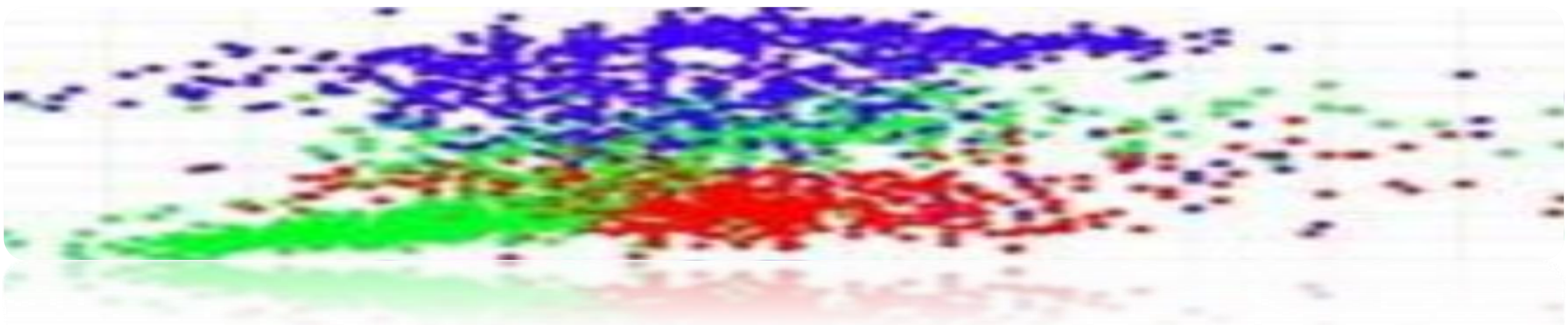


GRAPHICAL MODELS

Outline

- Bayesian networks
  - Generative models
  - Discriminative models
  - Markov chains
  - D-separation
  - Hidden Markov Models

- Undirected models
  - Factorization
  - Hammersley-Clifford Theorem
  - Moralization

Intro

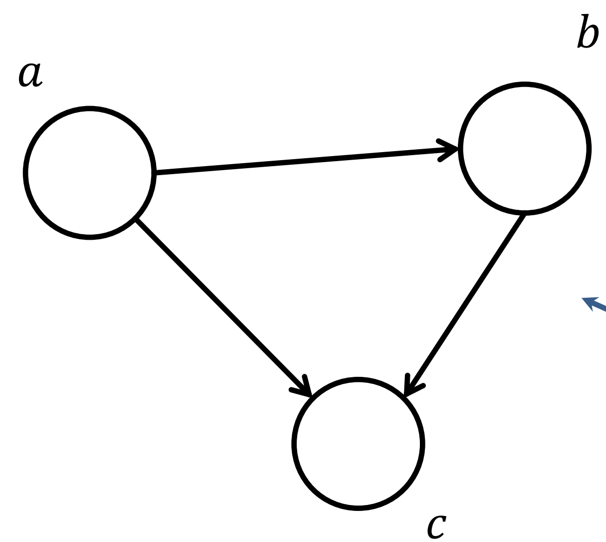
- What is a graphical model?
  - Graph $G(V, E)$
  - V: Set of random variables
  - $E(\subseteq V \times V)$: Set of dependence relationships

- Why graphical models?
  - Visualize the structure of probabilistic models
  - Obtain insights into properties of variables and their interdependencies (e.g. **conditional independence**, **causality**) through **graph-theoretic** means
  - Perform **probabilistic inference** through graphical manipulations

- Types of graphical models
  - **Directed** graphical models $\equiv$ **Bayesian networks** (coined by Judea Pearl 1985)
  - **Undirected** graphical models $\equiv$ **Markov random fields**

Bayesian networks (1)

- Represent **conditional dependencies** between random variables



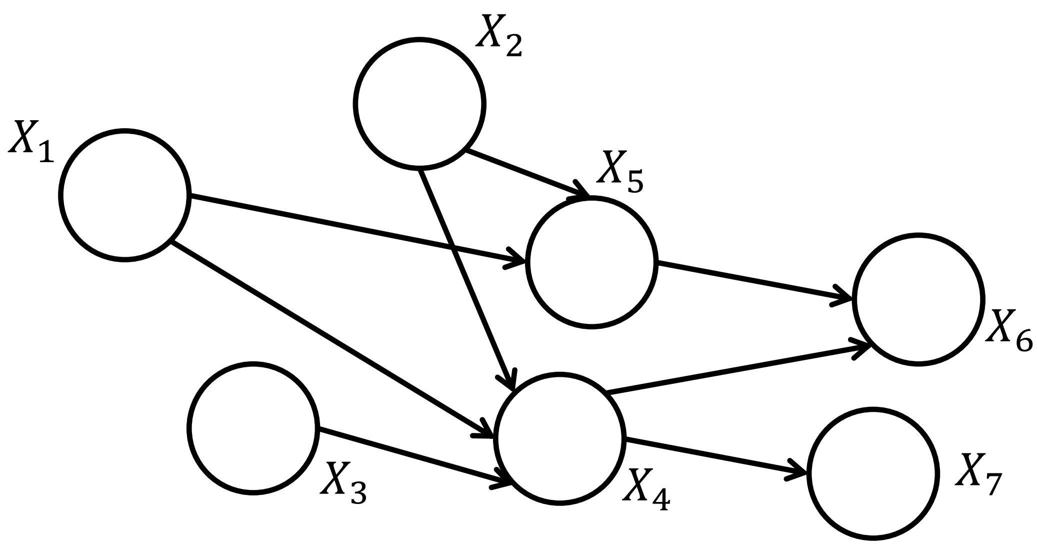
Example from
C. Bishop: PRML book

$$P(a,b,c) = P(c|a,b)P(a,b) = P(c|a,b)P(b|a)P(a)$$

$$P(X_1, \dots, X_k) = P(X_k|X_1, \dots, X_{k-1}) \dots P(X_2|X_1)P(X_1)$$

Bayesian networks (2)

- Represent **causal dependencies** between random variables
- **Local Markov Property**: Each variable is conditionally independent of its non-descendants given its parents

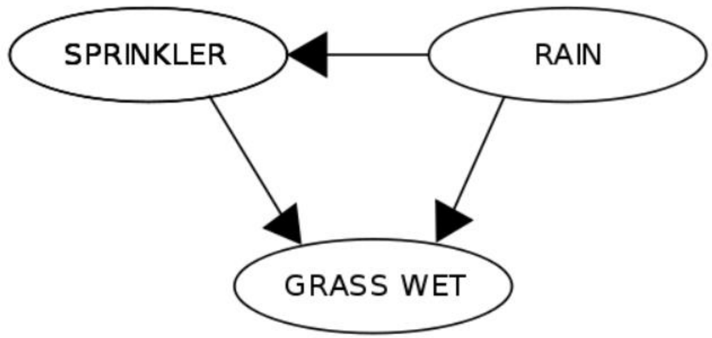


$$P(X_1, \dots, X_7) = P(X_1)P(X_2)P(X_3)P(X_4|X_1, X_2, X_3)P(X_5|X_1, X_2)P(X_6|X_4, X_5)P(X_7|X_4)$$

➔ Generally: $P(\mathbf{X}) = \prod_{i=1}^{k} P(X_i|par_i)$, where par_i denote the parents of X_i

Bayesian network example

RAIN	SPRINKLER: T	SPRINKLER: F
F	0.4	0.6
T	0.01	0.99



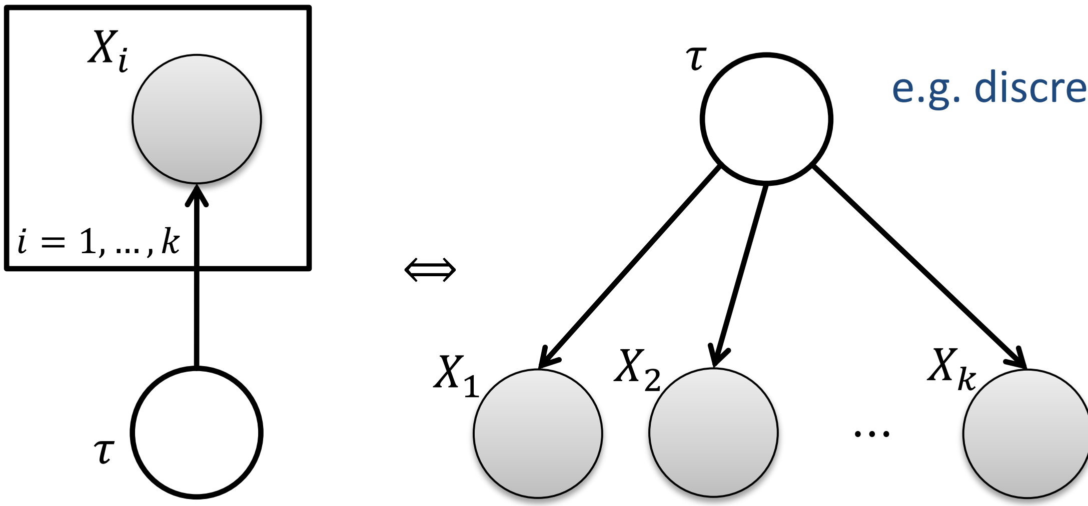
RAIN: T	RAIN: F
0.2	0.8

SPRINKLER	RAIN	GRASS WET: T	GRASS WET: F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

$$P(S,R,G) = P(R)P(S|R)P(G|S,R)$$

- What is the probability that the sprinkler was on given that the grass is wet?

Generative models



e.g. discrete topic variable

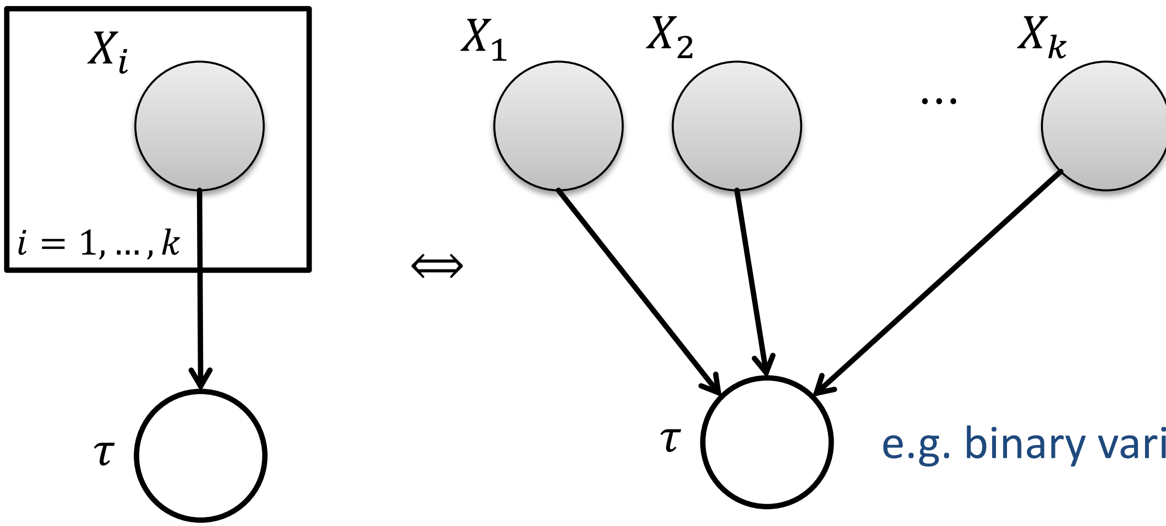
e.g. observed i.i.d. feature variables (binary)

$$P(X_1, \dots, X_k, \tau) = P(\tau) \prod_{i=1}^{k} P(X_i | \tau)$$

… we need to estimate $k + 1$ parameters

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Discriminative models



e.g. discrete feature variables

e.g. binary variable

$$P(\tau = 1 | X_1, \ldots, X_k)$$

… we need to estimate l^k parameters, if each X_i has l states

The trick is to parameterize each X_i with a weight β_i and estimate

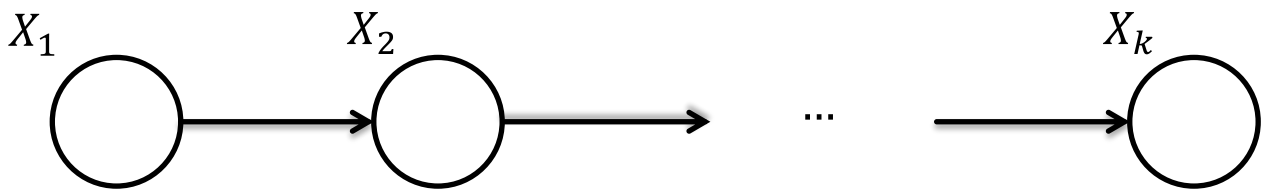
$$\frac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \cdots - \beta_k X_k)}$$

… we need to estimate $k + 1$ parameters

Markov chains

- Generally, for the joint distribution of k discrete variables $X_1, \dots, X_k$, with l states each, we need to estimate $l^k - 1$ parameters

- For k-node Markov chain: $l - 1 + (k-1)l(l-1)$ parameters



- A Markov chain is a Bayesian network in which each variable has at most one predecessor

Conditional independence

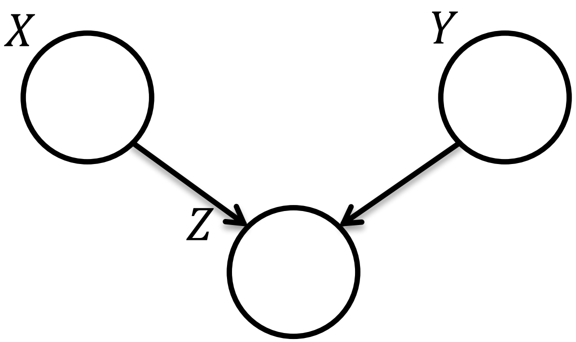
X is independent of Y given Z (we write: $X \perp Y|Z$)

$$\Leftrightarrow$$

$$P(X|Y,Z) = P(X|Z)$$

$$\Leftrightarrow$$

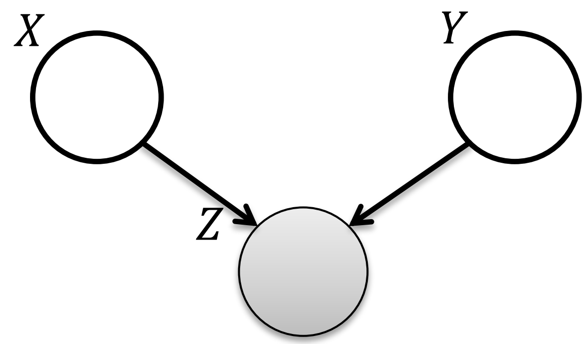
$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$



$$P(X,Y,Z) = P(X)P(Y)P(Z|X,Y)$$

$$\Leftrightarrow$$

$$P(X,Y) = P(X)P(Y) \Leftrightarrow X \perp Y|\emptyset$$



$$P(X,Y|Z) = P(X,Y,Z)/P(Z)$$

$$\Leftrightarrow$$

$$P(X,Y|Z) = P(X)P(Y)P(Z|X,Y)/P(Z)$$

$$\Longrightarrow X \not\perp Y|Z$$

Conditional independence example

- Example from C. Bishop, PRML book

  B: battery (0=flat, 1=full)

  F: fuel (0=empty, 1=full)

  R: Fuel reading (0=empty, 1=full)

  $P(B=1)=0.9$

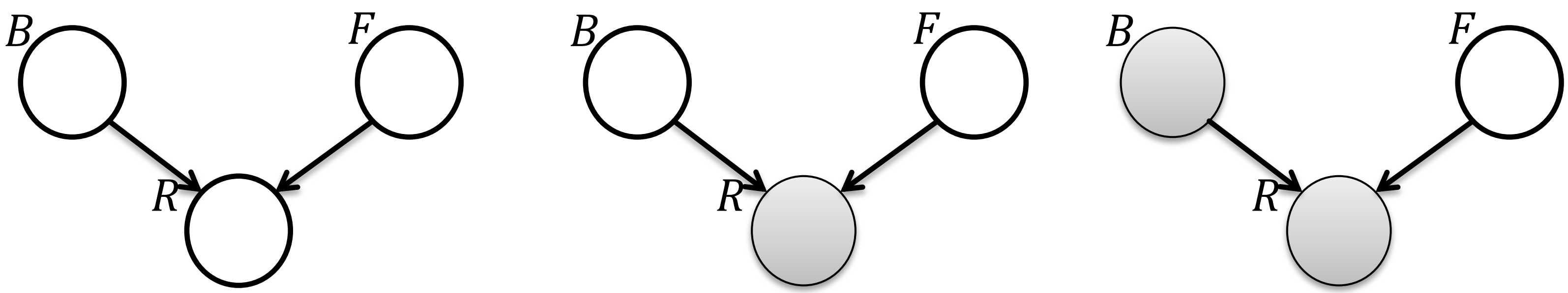
  $P(F=1)=0.9$

  $P(R=1|B=1,F=1)=0.8$

  $P(R=1|B=1,F=0)=0.2$

  $P(R=1|B=0,F=1)=0.2$

  $P(R=1|B=0,F=0)=0.1$



We can calculate:

$P(R=0)=\sum_{B\in\{0,1\}}\sum_{F\in\{0,1\}}P(R=0|B,F)P(B)P(F)=0.315$

$P(R=0|F=0)=\sum_{B\in\{0,1\}}P(R=0|B,F=0)P(B)=0.81$

$P(F=0|R=0)=\frac{P(R=0|F=0)P(F=0)}{P(R=0)}\approx 0.257$ (probability of $F=0$ increases)

$P(F=0|R=0,B=0)\approx 0.111$ (prob. of $F=0$ decreases, is "explained away")

D-separation

- Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ be disjoint subsets of nodes from a Bayesian network
- A path from $\boldsymbol{A}$ to $\boldsymbol{B}$ is **blocked** if it contains node v such that either

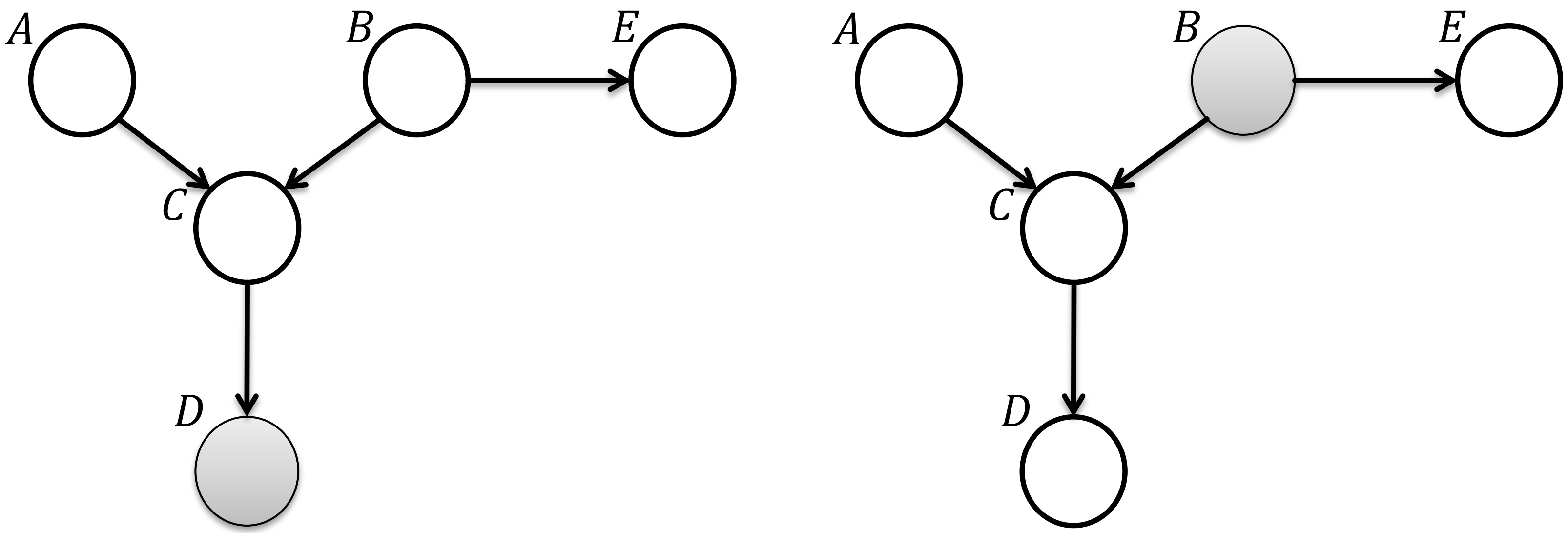
  a) arrows on the path meet head-to-tail ($\rightarrow v \rightarrow$) or tail-to tail ($\leftarrow v \rightarrow$) at v and v is from $\boldsymbol{C}$, or

  b) arrows on the path meet head-to-head at v ($\rightarrow v \leftarrow$) and neither v nor any of its descendants are in $\boldsymbol{C}$

- $\boldsymbol{A}$ is **d-separated** from $\boldsymbol{B}$ by $\boldsymbol{C}$ if all paths from $\boldsymbol{A}$ to $\boldsymbol{B}$ are **blocked**

- **Theorem:** If $\boldsymbol{A}$ is **d-separated** from $\boldsymbol{B}$ by $\boldsymbol{C}$ then $\boldsymbol{A} \perp \boldsymbol{B} | \boldsymbol{C}$

D-separation examples

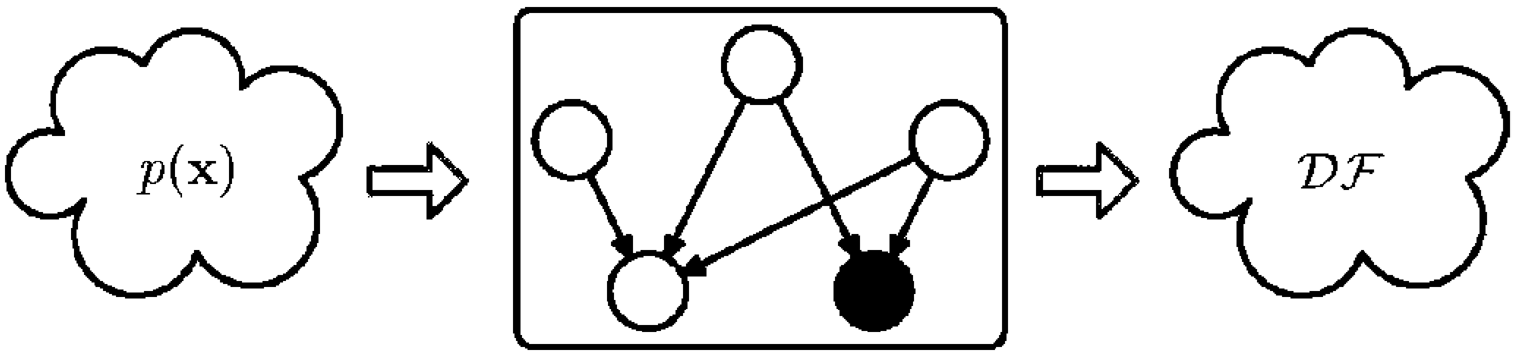


$A \perp E | D$ ✗

$A \perp E | C$ ✗

$A \perp E | B$ ✓

Bayesian networks as distribution filters

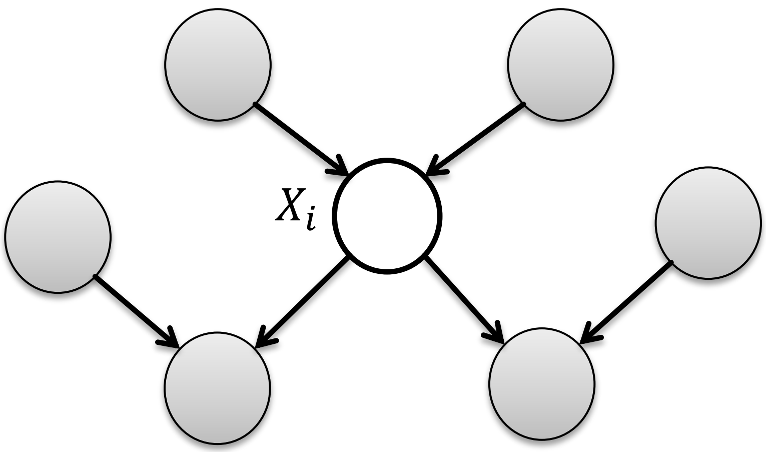


From C. Bishop: PRML book

- Any joint probability distribution $P(\mathbf{x})$ that factorizes according to the graphical model passes through the filter

- The set of distributions $P(\mathbf{x})$ that pass through the filter is denoted by $\mathcal{DF}$; this is exactly the set of distributions that respect all conditional independencies implied by the d-separation

Markov blanket



$$P\left(X_i \middle| \{X_{j\neq i}\}\right) = \frac{P(X_1, \dots, X_k)}{\int P(X_1, \dots, X_k)\, dX_i}$$

$$= \frac{\prod_j P(X_j | par_j)}{\int \prod_j P(X_j | par_j)\, dX_i}$$

Factors independent of X_i cancel out!

- In a Bayesian network, a variable is independent of all other variables given its Markov blanket

Hidden Markov models (HMMs)

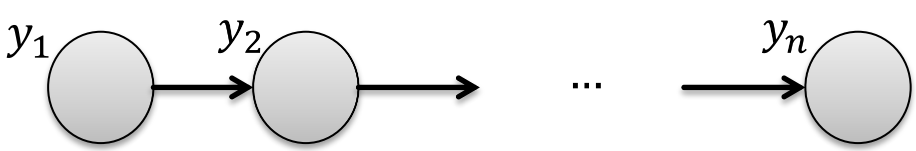
- Useful for explaining sequential data, e.g., arising from
  - Measurements of time series
  - Observations of sequential nucleotide base pairs along a DNA strand
  - Sequential tokens in a sentence/speech

- Simplest way to model the probability of an observed sequence $y_1, \dots, y_n$ is to assume pairwise independence between the observations:

$$P(\mathbf{y}) = \prod_{i=1}^{n} P(y_i)$$

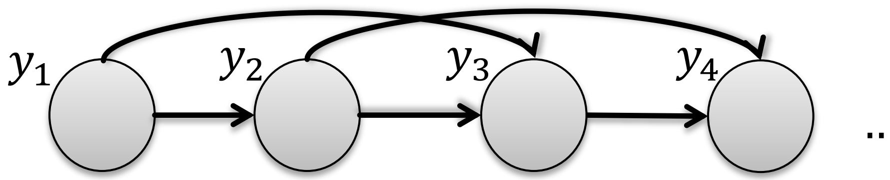
- Above assumption may be too strong; if it is relaxed such that every observation is dependent only on the (most) recent observations we obtain a Markov model

From Markov models to HMMs



$$P(\mathbf{y}) = P(y_1) \prod_{i=2}^{n} P(y_i | y_{i-1})$$

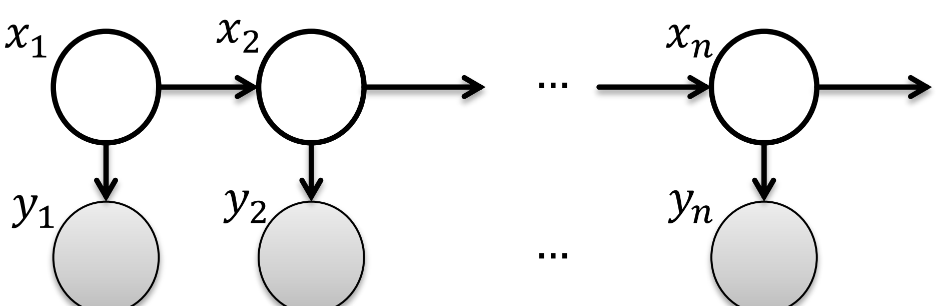
First-order Markov chain



$$P(\mathbf{y}) = P(y_1) P(y_2 | y_1) \prod_{i=3}^{n} P(y_i | y_{i-1}, y_{i-2})$$

Second-order Markov chain

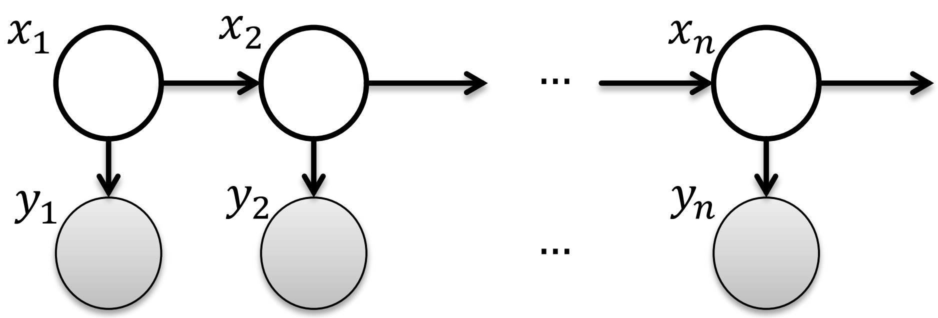
- More powerful model assumes latent variables (unknown states) for each y_i:



$$P(\mathbf{y}, \mathbf{x}) = P(x_1) \prod_{i=2}^{n} P(x_i | x_{i-1}) \prod_{i=1}^{n} P(y_i | x_i)$$

Extending Markov chains to HMMs

- If the latent variables are discrete the above model is an HMM



$x_{i+1} \perp x_{i-1} | x_i$

$$P(\mathbf{y}, \mathbf{x}) = P(x_1) \prod_{i=2}^{n} P(x_i | x_{i-1}) \prod_{i=1}^{n} P(y_i | x_i)$$

Emission probabilities

- Inference

  1) Find the most likely state at a given x_i: $\max_s P(x_i = s | y_1, \ldots, y_i)$

     (can be solved with the **forward-backward** algorithm)

  2) Find the most likely sequence of states: $\max_{\mathbf{s}} P(\mathbf{x} = \mathbf{s}, y_1, \ldots, y_i)$

     (can be solved with the **Viterbi** algorithm)

The Viterbi (dynamic programming) algorithm

- Input: observation space $O = \{o_1, \dots, o_N\}$, observations $Y = \{y_1, \dots, y_n\}$, state space $S = \{s_1, \dots, s_k\}$, $k \times k$-dim. state transition matrix $\boldsymbol{A}$, $k \times N$-dim. emission matrix $\boldsymbol{B}$, two $k \times n$-dim. matrices T_1, T_2, array $\boldsymbol{\pi}$ of size k (with prior probabilities for each state)

For each s_i do

$$T_1[i,1] = \ln(\pi_i * B_{i,y_1})\,;\; T_2[i,1] = 0;$$

For $i = 2 \dots n$ do

For each s_j do

$$T_1[j,i] = \max_l T_1[l, i-1] + \ln(A_{l,j} * B_{j,y_i})$$

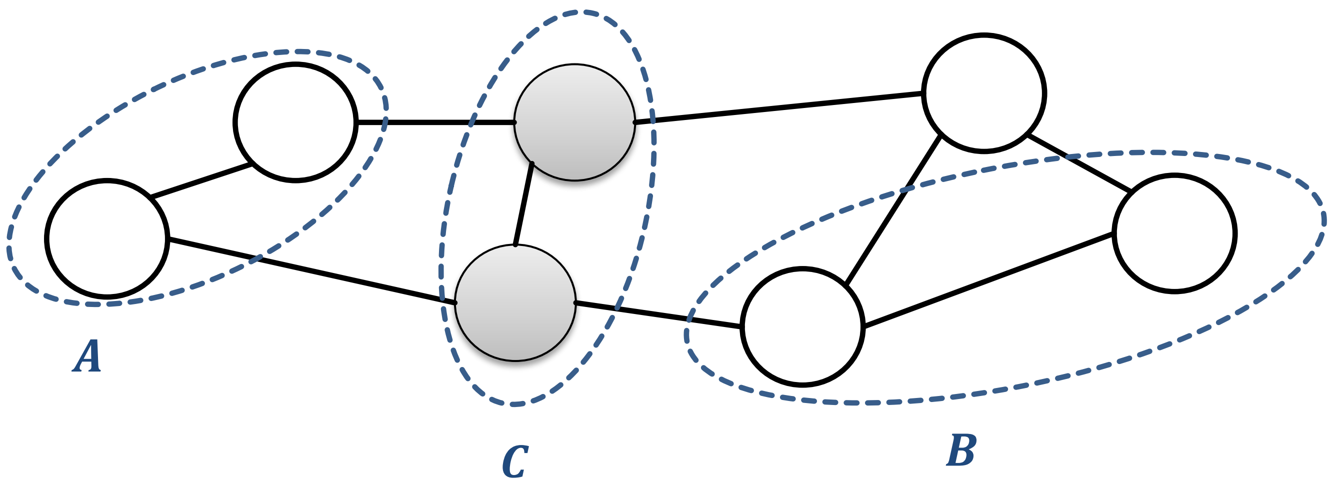
$$T_2[j,i] = \arg\max_l T_1[l, i-1] + \ln(A_{l,j} * B_{j,y_i})$$

Assign the right state label to each observation by using T_2

Complexity: $O(nk^2)$

Markov random fields

- Graph separation



The set of nodes $\boldsymbol{A}$ is independent of $\boldsymbol{B}$ given $\boldsymbol{C}$ if and only if all paths from $\boldsymbol{A}$ to $\boldsymbol{B}$ lead though $\boldsymbol{C}$; we write: $\boldsymbol{A} \perp \boldsymbol{B} | \boldsymbol{C}$

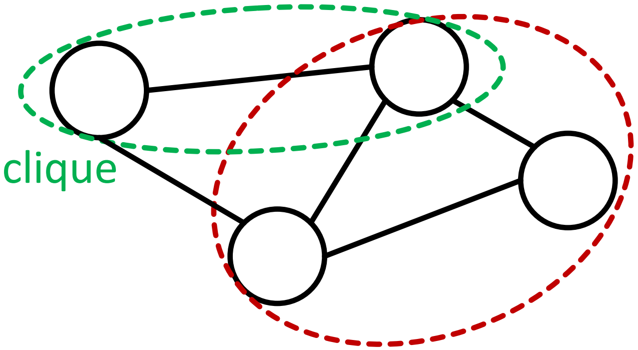
Markov blanket of variable node X is given by all direct neighbors $Nb(X)$ of X

$$P(X_i | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) = P(X_i | Nb(X_i))$$

➔ The joint distribution can be factorized according to above separation rule

Factorization through maximal cliques

- **Observation**: If two nodes X_i, X_j are not directly connected, they are independent given all other nodes ➔ they should **not** occur in same factor



Partition function

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{C \in MaxCliques} \psi_C(\boldsymbol{X}_C) \qquad Z = \sum_{\boldsymbol{X}} \prod_{C \in MC} \psi_C(\boldsymbol{X}_C)$$

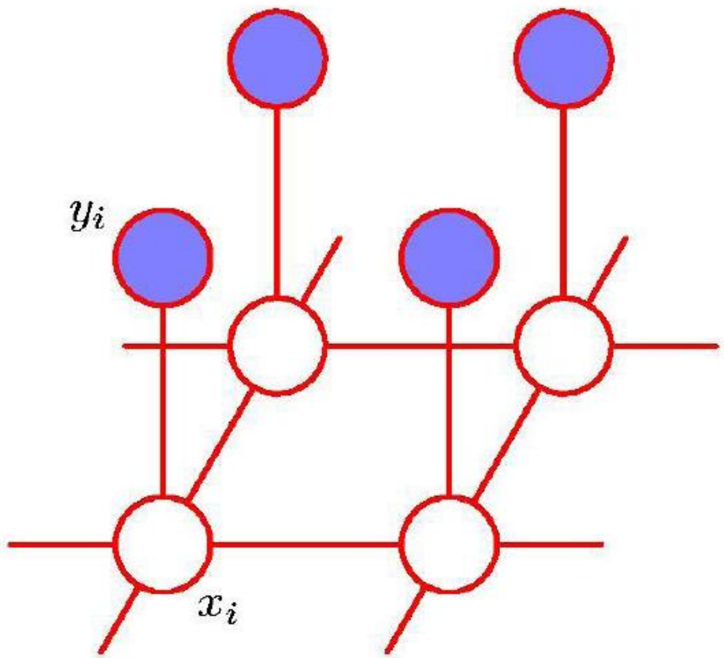
Energy function

$\psi_C(\boldsymbol{X}_C)$: **Potential** over clique C, typically: $\psi_C(\boldsymbol{X}_C) = \underbrace{\exp\left(-E(\boldsymbol{X}_C)\right)}_{\text{Boltzmann distribution}}$

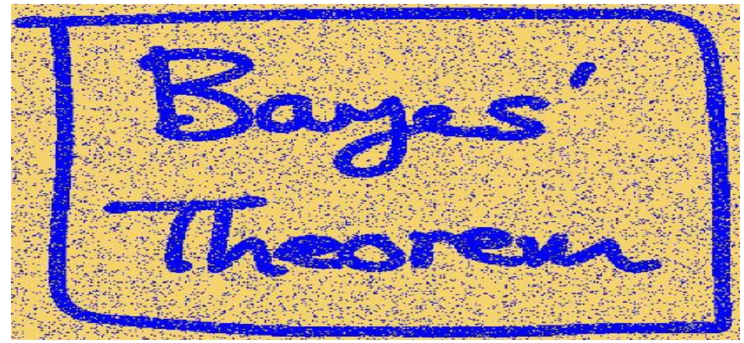
Boltzmann distribution

- **Hammersley-Clifford Theorem:** Factorization through graph separation and factorization through maximal cliques lead to the same sets of distributions

Example: Image de-noising

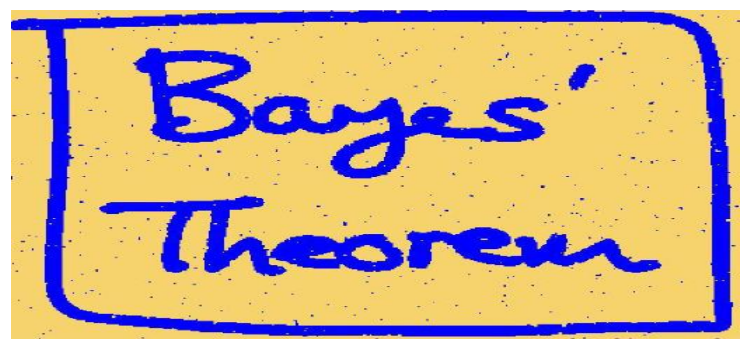


From C.Bishop: PRML



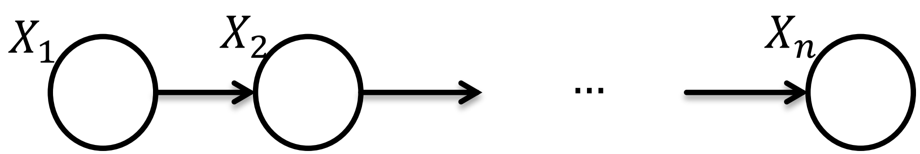
$$E(\mathbf{x},\mathbf{y}) = h\sum_i x_i - \beta\sum_{\{i,j\}} x_i x_j - \eta\sum_i x_i y_i$$

$$P(\mathbf{x},\mathbf{y}) = \frac{1}{Z}\exp\{-E(\mathbf{x},\mathbf{y})\}$$

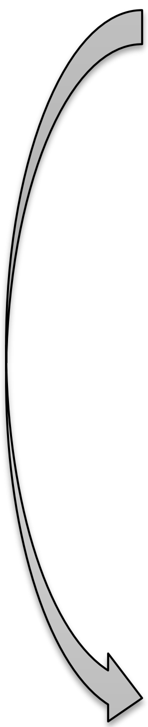


From directed to undirected chains

- Converting directed into undirected graphical models: chains



$$P(\mathbf{X}) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_n|X_{n-1})$$



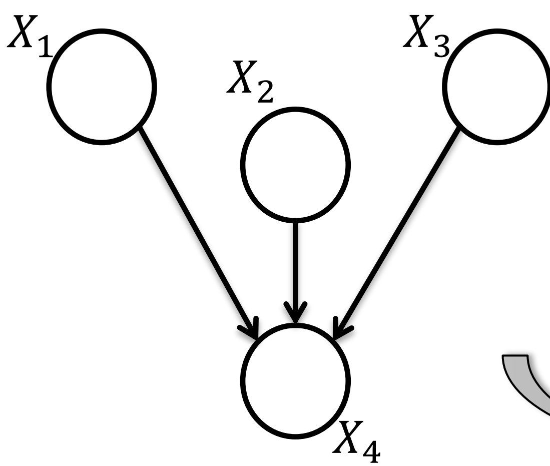
$$P(\mathbf{X}) = \frac{1}{Z}\psi_{1,2}(X_1, X_2)\psi_{2,3}(X_2, X_3)\ldots\psi_{n-1,n}(X_{n-1}, X_n)$$

What is the value of the partition function Z?

Moralization

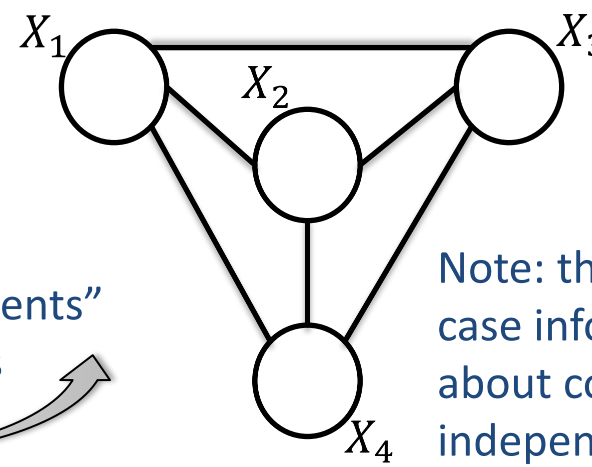
➢ Converting directed into undirected graphical models: general graphs

Take care that each conditional probability factor in the directed graph
is represented by at least one of the maximal cliques in the undirected graph!



“Marrying the parents”
& omitting arrows

moralization

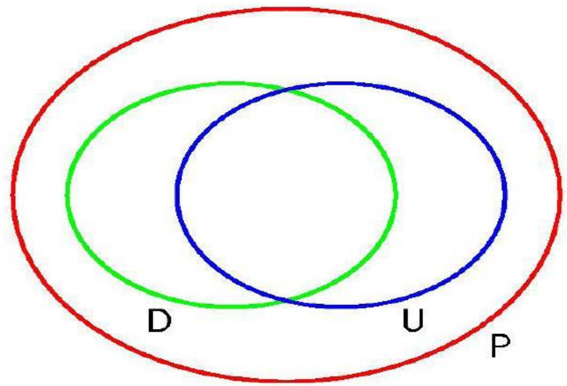


Note: that in this case information about conditional independence is lost!

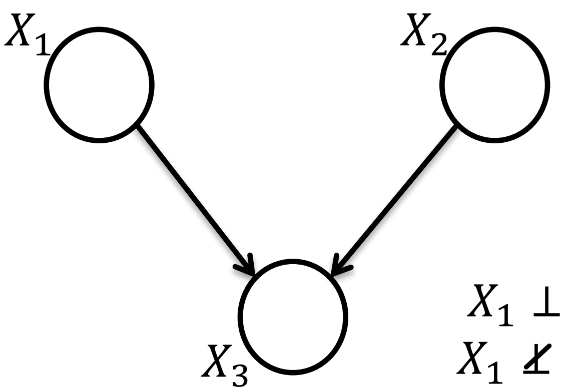
$$P(\mathbf{X}) = P(X_1)P(X_2)P(X_3)P(X_4|X_1,X_2,X_3) = \frac{1}{Z}\psi_{1,2,3,4}(X_1,X_2,X_3,X_4)$$

Theorem: No other technique of turning a directed into an undirected graph retains more independence information than moralization

Directed vs. undirected graphical models

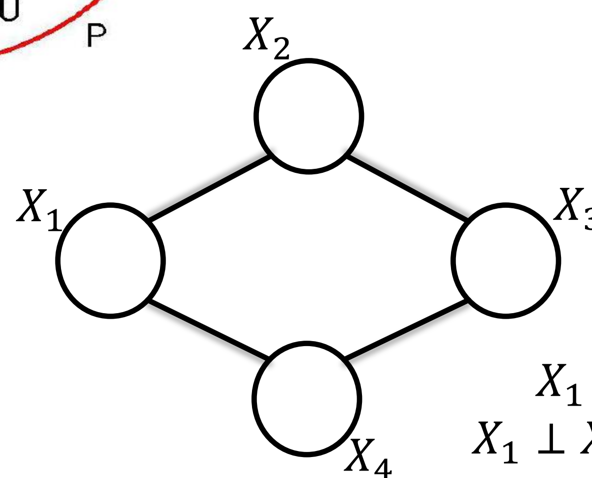


From C.Bishop: PRML



$$X_1 \perp X_2 | \emptyset$$
$$X_1 \not\perp X_2 | X_3$$

Independence properties cannot be represented by an undirected graph



$$X_1 \not\perp X_2 | \emptyset$$
$$X_1 \perp X_3 | \{X_2, X_3\}$$
$$X_2 \perp X_4 | \{X_1, X_2\}$$

Independence properties cannot be represented by a directed graph

Summary

- Bayesian networks (directed graphical models)
  - Local Markov property
  - D-separation
  - Distribution filters
  - Example: HMMs

- Markov random fields (undirected graphical models)
  - Graph separation
  - Hammersley-Clifford Theorem
  - Moralization