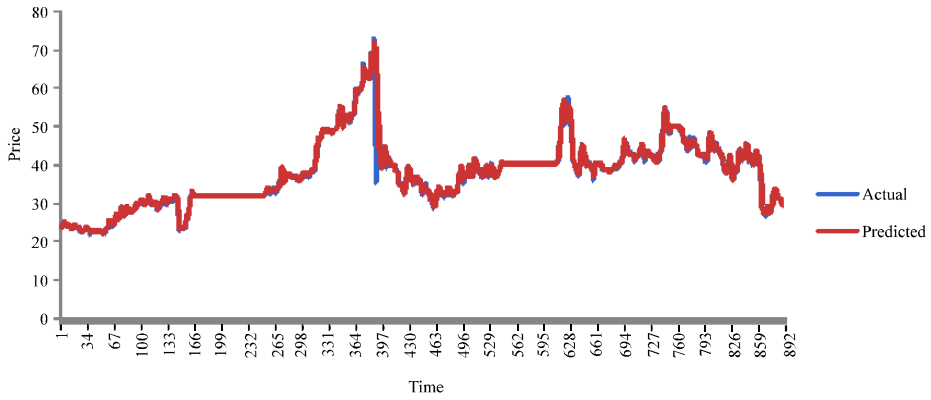

Fig. 3: Daily stock prices in the banking sector of the Nigerian Stock Exchange (red line represents predicted stock price while blue line represents actual stock price)

Table 1: Predicated daily stock price in the banking sector of the NSE

Actual	Predicted	R.E (%)
24	23.8489	0.629583
24.7	24.0614	2.585425
24.9	24.4768	1.699598
25	24.9410	0.236000
24.8	24.9793	-0.722980
24.45	24.6880	-0.973420
24.3	24.3934	-0.384360
23.99	24.0885	-0.410590
23.95	23.9330	0.070981
24.47	24.2088	1.067430
24.09	24.1513	-0.254460
23.8	23.9220	-0.512610
23.22	23.4166	-0.846680
23.6	23.4176	0.772881
23.42	23.3770	0.183604
23.6	23.4982	0.431356
24.49	24.1671	1.318497
24.3	24.3828	-0.340740
23.88	24.1404	-1.090450
23.94	24.0180	-0.325810
23.85	23.8900	-0.167710
23.86	23.8301	0.125314
23.73	23.7339	-0.016430
23	23.1971	-0.856960
22.98	22.9523	0.120540
22.99	22.8886	0.441061
23	22.9326	0.293043
23	22.9550	0.195652
23.1	23.0550	0.194805
23.2	23.1768	0.100000
23.78	23.6018	0.749369
23.7	23.7578	-0.243880
23.45	23.6338	-0.783800
23.3	23.4173	-0.503430
23.35	23.3440	0.025696
22.89	23.0174	-0.556570
22	22.2651	-1.205000
22.97	22.5771	1.710492
22.9	22.7748	0.546725
23	22.9519	0.209130
22.95	22.9895	-0.172110
22.91	22.9678	-0.252290
22.55	22.6986	-0.658980
22.95	22.8260	0.540305

Table 1: Continued

Actual	Predicted	R.E (%)
22.94	22.8994	0.176983
23	22.9894	0.046087
22.98	23.0266	-0.202790
22.94	23.0066	-0.290320
22.8	22.8641	-0.281140
22.51	22.6008	-0.403380
22.75	22.6411	0.478681
22.5	22.5232	-0.103110
22.35	22.3730	-0.102910
22.45	22.3671	0.369265
22.46	22.4187	0.183882
23.58	23.1687	1.744275
22.41	22.7752	-1.629630
23.06	22.9608	0.430182
23.7	23.5019	0.835865
24.8	24.4987	1.214919
25.68	25.5147	0.643692
25.08	25.5347	-1.813000
24.4	24.9159	-2.114340
24.7	24.7253	-0.102430
24.49	24.4938	-0.015520
24.5	24.4089	0.371837
25.03	24.7630	1.066720
25.4	25.2465	0.604331
26.24	26.0237	0.824314
27	26.8721	0.473704
27	27.2044	-0.757040
26.98	27.2338	-0.940700
26	26.5007	-1.925770
26.09	26.1648	-0.286700
26.17	26.0937	0.291555
27.39	26.8896	1.826944
28.75	28.2272	1.818435
28.98	29.0147	-0.119740
28.07	28.6229	-1.969720
27.5	27.8895	-1.416360
26.77	27.0194	-0.931640
27.5	27.1466	1.285091
28.24	27.8034	1.546034
29.22	28.8430	1.290212
28.99	29.1623	-0.594340
28.5	28.8644	-1.278600
28.31	28.5203	-0.742850
28.3	28.3238	-0.084100

Table 1: Continued

Actual	Predicted	R.E (%)
28.02	28.0612	-0.147040
28.08	27.9971	0.295228
28.05	27.9861	0.227807
27.95	27.9407	0.033274
27.91	27.9132	-0.011470
28.6	28.3646	0.823077
29.4	29.1204	0.951020
29.99	29.8659	0.413805
29.65	29.9393	-0.975720
29.75	29.9012	-0.508240
29.96	29.9926	-0.108810
29.99	30.0266	-0.122040

Mape (%); 0.068285

market which maybe in bull, even or bear state. It grows along time through different market state which are hidden states. The state of the market can be a Markovian process and are modeled in HMM.

Experimental outcome: Utilizing the functions provided by C++, this study develops an on-line algorithm of predicting hidden Markov model according to the analysis of section 2 and 3. It draws motivation from Johansen (2009) SMCTC: Sequential Monte Carlo in C++. The on-line prediction using SMC begins with states producing signals that follow the normal distribution. The number of hidden states in the Markov chain are defined as Bull (state 1), Even (state 2) and Bear (state 3). Figure 3 shows the predicted and actual daily stock prices and Table 1 shows predicted representational prices of the NSE and predicted errors.

The stock price is modeled in HMM and prediction is made based on available observations. Due to the strong statistical foundation of HMM and SMC method, it can predict similar pattern proficiently (Fig. 3). From Table 1, we can observe that the Mean Absolute Percentage Error (MAPE) is 0.068. Hence, the predictive exactness is high.

CONCLUSION

In this study, an online, sequential Monte Carlo method is applied for prediction in Hidden Markov model. A C++ (Sequential Monte Carlo in C++) template class library (Johansen, 2009) enabled us to develop an online, sequential Monte Carlo for the prediction.

The basic theory of HMM and SMC method was introduced. Then we approximated the density function with a set of random samples with associated weights.

Lastly, the data sets of daily stock prices in the banking sector of the Nigerian Stock Exchange (price index between the years 1st January 2005 to 31st) are analyzed, and experimental results revealed that the online algorithm is effective.

REFERENCES

Avitzour, D., 1995. Stochastic simulation Bayesian approach to multitarget tracking. Proc. IEEE Radar Sonar Navigation, 142: 41-44.

Carpenter, J., P. Clifford and P. Fearnhead, 1999. Improved particle filter for nonlinear problems. IEE Proc. Radar, Sonar Navigation, 146: 2-2.

Churchill, G.A., 1989. Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol., 51: 79-94.

Doucet, A. and A.M. Johansen, 2009. A Tutorial on Particle Filtering and Smoothing: Fifteen years Later. In: Handbook of Nonlinear, Crisan, D. and B. Rozovsky (Eds.). Oxford University Press, Oxford, pp: 4-6.

Doucet, A., J.F.G. de Freitas and N.J. Gordon, 2000. Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York.

Gordon, N.J., D.J. Salmond and A.F.M. Smith, 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proc. F, 140: 107-113.

Gordon, N.J., D.J. Salmon and C.M. Ewing, 1995. Bayesian state estimation for tracking and guidance using the bootsrap filter. J. Guidance Control Dyn., 18: 1434-1443.

Isard, M. and A. Blake, 1996. Contour Tracking by Stochastic Propagation of Conditional Density. In: Computer Vision, Buxton and R. Cipolla (Eds.). Springer Berlin, New York, pp: 343-356.

Johansen, A.M., 2009. Sequential monte carlo in C^{++}. J. Statistical Software, 30: 1-41.

Johansen, A.M., A. Doucet and M. Davy, 2008. Particle methods for maximum likelihood parameter estimation in latent variable models. Statistics Comput., 18: 47-57.

Krogh, A., M. Brown, S. Mian, K. Sjolander and D. Haussler, 1994. Proetein modeling using hidden markov models. J. Mol. Biol., 235: 1501-1531.

Liu, J.S. and R. Chen, 1995. Blind deconvolution via sequential imputations. J. Am. Statistical Assoc., 90: 567-576.

Liu, J.S., A.F. Neuwald and C.E. Lawrence, 1997. Markov structures in biological sequence alignment. Tech. Rep. Stanford Universty.

Pitt, M.K. and N. Shephard, 1997. Filtering via simulation: Auxiliar particle filters. J. Am. Statistical Assoc., 94: 590-599.

Rabiner, L.R., 1989. A tutirial on hidden markov models and selected applications in speech recognition. Proc. IEEE., 77: 257-286.

Sanjeev, M.A., S. Maskell, N. Gordon and T. Clapp, 2002. A tutorial on particle filter for on-line non-linear/non-guassian bayesian tracking. IEEE Trans. Signal Process., 50: 174-188.