Sequential data

Sequential data is everywhere.

Examples:
- spoken language (word production)
- written language (sentence level statistics)
- weather
- human movement
- stock market data

Sequential data

Graphical models for such data?

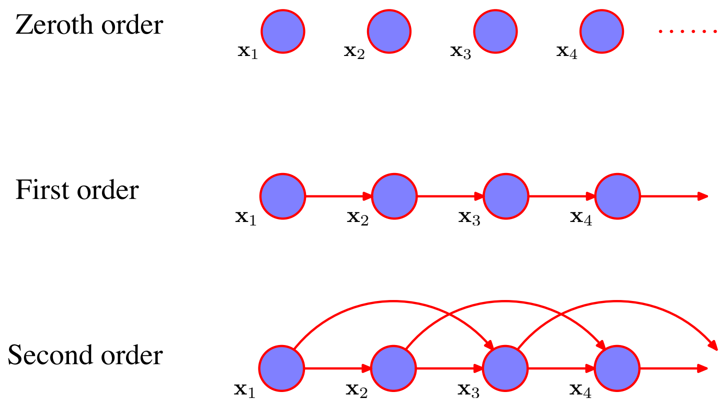
The complexity of the representation seems to increase with time.

Observations over time tend to depend on the past.

We can simply life by assuming that the distant past does not matter.

If we assume that history does not matter other than the immediate previous state, we have a first order Markov model.

Markov chains



Zeroth order

First order

Second order

Temporal statistical clustering

In sequence data, cluster membership can have temporal (or sequential) structure.

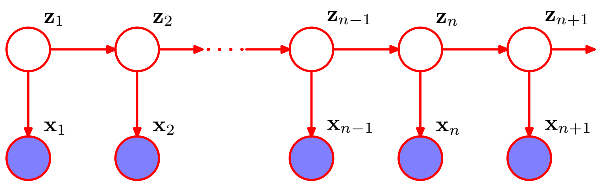
The data comes from the current cluster (as usual), but what is the next cluster?

Example, rain and sleet come from “stormy” and sunshine from “fair weather”.

But now, our hidden cluster variables depend on the past.

Temporal statistical clustering

Hidden Markov Model (HMM).



The particular state encodes the important part of history.

Temporal statistical clustering

$$p\left(x_n \middle| z_n\right)$$

Once you know your cluster, things are easy.

But, the cluster is now changing over time.

Markovian assumptions

As before, if the current state depends on only the previous state, we have a first order Markov model.

The basic HMM is like a mixture model, with the mixture components being used for the current observations depending only on the last mixture component.

$$z_{n+1} \perp z_{n-1} \,\big|\, z_n$$

Markovian assumptions

Represent each component as a “state”.

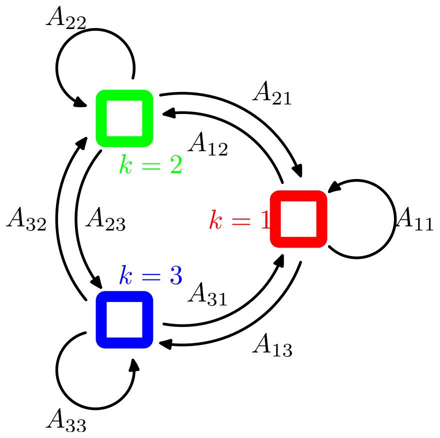
Then, for first order Markov models, this leads to the concept of “transition” probabilities.

$$A_{jk} \equiv p\left(z_{nk} = 1 \middle| z_{nj} = 1\right)$$

$$0 \le A_{jk} \le 1 \quad \text{and} \quad \sum_k A_{jk} = 1$$

Transition matrix representation

(Not a graphical model)



Starting state

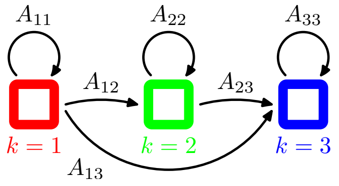
Our HMM will be a generative model, so we need to know how to start.

$$\pi_k \equiv p\left(z_{1k}=1\right)$$

with $0 \le \pi_k \le 1$ and $\sum_k \pi_k = 1$

Left to right HMM

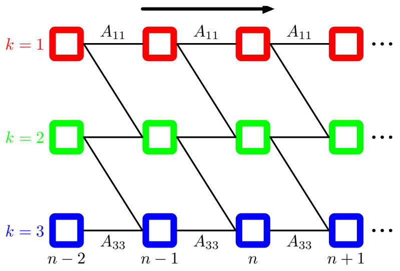
Constrain state number to increase



(State transition diagram)

Left to right HMM

Even more constrained, left to right HMM with single state jumps.



(Lattice diagram)

HMM parameter summary

$\theta = \{\pi, A, \phi\}$

π is probability over initial states

A is transition matrix

ϕ are the data emmission probabilities
(e.g., means of Gaussians)

Data distribution from an HMM

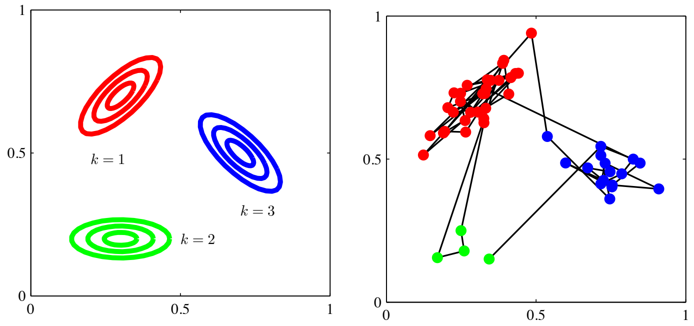
$p(X|\theta)$ is a marginalization over Z.

$p(X,Z|\theta) = ?$

Data distribution from an HMM

$$p(X,Z|\theta) = p(z_1|\pi)\left[\prod_{n=2}^{N} p(z_n|z_{n-1}, A)\right]\prod_{m=1}^{N} p(x_m|z_m, \phi)$$

Data distribution from an HMM



Transition probability to another state is 5%

Classic HMM computational problems

Given data, what is the HMM (**learning**).

Given an HMM, what is the **distribution over the state** variables.

Given an HMM, what is the most likely **state sequence** for some data?

Learning the HMM (sketch)

If we know the states, we can compute the parameters.

If we know the parameters, we can compute the states (if we know how to solve the second problem).

General EM algorithm

1. Choose initial values for $\theta^{(s=1)}$
   (can also do assignments, but then jump to M step).
2. E step: Evalute $p\left(Z \middle| X, \theta^{(s)}\right)$
3. M step: Evalute $\theta^{(s+1)} = \underset{\theta}{\arg\max}\left\{\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right)\right\}$

   where $\mathrm{Q}\left(\theta^{(s+1)}, \theta^{(s)}\right) = \sum_{Z} p\left(Z \middle| X, \theta^{(s)}\right) \log\left(p\left(X, Z \middle| \theta^{(s+1)}\right)\right)$
4. Check for convergence; If not done, goto 2.

★ At each step, our objective function is increases unless it is at a local maximum. It is important to check this is

EM for HMM (sketch)

In the simple clustering case (e.g., GMM), the E step was simple. For HMM it is a bit more involved.

The M step works a lot like the GMM. Consider it first.

EM for HMM (sketch)

$$p\left(X,Z\middle|\theta\right)=p\left(z_1\middle|\pi\right)\left[\prod_{n=2}^{N}p\left(z_n\middle|z_{n-1},A\right)\right]\prod_{n=1}^{N}p\left(x_n\middle|z_n,\phi\right)$$

$$=\prod_{k=1}^{K}\pi^{z_{1k}}\left[\prod_{n=2}^{N}\prod_{j=1}^{K}\prod_{k=1}^{K}\left(p\left(z_n\middle|z_{n-1},A\right)\right)^{z_{n-1,j}\bullet z_{n,k}}\right]\prod_{n=1}^{N}\prod_{k=1}^{K}\left(p\left(x_n\middle|z_n,\phi\right)\right)^{z_{nk}}$$

$$\log\left(p\left(X,Z\middle|\theta\right)\right)=$$

$$\sum_{k=1}^{K}z_{1k}\log\left(\pi\right)+\sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}z_{n-1,j}z_{n,k}\log\left(p\left(z_n\middle|z_{n-1},A\right)\right)+\sum_{n=1}^{N}\sum_{k}^{K}z_{nk}\log\left(p\left(x_m\middle|z_m,\phi\right)\right)$$

EM for HMM (sketch)

$$\log\left(p\left(X,Z\middle|\theta\right)\right)=$$

$$\sum_{k=1}^{K}z_{1k}\log\left(\pi\right)+\sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}z_{n-1,j}z_{n,k}\log\left(p\left(z_n\middle|z_{n-1},A\right)\right)+\sum_{n=1}^{N}\sum_{k}^{K}z_{nk}\log\left(p\left(x_m\middle|z_m,\phi\right)\right)$$

Now define

$$\gamma\left(z_n\right)=p\left(z_n\middle|X,\theta^{(s)}\right)$$

$$\xi\left(z_{n-1},z_n\right)=p\left(z_{n-1},z_n\middle|X,\theta^{(s)}\right)$$

EM for HMM (sketch)

$$\log\left(p\left(X,Z\middle|\theta\right)\right)=$$

$$\sum_{k=1}^{K}z_{1k}\log\left(\pi\right)+\sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}z_{n-1,j}z_{n,k}\log\left(p\left(z_n\middle|z_{n-1},A\right)\right)+\sum_{n=1}^{N}\sum_{k}^{K}z_{nk}\log\left(p\left(x_m\middle|z_m,\phi\right)\right)$$

By analogy with the GMM

$$Q\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_{z}p\left(Z\middle|\theta^{(s)}\right)\log\left(X,Z\middle|\theta^{(s+1)}\right)$$

$$=\sum_{k=1}^{K}\gamma\left(z_{1k}\right)\log\left(\pi\right)+\sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}\xi\left(z_{n-1,j}z_{n,k}\right)\log\left(p\left(z_n\middle|z_{n-1},A\right)\right)$$

$$+\sum_{n=1}^{N}\sum_{k}^{K}\gamma\left(z_{nk}\right)\log\left(p\left(x_m\middle|z_m,\phi\right)\right)$$

EM for HMM (sketch)

Doing the maximization using Lagrange multipliers

$$\pi_k=\frac{\gamma\left(z_{1k}\right)}{\sum_{k'}\gamma\left(z_{1k'}\right)}$$ (Much like the mixture model case)

$$A_{jk}=\frac{\sum_{n=2}\zeta\left(z_{n-1,j},z_{nk}\right)}{\sum_{k'}\sum_{n=2}\zeta\left(z_{n-1,j},z_{nk'}\right)}$$

EM for HMM (sketch)

The maximization of $p\left(x_n \middle| \phi\right)$ is exactly the same as the mixture model.

For example, if we have Gaussian emmisions, then

$$\mu_k = \frac{\sum_n x_n \gamma\left(z_{nk}\right)}{\sum_n \gamma\left(z_{nk}\right)}$$