

11-1-2011

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


Ahani E. Bridget

University of Lagos, bridgetk2002ng@yahoo.com

O. Abass

University of Lagos, Nigeria, Africa, olabass@unilag.edu.ng

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Bridget, Ahani E. and Abass, O. (2011) "A Sequential Monte Carlo Approach for Online Stock Market Prediction Using Hidden Markov Models," *Journal of Modern Applied Statistical Methods*: Vol. 10: Iss. 2, Article 25.

Available at: <http://digitalcommons.wayne.edu/jmasm/vol10/iss2/25>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized administrator of DigitalCommons@WayneState.

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

Ahani E. Bridget O. Abass
University of Lagos,
Nigeria, Africa

A sequential Monte Carlo (SMC) algorithm prediction approach is developed based on joint probability distribution in hidden Markov Models (HMM). SMC methods, a general class of Monte Carlo methods, are typically used for sampling from sequences of distributions and simple examples of these algorithms are found extensively throughout 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. Due to the problem involved in estimating the parameter of HMM, the HMM is represented in a state space model and the sequential Monte Carlo (SMC) method is used. Predictions are made using the SMC method in HMM and the corresponding on-line algorithm is developed. Daily stock price data from the banking sector of the Nigerian Stock Exchange (NSE) (price index between the years 1 January 2005 to 31 December 2008) are analyzed; experimental results reveal that the method proposed is effective.

Key words: Sequential Monte Carlo, hidden Markov model, state-space model, stock market.

Introduction

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

the state equation,

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

and the 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} to 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 & 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).

Using the functions provided by C++ to expand an on-line algorithm for predicting a hidden Markov model, this article utilizes Johansen (2009) SMCTC: Sequential Monte Carlo in C++. Further supports were derived from results on predicted and actual data of

Ahani Bridget is a Lecturer in the Department of Mathematics. Email her at: bridgetk2002ng@yahoo.com. O. Abass is a Professor in the Department of Computer Science. Email him at: olabass@unilag.edu.ng.

SEQUENTIAL MONTE CARLO APPROACH USING HIDDEN MARKOV MODELS

monthly national air passengers in America (Zhang, et al., 2007). Cheng, et al. (2003) applied SMC methodology to the problems of optimal filtering and smoothing in hidden Markov models and SMC have also stirred great interest in the engineering and statistical literature (see Doucet, et al., 2000, for a summary). SMC methods have been applied for resolving a marginal Maximum Likelihood problem (Johansen, 2008) and Gordon, et al. (1993) applied SMC to optimal filtering. Herein the SMC method is developed for prediction of state by estimating the probability $p(x_t|y_{1-t-1})$.

Hidden Markov Models (HMM)

Initially introduced and studied as far back as 1957 and into the early 1970's, HMM statistical methods have enjoyed more recent popularity. An HMM is a bivariate discrete-time process $\{X_k, Y_k\}_{k \geq 0}$ where $\{X_k\}_{k \geq 0}$ is a homogeneous Markov chain that is not directly observed, it can only be observed through $\{Y_k\}_{k \geq 0}$ that produces the observation. $\{Y_k\}_{k \geq 0}$, which 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 and Y_k can be of any dimension and of any domain, such as discrete, real or complex. K elements of X_k and Y_k for $k = 1, 2, \dots, K$ are collected to construct the vectors X_k and Y_k , respectively. Due to the Markov assumption, the probability of the current true state given the immediately previous one is conditionally independent of the other earlier states:

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

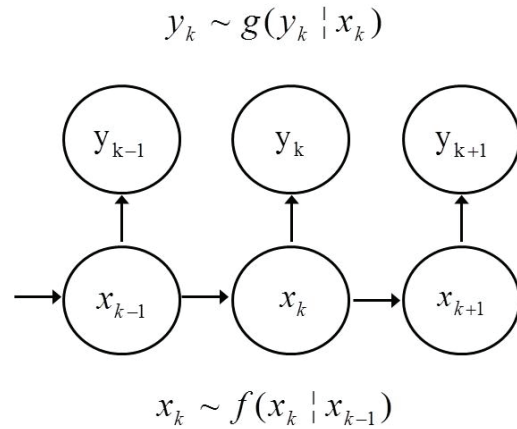
Similarly, the measurement at the k^{th} time step is dependent only upon the current state, thus it is conditionally independent of all other states given the current state:

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

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

$$p(x_0, \dots, x_k, y_1, \dots, y_k) = p(x_1)p(y_1|x_1) \prod_{k=2}^K p(x_k|x_{k-1})p(y_k|x_k)$$

which is reflected graphically as:



Given $p(x_{k-1} | y_{k-1})$, $p(x_k | y_k)$ can be found using the following prediction and update steps:

Prediction

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

Update

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

In this case numerical integration is used, 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 Sequential Monte Carlo Method (SMC).

Sequential Monte Carlo Methods (SMC)

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 main idea of the SMC method is to represent the posterior density function $p(x_{0:k-1} | y_{0:k-1})$ at time $k-1$ by samples and associated weights, $\{x_{0:k-1}^{(i)}, w_{0:k-1}^{(i)} | i=1, \dots, N\}$ 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 $\{x_{0:k-1}^{(i)}, w_{0:k-1}^{(i)} | i=1, \dots, N\}$ are samples and associated weights approximating the density function, then $p(x_{0:k-1} | y_{0:k-1})$, $\{x_{0:k-1}^{(i)}\}_{i=1}^N$ is a set of particles with associated weights $\{w_{0:k-1}^{(i)}\}_{i=1}^N$ with $\sum_{i=1:N} w_{k-1}^{(i)} = 1$, and the density function are approximated by:

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

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

1. Drawing samples $x_k^i \sim p(x_k | x_{k-1})$,

and

2. Updating the weight with the principle of importance sampling. (For details on SMC, see Doucel, et al., 2000; Sanjeev, 2002).

The particles are proliferated over time by Monte Carlo simulation to obtain 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.

Methodology

Procedural Functions

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 $p(x_k | x_{k-1}) = p(x_k)$ is defined.

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 weight 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 obtain $\left\{s_k^{(i)}, \frac{1}{N}\right\}_{i=1:N}$ 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 $\hat{s}_i = [\hat{s}_1^{(i)}, \dots, \hat{s}_k^{(i)}]$ is the trajectory for particle i and where $\tilde{w}_k^{(i)} = \hat{w}_k(s_k^{(i)})$ 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 k^{th} step, the density function estimate for $p(x_k | y_k)$ is

SEQUENTIAL MONTE CARLO APPROACH USING HIDDEN MARKOV MODELS

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

After the density function $\hat{p}(x_k | y_k)$ has been estimated, the observation prediction \hat{y}_k 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:

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

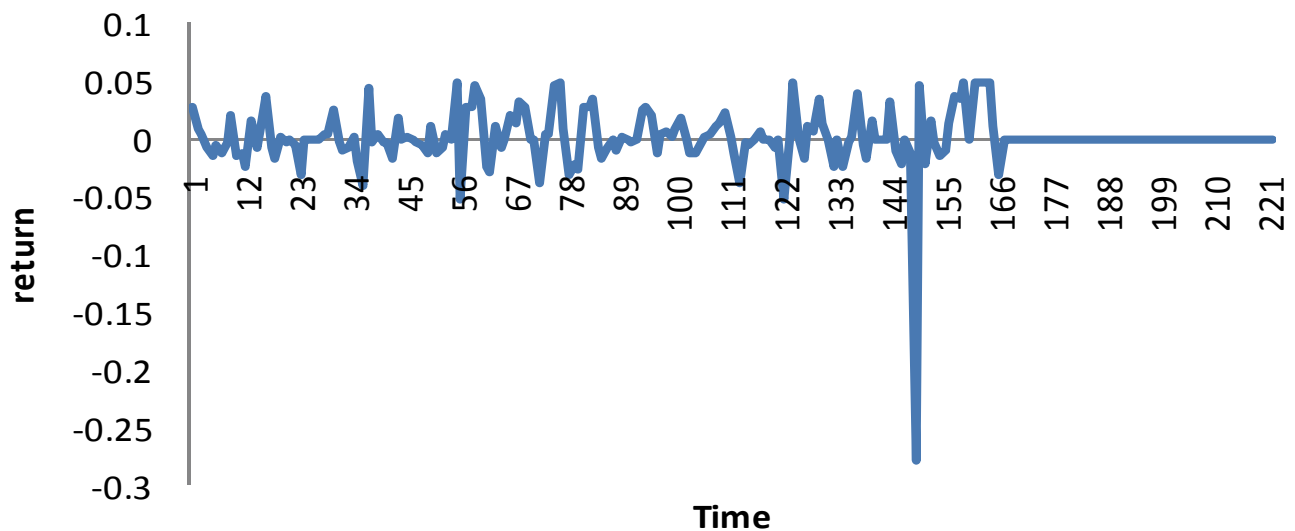
Data Description

The above method is applied to the data sets of daily stock prices in the banking sector of the Nigerian Stock Exchange for price indices between the years 1 January 2005 to 31 December 2008 (see www.cashcraft.com/pricemovement.asp and Figure 1). 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, the percentage change then gives the classification of the sequence of observation, where P_t is the price of an asset at time t , and the daily price relative/log return is calculated as $r_t = \log p_t / p_{t-1}$.

Stock prices regularly alter in stock markets as observed in the price index on Tuesday, 5 February 2006; it fell by more than 100% (see Figure 2). No infallible system exists 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 behavior. These decide the state of the market which may be in bull or bear state. It grows along time through different market states, which are hidden states. The state of the market can be a Markovian process and is modeled in HMM.

Figure 1: Daily Stock Prices in the Banking Sector of the Nigerian Stock Exchange (Price Index between the Years 1 January 2005 to 31 December 2008)



Results

Using the functions provided by C++, this study develops an on-line algorithm of predicting hidden Markov model (Johansen, 2009). The on-line prediction using SMC begins with states producing signals that follow the normal distribution. The numbers of hidden states in the Markov chain are defined as bull (state 1), even (state 2) and bear (state 3). Figure 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 the HMM and SMC methods, the model can predict similar patterns proficiently (see Figure 2). Table 1 shows that the mean absolute percentage error (MAPE) is 0.068, hence, the predictive exactness is high.

Conclusion

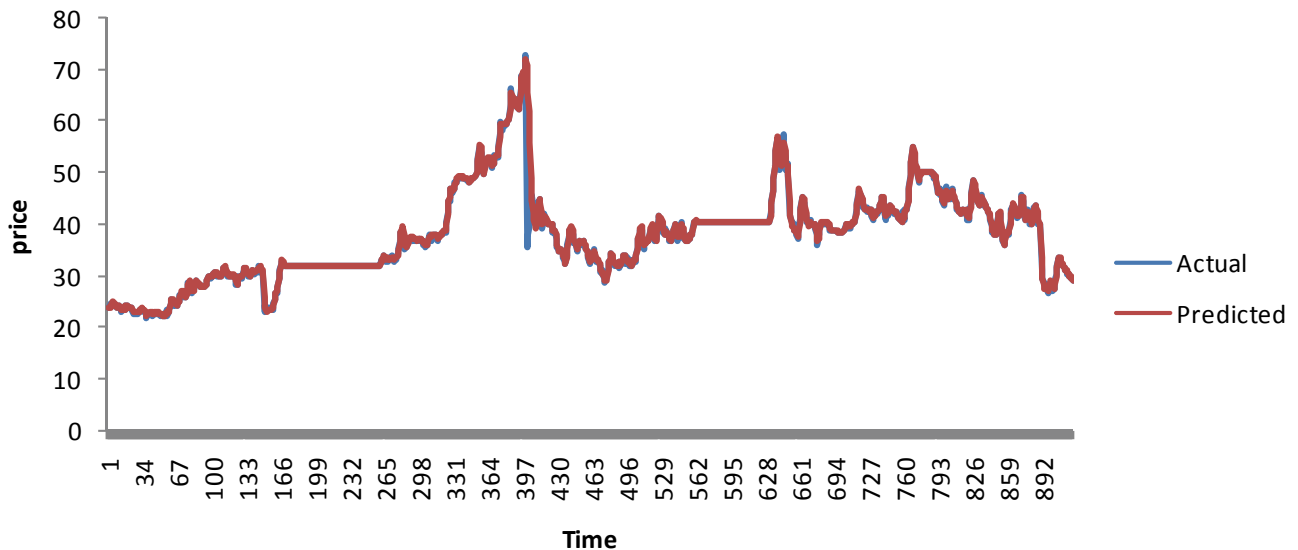
An online sequential Monte Carlo method is used to predict a hidden Markov model. A C++ (Sequential Monte Carlo in C++) template class

library (Johansen, 2009) enabled the development of an online, sequential Monte Carlo for prediction. HMM and SMC method were introduced and the density function with a set of random samples with associated weights was approximated. Lastly, the data sets of daily stock prices in the banking sector of the Nigerian Stock Exchange were analyzed and experimental results revealed that the online algorithm is effective.

References

- Avitzour, D. (1995). A Stochastic Simulation Bayesian Approach to Multitarget Tracking. *Proceedings of the IEEE on Radar, Sonar and Navigation*, 142, 41-44.
- Carpenter, J. Clifford, P., & Fearnhead, P. (1999). Improved Particle Filter for Nonlinear problems. *Proceedings of the IEEE on Radar, Sonar and Navigation*, 146, (1), 2-7.
- Doucet, A., & Johansen, A. M. (2008). A Tutorial on Particle Filtering and Smoothing. In *Oxford Handbook of Nonlinear Filtering*, D. Crisan & B. Rozovsky, Eds. Oxford University Press.

Figure 2: Daily Stock Prices in the Banking Sector of the Nigerian Stock Exchange (Red line represents predicted stock price, Blue line represents actual stock price)



SEQUENTIAL MONTE CARLO APPROACH USING HIDDEN MARKOV MODELS

Table 1: Predicted Daily Stock Price in the Banking Sector of the NSE

Actual	Predicted	R.E(%)	MAPE(%)	Actual	Predicted	R.E(%)
24	23.8489	0.629583	0.068285	22.75	22.6411	0.478681
24.7	24.0614	2.585425		22.5	22.5232	-0.10311
24.9	24.4768	1.699598		22.35	22.373	-0.10291
25	24.941	0.236		22.45	22.3671	0.369265
24.8	24.9793	-0.72298		22.46	22.4187	0.183882
24.45	24.688	-0.97342		23.58	23.1687	1.744275
24.3	24.3934	-0.38436		22.41	22.7752	-1.62963
23.99	24.0885	-0.41059		23.06	22.9608	0.430182
23.95	23.933	0.070981		23.7	23.5019	0.835865
24.47	24.2088	1.06743		24.8	24.4987	1.214919
24.09	24.1513	-0.25446		25.68	25.5147	0.643692
23.8	23.922	-0.51261		25.08	25.5347	-1.813
23.22	23.4166	-0.84668		24.4	24.9159	-2.11434
23.6	23.4176	0.772881		24.7	24.7253	-0.10243
23.42	23.377	0.183604		24.49	24.4938	-0.01552
23.6	23.4982	0.431356		24.5	24.4089	0.371837
24.49	24.1671	1.318497		25.03	24.763	1.06672
24.3	24.3828	-0.34074		25.4	25.2465	0.604331
23.88	24.1404	-1.09045		26.24	26.0237	0.824314
23.94	24.018	-0.32581		27	26.8721	0.473704
23.85	23.89	-0.16771		27	27.2044	-0.75704
23.86	23.8301	0.125314		26.98	27.2338	-0.9407
23.73	23.7339	-0.01643		26	26.5007	-1.92577
23	23.1971	-0.85696		26.09	26.1648	-0.2867
22.98	22.9523	0.12054		26.17	26.0937	0.291555
22.99	22.8886	0.441061		27.39	26.8896	1.826944
23	22.9326	0.293043		28.75	28.2272	1.818435
23	22.955	0.195652		28.98	29.0147	-0.11974
23.1	23.055	0.194805		28.07	28.6229	-1.96972
23.2	23.1768	0.1		27.5	27.8895	-1.41636
23.78	23.6018	0.749369		26.77	27.0194	-0.93164
23.7	23.7578	-0.24388		27.5	27.1466	1.285091
23.45	23.6338	-0.7838		28.24	27.8034	1.546034
23.3	23.4173	-0.50343		29.22	28.843	1.290212
23.35	23.344	0.025696		28.99	29.1623	-0.59434
22.89	23.0174	-0.55657		28.5	28.8644	-1.2786
22	22.2651	-1.205		28.31	28.5203	-0.74285
22.97	22.5771	1.710492		28.3	28.3238	-0.0841
22.9	22.7748	0.546725		28.02	28.0612	-0.14704
23	22.9519	0.20913		28.08	27.9971	0.295228
22.95	22.9895	-0.17211		28.05	27.9861	0.227807
22.91	22.9678	-0.25229		27.95	27.9407	0.033274
22.55	22.6986	-0.65898		27.91	27.9132	-0.01147
22.95	22.826	0.540305		28.6	28.3646	0.823077
22.94	22.8994	0.176983		29.4	29.1204	0.95102
23	22.9894	0.046087		29.99	29.8659	0.413805
22.98	23.0266	-0.20279		29.65	29.9393	-0.97572
22.94	23.0066	-0.29032		29.75	29.9012	-0.50824
22.8	22.8641	-0.28114	29.96	29.9926	-0.10881	
22.51	22.6008	-0.40338	29.99	30.0266	-0.12204	

- Churchill, G. A. (1989). Stochastic Models for Heterogeneous DNA Sequences. *Bulletin of Mathematical Biology*, 51, 79-94.
- Gordon, N. J., Salmond, D. J., & Smith, A. (1993). Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEEE Proceedings on Radar Signal Process*, 140(2), 107-113.
- Gordon, N. J., Salmon, D. J., & Ewing, C. M. (1995). Bayesian State Estimation for Tracking and Guidance Using the Bootstrap Filter. *Journal of Guidance, Control and Dynamics*, 18, 1434-1443.
- Isard, M., & Blake, A. (1996). Contour Tracking by Stochastic Propagation of Conditional Density. In *Computer Vision*, Buxton & R. Cipolla, Eds. New York: Springer.
- Doucet, A., de Freitas, J. F. G., & Gordon, N. J. (2000). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.
- Johansen, A. M., Doucet, A., & Davy, M. (2008). Particle methods for Maximum Likelihood Parameter Estimation in Latent Variable Models. *Statistics and Computing*, 18(1):47-57.
- Johansen, A. M. (2009). Sequential Monte Carlo in C++. *Journal of Statistical Software*, 30(6), <http://www.jstatsoft.org/>.
- Krogh, A., Brown, M., Mian, S., Sjolander, K., & Haussler, D. (1994). Protein Modeling Using Hidden Markov Models. *Journal of Molecular Biology*, 235, 1501-1531.
- Liu, J. S., & Chen, R. (1995). Blind Deconvolution via Sequential Imputations. *Journal of the American Statistical Association*, 90, 567-576.
- Liu, J. S., Neuwald, A. F., & Lawrence, C. E. (1997). *Markov Structures in Biological Sequence Alignment*. Technical Report, Stanford University.
- Pitt, M. K., & Shephard, N. (1997). *Filtering via simulation: Auxiliary particle filters*. www.nuff.ox.ac.uk/users/shephard.
- Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77, 257-286.
- Sanjeev, A., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filter for on-line non-linear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50, 174-188.