Sequential models

- In many settings data are sequential
  - Language data
  - Time indexed data
  - DNA

- We will study 2 sequential models — HMM & Kalman filter
  (later make EM concrete...)
- The HMM is a sequential generalization of the mixture model at time t
- Idea — the chosen mixture component depends on the component at time t-1.
  (Contrast to a mixture, where the components are assumed i.i.d.)

- The HMM models transitions between the components.

[Show data generated from this picture.]

HMM modeling assumptions

\[ p(z_{1:T}, x_{1:T}) = p(z_1)p(x_1 | z_1) \prod_{t=2}^{T} p(z_t | z_{t-1}) p(x_t | z_t) \]

\( Z_t \): discrete (K values)
\( X_t \): anything (e.g., \( \mathbb{R}^2 \))

(or group the x's all together...)
Parameters are:

- Data generating distributions $\Theta_k$ (also called emission probs)

$$p(x_t, l_t) = \prod_{k=1}^{K} \pi_{l_t} \theta_{x_t}$$

- As for mixtures, this can be anything.

- To be concrete, make $\Theta_k$ parameters to 2D Gaussians (note: MN in genetics)

- Transition matrix $A$

  - Probability of the next state given the current state.

  $$A_{i,k}$$

  - Recall that $Z_t$ is an indicator vector.

  - Thus,

  $$p(Z_{t+1} | Z_t) = \prod_{k=1}^{K} \prod_{j=1}^{K} \alpha_{j,k} \pi_{j} \pi_{k}$$

- Probability of the first state (to start things off)

  $$p(Z_1) = \prod_{k=1}^{K} \pi_k \prod_{k=1}^{K} Z_k$$

- Show a state transition diagram and emphasize it is not a G.M.

Applications

- Speech recognition
- Part of speech tagging
- Handwriting recognition
- DNA analysis
- Others...
Fitting an HMM

EM algorithm.

1. Write down the expected complete log likelihood

\[
\mathbb{E} \left[ \log p(\mathbf{Z}_{1:T}, \mathbf{X}_{1:T}) \right] = \mathbb{E} \left[ \sum_{t=1}^{T} \log \left( \prod_{k=1}^{K} p(z_t = k) \prod_{t=2}^{T} p(z_t | z_{t-1}) \prod_{t=1}^{T} p(x_t | z_t) \right) \right]
\]

\[
= \sum_{k=1}^{K} \mathbb{E}[Z_i] \pi_k^{(1)} \sum_{t=2}^{T} \sum_{j=1}^{K} \mathbb{E}[Z_{t-1} = j, Z_t = k] \log \pi_{j,k} + \sum_{t=1}^{T} \mathbb{E}[Z_t] \log p(x_t | \theta_k)
\]

All expectations are taken w.r.t. the posterior, \( p(\mathbf{Z}_{1:T} | \mathbf{X}_{1:T}) \)

And, since \( Z_t \) is an indicator vector,

\[
\mathbb{E}_x[Z_i] = p(Z_i = k | \mathbf{X}_{1:T}) \quad \text{and} \quad \mathbb{E}_x[Z_t] = p(Z_t = k | \mathbf{X}_{1:T})
\]

\[
\mathbb{E}_x[Z_{t-1} = j, Z_t = k] = p(Z_{t-1} = j, Z_t = k | \mathbf{X}_{1:T})
\]

2. M-step: Note as usual that the log complete likelihood decomposes.

\[
\pi_k = \mathbb{E}[Z_i] / \sum_{j} \mathbb{E}[Z_i] \quad \text{normalized expected counts}
\]

\[
a_{j,k} = \sum_{t=2}^{T} \mathbb{E}[Z_{t-1} = j, Z_t = k] / \sum_{t=2}^{T} \sum_{k=1}^{K} \mathbb{E}[Z_{t-1} = j, Z_t = k]
\]

\( ? \)

Note: This stays at \( j \).
\[ \Theta_k: \text{weighted maximum likelihood, as for mixture models.} \]

Gaussian — \[ \mu_k = \frac{\sum_{t=2}^{T} \mathbb{E}[Z_t^k] x_t}{\sum_{t=2}^{T} \mathbb{E}[Z_t^k]} \]

\[ \Sigma_k = \frac{\sum_{t=2}^{T} \mathbb{E}[Z_t^k] (x_t - \mu_k)(x_t - \mu_k)^T}{\sum_{t=2}^{T} \mathbb{E}[Z_t^k]} \]

Discrete (e.g., biology)—

\[ p(x_t | \theta_k) = \prod_{i=1}^{V} \frac{\theta_{ki}^{x_{ti}}}{\sum_{i=1}^{V} \theta_{ki}} \]

\[ \theta_{ki} = \frac{\sum_{t=2}^{T} \mathbb{E}[Z_t^k] x_{ti}}{\sum_{t=2}^{T} \mathbb{E}[Z_t^k]} \]

Others... Poisson, etc.

\textbf{E-step} — also used for prediction.

This is the interesting part

\[ \mathbb{E}[Z_t | x_{1:T}] = p(z_t | x_{1:T}) \]

\[ = p(z_t, x_{1:T}) / p(x_{1:T}) \]

\[ = \frac{\prod_{t=1}^{T} p(x_{1:t} | z_t) p(x_{t+1:T} | z_t)}{p(x_{1:T})} \]

We will compute these terms recursively —

\[ \alpha(z_t) \overset{\Delta}{=} p(x_{1:t}, z_t) \quad [K\text{-vector}] \]

\[ \beta(z_t) \overset{\Delta}{=} p(x_{t+1:T} | z_t) \quad [K\text{-vector}] \]

\[ = \alpha(z_t) \beta(z_t) / p(x_{1:T}) \]
• Also lets us compute:

\[
\mathbb{E}[z_{t-1} z_t | x_{1:T}] = p(z_{t-1}, z_t | x_{1:T}) \\
= p(x_{1:T}, z_{t-1}, z_t) / p(x_{1:T}) \\
= p(x_{1:t-1}, z_t) p(z_t | z_{t-1}) p(x_t | z_t) p(z_{t-1}) / p(x_{1:T}) \\
= \alpha(z_{t-1}) p(z_t | z_{t-1}) p(x_t | z_t) \beta(z_t) / p(x_{1:T})
\]

(Note - recursively applying this gives us any subsequence. But, we only need pairs for EM.)

• The likelihood is easy from these quantities, from \( \mathcal{X} \)

\[
\sum_{z_t} \alpha(z_t) \beta(z_t) / p(x_{1:T}) = 1 \Rightarrow p(x_{1:T}) = \sum_{z_t} \alpha(z_t) p(z_t)
\]

• We are left to compute \( \alpha, \beta \): Recursive algorithm. Sometimes called forward-backward, \( \alpha-\beta \), is an instance of propagation.

\[
\alpha(z_t) = p(z_t, x_t) \\
= \Pi_{z_{<t}} p(x_t | z_t) \quad \text{for slight abuse of notation.}
\]

\[
\alpha(z_{t+1}) = p(x_{1:t+1}, z_{t+1}) \\
= p(x_{1:t+1} | z_{t+1}) p(z_{t+1}) \\
= p(x_{1:t+1} | z_{t+1}) p(x_{t+1} | z_{t+1}) p(z_{t+1}) \\
= \sum_{z_{<t+1}} \alpha(z_{t+1}) p(x_{t+1} | z_{t+1}) p(z_{t+1}) \\
= \sum_{z_{<t+1}} \alpha(z_{t+1}) p(x_{t+1}) p(z_{t+1})
\]
\[
\sum_{z_t} p(x_{t+1} | z_t) p(z_{t+1} | z_t) p(z_t) p(x_{t+1} | z_{t+1}) [z_{t+1} \perp x_{t+1} | z_t] = \sum_{z_t} p(x_{t+1} | z_t) p(z_{t+1} | z_t) p(x_{t+1} | z_{t+1}) \alpha(z_t) \beta(z_{t+1}),
\]

emission prob.

Complexity for each step is \(O(K^2)\):

- for each value of \(z_{t+1}\), we sum over \(K\) elements.
- Complexity for the recursion is \(O(TK^2)\).

Turn to the \(\beta\) recursion, i.e., the backward step.

\[
\beta(z_t) = p(x_{t+1:T} | z_t) = \sum_{z_{t+1}} p(x_{t+1:T} | z_{t+1}, z_t) p(z_{t+1} | z_t)
\]

\[
= \sum_{z_{t+1}} p(x_{t+1:T} | z_{t+1}, z_t) p(z_{t+1} | z_t) [\text{chain rule}]
\]

\[
= \sum_{z_{t+1}} p(x_{t+2:T} | z_{t+1}) p(x_{t+1} | z_{t+1}) p(z_{t+1} | z_t)
\]

\[
\beta(z_{t+1}) \text{ emission } \alpha(z_{t+2:T} | z_{t+1}) \beta(z_{t+1})
\]

Setting \(\beta(z_T) = 1\) makes the recursion work.

(Not \(\beta(z_T)\) is meaningless because \(x_{T+1}\) does not exist.)

or \(\beta(z_{T-1}) = p(x_T | z_{T-1}) = \sum_{z_t} p(x_T | z_t) p(z_t | z_{T-1})\)
Can also compute the predictive dist:
\[ p(x_{T+1} | x_{1:T}) = \sum_{z_{T+1}} p(x_{T+1} | z_{T+1}) p(z_{T+1} | x_{1:T}) \]

Many variants of HMMs:

- **Hierarchical HMMs** - state transitions at different levels
- **Generalized HMMs** - random number of emissions per time point
- **Shadower** - HMM respecting a phylogenetic tree
- **Factorial HMMs** - \( \begin{pmatrix} 0 \\ 1 \end{pmatrix} \) factorial states.

"Infinite HMMs" - \( K \) is unbounded