Statistical Machine Learning

Christian Walder

Machine Learning Research Group
CSIRO Data61

and

College of Engineering and Computer Science
The Australian National University

Canberra
Semester One, 2020.

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Outlines

Part XXI

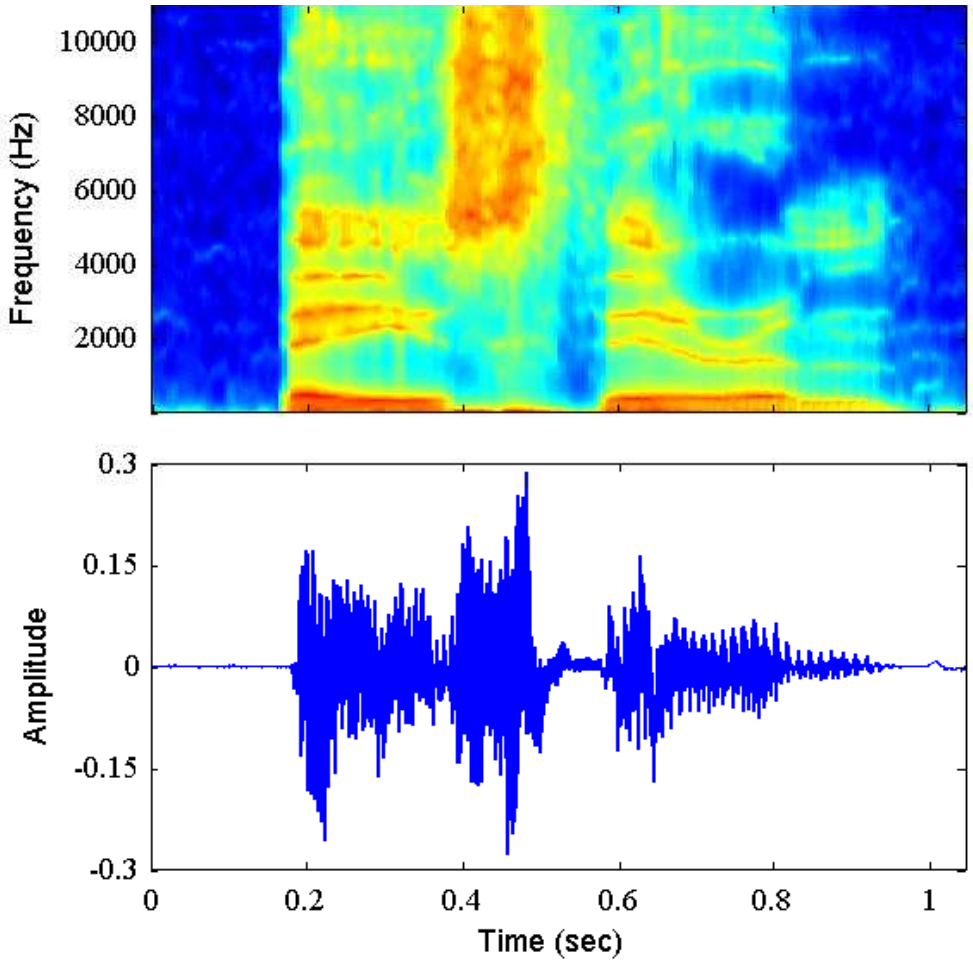
Sequential Data 1

Sequential Data

- Often data is not *i.i.d.*, *e.g.*
  - time series (finance, daily rainfall, speech, . . .),
  - DNA sequences,
  - Natural language text (English, Chinese, . . .).
- Current data may not be independent of previous data.
- The distribution of the observed data may change as the sequence progresses.
- Note: Use ‘past’ and ‘future’ to describe an order on the observations, but the concept of sequence is not restricted to temporal sequences.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Sequential Data Example



Data61 | CSIRO
The Australian National University

Stationary vs. Nonstationary Sequential Distributions

- Stationary:
  Data evolves in time, but the distribution from which it is drawn stays the same.
- Nonstationary:
  The generative distribution changes itself with time.

We will focus on the stationary case.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

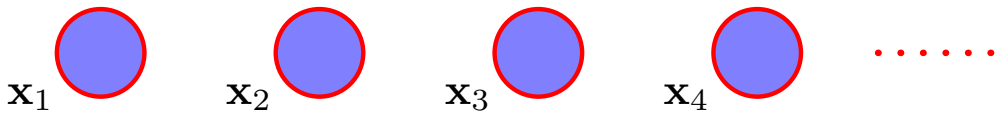
Sequence Modelling Challenges

- Goal:
  Predict the next value in a sequence.
- Assumption:
  Not all previous data equally influence the next value.
- Technical problem:
  How to store an ever growing history of observations?)
- Assume that recent observations are more likely to be informative for the prediction of the next value than less recent observations.
- Is this always a good assumption?

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Limitations of Modelling a Sequence as i.i.d.

- Example:
  Binary variable recording whether it rained or not on a day.
- Only information from such a model: frequency of rainfall.
- Can only use the frequency to predict whether it rains tomorrow or not. (Maybe not so bad for Canberra ;-)
- Usually:
  Observing whether it rained today helps to predict the weather for tomorrow.
- Need to relax the *i.i.d.* assumption to grasp this idea.



Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

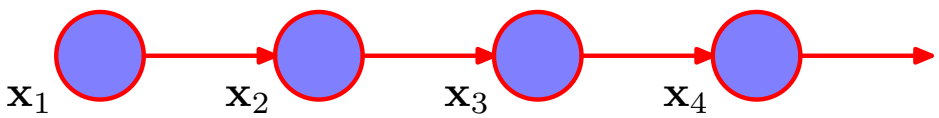
The Markov Model

- One of the simplest ways to relax the i.i.d. assumption.
- Use the product rule to exactly express the joint distribution

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

- Assume that each of the conditional expressions depends only on the most recent.
- First-order Markov chain

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



Markov Model: Forecasting

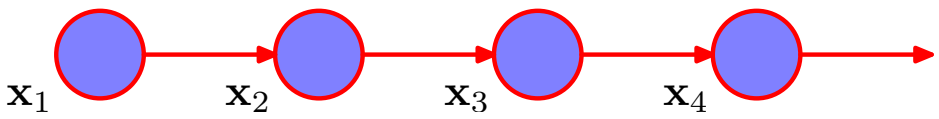
- Given the factorisation of the first-order Markov chain

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

- What is the conditional distribution for observation $\mathbf{x}_n$ given the previous observations?

$$\begin{aligned} p(\mathbf{x}_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})} \\ &= \frac{p(\mathbf{x}_1) \prod_{i=2}^{n} p(\mathbf{x}_i \mid \mathbf{x}_{i-1})}{p(\mathbf{x}_1) \prod_{j=2}^{n-1} p(\mathbf{x}_j \mid \mathbf{x}_{j-1})} \\ &= p(\mathbf{x}_n \mid \mathbf{x}_{n-1}) \end{aligned}$$

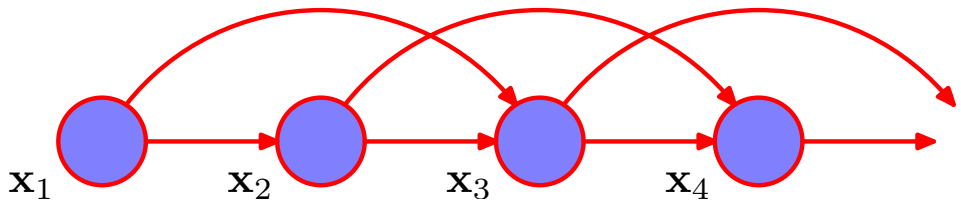
Also clear from the graphical D-separation:



Second-order Markov Chain

- Assume that, say, the trend in the previous observations provides important information in predicting the next value.
- For a trend, we need at least two previous observations.

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)\, p(\mathbf{x}_2 \mid \mathbf{x}_1) \prod_{n=3}^{N} p(\mathbf{x}_n \mid \mathbf{x}_{n-1}, \mathbf{x}_{n-2})$$



Data61 | CSIRO
The Australian National University

M^{th}-order Markov Chain

- Extend this idea to an M^{th}-order Markov chain in which the probability of each observation depends on the previous M observations

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)\, p(\mathbf{x}_2 \mid \mathbf{x}_1) \ldots p(\mathbf{x}_M \mid \mathbf{x}_1, \ldots \mathbf{x}_{M-1}) \times \prod_{n=M+1}^{N} p(\mathbf{x}_n \mid \mathbf{x}_{n-1}, \mathbf{x}_{n-2} \ldots \mathbf{x}_{n-M})$$

- What is a reasonable M?

M^{th}-order Markov Chain

Assuming K states for each variable, how many parameters does a M^{th}-order Markov chain have?

- $M = 0$: no Markov parameter, *i.i.d.* data
- $M = 1$: First-order Markov chain, $K - 1$ parameters for each of the K states of the previous observation. Number of parameters: $K(K-1)$.
- M : M^{th}-order Markov Chain, $K - 1$ parameters for each of the K states of the previous M obervation. Number of parameters: $K^M(K-1)$.
- Number of parameters grows exponentially with the order of the Markov chain. Impractical for larger M.

(Neglects the special case of the first few observations)

Extending the Model

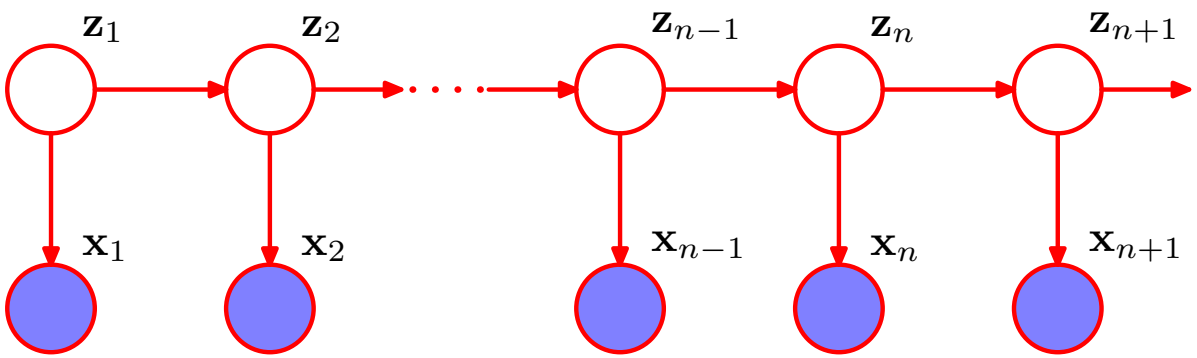
- Goal: We want a model which is NOT restricted to the Markov assumption to any order. BUT can be specified by a limited number of free parameters.
- Use the idea of latent variables to construct a rich class of models out of simple components.
- (Remember mixture of Gaussians.)

State Space Model

- For each observation $\mathbf{x}_n$, a latent variable $\mathbf{z}_n$ is added.
- The type and dimensionality of $\mathbf{z}_n$ can differ from $\mathbf{x}_n$.
- Assume Markovian latent variables, *i.e.*

$$\mathbf{z}_{n+1} \perp\!\!\!\perp \mathbf{z}_{n-1} \,|\, \mathbf{z}_n$$

- As a directed graphical model:



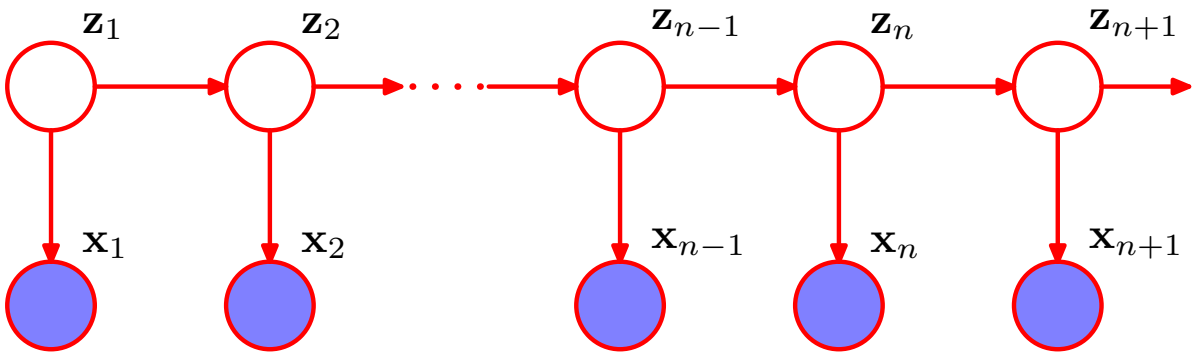
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

State Space Model - Basic Properties

- Joint distribution:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \left[\prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n)$$

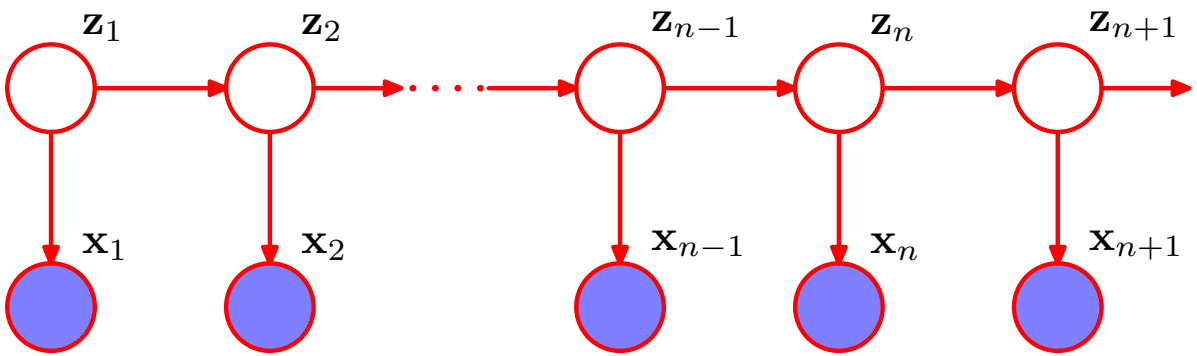
- D-separation:
  - all pairs of $\mathbf{x}_i$ and $\mathbf{x}_j$ are not D-separated.
  - Forecast $p(\mathbf{x}_{n+1} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)$ depends on the entire history.
  - The $\mathbf{x}_n$ are not (fixed order) Markovian.



State Space Model - Two Important Varieties

Two important models described by this graphical model.

- Hidden Markov Model or HMM:
  - Latent $\mathbf{z}_n$ are discrete.
  - Observed $\mathbf{x}_n$ may be discrete or continuous.
- Linear Dynamical System:
  - Latent and observed variables are continuous.
  - In particular, latent and observed variables are Gaussian with a linear-Gaussian dependence of the conditional distribution on their parents.



Data61 | CSIRO
The Australian National University

Hidden Markov Model

- State space model with discrete latent variables.

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N) = p(\mathbf{z}_1)\left[\prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1})\right]\prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n)$$

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

- Each step can be viewed as an extension of the mixture distribution model where each of the component densities is $p(\mathbf{x} \mid \mathbf{z})$. Choice of mixture component depends now on the previous state and is represented by $p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$.
- Latent variables are discrete multinomial variables $\mathbf{z}_n$ describing which component of the mixture is responsible for generating the corresponding observable $\mathbf{x}_n$.

Hidden Markov Model: Latent Dynamics

- Assume 1-in-K coding scheme.
- Each latent variable $\mathbf{z}_n$ has K different states.
- Initial latent node $\mathbf{z}_1$ has no parent, therefore the marginal distribution is given by a vector of probabilities $\boldsymbol{\pi}$ with $\pi_k = p(z_{1k} = 1)$ so that

$$p(\mathbf{z}_1 \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \qquad \sum_k \pi_k = 1$$

- The conditional distribution $p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$ is a table (matrix) with $K \times K$ entries, denoted by $\mathbf{A} \in [0,1]^{K \times K}$.
- The elements of $\mathbf{A}$ are called transition probabilities

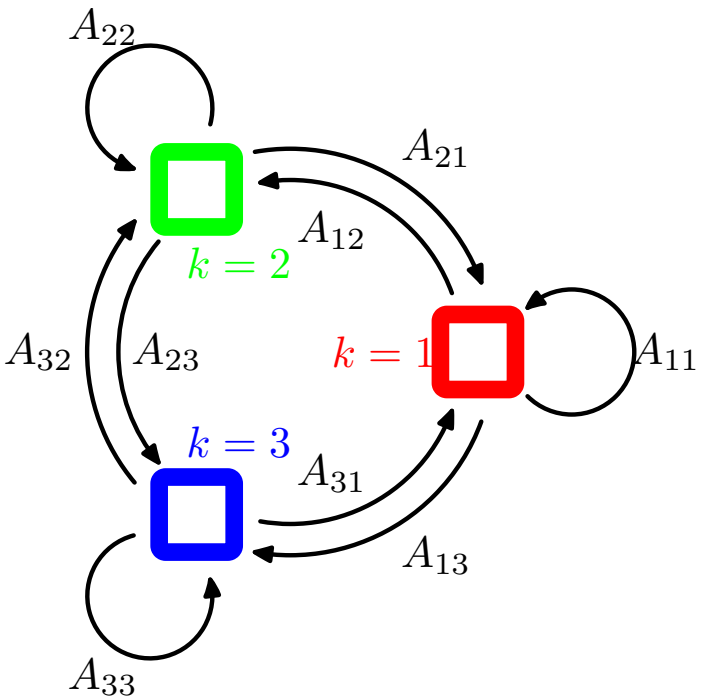
$$A_{jk} = p(z_{n,k} = 1 \mid z_{n-1,j} = 1).$$

satisfying $0 \leq A_{jk} \leq 1$ and $\sum_{k=1}^{K} A_{jk} = 1$.

Hidden Markov Model - Transition Diagram

- Given the state $\mathbf{z}_{n-1}$ at step $n-1$, now what is the next state $\mathbf{z}_n$?

$$p(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \mathbf{A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j}\, z_{nk}}$$

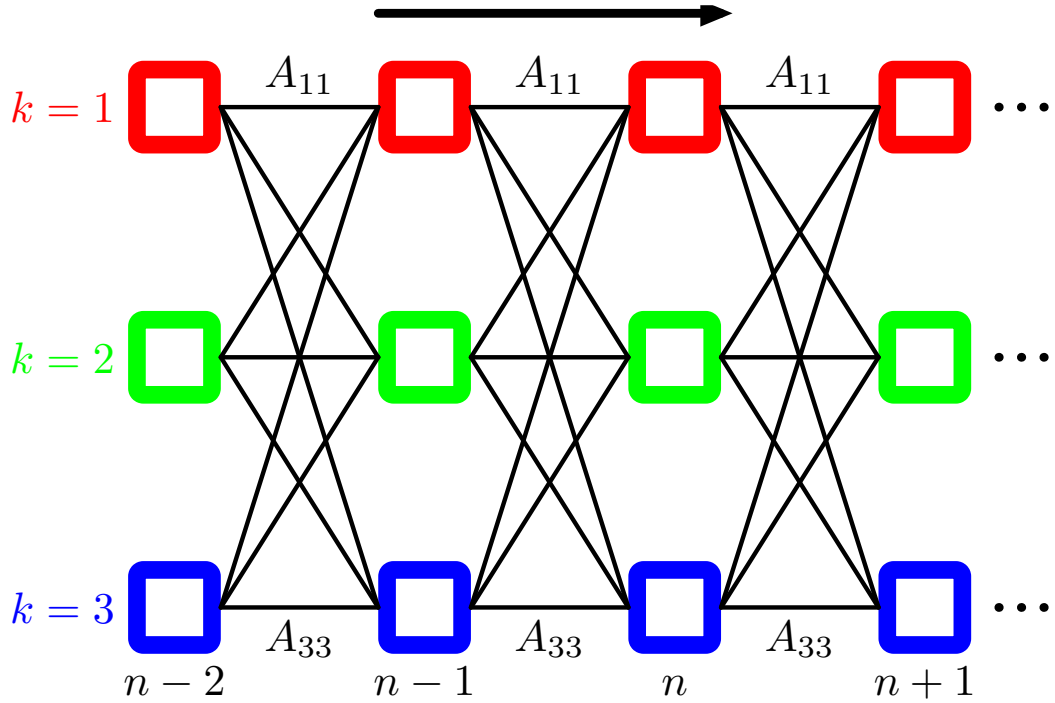


Transition Diagram for a model with three possible states. Black lines denote the elements of the transition matrix A_{jk}.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

HMM - Unfolded Transition Diagram

- Unfold the transition diagram over the steps to get a lattice, or trellis. (Note, the transitions $\mathbf{A}$ in each step can be different.)



Unfolded Transition Diagram for a model with 3 possible states. Each column corresponds to one latent variable $\mathbf{z}_n$.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Hidden Markov Model - Emission Probabilities

- Complete the HMM by defining the emission probabilities $p(\mathbf{x}_n \mid \mathbf{z}_n, \phi)$ where ϕ is a set of parameters governing the conditional distributions.
- As $\mathbf{x}_n$ is observed, and ϕ is given, $p(\mathbf{x}_n \mid \mathbf{z}_n, \phi)$ is a K-dimensional vector corresponding to the K possible states of the binary vector $\mathbf{z}_n$.
- Emission probabilities can be represented as

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

Hidden Markov Model - Homogeneous

- Remember: Transition probability in a Markov chain $T(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$.
- A Hidden Markov Model is called homogeneous if the transition probabilities are the same for all steps n

$$p(\mathbf{z}_n \mid \mathbf{z}_{n-1}) = p(\mathbf{z}_{n-1} \mid \mathbf{z}_{n-2}) \qquad \forall\, n = 3, \ldots, N$$

- Assume a homogeneous HMM in the following.

Data61 | CSIRO
The Australian National University

Homogeneous Hidden Markov Model

- Joint probability distribution over both latent and observed variables

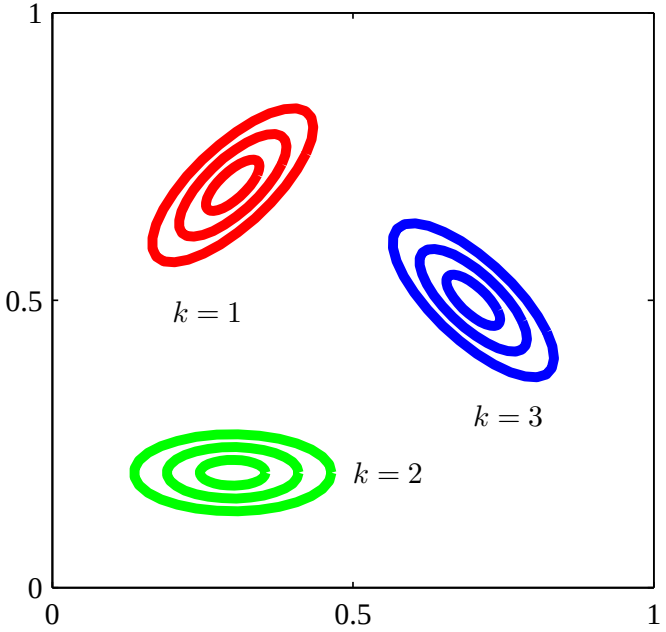
$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = p(\mathbf{z}_1 \mid \boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \mathbf{A}) \right] \prod_{m=1}^{N} p(\mathbf{x}_m \mid \mathbf{z}_m, \phi)$$

  where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$, and $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{A}, \phi\}$.

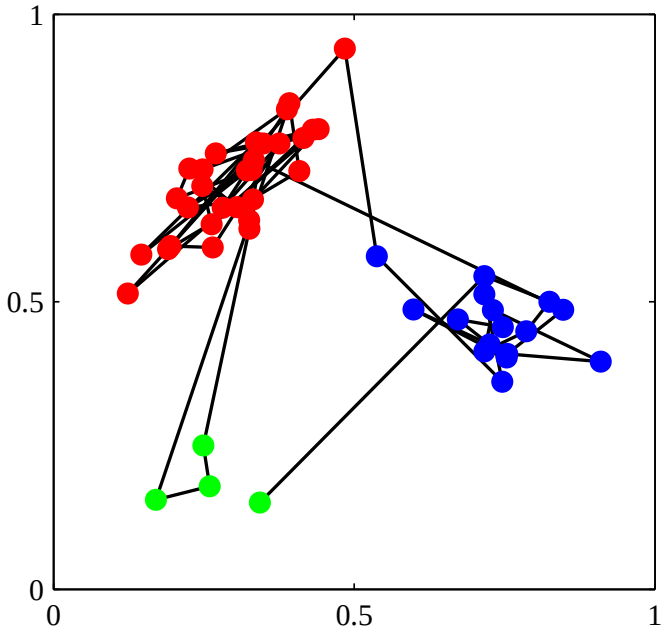
- Most of the discussion will be independent of the particular choice of emission probabilities (e.g. discrete tables, Gaussians, mixture of Gaussians).

Homogeneous HMM - Generative View

- Sampling from a hidden Markov model having a 3-state latent variable $\mathbf{z}$ and a Gaussian emission model $p(\mathbf{x} \mid \mathbf{z})$ where $\mathbf{x} \in \mathbb{R}^2$.



Contours of constant probability for the emission probabilities.



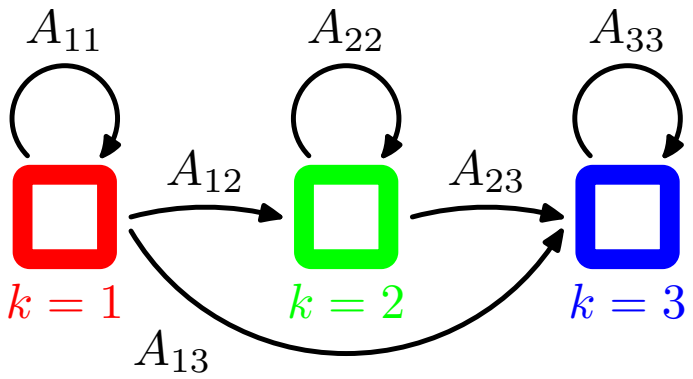
Sampling with 5% probability of change to each of the other states.

Statistical Machine Learning

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Hidden Markov Model - Variants

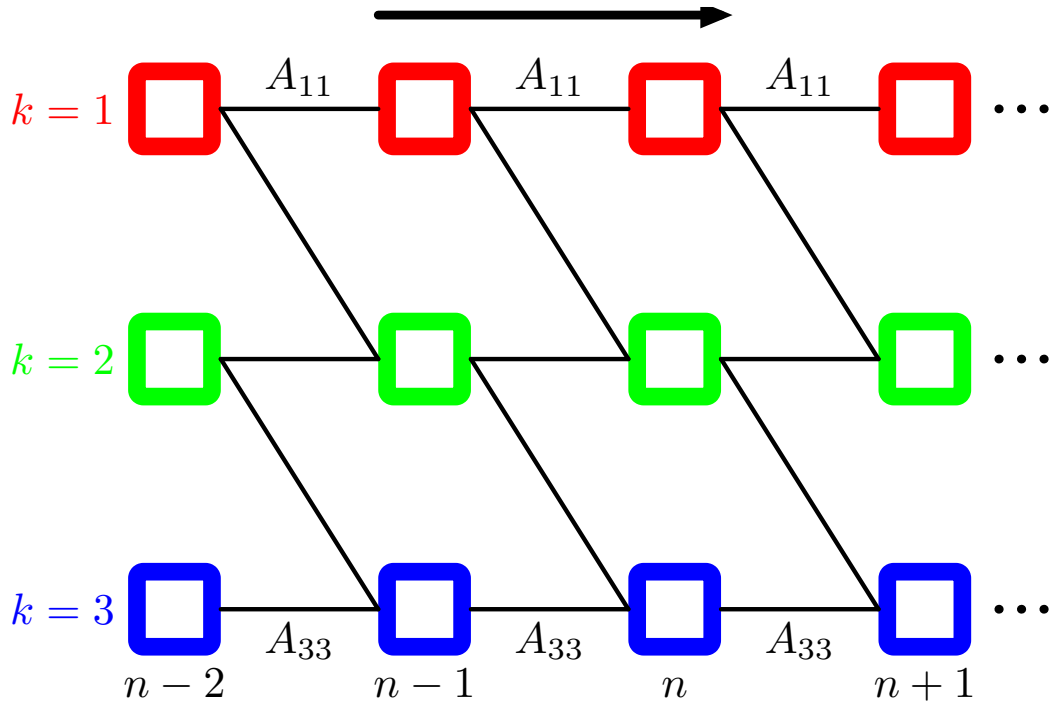
- Create variants of HMM by imposing constraints on the transition matrix $\mathbf{A}$.
- Left-to-right HMM : set all elements of $\mathbf{A}$ above the diagonal to zero, $A_{jk} = 0$ for $k < j$.
- Set the initial state probability $p(z_{11}) = 1$ and $p(z_{1j}) = 0$ for $j \neq 1$.
- Then, every sequence is constrained to start in state $k = 1$ and every state left can not be revisited again.



Data61 | CSIRO
The Australian National University

Hidden Markov Model - Variants

- Further restrict the transition matrix $\mathbf{A}$ to ensure that no large changes in the state space occur $A_{jk} = 0$ for $k > j + \Delta$ where Δ is the maximal change of state in one step.



Example of a HMM with $\Delta = 1$.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Hidden Markov Model - Handwritten Digits

- Digitize the trajectory online.
- $\mathbf{x}_i$ represent fixed width line segments at 16 different angles.
- Coincidentally also $K = 16$ states.
- $\Rightarrow$ 16 $\times$ 16 transition and emission parameter matrices.
- Trained with 45 examples of the digit '2'.
- Left-to-right transmission probability with $\Delta = 1$.
- Upper row: training samples lower row: model samples.