Graphical Models For Complex Health Data (P8124)

Daniel Malinsky

Columbia University
dsm2128@cumc.columbia.edu

Introductory Lecture

What is this course?

This is a course about **Graphical Models**.

GMs are statistical tools with applications in neuroscience, computational biology, epidemiology, economics, diagnostic systems, climate science, robotics, NLP, and other areas. We'll focus on some application domains relevant to public health, especially neuro/psych, genetics, epi, and image analysis.

This course will involve a fair bit of mathematical theory, some programming (in R), and reading/understanding of research papers.

Note: **this course is not about data visualization.** Not *that* sense of "graphical" (bar graphs, scatterplots, charts, etc.)

What is a graphical model?

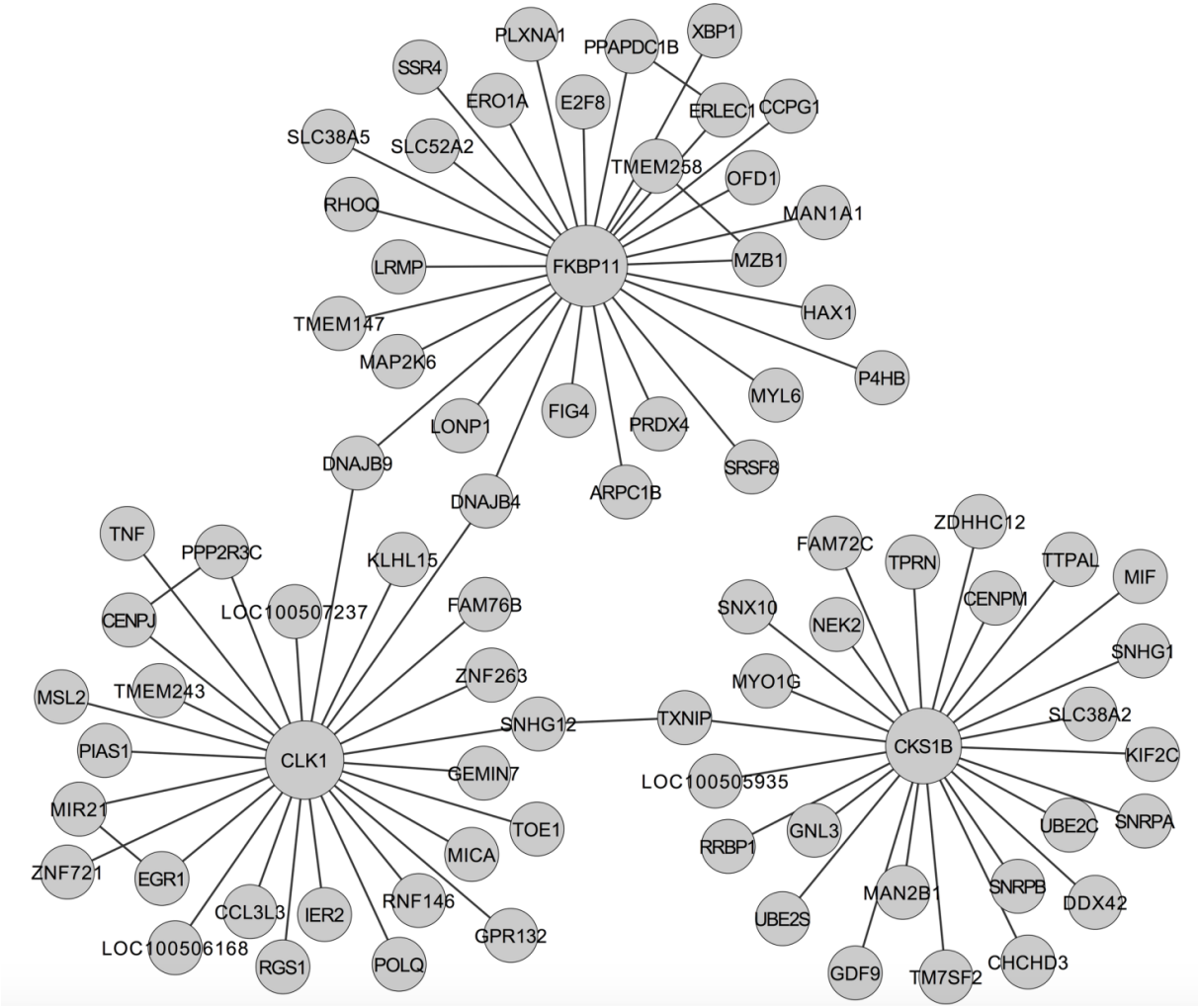


Figure: Gene associations related to asthma, an undirected graph from Wang et al. (2016).

What is a graphical model?

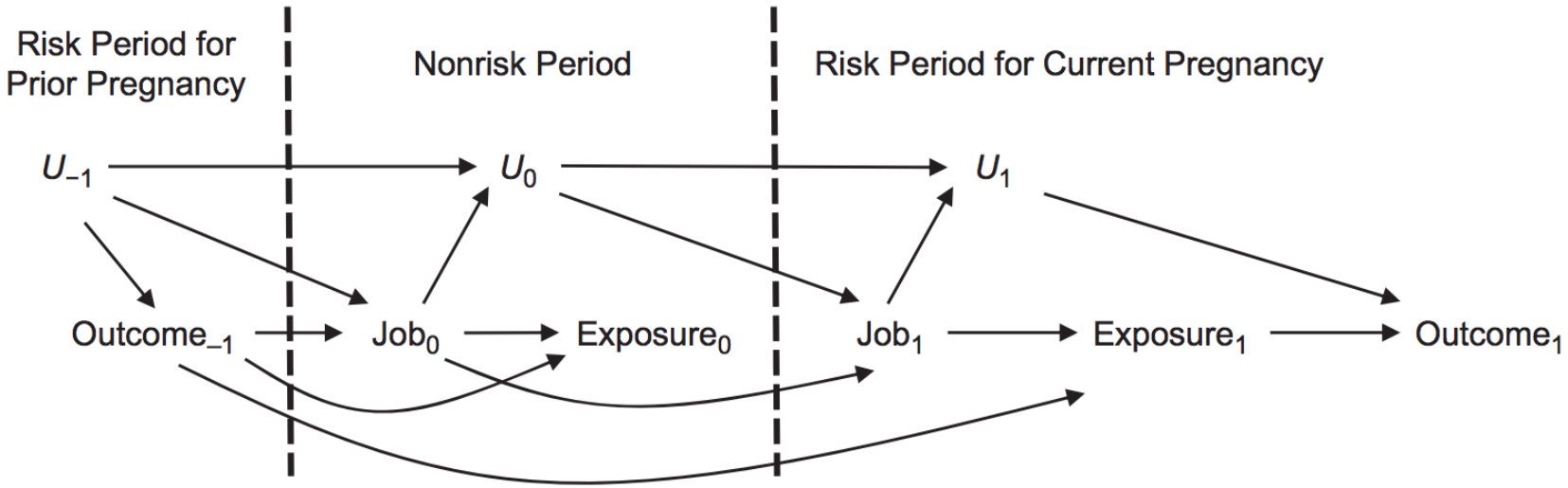


Figure: A directed acyclic graph from Johnson et al. (2019) that illustrates epidemiologic phenomena related to pregnancy risk.

What is a graphical model?

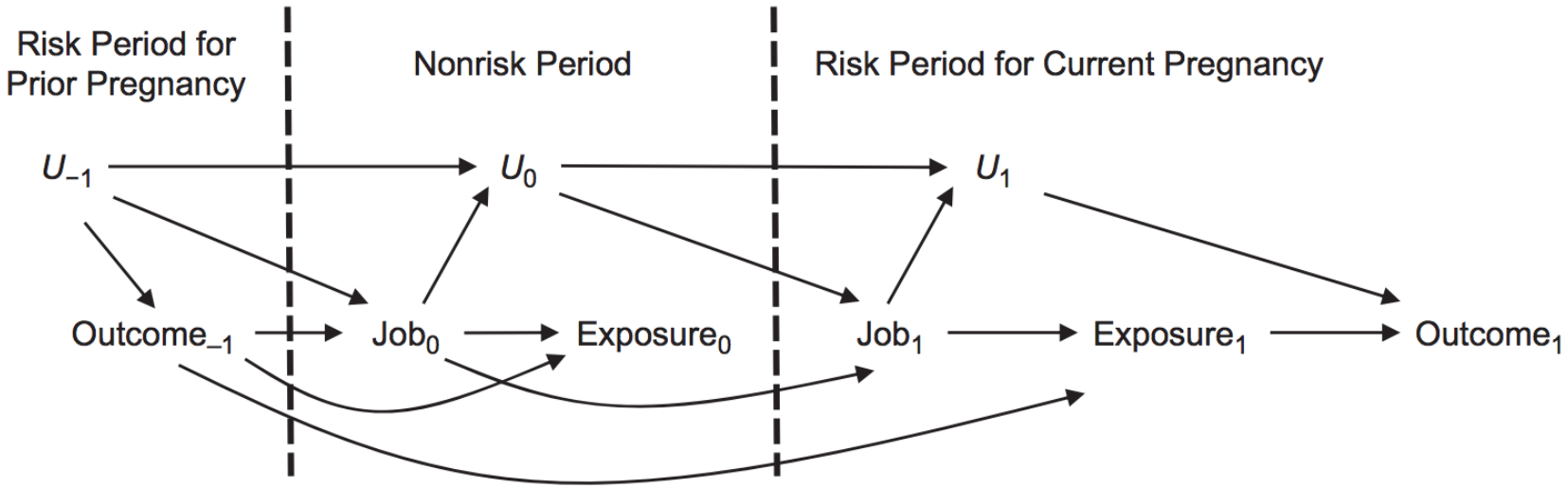


Figure: A directed acyclic graph from Johnson et al. (2019) that illustrates epidemiologic phenomena related to pregnancy risk.

- Vertices (nodes) correspond to random variables and edges (arrows) correspond to statistical dependence of some kind.

Logistics

Instructor: Daniel Malinsky (dsm2128@cumc.columbia.edu)

- Office Hours: Thursday 2-3pm

Teaching Assistant: Angel Garcia de la Garza (ag3914@cumc.columbia.edu)

- Office Hours: TBA

Text: **none required** but I'll be posting readings from several books on courseworks. See syllabus.

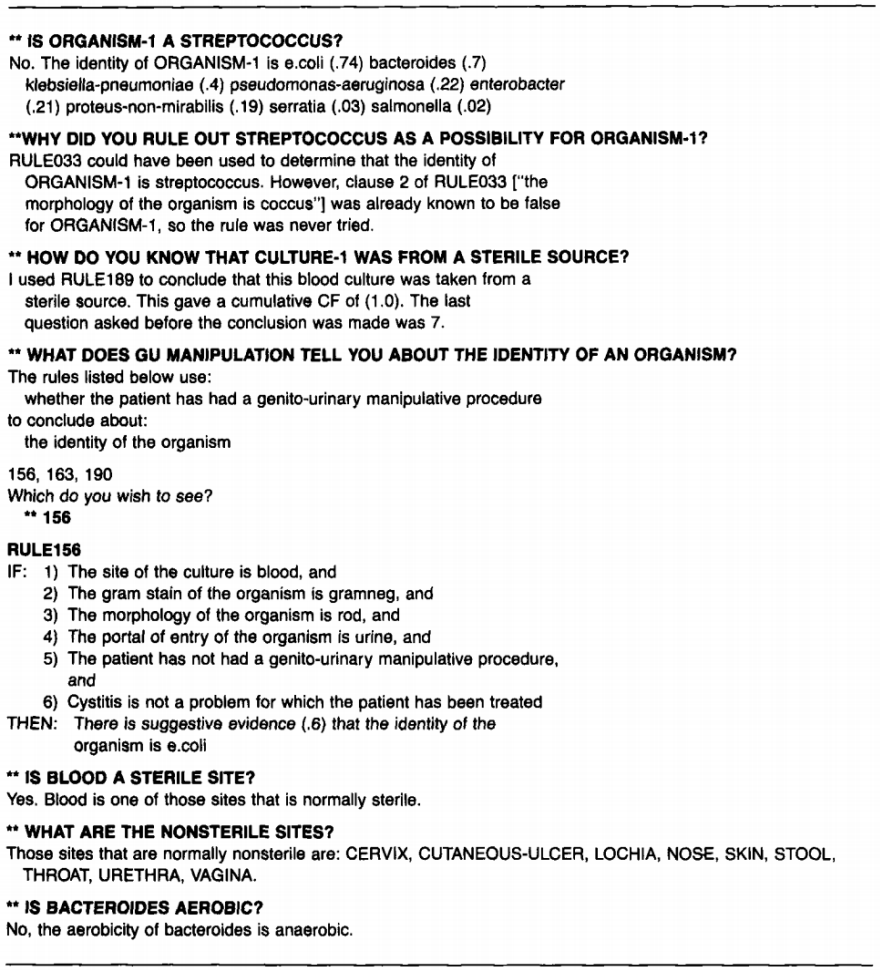
Course structure

- 4 homework assignments = 4x15=60% of your grade
  - Homeworks will have theory questions as well as some programming
- Final project is a data analysis project, worth 30-35% of your grade
  - Analyze a real data set using some/several of the techniques we talk about in class, write a 4-7 page research paper describing your procedure and the results (more details later)
  - This is a very open-ended, creative project. It is not something you can do (well) at the last-minute
  - 5% of grade will be your project proposal, to submit before starting the final
- Remaining 5% is class participation

Questions?

If not, let's get started...

Humble beginnings in expert systems



**** IS ORGANISM-1 A STREPTOCOCCUS?**
No. The identity of ORGANISM-1 is e.coli (.74) bacteroides (.7)
klebsiella-pneumoniae (.4) pseudomonas-aeruginosa (.22) enterobacter
(.21) proteus-non-mirabilis (.19) serratia (.03) salmonella (.02)

****WHY DID YOU RULE OUT STREPTOCOCCUS AS A POSSIBILITY FOR ORGANISM-1?**
RULE033 could have been used to determine that the identity of
ORGANISM-1 is streptococcus. However, clause 2 of RULE033 ["the
morphology of the organism is coccus"] was already known to be false
for ORGANISM-1, so the rule was never tried.

**** HOW DO YOU KNOW THAT CULTURE-1 WAS FROM A STERILE SOURCE?**
I used RULE189 to conclude that this blood culture was taken from a
sterile source. This gave a cumulative CF of (1.0). The last
question asked before the conclusion was made was 7.

**** WHAT DOES GU MANIPULATION TELL YOU ABOUT THE IDENTITY OF AN ORGANISM?**
The rules listed below use:
whether the patient has had a genito-urinary manipulative procedure
to conclude about:
the identity of the organism

156, 163, 190
Which do you wish to see?
**** 156**

RULE156
IF: 1) The site of the culture is blood, and
2) The gram stain of the organism is gramneg, and
3) The morphology of the organism is rod, and
4) The portal of entry of the organism is urine, and
5) The patient has not had a genito-urinary manipulative procedure,
and
6) Cystitis is not a problem for which the patient has been treated
THEN: *There is suggestive evidence (.6) that the identity of the*
organism is e.coli

**** IS BLOOD A STERILE SITE?**
Yes. Blood is one of those sites that is normally sterile.

**** WHAT ARE THE NONSTERILE SITES?**
Those sites that are normally nonsterile are: CERVIX, CUTANEOUS-ULCER, LOCHIA, NOSE, SKIN, STOOL,
THROAT, URETHRA, VAGINA.

**** IS BACTEROIDES AEROBIC?**
No, the aerobicity of bacteroides is anaerobic.

Figure: A rule-based expert system to diagnose bacterial infections, MYCIN

ca. 1960's

Humble beginnings in expert systems

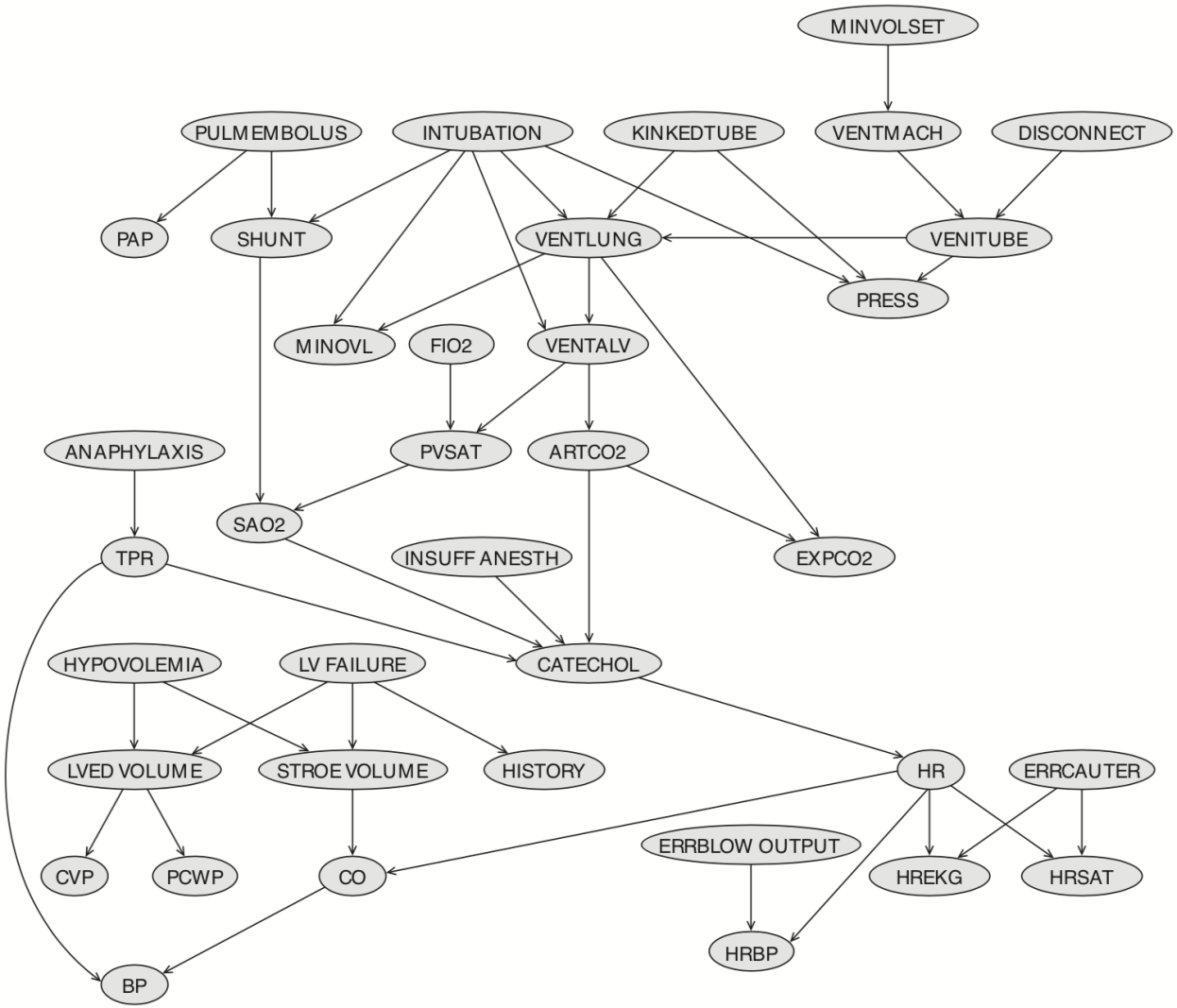
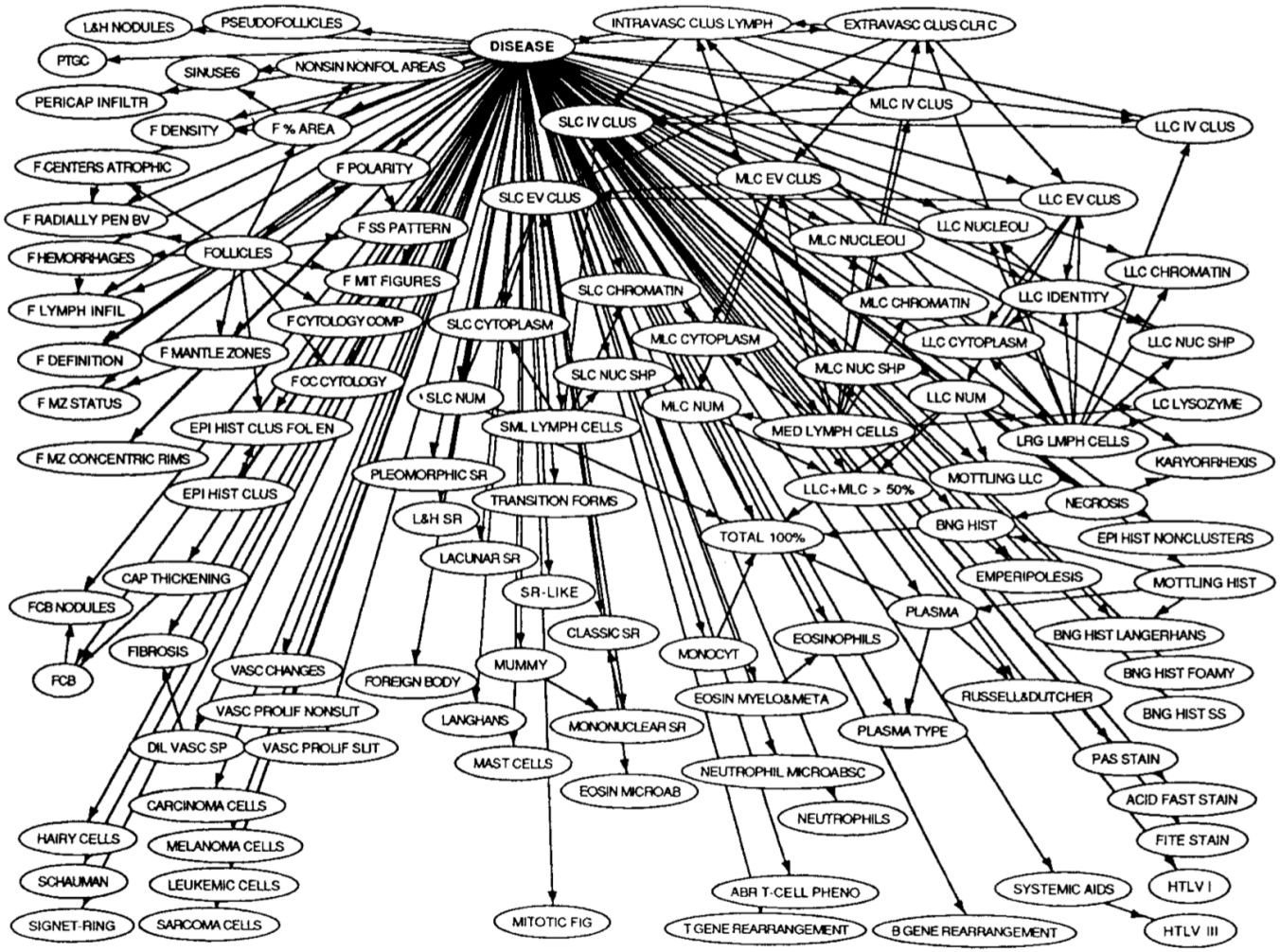


Figure 17.C.1 — The ICU-Alarm **Bayesian network.**

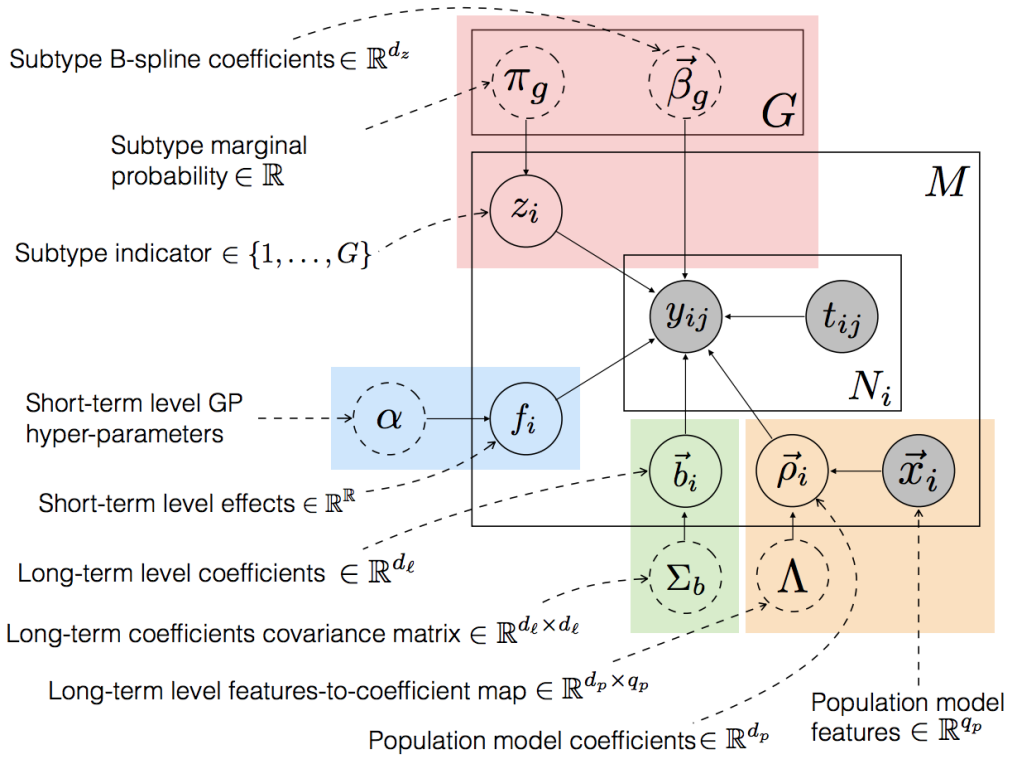
ca. 1980's

Humble beginnings in expert systems



ca. 1990, Pathfinder network w/ 100 symptoms and 60 diseases

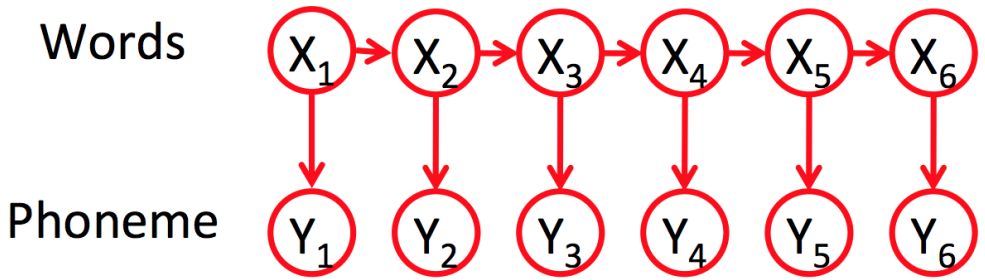
Modern applications



Bayesian/hierarchical model for predicting clinical disease trajectories (Schulam and Saria 2016)

Modern applications

Speech recognition



"He ate the cookies on the couch"

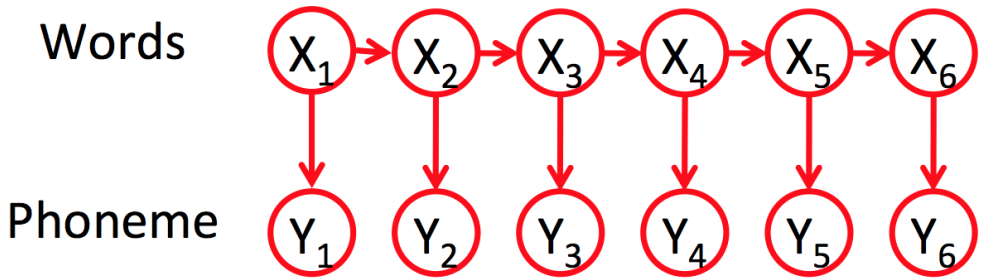
- Infer spoken words from audio signals
- "Hidden Markov Models"

6

(from slides by Andreas Krause)

Modern applications

Speech recognition



"He ate the cookies on the couch"

- Infer spoken words from audio signals
- "Hidden Markov Models"

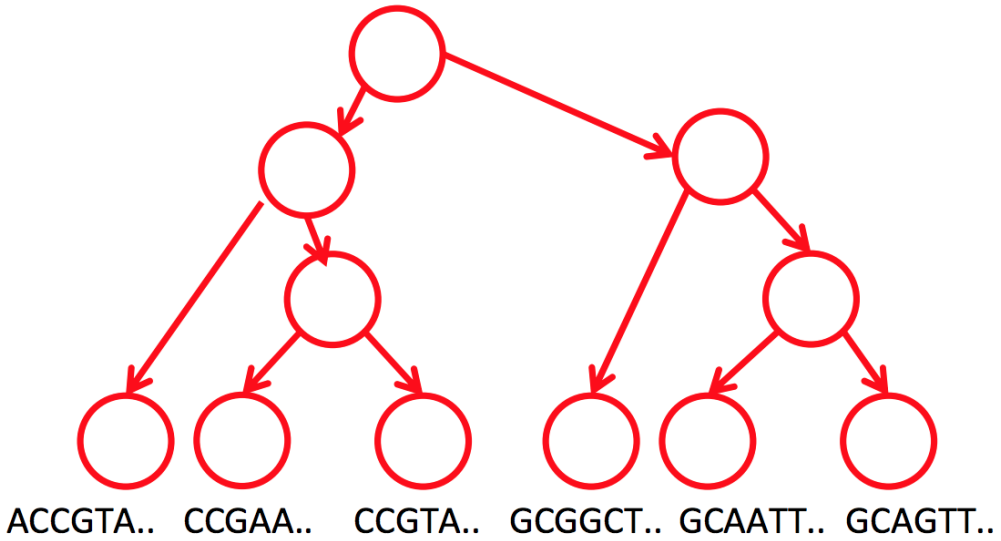
6

(from slides by Andreas Krause)

Modern applications

Evolutionary biology

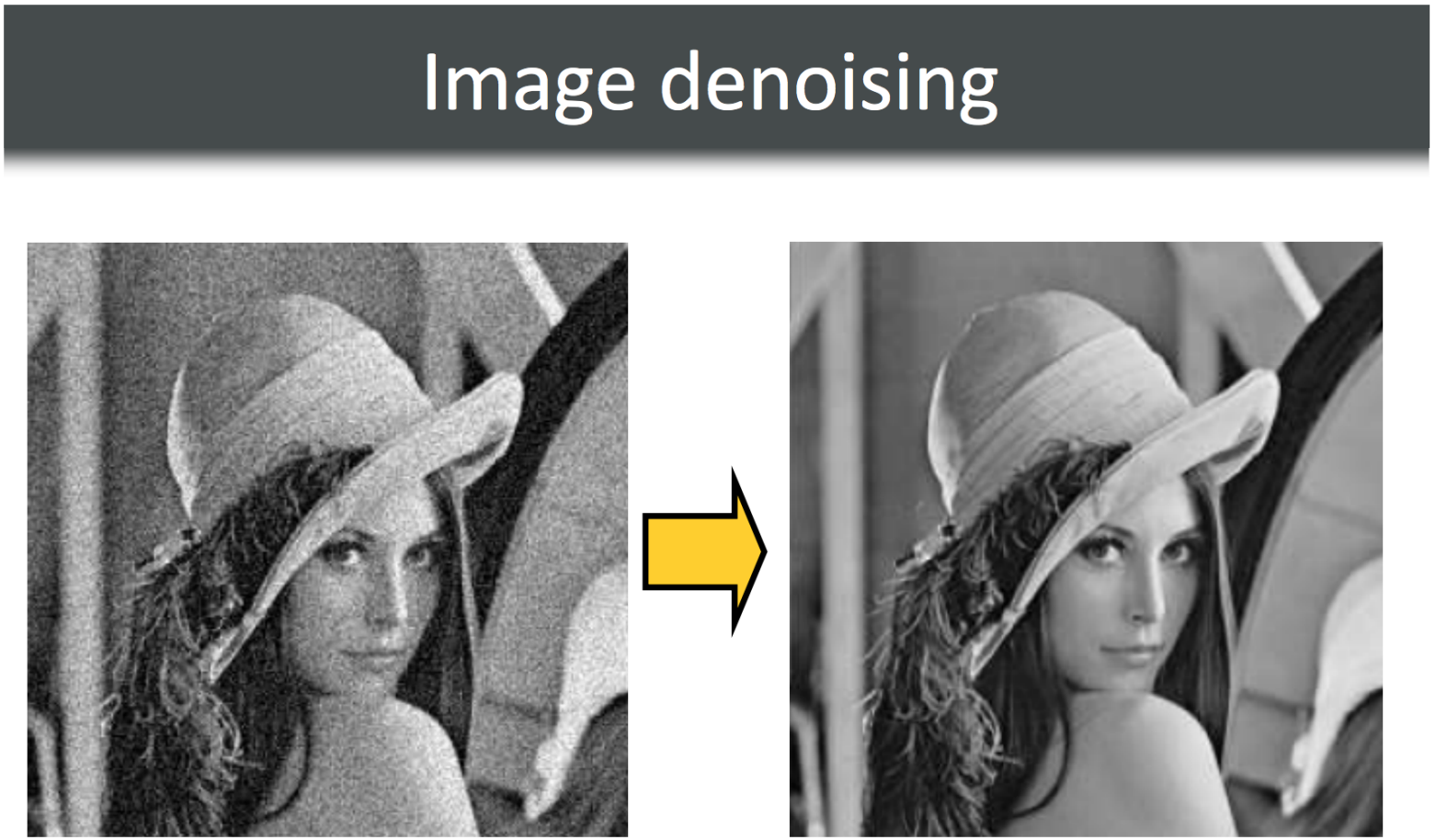
[Friedman et al.]



- Reconstruct phylogenetic tree from current species (and their DNA samples)

(from slides by Andreas Krause)

Modern applications

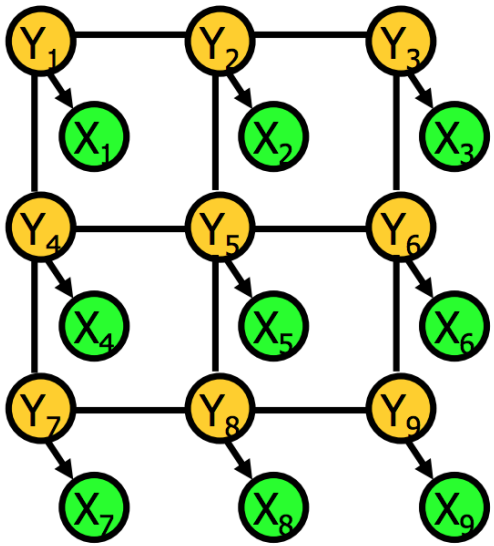


(from slides by Andreas Krause)

Modern applications

Image denoising

Markov Random Field



X_i: noisy pixels

Y_i: "true" pixels

12

(from slides by Andreas Krause)

What is a model?

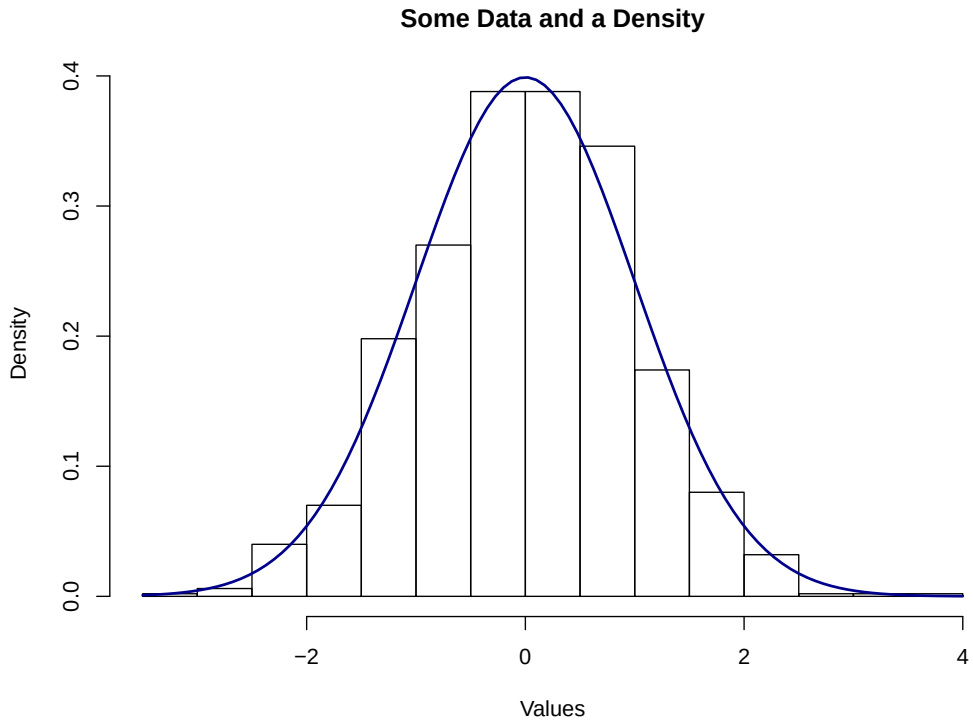
Informally: a model is a mathematical description of some system under study. A probabilistic (or statistical) model quantifies uncertainties in a way such that is useful for *prediction* and sometimes *control*.

What is a model?

Informally: a model is a mathematical description of some system under study. A probabilistic (or statistical) model quantifies uncertainties in a way such that is useful for *prediction* and sometimes *control*.

More formally: in probabilistic or statistical modeling we use probability distributions to summarize the chancy processes which generate observed data. A *model* is defined to be **a set of probability distributions** which may have generated the data.

Model example



$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$ or $X \sim N(\mu, \sigma^2)$.

For predicting that X falls into interval $[a, b]$, compute:
$P(a \leq X \leq b) = \int_a^b p(x)dx$.

Note: with parameters μ and σ^2 fixed to some specific values, e.g., 0 and 1 we call $N(0,1)$ a probability distribution (density). With μ and σ^2 allowed to vary over some set, e.g., $\mathbb{R}$ and $\mathbb{R}^+$, we say $N(\mu, \sigma^2)$ represents a *set* of distributions, or parametric *family*, or *model*.

In other words, the model is

$$\left\{ p(x) : p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \right\}$$

On notation

We will always assume that the random variables we are talking about are defined wrt some appropriate background probability space $(\Omega, \mathcal{S}, P)$.

X is a random variable or random vector, e.g., $X = (X_1, ..., X_p)'$

x is a specific value taken by a r.v., or a vector of values, $x = (x_1, ..., x_p)'$

$p(x)$ will be generically used to denote a probability distribution: a probability mass function (pmf) for discrete X or probability density function (pdf) for continuous X.[1] I'm going to assume you can tell which I mean from context. Formally $p(x)$ means $P(X = x)$ in the discrete case as well as the density function evaluated at x in the continuous case; I hope this causes no confusion!

[1] Some people use $f(x)$ for density.

Models

Continuous:

$\{p(x) : X \sim N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$

$\{p(x) : X \sim Unif(a, b), a \in \mathbb{R}, b \in \mathbb{R}, a < b\}$

$\{p(x) : X \sim Beta(\alpha, \beta), \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+\}$

$\{p(x) : \int (p''(x))^2 dx \leq c^2\}$ (for some c^2)

...

Discrete:

$\{p(x) : X \sim Bin(n, p), n \in \mathbb{N}, p \in [0, 1]\}$

$\{p(x) : X \sim Poisson(\lambda), \lambda \in \mathbb{R}^+\}$

$\{p(x) : X \sim Multinom(n, p_1, ..., p_k), n \in \mathbb{N}, p_i > 0, \sum_i p_i = 1\}$

...

Joint distributions

Consider two random variables, X and Y taking on values in $\{0, 1\}$. How many numbers (parameters) do we need to specify to fully describe the joint distribution $p(x, y)$?

We want to know $p(0, 0), p(0, 1), p(1, 0), p(1, 1)$. Easy. 3 numbers.

Joint distributions

Consider two random variables, X and Y taking on values in $\{0, 1\}$. How many numbers (parameters) do we need to specify to fully describe the joint distribution $p(x, y)$?

We want to know $p(0, 0), p(0, 1), p(1, 0), p(1, 1)$. Easy. 3 numbers.

Now consider three binary random variables, X, Y, Z. How many numbers do we need to specify to describe the joint distribution $p(x, y, z)$?

Joint distributions

Consider two random variables, X and Y taking on values in $\{0, 1\}$. How many numbers (parameters) do we need to specify to fully describe the joint distribution $p(x, y)$?

We want to know $p(0, 0), p(0, 1), p(1, 0), p(1, 1)$. Easy. 3 numbers.

Now consider three binary random variables, X, Y, Z. How many numbers do we need to specify to describe the joint distribution $p(x, y, z)$?

$2^3 - 1$

Joint distributions

Consider two random variables, X and Y taking on values in $\{0, 1\}$. How many numbers (parameters) do we need to specify to fully describe the joint distribution $p(x, y)$?

We want to know $p(0, 0), p(0, 1), p(1, 0), p(1, 1)$. Easy. 3 numbers.

Now consider three binary random variables, X, Y, Z. How many numbers do we need to specify to describe the joint distribution $p(x, y, z)$?

$2^3 - 1$

What about k binary random variables?

Joint distributions

Consider two random variables, X and Y taking on values in $\{0, 1\}$. How many numbers (parameters) do we need to specify to fully describe the joint distribution $p(x, y)$?

We want to know $p(0, 0), p(0, 1), p(1, 0), p(1, 1)$. Easy. 3 numbers.

Now consider three binary random variables, X, Y, Z. How many numbers do we need to specify to describe the joint distribution $p(x, y, z)$?

$2^3 - 1$

What about k binary random variables?

$2^k - 1$
$\Rightarrow$ This can quickly grow prohibitively big! Try asking a doctor to specify $2^{20} - 1$ probabilities for some related symptoms, test results, biometric measurements, or estimating all those parameters from only $N = 100$ data points. Try storing a table of $\sim 2^{1000000}$ numbers on your computer.

Marginalization

Say you're given some big joint distribution $p(x_1, ..., x_p)$, and you want to know the marginal $p(x_i)$ for some i.
$p(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_p} p(x_1, ..., x_p)$.

$\Rightarrow$ Requires summing over exponentially many values!

Conditioning

Say instead you want to know the conditional $p(x_i|x_j, x_k)$ for some i, j, k.

$p(x_i|x_j, x_k) = \frac{p(x_i, x_j, x_k)}{p(x_j, x_k)}$.

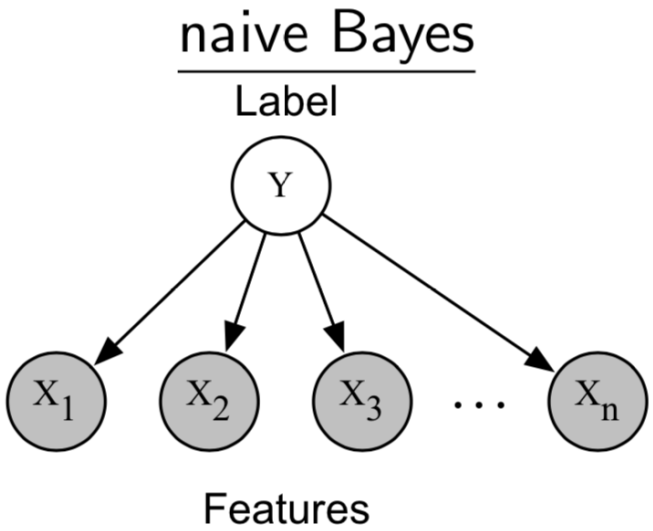
$\Rightarrow$ May require dealing with big joints and/or big sums!

Structuring the problem by independence

Instead of dealing directly with these unwieldly quantities, we can exploit conditional independence assumptions to do probabilistic reasoning more efficiently, compactly, and intuitively.

Conditional independence assumptions are represented with graphical models.

Naive Bayes model



$X_i \perp\!\!\!\perp X_{-i} | Y$

joint:
$p(y, x_1, ..., x_n) = p(y) \prod_{i=1}^{n} p(x_i|y)$

conditional:
$p(Y = 1|x_1, ..., x_n) = \frac{p(Y=1) \prod_{i=1}^{n} p(x_i|Y=1)}{\sum_{y \in \{0,1\}} p(Y=y) \prod_{i=1}^{n} p(x_i|Y=y)}$.

Graphical models

- A statistical model is a set of distributions $\mathcal{P}$.

Graphical models

- A statistical model is a set of distributions $\mathcal{P}$.
- A graphical model is a pair $(\mathcal{G}, \mathcal{P})$ where $\mathcal{G}$ is a graph and $\mathcal{P}$ is a set of distributions which *factorizes* according to $\mathcal{G}$.

Graphical models

- A statistical model is a set of distributions $\mathcal{P}$.

- A graphical model is a pair $(\mathcal{G}, \mathcal{P})$ where $\mathcal{G}$ is a graph and $\mathcal{P}$ is a set of distributions which *factorizes* according to $\mathcal{G}$.

A graph $\mathcal{G} = (V, E)$ where V is a set of vertices and E is a set of edges. Typically, the vertices index some random variables, i.e., let $V = (1, ..., p)$ with corresponding random variables $X_1, ..., X_p$.

The set of edges depends on the type of graph. In a directed graph, edges are **ordered** pairs of vertices $(V_1, V_2) \in E \subseteq V \times V$. In an undirected graph, edges are **unordered** pairs of vertices.

NB: Sometimes we are a bit informal with notation and just let $V \equiv X = (X_1, ..., X_p)$ when this causes no confusion.

Example: Bayesian Network

Figure: A DAG

$\mathcal{G}$ is a directed acyclic graph (DAG), that is a graph with only directed edges ($\rightarrow$) and no *cycles*, i.e., no paths from $X_i \rightarrow \cdots \rightarrow X_i$. Note: A DAG has at most one edge between any two vertices ("simple" graph).

Example: Bayesian Network

Figure: A DAG

$\mathcal{G}$ is a directed acyclic graph (DAG), that is a graph with only directed edges ($\rightarrow$) and no *cycles*, i.e., no paths from $X_i \rightarrow \cdots \rightarrow X_i$. Note: A DAG has at most one edge between any two vertices ("simple" graph).

$\mathcal{P}$ is the set of distributions $\{p(x) : p(x) = \prod_{i=1}^{p} p(x_i | \text{Pa}(X_i, \mathcal{G})\}$ where $X_j \in \text{Pa}(X_i, \mathcal{G})$ if $X_j \rightarrow X_i$ in $\mathcal{G}$.

$\text{Pa}(X_i, \mathcal{G})$ are called the parents of X_i in $\mathcal{G}$.

Example: Markov Random Field

Figure: An UG

$\mathcal{G}$ is a undirected graph (UG), that is a (simple) graph with only undirected edges ($-$).

Example: Markov Random Field

Figure: An UG

$\mathcal{G}$ is a undirected graph (UG), that is a (simple) graph with only undirected edges ($-$).

$\mathcal{P}$ is the set of distributions $\{p(x) : p(x) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\phi_C(x_C)\}$ where $\mathcal{C}$ is the set of cliques in $\mathcal{G}$, ϕ_C is a (non-negative) potential function, and Z is a normalizing constant.

We'll get into the exact definitions later!

Graphical models

- A statistical model is a set of distributions $\mathcal{P}$.

Graphical models

- A statistical model is a set of distributions $\mathcal{P}$.
- A graphical model is a pair $(\mathcal{G}, \mathcal{P})$ where $\mathcal{G}$ is a graph and $\mathcal{P}$ is a set of distributions which *factorizes* according to $\mathcal{G}$.

Graphical models

- A statistical model is a set of distributions $\mathcal{P}$.
- A graphical model is a pair $(\mathcal{G}, \mathcal{P})$ where $\mathcal{G}$ is a graph and $\mathcal{P}$ is a set of distributions which *factorizes* according to $\mathcal{G}$.
- A causal graphical model is a graphical model + *some causal interpretation*.

Graphical models

- A statistical model is a set of distributions $\mathcal{P}$.
- A graphical model is a pair $(\mathcal{G}, \mathcal{P})$ where $\mathcal{G}$ is a graph and $\mathcal{P}$ is a set of distributions which *factorizes* according to $\mathcal{G}$.
- A causal graphical model is a graphical model + *some causal interpretation*.

We'll discuss causal interpretations of graphical models later.

Representation, Inference, and Learning

Representation: what is the right kind of model or model class? What are the advantages or disadvantages of any given representation from the point of view of "accuracy" and/or complexity?

Representation, Inference, and Learning

Representation: what is the right kind of model or model class? What are the advantages or disadvantages of any given representation from the point of view of "accuracy" and/or complexity?

Inference: answering queries using the model.

- marginal inference: $p(x_i)$
- conditional inference: $p(y|x)$
- MAP (maximum a posteriori*) inference: $\arg\max_y p(y|x)$
- others...

Representation, Inference, and Learning

Representation: what is the right kind of model or model class? What are the advantages or disadvantages of any given representation from the point of view of "accuracy" and/or complexity?

Inference: answering queries using the model.

- marginal inference: $p(x_i)$
- conditional inference: $p(y|x)$
- MAP (maximum a posteriori*) inference: $\arg\max_y p(y|x)$
- others...

Learning: estimating graphical structure or model parameters from data

What makes a good representation?

- tractability
- transparency
- utility for achieving certain inference goals
- ability to encode relevant/background knowledge
- "accuracy" (whatever that may mean in context!)

What kind of representations are there?

- BNs, MRFs/CRFs
- Dynamic Bayesian Networks (DBNs), Hidden Markov Models (HMMs), tree-like structures (special cases)
- Latent variable models
- Cyclic graphs, Chain graphs, ADMGs, Ancestral Markov models
- Relational models, context-specific independence models (e.g., multinets), sum-product networks

$\Rightarrow$ graphical models are generally *conditional independence* models, so let's define conditional independence

Probability Review & Conditional Independence

Sample space Ω: The set of all the outcomes of a random experiment.

Set of events (or event space) $\mathcal{S}$: A set whose elements $A \in \mathcal{S}$ (called events) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment).

A probability measure P: A function $P : \mathcal{S} \to \mathbb{R}$ that satisfies the following properties:

- $P(A) \geq 0 \; \forall A \in \mathcal{S}$
- If $A_1, A_2, \ldots$ are disjoint events then $P(\cup_i A_i) = \sum_i P(A_i)$
- $P(\Omega) = 1$

A random variable X is a function $X : \Omega \to \mathbb{R}$ (continuous) or $\{0, 1\}$ (binary) or some other discrete set.

Probability Review & Conditional Independence

Joint probability of events A and B: $P(A, B)$

Probability Review & Conditional Independence

Joint probability of events A and B: $P(A, B)$

Chain rule: $P(A, B) = P(A|B)P(B)$

Probability Review & Conditional Independence

Joint probability of events A and B: $P(A, B)$

Chain rule: $P(A, B) = P(A|B)P(B)$

Conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$

Probability Review & Conditional Independence

Joint probability of events A and B: $P(A, B)$

Chain rule: $P(A, B) = P(A|B)P(B)$

Conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$

Bayes rule: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

Probability Review & Conditional Independence

Joint probability of events A and B: $P(A, B)$

Chain rule: $P(A, B) = P(A|B)P(B)$

Conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$

Bayes rule: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

Density function $p(x)$ s.t. $P(a \leq X \leq b) = \int_a^b p(x)dx$ (continuous case) or $p(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$ (discrete case)
(Note: we are going to always assume the relevant distributions and conditional distributions are well-defined and exist.)

Probability Review & Conditional Independence

Two events are independent iff $P(A_1 \cap A_2) = P(A_1)P(A_2)$ or equivalently $P(A_1|A_2) = P(A_1)$.

Two sets of (discrete) random variables X_1, X_2 are conditionally independent given Z iff
$P(X_1 = x_1, X_2 = x_2|Z = z) = P(X_1 = x_1|Z = z)P(X_2 = x_2|Z = z)$
$\forall x_1, x_2, z$ s.t. $P(Z = z) > 0$.

Probability Review & Conditional Independence

Two events are independent iff $P(A_1 \cap A_2) = P(A_1)P(A_2)$ or equivalently $P(A_1|A_2) = P(A_1)$.

Two sets of (discrete) random variables X_1, X_2 are conditionally independent given Z iff
$P(X_1 = x_1, X_2 = x_2|Z = z) = P(X_1 = x_1|Z = z)P(X_2 = x_2|Z = z)$
$\forall x_1, x_2, z$ s.t. $P(Z = z) > 0$.

For densities conditional independence corresponds to:
$p(x_1, x_2|z) = p(x_1|z)p(x_2|z)$
or equivalently
$p(x_1|x_2, z) = p(x_1|z)$.
(almost surely wrt P)

Notation: $X_1 \perp\!\!\!\perp X_2|Z$.

Probability Review & Conditional Independence

Two events are independent iff $P(A_1 \cap A_2) = P(A_1)P(A_2)$ or equivalently $P(A_1|A_2) = P(A_1)$.

Two sets of (discrete) random variables X_1, X_2 are conditionally independent given Z iff
$P(X_1 = x_1, X_2 = x_2|Z = z) = P(X_1 = x_1|Z = z)P(X_2 = x_2|Z = z)$
$\forall x_1, x_2, z$ s.t. $P(Z = z) > 0$.

For densities conditional independence corresponds to:
$p(x_1, x_2|z) = p(x_1|z)p(x_2|z)$
or equivalently
$p(x_1|x_2, z) = p(x_1|z)$.
(almost surely wrt P)

Notation: $X_1 \perp\!\!\!\perp X_2|Z$.

If Z is empty we say they are *marginally* independent and write $X_1 \perp\!\!\!\perp X_2$.

Probability Review & Conditional Independence

What does conditional independence capture? "Irrelevance of information"

Properties:

- $X_1 \perp\!\!\!\perp X_2 | Z \implies X_2 \perp\!\!\!\perp X_1 | Z$ (symmetry)
- $X_1 \perp\!\!\!\perp (X_2, X_3) | Z \implies X_1 \perp\!\!\!\perp X_2 | Z$ (decomposition)
- $X_1 \perp\!\!\!\perp (X_2, X_3) | Z \implies X_1 \perp\!\!\!\perp X_2 | (Z, X_3)$ (weak union)
- $X_1 \perp\!\!\!\perp X_2 | (Z, X_3)$ & $X_1 \perp\!\!\!\perp X_3 | Z \implies X_1 \perp\!\!\!\perp (X_2, X_3) | Z$ (contraction)

and for *positive* distributions, also:

- $X_1 \perp\!\!\!\perp X_2 | (Z, X_3)$ & $X_1 \perp\!\!\!\perp X_3 | (Z, X_2) \implies X_1 \perp\!\!\!\perp (X_2, X_3) | Z$ (intersection)

(together these are called the graphoid axioms; the first four are the semi-graphoid axioms)

Probability Review & Conditional Independence

- Symmetry: Knowing C, if learning A is irrelevant for predicting B, then learning B is irrelevant for predicting A.
- Decomposition: Knowing C, if learning A is irrelevant for predicting D, then learning A is irrelevant for predicting any part B of D.
- Weak union: If, knowing C, learning A is irrelevant for predicting D, then knowing also any part B of D, learning A is still irrelevant for predicting the rest of D.
- Contraction: If, knowing C, learning A is irrelevant for predicting D and, knowing both C and D, learning A is irrelevant for predicting B, then knowing C, learning A is irrelevant for predicting both B and D.

Probability Review & Conditional Independence

Correlation is *linear* (or "second order") dependence because it just about second-moments, and it is equivalent to dependence in linear models. (Conditional) independence does not equal (conditional) non-correlation, except in special cases.

$$cov(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]$$

$$\rho_{12} \equiv corr(X_1, X_2) = \frac{cov(X_1, X_2)}{\sigma_1 \sigma_2}$$

Probability Review & Conditional Independence

Correlation is *linear* (or "second order") dependence because it just about second-moments, and it is equivalent to dependence in linear models. (Conditional) independence does not equal (conditional) non-correlation, except in special cases.

$$cov(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]$$

$$\rho_{12} \equiv corr(X_1, X_2) = \frac{cov(X_1, X_2)}{\sigma_1 \sigma_2}$$

$X_1 \perp\!\!\!\perp X_2 \implies \rho_{12} = 0$ but
$\rho_{12} = 0$ DOES NOT IMPLY $X_1 \perp\!\!\!\perp X_2$ in general.
Same is true for *conditional* versions of these.