CPSC 540 Tutorial

Reza Babanezhad

rezababa@cs.ubc.ca

Motivation

- Action detection in video
    - You are given a video and asked to extract actions happening in it. What would you do?

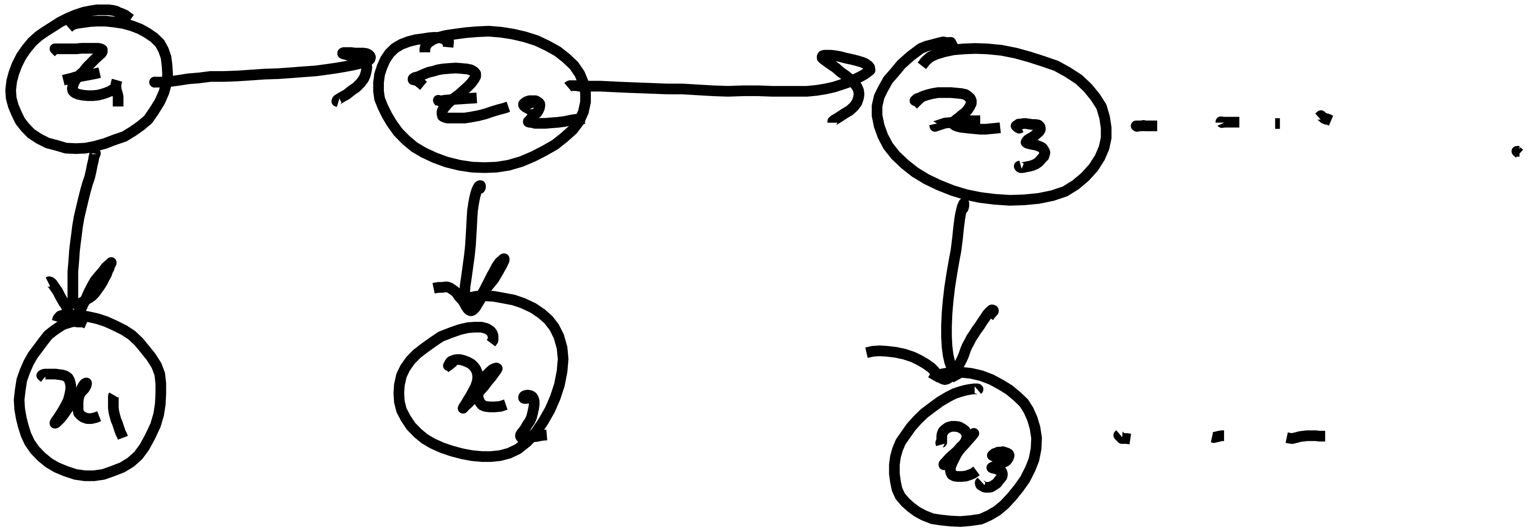
- One way to solve it: using HMM
    - Each frame is known variable and action happening in there is latent!

HMM

- HMM can be used to model sequential data or stochastic processes.
  - Including two models
    - A discrete time, discrete state Markov chain with hidden states
      $$z_t \in \{1,2,\ldots,K\}$$
    - An observation model $p(x_t|z_t)$
    - Observation model could be discrete or continuous
      - In discrete case, it could be observation matrix (we consider this case here)
        $$p(x_t = l|z_t = k) = B_{kl}$$
      - In continuous case, we can use conditional Gaussian.
        $$p(x_t|z_t = k;\theta) = N(x_t|\mu_k, \Sigma_k)$$

- Applications: Speech recognition, Activity recognition, Part of speech tagging, Gene finding, …

Model Parameters



$P(z_1 = k) = \pi_k \longrightarrow$ initial distribution

$P(z_t = k \mid z_{t-1} = j, \theta) = A_{jk} \longrightarrow$ Transition prob.

$P(x_t = l \mid z_t = k) = B_{kl} \longrightarrow$ observation prob.

Inference with HMM

- Given parameters, we can do the following inference in HMM
  - Filtering: computing the belief state (online) $P(z_t | x_{1:t})$
  - Smoothing: computing the state offline $P(z_t | x_{1:T})$
  - Fixed lag smoothing: $P(z_{t-l} | x_{1:t})$
  - Prediction: $P(z_{t+h} | x_{1:t})$.
  - MAP estimation: $\longrightarrow \arg\max_{z_{1:T}} P(z_{1:T} | x_{1:T})$
  - Posterior sampling: $z^s_{1:T} \sim P(z_{1:T} | x_{1:T})$
  - Probability of evidence: $\rightarrow P(x_{1:T}) = \sum_{z_{1:T}} P(x_{1:T}, z_{1:T})$

Filtering using Forward Algorithm : $P(z_t = j \mid x_{1:t}, \theta)$

$$P(z_t = j \mid x_{1:t-1}) = \sum_i P(z_t = j \mid z_{t-1} = i)\, P(z_{t-1} = i \mid x_{1:t-1})$$

$$\alpha_t(j) \equiv P(z_t = j \mid x_{1:t}) = P(z_t = j \mid x_t, x_{1:t-1})$$

$$= \frac{1}{Z_t} P(x_t \mid z_t = j, \cancel{x_{1:t-1}})\, P(z_t = j \mid x_{1:t-1})$$

$$Z_t = P(x_t \mid x_{1:t-1}) = \sum_j P(z_t = j \mid x_{1:t-1})\, P(x_t \mid z_t = j)$$

$$P(z_1 \mid x_1) = \frac{P(x_1 \mid z_1)\, P(z_1)}{\sum_{z_1} P(x_1 \mid z_1)\, P(z_1)}$$

Forward Backward algorithm for Smoothing

$$P(Z_t = j \mid x_{1:T}) = P(Z_t = j \mid x_{1:t}, x_{t+1:T}) = \frac{P(Z_t, x_{t+1:T} \mid x_{1:t})}{P(x_{1:T})}$$

$$\propto P(Z_t = j, x_{t+1:T} / x_{1:t}) = P(x_{t+1:T} \mid Z_t, \cancel{x_{1:t}}) P(Z_t \mid x_{1:t})$$

We know how to compute $\alpha_t(j) = P(Z_t = j \mid x_{1:t})$

let $\beta_t(j) = P(x_{t+1:T} \mid Z_t = j) = \sum_i P(Z_{t+1} = i, x_{t+1:T} \mid Z_t = j)$

$$= \sum P(Z_{t+1} = i, x_{t+1}, x_{t+2:T} \mid Z_t = j)$$

$$\beta_t(j) = \sum_i P(x_{t+2:T} \mid z_{t+1}=i, \cancel{z_t=j}, \cancel{x_{t+1}})\, P(x_{t+1}, z_{t+1}=i \mid z_t=j)$$

$$= \sum_i \underbrace{P(x_{t+2:T} \mid z_{t+1}=i)}_{\beta_{t+1}(i)}\, \underbrace{P(x_{t+1} \mid z_{t+1}=i)}_{B_{i\cdot}}\, \underbrace{P(z_{t+1}=i \mid z_t=j)}_{A_{ji}}$$

Base case : $\beta_T(j) = P(x_T \mid z_T = j) = B_{j\cdot}$

So $P(z_t \mid X_{1:T}) \propto \alpha_t(j)\beta_t(j)$

Two-slice smoothed marginals

Two slice marginal : $\xi_{t,t+1}(i,j) \triangleq P(z_t=i, z_{t+1}=j \mid x_{1:T})$

$$\xi_{t,t+1}(i,j) = P(z_t, z_{t+1} \mid x_{t+1:T}, x_{1:t})$$

$$\propto P(x_{t+1:T}, z_t, z_{t+1} \mid x_{1:t})$$

$$\propto P(x_{t+1:T} \mid z_t, z_{t+1}, x_{1:t})\, P(z_t, z_{t+1} \mid x_{1:t})$$

$$\propto P(x_{t+1:T} \mid z_{t+1})\, \overbrace{P(z_{t+1} \mid z_t)}^{A_{ij}}\, \overbrace{P(z_t \mid x_{1:t})}^{\alpha_t(i)}.$$

$$P(x_{t+1:T} \mid z_{t+1}) = P(x_{t+1}, x_{t+2:T} \mid z_{t+1})$$

$$= \underbrace{P(x_{t+1} \mid z_{t+1})}_{B_{j\cdot}} \underbrace{P(x_{t+2:T} \mid z_{t+1})}_{\beta_{t+1}(j)}$$

So $\xi_{t,t+1}^{(i,j)} \propto \alpha_t(j)\, B_{j\cdot}\, \beta_{t+1}(j)\, A_{ij}$

The Viterbi Algorithm

goal : $Z^* = \underset{Z_{1:T}}{\arg\max}\, P(Z_{1:T} | X_{1:T})$

$\delta_t(j) \triangleq \max_{Z_{1:t-1}} P(Z_{1:t-1}, Z_t = j | X_{1:t}) \rightarrow *$

$P(Z_{1:t-1}, Z_t = j | X_{1:t-1}, X_t) \propto \underbrace{P(X_t | Z_t)}_{B_j} P(Z_{1:t-1}, Z_t = j | X_{1:t-1})$

$P(Z_t = j, Z_{1:t-1} | X_{1:t-1}) = \underbrace{P(Z_t | Z_{t-1})}_{A_j} P(Z_{1:t-1} | X_{1:t-1}) \rightarrow *$

$$(*) \rightarrow \delta_t(j) = \max_{z_{1:t-1}} P(z_{1:t-1} \mid x_{1:t-1}) A_{\cdot j} B_{j\cdot}$$

$$\max_{z_{1:t-2}} \max_{i} P(z_{1:t-2}, z_{t-1} = i \mid x_{1:t-1}) = \max_i \delta_{t-1}(i)$$

So : $\delta_t(j) = \max_i \delta_{t-1}(i) A_{ij} B_{j\cdot}$

To keep track of most likely previous state

$$a_t(j) = \arg\max_i \delta_{t-1}(i) A_{ij} B_{j\cdot}$$

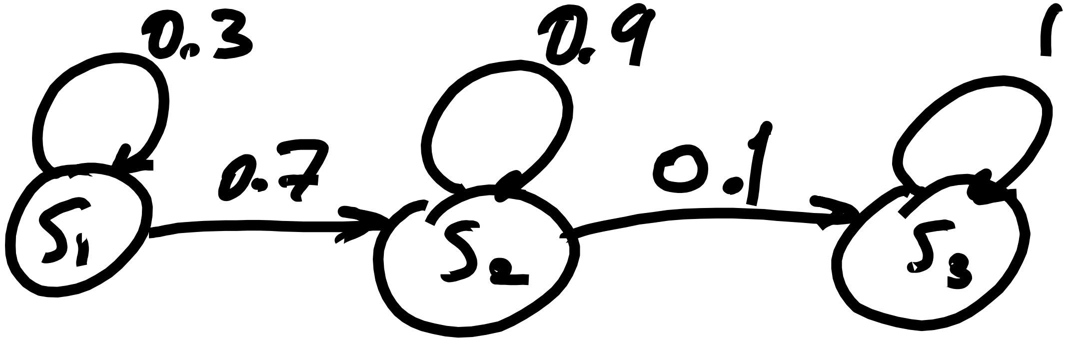
Base case : $\delta_1(j) = \pi_1 B_{j o}$

the algorithm terminates when we compute $Z_T^* = \arg\max_i \delta_T(i)$

To compute moste probable sequence using trace back

$$Z_t^* = a_{t+1}(Z_{t+1}^*)$$

Example



state / obs.	S_1	S_2	S_3
C_1	0.5	0	0
C_2	0.3	0	0
C_3	0.2	0.2	0
C_4	0	0.7	0.1
C_5	0	0.1	0
C_6	0	0	0.5
C_7	0	0	0.4

$A_{11} = 0.3, \; A_{12} = 0.7, \; A_{21} = 0 \ldots$

$$A = \begin{array}{c c} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0 & 0.9 & 0.1 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

Data: C1 C3 C4 C6 $= x_{1:4}$

$\pi = (1, 0, 0)$

$\delta_1(1) = 1 \times 0.5 = 0.5$, $\delta_1(2) = 0$, $\delta_1(3) = 0$ $\leftarrow$ $t = 1$

$\delta_2(1) = \max_i \delta_1(i)\, A(i,1)\, B_{1C3}$ $\leftarrow$ $t = 2$

for $i = 1$ $\delta_1(i) = 0.5$ but for $i = 2, 3$ $\delta_i = 0$

so $\delta_2(1) = \delta_1(1) * A(1,1) * B_{1C3} = 0.5 * 0.2 * 0.3 = 0.03$

$\delta_2(2)$ like $\delta_2(1)$ just $\delta_1(1) \neq 0$

so $\delta_2(2) = \delta_1(1)\, A(1,2)\, B_{2C3} = 0.5 * 0.7 * 0.2 = 0.07$

like before $\delta_2(3) = \delta_1(1)\, \underset{0}{A(1,3)}\, \underset{0}{B_{3C_3}} = 0$

So $a_2(1) = 1$, $a_2(2) = 1$ $a_2(3) = 1$ or 2 or 3

Now

$t = 3$

since $B_{1C4} = 0 \Rightarrow \delta_3(1) = \max_i \delta_2(i)\, A(i,1)\, B_{1C4} = 0$

$$\delta_3(2) = \max \begin{cases} \delta_2(1)\, A(1,2)\, B_{2C4} = 0.03 \times 0.7 \times 0.7 = 0.0147 \\ \delta_2(2)\, A(2,2)\, B_{2C4} = 0.07 \times 0.9 \times 0.7 = 0.0441 \\ \underset{0}{\delta_2(3)}\, A(3,2)\, B_{2C4} = 0 \end{cases}$$

$\delta_3(2) = 0.0441$

$$\delta_3(3) = \max \begin{cases} \delta_2(1)\, A(1,3)\, B_{3C4} = 0 \\ \delta_2(2)\, A(2,3)\, B_{3C4} = 0.07 \times 0.1 \times 0.1 = 0.0007 \\ \delta_2(3)\, A(3,3)\, B_{3C4} = 0 \end{cases}$$

$a_3(1) = 1, 2, 3$, $a_3(2) = 2$, $a_3(3) = 2$

$t = 4$:

Since $B_{1C6} = 0 \Longrightarrow \delta_4(1) = 0$

Since $B_{2C6} = 0 \Longrightarrow \delta_4(2) = 0$

$$\delta_4(3) = \max \begin{cases} \delta_3(1)\ A(1,3)\ B_{3c6} = 0 \\ \delta_3(2)\ A(2,3)\ B_{3c6} = 0.0441 \times 0.1 \times 0.5 = 0.002205 \\ \delta_3(3)\ A(3,3)\ B_{3c6} = 0.007 \times 1 \times 0.5 = 0.0035 \end{cases}$$

$$\delta_4(3) = 0.0035$$

$a_4(1) = 1, 2, 3$, $a_4(2) = 1, 2, 3$, $a_4(3) = 2$

$Z^* = 3$, $a_4(3) = 2$, $a_3(2) = 2$, $a_2(2) = 1$, $a_1(1) = 1$

$$Z_{1234} = (1, 2, 2, 3)$$

Learning for HMM

- To do the inference we need to estimate the parameters of model
- When hidden variables are observed: MLE
- When hidden variables are not observed: EM

Fully observed case

A data sample from HMM : $x^i_{1:T}$, $z^i_{1:T}$, $\theta = (\pi, A, B)$

$$P(X, Z|\theta) = \prod_{i=1}^{N} P(x^i, z^i|\theta) = \prod_{i=1}^{N} P(x^i|z^i, \theta) P(z^i|\theta)$$

$$= \prod_{i=1}^{N} \left[\prod_{t=1}^{T_i} P(x^i_t|z^i, \theta) \prod_{t=2}^{T_i} P(z^i_t|z^i_{t-1}, \theta) P(z^i_1|\theta) \right]$$

$$\sum_{l} B_{jl} = 1 , \quad \sum_{j} \pi_j = 1 \quad \sum_{j} A_{ij} = 1$$

$$\log P(X, Z \mid \theta) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P(x_t^i \mid z_t^i, \theta) + \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P(z_t^i \mid z_{t-1}^i, \theta)$$

$$+ \sum_{i=1}^{N} \log P(z_1^i \mid \theta)$$

$B_{jl} = P(x_t = l \mid z_t = j)$ let $j \in \{1 .. K\}$, $l \in \{1, .. S\}$

$$P(x_t^i \mid z_t^i, \theta) = \prod_{j=1}^{K} \prod_{l=1}^{S} \left(P(x_t^i = l \mid z_t^i = j, \theta) \right)^{I(x_t^i = l,\, z_t^i = j)}$$

$$\log P(x_t^i \mid z_t^i, \theta) = \sum_{j}^{K} \sum_{l}^{S} I(x_t^i = l, z_t^i = j) \log B_{jl}$$

$$\log P(z_t^i \mid z_{t-1}^i, \theta) = \sum_i \sum_{t=2} \sum_j \sum_k \mathbb{I}(z_t^i = k, z_{t-1}^i = j) \log A_{jk}$$

$$\log P(z_1^i \mid \theta) = \sum_i \sum_{j=1}^{K} \mathbb{I}(z_1^i = j) \log \pi_j$$

$$\frac{\partial \log P(X, Z \mid \theta)}{\partial B_{jl}} = \frac{\partial \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T_i} \mathbb{I}(x_t^i = l, z_t^i = j) \log B_{jl} + \lambda_j \left(1 - \sum_l B_{jl}\right) \right\}}{\partial B_{jl}}$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T_i} \frac{\mathbb{I}(x_t^i = l, z_t^i = j)}{B_{jl}} - \lambda_j = 0$$

$$\Rightarrow B_{jl} = \frac{\sum_i \sum_t \mathbb{I}(x_t^i = l, z_t^i = j)}{\lambda_j}$$

$$\sum_l B_{jl} = 1 \Rightarrow \lambda_j = \sum_l \sum_i \sum_t \mathbb{I}(x_t^i = l, z_t^i = j)$$

In similar way, we can derive for π_j and A_{ij}

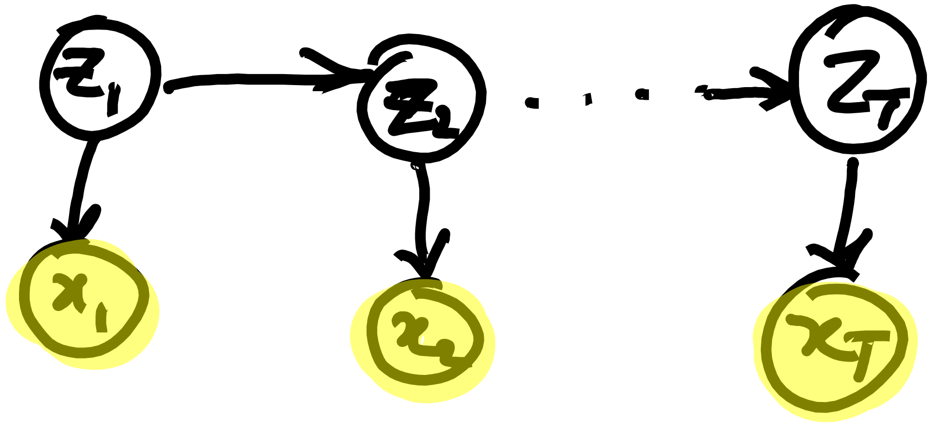
$$A_{jk} = \frac{\sum_i^N \sum_{t=2}^{T_i} \mathbb{I}(z_t^i = k, z_{t-1}^i = j)}{\sum_k \sum_i \sum_{t=2}^{T_i} \mathbb{I}(z_t^i = k, z_{t-1}^i = j)}$$

$$\pi_j = \frac{\sum_i \mathbb{I}(z_1^i = j)}{\sum_{j=1}^K \sum_{i=1}^N \mathbb{I}(z_1^i = j) = N}$$

Partially observed case

Observed data : $X_{1:T}$

Hidden data : $Z_{1:T}$



likelihood: $P(X \mid \theta) = ?$

X_i's are not independent!

Use EM to estimate parameter and make model simpler

or likelihood

E-step:

$$\log P(X|\theta) = \log \int P(X,Z|\theta)\,dZ = \log \int \frac{P(X,Z|\theta)}{q(Z|\theta^t)}\, q(Z|\theta^t)\,dZ$$

$$\geq E_Z \log P(X,Z|\theta) - \text{Const (w.r.t. } \theta)$$

$$Q(\theta,\theta^t) = E_Z \log P(X,Z|\theta) = E_Z \log P(X|Z,\theta) + E_Z \log P(Z|\theta)$$

$$P(X|Z,\theta) = \prod_{i=1}^{N} \prod_{t=1}^{T_i} \prod_{j=1}^{K} \prod_{l=1}^{S} P(x_t^i = l \,|\, z_t^i = j)^{\mathbb{I}(x_t^i = l,\, z_t^i = j)}$$

$$P(Z|\theta) = \prod_{i}^{N} \prod_{t=2}^{T_i} \prod_{j=1}^{K} \prod_{k=1}^{K} P(z_t^i = k \mid z_{t-1}^i = j; \theta)^{\mathbb{I}(x_t^i = k,\, z_{t-1}^i = j)}$$

$$\times \prod_{i=1}^{N} \prod_{j=1}^{K} P(z_1^i = j \mid \theta)^{\mathbb{I}(z_1^i = j)}$$

$$Q(\theta|\theta^{t}) = E_{z} \sum_{i} \sum_{t} \sum_{j} \sum_{l} \mathbb{I}(z_{t}^{i} = j)\, \mathbb{I}(x_{t}^{i} = l) \log B_{jl}$$

$$+ E_{z} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbb{I}(z_{t}^{i} = k, z_{t-1}^{i} = j) \log A_{jk}$$

$$+ E_{z} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{I}(z_{1}^{i} = j) \log \pi_{j}$$

$$Q(\theta|\theta^t) = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{j=1}^{K}\sum_{l=1}^{S} p(z_t^i|X^i,\theta^t)\,\mathbb{I}(x_t^i=l)\log B_{jl}$$

$$+ \sum_{i=1}^{N}\sum_{t=2}^{T_i}\sum_{j=1}^{K}\sum_{k=1}^{K} p(z_t^i=k, z_{t-1}^i=j|X^i,\theta^t)\log A_{jk}$$

$$+ \sum_{i=1}^{N}\sum_{j=1}^{K} p(z_1^i=j|X^i,\theta^t)\log \pi_j$$

$$P(z_t^i = j \mid x^i; \theta^t) = \gamma_{i,t}(j) \quad (\text{Smoothing})!$$

$$P(z_t^i = k, z_{t-1}^i = j \mid x^i, \theta) = \xi_{i,t}(j,k) \quad (\text{Twoslice smooth marginals})$$

$$P(z_1^i = j \mid x^i, \theta) = \frac{P(x^i \mid z_1^i = j) P(z_1^i = j \mid \theta)}{\sum_j P(x^i \mid z_1^i = j) P(z_1^i = j \mid \theta)} = \eta_i(j)$$

So:

$$Q(\theta \mid \theta^t) = \sum_i \sum_t \sum_j \sum_l \gamma_{i,t}(j) I(x_t^i = l) \log B_{jl} + \sum_i \sum_t \sum_j \sum_k \xi_{i,t}(j,k) \log A_{jk} + \sum_i \sum_j \eta_i(j) \log \pi_j$$

M-step:

We compute for B_{jl}. Other parameters can be calculated similarly.

Extra constraint: $\forall j \quad \sum_l B_{jl} = 1, \ \sum_k A_{jk} = 1, \ \sum_j \pi_j = 1$

So: $$\frac{\partial Q}{\partial B_{jl}} = \frac{\partial}{\partial B_{jl}} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j) I(x_t^i = l) \log B_{jl} + \lambda_j \left(1 - \sum_l B_{jl}\right) + \text{Cons} \right\}$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T_i} \frac{\gamma_{i,t}(j) I(x_t^i = l)}{B_{jl}} - \lambda_j = 0 \Rightarrow B_{jl} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j) I(x_t^i = l)}{\lambda_j}$$

$$\sum_{l} B_{jl} = 1 \Rightarrow \quad \lambda_j = \sum_{l} \sum_{i} \sum_{t} \gamma_{it}(j)\, I(x_t^i = l) = N_j$$

Exercise: drive the updates for π_j, A_{jk}?

Exercise: drive MLE and E_M when $p(x|z)$ is Gussian?