CO902
Probabilistic and statistical inference

Lecture 8

Tom Nichols
Department of Statistics &
Warwick Manufacturing Group

t.e.nichols@warwick.ac.uk

Outline of course

- A. Basics: Probability, random variables (RVs), common distributions, introduction to statistical inference
- B. Supervised learning: Regression, classification, including high-dimensional issues and Bayesian approaches
- C. Unsupervised learning: Dimensionality reduction, clustering and mixture models
- **D. Networks: Probabilistic graphical models, learning in graphical models, inferring network structure**

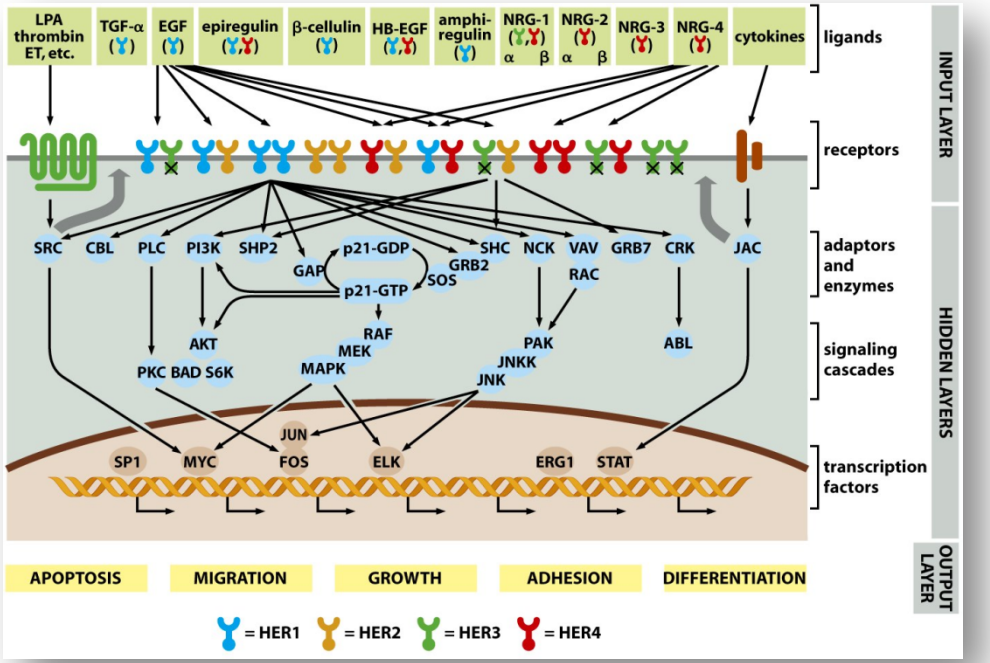
Today

- Probabilistic models on graphs: “Probabilistic graphical models”
  - Models on graphs
  - Definition of a probabilistic graphical model
  - Conditional independence and d-separation
  - Structural inference

Some examples…

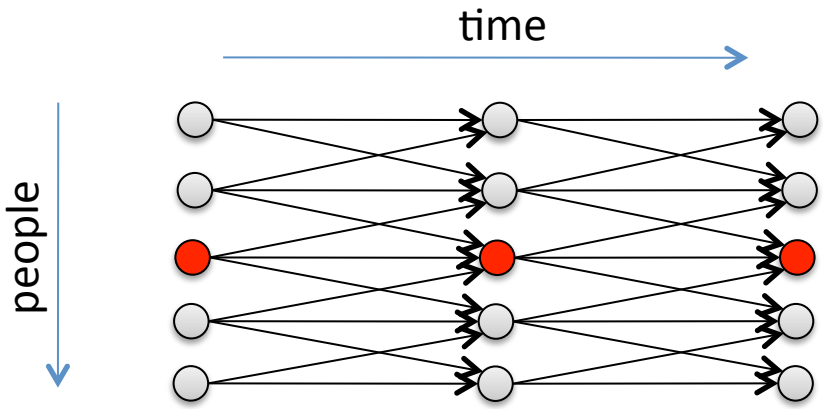
- **Biological networks**
- In recent years, movement towards thinking about molecular biology not just in terms of single gene/proteins, or pair-wise relationships between them, but in terms of **multiple interacting components**
- **Networks of molecular players** have key role in every aspect of molecular biology, from development and growth, to day-to-day functioning, to diseases like cancer
- Example: Epidermal growth factor receptor (EGFR) system

A biological network



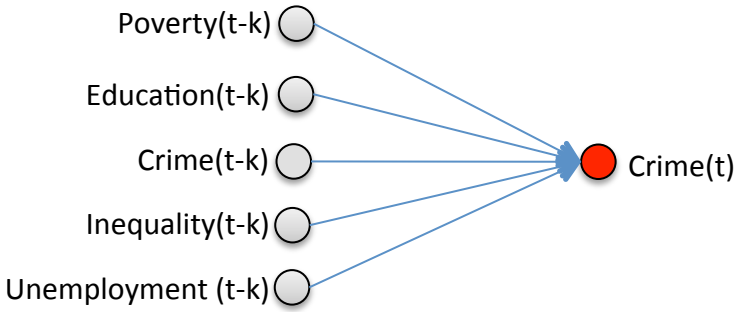
- EGFR system
  - Key role in breast and lung cancers

Social networks in health/economics/business



- Maybe interested in how health status is influenced, through time, by social network effects
- Or how chance of being married depends on marital status of people you know

Influences on crime rates



- Or how various factors (jointly) influence (e.g.) crime rates
- Or, in a richer model, how they influence *each other*

- Similarly, public health factors, economic variables etc.

RVs and graphs

- In all three cases, we're starting to use *graphs* to reason, in an intuitive fashion, about *systems of random variables*
- *Graph G = { V(G), E(G) }* is an object consisting of
  - A set *V* of vertices, and
  - Set *E* of edges, these are pairs of vertices
- **Probabilistic graphical models** are about formalizing the intuitive link between graphs and probability models
- Idea is:
  - Graph tells you how variables are related
  - Conditional probability distributions + parameters then give a "full model"

Directed graphical model

- A **directed graphical model** or **Bayesian network** consists of
  - (Directed acyclic) graph *G*
  - Parameters **θ**
- Graph *G* implies:

$$p(X_1 \dots X_p \mid G, \boldsymbol{\theta}) = \prod_{i=1}^{p} p(X_i \mid \mathbf{Pa}_G(X_i), \boldsymbol{\theta})$$

- This equation is the *definition* of the model
- Effectively a structuring of how the variables depend on each other

Directed graphical model

$$p(X_1 \ldots X_p \mid G, \boldsymbol{\theta}) = \prod_{i=1}^{p} p(X_i \mid \mathbf{Pa}_G(X_i), \boldsymbol{\theta})$$

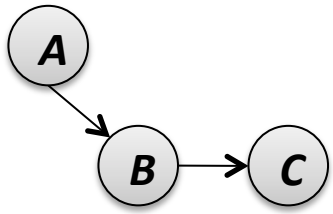
"Global" joint *"Local" terms*

- What this is doing is giving a compact description of the *global* joint as product of *local* terms, without being as strong a total independence
- The full set of parameters then specify those local terms

Example: MC

$$p(X_1 \dots X_p \mid G, \boldsymbol{\theta}) \;=\; \prod_{i=1}^{p} p(X_i \mid \mathbf{Pa}_G(X_i), \boldsymbol{\theta})$$

- Example:



"Global" joint

$$P(A, B, C)$$
$$= \; P(A) \times P(B \mid A) \times P(C \mid B)$$

Product of "local" terms

- This is a Markov chain: e.g. binary case
- **Q: is C independent of A? Is it independent *given* B?**

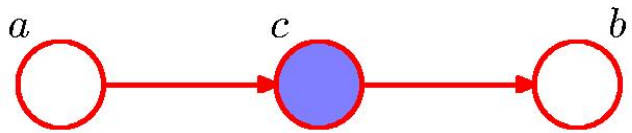
Conditional independence

- Graphs like this imply conditional independence statements between variables
- What *is* conditional independence?
- Definition: A is conditionally independent of B given C, iff

$$P(A,B \mid C) = P(A \mid C)\, P(B \mid C)$$

- Intuitively: once you know C, knowing B doesn't help you guess A.
- E.g.:
  - A = individual's genetic sequence (coding for proteins),
    C = status of certain disease-causing proteins,
    B = disease caused by those proteins
- Consider 3 graph motifs…

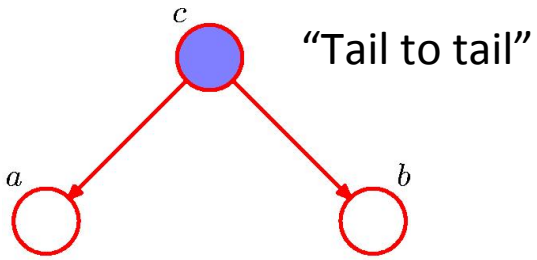
Conditional independence from graphs: case I



"Head to tail"

- Example we've seen already
- Note that a,b are not *marginally* independent!

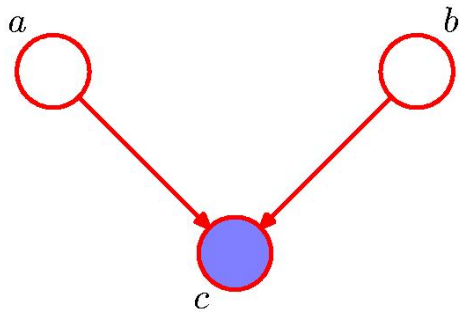
Conditional independence from graphs: case II



"Tail to tail"

- E.g. one gene (plus "nurture" factors) may lead to *both* cancer and diabetes
  - Cancer and diabetes are then certainly not marginally independent
  - But *given the presence of the gene*, they are

Conditional independence from graphs: case III



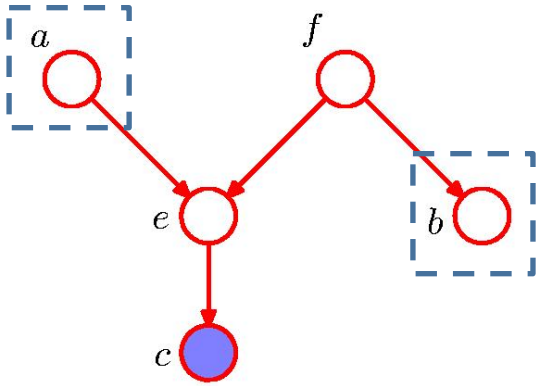
"Head to head"

- Are they marginally independent?
  - a & b marginally independent
  - But *given* c, a & b conditionally *dependent*!
- Example of "Berkson's paradox"

d-separation

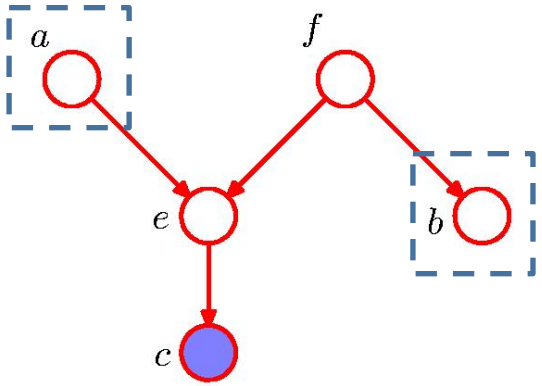
- Generalizing from these three cases gets us to a general rule for "reading off" c.i. statements from directed graphical models
- Algorithm for determining d-separation
  - A, B, C are non-intersecting sets of nodes (need not cover whole graph). Wish to ascertain whether A is c.i. of B given C
  - Consider all possible paths from (any node in) A to (any node in) B
  - Any such path said to be *blocked* if it includes at least one node s.t.:
    - Edges in path meet in C *and* it is a "head-to-tail" (or "tail-to-head") or "tail-to-tail" connection
    - Arrows meet "head-to-head" at node, and *neither the node, nor any of its descendants are in C*
  - If *all* paths are blocked, A is said to be d-separated from B by C, and the property holds… i.e. A is c.i. of B given C

d-separation: example I



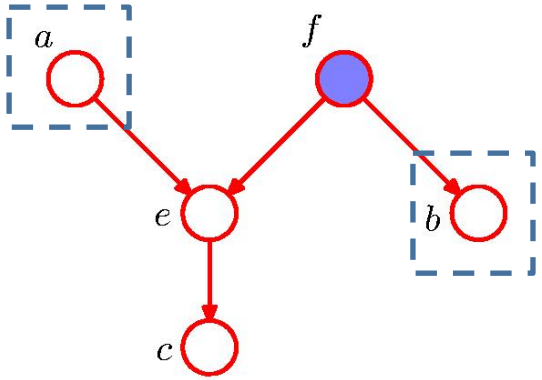
- **Q: is a is c.i. of b given c ?**

d-separation: example I



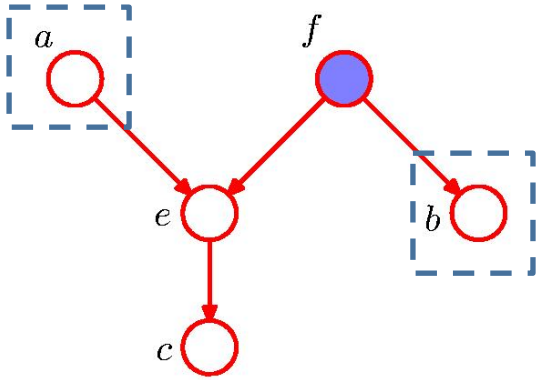
- **<u>Q: is a is c.i. of b given c ?</u>**
- No, “head-to-head” at e, and c is descendant of e

d-separation: example II



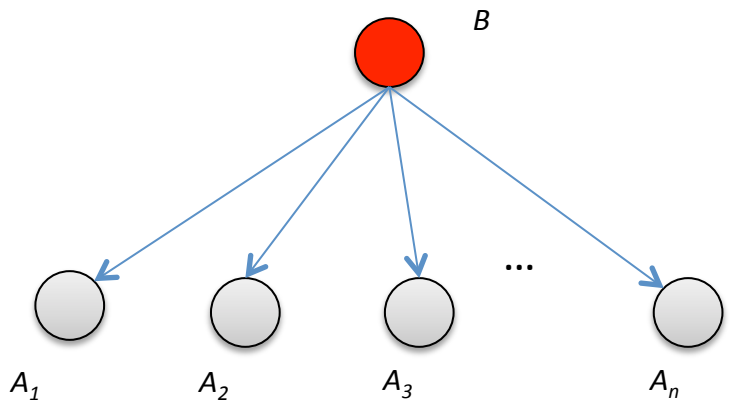
- **Q: Is a c.i. of b given f ?**

d-separation: example II



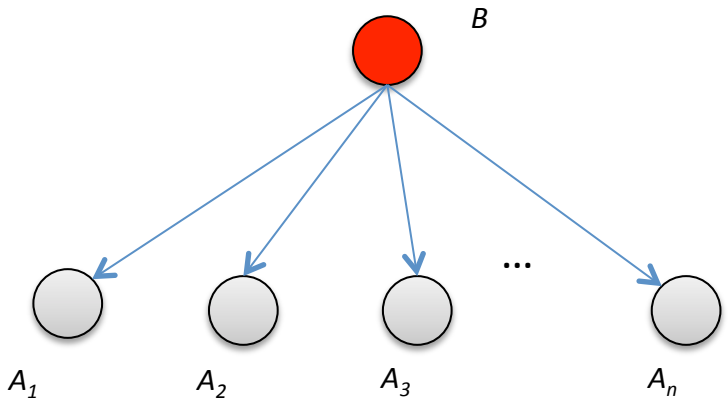
- **Q: Is a c.i. of b given f ?**
- Yes: meets "tail-to-tail" at *f*
  - Have a "head-to-head" (at *e*), but isn't in nor doesn't project to *f*

d-separation: example III



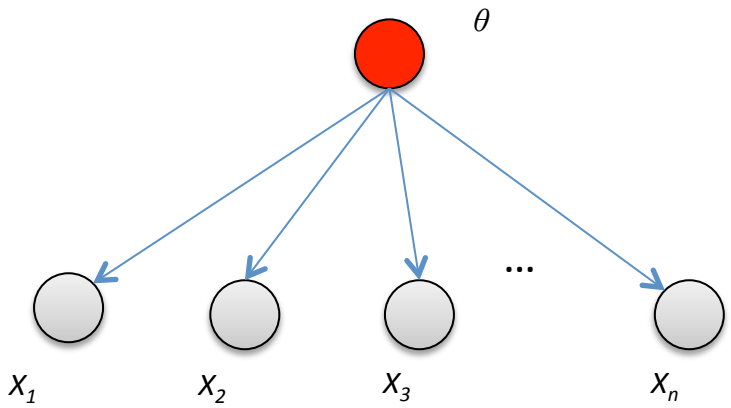
- **Q: is A_i c.i. of A_j given B?**

d-separation: example III



- **Q: is A_i c.i. of A_j given B?**
- Tail-to-tail in B, and no "head-to-head"s to consider so
  Yes!

d-separation: example III

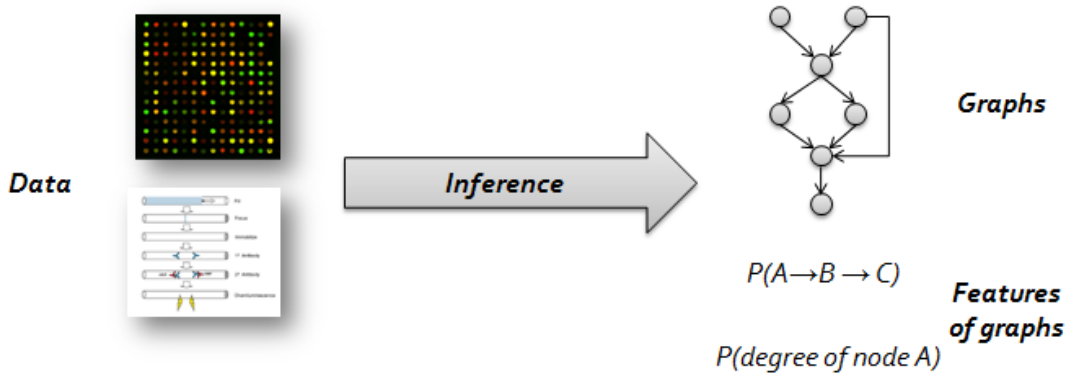


- This is just the Bernoulli (or other iid) model!
- Here, clearly the previous tosses tell you *something* about the next one, they're not independent in *that* sense - but the idea is that *once you know the parameter*, you can't get a better guess by knowing what the last toss was
- i.i.d. can be thought of as meaning c.i. given parameters: "all independence is conditional"
- What does classification look like?

Conditional independence in models

- *So* even iid type independence is conditional on *something*
- Almost every model we've seen has some sort of conditional independence
- E.g.:
  - Bernoulli model
  - Classifier
  - Markov Chain – 1st order, 2nd order

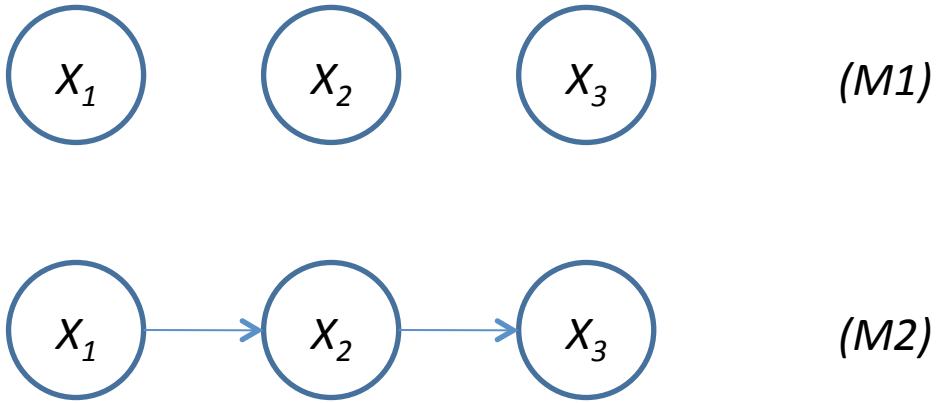
Structure learning



- Often the case that the graph G itself is of interest
  - Want to say something about *how* variables are related
- Task often referred to as *structure learning* or *network inference*
- Examples:
  - Interplay between public health factors and outcomes
  - Protein signalling network
  - Social and demographic factors in crime

Model selection

- Consider comparing M1 vs. M2 for some data:



- Model selection question is, given some data ***X***, which of M1 and M2 are better?
- Let's write down posterior probability of each model….

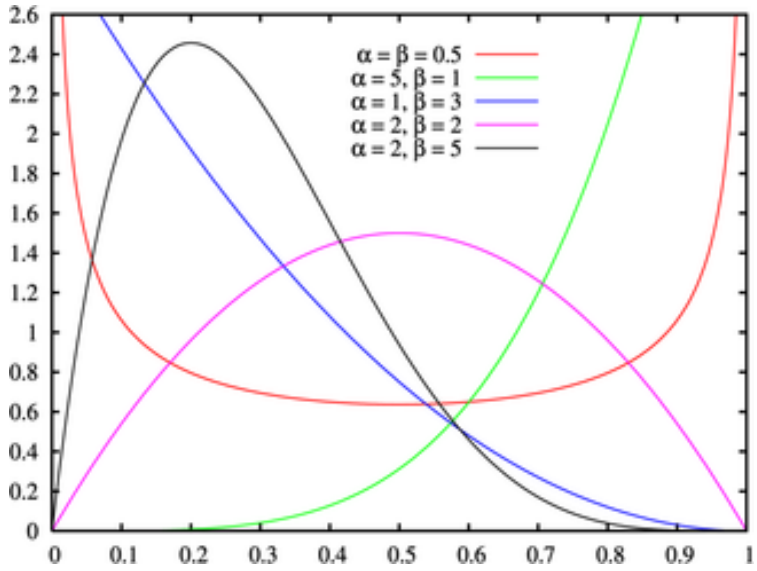
Model posterior

- Write down P(M1 | **X**) and P(M2 | **X**)
- Consider *p(***X** | *M2)*
  - This is still not a likelihood function, there are no parameters
  - We can introduce these and integrate them out to get a *marginal likelihood*

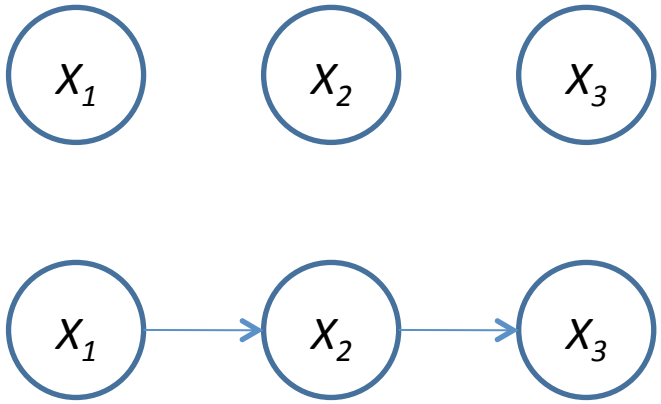
- What does this look like for iid binary data vectors, i.e. everything Bernoulli?
- Recall Bayesian analysis for Bernoulli…

Beta pdf

$$
\begin{aligned}
p(x \mid \alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\
x &\in [0,1] \\
\alpha, \beta &> 0 \\
\Gamma(x) &= \int_0^\infty t^{x-1} e^{-t}\, \mathrm{d}t \\
x &> 0
\end{aligned}
$$



Structure learning



- These just examples of conditional independence graphs
  - We can do the same for *any* graph *G*
- OK, a graph *G* does *not* specify a full model
  - Only gives only conditional independence structure of variables
  - But, in principle, can integrate out parameters of local conditionals, and get a marginal likelihood for the graph

Structure learning

- **Posterior distribution over graphs**:

$$P(G \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid G)P(G)}{\sum_{G \in \mathcal{G}} p(\mathbf{X} \mid G)P(G)}$$

- Terms:
  - $P(G)$ *prior distribution* on graphs
  - $p(\mathbf{X} \mid G)$ *marginal likelihood*
  - Denominator is sum over graphs under consideration

Marginal likelihood

- Graph *G* does *not* specify a full model:
  - Only conditional independence structure of variables
  - But not parameters of local conditionals
- To get **marginal likelihood** we must marginalize over parameters Θ :

$$p(\mathbf{X} \mid G) = \int p(\mathbf{X} \mid G, \mathbf{\Theta}) p(\mathbf{\Theta} \mid G) \, \mathrm{d}\mathbf{\Theta}$$

- In general, integral won't be easy to evaluate, but closed-form expressions exist for certain local conditionals and conjugate priors

Selecting a graph

- For closed form marginal likelihood and given a graph prior *P(G)*, we can quite easily evaluate posterior up to proportionality:

$$P(G \mid \mathbf{X}) \quad \propto \quad p(\mathbf{X} \mid G) \times P(G)$$

- This provides a probabilistically correct "score" for graph *G*, taking into account
  - fit to data
  - model complexity
  - as well as prior knowledge encoded in *P(G)*

Network inference

- Data-driven network inference is a topic very much at the “research frontier”
- This is just the “tip of the iceberg”. Many open areas include:
  - Dealing with large spaces of graphs, including MCMC etc.
  - Priors on graphs
  - Desktop compute power (& GPU’s!) means this sorts of analyses are now really practical and not just day-dreaming

Conclusions

- Graphical Models
  - Just a covariance model for multivariate data
  - Visual, concise way to represent conditional independences
  - Limited to Directed Acyclic Graphs
- Learning Graphical Structure
  - Integrate out parameters, compare structures
  - Bleeding edge area
    - Space of models just so large
      - 8 nodes – 1+ billion possible DAGs! (OEIS A003024)