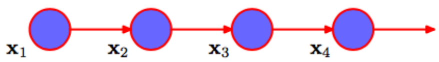
13 Sequential Data

1. **Stationary sequential distribution**
2. **Markov Model** assume future prediction depends on the most recent observations
3. **State Space Model**

Definition. ***Markov Model*** *models observations that are not i.i.d. If each of conditional distribution is independent of all previous observations except the most recent, then we obtain first-order Markov chain,*

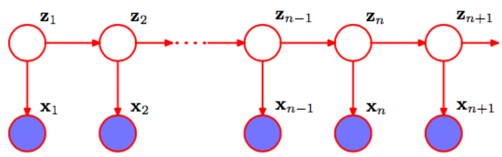


$$p(\mathbf{x}_1, \cdots, \mathbf{x}_N) \overset{product\,rule}{=} \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{x}_1, \cdots, \mathbf{x}_{n-1}) \overset{markov}{=} p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1})$$

the distribution of prediction depends only on value of immediate preceding observation

$$p(\mathbf{x}_n|\mathbf{x}_1, \cdots, \mathbf{x}_{n-1}) = p(\mathbf{x}_n|\mathbf{x}_{n-1})$$

where $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ is fixed and hence the model is ***stationary****. We can extend to M^{th} order Markov chain in which the conditional distribution for a particular variable depends on the previous M variables. However we have exponentially more parameters to maintain. So instead we introduce latent variables to the Markov chain, with each observation conditioned on the state of the corresponding latent variable. This gives rise to* ***hidden markov model*** *if latent variable is discrete and* ***linear dynamic systems*** *if both latent and observed variables are Gaussian.*



$$p(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{z}_1, \cdots, \mathbf{z}_N) = p(\mathbf{z}_1) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

and we note $\mathbf{z}_{n+1} \perp\!\!\!\perp \mathbf{z}_{n-1} \,|\, \mathbf{z}_n$. Also, there is always a nonblocking path connecting any observation. So any prediction for $\mathbf{x}_{n+1}$ does not exhibit conditional independence property, so depends on all previous observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$

Definition. ***Hidden Markov Model***

1. ***Model Formulation*** *Like mixture model where choice of mixture component for each observation not selected independently but depends on choice of component for previous observation. Latent variables are discrete multinomial varaible* $\mathbf{z}_n$ *describing which component of mixture is responsible for generating the corresponding observations. We assume a constant* ***transition probability*** $\mathbf{A}$ *cross all hidden states*

$$A_{jk} = p(z_{nk} = 1 | z_{n-1,j} = 1) \qquad 0 \leq A_{jk} \leq 1 \quad \sum_k A_{jk} = 1$$

i.e. transition probability from picking j*-th component to picking* k*-th component. We define* $\boldsymbol{\pi}$ *be* ***initial probability*** *for* $\mathbf{z}_1$ *since it does not have parent node, in other words* $\pi_k = p(z_{1k} = 1)$*. Therefore we can define probability distribution for edges connecting hidden states* $p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A})$ *and* $p(\mathbf{z}_1|\boldsymbol{\pi})$

$$p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) = \prod_{k=1}^K \prod_{j=1}^K (A_{jk})^{z_{n-1,j} z_{nk}} \qquad p(\mathbf{z}_1|\boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{z_{1k}}$$

We define ***emission probabilities*** *as* $p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi})$ *be responsible for converting state paths into a sequence of observable varaibles.* $\boldsymbol{\phi}$ *is a set of parameters governing the distribution that is constant cross all emission probabilities under a homogeneous model. The emission probability consists of* K *possible different distributions corresponding to* K *possible states of* $\mathbf{z}_n$

$$p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi}) = \prod_{k=1}^K p(\mathbf{x}_n|\boldsymbol{\phi}_k)^{z_{nk}}$$

Therefore the ***joint distribution*** *of both observed and latent varaibles is given by*

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \prod_{n=2}^N p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \prod_{m=1}^N p(\mathbf{x}_m|\mathbf{z}_m, \boldsymbol{\phi})$$

given $\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi}\}$.

2. ***A Generative View*** *we can treat HMM as follows*

   (a) *pick initial latent variable* $\mathbf{z}_1$ *given* $\boldsymbol{\pi}$*, then sample* $\mathbf{x}_1$

   (b) *choose next latent variable using* $\mathbf{A}$*, then sample from the emission probabilities*

3. ***MLE for HMM*** *Given dataset* $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ *want to determine parameter* $\boldsymbol{\Theta}$ *with maximum likelihood. Want to maximize the likelihood*

$$p(\mathbf{X}|\boldsymbol{\Theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})$$

which does not factor over n easily. Have to consider K^N different possible Zs. Not feasible. We solve this by rearranging the summation such that the cost scales linearly instead of exponentially

$$
\begin{aligned}
p(\mathbf{X}|\mathbf{\Theta}) &= \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\mathbf{\Theta}) \\
&= \sum_{\mathbf{z}_1,\cdots,\mathbf{z}_n} p(\mathbf{z}_1, \mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_1} p(\mathbf{z}_1) p(\mathbf{x}_1|\mathbf{z}_1) \sum_{\mathbf{z}_2} p(\mathbf{z}_2|\mathbf{z}_1) p(\mathbf{x}_2|\mathbf{z}_2) \cdots \sum_{\mathbf{z}_N} p(\mathbf{z}_N|\mathbf{z}_{n-1}) p(\mathbf{x}_N|\mathbf{z}_N)
\end{aligned}
$$

The emission probability maybe very complex leading to no closed form solution for the likelihood function. We us EM algorithm to solve the problem