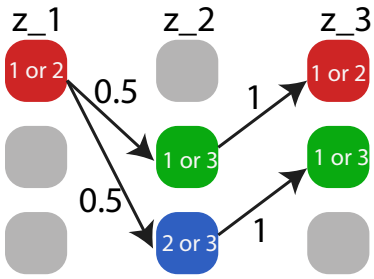
DS-GA 3001.001 Special Topics in Data Science: Probabilistic Time Series Analysis
Homework 3

Due date: Oct 25, by 6pm

Problem 1. (15p)
Consider the HMM with K=3 latent states and discrete observations $\{1, 2, 3\}$, with parameters specified by: initial distribution $\pi = [1,\ 0,\ 0]$, transition matrix $\mathbf{A} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, where $A_{ij} = \mathrm{P}(z_{t+1} = j | z_t = i)$ and likelihood $\mathrm{P}(x_t|z_t)$ described by matrix entries B_{xz}: $\mathbf{B} = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}$.
Write down all possible state sequences consistent with observations a) 1, 2, 3 and b) 1, 3, 1.

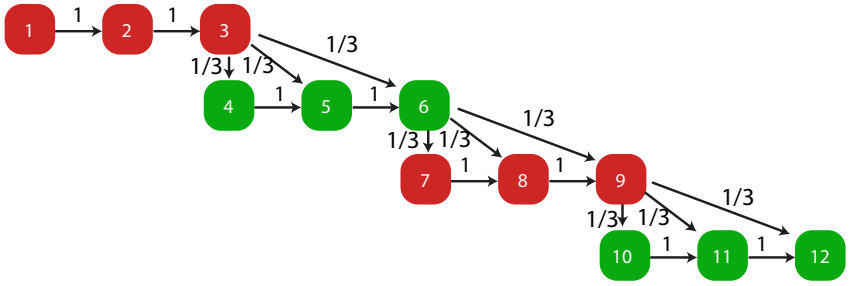
Solution:



Given the initial conditions distribution and the transition probabilities, there are only 2 possible sequences of latent states of length 3: $z_1 = 1, z_2 = 2, z_3 = 1$ and $z_1 = 1, z_2 = 3, z_3 = 2$, each occurring with probability 0.5 (states in gray are impossible under the model, white text marks possible observations). We just need to check which of these are consistent with the observations. For observation sequence 1,2,3: $x_2 = 2$ could only have occurred for $z_2 = 3$, which makes the only possible valid latent sequence $z = [1, 3, 2]$. Similarly, the observation sequence 1,3,1 could have arisen from the either latent sequence $z = [1, 2, 1]$ or $z = [1, 3, 2]$.

Problem 2. (15p)
Construct an HMM that generates the observation sequence $A^{k_1}C^{k_2}A^{k_3}C^{k_4}$ where A^{k_1} denotes k_1 repeats of symbol A and the number of repeats k_i are drawn from the set $\{1, 2, 3\}$ with equal probability.

Solution:
There are several options for doing this, and any valid solution is fine. Here's a solution with 12 states (state 12 also serves as terminal state): the initial probability is $\pi = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0, 0, 0, 0, 0, 0, 0]$, the observations are deterministic: 'A' for red states, 'C' for green, with transition probability matrix A given by:



Problem 3. (20p)
Implement EM for an HMM model with K states and gaussian observations (full derivations in handout). Use this code to fit the weekly S&P 500 returns data (data/sp500w.csv) for K = 2 vs. K = 3 and compare the two results.
Hint: Use Example 6.17 from tsa4 textbook as guideline for plots and interpretation.
Solution:
The exact solution will depend on the procedure used for setting up initial conditions, but it will be similar to the one in the book, with one component of high variance and 1 or 2 components for the rest; the contribution of the 3rd component is relatively minor, so we may consider dropping it. To decide, one should check the final log likelihood of the two variants and select the better one. My implementation initializes the observation model using a gaussian mixture, and converges in 20-30 iterations; and model comparison weakly favors the K=2 solution. Plot shows parameters for K=2, with mean in dashed lines and shaded regions for 95% confidence intervals.