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

By
Ahani, E., and Abass O

Abstract
This paper attempts an application of a sequential Monte Carlo (SMC) algorithm to perform online 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 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. Secondly, we make the prediction using SMC method by developing 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.

Keywords: Sequential Monte Carlo, Hidden Markov Model, State-Space model, Stock Market

Authors:
Ahani, E., briggetk2002ng@yahoo.com
Abass O, olabass@unilag.edu.ng

Editor: Tsionas, Efthymios G., tsionas@aueb.gr
1.0 Introduction

Hidden Markov models are convenient means to statistically model a process that varies in time. The state space model (Doucet & Johansen, 2008) of a hidden Markov model is represented by the following two equations:

(State equation) \[ X_t \mid (X_{t-1} = x_{t-1}) \sim f(x_t \mid x_{t-1}) \] \[ (1) \]

(Observation equation) \[ Y_t \mid (X_t = x_t) \sim g(y_t \mid x_t) \] \[ (2) \]

The state variables \( x_t \) and observations \( y_t \) may be continuous-valued, discrete-valued, or a combination of the two, \( f(x_t \mid x_{t-1}) \) which indicates the probability density, associated with moving from \( x_{t-1} \) to \( x_t \) and \( g(y_t \mid x_t) \) 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; Gordon, et al 1995), the image characteristics in computer vision (Isard & Blake 1996), the gene indicator in a DNA sequence analysis (Churchill 1989), or the underlying volatility in an economical time series (Pitt & Shephard 1997). Hidden Markov Models represent 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 paper 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 Zhang et al. (2007). Cheng et al (2003) 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 (see Doucet, et al (2000) for a summary of the state-of-the-art). Lately, in (Johansen, 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 \( p(x_t \mid y_{t-1}) \).
The outline of the remainder of this article is arranged thus: Section 2 delved into a concise analysis of basic theory of HMM and SMC method respectively. The demonstration of the approximation of the probability density function using SMC method in HMM occupies Section 3; the corresponding on-line algorithm is also developed here. In section 4, the application of the estimation algorithm to the data 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) are analyzed and results generated. Section 5 concludes the work.

2.0 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 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.

We collect \( K \) elements of \( X_k \) and \( Y_k \) for \( k = 1, 2, \cdots, K \) 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:
\[ p(x_0, \ldots, x_k, y_1, \ldots, y_k) = p(x_1) p(y_1 \mid x_1) \prod_{k=2}^{K} p(x_k \mid x_{k-1}) p(y_k \mid x_k) \]

which is reflected graphically below

\[ y_k \sim g(y_k \mid x_k) \]

\[ x_k \sim f(x_k \mid x_{k-1}) \]

Given \( p(x_{k-1} \mid y_{k-1}) \), we can find \( p(x_k \mid y_k) \) using the following prediction and update steps

\[
p(X_k \mid Y_{1:k-1}) = \int p(X_k \mid X_{k-1}) p(X_{k-1} \mid Y_{1:k-1}) dx_{k-1}
\]

updating: \( p(X_k \mid Y_{1:k}) = \frac{p(Y_k \mid X_k)p(X_k \mid Y_{1:k-1})}{\int p(Y_k \mid X_k)p(X_k \mid Y_{1:k-1}) dx_k} \)

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.

**2.1 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 \( p(x_{0:k-1} \mid y_{0:k-1}) \) at time \( k-1 \) by samples and associated weights, \( \{x_{0:k-1}, w_{0:k-1}^{(i)} | i = 1, \cdots, N \} \) and to compute estimates based on 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 \( \{x_{0:k-1}^{(i)}, w_{0:k-1}^{(i)} \mid i = 1, \cdots, N \} \) be samples and associated weights approximating the density function \( p(x_{0:k-1} \mid y_{0:k-1}) \), \( \{x_{0:k-1}^{(i)} \mid i = 1, \cdots, N \} \) is a set of particles with associated weights \( \{w_{0:k-1}^{(i)} \mid i = 1, \cdots, N \} \) with \( \sum_{i=1}^{N} w_{k-1}^{(i)} = 1 \), then the density function are approximated by

\[
p(x_{0:k-1} \mid y_{0:k-1}) \approx \sum_{i=1}^{N} w_{k-1}^{(i)} \delta(x_{k-1} - x_{k-1}^{(i)})
\]

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

1. drawing samples \( x_k^i \sim p(x_k \mid x_{k-1}) \).  
2. updating weight with the principle of importance sampling. (For details of SMC, see Doucel et al, 2000, Sanjeev, 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.

### 3.0 Procedural Functions

This is how it works. We consider a particular algorithm for the SMC, known also as the sampling importance resampling (SIR) (Gordon, 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 \( p(x_k \mid x_{k-1}) = p(x_k) \)

**Prediction (for step k):**

Draw \( N \) samples from the distribution \( p(x_k \mid x_{k-1} = s_{k-1}^{(i)}) \forall i \) to form the particles \( \{s_k^{(i)}, w_k^{(i)} \mid i = 1, \cdots, N \} \). The weights is \( \tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_i w_k^{(i)}} \) where \( \hat{w}_k^{(i)} \) is calculated from the conditional PDF \( p(y_k \mid x_k = \hat{s}_k^{(i)}) \), given observation \( Y_k \),
Resample (for step k):
Resample the random measure \( \{ \tilde{s}_k^{(i)}, \tilde{w}_k^{(i)} \}_{i=1}^N \) obtained in the prediction procedure to get
\[
\left\{ \tilde{s}_k^{(i)}, \frac{1}{N} \right\}_{i=1}^N
\]
which has uniform weights.

The importance of the prediction step is clear by establishing the following results.
Using an importance function \( q(x_k \mid y_k) \) satisfying the property
\[
q(x_k \mid x_{k-1}, y_k) = q(x_k \mid x_{k-1}, Y_k)
\]
\( \{ \tilde{s}_k^{(i)}, \tilde{w}_k^{(i)} \}_{i=1}^N \) is the random measure for estimating \( p(x_k \mid y_k) \), where \( \tilde{s}_i = [\tilde{s}_1^{(i)}, \ldots, \tilde{s}_k] \) is the trajectory for particle \( i \) and where \( \tilde{w}_k^{(i)} = \tilde{w}_k^{(i)}(\tilde{s}_k^{(i)}) \) is the normalized weights of particle \( i \) at time \( k \) which can be calculated recursively.

Let \( \tilde{w}_k^{(i)} = \tilde{w}_k^{(i)}(\tilde{s}_k^{(i)}) \) According to our argument, at the \( k^{th} \) step, the density function estimate for \( p(x_k \mid y_k) \) is
\[
p(\hat{x}_k \mid y_k) = \sum_{i=1}^N \tilde{w}_k^{(i)}(\tilde{s}_k^{(i)} - \hat{s}_k^{(i)})
\]
After the density function \( \hat{p}(x_k \mid y_k) \) has been estimated, the observation prediction \( \hat{y}_k \) with some samples with associated weights can be made. Accordingly, \( p(\hat{y}_k \mid y_{k-1}) \) are approximated by a new set of samples \( \{ \tilde{y}_k^{(i)}, \tilde{w}_{k-1}^{(i)} \}_{i=1}^N \) and the observation prediction equation is
\[
\hat{p}(\hat{y}_k \mid y_k) = \sum_{i=1}^N \tilde{w}_k^{(i)}(y_k - \tilde{y}_k^{(i)}).
\]

3.2 Data description
The above 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). (see www.cashcraft.com/pricemovement.asp). 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.

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 $r_t$ is calculated as $r_t = \log \frac{P_t}{P_{t-1}}$.

Fig. 1.1 *daily stock prices in the banking sector of the Nigerian Stock Exchange (price index between the years 1st January 2005 to 31st December 2008)*

Regularly, stock prices alter in stock markets as seen in the price index on Tuesday, February 5th 2006; it fell by more than 100% (see Fig. 2). 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 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.

4.0 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 & 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);
Fig. 2 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, refer to Fig. 2. From Table 1, we can observe that the mean absolute Percentage error (MAPE) is 0.068. Hence, the predictive exactness is high.

**Fig 2.** 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: Predicted daily stock price in the banking sector of the NSE

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Prob</th>
<th>R.E(%)</th>
<th>MAPE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>23.8489</td>
<td>0.49487</td>
<td>0.629583</td>
<td>0.068285</td>
</tr>
<tr>
<td>24.7</td>
<td>24.0614</td>
<td>0.3666</td>
<td>2.585425</td>
<td></td>
</tr>
<tr>
<td>24.9</td>
<td>24.4768</td>
<td>0.577916</td>
<td>1.699598</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>24.941</td>
<td>0.682734</td>
<td>0.236</td>
<td></td>
</tr>
<tr>
<td>24.8</td>
<td>24.9793</td>
<td>0.679789</td>
<td>-0.72298</td>
<td></td>
</tr>
<tr>
<td>24.45</td>
<td>24.688</td>
<td>0.701073</td>
<td>-0.97342</td>
<td></td>
</tr>
<tr>
<td>24.3</td>
<td>24.3934</td>
<td>0.703905</td>
<td>-0.38436</td>
<td></td>
</tr>
<tr>
<td>23.99</td>
<td>24.0885</td>
<td>0.706452</td>
<td>-0.41059</td>
<td></td>
</tr>
<tr>
<td>23.95</td>
<td>23.933</td>
<td>0.684708</td>
<td>0.070981</td>
<td></td>
</tr>
<tr>
<td>24.47</td>
<td>24.2088</td>
<td>0.705746</td>
<td>1.06743</td>
<td></td>
</tr>
<tr>
<td>24.09</td>
<td>24.1513</td>
<td>0.674846</td>
<td>-0.25446</td>
<td></td>
</tr>
<tr>
<td>23.8</td>
<td>23.922</td>
<td>0.685373</td>
<td>-0.51261</td>
<td></td>
</tr>
<tr>
<td>23.22</td>
<td>23.4166</td>
<td>0.70195</td>
<td>-0.84668</td>
<td></td>
</tr>
<tr>
<td>23.6</td>
<td>23.4176</td>
<td>0.698992</td>
<td>0.772881</td>
<td></td>
</tr>
<tr>
<td>23.42</td>
<td>23.377</td>
<td>0.673668</td>
<td>0.183604</td>
<td></td>
</tr>
<tr>
<td>23.6</td>
<td>23.4982</td>
<td>0.682281</td>
<td>0.431356</td>
<td></td>
</tr>
<tr>
<td>24.49</td>
<td>24.1671</td>
<td>0.677155</td>
<td>1.318497</td>
<td></td>
</tr>
<tr>
<td>24.3</td>
<td>24.3828</td>
<td>0.68817</td>
<td>-0.34074</td>
<td></td>
</tr>
<tr>
<td>23.88</td>
<td>24.1404</td>
<td>0.699763</td>
<td>-1.09045</td>
<td></td>
</tr>
<tr>
<td>23.94</td>
<td>24.018</td>
<td>0.685449</td>
<td>-0.32581</td>
<td></td>
</tr>
<tr>
<td>23.85</td>
<td>23.89</td>
<td>0.673404</td>
<td>-0.16771</td>
<td></td>
</tr>
<tr>
<td>23.86</td>
<td>23.8301</td>
<td>0.682303</td>
<td>0.125314</td>
<td></td>
</tr>
<tr>
<td>23.73</td>
<td>23.7339</td>
<td>0.683259</td>
<td>-0.01643</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>23.1971</td>
<td>0.668425</td>
<td>-0.85696</td>
<td></td>
</tr>
<tr>
<td>22.98</td>
<td>22.9523</td>
<td>0.69233</td>
<td>0.12054</td>
<td></td>
</tr>
<tr>
<td>22.99</td>
<td>22.8886</td>
<td>0.678409</td>
<td>0.441061</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>22.9326</td>
<td>0.674473</td>
<td>0.293043</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>22.955</td>
<td>0.686205</td>
<td>0.195652</td>
<td></td>
</tr>
<tr>
<td>23.1</td>
<td>23.055</td>
<td>0.694669</td>
<td>0.194805</td>
<td></td>
</tr>
<tr>
<td>23.2</td>
<td>23.1768</td>
<td>0.688846</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>23.78</td>
<td>23.6018</td>
<td>0.700951</td>
<td>0.749369</td>
<td></td>
</tr>
<tr>
<td>23.7</td>
<td>23.7578</td>
<td>0.670416</td>
<td>-0.24388</td>
<td></td>
</tr>
<tr>
<td>23.45</td>
<td>23.6338</td>
<td>0.675382</td>
<td>-0.7838</td>
<td></td>
</tr>
<tr>
<td>23.3</td>
<td>23.4173</td>
<td>0.691672</td>
<td>-0.50343</td>
<td></td>
</tr>
<tr>
<td>23.35</td>
<td>23.344</td>
<td>0.704657</td>
<td>0.025696</td>
<td></td>
</tr>
<tr>
<td>22.89</td>
<td>23.0174</td>
<td>0.682789</td>
<td>-0.55657</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>22.2651</td>
<td>0.682968</td>
<td>-1.205</td>
<td></td>
</tr>
<tr>
<td>22.97</td>
<td>22.5771</td>
<td>0.702312</td>
<td>1.710492</td>
<td></td>
</tr>
<tr>
<td>22.9</td>
<td>22.7748</td>
<td>0.687292</td>
<td>0.546725</td>
<td></td>
</tr>
</tbody>
</table>
5.0 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.

REFERENCE


