AMS 263 - Final Project

Tracy Holsclaw

June 2010

1 Poisson Hidden Markov Model

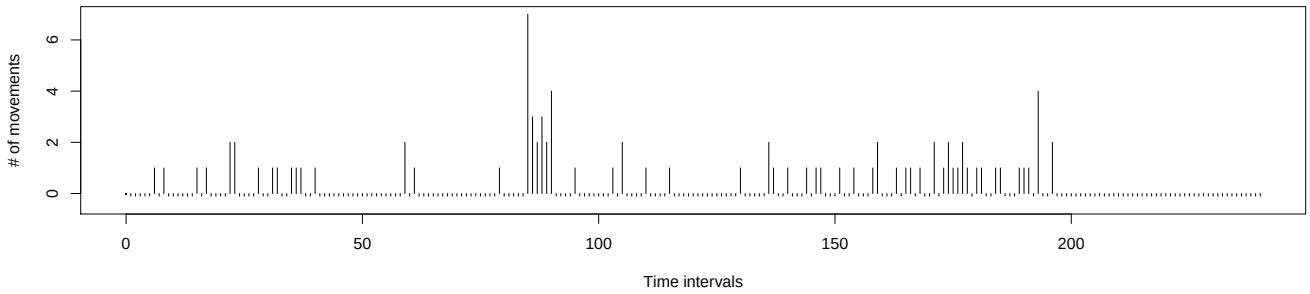
The count data $\mathbf{y}=\{y_1, y_2, ..., y_T\}$ i.This will be a Bayesian formulation for a Markov dependent mixture of K Poisson distributions, specified as:

The Poisson distribution is defined as $f(y_t|\lambda_k) = \frac{\lambda_k^{y_t} e^{-\lambda_t}}{y_t!}$, $t = 1, ..., 240$ and the Gamma distribution in the following parameterization $\Gamma(shape, rate)$.

$$
\begin{aligned}
y_t|z_t, (\lambda_1, ..., \lambda_K) &\overset{ind.}{\sim} Poisson(y_t|\lambda_{z_t}), t = 1, ..., T \\
z = (z_2, .., z_T)|Q &\sim \prod_{t=2}^{T} Pr(z_{t+1}|z_t; Q) = \prod_{t=2}^{T} q_{z_t, z_{t+1}} \\
\lambda_j &\overset{ind}{\sim} \Gamma(c_j, d_j), j = 1, ..., K \\
q_i &\overset{ind}{\sim} Dirichlet(a_{i1}, ..., a_{iK}), i = 1, ..., K
\end{aligned}
$$

where $q_i = (q_{i1}, ..., q_{iK})$ is the i^th row ($\sum_{j=1}^{K} q_{ij} = 1$) of the transition matrix Q. The first hidden state will be fixed, such as, $Pr(z_1 = 1) = 1$.

The data will be observed movement counts of a fetal lamb by ultrasound over 240 consecutive 5 second intervals. There could be hidden states of a relaxed states and excited states, which we will assume to be Markovian and depend on the previous state. And the observations over intervals will be assumed to be independent. The data will come from the paper by Puterman (1992).



(a) Data

Figure 1: Fetal Lamb Movement

We will show the posteriors for K=2 hidden states (z_1,z_2) but this can easily be extended to other numbers of states.
Full conditional:
$\sum_{(z_2,...,z_T)\in\{1,2\}} \left(Pois(y_1|\lambda_1) \prod_{t=1}^{T-1} q_{z_{t-1},z_t} Pois(y_t|\lambda_{z_t}) \right) \Gamma(\lambda_k|a_k,b_k) Dir(q_i|\alpha_{i1},\alpha_{i2})$

Posteriors:

$$\begin{aligned}
\lambda_k|... &= \sum_{(z_2,...,z_T)\in\{1,2\}} \left(Pois(y_1|\lambda_1) \prod_{t=2}^{T} q_{z_{t-1},z_t} Pois(y_t|\lambda_{z_t}) \right) \Gamma(\lambda_k|a_k,b_k) Dir(q_i|\alpha_{i1},\alpha_{i2}) \\
&\propto \left(Pois(y_1|\lambda_1) \prod_{t=2}^{T} Pois(y_t|\lambda_{z_t}) \right) \lambda_k^{a_k+1} e^{-b_k\lambda_k} \\
&\propto \left(e^{-\sum \lambda_k} \sum \lambda_k^{y_t} I_{z_t=k} \right) \lambda_k^{a_k+1} e^{-b_k\lambda_k} \\
&\propto \left(e^{-\sum \lambda_k I_{z_t=k} + b_k\lambda_k} \lambda_k^{\sum y_t I_{z_t=k} + a_k + 1} \right) \\
&\sim \Gamma(a_k + \sum y_t I_{z_t=k}, b_k + \sum I_{z_t=k})
\end{aligned}$$

Let $n_{ij} = \sum_{t=1}^{T-1} 1_{\{z_t=i, z_{t+1}=j\}}$

$$\begin{aligned}
q_i|... &= \sum_{(z_2,...,z_T)\in\{1,2\}} \left(Pois(y_1|\lambda_1) \prod_{t=2}^{T} q_{z_{t-1},z_t} Pois(y_t|\lambda_{z_t}) \right) \Gamma(\lambda_k|a_k,b_k) Dir(q_i|\alpha_{i1},\alpha_{i2}) \\
&\propto \left(\prod_{t=2}^{T} q_{z_{t-1},z_t} \right) Dir(q_i|\alpha_{i1},\alpha_{i2}) \\
&\propto q_{i1}^{n_{i1}} q_{i2}^{n_{i2}} Dir(q_i|\alpha_{i1},\alpha_{i2}) \\
&\propto q_{i1}^{n_{i1}} q_{i2}^{n_{i2}} q_{i1}^{\alpha_{i1}-1} q_{i2}^{\alpha_{i2}-1} \\
&\propto q_{i1}^{n_{i1}+\alpha_{i1}-1} q_{i2}^{n_{i2}+\alpha_{i2}-1} \\
&\sim Dir(n_{i1}+\alpha_{i1}, n_{i2}+\alpha_{i2})
\end{aligned}$$

This needs an extra subscript we will call m for the iteration of the MCMC:

$$\begin{aligned}
z_t^m|... &= \sum_{(z_2,...,z_T)\in\{1,2\}} \left(Pois(y_1|\lambda_{z_1}) \prod_{t=2}^{T} q_{z_{t-1}^m, z_t^{m-1}} Pois(y_t|\lambda_{z_t^m}) \right) \Gamma(\lambda_k|a_k,b_k) Dir(q_i|\alpha_{i1},\alpha_{i2}) \\
&\propto \sum_{(z_2,...,z_T)\in\{1,2\}} \left(Pois(y_1|\lambda_{z_1}) \prod_{t=2}^{T} q_{z_{t-1}^m, z_t^{m-1}} Pois(y_t|\lambda_{z_t^m}) \right) \\
&\propto q_{z_{t-1}^m,j} q_{j,z_{t+1}^{m-1}} Pois(y_t|\lambda_j) \\
&\propto \frac{q_{z_{t-1}^m,j} q_{j,z_{t+1}^{m-1}} Pois(y_t|\lambda_j)}{\sum_{k=1}^{K} q_{z_{t-1}^m,k} q_{k,z_{t+1}^{m-1}} Pois(y_t|\lambda_k)} \\
&\sim Bin\left(1, \frac{q_{z_{t-1}^m,j} q_{j,z_{t+1}^{m-1}} Pois(y_t|\lambda_j)}{\sum_{k=1}^{K} q_{z_{t-1}^m,k} q_{k,z_{t+1}^{m-1}} Pois(y_t|\lambda_k)}\right)
\end{aligned}$$

We will run both K=2 and K=3 models. The priors will be similar to those in Chib(1996). The MCMC was run 11,000 iterations and the first 1000 were discarded.

Table 1: Summary

Model	Parameters	Priors	Posteriors	Model Comparison
K=2	λ_1	$\Gamma(1,2)$	0.23	AIC = 362.96
	λ_2	$\Gamma(2,1)$	2.40	BIC = **376.88**
	q_{11}, q_{12}	$Dir(3,1)$	(0.97,0.03)	DIC =**341.81**
	q_{21}, q_{22}	$Dir(0.5,0.5)$	(0.33,0.66)	Post=0.00
				PPQL1=198.92
				PPQL2=178.54
				PPLL1=**246.68**
				PPLL2=167.77
K=3	λ_1	$\Gamma(1,2)$	0.08	AIC = **350.20**
	λ_2	$\Gamma(2,1)$	0.63	BIC = 381.52
	λ_3	$\Gamma(3,1)$	3.25	DIC = 386.36
	q_{11}, q_{12}, q_{13}	$Dir(3,1,0.1)$	(0.93,0.06,0.006)	Post=**0.999**
	q_{21}, q_{22}, q_{23}	$Dir(0.5,0.5,0.5)$	(0.09,0.89,0.03)	PPQL1=**189.90**
	q_{31}, q_{32}, q_{33}	$Dir(1,0.1,3)$	(0.23,0.02,0.75)	PPQL2=**173.15**
				PPLL1=251.87
				PPLL2=**157.47**

Predicted: $y_{rep} \sim \pi_1 Pois(y_t|\lambda_1) + \pi_2 Pois(y_t|\lambda_2)$, where π_1 and π_2 are calculated from the posteriors of q_{ij}.
We have the stationary distribution: $\pi = \pi Q$ giving: $\pi_1 = \frac{q_2}{1-q_1+q_2}$ and $\pi_2 = 1 - \pi_1$ for K=2. These will be used to create predictive samples for y and used in the model comparison. This seems like a good way to get predictions but it bothers me a little that it does not directly account for z_i but then again we would never know if we would be in a rest or active state when looking directly at the data.

There is a second way to do predicted values here. I could use the current values of z's to know which Poisson distribution to draw each y_{pred} from. I will include plots for both methods (see Fig(7))
Model comparison can be used to choose the number of parameters needed in the model, specifically K in this case. Here will be some standard notation for this section:

- p is the number of parameters in the model (K=2 then p=4, K=3 p=9). We have $K^2 - K$ parameters in the Q matrix and K parameters in λ.
- n is the number of observations (n=240)
- l is the log-likelihood computed from the iterative method in Scott(2002). The likelihood computed directly becomes too small for computational methods:
  $L_t(k) = Pois(y_t|\lambda_k) \sum_0 r = 0^{K-1} q(r,k) l_{t-1}(r)$.
  However, the log-likelihood l_t is stable can be computes as follows:
  $\pi_t(k|\lambda_k) = L_t(k)/l_t$
  $M_t = max_k\{\ln Pois(y_t|\lambda_k) \sum_{r=0} q(r,k)\pi_{t-1}(r|\lambda_k)\}$
  $l_t = log(l_{t-1}) + M_t + \log\left(\sum_{s=0}^{K-1} \exp\left[\log Pois(y_t|\lambda_s) + \log\left(\sum_{r=0}^{K-1} \pi_{t-1}(r|\lambda_s) q(r,s)\right) - M_t\right]\right)$
  $l = l_n$ is the recursive log-likelihood of interest we will use.

Types of model comparison (the model with the minimum of all these values is the favored one):

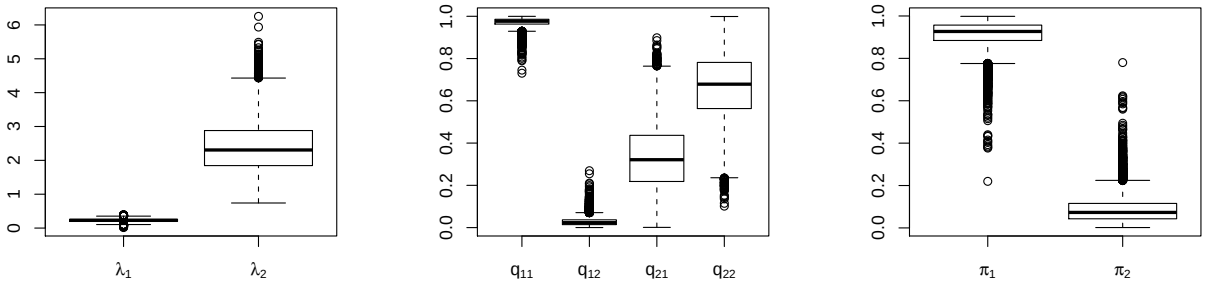
- $BIC = -2max(l) + p\ln(n)$

- $AIC = -2max(l) + 2p$
- $DIC = 2\overline{D} - \widehat{D}$ where $\overline{D} = \sum_{m=0}^{M}(-2l)/M$ and $\widehat{D} = -2\sum_{r=1}^{K}\log \pi_r^{MAP} Pois(y|\lambda_r^{MAP})$ where π is the stationary distribution. The $\widehat{D}$ is the density equation used when obtaining predicted values for y.
- The posterior method (we will denote as Post) described in Scott(2002) for choosing between any number of K (up to K_max=240 of course) can be done for independent Gibb samples of each model. We want a posterior for K, which is now a variable but is set in each of K Gibbs samplers. We will need a prior for K, which we will set to be a Uniform over the space of K, so it will cancel out of all the equations. $p(K|y) = \int p(K|y,\lambda)p(\lambda|y)d\lambda \approx 1/M\sum_{j=1}^{M} p(K|y,\lambda^{(j)})$, where M is the number of iterations in the Gibbs sampler for K_j. And $p(K|y,\lambda^{(j)}) \propto p(y|\lambda^{(j)}),K)p(K)$
  Basically this comes down to evaluating the likelihood at each step (obtained from the recursive log-likelihood method.) We get the average likelihood over all runs for each K: $1/M\sum_{j=1}^{M} L(y|\lambda_K)$, and then scale them to add to one because there is a finite number of models being compared.
- Posterior Predicted Quadratic Loss (PPQL)- this requires a loss function here were will use quadratic loss. This method is based on the predicted values of y from every iteration.
  $min_a(E[(a - y_{pred})^2 + k(a - data)]) = k/(1+k)(\mu_{pred} - data)^2 + \sigma^2_{pred}$
  where k is a constant that must be set, (several values of k are tried and we will display results for k=10.) They are denoted PPQL1 and PPQL2; they use the two different methods of obtaining y_{pred}. The first method uses the stationary distribution and assumes all intervals are the same, whereas, the second method uses the current z values to assign which distribution to draw y_{pred} from.
- Posterior Predicted Linear Loss (PPLL) - this is the same as the quadratic loss but uses a linear loss function instead.
  $min_a(E[|a - y_{pred}| + k|a - data|]) = k/(1+k)[2(\mu_{pred} - data)]$

The results of model comparison can be seen in Table 1 and the favored method is in bold. My BIC values are bit higher than Scott (2002) because my log-likelihood is slightly different. But BIC is picking the model with K=2 over the model with K=3, just like in Scott(2002). Using the table in Scott(2002) I calculate the AIC to be 309.4 (K=2) and 299.4 (K=3) which means that K=3 model is favored. BIC and AIC are typically both run on models for comparison reasons, since one is too liberal and one tends to be too conservative. In this case the AIC and BIC do not agrees which means that there can be no strong conclusion drawn from these methods. Scott(2002) makes the point that the posteriors are not Normal enough for the BIC to work properly. It would be good to do some other model comparison tests.

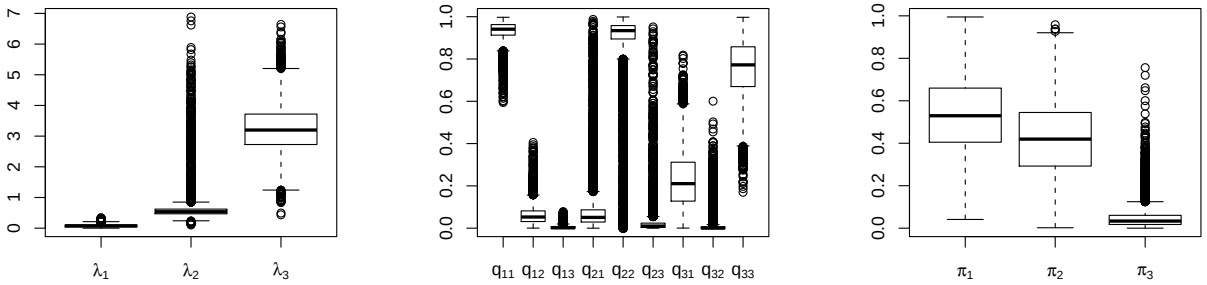
I also ran the posterior model probabilities as specified in Scott(2002). I got different values than they did but the same conclusion that K=3 model is chosen. Scott(2002) is averaging over many models and I am only averaging over two models, which will account for a difference.

The posterior predicted quadratic loss and linear loss choice different models as well. The second method of getting y_{pred} chooses model K=3, I think this is the better way of obtaining predicted values based on the information provided by the hidden z states. I think model K=3 is best, as Scott(2002) points out the assumptions of the BIC are not valid here. Scott(2002) concludes from the Gibbs sampling method that K=3 is best.



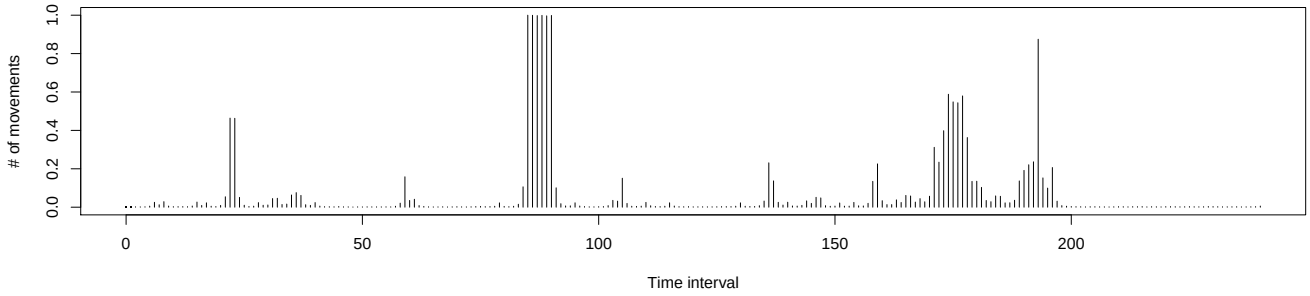
(a) (b) (c)

Figure 2: K=2 Posterior Plots

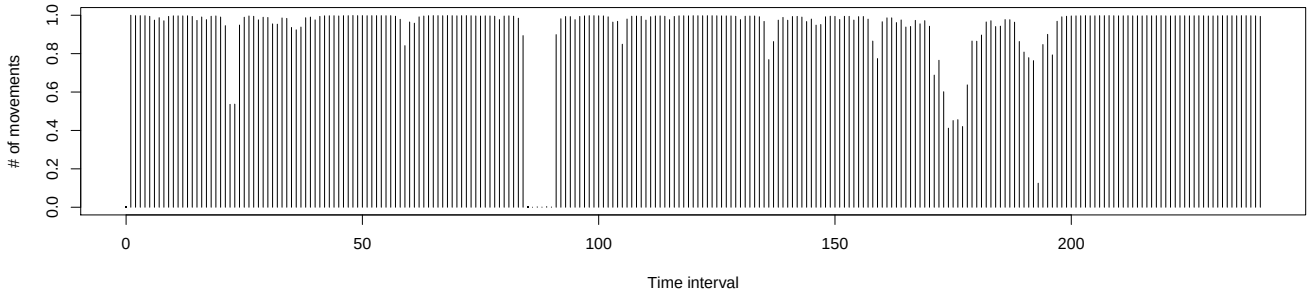


(a) (b) (c)

Figure 3: K=3 Posterior Plots

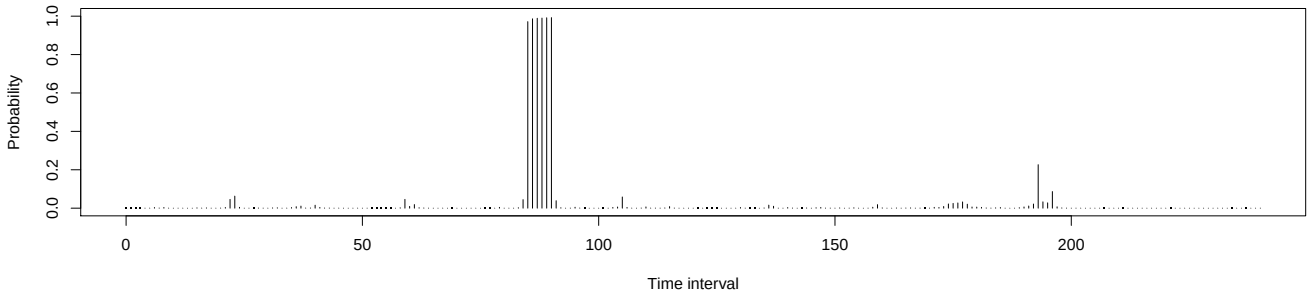


(a) Excited State

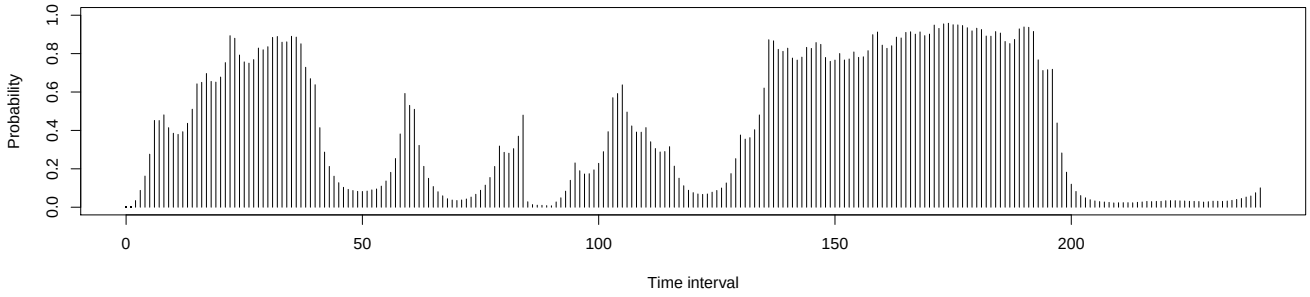


(b) Relaxed State

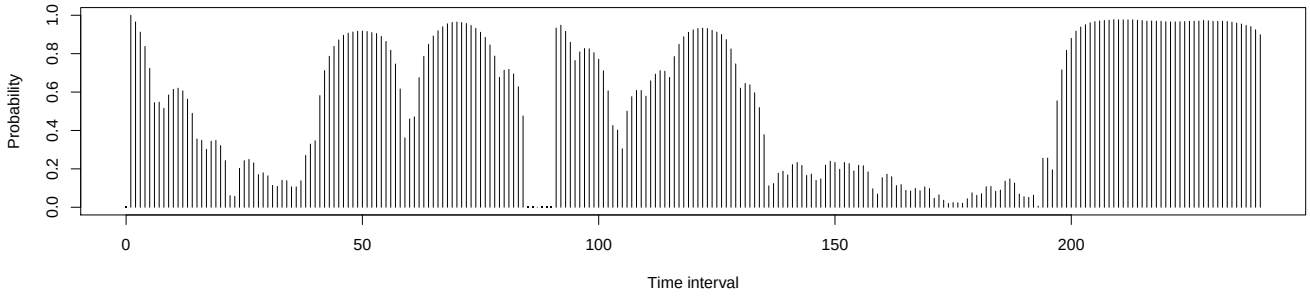
Figure 4: K=2 Probability of z



(a) Excited State

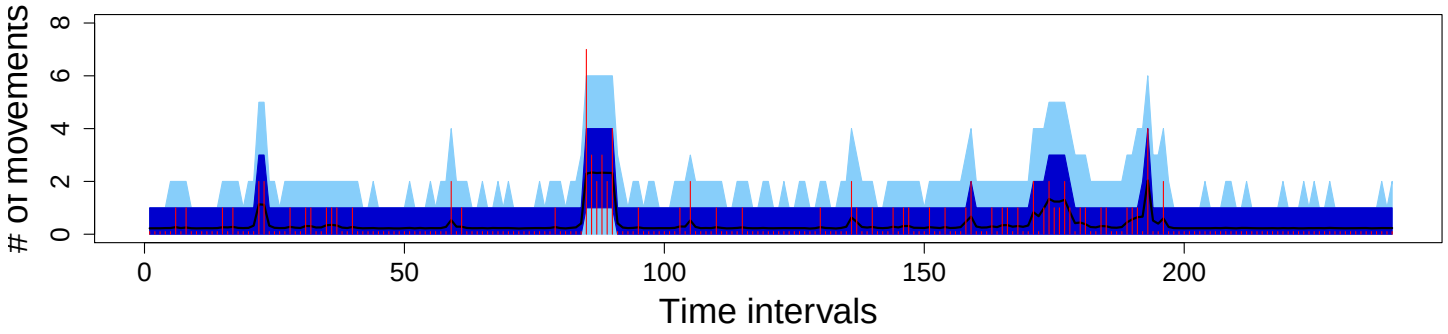


(b) Transition State

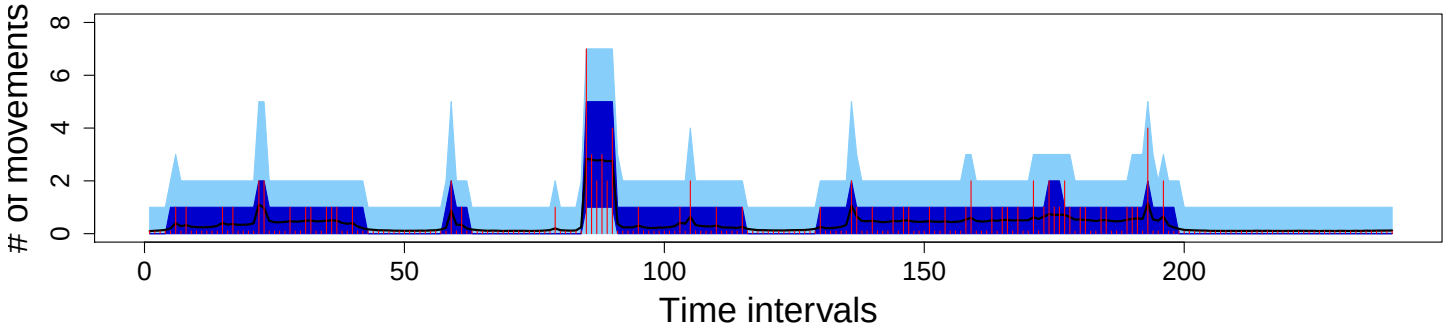


(c) Relaxed State

Figure 5: K=3 Probability of z

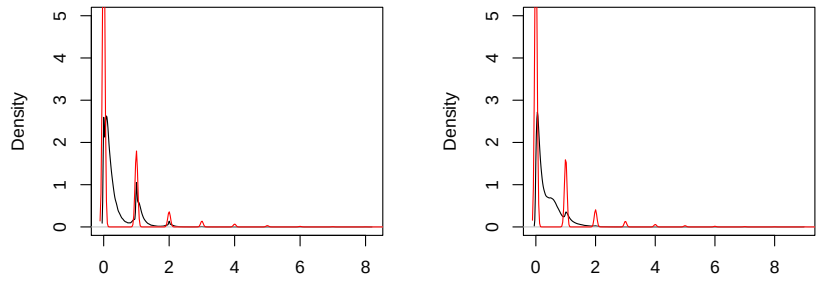


(a) K=2



(b) K=3

Figure 6: Posterior Predictive - 95% PI-light blue, 68%PI-blue, mean-black, data-red



(a) K=2 (b) K=3

Figure 7: Posterior Predictive — stationary probabilities method (black) and using z value method (red)

Bibliography:

Cox, D.R. and Miller, H.D. The Theory of Stochastic Process. Chapter 4 (for problem 3)

Chib, S.(1996). "Calculating posterior distributions and modal estimates in Markov mixture models." Journal of Econometrics, 75, 79-97.

Leroux, B.G. and Puterman, M.L. (1992). "Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models." Biometrics, 48, 545-558.

Scott, S.L. (2002). "Bayesian methods for hidden Markov models: Recursive computing in the 21st Century." Journal of the American Statistical Association, 97, 337-351.