ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Bayesian networks and gene regulation

Niko Beerenwinkel

Outline

- Probabilities
- Statistical inference
- Gene regulation
- Bayesian networks
- Learning, Marginal likelihood

Probability distributions

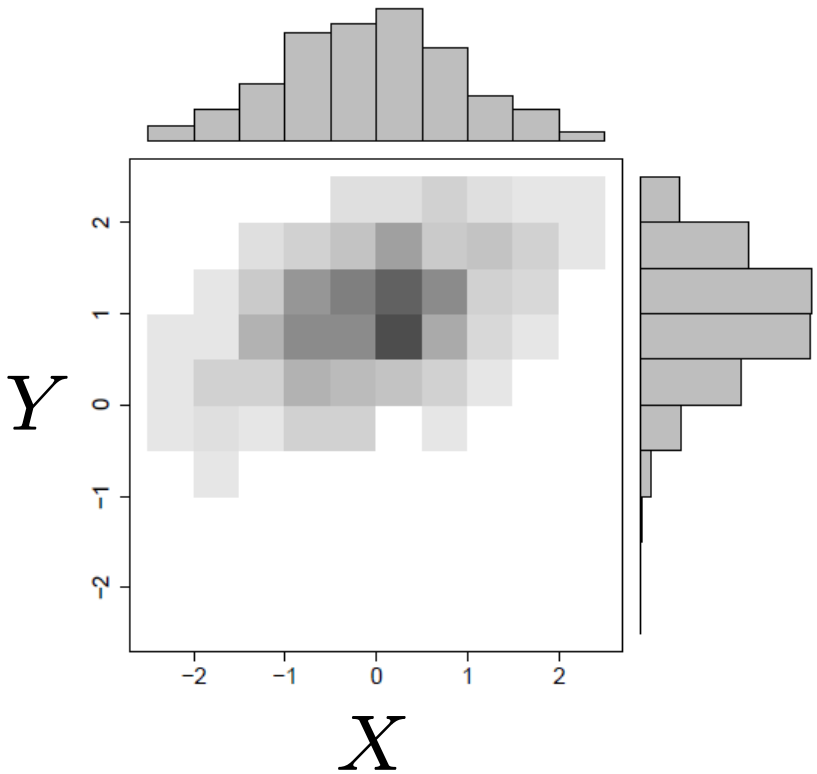
- Let X be a random variable (discrete or continuous) with probability distribution P(X).
- The *joint probability* of X and Y is denoted P(X,Y).
- The *marginal probabilities are*, in the discrete case,

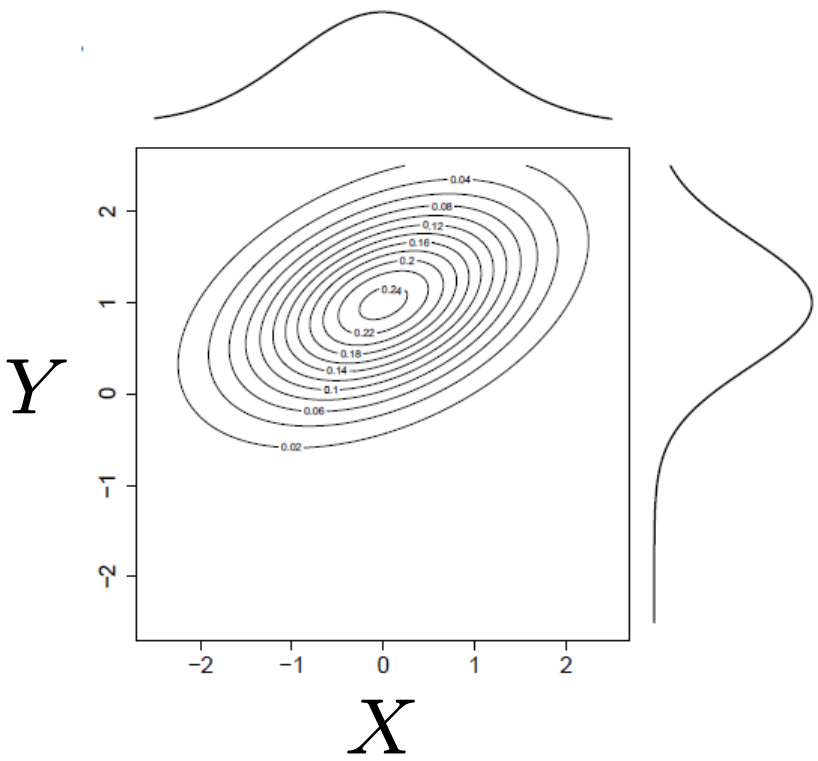
$$P(X) = \sum_Y P(X,Y), \qquad P(Y) = \sum_X P(X,Y)$$

and, in the continuous case,

$$P(X) = \int_Y P(X,Y)\,dY, \qquad P(Y) = \int_X P(X,Y)\,dX$$

Marginalization





Conditional probabilities

- The *conditional probability* of X given Y is

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

- Example:
  - Let G indicate overexpression of a certain oncogene.
  - Let C indicate the presence of a tumor.
  - P(G, C) is the probability of oncogene overexpression *and* the person suffering from a tumor.
  - P(G | C) = Prob. of oncogene overexpression in tumor patients (can be assessed by counting).
  - P(C | G) = Prob. of cancer given gene expression measurement (might be difficult to assess).

Bayes' theorem

- Because P(G, C) = P(G | C) P(C) = P(C | G) P(G),

$$P(C \mid G) = \frac{P(G \mid C)P(C)}{P(G)}$$

- The diagnostic conditional probability P(C | G) can be computed without determining it explicitly.

Statistical inference

- Let X be the outcome of a coin tossing experiment
- θ = P(X = heads) is the model parameter
- We want to estimate θ from the data $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$, where each $x^{(i)}$ is an observation of a coin toss ("heads" or "tails").

- *Frequentist approach*: Find best guess of θ, usually invoking maximum likelihood

- *Bayesian approach*: Regard θ as a random variable and estimate its posterior $P(\theta \mid \mathcal{D})$

Likelihood function

- The likelihood is the probability of the data given the model,

$$L(\theta) = P(\mathcal{D} \mid \theta)$$

- For the coin tossing experiment, with *k* the number of heads observed,

$$\begin{aligned} P(\mathcal{D} \mid \theta) &= \binom{N}{k} \prod_{i=1}^{N} P(X = x^{(i)} \mid \theta) \\ &\propto \prod_{i=1}^{N} \theta^{I\{x^{(i)}=\text{heads}\}} (1-\theta)^{I\{x^{(i)}=\text{tails}\}} \\ &= \theta^{k} (1-\theta)^{N-k} \end{aligned}$$

Maximum likelihood (ML)

- ML estimates are consistent and asymptotically unbiased.
- To find the MLE, we maximize the log-likelihood

$$\ell(\theta) = \log L(\theta) = \log P(\mathcal{D} \mid \theta)$$

- For the coin tossing model, we find

$$\ell(\theta) = k \log \theta + (N - k) \log(1 - \theta) + C$$

where C is a constant that does not depend on θ. Hence

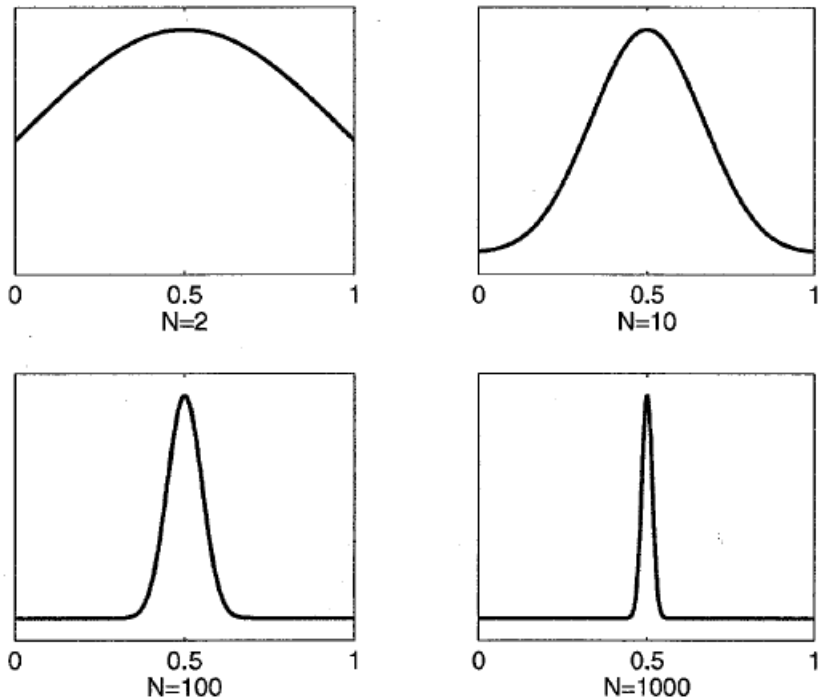
$$\frac{d\ell(\theta)}{d\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{k}{N}$$

The frequentist paradigm

$$\theta \longrightarrow \boxed{\text{Data}} \longrightarrow \hat{\theta}$$

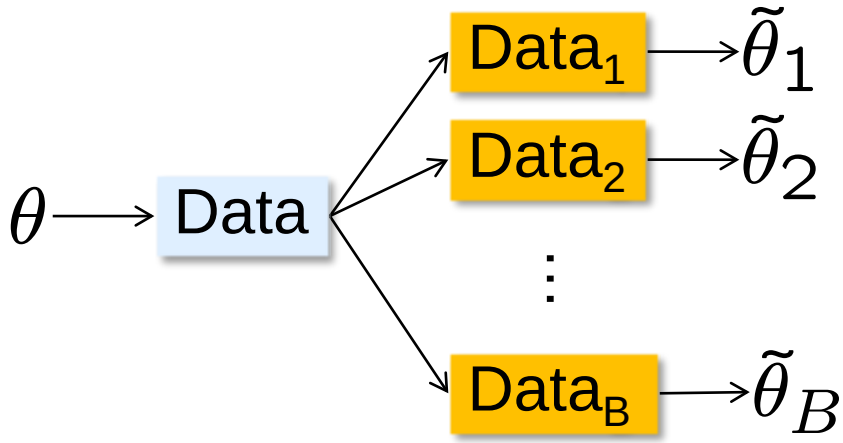
But how sure can we be about the MLE?

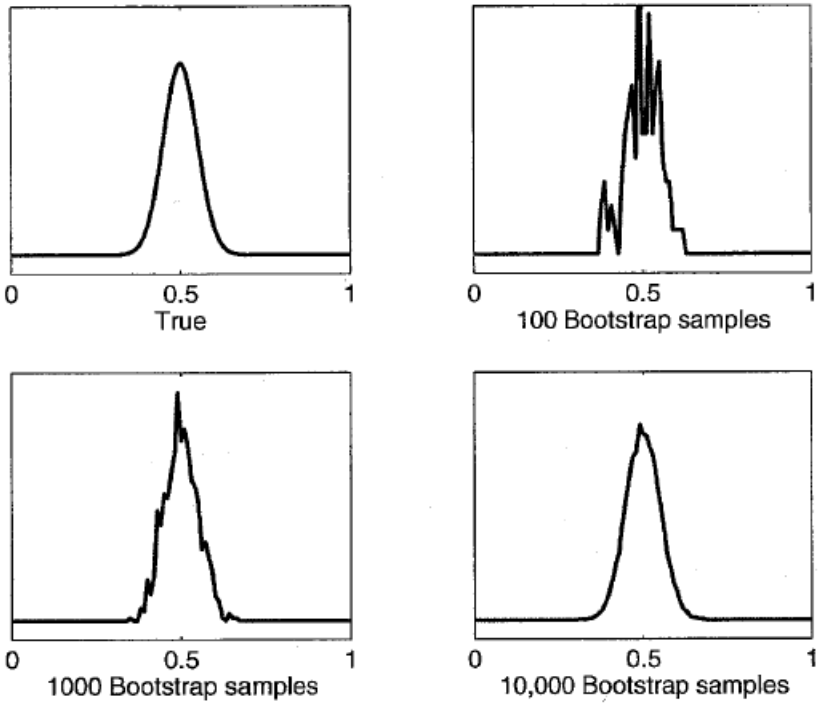
$$\theta \rightarrow \boxed{\text{Data}_1} \rightarrow \hat{\theta}_1$$
$$\theta \rightarrow \boxed{\text{Data}_2} \rightarrow \hat{\theta}_2$$
$$\vdots$$
$$\theta \rightarrow \boxed{\text{Data}_M} \rightarrow \hat{\theta}_M$$



The bootstrap

- If we cannot repeat the experiment, resample from $\mathcal{D}$

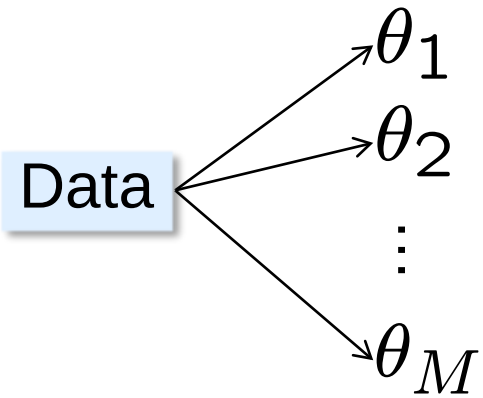




$N = 100$

The Bayesian paradigm

- We obtain P(θ | $\mathcal{D}$) directly from the observed data $\mathcal{D}$ using Bayes' theorem:



posterior

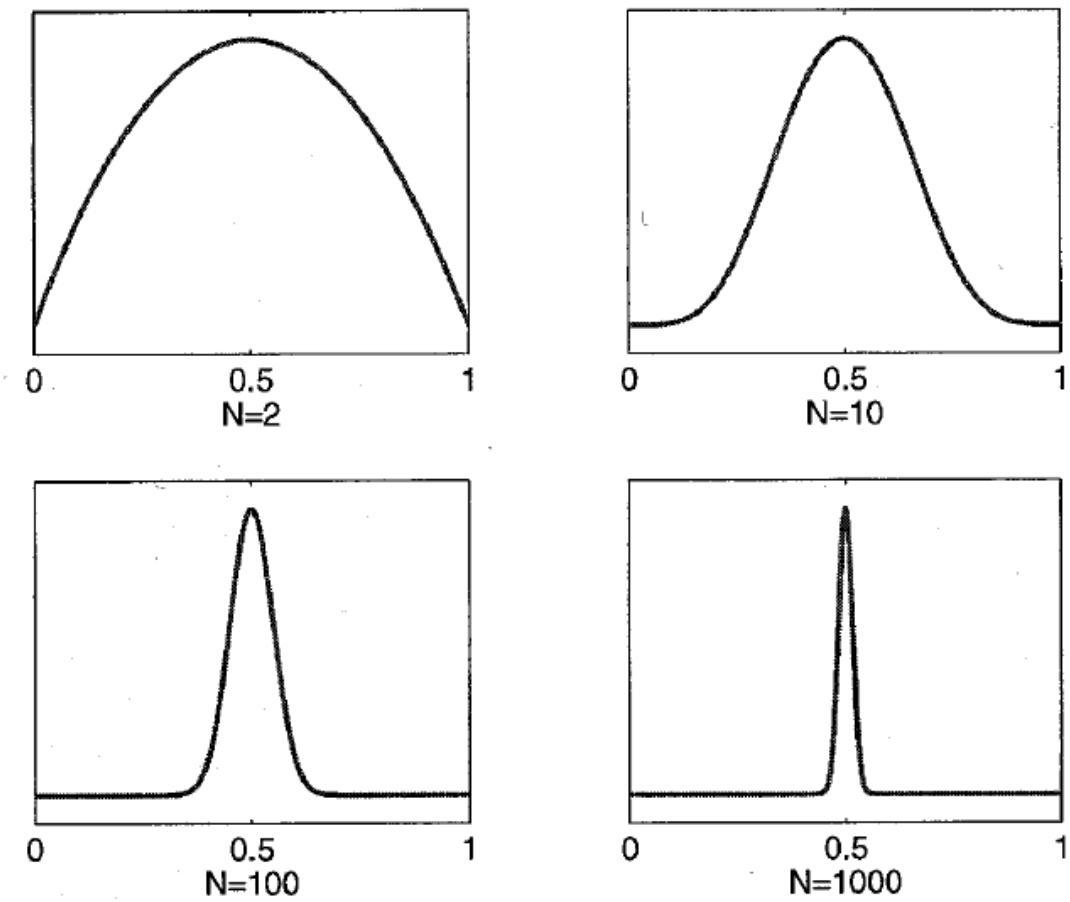
likelihood

prior

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

constant

Posterior of θ for a uniform prior



Prior

- The prior $P(\theta)$ is our *a priori* believe in θ.
  It reflects domain-specific knowledge.
- For an uninformative prior, any observation $x^{(i)}$ is equally likely *a priori*.
- A conjugate prior is one that is invariant (with respect to the distribution family) under multiplication with the likelihood, i.e., the posterior belongs to the same family as the prior.
- Conjugate priors are mathematically convenient, because the posterior can be calculated analytically.

Example: prior for the coin tossing model

- The coin tossing model has a binomial likelihood:

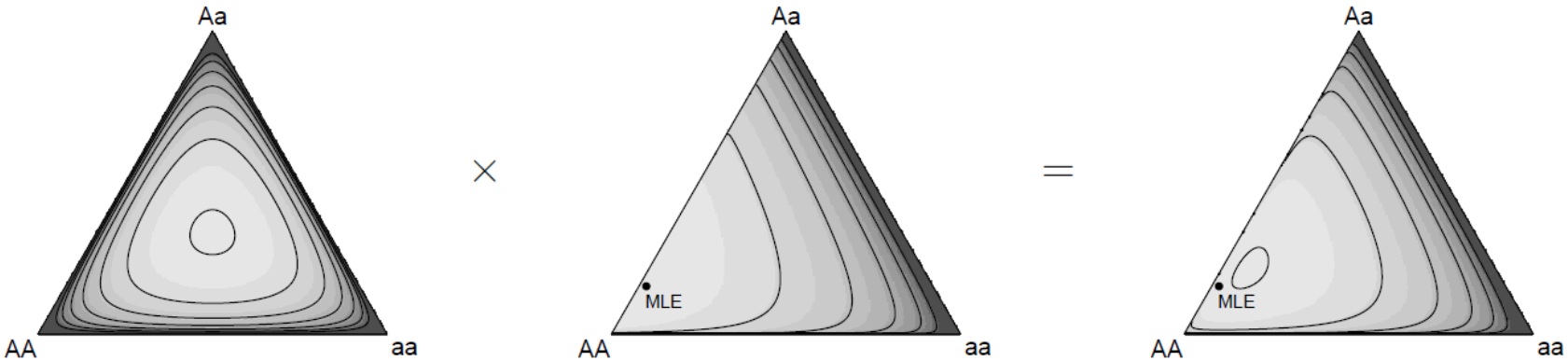
$$P(\mathcal{D} \mid \theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

- The beta distribution, Beta(θ | α, β) with hyperparameters α and β, is conjugate to the binomial:

$$P(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid k+\alpha,\, N-k+\beta)$$

Dirichlet prior

- The Dirichlet prior is conjugate to the multinomial likelihood:



$$T \sim \mathrm{Dir}(\alpha) \qquad X \mid T = \theta \sim \mathrm{Mult}(n, \theta_1, \ldots, \theta_K) \qquad \theta \mid X = x \sim \mathrm{Dir}(\alpha + x)$$

where the Dirichlet pdf and the multinomial pmf are, resp.,

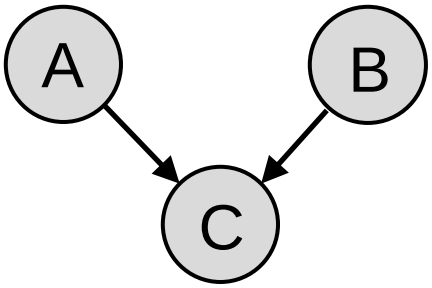
$$f(\theta_1, \ldots, \theta_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1} \qquad P(X = x) = \frac{n!}{x_1! \cdots x_K!} \theta_1^{x_1} \cdots \theta_K^{x_K}$$

Graphical models philosophy

Biology	Graph	Probabilistic model

Example: Gene regulation

Players:
genes A, B, C
Relationships:
"A regulates C"
"B regulates C"

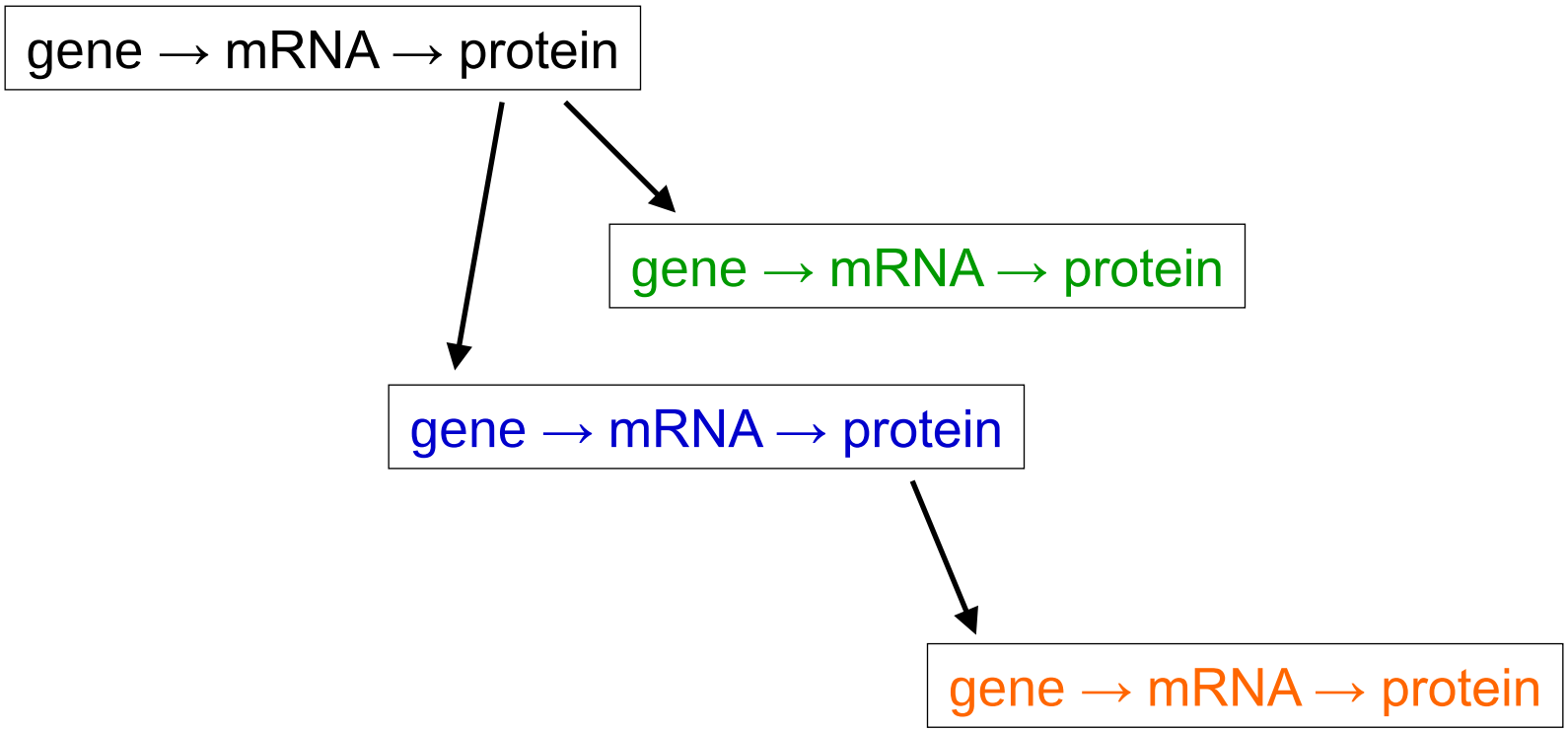


$$P(A,B,C) = P(A)\,P(B)\,P(C|A,B)$$

Biological players ⟷ Vertices ⟷ Random variables

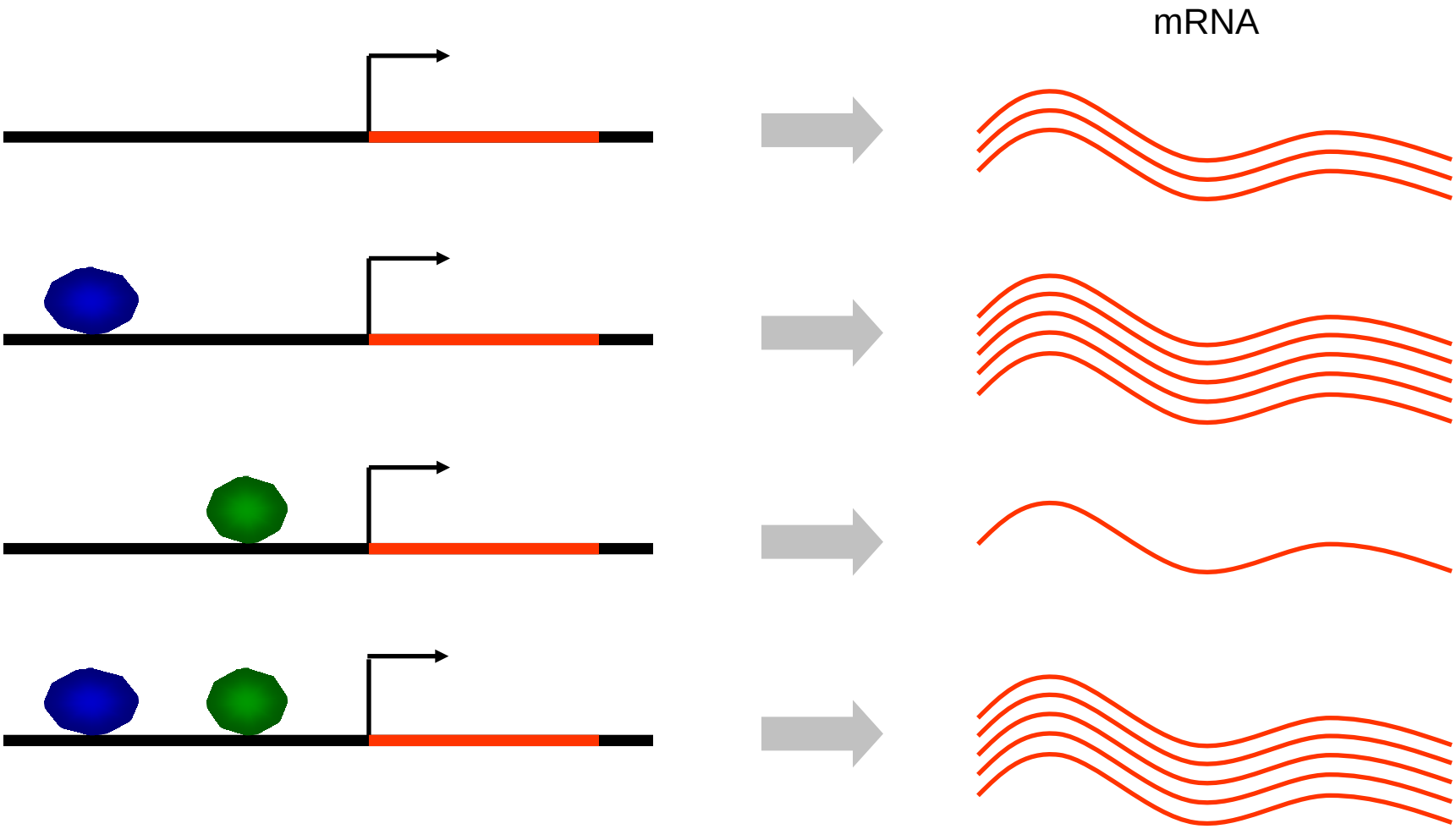
(Causal) Relationships ⟷ Edges ⟷ Statistical dependencies

Gene regulation

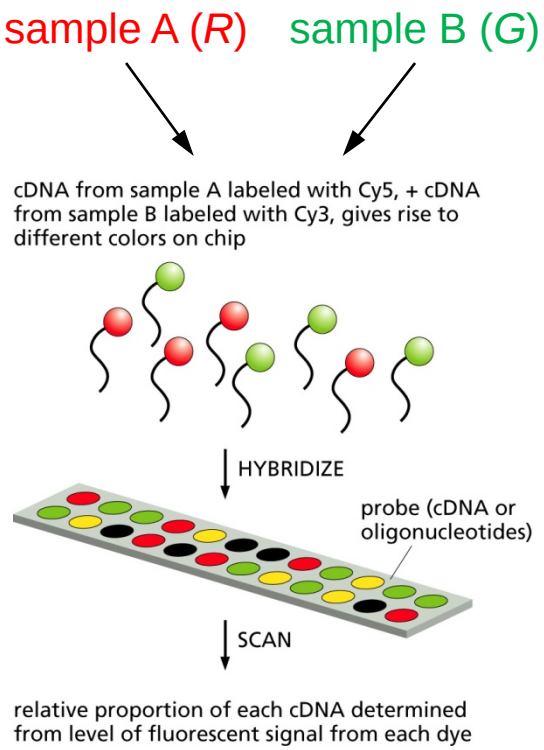


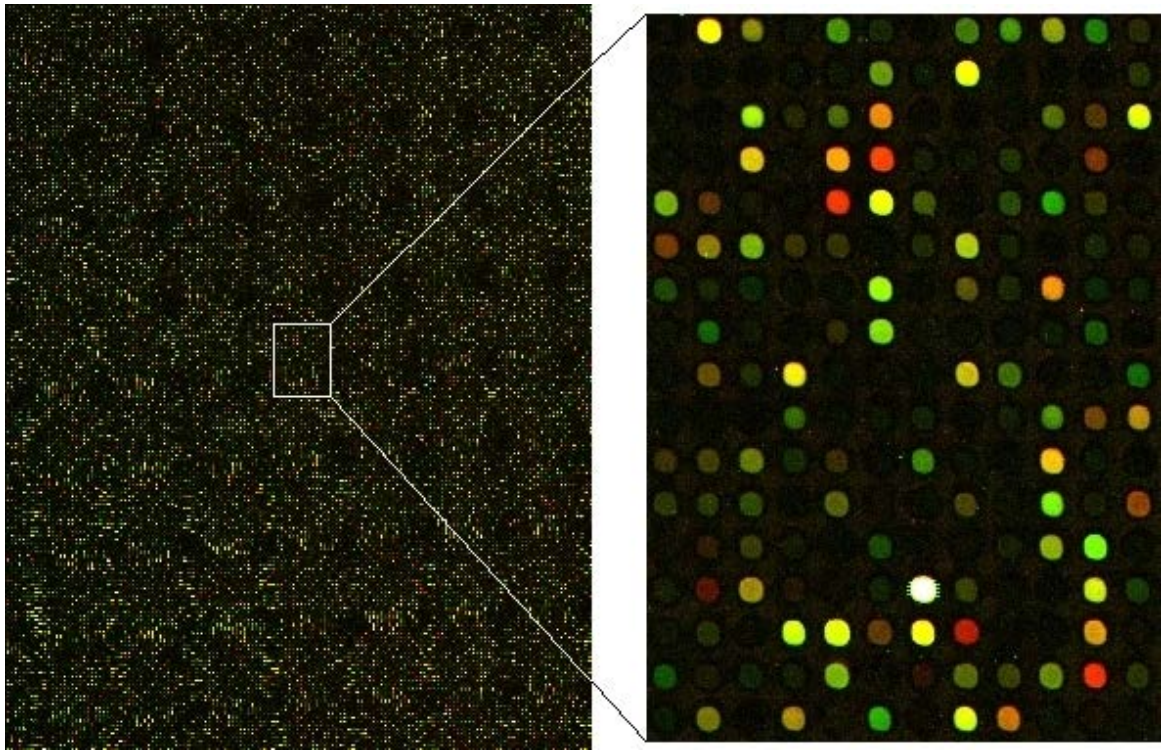
- Proteins can increase or decrease the rate of transcription of another gene by binding to the promoter region. These proteins are called *transcription factors*.

Transcriptional regulation



Two-color cDNA microarray

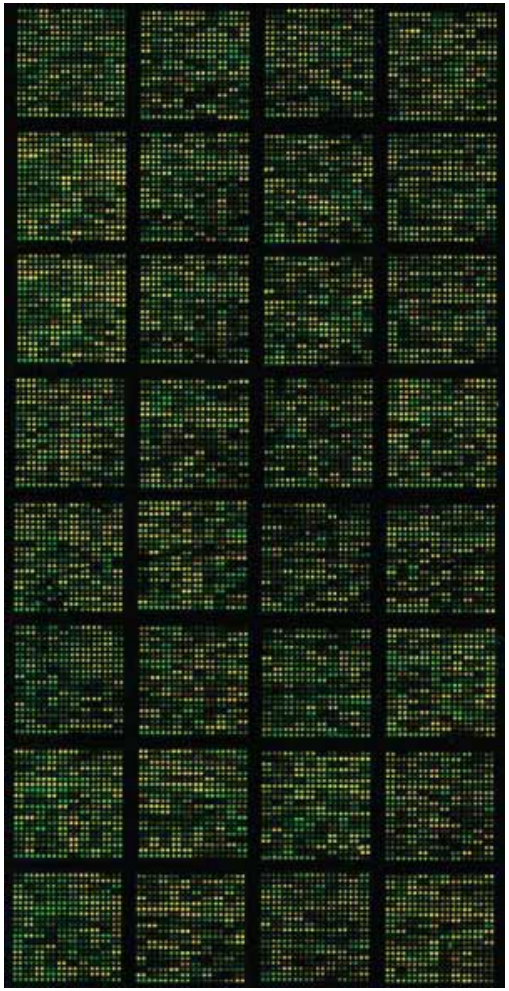




Microarrays measure gene expression

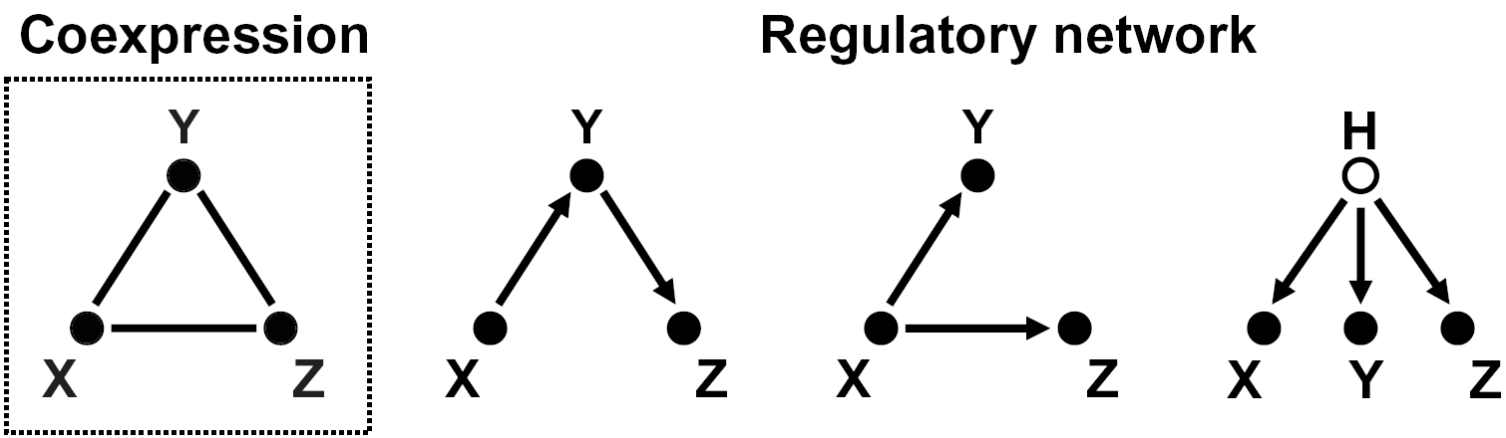
- 2D array of DNA sequences from thousands of genes
- Each spot has many copies of same gene
- mRNAs from a sample are allowed to hybridize. The number of hybridizations per spot is a measure of the number of mRNAs in the sample.
- Microarray data requires careful normalization before further analysis.

- Alternatively, the mRNA can be directly sequenced and counted (RNA-seq).



Correlation versus causation

- Suppose three genes are regulated as X → Y → Z.
- Then X and Z are correlated, but do not interact directly.



All three regulatory networks can give rise to the same coexpression pattern!

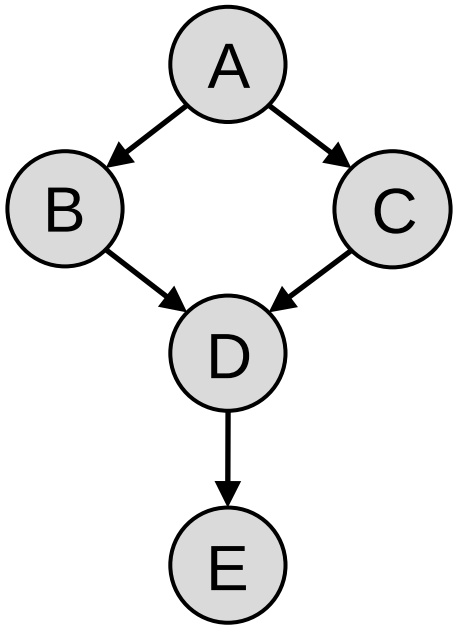
Bayesian networks

- A Bayesian network (BN) for $X = (X_1, \ldots, X_L)$ consists of
  - a directed acyclic graph (DAG) $G = (V, E)$, where $V = \{1, \ldots, L\}$
  - local probability distributions (LPDs), one for each vertex.
- The BN is defined as the family of distributions for which the joint probability factors into conditional probabilities as

$$P(X_1, \ldots, X_L) = \prod_{n=1}^{L} P\left(X_n \mid X_{\mathrm{pa}(n)}\right)$$

where pa(n) denotes the set of parents of vertex n in G, i.e., $X_{pa(n)} = (X_1, \ldots, X_k)$ if $\{1, \ldots, k\}$ are the parents of n in G.

Example



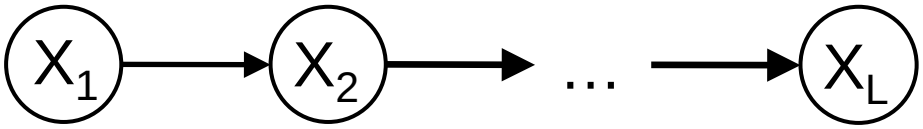
$$P(A, B, C, D, E) =$$
$$P(A)P(B \mid A)P(C \mid A)P(D \mid B, C)P(E \mid D)$$

Discrete variables

- If each X_n has K possible states [K] = {1, …, K}, then

$$\left(P(X_n = a \mid X_{\mathrm{pa}(n)} = b)\right)_{a \in [K],\, b \in [K]^{\mathrm{pa}(n)}}$$

  has $(K - 1) \times K^{|\mathrm{pa}(n)|}$ free parameters.
- If G is fully connected, the maximal number of $K^L - 1$ parameters is attained (exponential in L).
- If all X_n are independent (no edges), we have $L(K - 1)$ parameters.
- For the chain, we find $(K - 1) + (L - 1)K(K - 1)$ free parameters, $O(LK^2)$



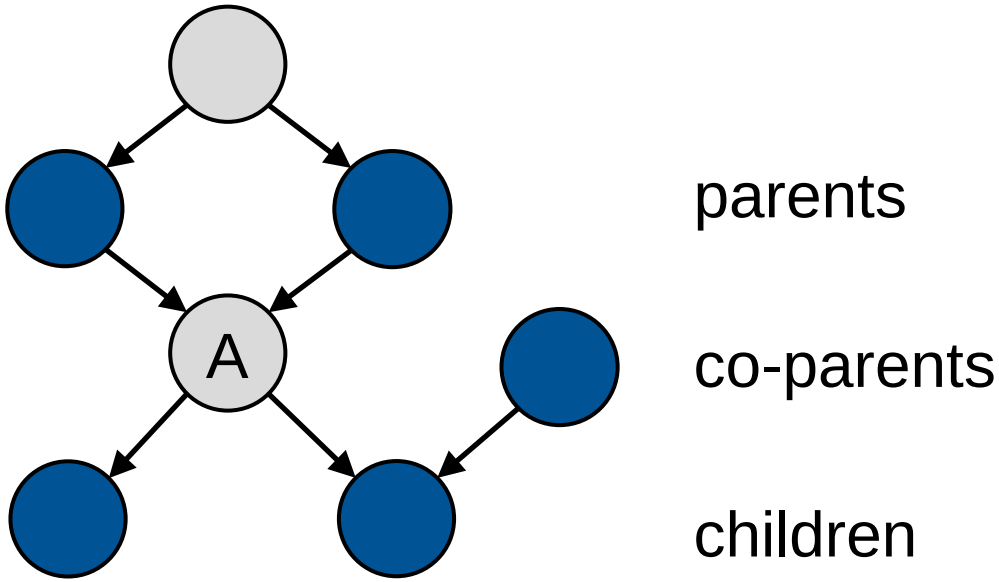
Linear Gaussian models

- Linear-Gaussian models are defined by the continuous LPD

$$P(X_n \mid X_{\mathsf{pa}(n)}) = \mathsf{Norm}(b_n + w_n^t \cdot X_{\mathsf{pa}(n)}, v_n)$$

  with parameters b_n, w_n for the mean, and variance v_n.
- There are recursive formulas for the expectation and covariance of $(X_1, \ldots, X_L)$.
- The number of parameters increases linearly with the number of parents.
- Only linear relationships can be modeled.

Markov blanket



- The Markov blanket (MB) of a vertex is the set of its parents, co-parents, and children.
- The BN factorization is equivalent to

$$P(X_k \mid X_n, n \neq k) = P(X_k \mid X_{\mathrm{MB}(k)}) \quad \forall k$$

Conditional independence

- We say that A and B are conditionally independent given C, and write

$$A \perp B \mid C$$

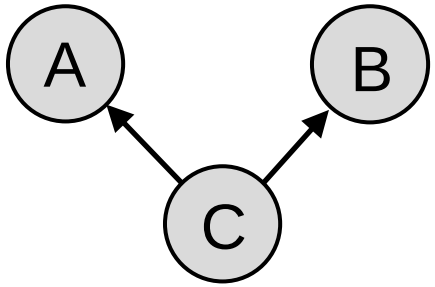
if $$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

- A, B, and C can be subsets of random variables.

- If C = $\emptyset$, we say that A and B are (marginally) independent,

$$A \perp B$$

Example



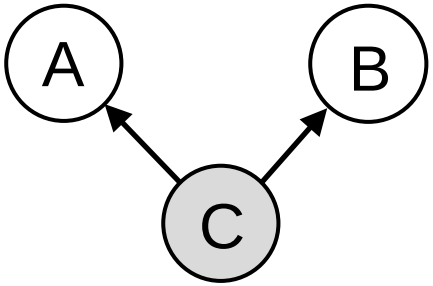
$$P(A,B,C) = P(A \mid C)P(B \mid C)P(C)$$

$$\begin{aligned} P(A,B \mid C) &= \frac{P(A,B,C)}{P(C)} \\ &= P(A \mid C)P(B \mid C) \\ &\Rightarrow \quad A \perp B \mid C \end{aligned}$$

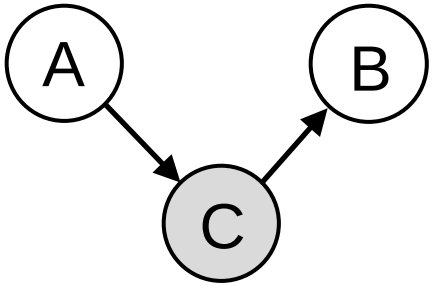
- However, in general, $A \not\perp B$

Three basic examples

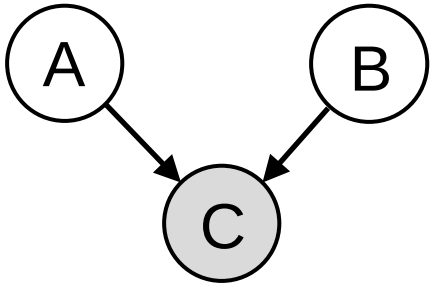
explaining away



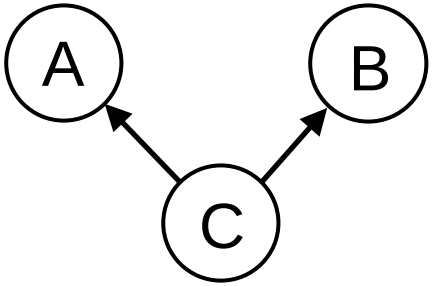
$$A \perp B \mid C$$



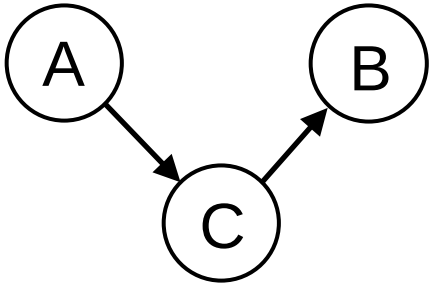
$$A \perp B \mid C$$



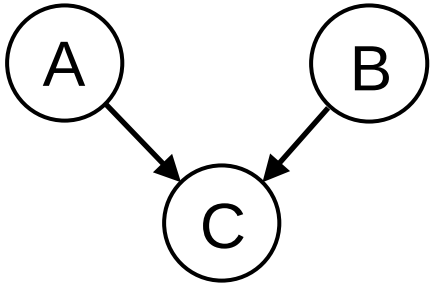
$$A \not\perp B \mid C$$



$$A \not\perp B$$



$$A \not\perp B$$



$$A \perp B$$

Learning Bayesian networks from data

- Learning a BN (G, θ) from data $\mathcal{D}$ involves two steps:

1. Find the maximum a posteriori (MAP) estimate of the network structure G,

$$G^* = \underset{G}{\operatorname{argmax}}\, P(G \mid \mathcal{D})$$

2. Given the optimal network structure G*, find the MAP estimate of the parameters θ,

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, P(\theta \mid G^*, \mathcal{D})$$

Marginal likelihood

- Applying Bayes' theorem we find for the posterior,

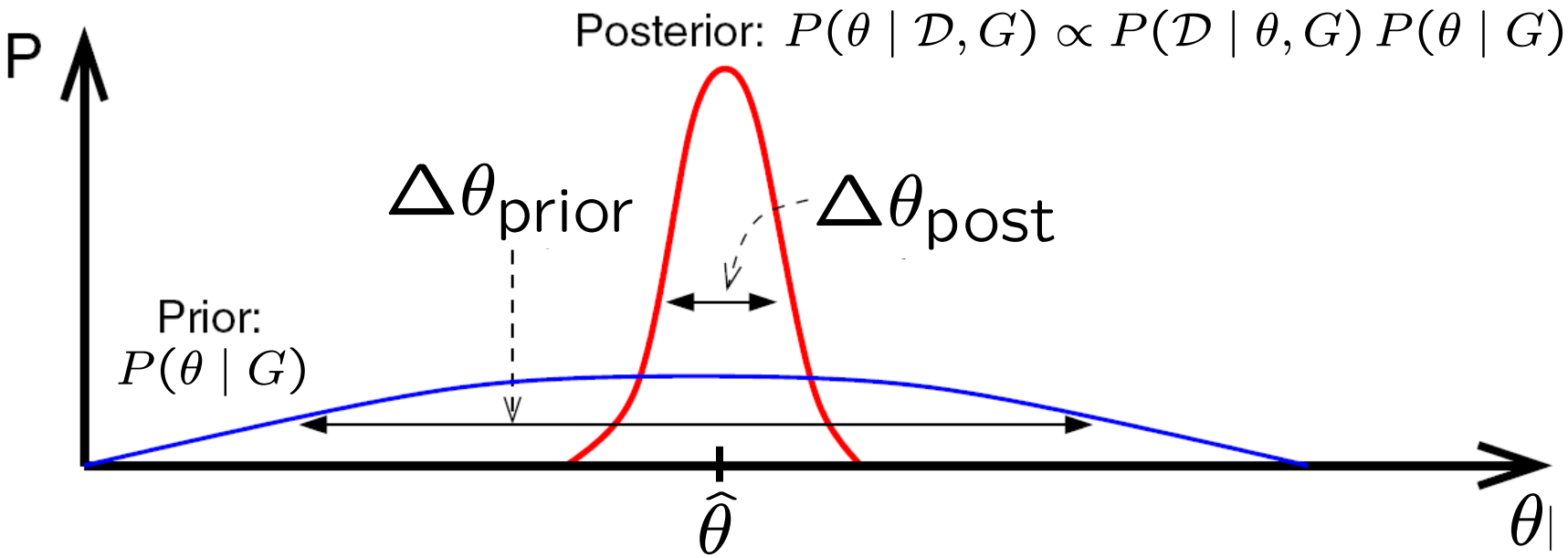
$$P(G \mid \mathcal{D}) \propto P(\mathcal{D} \mid G) P(G)$$

where

$$\begin{aligned} P(\mathcal{D} \mid G) &= \int P(\mathcal{D}, \theta \mid G) d\theta \\ &= \int P(\mathcal{D} \mid \theta, G) P(\theta \mid G) d\theta \end{aligned}$$

is the marginal likelihood.

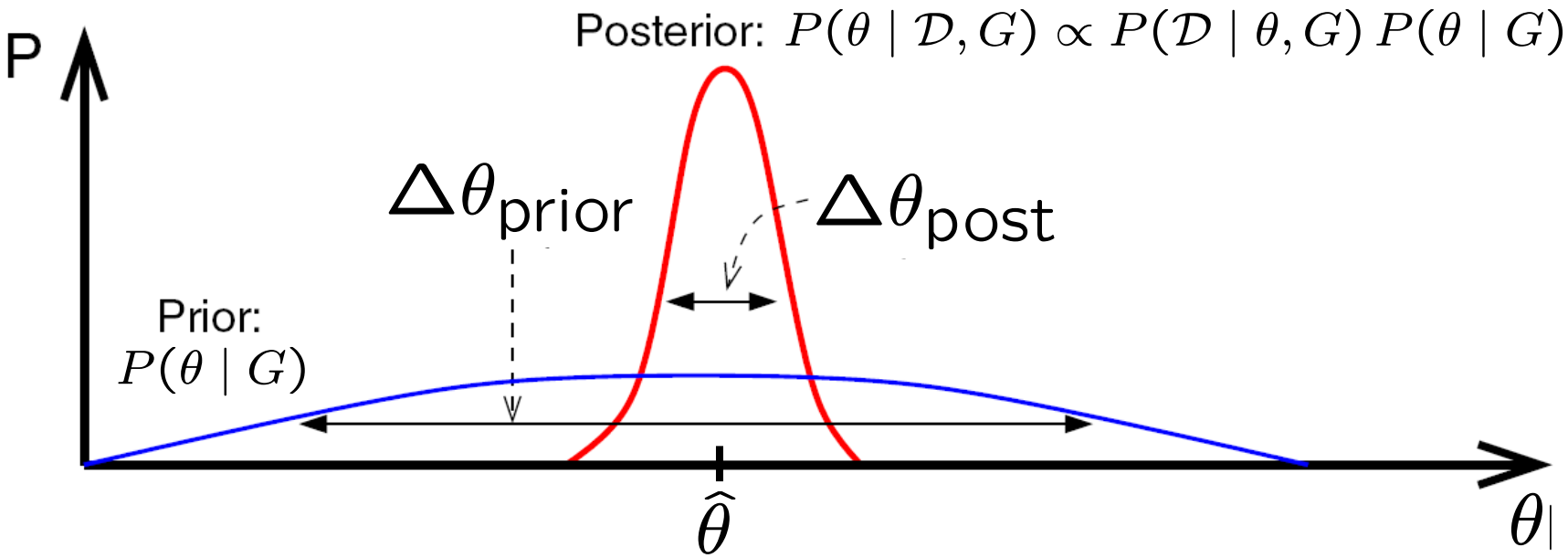
Marginal likelihood: flat prior and unimodal posterior



$$P(\mathcal{D} \mid G) = \int P(\mathcal{D} \mid \theta, G) P(\theta \mid G) d\theta$$

$$\approx P(\mathcal{D} \mid \hat{\theta}, G) \frac{\Delta\theta_{\text{post}}}{\Delta\theta_{\text{prior}}}$$

The marginal likelihood penalizes complexity



$$P(\mathcal{D} \mid G) \approx \underbrace{P(\mathcal{D} \mid \hat{\theta}, G)}_{\text{Likelihood at MLE: large if model fits well}} \underbrace{\frac{\Delta\theta_{\text{post}}}{\Delta\theta_{\text{prior}}}}_{\text{Occam factor: small if model is overfitted}}$$

Summary

- The two most popular techniques for statistical inference are maximum likelihood and Bayes. They differ conceptually, but mathematically they are closely related.
- Bayesian networks are probabilistic directed graphical models for a random vector $X = (X_1, ..., X_n)$, where the graph defines a factorization of the joint probability $P(X)$.
- Bayesian networks can be used to model biological networks, for example, gene regulatory networks.
- Learning a Bayesian network from observed data involves computing the marginal likelihood.

References

- Bishop CM. **Pattern Recognition and Machine Learning**. Section 8.1.
- Husmeier D, Dybowski R, Roberts S (eds.). **Probabilistic Modeling in Bioinformatics and Medical Informatics**. Chapters 1, 2.
- Beerenwinkel N and Siebourg J. **Statistics, probability, and computational science.** In Maria Anisimova, editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*, chapter 3, pages 77–110. Springer, New York, 2012. DOI: 10.1007/978-1-61779-582-4_3. Sections 1, 2, 7.