SISR based Hidden State Estimation of HMMs with transition density function

Longteng Li, Chengwen Zhu, Xiaoyan Cai, Chi Zhang, Chuizhen Zeng
Wuhan Ordnancy Non-commissioned Officer Academy of PLA
Wuhan, China
e-mail:chengwen_zhu@yeah.net

***Abstract*—Traditional Viterbi algorithm cannot be generally effective. Regarding the hidden state estimates of HMM as a Bayes filtering problem, the Sequential Importance Sampling with Resampling algorithm could get an approximate of its Bayes estimates. Its performance reached or even exceeds the Viterbi algorithm while lower dependence on the model, having a wider range of adaptation.**

Keywords- HMM; MAP; SISR

I. INTRODUCTION

HMM (Hidden Markov Models), which were brought forward by Baum and others in the late sixties of the twentieth century, are the most successful statistical modeling ideas that have came up in the last forty years. It has been widely used in many different areas such as speech recognition, anomaly detection and computational biology. The use of hidden (or unobservable) status makes the model generic enough to handle a variety of complex real-world time series, while the relatively simple prior dependence structure still allows for the use of efficient computational procedures.

Theoretically speaking, HMM need address three issues: identification problems, hidden state estimation and parameter estimation problems. They are issues form the theoretical basis of HMM, and are often inseparable in practice. Hidden state estimates of HMM is that to get the best estimates of the hidden state based on observations and model parameters. The classical algorithm is Viterbi algorithm, besides, the maximum a posteriori based algorithm, the Bayesian estimates based algorithm, the filtering and interpolation based algorithm Etc. are also commonly used. However, the form of HMMs is large difference in different application fields, and the algorithms may be quite different in the form of expression. Reference [1] use transition kernel to define a HMM which can contain most applications of HMM, called HMM with transition kernel, if its transition kernel has a density function, we treat it as HMM with transition density function.

This dissertation take full advantage of the special probability structure of HMM, derived the analytical expression of the MAP estimation of hidden state of HMM with transition density function. When the value space of hidden state is limited or the HMM is equivalent to a linear state space model, we are able to get the MAP estimation analytically. But in most cases, we couldn't get the MAP estimation analytically, or the complexity of calculation is too high. Therefore, we regard hidden state estimation problems of HMM as Bayes filtering problems, and Sequential Importance Sampling with Resampling will be used to obtain approximations of its Bayes Solution.

II. HMM WITH TRANSITION DENSITY FUNCTION

Let $(\mathbb{X},\mathcal{X})$ and $(\mathbb{Y},\mathcal{Y})$ be two measurable spaces. Q is a Markov kernel on $(\mathbb{X},\mathcal{X})$, G is a transition kernel from $(\mathbb{X},\mathcal{X})$ to $(\mathbb{Y},\mathcal{Y})$, ν is a probability measure on $(\mathbb{X},\mathcal{X})$, T is a Markov transition kernel contented formula(1). Then The Markov chain $\{(X_k,Y_k)\}_{k\geq 0}$ with transition kernel T and initial distribution $\nu\otimes G$ is called a hidden Markov model, simply referred to as HMM.

$$T[(x,y),C]=\iint_D Q(x,dx')G(x',dy'),\ (x,y)\in\mathbb{X}\times\mathbb{Y},C\in\mathcal{X}\otimes\mathcal{Y} \quad (1)$$

The integration region in formula (1) is:

$$D=C\cap\{(x,y):G(x,\{y\})\neq 0\}.$$

If there exists a probability measure μ on $(\mathbb{Y},\mathcal{Y})$, a probability measure λ on $(\mathbb{X},\mathcal{X})$, such that $\mu\ll\lambda$ and $\forall x\in\mathbb{X}$, $G(x,\cdot)\ll\mu$, $Q(x,\cdot)\ll\lambda$, then the transition kernel T must have a density function and there must exists transition density function $q(x,\cdot)$ and $g(x,\cdot)$ that $\forall A\in\mathcal{X}$, $\forall B\in\mathcal{Y}$, $Q(x,A)=\int_A q(x,x')\lambda(dx')$, $G(x,B)=\int_B g(x,y)\mu(dy)$,and the transition kernel T can be written as:

$$T[(x,y),C]=\iint_D q(x,x')g(x',y')\lambda\otimes\mu(d(x',y')) \quad (2)$$

In formula (2) $(x,y)\in\mathbb{X}\times\mathbb{Y},C\in\mathcal{X}\otimes\mathcal{Y},D=C\cap\{(x,y):g(x,y)\neq 0\}$.

$t[(x,y),(x',y')]\triangleq q(x,x')g(y,y')$ is called the transition density function of T. If the transition kernel of a HMM has transition density function, said the HMM has transition density function. This dissertation only discusses HMMs which has a transition density function, and will no longer special instructions below.

III. MAP ESTIMATION OF HMM HIDDEN STATE

The core of make judgments about HMM hidden state is obtain the joint conditional distribution of it at the condition of given observations. $\forall 0\leq k\leq l$ and $n>0$, definition $\phi_{k:l|n}(y_{0:n},\cdot)$ as the conditional distribution of $X_{k:l}$ given $\{Y_{0:n}=y_{0:n}\}$. For HMMs with transition density function, $\phi_{k:l|n}(y_{0:n},\cdot)$ is the transition kernel from $\mathbb{Y}^{n+1}$ to $\mathbb{X}^{l-k+1}$, and it must has a transition density function $\varphi_{k:l|n}(x_{k:l}\mid y_{0:n})$. Accordingly, we get the maximum a posteriori estimation of a HMM hidden state:

Single point optimal:

$$\hat{x}_k=\arg\max_{x_k}\varphi_{k|n}(x_k\mid y_{0:n}),\ k=0,\cdots,n \quad (3)$$

Path optimal:

$$\hat{x}_{0:n} = \arg\max_{x_{0:n}} \varphi_{o:n|n}(x_{0:n} \mid y_{0:n}) \quad (4)$$

The commonly used Viterbi algorithm is based on the path optimal principles of formula (4). When $\mathbb{X}$ is limited or the HMM is equivalent to a linear state-space model, it is easy to get the MAP estimation of formula (3) and (4) analytically. But normally, HMM is equivalent to a non-linear and non-Gaussian state space model, the analytical solution of formula (3) and (4) is not able to achieve. One solution is to adopt the approximate calculation; another solution is to abandon the MAP criterion. This dissertation chooses the second, use the SISR algorithm to get the hidden state's Bayes estimate of HMM.

IV. SISR BASED SINGLE POINT OPTIMA ESTIMATES OF HMM HIDDEN STATES

According to the definition of HMM with transition density function, likelihood of the observation $\{Y_{0:n} = y_{0:n}\}$ is:

$$L(y_{0:n}) = \int \cdots \int \pi_0(x_0) g(x_0, y_0) \prod_{i=1}^{n} q(x_{i-1}, x_i) g(x_i, y_i) \lambda^{\otimes n+1}(dx_{0:n}) \quad (5)$$

In formula (5), $\lambda^{\otimes n+1}$ is the product measure of $(n+1)$ th λ.

According to Theorem 3.1 of literature [2], if the observations $Y_{0:n} = y_{0:n}$ make $L(y_{0:n}) > 0$, then $\forall f \in \mathcal{F}_b(\mathbb{X}^{n+1})$,

$$\phi_{o:n|n}(y_{0:n}, f(x_{0:n})) = L(y_{0:n})^{-1} \int \cdots \int f(x_{0:n}) \pi_0(x_0) g(x_0, y_0) \prod_{i=1}^{n} q(x_{i-1}, x_i) g(x_i, y_i) \lambda^{\otimes n+1}(dx_{0:n})$$

The corresponding density function is:

$$\varphi_{o:n|n}(x_{0:n} \mid y_{0:n}) = L(y_{0:n})^{-1} \pi_0(x_0) g(x_0, y_0) \prod_{i=1}^{n} q(x_{i-1}, x_i) g(x_i, y_i) \quad (6)$$

Follow formula (6),

$$\varphi_{o:n|n}(x_{0:n} \mid y_{0:n}) \propto \pi_0(x_0) g(x_0, y_0) \prod_{i=1}^{n} q(x_{i-1}, x_i) g(x_i, y_i),$$

$$\varphi_{0|0}(x_0 \mid y_0) = L(y_0)^{-1} \pi_0(x_0) g(x_0, y_0) \propto \pi_0(x_0) g(x_0, y_0).$$

The first step of SISR is to obtain a number of random samples of the posterior density function $\varphi_{o:n|n}(x_{0:n} \mid y_{0:n})$. Foremost, you need select a proposal distribution, it's distribution density $\rho(x_{0:n} \mid y_{0:n})$ has the following form:

$$\rho(x_{0:n} \mid y_{0:n}) = \rho(x_n \mid y_n, x_{n-1}) \rho(x_{0:n-1} \mid y_{0:n-1}) = \rho(x_0 \mid y_0) \prod_{k=1}^{n} \rho(x_k \mid y_k, x_{k-1})$$

If the $X_{0:n-1}^{(i)}$ sampling from $\rho(x_{0:n-1} \mid y_{0:n-1})$ and $X_n^{(i)}$ from $\rho(x_n \mid y_n, X_{n-1}^{(i)})$, the $X_{0:n}^{(i)} = (X_{0:n-1}^{(i)}, X_n^{(i)})$ is just the sample of $\rho(x_{0:n} \mid y_{0:n})$. According to formula (6),

$$L(y_{0:n}) \varphi_{o:n|n}(x_{0:n} \mid y_{0:n}) = L(y_0) \varphi_{o|o}(x_0 \mid y_0) \prod_{k=1}^{n} q(x_{k-1}, x_k) g(x_k, y_k).$$

Therefore, the weight of the sample $X_{0:n}^{(i)}$ from $\rho(x_{0:n} \mid y_{0:n})$ is:

$$w_n^{(i)} \triangleq \frac{L(y_{0:n}) \varphi_{o:n|n}(X_{0:n}^{(i)} \mid y_{0:n})}{\rho_{o:n|n}(X_{0:n}^{(i)} \mid y_{0:n})} = w_{n-1}^{(i)} \frac{q(X_{n-1}^{(i)}, X_n^{(i)}) g(X_n^{(i)}, y_n)}{\rho(X_n^{(i)} \mid y_n, X_{n-1}^{(i)})}.$$

Using the independent and identically distributed samples $X_{0:n}^{(i)}$, $i = 1, \cdots, N$ of $\rho(x_{0:n} \mid y_{0:n})$, we can estimate the conditional density $\varphi_{o:n|n}(x_{0:n} \mid y_{0:n})$, thereby given estimate of the amount of various posteriori. So, the key is finding a suitable proposal distribution and sampling from it.

Suppose our goal is $\hat{\varphi}_{o:n|n}(x_{0:n} \mid y_{0:n})$, the optimal proposal distribution shall be that the weight variance of moment k is minimum.

At the condition of know $X_{0:k-1}^{(i)}$ and $Y_{0:k} = y_{0:k}$, the variance of weight $\omega_k^{(i)}$ is:

$$\begin{aligned}
&Var_{\rho(X_k^{(i)} \mid y_k, X_{k-1}^{(i)})} \omega_k^{(i)} \\
&= \left(\omega_{k-1}^{(i)}\right)^2 \left[E_{\rho(X_k^{(i)} \mid y_k, X_{k-1}^{(i)})} \left(\frac{q(X_{k-1}^{(i)}, x_k) g(x_k, y_k)}{\rho(x_k \mid y_k, X_{k-1}^{(i)})} \right)^2 - \left(E_{\rho(X_k^{(i)} \mid y_k, X_{k-1}^{(i)})} \frac{q(X_{k-1}^{(i)}, x_k) g(x_k, y_k)}{\rho(x_k \mid y_k, X_{k-1}^{(i)})} \right)^2 \right] \\
&= \left(\omega_{k-1}^{(i)}\right)^2 \left[\int \frac{\left(q(X_{k-1}^{(i)}, x_k) g(x_k, y_k)\right)^2}{\rho(x_k \mid y_k, X_{k-1}^{(i)})} \lambda(dx_k) - \left(\int q(X_{k-1}^{(i)}, x_k) g(x_k, y_k) \lambda(dx_k) \right)^2 \right].
\end{aligned}$$

If $\rho(x_k \mid y_k, X_{k-1}^{(i)}) = \frac{q(X_{k-1}^{(i)}, x_k) g(x_k, y_k)}{\int q(X_{k-1}^{(i)}, x_k) g(x_k, y_k) \lambda(dx_k)}$, then $Var_{\rho(X_k^{(i)} \mid y_k, X_{k-1}^{(i)})} \omega_k^{(i)} = 0$, so the optimal proposal distribution is:

$$\rho(x_0 \mid y_0) = \varphi_{0|0}(x_0 \mid y_0) \propto \pi_0(x_0) g(x_0, y_0),$$

$$\rho(x_k \mid y_k, X_{k-1}^{(i)}) = \frac{q(X_{k-1}^{(i)}, x_k) g(x_k, y_k)}{\int q(X_{k-1}^{(i)}, x_k) g(x_k, y_k) \lambda(dx_k)} \propto q(X_{k-1}^{(i)}, x_k) g(x_k, y_k)$$

The weights

$$\omega_0^{(i)} = \int \pi_0(x_0) g(x_0, y_0) dx_0,$$

$$\omega_k^{(i)} = \omega_{k-1}^{(i)} \int q(X_{k-1}^{(i)}, x_k) g(x_k, y_k) \lambda(dx_k).$$

However, in the implementation of the process of the SIS algorithm, with the increase of k, weights of the vast majority of particles will tend to 0. These particles are almost zero contribution to the estimator. To overcome this degradation phenomenon, Gordon proposed resampling technology, to reduce the particles with small weight and copy the particles with large weight.

Define $N_{eff} = \left[\sum_{i=1}^{N} (W_k^{(i)})^2 \right]^{-1}$, when $W_k^{(1)} = \cdots = W_k^{(N)} = \frac{1}{N}$, N_{eff} reaches its maximum N when one of $W_k^{(1)}, \cdots, W_k^{(N)}$ equal to 1 and others equal to 0, N_{eff} reaches its minimum 1. Therefore, N_{eff} can be used as the measure of valid samples. Given beforehand a threshold value N_{thres}, if $N_{eff} < N_{thres}$, execute the re-sampling process. Summarized as Sequential Importance Sampling with resampling algorithm (SISR):

(i) For $k = 0$

Sampling

Get independent samples $\tilde{X}_0^{(i)}$, $i = 1, \cdots, N$ from $\rho(x_0 \mid y_0)$

Calculate the weights $w_0^{(i)} = \frac{\pi_0(\tilde{X}_0^{(i)}) g(\tilde{X}_0^{(i)}, y_0)}{\rho(\tilde{X}_0^{(i)} \mid y_0)}$, and the normalized weights $\tilde{W}_0^{(i)} = \frac{w_0^{(i)}}{\sum_{i=1}^{N} w_0^{(i)}}$, $i = 1, \cdots, N$.

Calculate $N_{eff} = \left[\sum_{i=1}^{N} (\tilde{W}_0^{(i)})^2 \right]^{-1}$.

Resampling

If $N_{eff} \geq N_{thres}$, do not resampling, $W_0^{(i)} = \tilde{W}_0^{(i)}$, $X_0^{(i)} = \tilde{X}_0^{(i)}$, $i = 1, \cdots, N$.

Else, Get independent samples $I_0^{(i)}, i = 1, \cdots, N$ from $\{1, \cdots, N\}$ makes $P(I_0^{(i)} = l) = \tilde{W}_0^{(l)}, l = 1, \cdots, N$.

set $W_0^{(i)} = w_0^{(i)} = \frac{1}{N}, i = 1, \cdots, N$, update the tracks as $X_0^{(I_0^{(i)})}$, $i = 1, \cdots, N$.

(ii) For $k=1,2,\cdots,n$

Sampling

Get independent samples $\tilde{X}_k^{(i)}$ from $\rho(x_k \mid y_k, X_{k-1}^{(i)})$

calculate the weights

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{q(X_{k-1}^{(i)}, \tilde{X}_k^{(i)}) g(\tilde{X}_k^{(i)}, y_k)}{\rho(\tilde{X}_k^{(i)} \mid y_k, X_{k-1}^{(i)})},$$

and the normalized weights $\tilde{W}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^N w_k^{(i)}}$, $i=1,\cdots,N$

calculate $N_{eff} = \left[\sum_{i=1}^N (\tilde{W}_k^{(i)})^2\right]^{-1}$.

Resampling

When $k=n$ or $N_{eff} \geq N_{thres}$, update the tracks as $X_{0:k}^{(i)} = (X_{0:k-1}^{(i)}, \tilde{X}_k^{(i)})$, weights as $W_k^{(i)} = \tilde{W}_k^{(i)}$, $i=1,\cdots,N$

Else, get independent samples $I_k^{(i)}, i=1,\cdots,N$ from $\{1,\cdots,N\}$ makes $P(I_k^{(i)}=l) = \tilde{W}_k^{(l)}, l=1,\cdots,N$.

Update the tracks as $X_{0:k}^{(i)} = (X_{0:k-1}^{(I_k^{(i)})}, \tilde{X}_k^{(I_k^{(i)})})$, weights as $W_k^{(i)} = w_k^{(i)} = \frac{1}{N}, i=1,\cdots,N$.

SISR algorithm samples obtained estimates of the joint posterior density

Use the samples $X_{0:n}^{(i)}$, $i=1,\cdots,N$ of SISR algorithm, we obtained the estimates of $\varphi_{o:n|n}(x_{0:n} \mid y_{0:n})$:

$$\hat{\varphi}_{o:n|n}(x_{0:n} \mid y_{0:n}) = \sum_{i=1}^N W_n^{(i)} \delta_{X_{0:n}^{(i)}}(x_{0:n}).$$

According to formula (6)

$$\begin{aligned}
&\varphi_{k|n}(x_k \mid y_{0:n}) \\
&= L(y_{0:n})^{-1} \int \cdots \int \pi_0(x_0) g(x_0, y_0) \prod_{i=1}^n q(x_{i-1}, x_i) g(x_i, y_i) \lambda^{\otimes n}(dx_{0:k-1}, dx_{k+1:n}) \\
&= \varphi_{k|k}(x_k \mid y_{0:k}) \int \varphi_{k+1|n}(x_{k+1} \mid y_{0:n}) \times \frac{q(x_k, x_{k+1})}{\int q(x_k, x_{k+1}) \varphi_{k|k}(x_k \mid y_{0:k}) \lambda(dx_k)} \lambda(dx_{k+1}).
\end{aligned}$$

Therefore, when $k<n$, the estimates of $\varphi_{k|n}(x_k \mid y_{0:n})$ is:

$$\begin{aligned}
&\hat{\varphi}_{k|n}(x_k \mid y_{0:n}) \\
&= \hat{\varphi}_{k|k}(x_k \mid y_{0:k}) \int \hat{\varphi}_{k+1|n}(x_{k+1} \mid y_{0:n}) \times \frac{q(x_k, x_{k+1})}{\int q(x_k, x_{k+1}) \hat{\varphi}_{k|k}(x_k \mid y_{0:k}) \lambda(dx_k)} \lambda(dx_{k+1}) \\
&= \sum_{i=1}^N \tilde{W}_k^{(i)} \left(\sum_{j=1}^N W_{k+1|n}^{(j)} \frac{q(x_k, \tilde{X}_{k+1}^{(j)})}{\sum_{l=1}^N \tilde{W}_k^{(l)} q(\tilde{X}_k^{(l)}, \tilde{X}_{k+1}^{(j)})} \right) \delta_{\tilde{X}_k^{(i)}}(x_k).
\end{aligned}$$

Assume $\hat{\varphi}_{k|n}(x_k \mid y_{0:n}) = \sum_{i=1}^N W_{k|n}^{(i)} \delta_{\tilde{X}_k^{(i)}}(x_k)$, (7)

Then $W_{n|n}^{(i)} = W_n^{(i)}$,

$$W_{k|n}^{(i)} = \tilde{W}_k^{(i)} \sum_{j=1}^N W_{k+1|n}^{(j)} \frac{q(\tilde{X}_k^{(i)}, \tilde{X}_{k+1}^{(j)})}{\sum_{l=1}^N \tilde{W}_k^{(l)} q(\tilde{X}_k^{(l)}, \tilde{X}_{k+1}^{(j)})}. \quad (8)$$

Firstly, $W_{n|n}^{(i)} = W_n^{(i)} = \tilde{W}_n^{(i)}$. For $k=n-1,\cdots,0$, obtain the weight $W_{k|n}^{(i)}$ by formula (8) recursively; obtain the estimate of $\varphi_{k|n}(x_k \mid y_{0:n})(k=0,\cdots,n)$ by formula (7). Then, we can get the estimates of HMM hidden state by formula (3).

Written as algorithm is:

SISR based single point optima estimates of HMM hidden states

（1） obtain the estimates of $\varphi_{k|n}(x_k \mid y_{0:n})(k=0,\cdots,n)$ by SISR algorithm:

$$\hat{\varphi}_{k|n}(x_k \mid y_{0:n}) = \sum_{i=1}^N W_{k|n}^{(i)} \delta_{\tilde{X}_k^{(i)}}(x_k).$$

（2） obtain the estimates of hidden states:

$$\hat{x}_k = \arg\max_{x_k} \varphi_{k|n}(x_k \mid y_{0:n}), \quad k=0,\cdots,n.$$

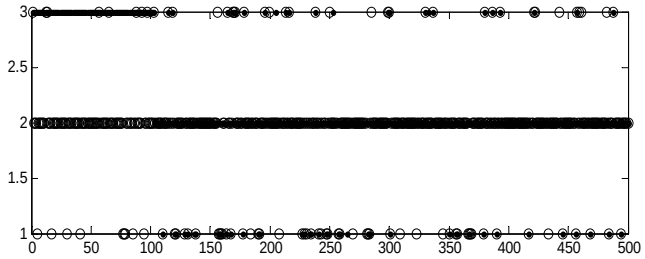
V. SIMULATION TESTS AND CONCLUSIONS

Generate a group of hidden status and corresponding observations with a length of 500 ($n=499$) from a HMM whose $\mathbb{X}=\{1,2,3\}$, $\pi_0 = (0.1 \ 0.8 \ 0.1)$, $g(x_i, y_i) = \frac{1}{\sqrt{2\pi}\sigma_{x_i}} \exp\left\{-\frac{(y_i-\mu_{x_i})^2}{2\sigma_{x_i}^2}\right\}$, $\mu = (-3 \ 0 \ 3)$,

$\Sigma = (\sqrt{2} \ 1 \ \sqrt{2})$, $Q = \begin{pmatrix} 0.2 & 0.7 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.7 & 0.2 \end{pmatrix}$.

Assume Unknown the hidden states, need to get its estimation by the observations and model parameters.

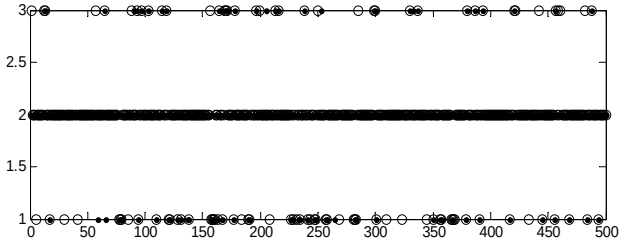
The single point optima estimations of $X_k(k=0,\cdots,n)$ is shown below,



Figer 1 Single point optima estimations of hidden state

There are 123 error estimates In figer1, where " o " signify the real hidden states, " • "signify the single point optima estimations of formula (3).

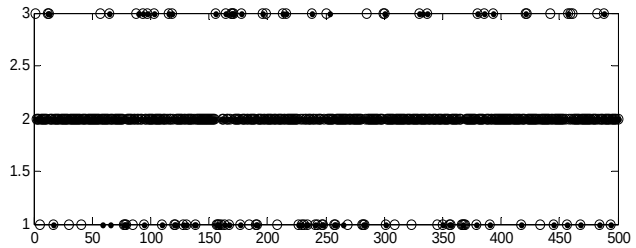
The Viterbi estimations of $X_k(k=0,\cdots,n)$ is shown below:



Figer 2 Viterbi estimations of hidden state

There are 39 error estimates In figer2, where " o " signify the real hidden states, " • "signify the Viterbi estimations.

The SISR based single point optima estimations of $X_k(k=0,\cdots,n)$ is shown below:



Figer 3 SISR based single point optima estimations

There are 38 error estimates In figer3, where " o " signify the real hidden states, " • "signify the SISR based single point optima estimations.

The results above showing, the introduction of SISR algorithm has greatly improved the accuracy of single point optima estimations, reached the level of the Viterbi algorithm. Despite the cost of the calculation and time, the SISR based algorithm does not set any condition of the model structure and has a wider range of applications. When Viterbi algorithm can not be achieved, or less demanding of time, the SISR based algorithm might be a good choice.

REFERENCES

[1] Olivier Cappé, Eric Moulines, Tobias Rydén. Inference in hidden markov models. Springer, 2005.

[2] Zhu Chengwen, Li Bing, Hu Kui, Pang Kui. Particle filters for HMM state inference. Computer Engineering and Applications (2012),8,165-167.

[3] R.Rosales , MCMC for hidden markov models incorporating aggregation of states and filtering. Bulletin of Mathematical Biology(2004),66:1173-1199.

[4] Scott.S.L, Bayesian methods for hidden markov models: recursive computing in the 21st century. J.Am.stat.Assoc.(2002), 97:337-351.

[5] Cappé, O. Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation. Monte Carlo Methods Appl., 7, 81–92.

[6] Christophe Andrieu, Johannes Thoms. A tutorial on adaptive MCMC. Stat Comput(2008), 18: 343-374.

[7] Robert, C. P., Celeux, G. and Diebolt, J. Bayesian estimation of hidden Markov chains: A stochastic implementation. Statist. Probab. Lett. (1993), 16, 77–83.

[8] Jun S. Liu. monte carlo strategies in scientific computing. Springer, 2001.