Machine Learning for Data Science (CS4786)
Lecture 18

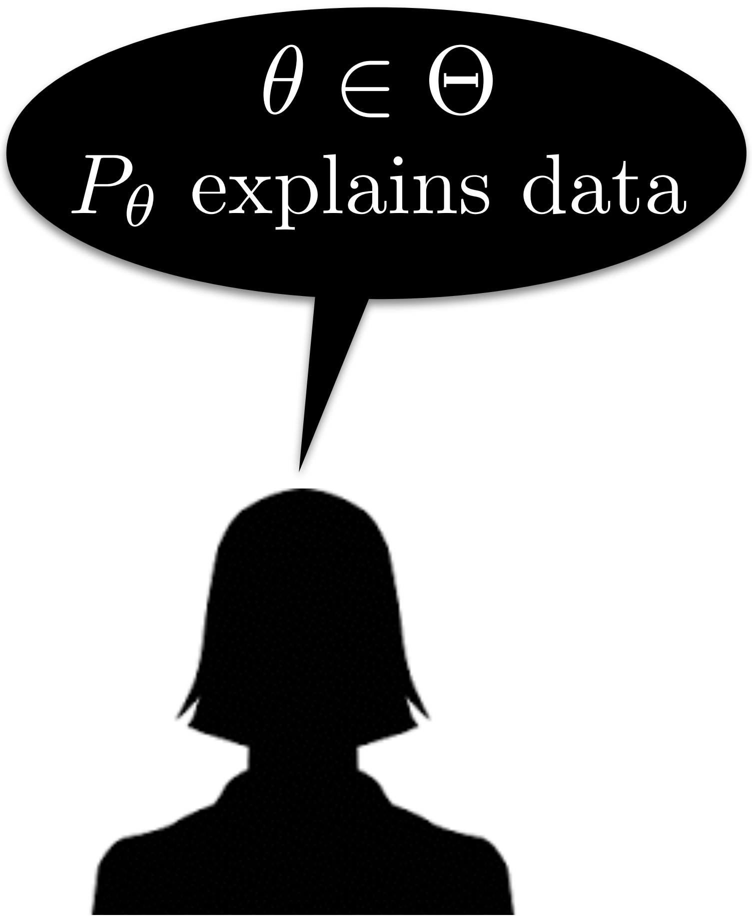
Graphical Models

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2017fa/

Probabilistic Model

Data

Probabilistic Model

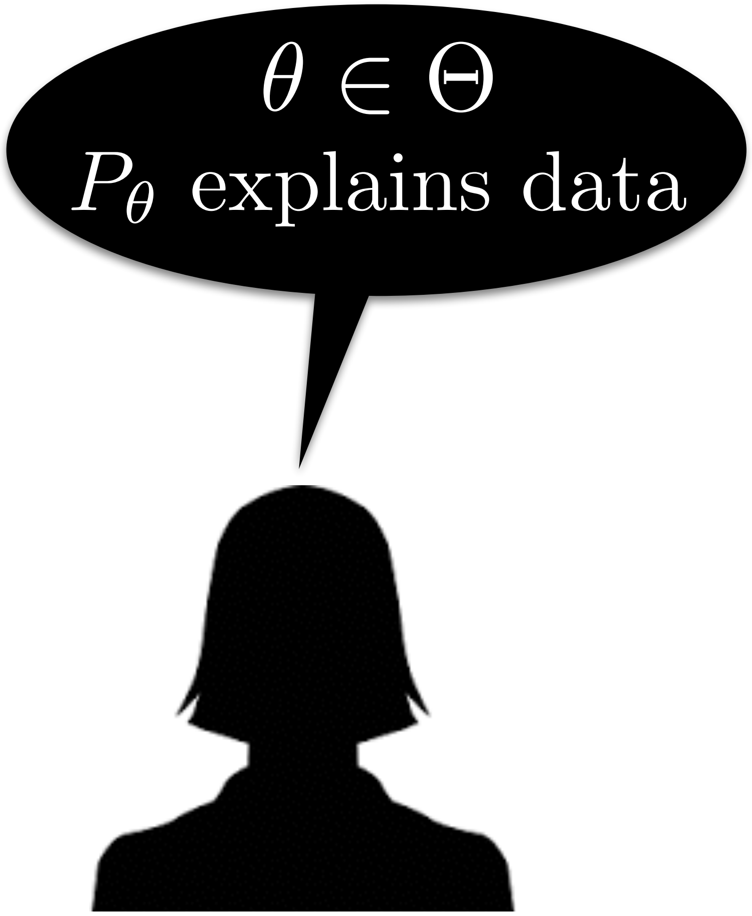


Data

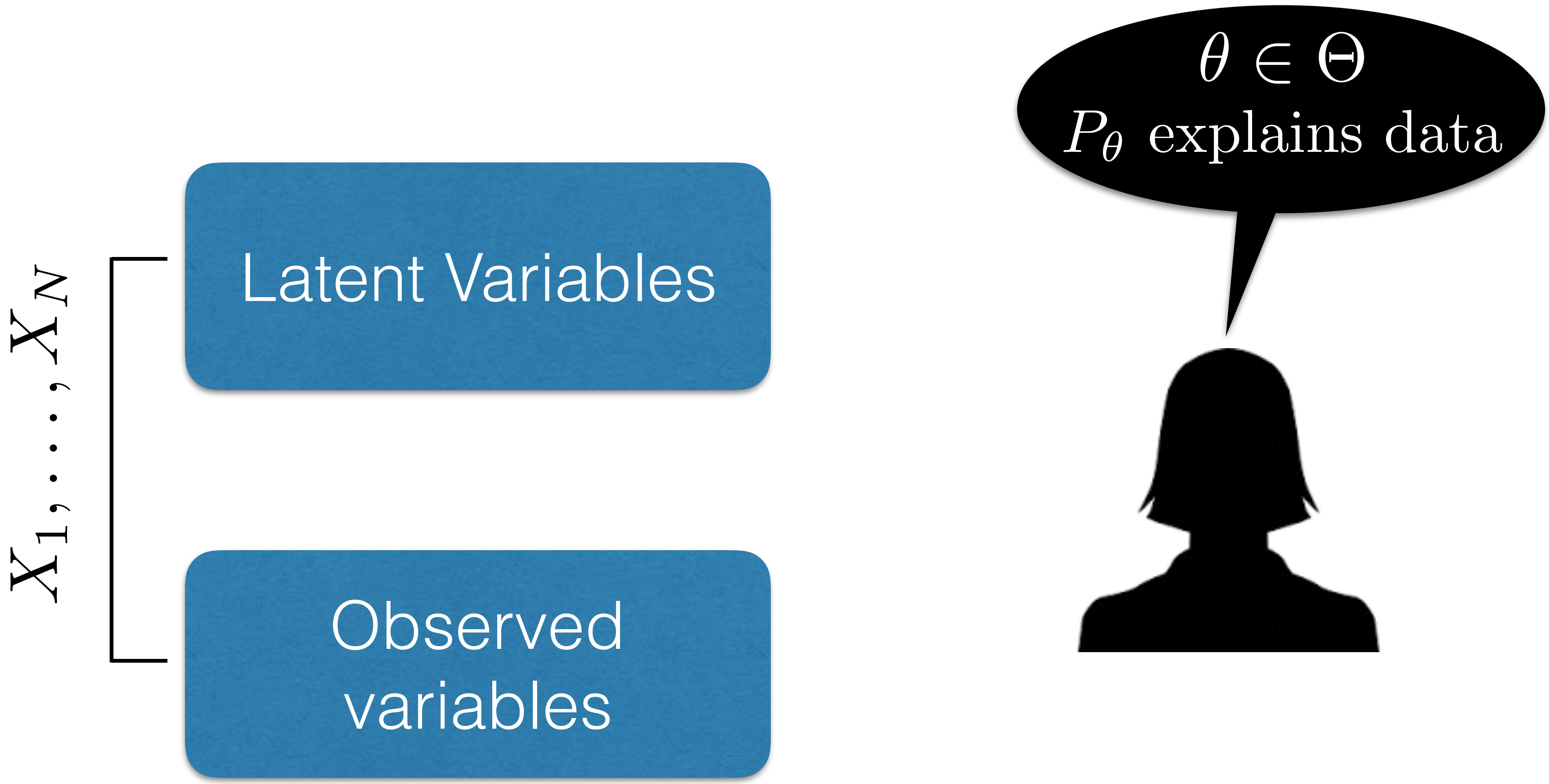
Probabilistic Model

Latent Variables

Observed variables



Probabilistic Model



Graphical Models

- Abstract away the parameterization specifics
- Focus on relationship between random variables

Relationship Between Variables

Let $X = (X_1, \ldots, X_N)$ be the random variables of our model (both latent and observed)

- Joint probability distribution over variable can be complex esp. if we have many complexly related variables
- Can we represent relation between variables in conceptually simpler fashion?
- We often have prior knowledge about the dependencies (or conditional (in)dependencies) between variables

Graphical Models

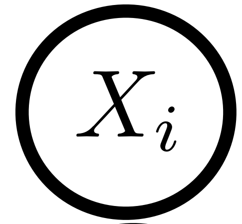
- A graph whose nodes are variables $X_1, \ldots, X_N$
- Graphs are an intuitive way of representing relationships between large number of variables
- Allows us to abstract out the parametric form that depends on θ and the basic relationship between the random variables.

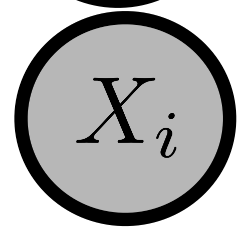
Graphical Models

- A graph whose nodes are variables $X_1, \ldots, X_N$
- Graphs are an intuitive way of representing relationships between large number of variables
- Allows us to abstract out the parametric form that depends on θ and the basic relationship between the random variables.

Draw a picture for the generative story
that explains what generates what.

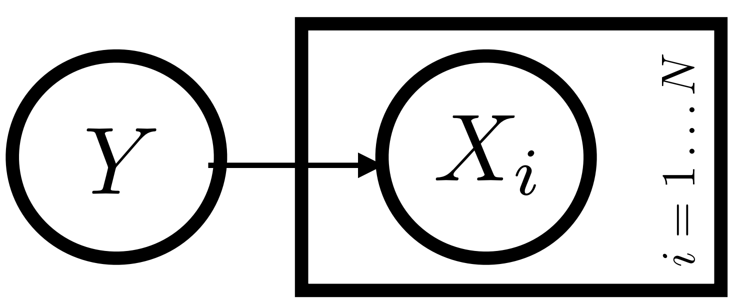
Graphical Models

- Variables X_i is written as 

if X_i is observed
- Variables X_i is written as 

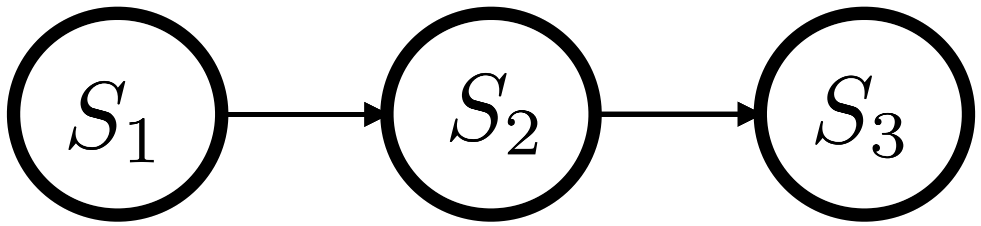
if X_i is latent
- Parameters are often left out (its understood and not explicitly written out). If present they don't have bounding objects

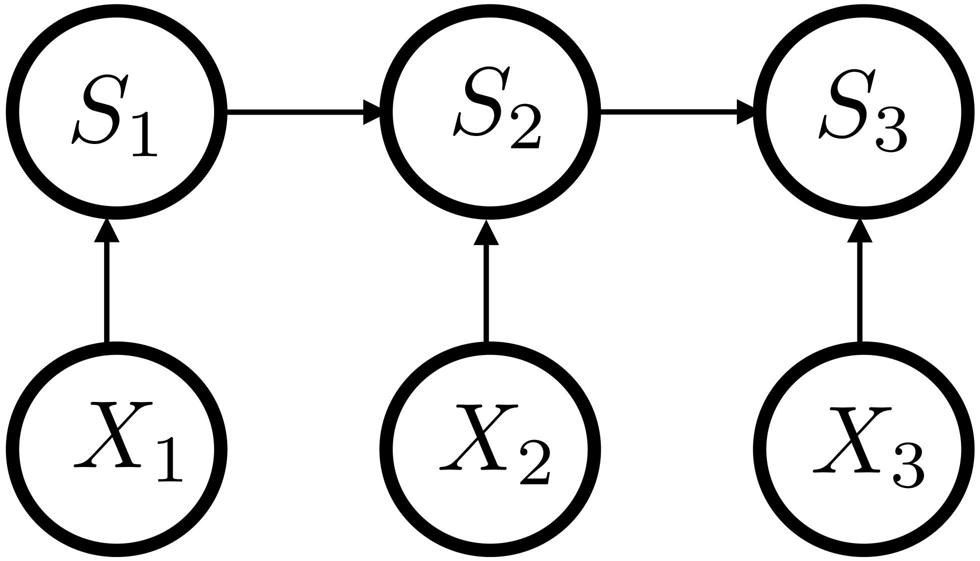
- An directed edge $\longrightarrow$ is drawn connecting every parent to its child (from parent to child)



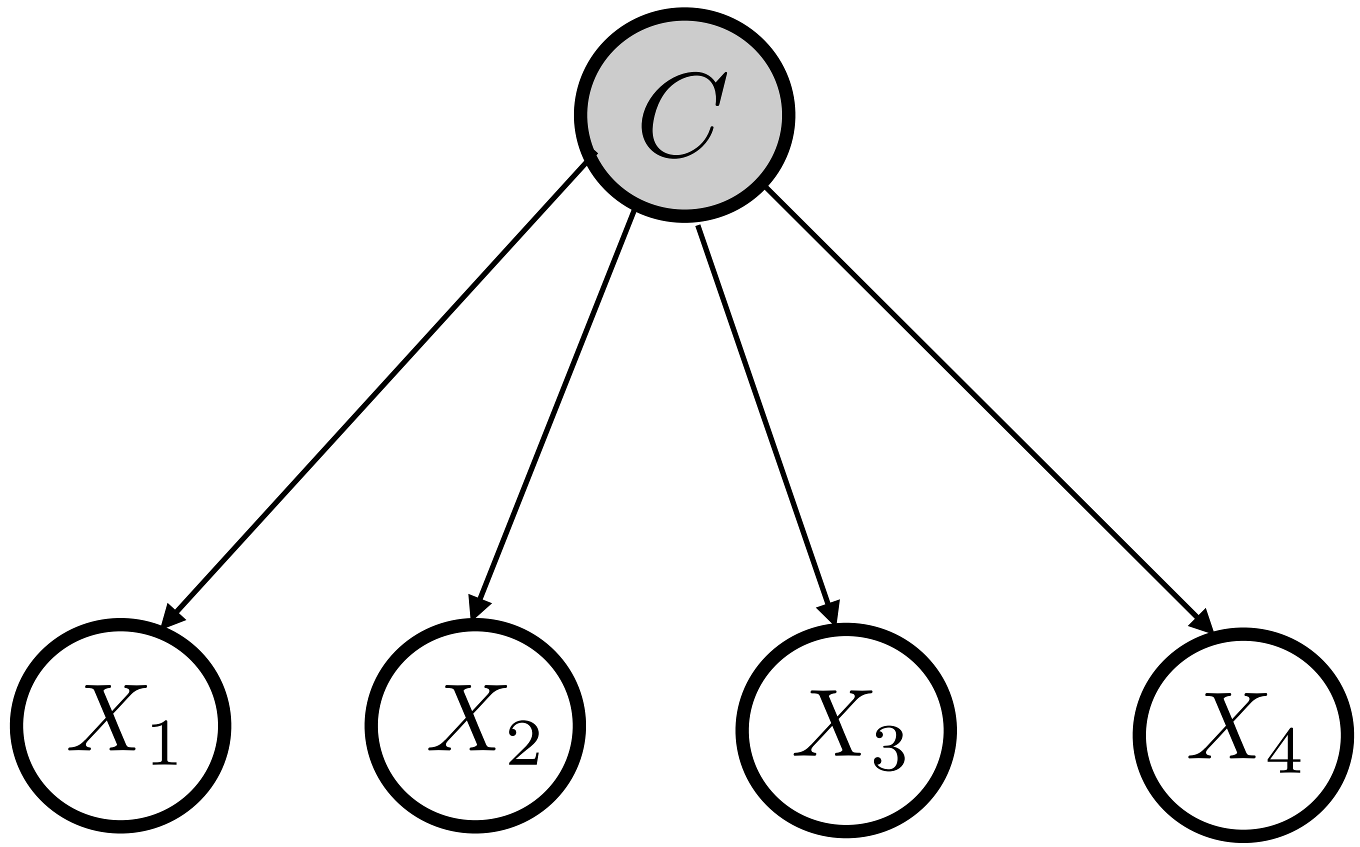
$X_1 \ldots X_N$ drawn repeatedly from $P(Y|X)$

Example: Sum of Coin Flips



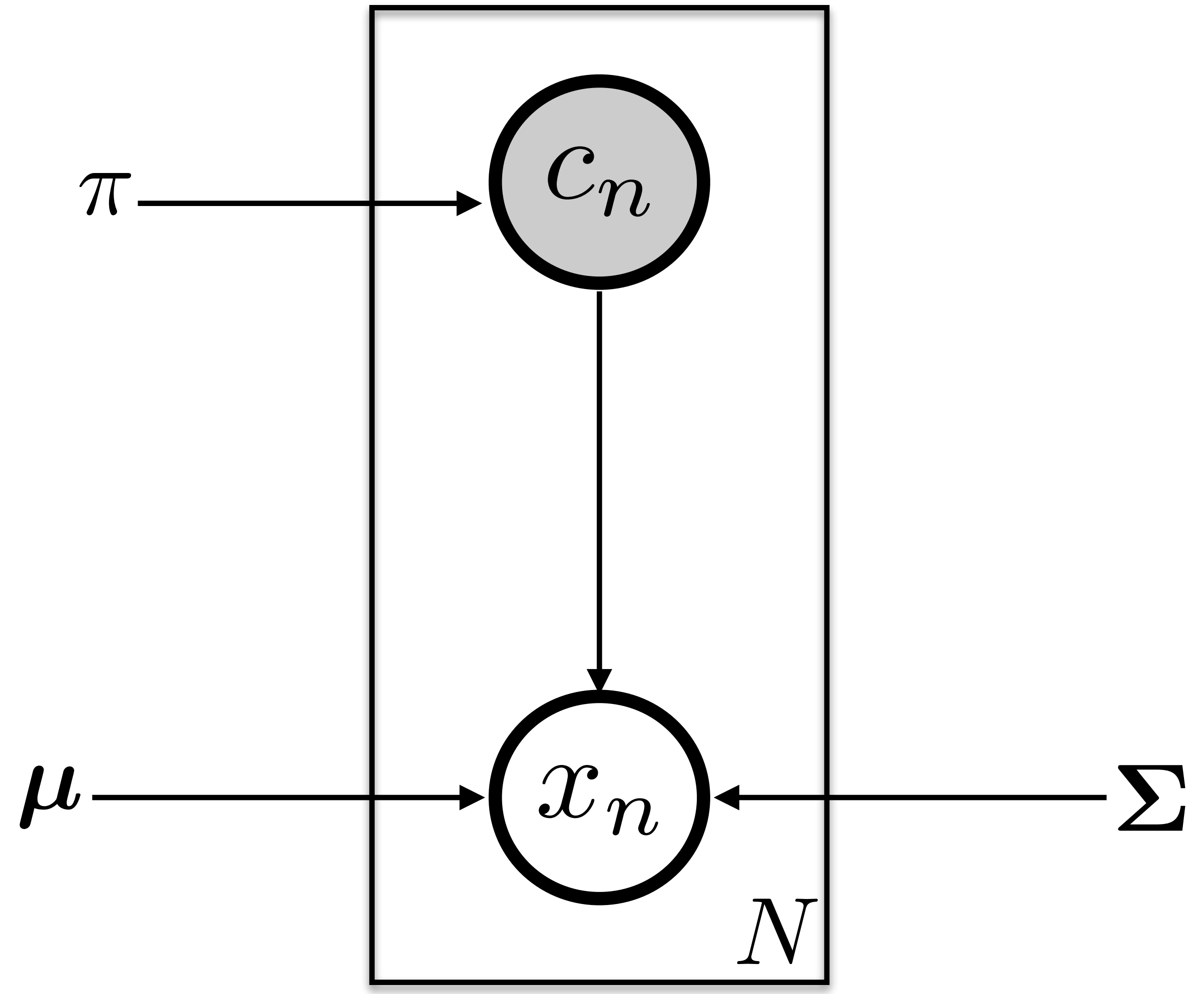


Example: Naive Bayes Classifier



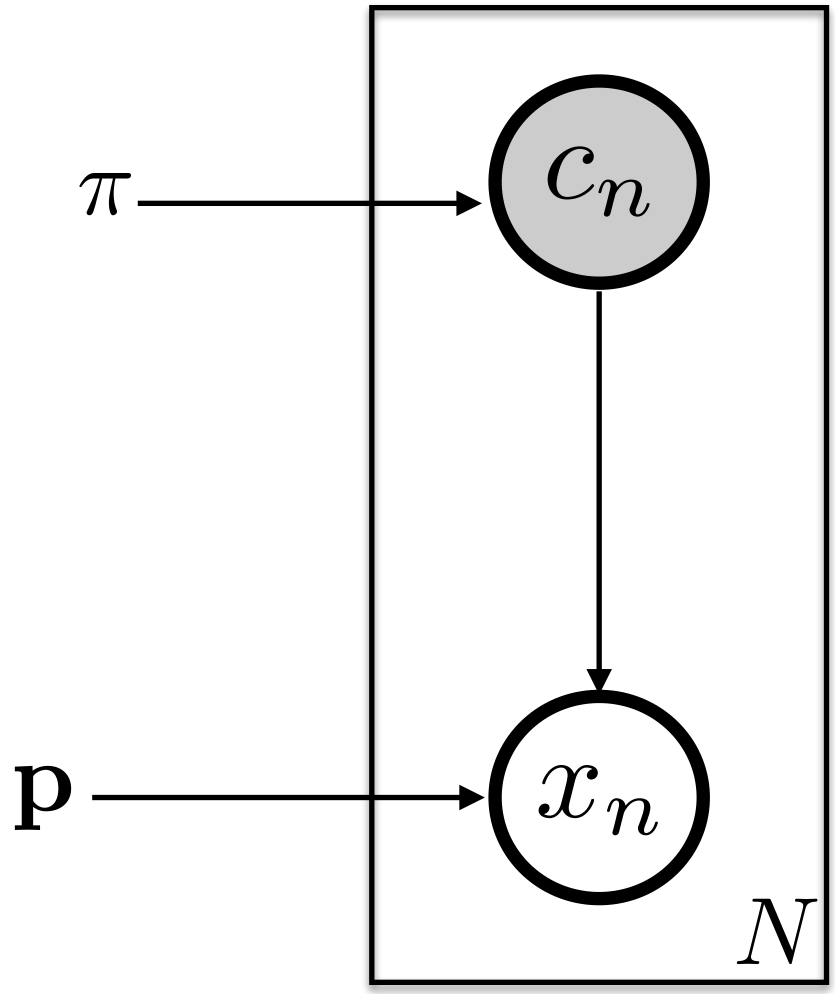
Eg. Spam classification

Example: Mixture Models

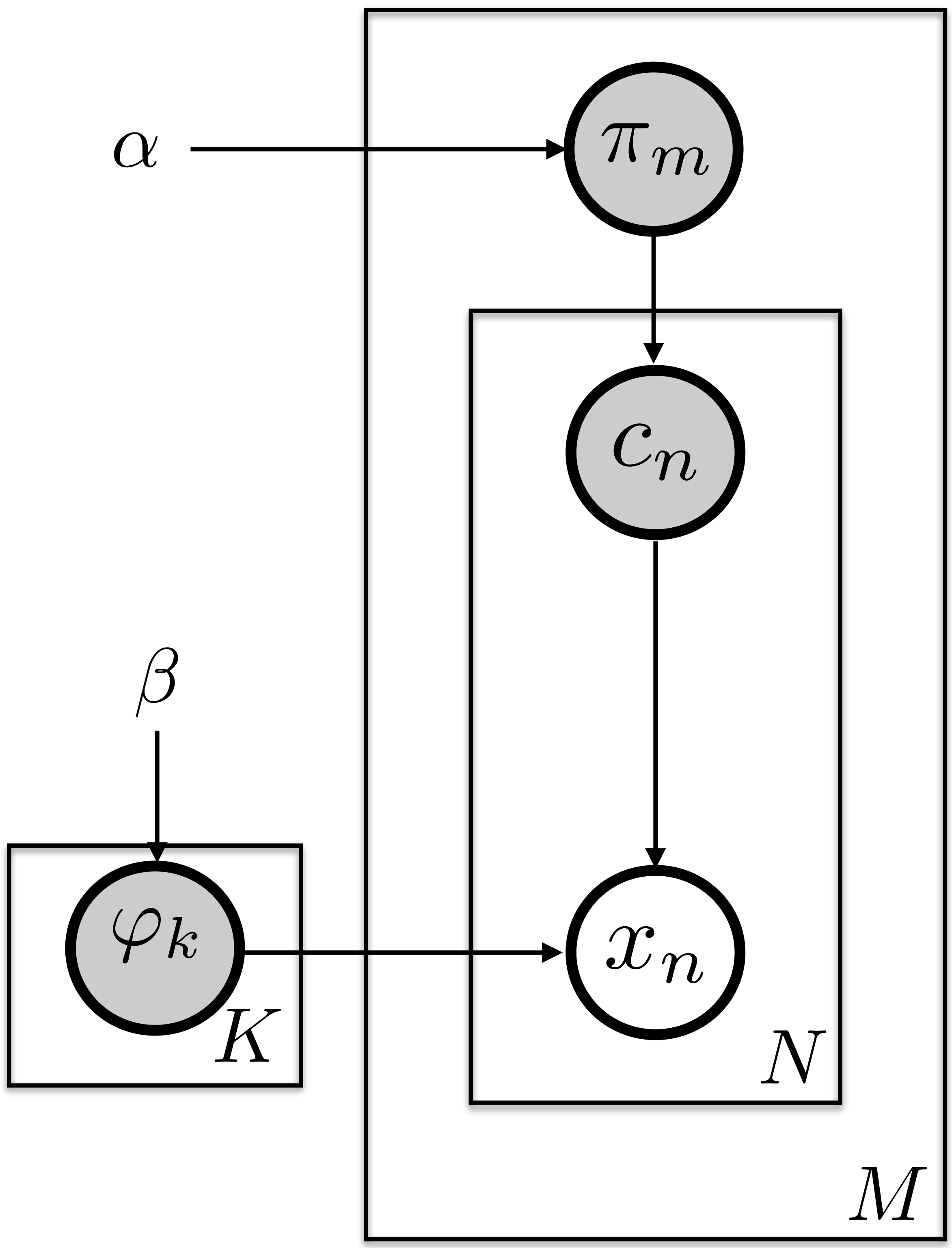


Eg. Clustering

Mixture of Multinomials

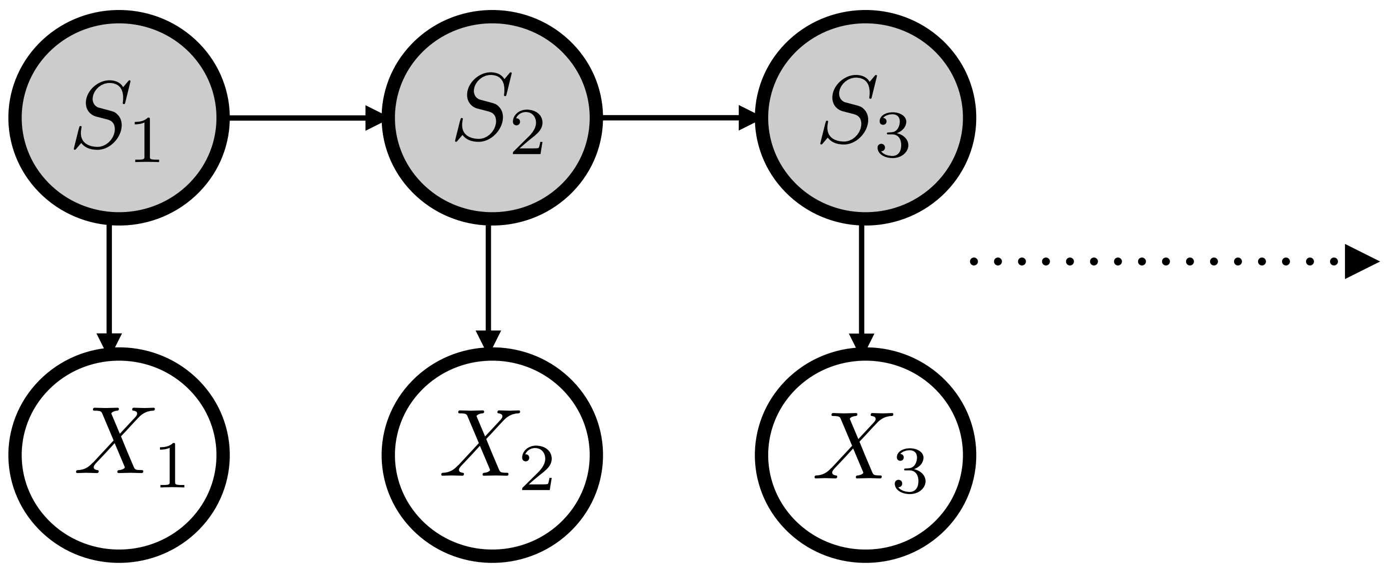


EXAMPLE: LATENT DIRICHLET ALLOCATION



Eg. Topic modelling

EXAMPLE: HIDDEN MARKOV MODEL



Eg. Speech recognition

Bayesian Networks

Bayesian Networks

- Directed acyclic graph G = (V,E) **(graph with no directed cycle)**

Bayesian Networks

- Directed acyclic graph G = (V,E) **(graph with no directed cycle)**
  - Edges going from parent nodes to child nodes

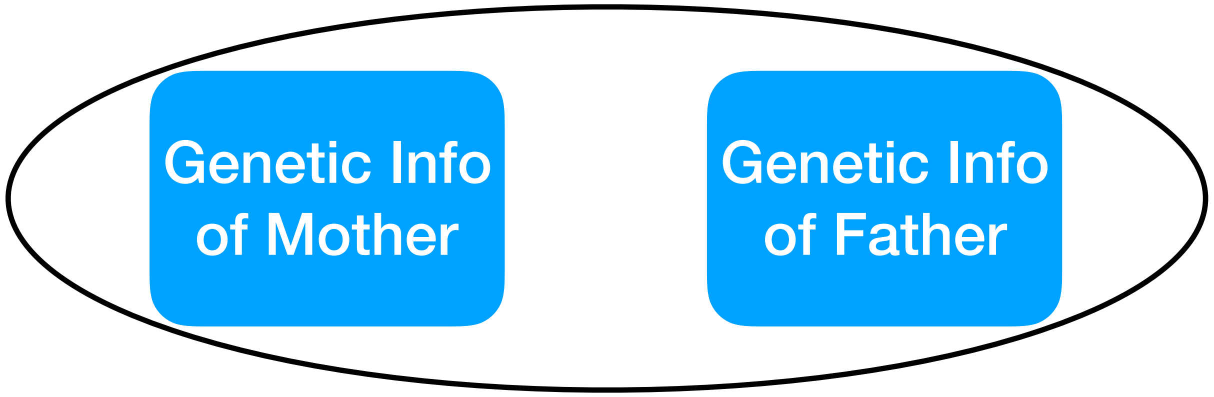
Bayesian Networks

- Directed acyclic graph $G = (V,E)$ **(graph with no directed cycle)**
  - Edges going from parent nodes to child nodes
  - Direction indicates parent "generates" child

Example: CI and MI

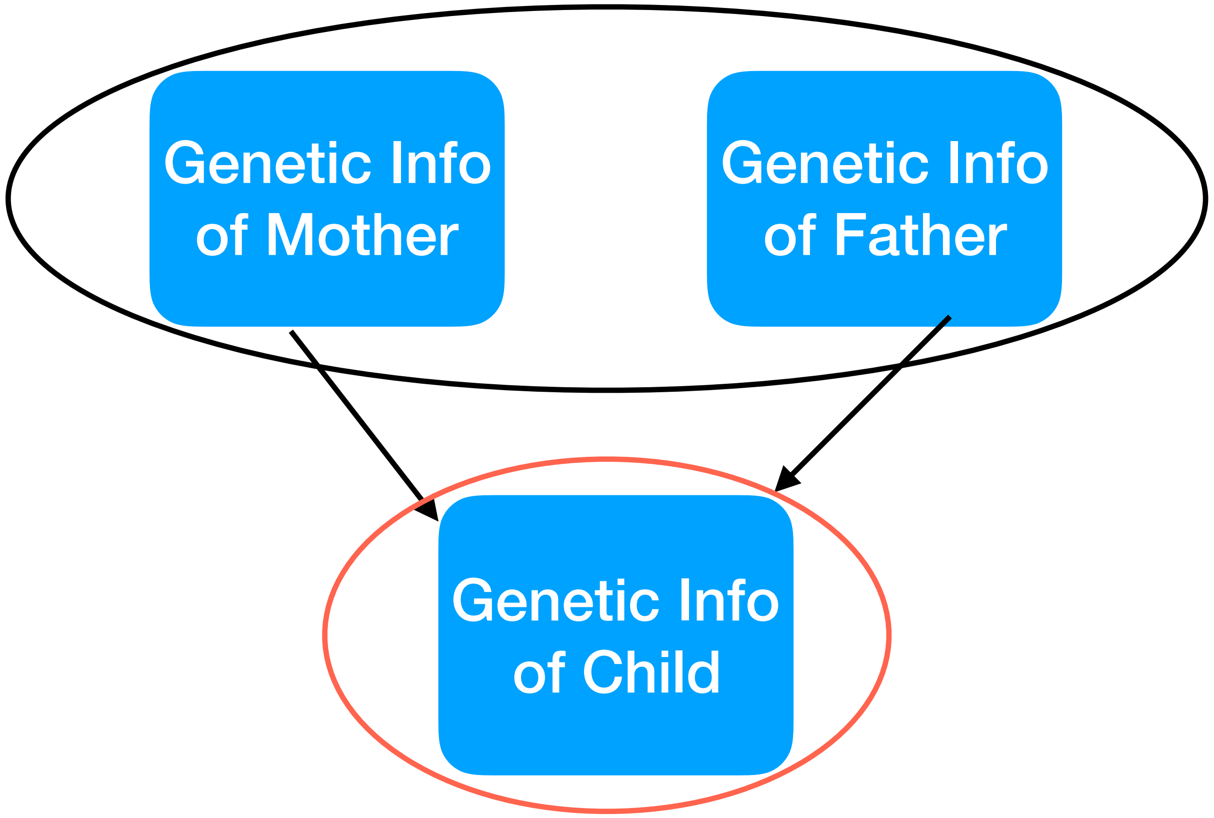
Example: CI and MI

Marginally independent



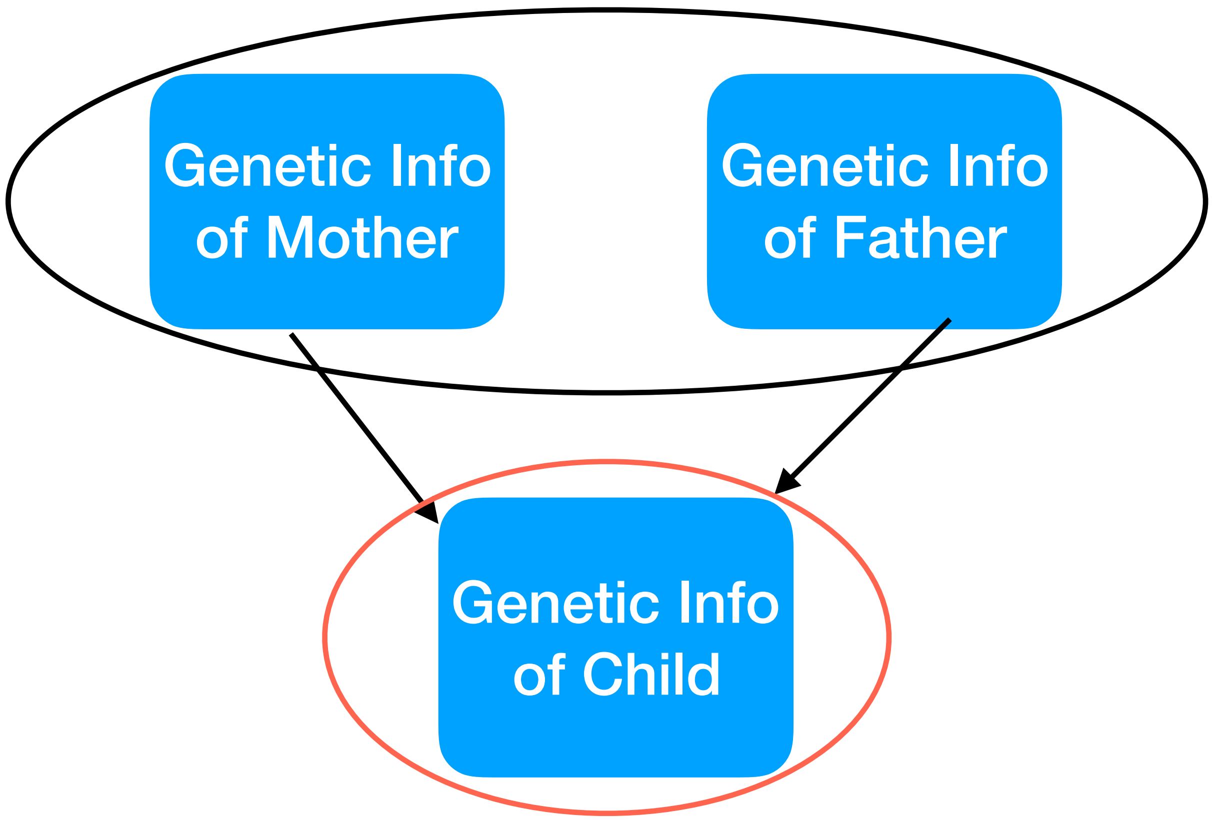
Example: CI and MI

Marginally independent
but Conditionally dependent
given child

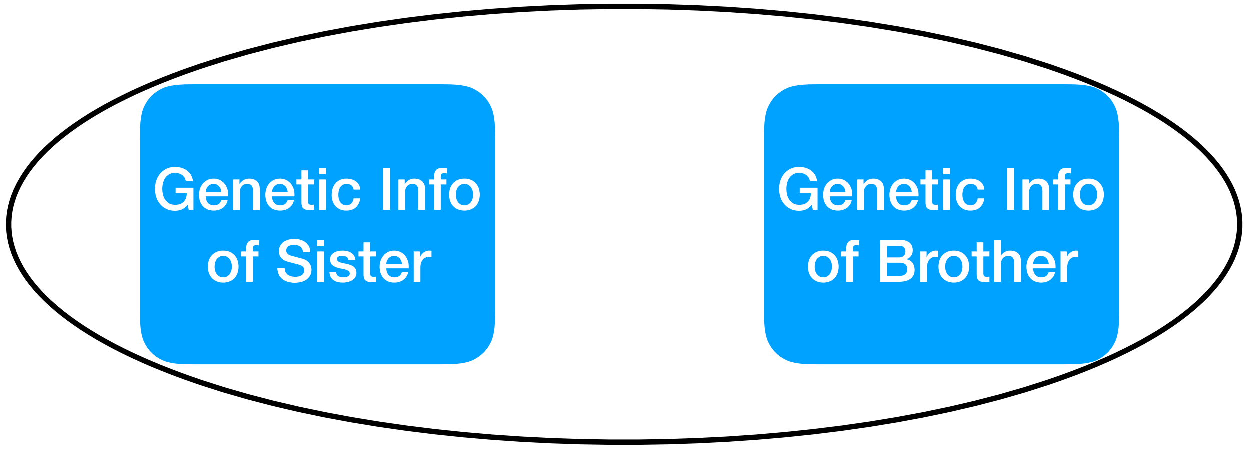


Example: CI and MI

Marginally independent
but Conditionally dependent
given child

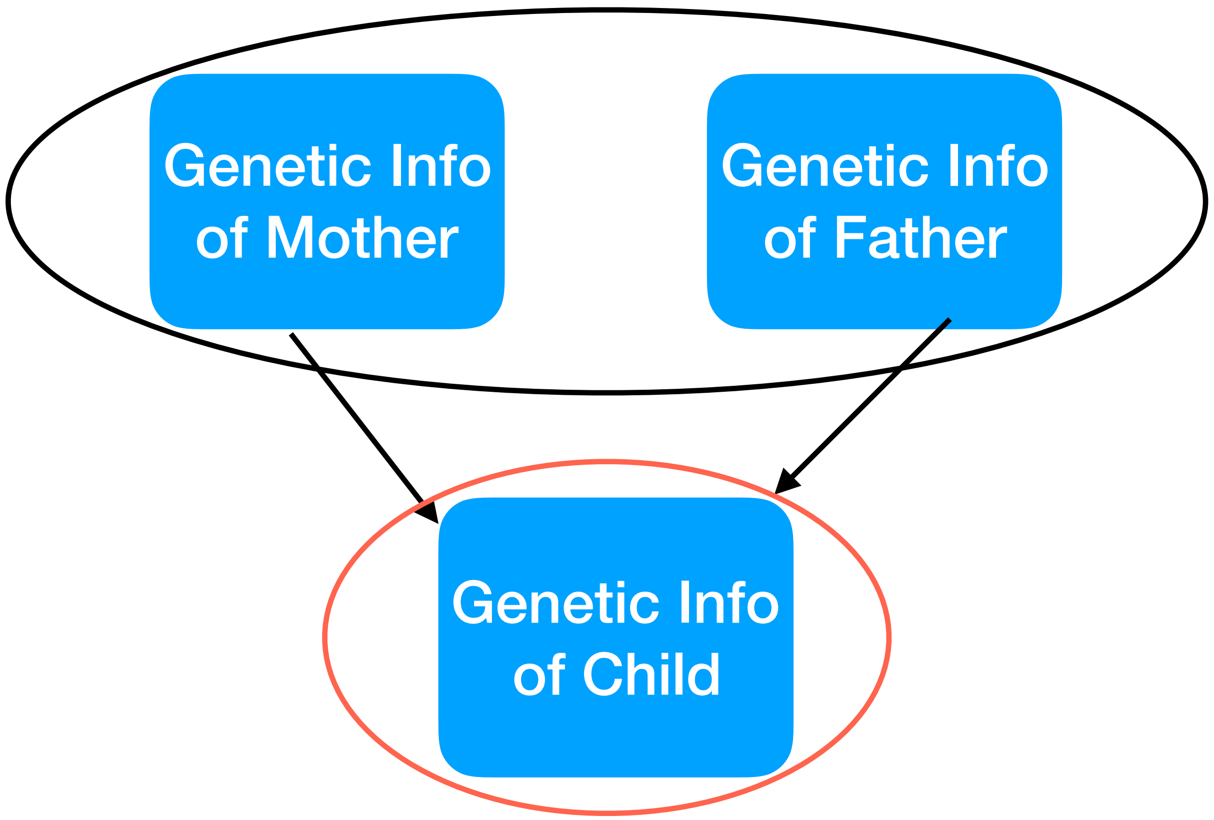


Marginally dependent

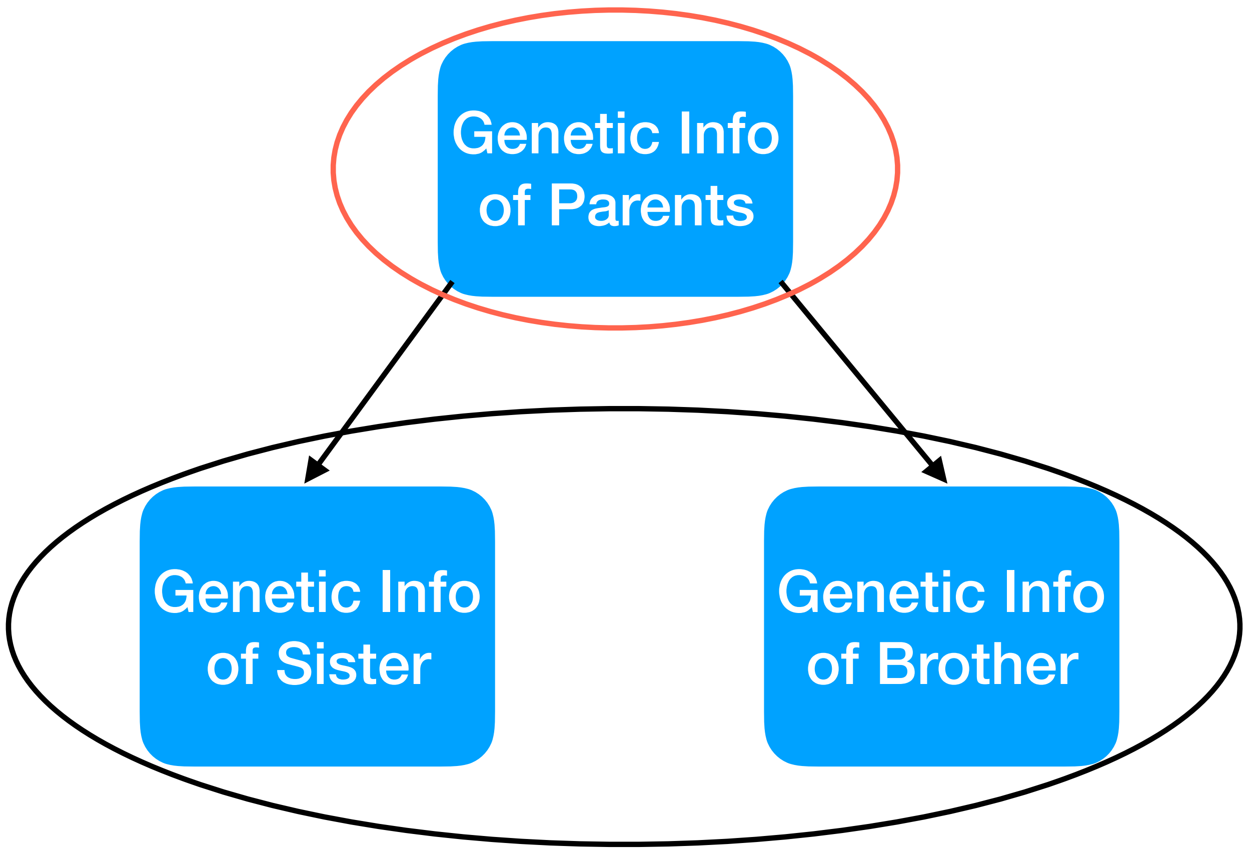


EXAMPLE: CI AND MI

Marginally independent
but Conditionally dependent
given child



Marginally dependent
but Conditionally independent
given Parent



Conditional and Marginal Independence

- Conditional independence
  - X_i is conditionally independent of X_j given $A \subset \{X_1, \ldots, X_N\}$:

$$X_i \perp X_j | A \Leftrightarrow P_\theta(X_i, X_j | A) = P_\theta(X_i|A) \times P_\theta(X_j|A)$$
$$\Leftrightarrow P_\theta(X_i | X_j, A) = P_\theta(X_i|A)$$

- Marginal independence:

$$X_i \perp X_j | \varnothing \Leftrightarrow P_\theta(X_i, X_j) = P_\theta(X_i) P_\theta(X_j)$$

Bayesian Networks

Bayesian Networks

- Directed acyclic graph G = (V,E) **(graph with no directed cycle)**
  - Edges going from parent nodes to child nodes
  - Direction indicates parent “generates” child

Bayesian Networks

- Directed acyclic graph G = (V,E) **(graph with no directed cycle)**
  - Edges going from parent nodes to child nodes
  - Direction indicates parent “generates” child
- Local Markov Property: Each node conditionally independent of its non-descendants given its parents

BAYESIAN NETWORKS

- Directed acyclic graph G = (V,E) **(graph with no directed cycle)**
  - Edges going from parent nodes to child nodes
  - Direction indicates parent "generates" child
- Local Markov Property: Each node conditionally independent of its non-descendants given its parents

Joint probability factorizes as:

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | \text{Parents}(X_i))$$

Local Markov Property

- Each variable is conditionally independent of its non-descendants given its parents
- Any joint distribution satisfying the local markov property w.r.t. graph factorizes over the graph

Local Markov Property

- Each variable is conditionally independent of its non-descendants given its parents
- Any joint distribution satisfying the local markov property w.r.t. graph factorizes over the graph

Why?

Factorizing Joint Probability

- Fact about DAG: we obtain an ordering of nodes (called topological sort) such that for every directed edge between X_i to X_j, X_i appears before X_j in sorted order.

Factorizing Joint Probability

- Fact about DAG: we obtain an ordering of nodes (called topological sort) such that for every directed edge between X_i to X_j, X_i appears before X_j in sorted order.
- Assume nodes are arranged according to some topological sort

Factorizing Joint Probability

- Fact about DAG: we obtain an ordering of nodes (called topological sort) such that for every directed edge between X_i to X_j, X_i appears before X_j in sorted order.
- Assume nodes are arranged according to some topological sort
- For any distribution we have:

$$P_\theta(X_1, \ldots, X_N) = \prod_{i=1}^{N} P_\theta(X_i | X_1, \ldots, X_{i-1})$$

Factorizing Joint Probability

- Fact about DAG: we obtain an ordering of nodes (called topological sort) such that for every directed edge between X_i to X_j, X_i appears before X_j in sorted order.
- Assume nodes are arranged according to some topological sort
- For any distribution we have:

$$
\begin{aligned}
P_\theta(X_1, \ldots, X_N) &= \prod_{i=1}^{N} P_\theta(X_i | X_1, \ldots, X_{i-1}) \\
&= \prod_{i=1}^{N} P_\theta(X_i | \text{Parents}(X_i))
\end{aligned}
$$

(Local Markov Property)

Bayesian Networks

Bayesian Networks

- Bayes net: directed acyclic graph + P(node|parents)

Bayesian Networks

- Bayes net: directed acyclic graph + P(node|parents)
- Directed acyclic graph G = (V,E)
  - Edges going from parent nodes to child nodes
  - Direction indicates parent "generates" child

Bayesian Networks

- Bayes net: directed acyclic graph + P(node|parents)
- Directed acyclic graph G = (V,E)
  - Edges going from parent nodes to child nodes
  - Direction indicates parent “generates” child
- Provide conditional probability table/distribution P(node|parents)

Bayesian Networks

- Bayes net: directed acyclic graph + P(node|parents)
- Directed acyclic graph G = (V,E)
  - Edges going from parent nodes to child nodes
  - Direction indicates parent “generates” child
- Provide conditional probability table/distribution P(node|parents)

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | \text{Parents}(X_i))$$

Representational Power

- Not all joint distributions can be represented by Bayesian Networks
- Eg. $X_1 \perp X_4 \mid X_3, X_2$ and $X_3 \perp X_2 \mid X_1, X_4$
  This dependence can never be captured by a bayesian network, Why?

Representational Power

- Not all joint distributions can be represented by Bayesian Networks
- Eg. $X_1 \perp X_4 \mid X_3, X_2$ and $X_3 \perp X_2 \mid X_1, X_4$
  This dependence can never be captured by a bayesian network, Why?

Which distributions can be represented by Bayesian networks?

Two main questions

Two main questions

- Learning/estimation: Given observations, can we learn the parameters for the graphical model ?

Two main questions

- Learning/estimation: Given observations, can we learn the parameters for the graphical model ?
- Inference: Given model parameters, can we answer queries about variables in the model

Two main questions

- Learning/estimation: Given observations, can we learn the parameters for the graphical model ?
- Inference: Given model parameters, can we answer queries about variables in the model
    - Eg. what is the most likely value of a latent variable given observations

Two main questions

- Learning/estimation: Given observations, can we learn the parameters for the graphical model ?
- Inference: Given model parameters, can we answer queries about variables in the model
  - Eg. what is the most likely value of a latent variable given observations
  - Eg. What is the distribution of a particular variable conditioned on others

Two main questions

- Learning/estimation: Given observations, can we learn the parameters for the graphical model ?

- Inference: Given model parameters, can we answer queries about variables in the model

  - Eg. what is the most likely value of a latent variable given observations

  - Eg. What is the distribution of a particular variable conditioned on others

E-step in EM is inference of Latent given observed

Inference in Graphical Models

Given parameters of a graphical model, we can answer any questions about distributions of variables in the model

Example queries:

1. What is the probability of a given assignment for a subset of variables (marginal)?

2. What is the probability of a particular assignment of a subset of variables given observed values (evidence) of some subset of the variables (conditional)?

3. Given observed values (evidence) of some subset of variables what is the most likely assignment for a given subset of variables?

Inference in Graphical Models

Given parameters of a graphical model, we can answer any questions about distributions of variables in the model
Example queries:

1. What is the probability of a given assignment for a subset of variables (marginal)?
2. What is the probability of a particular assignment of a subset of variables given observed values (evidence) of some subset of the variables (conditional)?
3. Given observed values (evidence) of some subset of variables what is the most likely assignment for a given subset of variables?

Suffices to be able to compute joint probability

INFERENCE IN GRAPHICAL MODELS

Given parameters of a graphical model, we can answer any questions about distributions of variables in the model

Example queries:

1. What is the probability of a given assignment for a subset of variables (marginal)?
2. What is the probability of a particular assignment of a subset of variables given observed values (evidence) of some subset of the variables (conditional)?
3. Given observed values (evidence) of some subset of variables what is the most likely assignment for a given subset of variables?

Suffices to be able to compute joint probability

Why?

Can compute any marginal from joint :

$$P(A=a, B=b, C=c) = \sum_{d} P(A=a, B=b, C=c, D=d)$$

Can compute any marginal from joint :

$$P(A = a, B = b, C = c) = \sum_{d} P(A = a, B = b, C = c, D = d)$$

Can compute any conditional from marginal :

$$P(A = a | B = b, C = c) = \frac{P(A = a, B = b, C = c)}{P(B = b, C = c)}$$

Can compute any marginal from joint :

$$P(A=a,B=b,C=c)=\sum_{d} P(A=a,B=b,C=c,D=d)$$

Can compute any conditional from marginal :

$$P(A=a|B=b,C=c)=\frac{P(A=a,B=b,C=c)}{P(B=b,C=c)}$$

For Bayesian Networks P(node|Parents) completely defines joint.

Next class

- Start with example of Hidden Markov Model (HMM)