The parameter's MCMC estimation of HMMs with transition density function

Chengwen Zhu, Yu Ge, Lina Lu, Zhang Tian, Chuizhen Zeng
Wuhan Ordnancy Non-commissioned Officer Academy of PLA
Wuhan, China
e-mail:chengwen_zhu@yeah.net

***Abstract*—The parameter estimation of HMM is critical to all its applications. The classic B-W algorithm is not flexible with the initial parameters and is easy to fall into the local optimal solution. Bayes estimation of it makes posterior risk minimization, and make full use of the experience, history information and other information other than samples, is useful in many cases. Employs the great computational power of MCMC, the MCMC estimation of HMM parameter can be more effective.**

Keywords-HMM; Gibbs Sampling; Conjugate Priors

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. The parameter estimation, that via the statistical calculations of sample set to adjust the model parameters to find the most suitable parameters, is the most difficult one compared to the other two, and generally, there is no best way to solve it. Usually, the structure of HMM is very complex, this limits the use of the least square method and moment method which are most commonly used in statistics. But we can use maximum likelihood method, transforming parameter estimation into seeking the extreme points of the Log-likelihood function.

The B-W algorithm using the EM algorithm to seeking the extreme points of the Log-likelihood function, lies on the calculation of the posterior distribution of hidden status .The E steps and M steps of B-W algorithm all have analytical solutions (standard EM algorithm) for HMMs with simple structure, but it's difficult and can only be solved by approximate calculation in general, such as MCEM (E step using the Monte Carlo method), SAEM (E step stochastic approximation method), SEM (Stochastic EM).

To treat HMM parameter estimation as a statistical decision problem, maximum likelihood is not the only option; the posterior risk minimization (often referred to as the Bayes decision criteria) is also commonly used. Usually, Bayes estimates can be attributed to integral calculation of the posterior distribution. The integration can be calculated directly or using the normal approximation, numerical integration, static Monte Carlo method if the posterior distribution is relatively simple. But when posterior distribution is a complex, high-dimensional, non-standard, all these methods are difficult to implement. Markov chain Monte Carlo methods (MCMC) can solve this effectively.

This dissertation discusses the parameter estimation of HMMs with transition density function from the point of view of Bayes statistics, using Bayes posterior risk minimization criterion instead of the maximum likelihood criteria, make a full use of the priori information, and employs the great computational power of MCMC, programming a MCMC algorithm of HMM parameters estimation.

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$(\mathbb{Y},\mathcal{Y})$, V 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.

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

The integration region in formula (1) is:

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

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:

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

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

$t[(x,y),(x',y')]\triangleq q(x,x')g(y,y')$ 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

Published by Atlantis Press, Paris, France.

which has a transition density function, and will no longer special instructions below.

III. HMM PARAMETER'S PRIORI DISTRIBUTION

Assume $\mathbb{X}$ is finite and $g(x_i, y_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma_{x_i}} \exp\left\{ -\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} \right\}$, in mind $\mathbb{X}=\{1,\cdots,r\}$, $\mu = (\mu_1, \mu_2, \cdots, \mu_r)$, $\Sigma = (\sigma_1, \sigma_2, \cdots, \sigma_r)$,use the notation π_0 to denote the probability density function of the initial state X_0 (with respect to ν). Then the HMM can be signified by its parameters $\theta = (\pi_0, Q, \mu, \Sigma)$. In mind θ 's Value space as Θ .

Given θ ,the density function of $(X_{0:n}, Y_{0:n})$ (respect to $\nu \otimes \lambda^{\otimes n} \otimes \mu^{\otimes(n+1)}$)is:

$$f_n(x_{0:n}, y_{0:n}; \theta) = \pi_0(x_0; \theta) g(x_0, y_0; \theta) \prod_{i=1}^{n} q(x_{i-1}, x_i; \theta) g(x_i, y_i; \theta) \qquad (3)$$

The greatest difference of Bayes statistics to classical statistical is it treat parameters as random variables, obtained the posterior distribution by the data and priori knowledge, and then makes a variety of statistical inference. Set the prior distribution density of θ as $\phi(\theta)$, according to Bayes formula, the posterior distribution density of θ under the condition that $(X_{0:n}, Y_{0:n})$ is known is:

$$\varphi(\theta \mid x_{0:n}, y_{0:n}) = \frac{\phi(\theta) f_n(x_{0:n}, y_{0:n}; \theta)}{\int \phi(\theta) f_n(x_{0:n}, y_{0:n}; \theta) d\theta} \propto \phi(\theta) f_n(x_{0:n}, y_{0:n}; \theta) \qquad (4)$$

In order to simplify the calculation, assume $\pi_0, Q_i, \mu_i, \sigma_i$ ($Q = (Q_1^T, Q_2^T, \cdots, Q_r^T)^T$) are independent of each other, and select the prior distribution as its conjugate prior distribution, that:

$\pi_0 = (\pi_0(1), \cdots, \pi_0(r))$ obeying Dirichlet distribution $D(b_1, \cdots, b_r)$;

$Q_i = (q(i,1), \cdots, q(i,r)), i = 1, \cdots, r$ obeying Dirichlet distribution $D(e_{i1}, \cdots, e_{ir})$;

μ_i obeying normal distribution $N(m_i, s_i{}^2)$;

σ_i^2 obeying Inverse gamma distribution $IG(u_i, w_i)$.

IV. HMM PARAMETER'S MCMC ESTIMATION

In the case of known $(X_{0:n}, Y_{0:n})$, we can yields the posterior distribution of the parameters θ by formula (3)、 (4), and then obtain its Bayes estimation. But we do not know the hidden state $X_{0:n}$, this is a typical problem of missing data, and can use augmented sampling methods to solve this kind of problems. Set the joint conditional density of $X_{0:n}$ and θ about $\{Y_{0:n} = y_{0:n}\}$ as $\varphi(x_{0:n}, \theta \mid y_{0:n})$, if we can obtain θ 's conditional density $\varphi(\theta \mid x_{0:n}, y_{0:n})$ and $X_{0:n}$'s conditional density $\varphi(x_{0:n} \mid y_{0:n}, \theta)$, then we can estimate θ . Similar to formula (4), according to Bayes formula:

$$\varphi(x_{0:n}, \theta \mid y_{0:n}) = \frac{\varphi(\theta) f_n(x_{0:n}, y_{0:n}; \theta)}{\sum_{x_{0:n}} \int_{\Theta} \varphi(\theta) f_n(x_{0:n}, y_{0:n}; \theta) d\theta} \qquad (6)$$

According to formula (3), (4)

$$\varphi(\theta \mid x_{0:n}, y_{0:n})$$
$$\propto \phi(\theta) \pi_0(x_0; \theta) g(x_0, y_0; \theta) \prod_{i=1}^{n} q(x_{i-1}, x_i; \theta) g(x_i, y_i; \theta) \qquad (7)$$

$$\varphi(x_{0:n} \mid y_{0:n}, \theta) = \varphi_{0:n|n}(x_{0:n} \mid y_{0:n}; \theta)$$
$$\propto \pi_0(x_0; \theta) g(x_0, y_0; \theta) \prod_{i=1}^{n} q(x_{i-1}, x_i; \theta) g(x_i, y_i; \theta) \quad (8)$$

Use formula (7)、 (8), obtained the MCMC sampling process for HMM parameters estimation:

(1) Given initial $\theta^{(0)}$, $x_{0:n}^{(0)}$, and sign $t = 1$;

(2) Sample $x_{0:n}^{(t)}$ from $\varphi(x_{0:n} \mid y_{0:n}, \theta^{(t-1)})$;

(3) Sample $\theta^{(t)}$ from $\varphi(\theta \mid x_{0:n}^{(t)}, y_{0:n})$, and then set $t = t + 1$, return(2).

Obviously

$$\varphi(x_{0:n} \mid y_{0:n}, \theta^{(t-1)}) = \varphi(x_n \mid y_{0:n}, \theta^{(t-1)}) \prod_{k=1}^{n} \varphi(x_{n-k} \mid x_{(n-k+1):n}, y_{0:n}, \theta^{(t-1)}) \quad (9)$$

If we can obtain samples from $\varphi(x_{n-k} \mid x_{(n-k+1):n}, y_{0:n}, \theta^{(t-1)})$, $k = 0, \cdots, n-1$ and $\varphi(x_n \mid y_{0:n}, \theta^{(t-1)})$, the problem has been solved. In formula (9)

$$\varphi(x_k \mid x_{(k+1):n}, y_{0:n}, \theta^{(t-1)})$$
$$= \sum_{x_0, \cdots, x_{k-1}} \varphi(x_0, \cdots, x_k \mid x_{(k+1):n}, y_{0:n}, \theta^{(t-1)})$$
$$\propto \sum_{x_0, \cdots, x_{k-1}} \varphi(x_0, \cdots, x_k, x_{(k+1):n} \mid y_{0:n}, \theta^{(t-1)})$$
$$\propto q(x_k, x_{k+1}; \theta^{(t-1)}) g(x_{k+1}, y_{k+1}; \theta^{(t-1)}) \varphi_{k|k}(x_k \mid y_{0:k}; \theta^{(t-1)})$$
$$\propto \varphi_{k,k+1|k+1}(x_k, x_{k+1} \mid y_{0:k+1}; \theta^{(t-1)})$$

$$\varphi(x_n \mid y_{0:n}, \theta^{(t-1)}) = \sum_{x_0, \cdots, x_{n-1}} \varphi(x_0, \cdots, x_n \mid y_{0:n}, \theta^{(t-1)})$$
$$= \varphi_{n|n}(x_n \mid y_{0:n}; \theta^{(t-1)})$$

Employ algorithm A In reference [4], $\varphi_{k,k+1|k+1}(x_k, x_{k+1} \mid y_{0:k+1}; \theta^{(t-1)})$, $k = 0, \cdots, n-1$ and $\varphi_{n|n}(x_n \mid y_{0:n}; \theta^{(t-1)})$ can be obtained recursively, and then we can easily get samples form $\varphi(x_n \mid y_{0:n}, \theta^{(t-1)})$ and $\varphi(x_{n-k} \mid x_{(n-k+1):n}, y_{0:n}, \theta^{(t-1)})$, $k = 0, \cdots, n-1$, that is just the sample $x_{0:n}^{(t)}$ from $\varphi(x_{0:n} \mid y_{0:n}, \theta^{(t-1)})$.

Combined with algorithm A in reference [4], given the Forward-Backward algorithm here:

(i) Forward calculation

$c_0(\theta^{(t-1)}) = L(y_0; \theta^{(t-1)}) = \sum_{x_0} \pi_0(x_0; \theta^{(t-1)}) g(x_0, y_0; \theta^{(t-1)})$,

$\varphi_{0|0}(x_0 \mid y_0; \theta^{(t-1)}) = c_0^{-1} \pi_0(x_0; \theta^{(t-1)}) g(x_0, y_0; \theta^{(t-1)})$, $x_0 \in \mathbb{X}$;

For $k = 0, \cdots, n-1$, calculated the following expression in turn

$c_{k+1}(\theta^{(t-1)}) = \sum_{x_k} \sum_{x_{k+1}} q(x_k, x_{k+1}; \theta^{(t-1)}) g(x_{k+1}, y_{k+1}; \theta^{(t-1)}) \varphi_{k|k}(x_k \mid y_{0:k}; \theta^{(t-1)})$,

$\varphi_{k+1|k+1}(x_{k+1} \mid y_{0:k+1}; \theta^{(t-1)}) = (c_{k+1})^{-1} \sum_{x_k} \varphi_{k|k}(x_k \mid y_{0:k}; \theta^{(t-1)}) q(x_k, x_{k+1}; \theta^{(t-1)}) g(x_{k+1}, y_{k+1}; \theta^{(t-1)})$,

$\varphi_{k,k+1|k+1}(x_k, x_{k+1} \mid y_{0:k+1}; \theta^{(t-1)}) = (c_{k+1})^{-1} \varphi_{k|k}(x_k \mid y_{0:k}; \theta^{(t-1)}) q(x_k, x_{k+1}; \theta^{(t-1)}) g(x_{k+1}, y_{k+1}; \theta^{(t-1)})$,

where $x_k, x_{k+1} \in \mathbb{X}$;

(ii) Backward sampling

Sample $x_n^{(t)}$ from $\varphi_{n|n}(x_n \mid y_{0:n}; \theta^{(t-1)})$;

For $k = n-1,\cdots,0$, sample $x_k^{(t)}$ from $\varphi_{k,k+1|k+1}(x_k, x_{k+1}^{(t)} \mid y_{0:k+1}; \theta^{(t-1)})$ in turn.

Note $p(\theta) = p(\pi_0, Q, \mu, \Sigma) = \varphi(\theta \mid x_{0:n}^{(t)}, y_{0:n})$. Because the dimensions of θ is relatively high, it's difficult to sample $\theta^{(t)}$ from $p(\theta)$ directly. As we had assumed $\pi_0, Q_i, \mu_i, \sigma_i$ are independent of each other, Gibbs sampling can be used:

(i) Sample $\pi_0^{(t)}$ from $p(\pi_0 \mid Q^{(t-1)}, \mu^{(t-1)}, \Sigma^{(t-1)})$;

(ii) For $i = 1,\cdots,r$, Sample $Q_i^{(t)}$ from

$$p(Q_i \mid \pi_0^{(t)}, Q_1^{(t)},\cdots,Q_{i-1}^{(t)}, Q_{i+1}^{(t-1)},\cdots,Q_r^{(t-1)}, \mu^{(t-1)}, \Sigma^{(t-1)});$$

(iii) For $i = 1,\cdots,r$, Sample $\mu_i^{(t)}$ from

$$p(\mu_i \mid \pi_0^{(t)}, Q^{(t)}, \mu_1^{(t)},\cdots,\mu_{i-1}^{(t)}, \mu_{i+1}^{(t-1)},\cdots,\mu_r^{(t-1)}, \Sigma^{(t-1)});$$

(iv) For $i = 1,\cdots,r$, Sample $\sigma_i^{(t)}$ from

$$p(\sigma_i^2 \mid \pi_0^{(t)}, Q^{(t)}, \mu^{(t)}, \sigma_1^{(t)},\cdots,\sigma_{i-1}^{(t)}, \sigma_{i+1}^{(t-1)},\cdots,\sigma_r^{(t-1)}).$$

As we have chosen conjugate priors, the conditional distribution of the above is easily obtained, after a simple calculation we can get the following conclusions:

(i) $\pi_0^{(t)} \sim D(b_1 + \delta_{x_0^{(t)},1},\cdots,b_r + \delta_{x_0^{(t)},r})$, $\delta_{x_0^{(t)},i} = \begin{cases} 1, & x_0^{(t)} = i \\ 0, & x_0^{(t)} \neq i \end{cases}$;

(ii) $Q_i^{(t)} \sim D(e_{i1} + n_{i1},\cdots,e_{ir} + n_{ir})$, $n_{ij} = \sum_{k=0}^{n-1} \mathrm{I}_{\{i\}}(x_k^{(t)})\mathrm{I}_{\{j\}}(x_{k+1}^{(t)})$;

(iii)

$\mu_i^{(t)} \sim N(\tilde{m}_i, \tilde{s}_i^2)$, $\tilde{m}_i = \dfrac{s_i^{\,2}\sum_{k=0}^{n} y_k \mathrm{I}_{\{i\}}(x_k^{(t)}) + m_i(\sigma_i^{(t-1)})^2}{s_i^{\,2}\sum_{k=0}^{n}\mathrm{I}_{\{i\}}(x_k^{(t)}) + (\sigma_i^{(t-1)})^2}$,

$\tilde{s}_i^{\,2} = \dfrac{s_i^{\,2}(\sigma_i^{(t-1)})^2}{s_i^{\,2}\sum_{k=0}^{n}\mathrm{I}_{\{i\}}(x_k^{(t)}) + (\sigma_i^{(t-1)})^2}$;

(iv)

$(\sigma_i^{(t)})^2 \sim IG(\tilde{u}_i, \tilde{w}_i)$, $\tilde{w}_i = w_i + \frac{1}{2}\sum_{k=0}^{n}(y_k - \mu_i^{(t)})^2 \mathrm{I}_{\{i\}}(x_k^{(t)})$,

$\tilde{u}_i = u_i + \frac{1}{2}\sum_{k=0}^{n}\mathrm{I}_{\{i\}}(x_k^{(t)})$.

Up to now, we have solved the sampling of step (2) and (3), and are able to get HMM parameter's estimates.

HMM parameter's MCMC estimate:

(i) Given the initials $\left(x_{0:n}^{(0)}, \theta^{(0)}\right)$, and the length T of the chain need to generate;

(ii) For $t = 1,2,\cdots,T$, sample $x_{0:n}^{(t)}$ from $\varphi(x_{0:n} \mid y_{0:n}, \theta^{(t-1)})$ and sample $\theta^{(t)}$ from $\varphi(\theta \mid x_{0:n}^{(t)}, y_{0:n})$;

(iii) Obtain a certain parameter θ_i 's Bayes estimation $\hat{\theta}_{nm} = \dfrac{1}{n-m}\sum_{t=m+1}^{n}\theta_i^{(t)}$

It's easy to prove that $\left\{(x_{0:n}^{(t)}, \theta^{(t)}), t = 1,2,\cdots,T\right\}$ yield from the algorithm above is a Markov chain. From the sampling process of the algorithm, we know the transition kernel from $(x_{0:n}^{(t)}, \theta^{(t)})$ to $(x_{0:n}^{(t+1)}, \theta^{(t+1)})$ is:

$$\varphi(\theta^{(t+1)} \mid x_{0:n}^{(t+1)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)}).$$

As we selected θ 's conjugate prior, the algorithm marginal distribution of $\theta^{(t+1)}$ which produced in the iterative process has a fixed form, so

$$\sum_{x_{0:n}^{(t)}}\int_\Theta \varphi(\theta^{(t+1)} \mid x_{0:n}^{(t+1)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)})\varphi(x_{0:n}^{(t)}, \theta^{(t)} \mid y_{0:n})d\theta^{(t)}$$
$$= \int_\Theta \varphi(\theta^{(t+1)} \mid x_{0:n}^{(t+1)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)})\varphi(\theta^{(t)} \mid y_{0:n})d\theta^{(t)}$$
$$= \varphi(\theta^{(t+1)} \mid x_{0:n}^{(t+1)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}) = \varphi(x_{0:n}^{(t+1)}, \theta^{(t+1)} \mid y_{0:n}) \quad (10)$$

Formula (10) told us $\varphi(x_{0:n}, \theta \mid y_{0:n})$ is the stationary distribution of transition kernel $\varphi(\theta^{(t+1)} \mid x_{0:n}^{(t+1)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)})$, which means $\varphi(x_{0:n}, \theta \mid y_{0:n})$ is the stationary distribution of Markov chain $\left\{(x_{0:n}^{(t)}, \theta^{(t)}), t = 1,2,\cdots,T\right\}$. It is difficult to consider the convergence properties of $\left\{(x_{0:n}^{(t)}, \theta^{(t)}), t = 1,2,\cdots,T\right\}$ directly. However, the algorithm uses augmented sampling methods, so only need to establish the convergence properties of one of the boundary chains, to obtain convergence properties of the joint chain. Here consider the convergence properties of the boundary chain $\left\{x_{0:n}^{(t)}, t = 1,2,\cdots,T\right\}$.

The marginal distribution density function of $X_{0:n}$ about $\{Y_{0:n} = y_{0:n}\}$ is:

$$\varphi_{x|y}(x_{0:n} \mid y_{0:n}) = \int_\Theta \varphi(x_{0:n}, \theta \mid y_{0:n})d\theta$$
$$= \int_\Theta \frac{\phi(\theta) f_n(x_{0:n}, y_{0:n}; \theta)}{\sum_{x_{0:n}}\int_\Theta \phi(\theta) f_n(x_{0:n}, y_{0:n}; \theta)d\theta}d\theta . \quad (11)$$

And the transition kernel from $X_{0:n}^{(t)}$ to $X_{0:n}^{(t+1)}$ is:

$$K_x(x_{0:n}^{(t+1)} \mid x_{0:n}^{(t)}) = \int_\Theta \varphi(\theta^{(t)} \mid x_{0:n}^{(t)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)})d\theta^{(t)} . \quad (12)$$

According to formula (11)、(12)

$$\sum_{x_{0:n}^{(t)}} K_x(x_{0:n}^{(t+1)} \mid x_{0:n}^{(t)})\varphi_{x|y}(x_{0:n}^{(t)} \mid y_{0:n})$$
$$= \sum_{x_{0:n}^{(t)}}\int_\Theta \varphi(\theta^{(t)} \mid x_{0:n}^{(t)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)})d\theta^{(t)}$$
$$\frac{\int_\Theta \phi(\theta) f_n(x_{0:n}^{(t)}, y_{0:n}; \theta)d\theta}{\sum_{x_{0:n}}\int_\Theta \phi(\theta) f_n(x_{0:n}, y_{0:n}; \theta)d\theta}$$
$$= \frac{\int_\Theta \phi(\theta^{(t)}) f_n(x_{0:n}^{(t+1)}, y_{0:n}; \theta^{(t)})d\theta^{(t)}}{\sum_{x_{0:n}}\int_\Theta \phi(\theta) f_n(x_{0:n}, y_{0:n}; \theta)d\theta}$$
$$= \varphi_{x|y}(x_{0:n}^{(t+1)} \mid y_{0:n}) .$$

So $\varphi_{x|y}(x_{0:n} \mid y_{0:n})$ is the stationary distribution of transition kernel $K_x(x_{0:n}^{(t+1)} \mid x_{0:n}^{(t)})$, which means $\varphi_{x|y}(x_{0:n} \mid y_{0:n})$ is the stationary distribution of boundary chain $\left\{x_{0:n}^{(t)}, t = 1,2,\cdots,T\right\}$.

Note $\varphi_{x|y}^{(t)}(x_{0:n} \mid y_{0:n})$ as $x_{0:n}^{(t)}$'s marginal distribution, if $\left\{x_{0:n}^{(t)}, t = 1,2,\cdots,T\right\}$ is irreducible and aperiodic, $\varphi_{x|y}^{(t)}(x_{0:n} \mid y_{0:n})$ will convergence to stationary distributions at Geometric rate in variation distance. That means, there exist $0 < r < 1$ and $c > 0$, cause

$$\left\|\varphi_{x|y}^{(t)}(\cdot \mid y_{0:n}) - \varphi_{x|y}(\cdot \mid y_{0:n})\right\| \triangleq \sum_{x_{0:n}}\left|\varphi_{x|y}^{(t)}(x_{0:n} \mid y_{0:n}) - \varphi_{x|y}(x_{0:n} \mid y_{0:n})\right| \le cr^t .$$

So it is only need to prove the boundary chain $\left\{x_{0:n}^{(t)}, t = 1,2,\cdots,T\right\}$ is irreducible and aperiodic, to prove the convergence of the algorithm.

For any given $x_{0:n}^{(t-1)}, x_{0:n}^{(t)} \in \mathbb{X}^{n+1}$, According to formula (7)、(8)

$$\varphi(\theta^{(t)} \mid x_{0:n}^{(t)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)})$$
$$\propto \left[\phi(\theta^{(t)})\pi_0(x_0^{(t)};\theta^{(t)})g(x_0^{(t)}, y_0;\theta^{(t)})\prod_{i=1}^{n} q(x_{i-1}^{(t)}, x_i^{(t)};\theta^{(t)})g(x_i^{(t)}, y_i;\theta^{(t)})\right]$$
$$\times\pi_0(x_0^{(t+1)};\theta^{(t)})g(x_0^{(t+1)}, y_0;\theta^{(t)})\prod_{i=1}^{n} q(x_{i-1}^{(t+1)}, x_i^{(t+1)};\theta^{(t)})g(x_i^{(t+1)}, y_i;\theta^{(t)}) .$$

Obvious, $\forall\theta^{(t)} \in \Theta$ and $\phi(\theta^{(t)}) > 0$, $\prod_{i=0}^{n} g(x_i^{(t)}, y_i;\theta^{(t)})g(x_0^{(t+1)}, y_0;\theta^{(t)}) > 0$. So the measure of the collection $\Theta^0 = \left\{\theta^{(t)} : \varphi(\theta^{(t)} \mid x_{0:n}^{(t)}, y_{0:n})\varphi(x_{0:n}^{(t+1)} \mid y_{0:n}, \theta^{(t)}) > 0\right\}$ must be greater than 0, and then $K_x(x_{0:n}^{(t+1)} \mid x_{0:n}^{(t)}) > 0$. Therefore, any two state of $\mathbb{X}^{n+1}$ is one step reachable, so $\left\{x_{0:n}^{(t)}, t = 1, 2, \cdots, T\right\}$ must be irreducible and aperiodic.

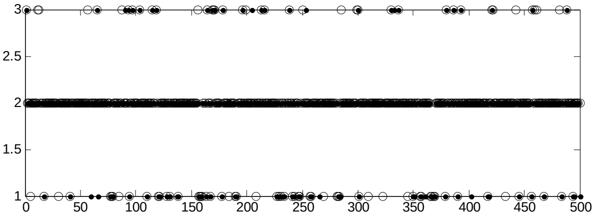
The above described boundary chain $\left\{x_{0:n}^{(t)}, t = 1, 2, \cdots, T\right\}$ is convergence, so the join chain $\left\{(x_{0:n}^{(t)}, \theta^{(t)}), t = 1, 2, \cdots, T\right\}$ is convergence. Thus prove the reasonableness of the algorithm.

V. SIMULATION TESTS AND CONCLUSIONS

Now we test the algorithm proposed in the previous section. Generate a group of hidden status and corresponding observations with a length of 500 ($n = 499$) from a HMM where $\mathbb{X}=\{1,2,3\}$, $\pi_0 = (0.1\ \ 0.8\ \ 0.1)$, $\mu = (-3\ \ 0\ \ 3)$, $\Sigma = (\sqrt{2}\ \ 1\ \ \sqrt{2})$, $Q = \left((0.2\ \ 0.1\ \ 0.1)^T, (0.7\ \ 0.8\ \ 0.7)^T, (0.1\ \ 0.1\ \ 0.2)^T\right)$.

Assume the hidden status and parameters are unknown, we need to obtain an estimate of the parameters. However, it is difficult to measure the gap between the estimated parameters and real parameters, as it is relatively complex. Therefore, we first obtain parameter estimation and then substituted into the Viterbi algorithm to get hidden status's estimation, and then make a proper evaluation.

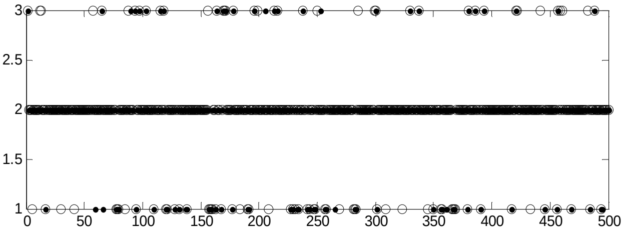
First, we get HMM parameters estimates by B-W algorithm, and then substituted it into the Viterbi algorithm; we get the estimates of $X_k(k = 0, \cdots, n)$, as shown in the following figure.



Figer1 Hidden status estimates of "B-W algorithm +Viterbi algorithm"

There are 44 error estimates In figer1, where " o " signify the real hidden status, " • "signify the Hidden status estimates of "B-W algorithm +Viterbi algorithm".

Although section 4 had described theoretically the convergence of the algorithm, however, the speed of convergence and the time needed to reach equilibrium is still unknown. A variety of methods has been proposed in Literatures. Here we avoid the theoretical discussion, just generate several chains from different initial points to determine the time needed to reach equilibrium by observing the trajectory of these chains. Finally get the MCMC estimation of our HMM parameters, substituted it into the Viterbi algorithm, we get the estimates of $X_k(k = 0, \cdots, n)$, as shown in the following figure.



Figer2 Hidden status estimates of "HMM parameter's MCMC estimate +Viterbi algorithm"

There are 40 error estimates in figer2, where " o " signify the real hidden status, " • "signify the Hidden status estimates of "HMM parameter's MCMC estimate +Viterbi algorithm".

From the simulation results of the above, we could see that "HMM parameter's MCMC estimate +Viterbi algorithm" is effective than "B-W algorithm +Viterbi algorithm". However, due to the use of MCMC, the time cost of the former is higher. The HMM of the present example is relatively simple, the B-W algorithm will be able to achieve good results, but in many cases B-W algorithms reach such good effects. When the data dimension is high, even be able to perform the B-W algorithm, the time cost is quite high, while the MCMC estimates has advantages, because of the MCMC method's strengths is dealing high-dimensional problems.

REFERENCES

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

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

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

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

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

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

[7] Jun S. Liu. Monte Carlo strategies in scientific computing. Springer, 2001.