R ladies hands-on HMM

Emmeke Aarts

Exercise 1

In this first exercise we will simulate data from a hidden Markov model, and subsequently analyse it. We will assume that the observations are generated by a normal distribution, for which the parameters depend on which state is active. Note that the used code can easily be adapted to generate and analyse data with more than 2 states, or another conditional distribution, like Poisson or categorical.

a)

Simulate a 2-state normal HMM with 100 observations using the code below. Start with a model that is easy to estimate for the HMM algorithm: conditional distributions that do not or only partly overlap and high self-transitions. For example,

- mu equals 3 and 10
- sd equals 1.5 and 3
- self transition probabilities of .8 for both states (and hence, a probability of .2 to switch to the other state)

```
norm.HMM.Generate <- function(n, m, mu, sd, gamma1, delta=NULL){
  if(m != length(gamma1[,1]) | m != length(gamma1[1,])){
   stop("The number of states given by m does not match the number of rows
        and/or colums of the transition probability matrix gamma")
  }
  state <- numeric(n)
      # first attribute the first observation in the sequence
      if (is.null(delta)) {
        state[1] <- sample(1:m, 1, prob = solve(t(diag(m) - gamma1 + 1), rep(1, m)))
      } else {
           state[1] <- sample(1:m, 1, prob = delta)
      }
    # conditional on the first observation, sample the following observations
      for (i in 2:n){
        state[i] <- sample(1:m, 1, prob = gamma1[state[i-1],])
      }
      x <- rnorm(n, mean = mu[state], sd = sd[state])
      return(list(state = state, seq.x = x))
    }

sample.data <- norm.HMM.Generate(n = 100, m = 2, mu = c(3, 10), sd = c(1.5, 3),
               gamma1 =   matrix(c(.8, .2, .2, .8), byrow = T, ncol = 2))
```

b)

Plot the data over time. A line plot gives the most intuitive visualization of the outcome here.

c)

Analyse the simulated data applying the HMM algorithm from the `depmixS4` package.

d)

Inspect conditional distributions and compare to true values.

e)

Inspect the transition matrix and compare to true values

f)

Use the Viterbi algorithm from the `depmixS4` package to recover the most likely state sequence

g)

Add the estimated and the true states to the plot that depicts the observations over time (that you made in 1a)

h)

Compare the plotted estimated and true states to each other, and to the simulated observations over time

i)

Play around with simulating the data and running the HMM: does it recover the hidden states correctly each time? How about if you make the simulated data sequence longer or shorter? And if you change the conditional distribution or the transition matrix?

Exercise 2

In this exercise, we will use example data from the book *Hidden Markov Models for Time Series - An Introduction Using R* (2009) by W. Zucchini and I.L. MacDonald. The data concerns the number of major earthquakes in the world from 1900-2006, and is assumed to be Poisson distributed.

a)

The data is as follows:

```
earthq.data <- c(13, 14, 8, 10, 16, 26, 32, 27, 18, 32, 36, 24, 22, 23, 22, 18, 25,
                 21, 21, 14, 8, 11, 14, 23, 18, 17, 19, 20, 22, 19, 13, 26, 13, 14,
                 22, 24, 21, 22, 26, 21, 23, 24, 27, 41, 31, 27, 35, 26, 28, 36,
                 39, 21, 17, 22, 17, 19, 15, 34, 10, 15, 22, 18, 15, 20, 15, 22,
                 19, 16, 30, 27, 29, 23, 20, 16, 21, 21, 25, 16, 18, 15, 18, 14,
                 10, 15, 8, 15, 6, 11, 8, 7, 18, 16, 13, 12, 13, 20, 15, 16, 12,
                 18, 15, 16, 13, 15, 16, 11, 11)
```

Plot the data over time.

b)

Fit a 2 state, 3 state and 4 state HMM using the `depmixS4` package. Note that depmix returns the intercept for the Poisson distribution instead of lambda. To get lambda, take the exponent of the value under intercept.

c)

Compare the different models on their conditional distributions

d)

Which model best fits the data?

e)

Obtain the most likely state sequence over time using the best fitting model, and add it to the plot that contains the data over time.

Exercise 3

Next, we will use a fictional example inspired on a collaboration with dr. William Hale. We have the data on non-verbal communication between a therapist and its patient. For both, we collected:

- looking behaviour: the person either does or does not look at the other person
- vocalizing behaviour: the person is speaking, back channelling, or silent.

The data is in the file patient-therapist.csv.

a)

Plot the data over time.

b)

Fit a 1 state, 2 state, 3 state, and 4 state HMM

c)

Inspect the composition of the states in the different models, and give a theoretical interpretation to them. Do each of the states make sense?

d)

Which model would you choose to represent the data?

e)

Obtain the most likely state sequence over time, and add it to the plot that contains the data over time