EQ2340 Pattern Recognition: Report 1 (HMM Signal Source)

Olga Mikheeva
890312-7085
olgamik@kth.se

Diego A. Mancevo
890526-5834
damdc@kth.se

September 2016

A1.1 HMM Random Source

We implemented functions **@DiscreteD/rand**, **@MarkovChain/rand** and **@HMM/rand**. Those files will be provided with the report.

A1.2 Verify the MarkovChain and HMM Sources

To verify HHM generation we use the infinite-duration HMM with two states and Gaussian-distributed emissions.

$$q = \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix}; A = \begin{bmatrix} 0.99 & 0.1 \\ 0.03 & 0.97 \end{bmatrix}; B = \begin{bmatrix} b_1(x) \\ b_2(x) \end{bmatrix};$$
$$b_1(x) \sim \mathcal{N}(0,1); b_2(x) \sim \mathcal{N}(3,4);$$

1.

We can calculate state probability distribution at time t $P(S_t = j)$ by performing the following matrix multiplication operations:

$$P(S_t = j) = q^T * A^{t-1}$$

$P(S_1 = j) = q^T = [0.75, 0.25]$

$$P(S_2 = j) = q^T * A = \underbrace{\begin{bmatrix} 0.75 & 0.25 \end{bmatrix}}_{q^T} \underbrace{\begin{bmatrix} 0.99 & 0.01 \\ 0.03 & 0.97 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 0.75 & 0.25 \end{bmatrix}}_{q^T}$$

Notice however that $q^T * A = q^T$ meaning HMM is stationary and

$$P(S_t) = q^T, \forall t$$

2.

Using our Markov Chain *rand* function we can produce a sequence of $T = 10000$. The proportion of time spent on either state is shown below:

State 1:	0.7348
State 2:	0.2652

3.

We also calculate the expected value and variance for a given emission at time t, namely, $E[X_t]$ and $Var(X_t)$ as follows:

$$E[X_t] = P(S_t = 1)E[X|S_t = 1] + P(S_t = 2)E[X|S_t = 2] = 0.75 \cdot 0 + 0.25 \cdot 3 = 0.75$$

$$Var(X_t) = w_1 \cdot \sigma_1^2 + w_2 \cdot \sigma_2^2 + w_1 \cdot (\mu_1 - \mu)^2 + w_2 \cdot (\mu_2 - \mu)^2$$

$$= 0.75 \cdot 1^2 + 0.25 \cdot 2^2 + 0.75 \cdot (0 - 0.75)^2 + 0.25 \cdot (3 - 0.75)^2 = 3.4375$$

Where:

$$w_1 = P(S_t = 1); X|S_t = 1 \sim N(\mu_1, \sigma_1^2)$$
$$w_2 = P(S_t = 2); X|S_t = 2 \sim N(\mu_2, \sigma_2^2)$$

and:

$$\mu = E[X_t]$$

We can also produce an estimate for the mean and variance of the model's emissions by producing a sequence of length $T = 10000$ so as to verify that our calculations above agree with our model's output:

Mean	0.7342
Variance	3.4017

4.

Now we present a plot of a series of 500 contiguous samples X_t from our HMM:

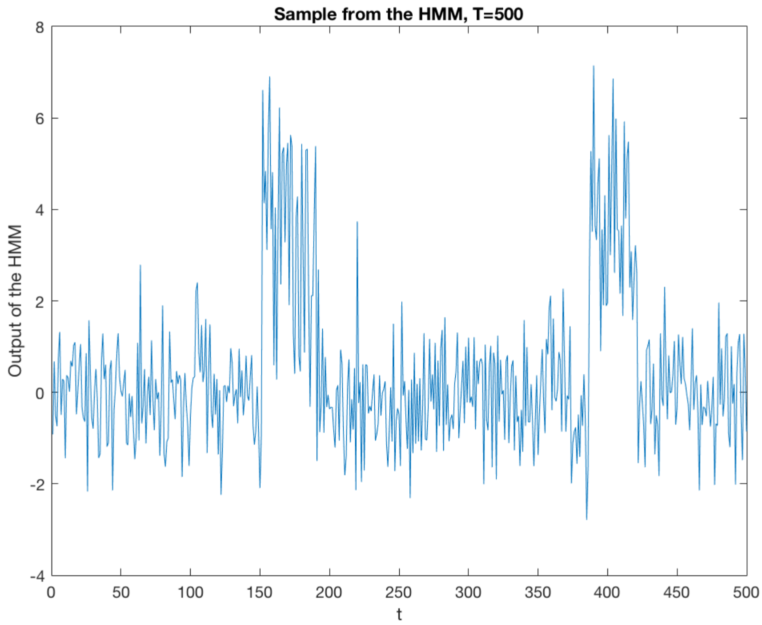

Figure 1: First HMM

Notice that in figure 1 above we can almost tell exactly when the HMM finds itself on State 1 or State 2 given the output values on the y-axis. For example, whenever the output values are close to 0 (e.g. within two standard deviations - i.e. +/- 2 as the standard deviation in this state equals one) we can be almost sure that the model is in state 1 - such is the case for the first 150 observations or so.

5.

Now lets produce some output from a second HMM where both emission distributions share the same mean (i.e. $\mu_1 = \mu_2 = 0$):

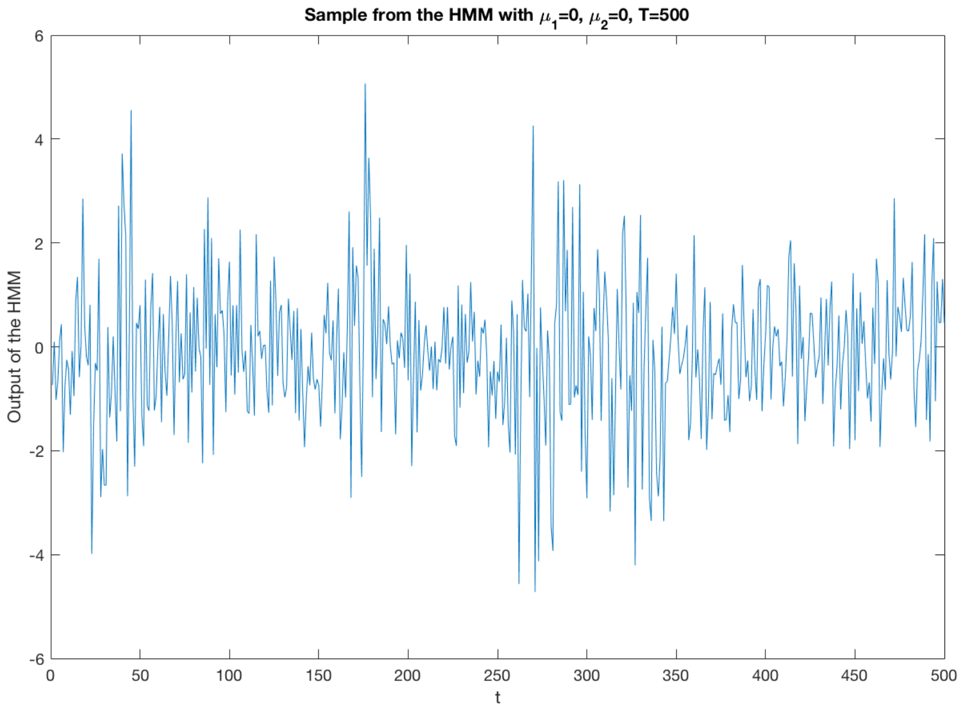

Figure 2: Second HMM

Notice that on figure 2 above it is not as clear whenever the model finds itself on State 1 or in State 2. We would argue however that it is still possible to state a sequence of Hidden states $\underline{S}$ as the variances from each emission distribution are not equal. Intuitively values that are much farther away from the mean are more likely to have come from State 2 as the emission distribution of this state has a much larger variance.

6.

We now turn our attention towards the implementation of our finite duration HMM. To test the model we can calculate the probability that the process will have a sequence length of 5 emissions or less. We use the initial state probabilities q and transition matrix A below:

$$q = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}; A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

$$P(\text{sequence length} \leq 5) = \sum_{t=1}^{5} 0.3 \cdot 0.7^{t-1} = 0.83$$

We then generate 1000 samples (whose maximum length is set to 100) and produce a normalized cumulative histogram with the sequence length on the x-axis:

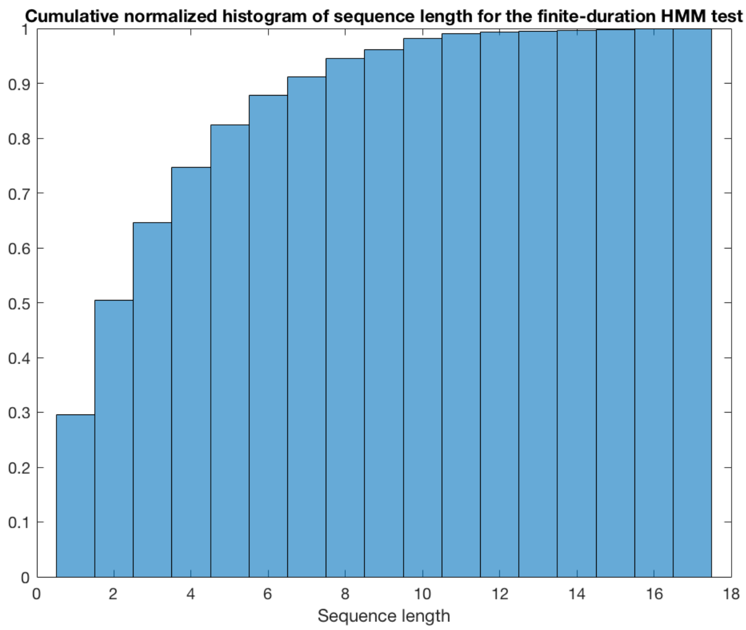

Figure 3: Normalized Cumulative Histogram, Finite Duration HMM

Notice that fifth bar of our histogram in figure 3 stands at around 0.83 which is exactly the value we had calculated previously confirming that our implementation is accurate.

7.

Now we proceed to check that our implementation's state-conditional output distributions can generate random vectors correctly. We define a new HMM model with initial state and trasition probabilities as given by q and A below:

$$q = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}; A = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

We also let both state-conditional outputs sample from a multivariate Gaussian as defined below:

$$b_i \sim \mathcal{N}\left(\begin{bmatrix} 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} \right); i = \{1, 2\}$$

Now we take one sample of length 1000 and use the observations to estimate the mean and covariance of the observations.

Estimated mean:

$$\hat{\mu} = \begin{bmatrix} 4.9330 \\ -0.0808 \end{bmatrix}$$

Estimated covariance:

$$\hat{C} = \begin{bmatrix} 1.8716 & 0.9737 \\ 0.9737 & 4.3220 \end{bmatrix}$$

Notice that the mean and co-variance estimates are very close to the theoretical mean and co-variance thus corroborating correctness of our implementation.

Appendix: MatLab Code

Normalized Cumulative Histogram of Sequence Length for Finite Duration HMM

```
rand('seed', 1);
mc = MarkovChain([0.5 0.5], [0.5 0.2 0.3; 0.2 0.5 0.3]);
g1=GaussD('Mean',0,'StDev',1);
g2=GaussD('Mean',3,'StDev',2);
h = HMM(mc, [g1,g2]);
l = zeros(1000,1);
for i=1:1000
    l(i,1) = length(rand(h,100));
end
%histogram(l,'Normalization', 'pdf')
histogram(l,'Normalization', 'cdf')
xlabel('Sequence length')
title('Cumulative normalized histogram of sequence length for the finite-duration HMM test')
```

Vector-valued test

```
mc = MarkovChain([0.5 0.5], [0.6 0.4; 0.4 0.6]);
g1=GaussD('Mean',[5 0],'Covariance',[2 1; 1 4]);
g2=GaussD('Mean',[5 0],'Covariance',[2 1; 1 4]);
h = HMM(mc, [g1,g2]);
seq = rand(h,1000);
mean(seq,2)
cov(seq(1,:),seq(2,:))
```