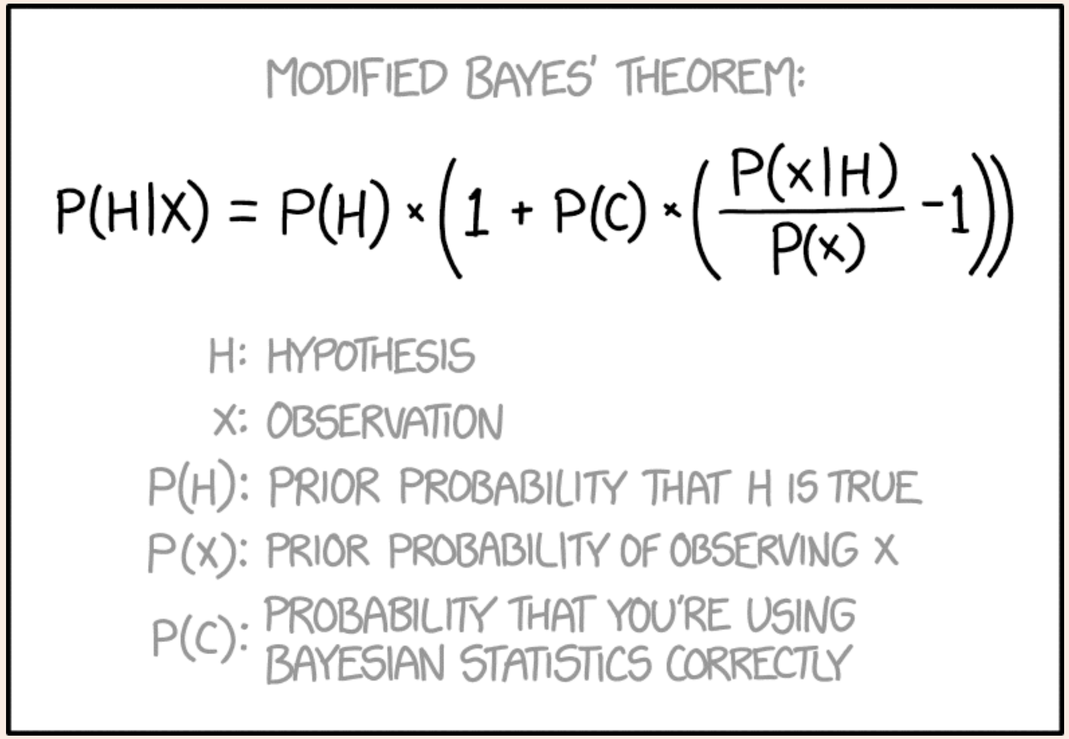
PROBABILISTIC MODELS

A. Bayesian Inference

B. Graphical Models



https://xkcd.com/2059/

A. BAYESIAN INFERENCE

Probably not Thomas Bayes

Certainly not Thomas Bayes

Previously

- Week 2: Maximum Likelihood, MAP estimates, regularisation terms
- Week 4: Maximum margin classifiers, generalization bounds

Revisit the height estimation

Random Variable

Not an estimator yet

$$\max_{\theta} p_{\theta}(D)$$

$$\max_{\theta} p(\theta|D)$$

$$p(\theta|D)$$

MAXIMUM LIKELIHOOD

MAXIMUM A POSTERIORI

POSTERIOR

Bayesian Inference

$$p(\theta|D)$$

Treat the parameters of the model as random variables
and update their distributions when observing data

Bayes Rule

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

Likelihood Prior

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

(Polynomial) Regression: Likelihood

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2 \mathbb{I})$$

Let:

- $\mathbf{X}$ be the N x M feature matrix $\quad \mathbf{x}^\top = [1, x, x^2, x^3, x^4, x^5, x^6, x^7]$
- $\mathbf{w}$ be the M x 1 weight vector
- Assume we know the measurement noise

https://jkrijthe.shinyapps.io/bayesian-poly/

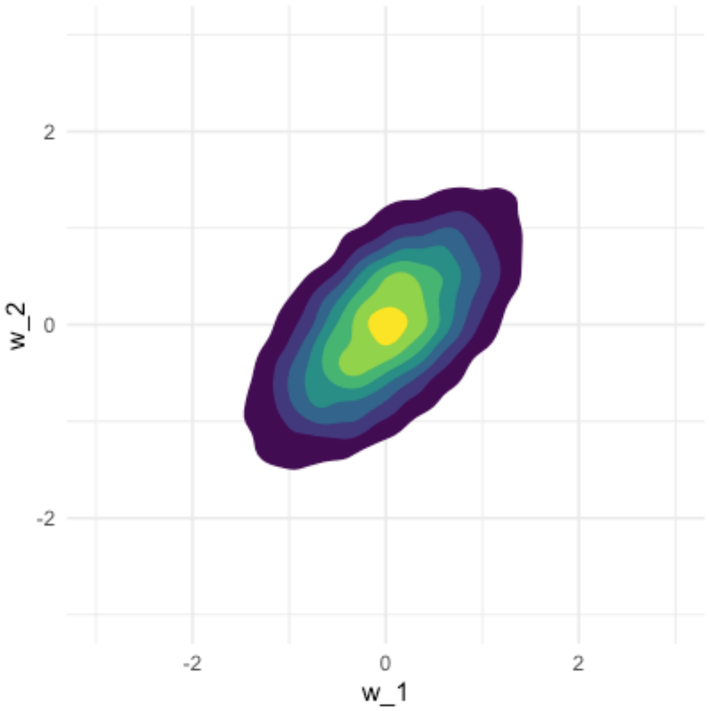
Prior

$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbb{I})$$

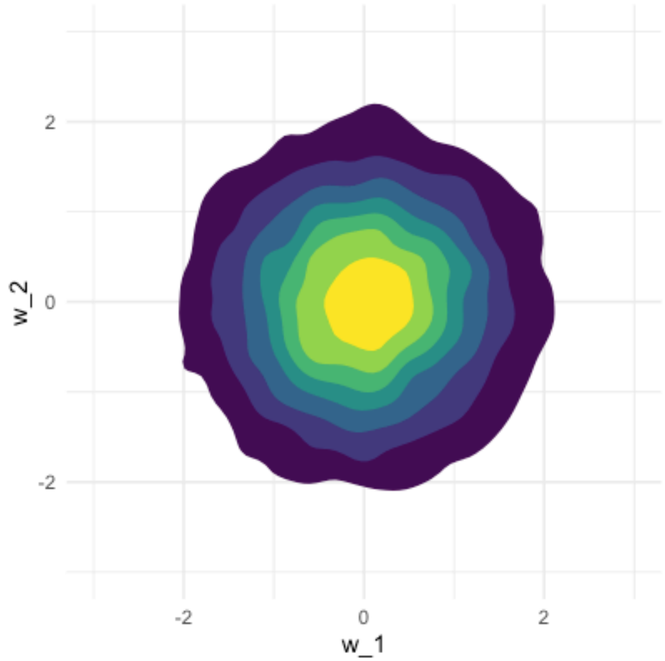
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I})$$

Question: What does the prior prob. look like?

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I})$$



Plot A



Plot B

Prior

$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbb{I})$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I})$$

Don't know anything? Perhaps consider the prior predictive distribution:

$$p(y^{\text{new}}) = \int p(y^{\text{new}}|\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

Getting to the Posterior

$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbb{I})$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I})$$

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})} = \frac{\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbb{I})\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I})}{Z}$$

$$\log p(\mathbf{w}|\mathbf{y}) = C_1 - \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + C_2 - \frac{1}{\alpha}\mathbf{w}^\top\mathbf{w} - Z$$

$$\frac{2}{\sigma^2}\mathbf{y}^\top\mathbf{X}\mathbf{w} - \frac{1}{\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \frac{1}{\alpha}\mathbf{w}^\top\mathbf{w} + C_1 + C_2 - Z - \frac{1}{\sigma^2}\mathbf{y}^\top\mathbf{y}$$

Perhaps posterior is normal? Form: $-(\mathbf{w} - \mathbf{m})^\top\mathbf{S}(\mathbf{w} - \mathbf{m})$

$$\mathbf{S} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I}$$

$$\mathbf{m} = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1}\mathbf{X}^\top\mathbf{y}, (\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1})$$

Posterior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbb{I}) \xrightarrow{\text{Observe Data}} p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}^{-1})$$

$$\mathbf{m} = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

$$\mathbf{S} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I}$$

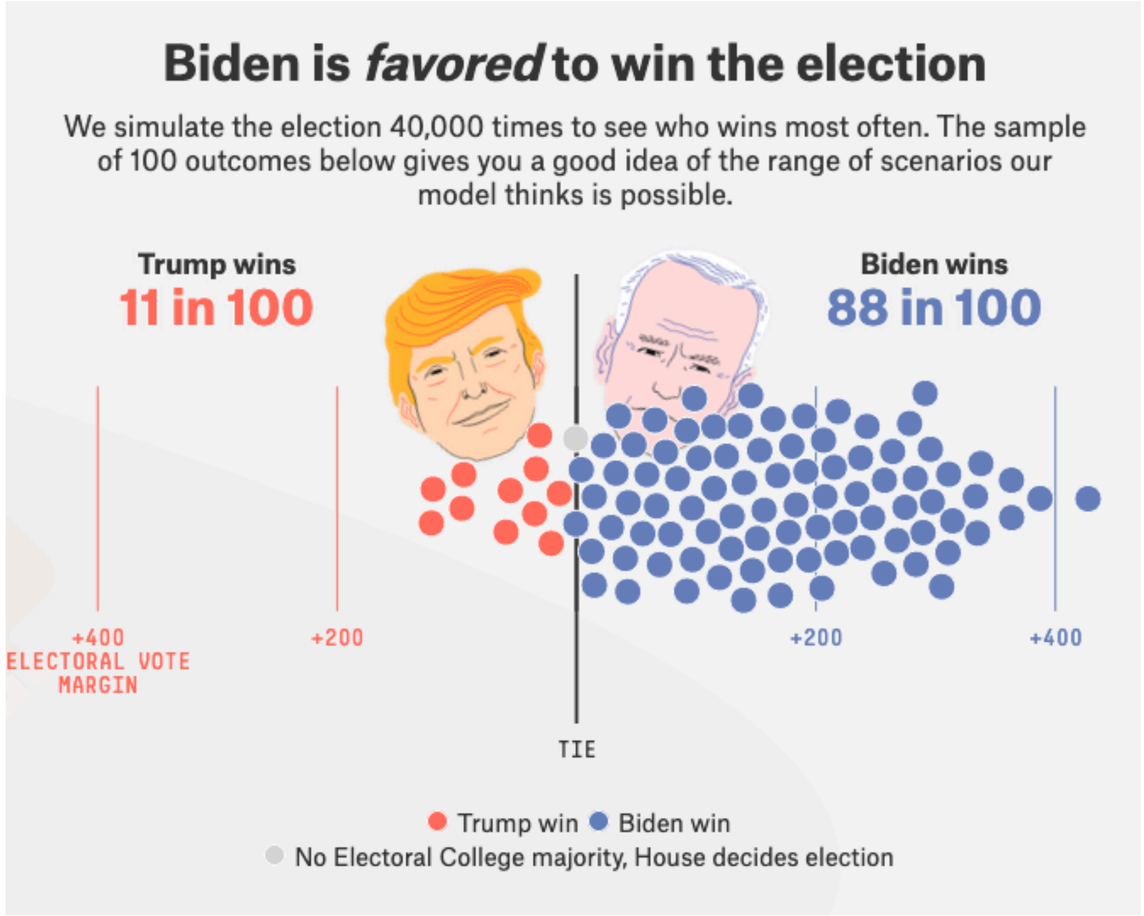
- What is the MAP estimate?
- Credible intervals vs. Confidence intervals

Posterior Predictive Distribution

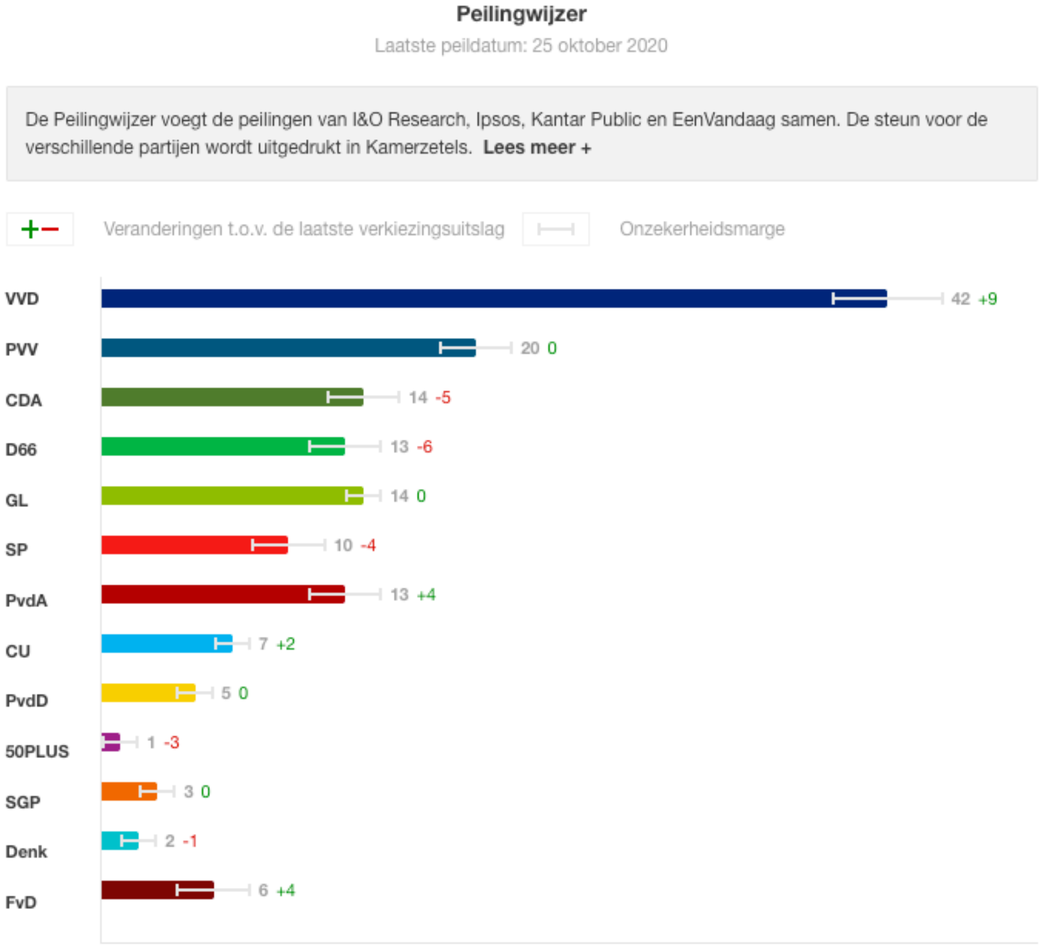
$$p(y^{\text{new}}|\mathbf{y}) = \int p(y^{\text{new}}|\mathbf{w})p(\mathbf{w}|\mathbf{y})d\mathbf{w}$$
$$= \mathcal{N}(y_{\text{new}}|\mathbf{x}_{\text{new}}^\top \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1}\mathbf{X}^\top\mathbf{y}, \mathbf{x}_{\text{new}}^\top(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbb{I})^{-1}\mathbf{x}_{\text{new}} + \sigma^2)$$

- *Important use-case*: model checking: posterior predictive checking
- What about decisions?
  - Use decision theory: minimise the expected loss.
  - Why are all these estimates so similar here? (symmetry)

Application: Election Forecasting



fivethirtyeight.com



peilingwijzer.nl

Comparison ML, MAP, Full Bayes

$$\max_{\theta} p_{\theta}(D)$$

MAXIMUM LIKELIHOOD

$$\max_{\theta} p(\theta|D)$$

MAXIMUM A POSTERIORI

$$p(\theta|D)$$

POSTERIOR

Arbitrary distributions/losses

Unfortunately, not every distribution is conjugate and symmetric…

So the minimal expected loss will not always coincide with the MAP solution

See examples in exercise session

Why Bayes

- "Rational belief is governed by the laws of probability"?
  - See discussion surrounding "Cox's Theorem"
- Elegant way to incorporate new information, sequential updating
- "Automatic" complexity control
- "Automatic" uncertainty estimates

Problems?

- Choosing the prior? (Choosing the model?)
- At what level should the prior come in?
- Do you even want to have Bayesian probabilities?
- Efficient inference

How to do it?

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Conjugacy (mainly considered here): closed form solutions to calculate posterior
- Sampling (Gibbs, Metropolis-Hastings, Hamiltonian Monte Carlo, Particle Filters): generate samples from the posterior
- (Variational) Approximations: approximate the posterior using a simpler distribution

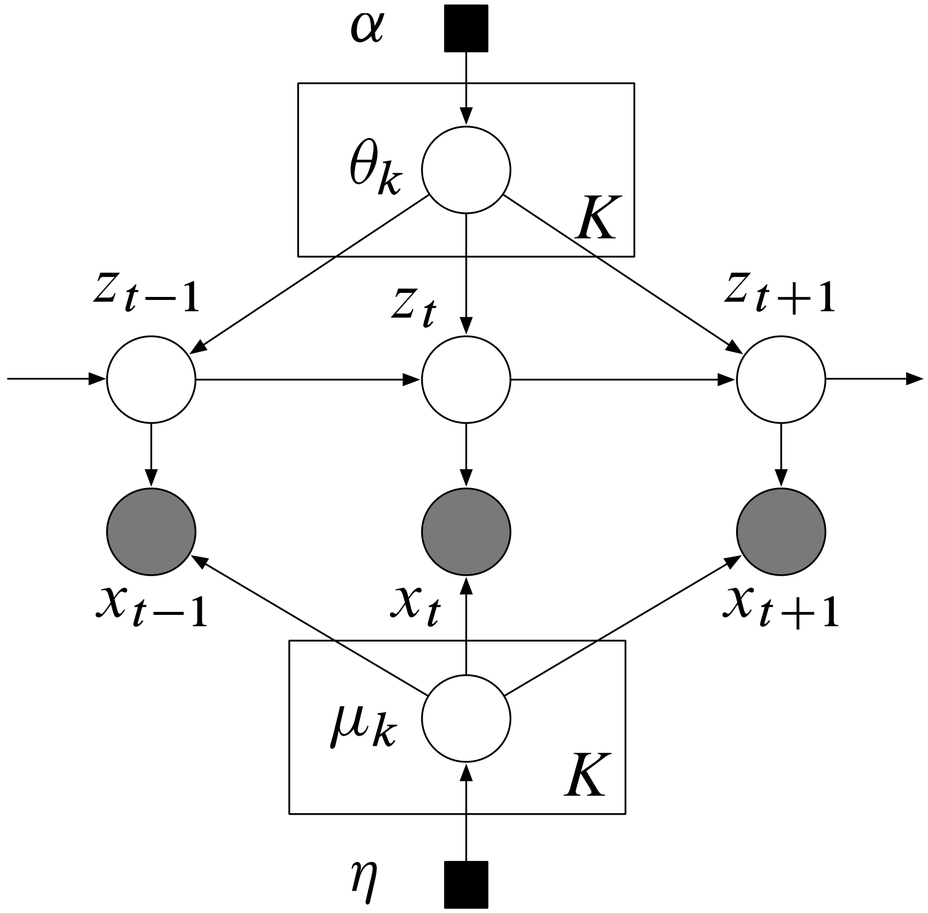
Interpretations of Bayesian Inference

- Subjective, objective, default, hypothetico-deductive, empirical
- Non-Bayesian: way to construct models with good (frequentist) properties

Conclusion Part A

- Bayesian inference treats unknown parameters as random variables. We update these variables based on the observed data, using the likelihood model
- By taking into account different sources of informations, and incorporating all uncertainties in the inference, we may be able to build good predictions/estimates

B. PROBABILISTIC GRAPHICAL MODELS



$$\theta_k \sim \text{Dirichlet}_K(\alpha)$$

$$z_t \sim \text{Discrete}(\theta_{z_{t-1}})$$

$$x_t \sim p(\cdot \mid \mu_{z_t})$$

$$\mu_k \sim p(\cdot \mid \eta)$$

D. Blei, "Build, Compute, Critique, Repeat:Data Analysis with Latent Variable Models"

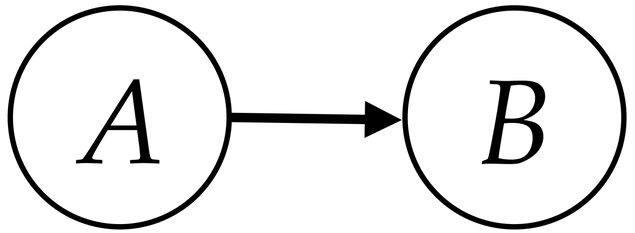
Modeling the Joint Distribution

$$P(X_1, X_2, ..., X_M)$$

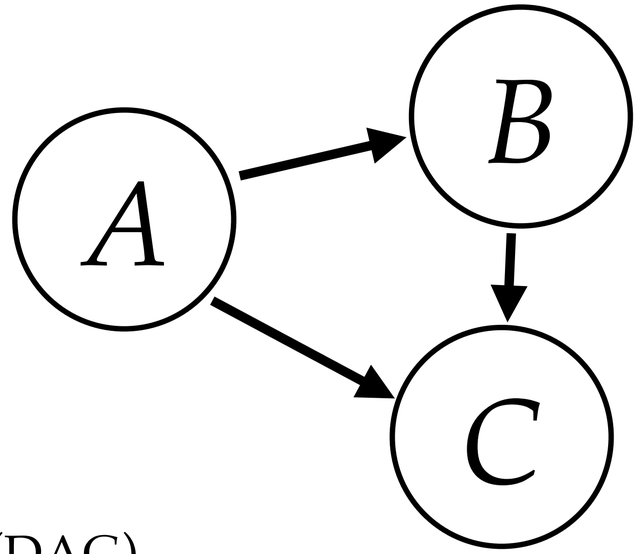
- Discrete: D^M - 1 values?
- Can we decompose it somehow to
  - represent it more efficiently?
  - reason about relations between variables?
  - do efficient inference?
  - incorporate our knowledge of *P*?
  - construct *P* from simpler components?

Factorising a PDF

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

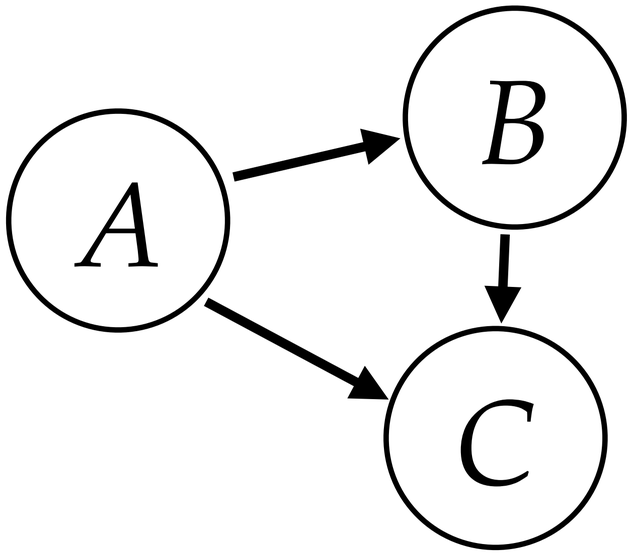


$$P(A,B,C) = P(A)\,P(B \mid A)\,P(C \mid A,B)$$



- Graphical representation: Directed Acyclic Graph (DAG)
- Graph captures the independence structure of a whole family of distributions
- Can we simplify the model by assumes some links are not there?

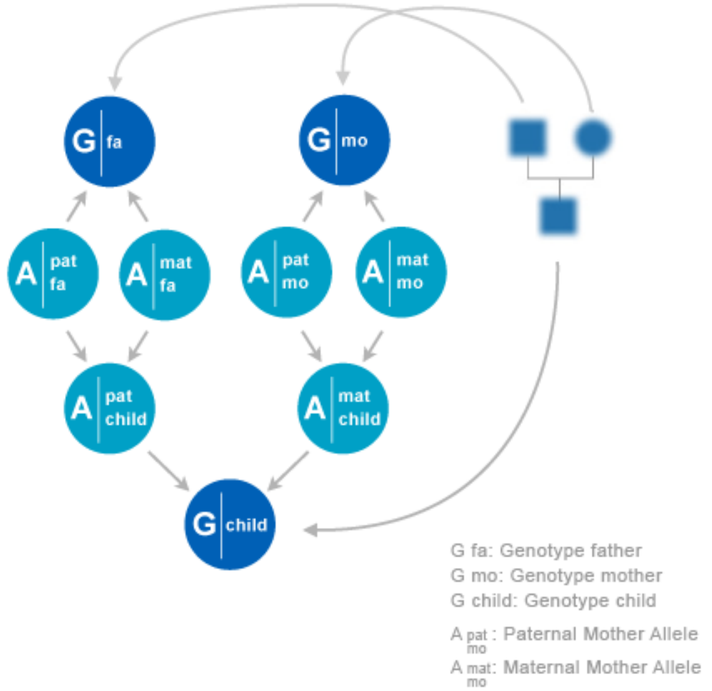
Directed Graphs: Bayesian Networks



- Undirected vs. Directed
- Bayesian because they use Bayes rule, not (necessarily) because parameters are represented by random variables.

Application: Bonaparte DVI

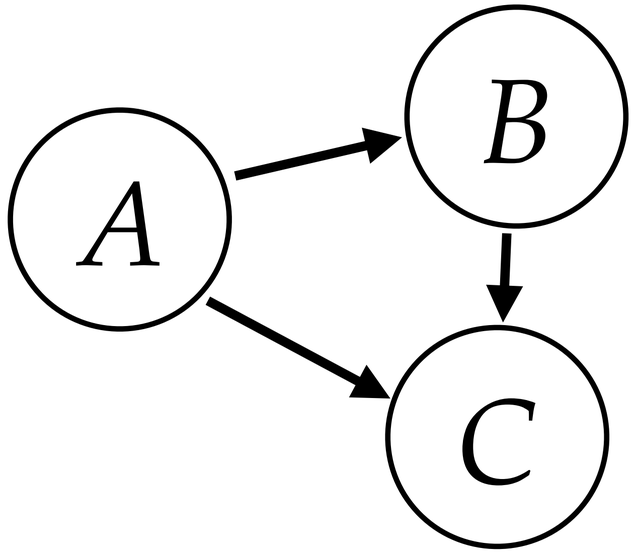
Afriqiyah Airways crash at Tripoli Airport (2010)



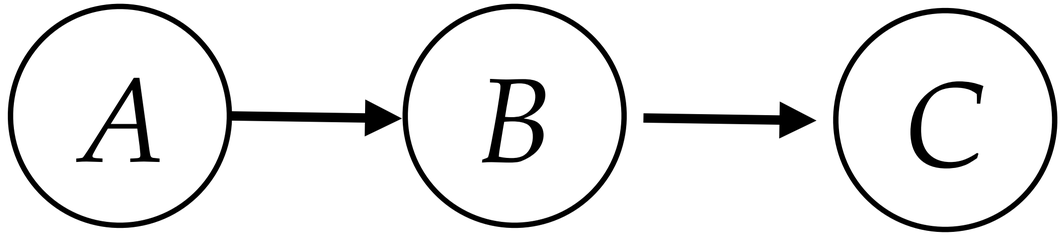
See: https://www.bonaparte-dvi.com/

Adding assumption: removing links

Any P(A,B,C) adheres to this:



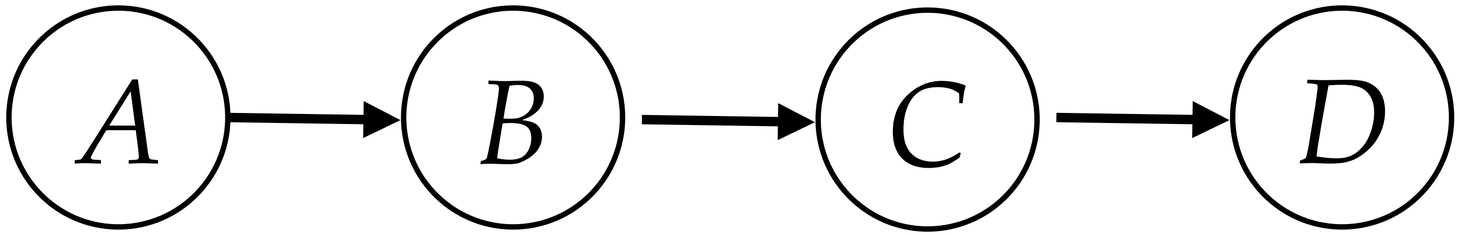
Not every P(A,B,C) adheres to this:



Example: Efficient Representation

M binary variables

In general, we need $2^M - 1$ numbers to represent this distribution



When we can factorize as a chain: 1 + 2(M-1)

(Conditional) Independence

Independence

$$p(A \mid B) = p(A) \quad \text{or} \quad p(A,B) = p(A)\, p(B)$$

Observing B will not give me additional information about A

$$A \perp B$$

Conditional Independence

$$p(A \mid B, C) = p(A \mid C) \quad \text{or} \quad p(A, B \mid C) = p(A \mid C)\, p(B \mid C)$$

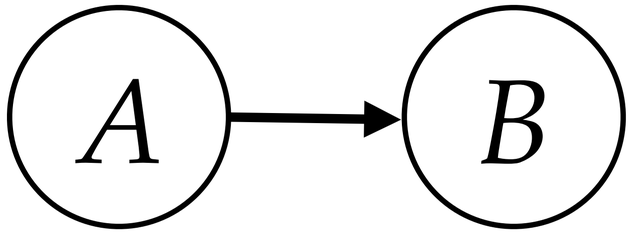
Given that I have observed C, observing B will not give me additional information about A

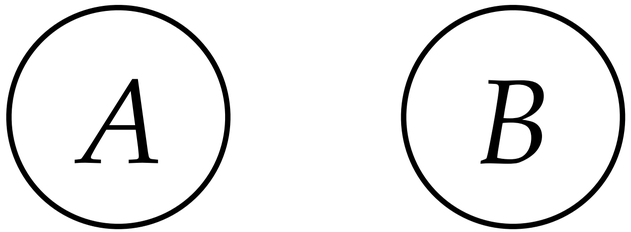
$$A \perp B \mid C$$

Checking Independence

- We can check for independence by starting with P(A,B,…,Z) and using algebraic manipulations to prove the independence holds.
- Cumbersome… can we find a way to check independences using the graph?

Independence: 2 Variables





$p(B \mid A)\ p(A)$

(Or vice versa)

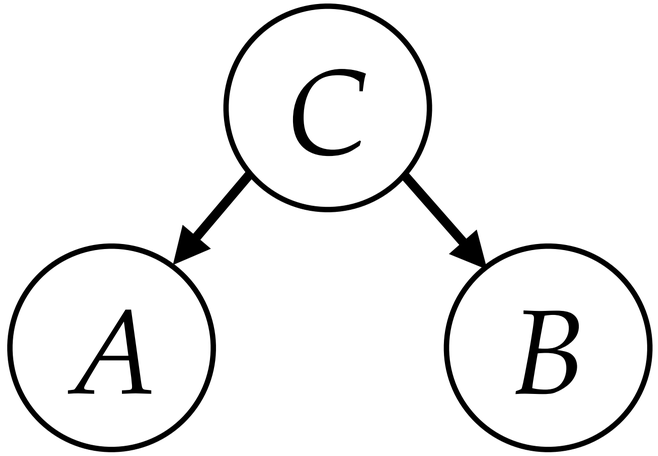
$A \not\perp B$

$p(B)\ p(A)$

$A \perp B$

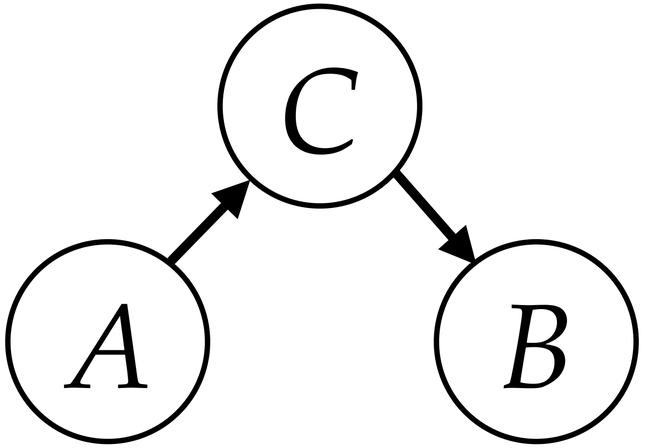
Only two variables, so no possible conditional independences

(Conditional) Independence: 3 Variables



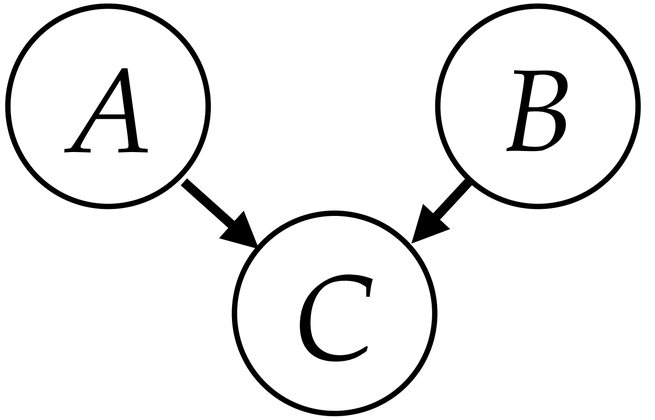
"Tail to tail"

p(A | C) p(B | C) p(C)



Chain

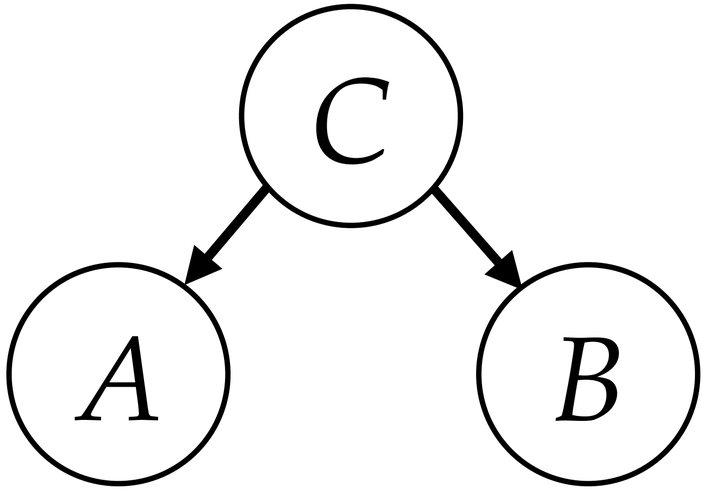
p(A) p(C | A) p(B | C)



Collider
(Explaining away)

p(A) p(B) p(C | A,B)

"Tail-to-Tail"/"Common cause"



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

$$p(A,B) = \sum_C p(A|C)p(B|C)p(C) \neq p(A)p(B)$$

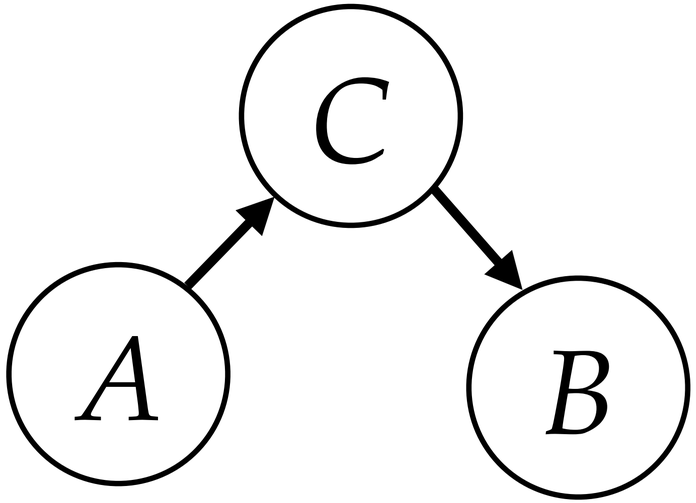
$$A \not\perp B$$

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

$$A \perp B \mid C$$

Interpretation: if we don't observe C, observing A gives information on C, which gives information about C. If we do observe C, all the information about C is already present, and observing A adds nothing to our knowledge of B

Chain



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

$$p(A,B) = p(A)\sum_{C} p(B|C)p(C|A) = p(A)p(B|A)$$

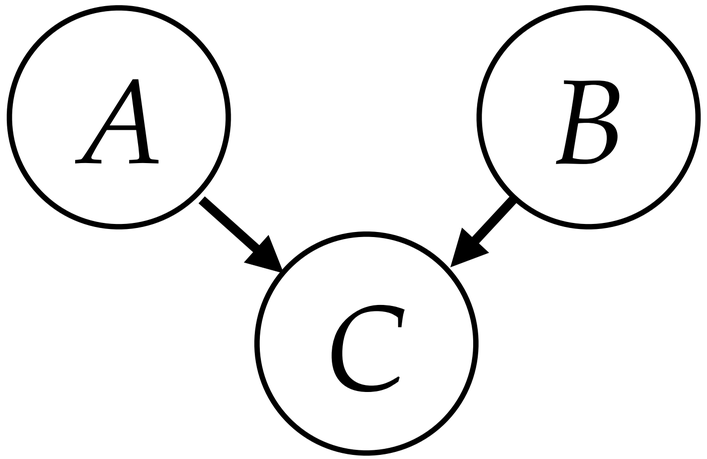
$$A \not\perp B$$

Interpretation: if we don't observe C, observing A gives information on C, which gives information about B. If we do observe C, all the information about B is already present.

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

$$A \perp B \mid C$$

Collider



$$p(A,B) = p(A)p(B)\sum_C p(C|A,B) = p(A)p(B)$$

$$A \perp B$$

$$p(A,B|C) = \frac{1}{p(C)}p(A)p(B)p(C|A,B) \neq p(A|C)p(B|C)$$

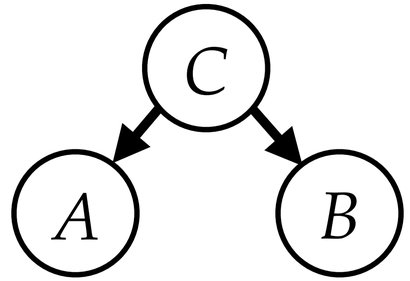
$$A \not\perp B \mid C$$

If we don't observe C, knowing A does not tell me what information B provided to C. If we do observe C, knowing A gives me information on what B must have been to explain the value of B (*explaining away*)

(Conditional) Independence: N variables

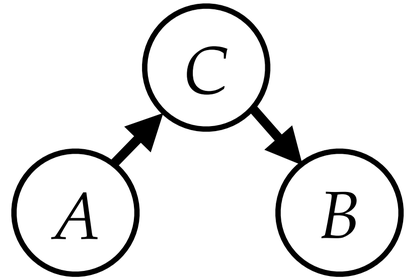
Let's use the insights so far to construct a general criterion.

Here: A, B, C can be (sets of) variables in the graph



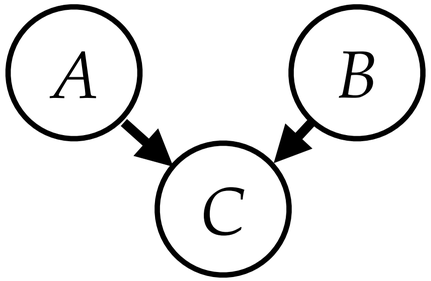
Blocking: We consider a path between A and B blocked given C if

1. a non-collider node on the path is in C
2. there is a collider node on the path, and neither it or any of its descendants are in C

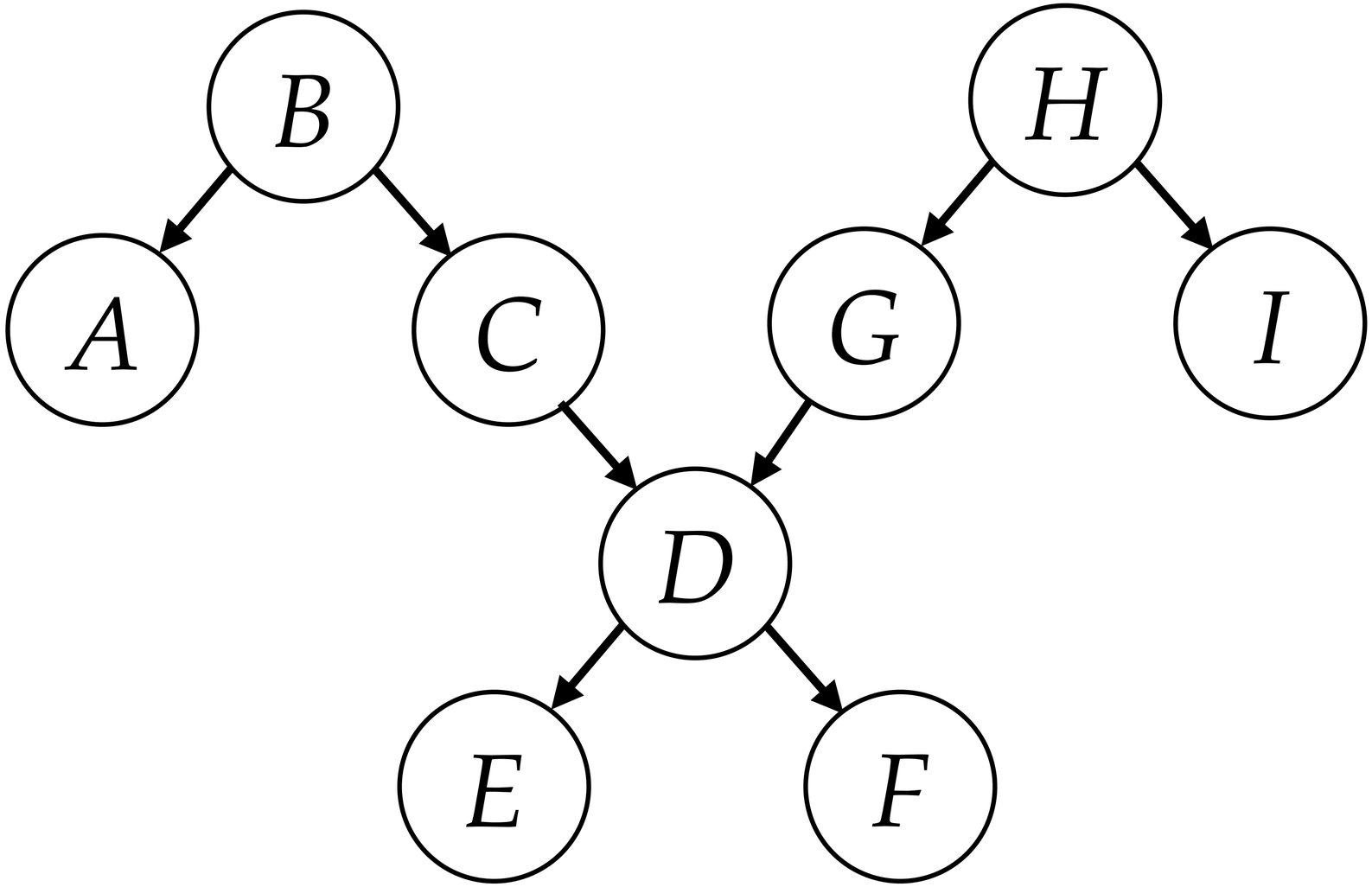


Directional-separation (D-separation)

If every path between A and B is blocked then $A \perp B \mid C$



Practice Examples



Which of these are true:

$B \perp E$

$B \perp E \mid C$

$B \perp I$

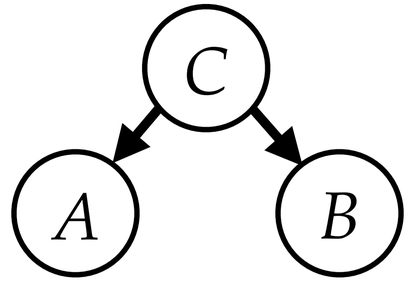
$B \perp I \mid E$

$B \perp I \mid \{D, H\}$

(Conditional) Independence: N variables

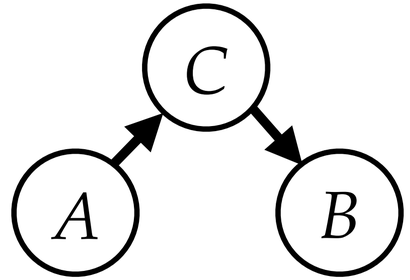
Let's use the insights so far to construct a general criterion.

Here: A, B, C can be (sets of) variables in the graph



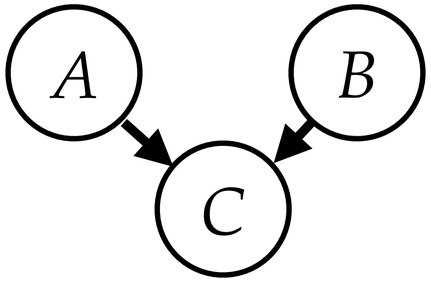
Blocking: We consider a path between A and B blocked given C if

1. a non-collider node on the path is in C
2. there is a collider node on the path, and neither it or any of its descendants are in C



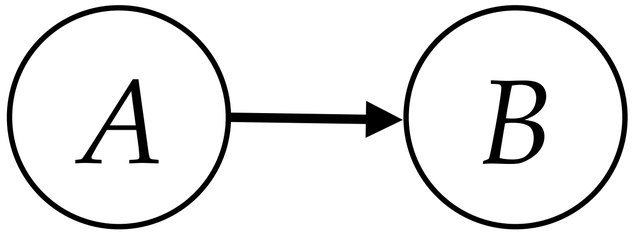
Directional-separation (D-separation)

If every path between A and B is blocked then $A \perp B \mid C$

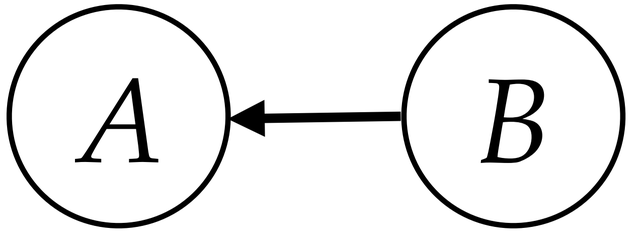


Note on Causality

- Often naturally follows from how we construct these graphs, but it is an additional assumption!
- Can have a correct probabilistic model, but not causal
- Important when we want to reason about the effect of actions



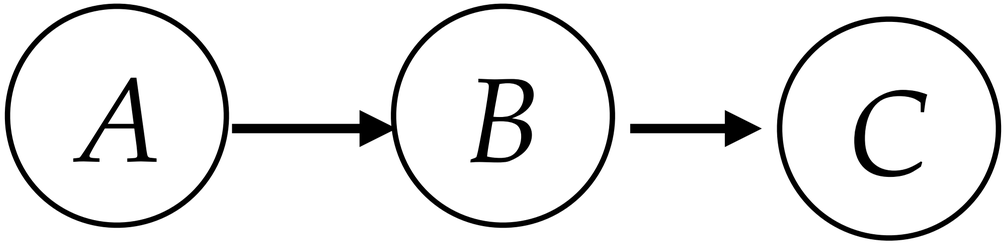
CAUSAL TRUTH



SAME INDEPENDENCIES

Inference in Bayesian Networks

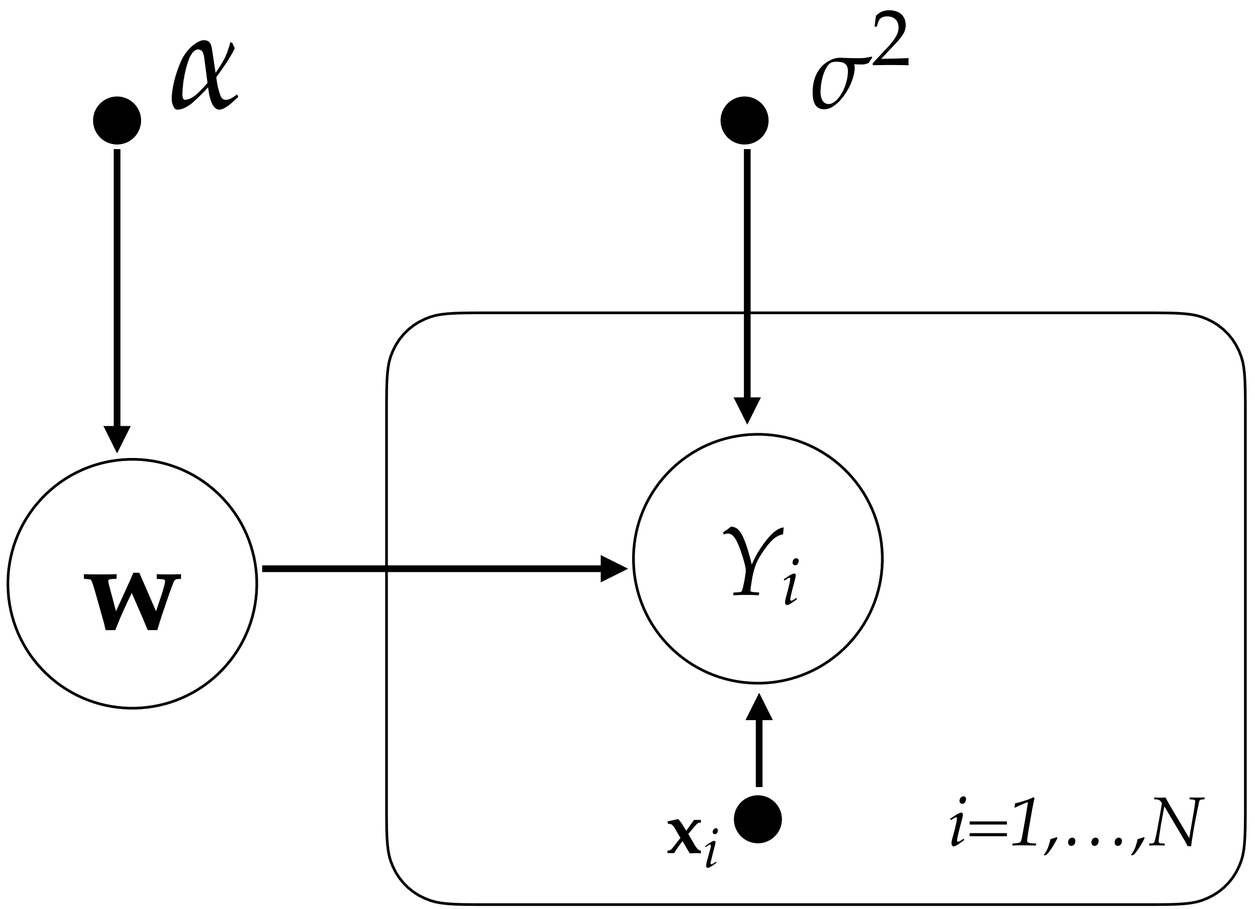
- **Ancestral sampling**: marginalizing is easy, what about conditioning?
- **Message passing algorithms:** passing "messages" over the graph:
  - Suppose we want p(A), p(C)
  - Naive: marginalizing the joint distribution
  - Use the structure to go from D^M to $M D^2$



Learning the Network Structure

- We might not know the structure, but we can try to learn it
- Fewer links -> lower complexity
- Hard problem. Ambiguous solutions?

Machine Learning Models as PGMs



Conclusion

- Bayesian Inference treats (all) parameters in the model as a random variables, and deals with consistently updating these random variables in the light of evidence.
- Bayesian networks are tools to represent, reason and do inference for joint probability distributions. D-separation is a visual technique to reason about (in)dependencies among variables.
- Next lecture: clustering (latent variable in the PGM?)