ECE521 Lecture 18

Graphical Models

Hidden Markov Models

UNIVERSITY OF TORONTO

Outline

- **Graphical models**
  - Conditional independence
  - Conditional independence after marginalization
- Sequence models
  - hidden Markov models

Graphical models

- There are two ways to represent a joint probability distributions over some random variables.
  - Express the joint distribution as a product of conditional distributions

e.g. $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$

  - Energy-based representation: product of potential functions (or factors)

e.g. $p(x_1, x_2, x_3) = \frac{1}{Z}\psi_1(x_1, x_2)\psi_2(x_1, x_3) \quad Z = \sum_{x_1, x_2, x_3} \psi_1(x_1, x_2)\psi_2(x_1, x_3)$

- The conditional independence between the random variables are explicitly captured by the **graphical models**.

Graphical models

$$p(x_1, x_2, x_3) = \frac{1}{Z}\psi_1(x_1, x_2)\psi_2(x_1, x_3)$$

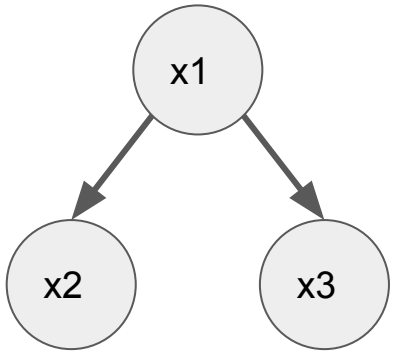
$$Z = \sum_{x_1, x_2, x_3} \psi_1(x_1, x_2)\psi_2(x_1, x_3)$$

- Advantages and disadvantages of the energy-based representation:
  - **Good**: it is very easy to define a proper probability distribution using almost any potential functions. We can turn any loss function into a probability distribution through the energy-based representation $p(\mathbf{x}) = \frac{1}{Z}e^{-\mathcal{L}(\mathbf{x})}$
  - **Bad**: evaluation of the PDF involves computing the normalization constant Z (or the partition function) that can be very expensive, ie. summation over all the random variable states. (If we are clever about the conditional independence, computing this summation can be efficient.)
  - **Bad**: learning using MLE over the energy-based representation requires taking the gradient of the normalization constant Z w.r.t. the model parameters. (Similarly, we can perform the exact gradient-based learning efficiently for some graphical models.)

Graphical models

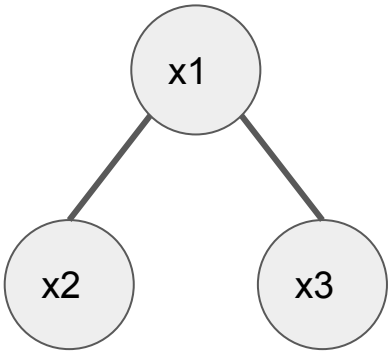
- Bayesian networks represent the product of conditionals. Markov random fields (and factor graph) represent the sum of energies.

Bayesian networks



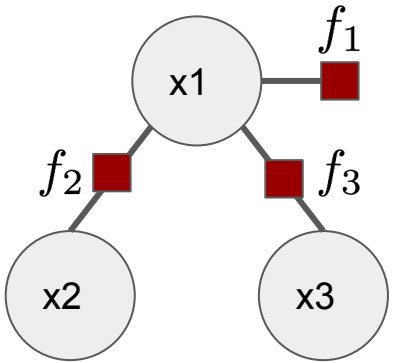
$$p(x_1)p(x_2|x_1)p(x_3|x_1)$$

Markov random fields



$$\frac{1}{Z}\psi_1(x_1,x_2)\psi_2(x_1,x_3)$$

Factor graph



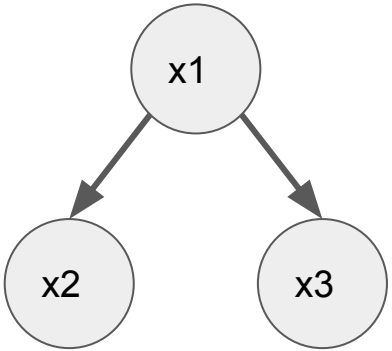
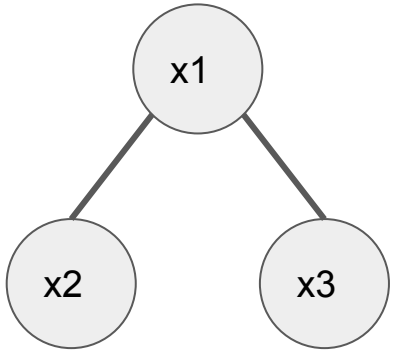
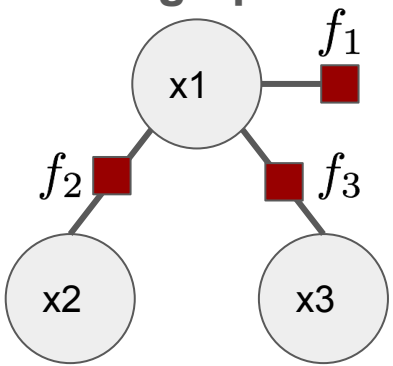
$$\frac{1}{Z}f_1(x_1)f_2(x_2,x_1)f_3(x_1,x_3)$$

Graphical models

- Conversion between the graphical model types:

naive Bayes model

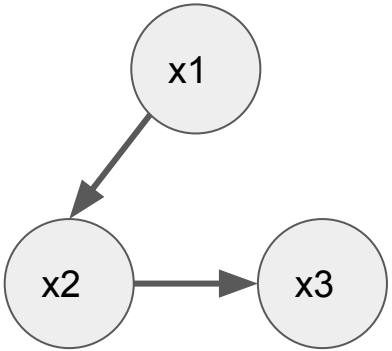
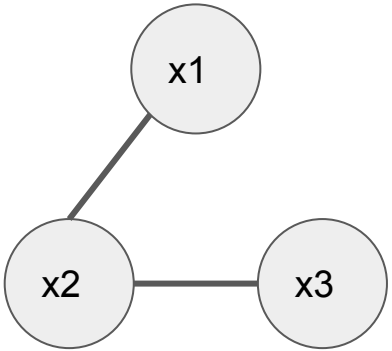
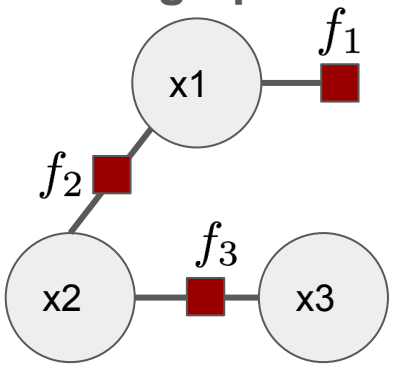
mixture of Gaussians with diagonal Gaussians

	Bayesian networks	Markov random fields	Factor graph
			
conditional independence:	$x_2 \perp\!\!\!\perp x_3 \mid x_1$	$x_2 \perp\!\!\!\perp x_3 \mid x_1$	$x_2 \perp\!\!\!\perp x_3 \mid x_1$
marginal independence:	$x_2 \not\perp\!\!\!\perp x_3$	$x_2 \not\perp\!\!\!\perp x_3$	$x_2 \not\perp\!\!\!\perp x_3$

Graphical models

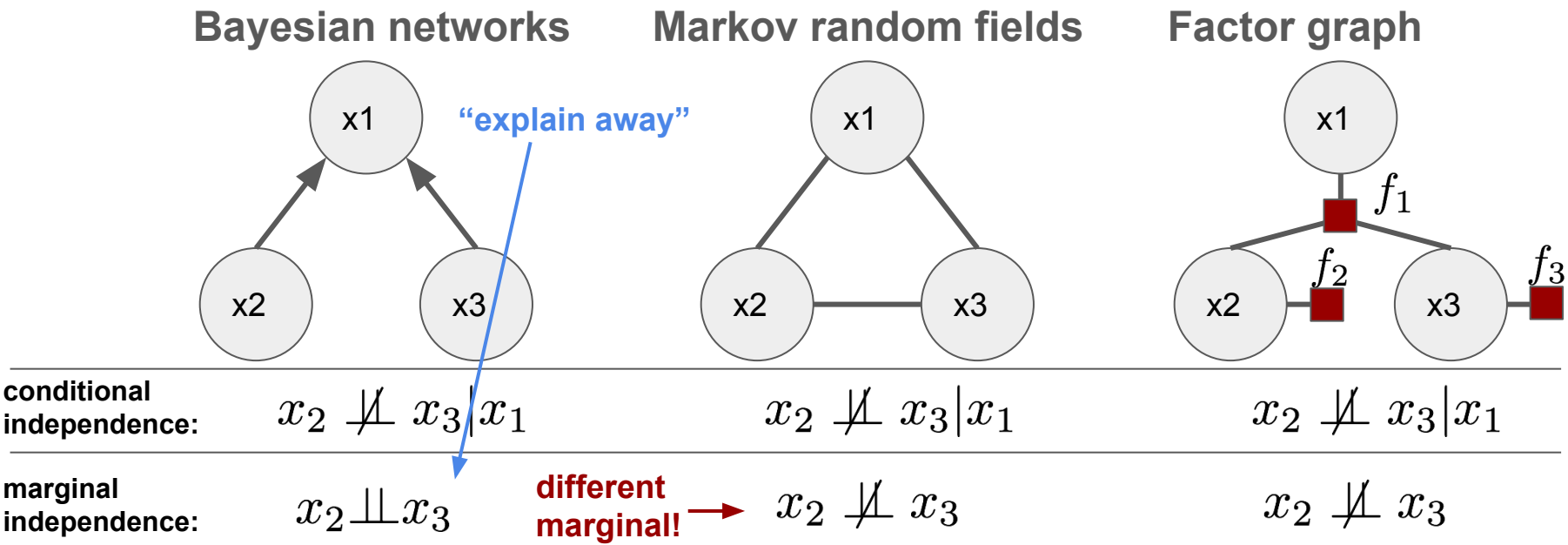
Markov chain models

- Conversion between the graphical model types:

	Bayesian networks	Markov random fields	Factor graph
			
conditional independence:	$x_1 \perp\!\!\!\perp x_3 \mid x_2$	$x_1 \perp\!\!\!\perp x_3 \mid x_2$	$x_1 \perp\!\!\!\perp x_3 \mid x_2$
marginal independence:	$x_1 \not\perp\!\!\!\perp x_3$	$x_1 \not\perp\!\!\!\perp x_3$	$x_1 \not\perp\!\!\!\perp x_3$

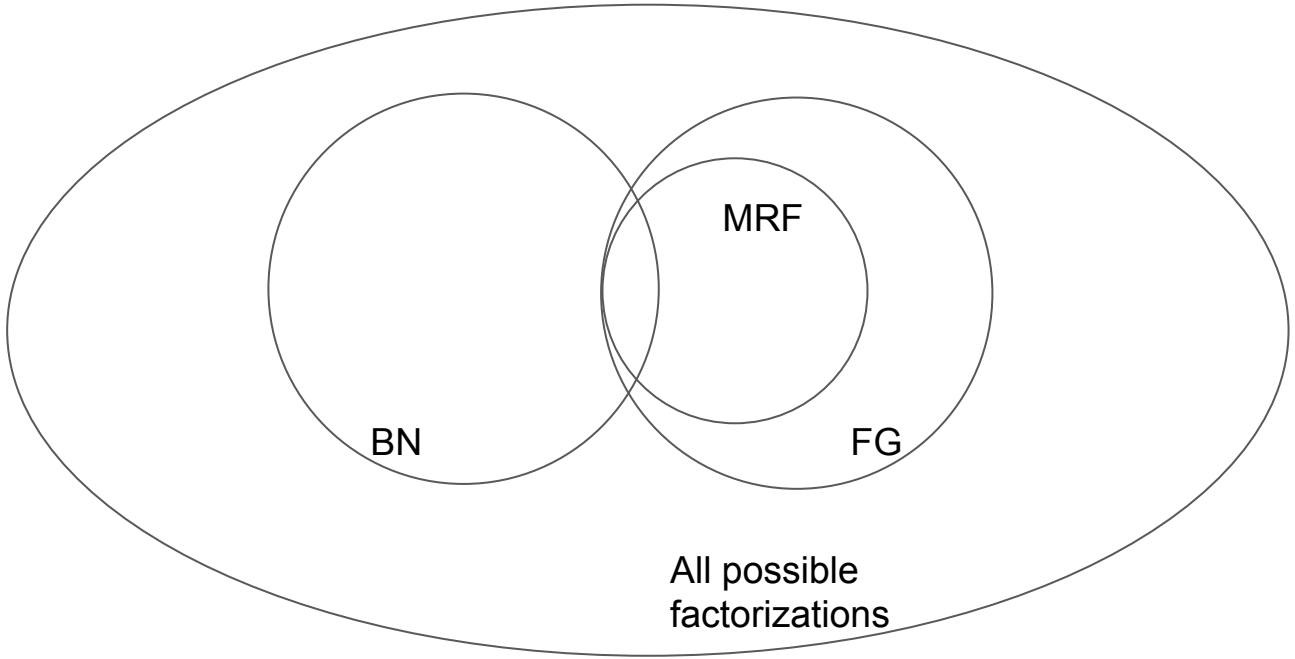
Graphical models

- Conversion between the graphical model types:



Graphical models

- Representing factorizations of the joint distributions using graphical models:



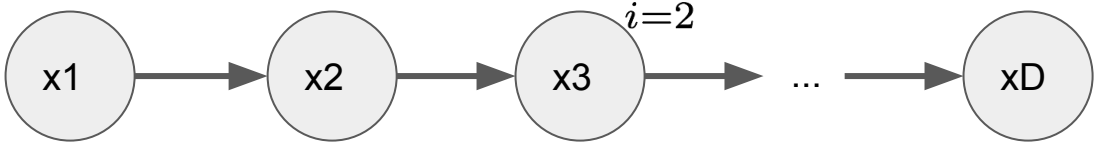
Outline

- Graphical models
  - Conditional independence
  - Conditional independence after marginalization
- **Sequence models**
  - hidden Markov models

Markov Models

- Remember the Markov chain model for a sequence (an ordered list) of observed random variables that only depends on the most immediate previous observation:
  - e.g. the stock price today are conditionally independent of the entire history given yesterday.

$$p(x_1, \ldots, x_D) = p(x_1) \prod_{i=2}^{D} p(x_d | x_{d-1})$$

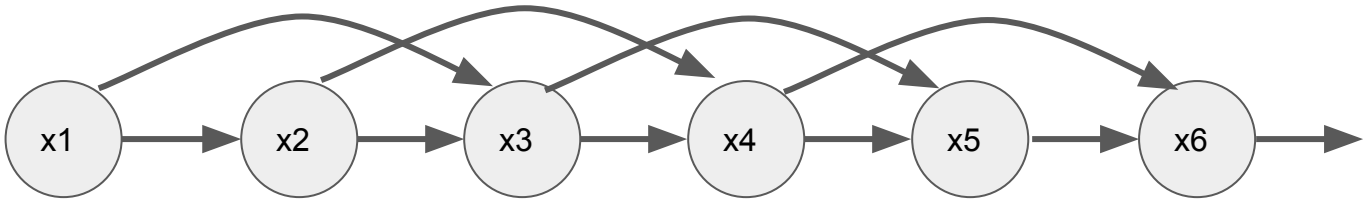


$$\forall n < m - 1 : \quad x_n \perp\!\!\!\perp x_m | x_{m-1}$$

- Although this is memory efficient and cheap to compute, the conditional independence assumptions on observations are too restrictive.

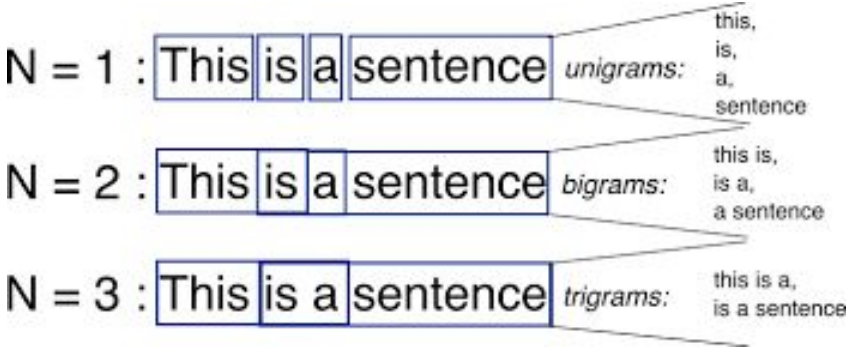
High order Markov models

- One solution to incorporate long term dependency is to construct a **Markov model of degree N** that depends on the N-1 most immediate previous observations.
  - Definition: $\forall n < m - N - 1, \quad x_n \perp\!\!\!\perp x_m | x_{m-1}, \ldots, x_{m-N-1}$
  - e.g. degree 3: $x_1 \perp\!\!\!\perp x_4 | x_2, x_3$
    $x_1 \not\perp\!\!\!\perp x_4 | x_3$



High order Markov models

- Sequence (language) modeling is the one of the fundamental tasks in natural language processing (NLP). The chain structure graphical models, i.e. Markov models, ideally captures the causal process in generating words. Namely, we generate(write) the current word given the previous words.
- One of the simplest language model is the N-grams model, which are the N degree Markov models. E.g:



$$p(w_1, \ldots, w_4) = p(w_1)p(w_2)p(w_3)p(w_4)$$

$$p(w_1, \ldots, w_4) = p(w_1)p(w_2|w_1)p(w_3|w_2)p(w_4|w_3)$$

$$p(w_1, \ldots, w_4) = p(w_1, w_2)p(w_3|w_2, w_1)p(w_4|w_3, w_2)$$

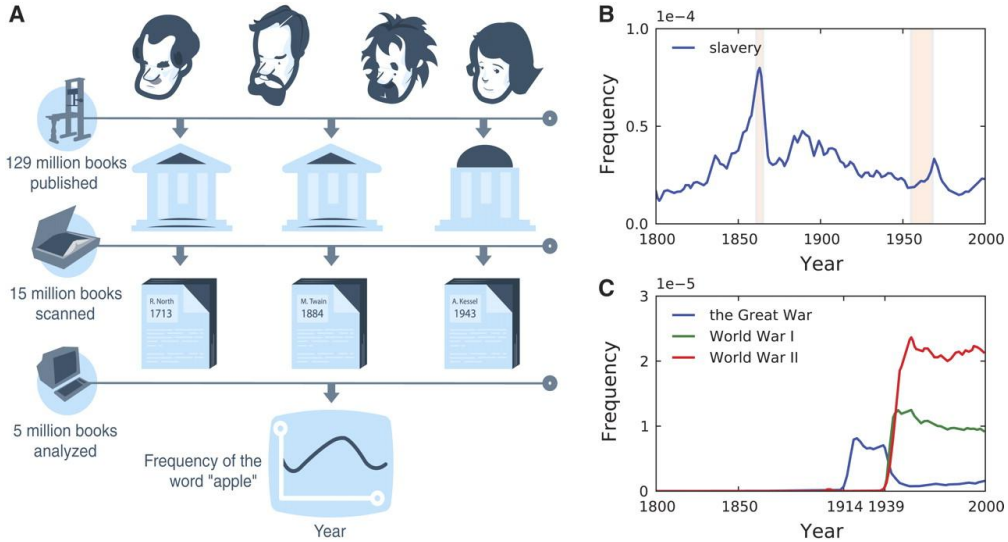
High order Markov models

- N-grams model defines a frequency count table for all of the N-tuple words appeared in a dataset. (e.g. wikipedia or millions of printed books)

previous words		**current word**	**Frequency**
w_k-2	w_k-1	w_k	P(w_k | w_k-1, w_k-2)
this	is	a	0.0002
is	a	sentence	1e-8
...	...	...	...

High order Markov models

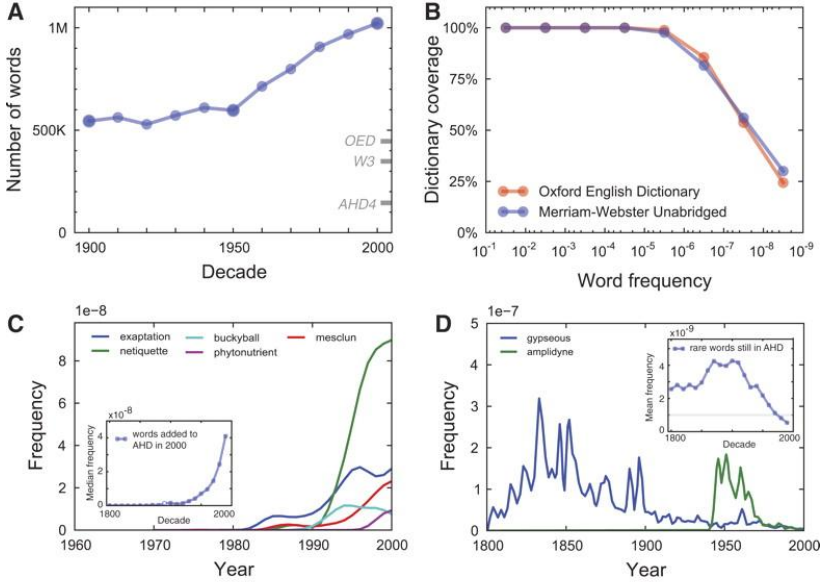
- Google N-grams viewer, one of the largest N-grams models in existence.
  - Application: study language and culture evolution over time

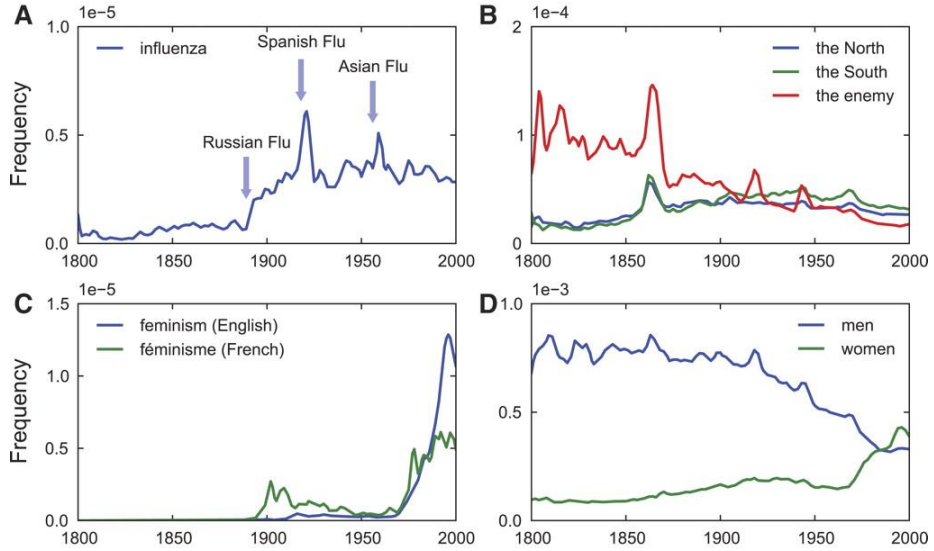


Source: Quantitative Analysis of Culture Using Millions of Digitized Books, 2011

High order Markov models

- Google N-grams viewer, one of the largest N-grams models in existence.
  - Application: study language and culture evolution over time





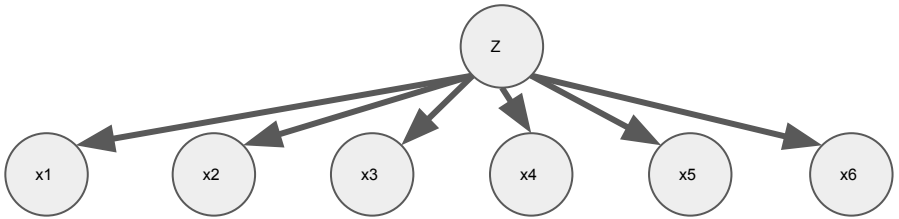
Source: Quantitative Analysis of Culture Using Millions of Digitized Books, 2011

High order Markov models

- High order Markov models can model any “close” range dependencies in a sequence of random variables very well.
- **Problem**: computational costs grow exponentially with the degree N. It becomes less practical to model any long range dependencies with high order Markov models.
- So far we have tried to directly model dependencies in an observation sequence. One idea to model more interesting and complicated dependencies is to incorporate latent variables into the sequence modelling.

Hidden Markov models

- Simplest latent variable model that captures long range dependencies:



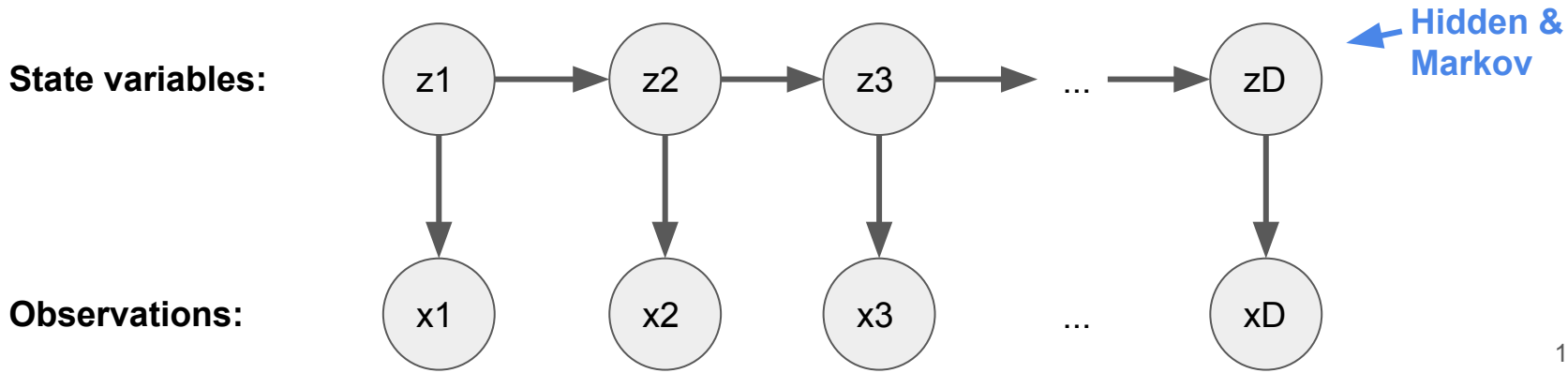
naive Bayes

- **Problem**: the conditional independence assumptions on all observations are too restrictive. Naive Bayes is non-Markovian (i.e. the value of x3 also depends on x4, x5,...) and can not model the sequential nature of the input data. (future depends on the past)

- **Solution:** construct a hidden Markov chain over the latent variables.

Hidden Markov models

- Formally, a hidden Markov model is defined on a sequence of observed random variables $\{x_1, \ldots, x_D\}$ through its corresponding latent sequence $\{z_1, \ldots, z_D\}$.



Hidden Markov models

- Conditional independence:
  - The hidden states are connected through a degree 2 Markov chain $\forall n < m-1: \quad z_n \perp\!\!\!\perp z_m | z_{m-1}$
  - The future observations and the past observations are conditionally independent given the current hidden state $\{x_1, \ldots, x_{m-1}\} \perp\!\!\!\perp \{x_{m+1}, \ldots, x_D\} | z_m$
  - The present is **dependent** on the entire past observations. $x_{m-2} \not\perp\!\!\!\perp x_m | x_{m-1}$