ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Probabilistic graphical models

Niko Beerenwinkel

Schedule

- Date: Fri, 17 Jan 2020
- Room: HG D 5.1

- Lecture: 13:00 – 14:45
- Tutorial: 15:00 – 16:00 (Simon Dirmeier)

Outline

- Statistical inference
- Bayesian networks
- Conditional independence
- Inference
- Learning
- Dynamic Bayesian networks

Probability distributions

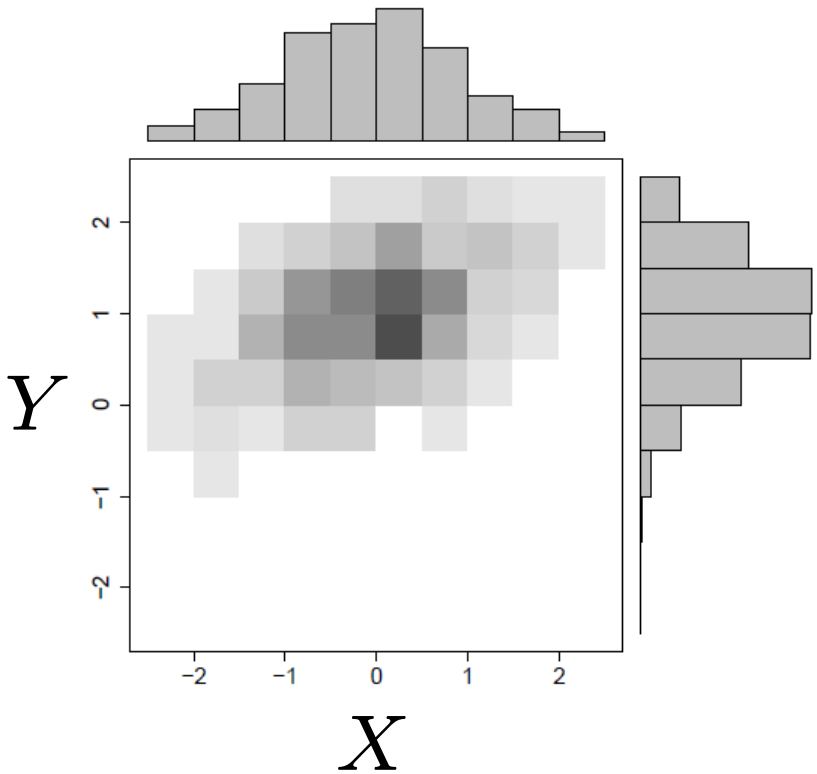
- Let X be a random variable (discrete or continuous) with probability distribution P(X).
- The *joint probability* of X and Y is denoted P(X,Y).
- The *marginal probabilities are*, in the discrete case,

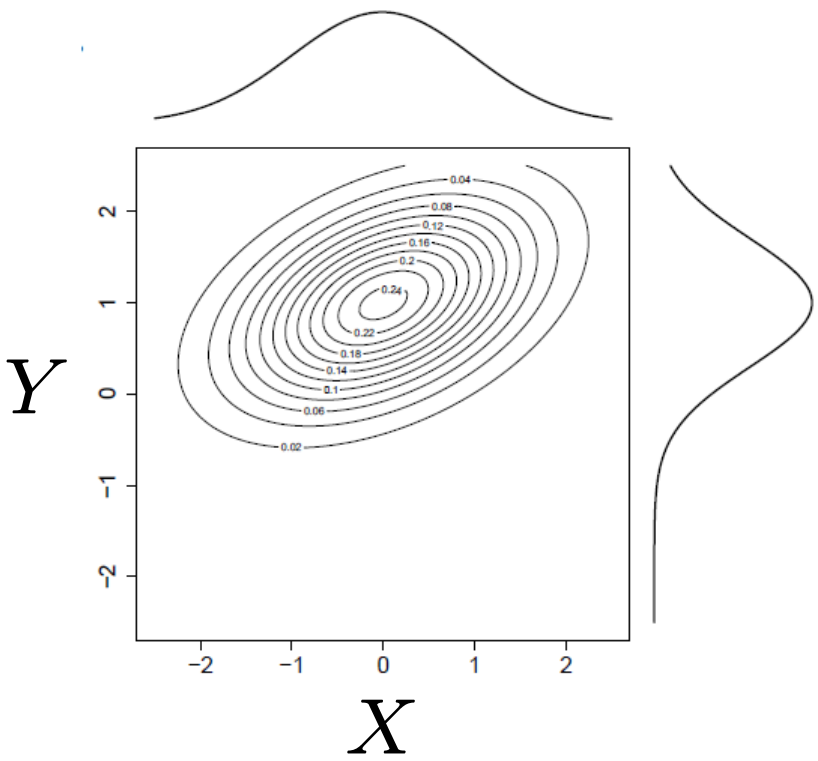
$$P(X) = \sum_Y P(X,Y), \qquad P(Y) = \sum_X P(X,Y)$$

and, in the continuous case,

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

Marginalization





Conditional independence

- The *conditional probability* of *Y* given *X* is

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

- *X* and *Y* are independent, if P(*X*,*Y*) = P(*X*) P(*Y*).
- *X* and *Y* are *conditionally independent given Z*, if

P(*X*,*Y* | *Z*) = P(*X* | *Z*) P(*Y* | *Z*), or equivalently, if

P(*X* | *Z*) = P(*X* | *Y*, *Z*).

Bayes theorem

- Bayes' theorem states that

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

  where P(Y | X) is the posterior and P(Y) is the prior probability.

- If $y_1, \ldots, y_n$ are the disjoint outcomes of Y, then for any r.v. X, $\mathrm{P}(X) = \sum_Y \mathrm{P}(X, Y) = \sum_{i=1,\ldots,n} \mathrm{P}(X \mid Y = y_i)\, \mathrm{P}(Y = y_i)$ and hence

$$\mathrm{P}(Y \mid X) = \frac{\mathrm{P}(X \mid Y)\mathrm{P}(Y)}{\sum_i \mathrm{P}(X \mid y_i)\mathrm{P}(y_i)}$$

Example: A diagnostic test

- $D = 1$ indicates disease (else $D = 0$)
- $T = 1$ indicates a positive test result (else $T = 0$)
- The disease is rare. It has a prevalence of 0.5% in the population, $\mathrm{P}(D = 1) = 0.005$ (prior disease probability).
- The test has a false positive rate of 5%, $\mathrm{P}(T = 1 \mid D = 0) = 0.05$, and a true positive rate of 90%, $\mathrm{P}(T = 1 \mid D = 1) = 0.9$.
- We are interested in the posterior probability of disease given a positive test result:

$$\begin{aligned}\mathrm{P}(D = 1 \mid T = 1) &= \frac{\mathrm{P}(T = 1 \mid D = 1)\mathrm{P}(D = 1)}{\mathrm{P}(T = 1 \mid D = 0)\mathrm{P}(D = 0) + \mathrm{P}(T = 1 \mid D = 1)\mathrm{P}(D = 1)} \\ &= 0.083\end{aligned}$$

→ only 8% of the positively tested persons actually have the disease!

Statistical inference

- Let Y be the outcome of a coin tossing experiment.
- θ = P(Y = heads) is the model parameter.
- We want to estimate θ from the data $\mathcal{D} = \{y_1, \ldots, y_N\}$, where each y_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,

$$\begin{aligned} P(\mathcal{D} \mid \theta) &= \binom{N}{k} \prod_{i=1}^{N} P(Y = y_i \mid \theta) \\ &= \binom{N}{k} \theta^k (1-\theta)^{N-k} \\ &\propto \theta^k (1-\theta)^{N-k} \end{aligned}$$

Maximum likelihood (ML)

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

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

- 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

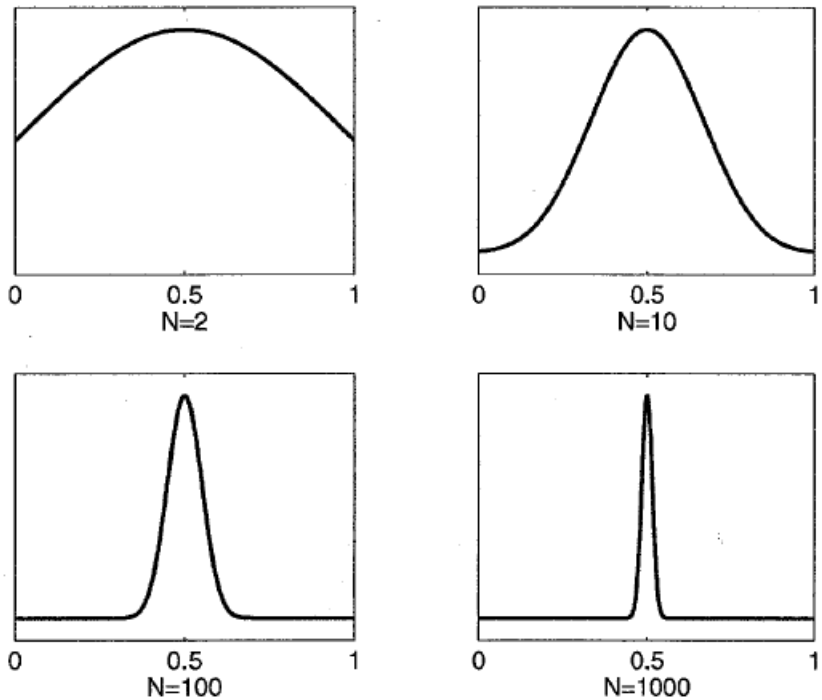
$$\frac{d\ell(\theta)}{d\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{k}{N}$$

The frequentist paradigm

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

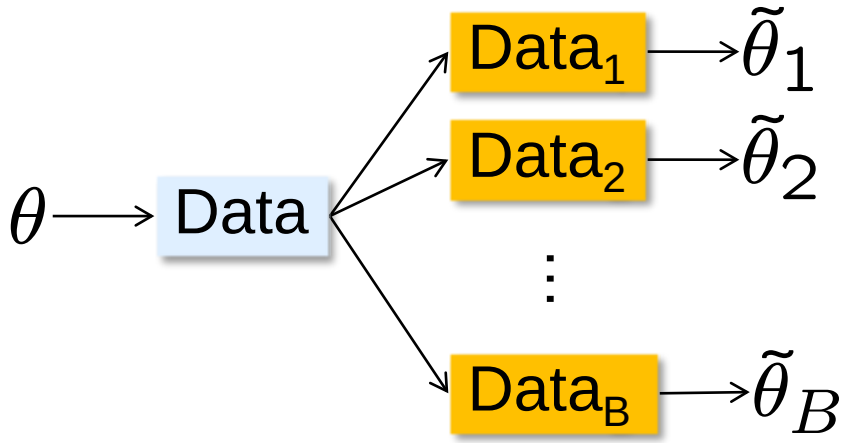
But how sure can we be about the MLE?

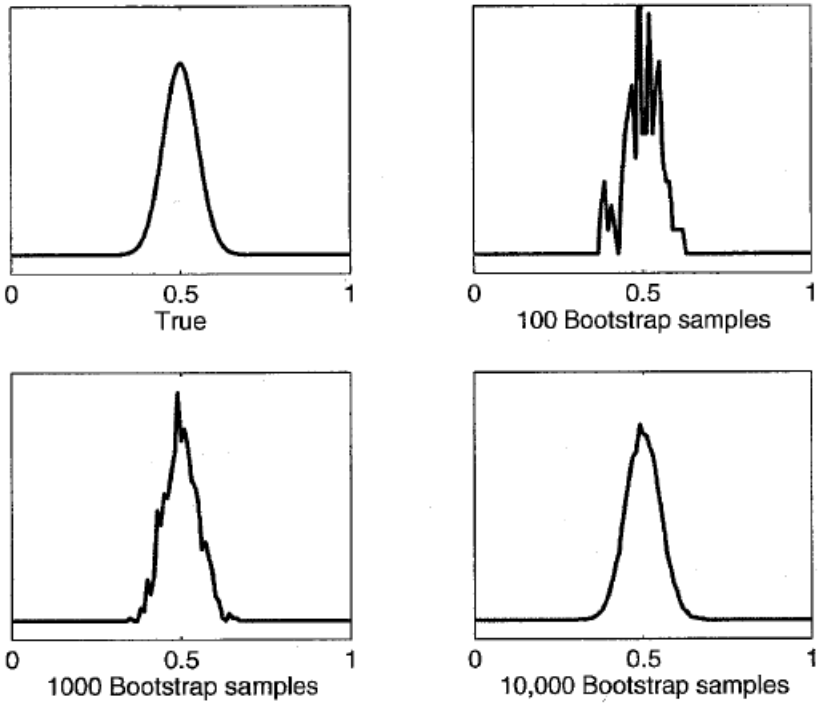
$$\theta \rightarrow \boxed{\text{Data}_1} \rightarrow \hat{\theta}_1$$
$$\theta \rightarrow \boxed{\text{Data}_2} \rightarrow \hat{\theta}_2$$
$$\vdots$$
$$\theta \rightarrow \boxed{\text{Data}_M} \rightarrow \hat{\theta}_M$$



The bootstrap

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

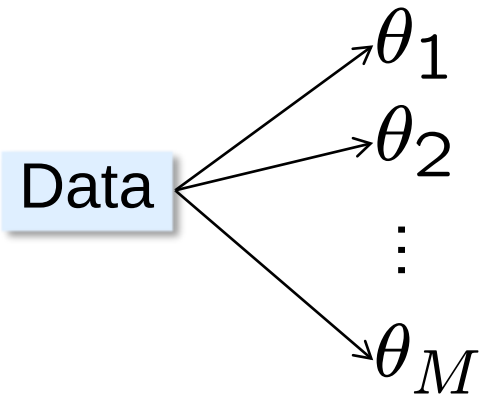




$N = 100$

The Bayesian paradigm

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



posterior

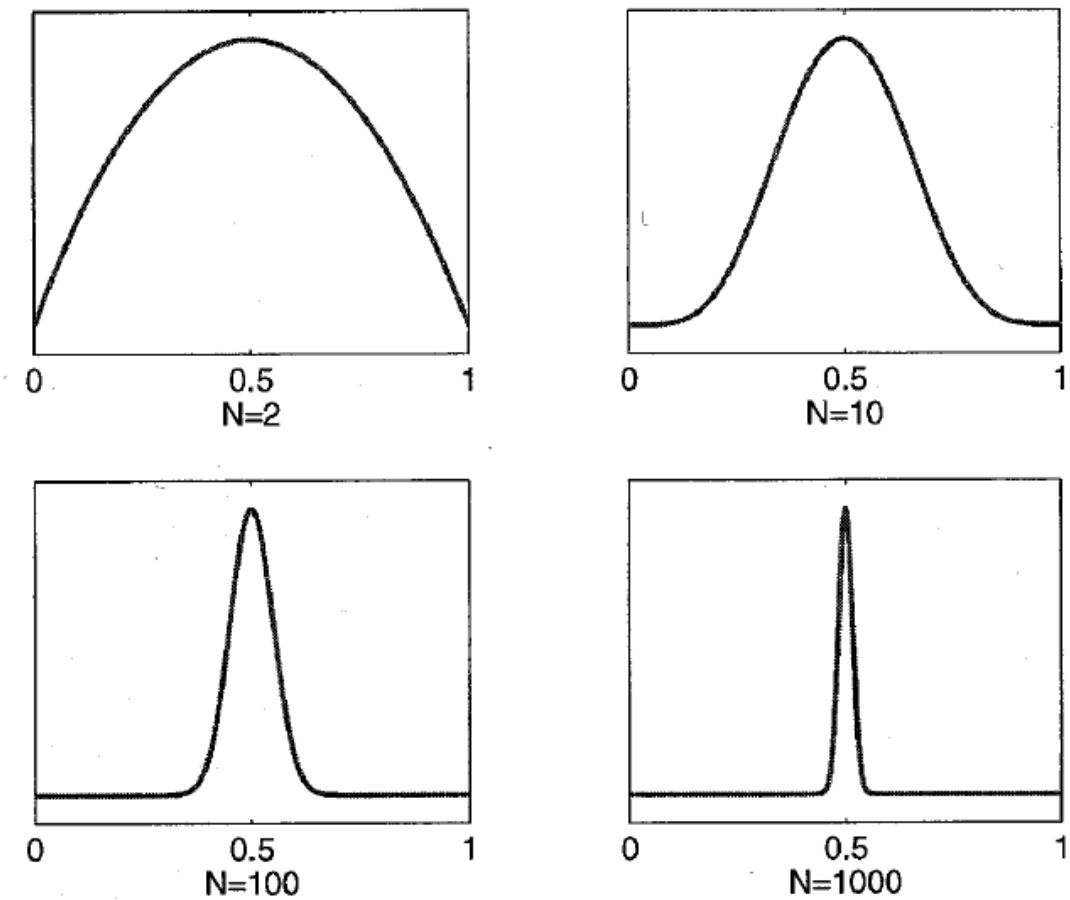
likelihood

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(θ) is our *a priori* believe in θ.
  It reflects domain-specific knowledge.
- For an uninformative prior, any observation y_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:

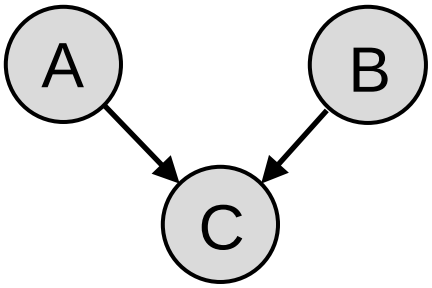
$$P(\theta \mid \mathcal{D}) = \mathrm{Beta}(\theta \mid k+\alpha,\, N-k+\beta)$$

Graphical models philosophy

Biology	Graph	Probabilistic model

Example: Gene regulation

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

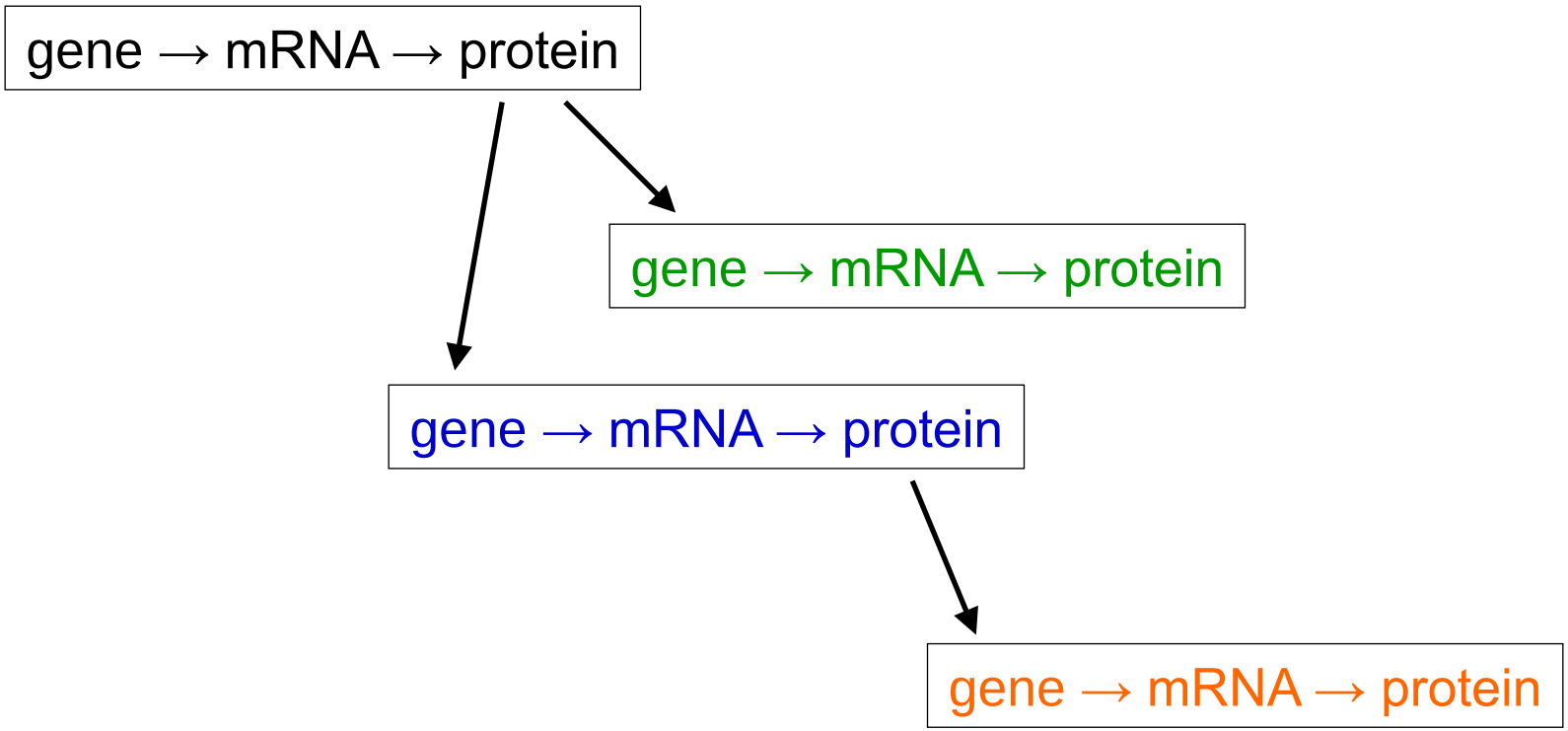


$$P(A,B,C) = P(A)\,P(B)\,P(C|A,B)$$

Biological players ⟷ Vertices ⟷ Random variables

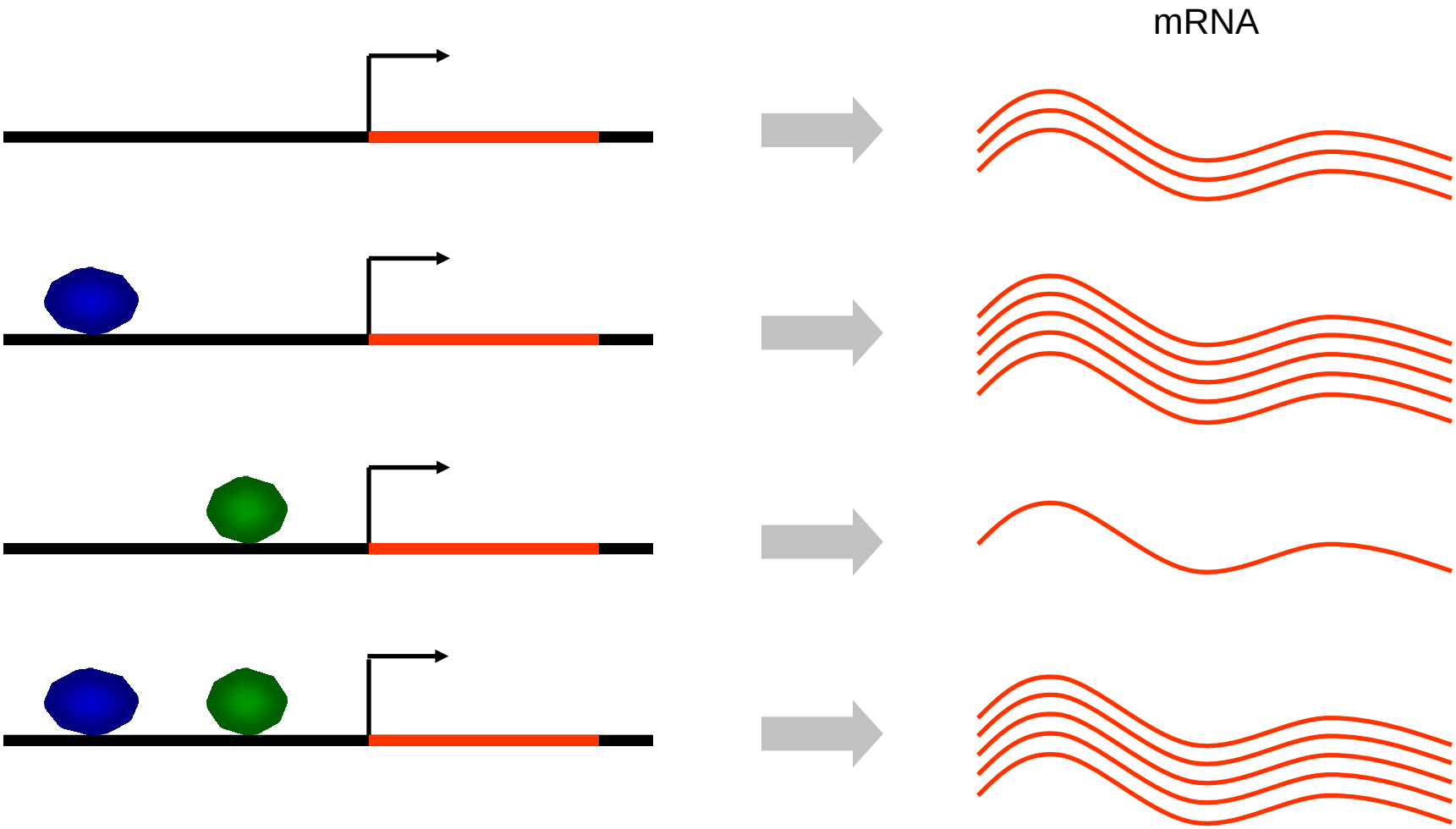
(Causal) Relationships ⟷ Edges ⟷ Statistical dependencies

Gene regulation



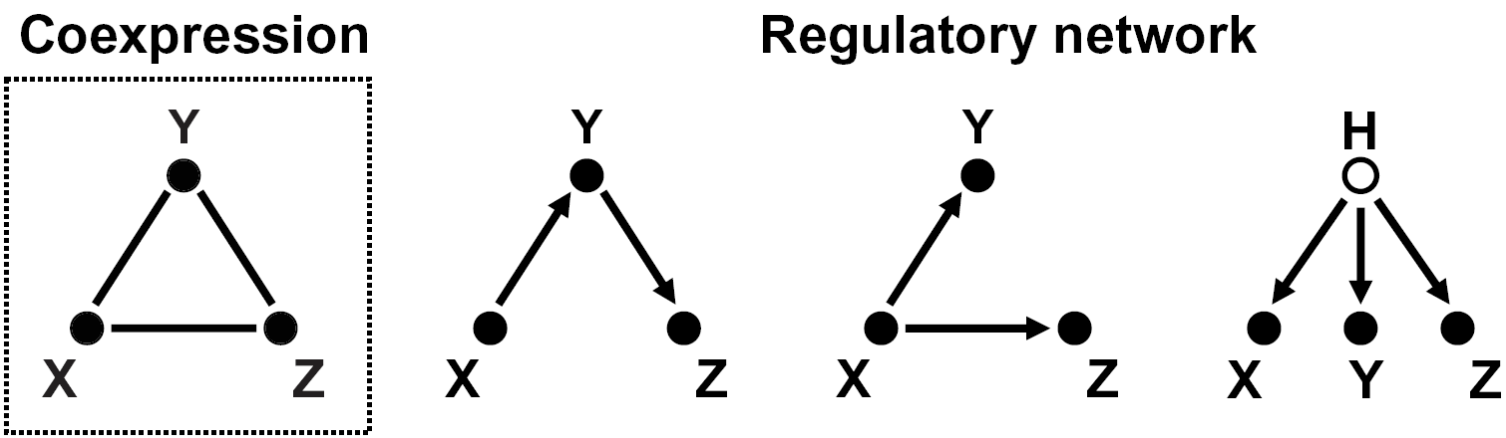
- Proteins can increase or decrease the rate of transcription of another gene by binding to the promoter region. These proteins are called *transcription factors*.

Transcriptional regulation



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!

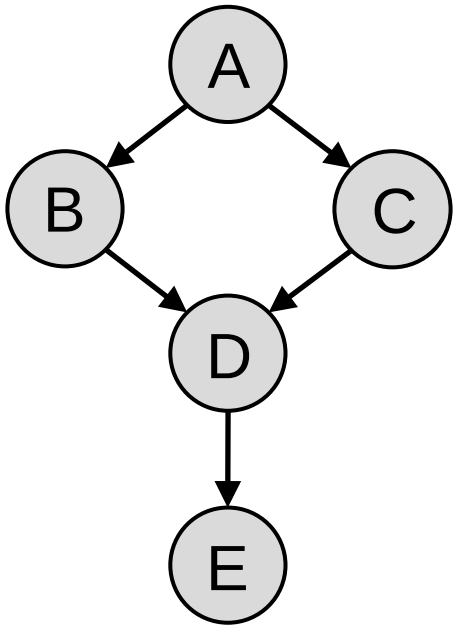
Bayesian networks

- A Bayesian network (BN) for $X = (X_1, \ldots, X_n)$ consists of
  - a directed acyclic graph (DAG) $G = (V, E)$, where $V = \{1, \ldots, n\}$
  - 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

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

where pa(i) denotes the set of parents of i in G, i.e., $X_{pa(i)} = (X_{j_1}, \ldots, X_{j_k})$ if $pa(i) = \{j_1, \ldots, j_k\}$ are the k parents of i in G.

Example



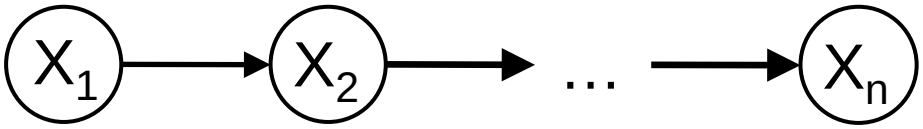
$$P(A, B, C, D, E) =$$
$$P(A)P(B \mid A)P(C \mid A)P(D \mid B, C)P(E \mid D)$$

Discrete variables

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

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

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



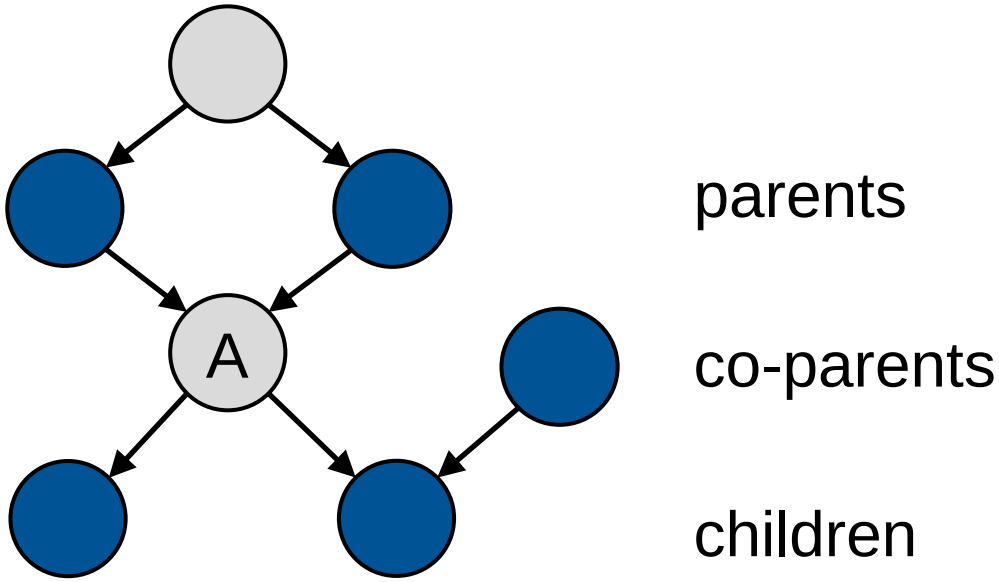
Linear Gaussian models

- Linear-Gaussian models are defined by the continuous LPD

$$P(X_i \mid X_{\mathsf{pa}(i)}) = \mathsf{Norm}(b_i + w_i^t \cdot X_{\mathsf{pa}(i)}, v_i)$$

  with parameters b_i, w_i specifying the mean, and variance v_i.
- There are recursive formulas for the expectation and covariance of $(X_1, \ldots, X_n)$.
- The number of parameters increases linearly with the number of parents.
- Only linear relationships can be modeled.

Markov blanket



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

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

Conditional independence

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

$$A \perp B \mid C$$

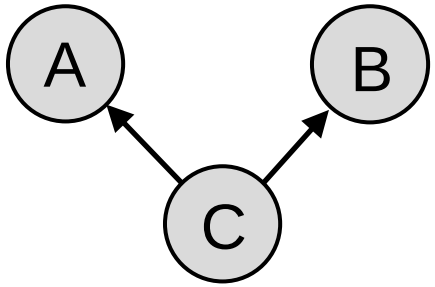
if $$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

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

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

$$A \perp B$$

Example



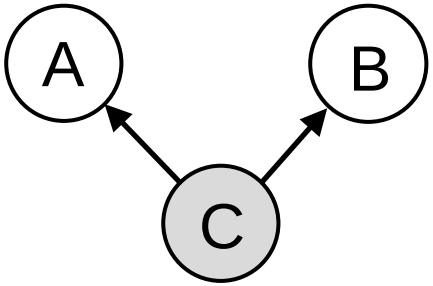
$$P(A,B,C) = P(A \mid C)P(B \mid C)P(C)$$

$$P(A,B \mid C) = \frac{P(A,B,C)}{P(C)} = P(A \mid C)P(B \mid C) \Rightarrow A \perp B \mid C$$

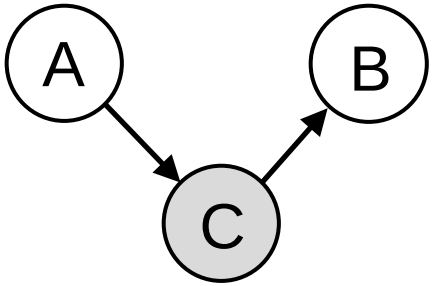
- However, in general, $A \not\perp B$

Three basic examples

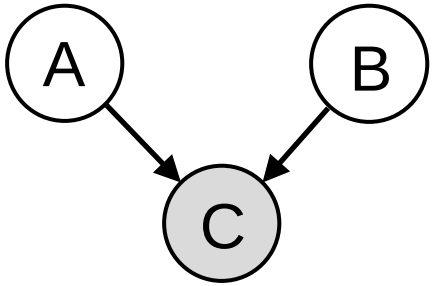
"explaining away"



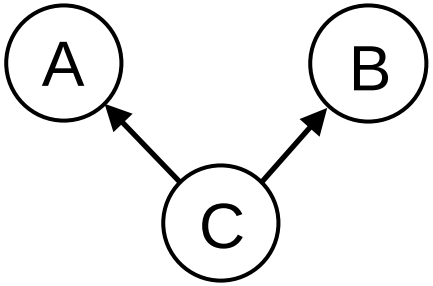
$$A \perp B \mid C$$



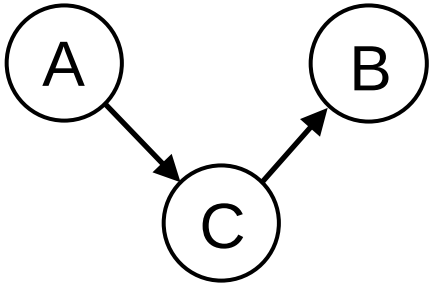
$$A \perp B \mid C$$



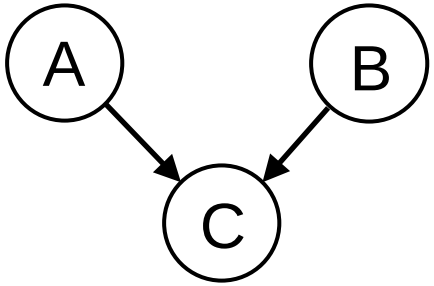
$$A \not\perp B \mid C$$



$$A \not\perp B$$



$$A \not\perp B$$



$$A \perp B$$

Inference

- Inference in graphical models refers to computing marginal probabilities:

$$P(X_i : i \in S) = \sum_{\{X_k : k \notin S\}} P(X_1, \ldots, X_n)$$

- For example,

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

- These computations can be organized efficiently along the structure of the (factor) graph, a procedure known as message passing or belief propagation.

Sum-product algorithm

- Basic idea: ab + ac = a(b + c), distributive law
- Example: Consider the chain $W \to X \to Y \to Z$:

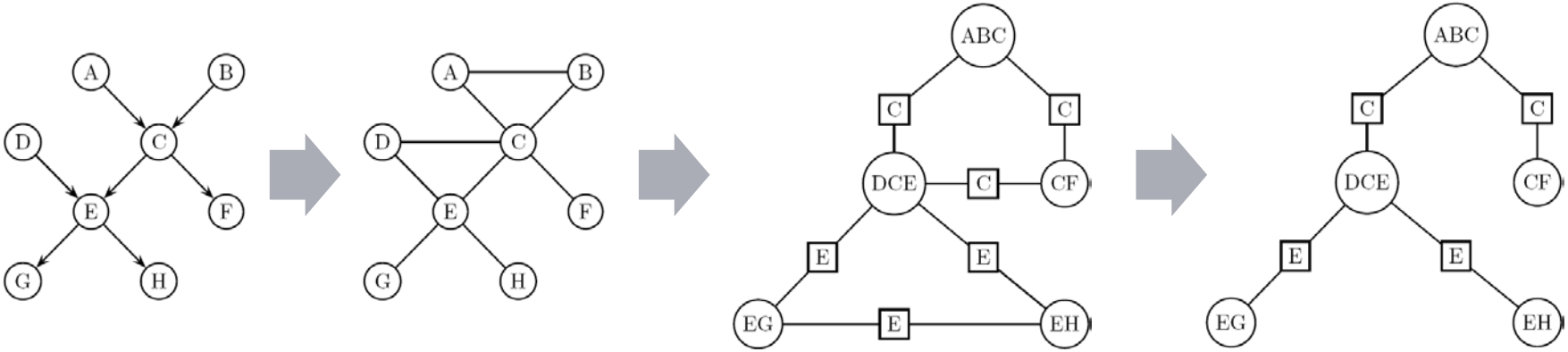
$$
\begin{aligned}
P(Z) &= \sum_{W,X,Y} P(W,X,Y,Z) \qquad O(K^n) \\
&= \sum_{W,X,Y} P(Z \mid Y) P(Y \mid X) P(X \mid W) P(W) \\
&= \sum_{X,Y} P(Z \mid Y) P(Y \mid X) \sum_{W} P(X \mid W) P(W) \\
&= \sum_{Y} P(Z \mid Y) \Big[\sum_{X} P(Y \mid X) \Big[\sum_{W} P(X \mid W) P(W) \Big] \Big]
\end{aligned}
$$

$O(nK^2)$

Junction tree algorithm

- In general, message passing is applied to the junction tree

DAG → moralize, triangulate → Junction tree → assign potentials, pass messages → Clique marginals



DAG | Moral graph | Junction graph | Junction tree

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,

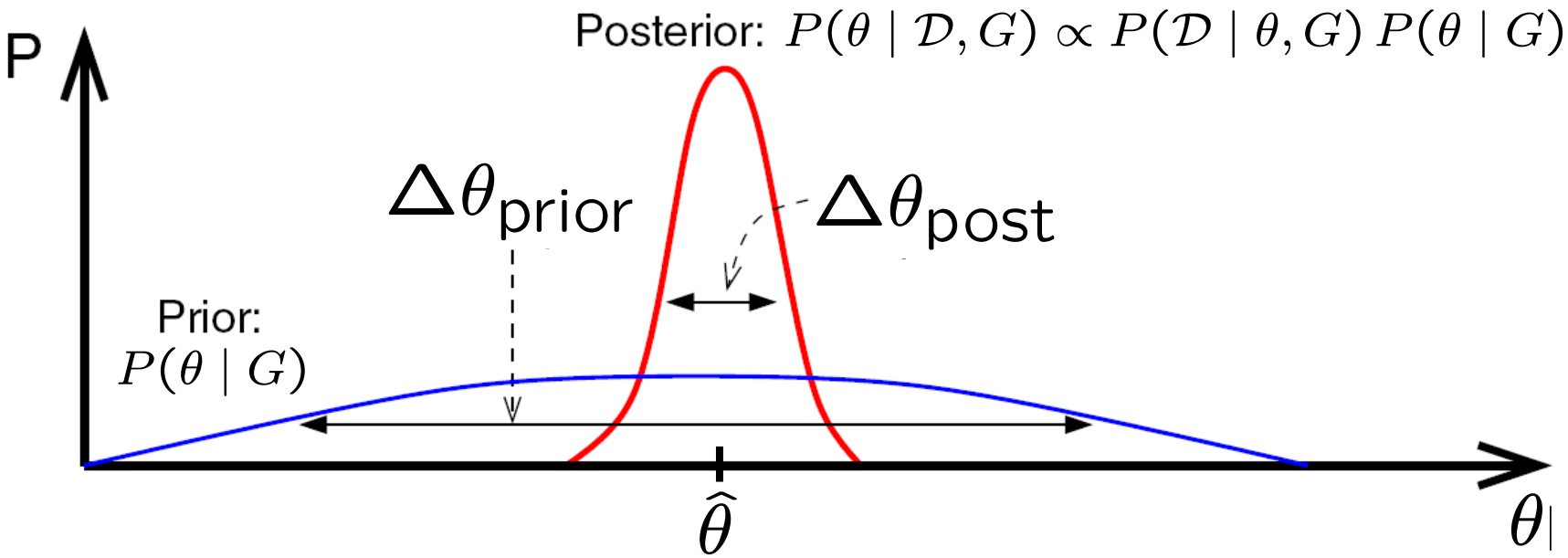
$$P(G \mid \mathcal{D}) \propto P(\mathcal{D} \mid G) P(G)$$

where

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

is the marginal likelihood.

Marginal likelihood: flat prior and unimodal posterior



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

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

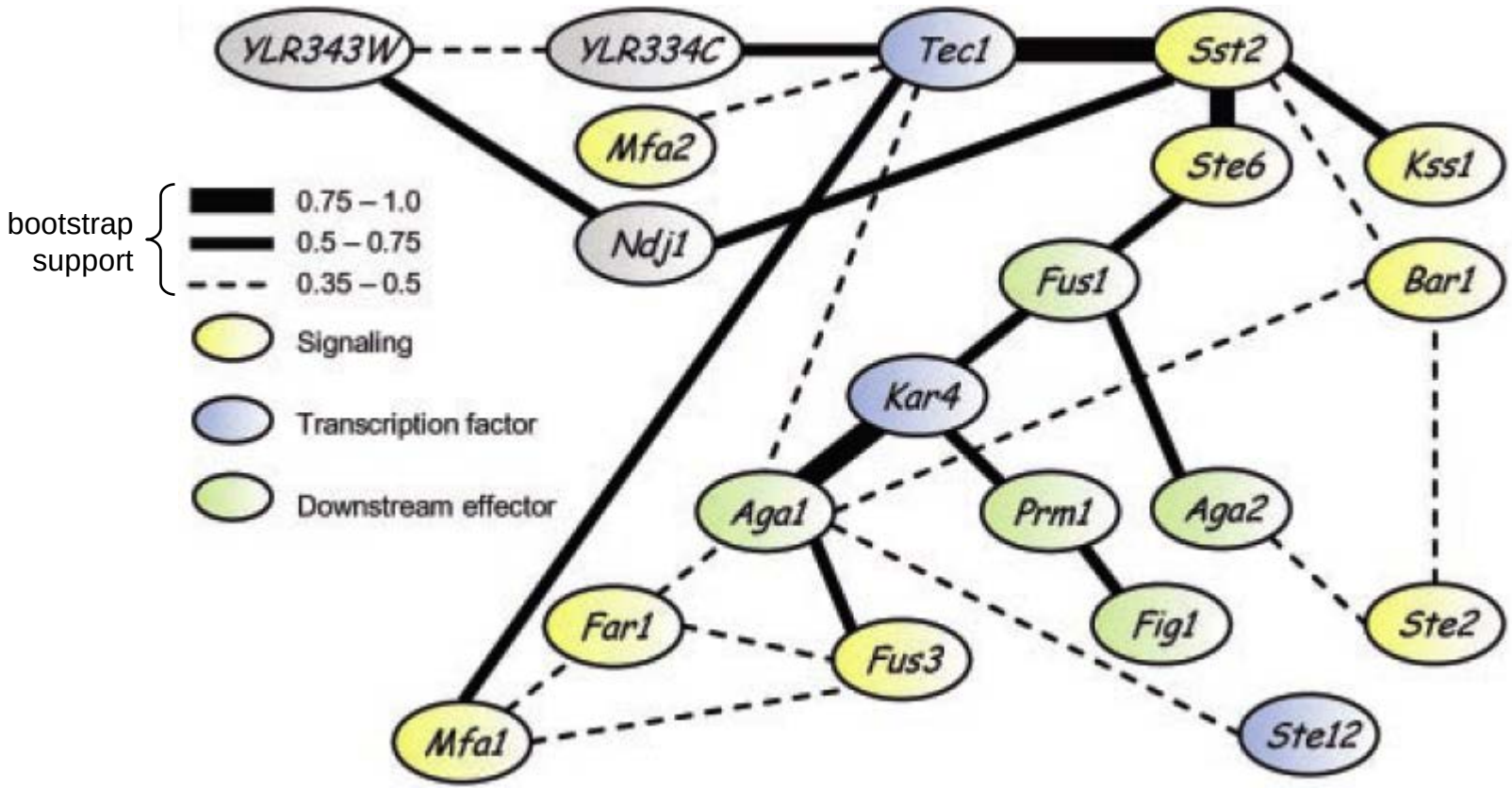
Bayesian information criterion (BIC)

- Under certain assumptions (including unimodal likelihood),

$$\log P(\mathcal{D} \mid G) \approx \underbrace{\log P(\mathcal{D} \mid \hat{\theta}, G)}_{\text{MLE score}} \underbrace{- \frac{\nu}{2} \log N}_{\text{penalty / regularization}}$$

  where ν is the number of free parameters of the model.
- The regularization term penalizes model complexity.

Example: Yeast mating pathway



Friedman (2004) Science 303:799

Bayesian learning of network structure

- MAP learning: $G^* = \underset{G}{\operatorname{argmax}}\, P(G \mid \mathcal{D})$
- Inference of the full posterior by sampling:

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

where

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

is the marginal likelihood

Markov Chain Monte Carlo (MCMC)

- Idea: Construct a Markov chain that converges to the true unknown distribution. Metropolis-Hastings:
- Start with a random DAG $G^{(0)}$
- For $n = 1, \ldots, N$
  - Generate a new DAG $G^{(n)}$ from a proposal distribution Q,

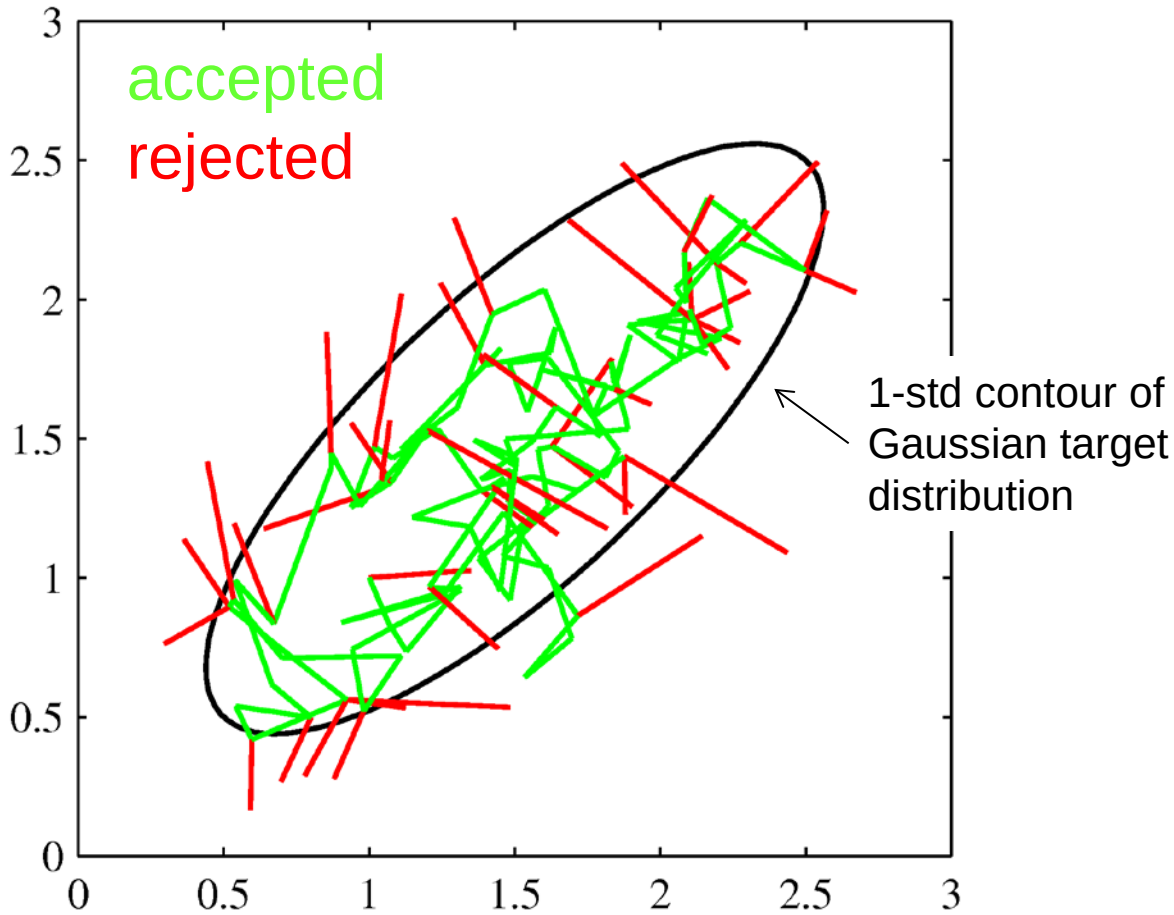
$$G^{(n)} \sim Q\left(G^{(n)} \mid G^{(n-1)}\right)$$

  - Accept the new graph with acceptance probability

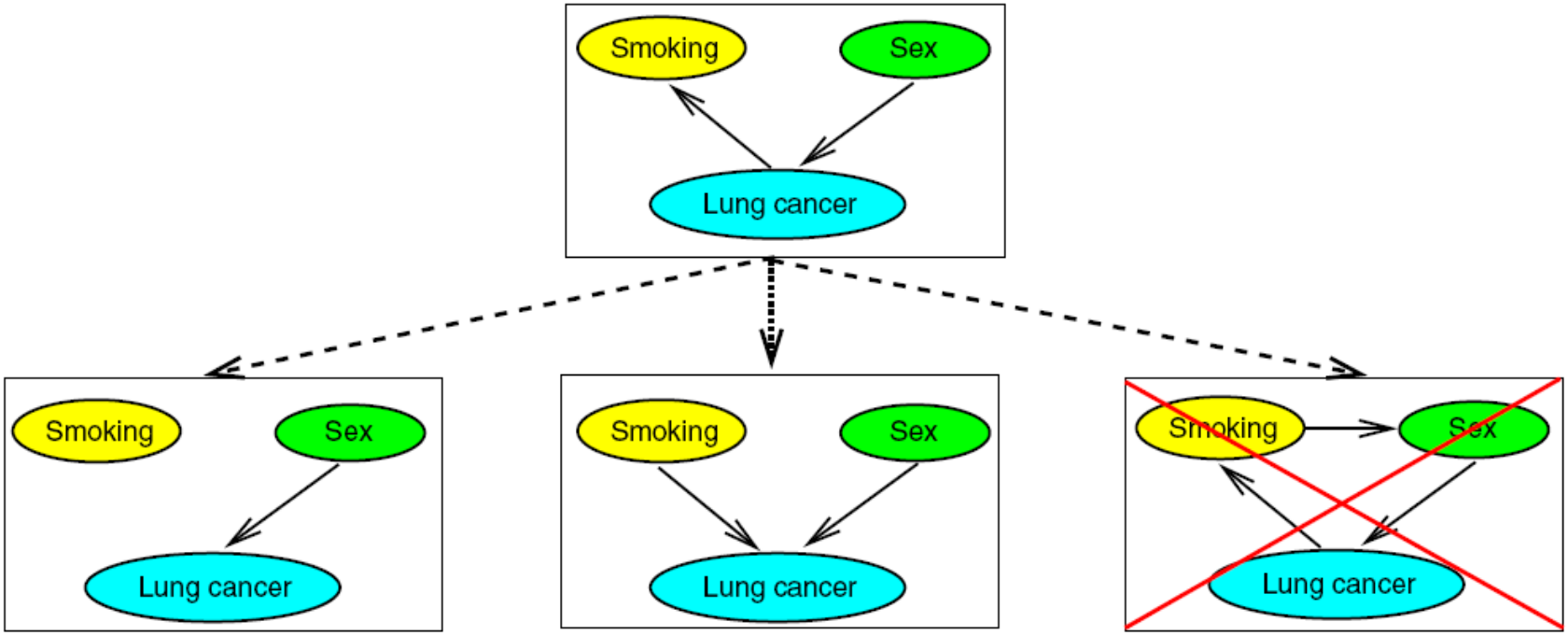
$$A\left(G^{(n)} \mid G^{(n-1)}\right) = \min\left\{\frac{P(\mathcal{D} \mid G^{(n)})P(G^{(n)})Q(G^{(n-1)} \mid G^{(n)})}{P(\mathcal{D} \mid G^{(n-1)})P(G^{(n-1)})Q(G^{(n)} \mid G^{(n-1)})}, 1\right\}$$

  otherwise, leave the value unchanged, $G^{(n)} = G^{(n-1)}$.
- Under certain conditions, this Markov chain converges to a stationary distribution, the target distribution $P(G \mid \mathcal{D})$

Example: Q = isotropic Gaussians



Elementary MCMC moves for DAGs

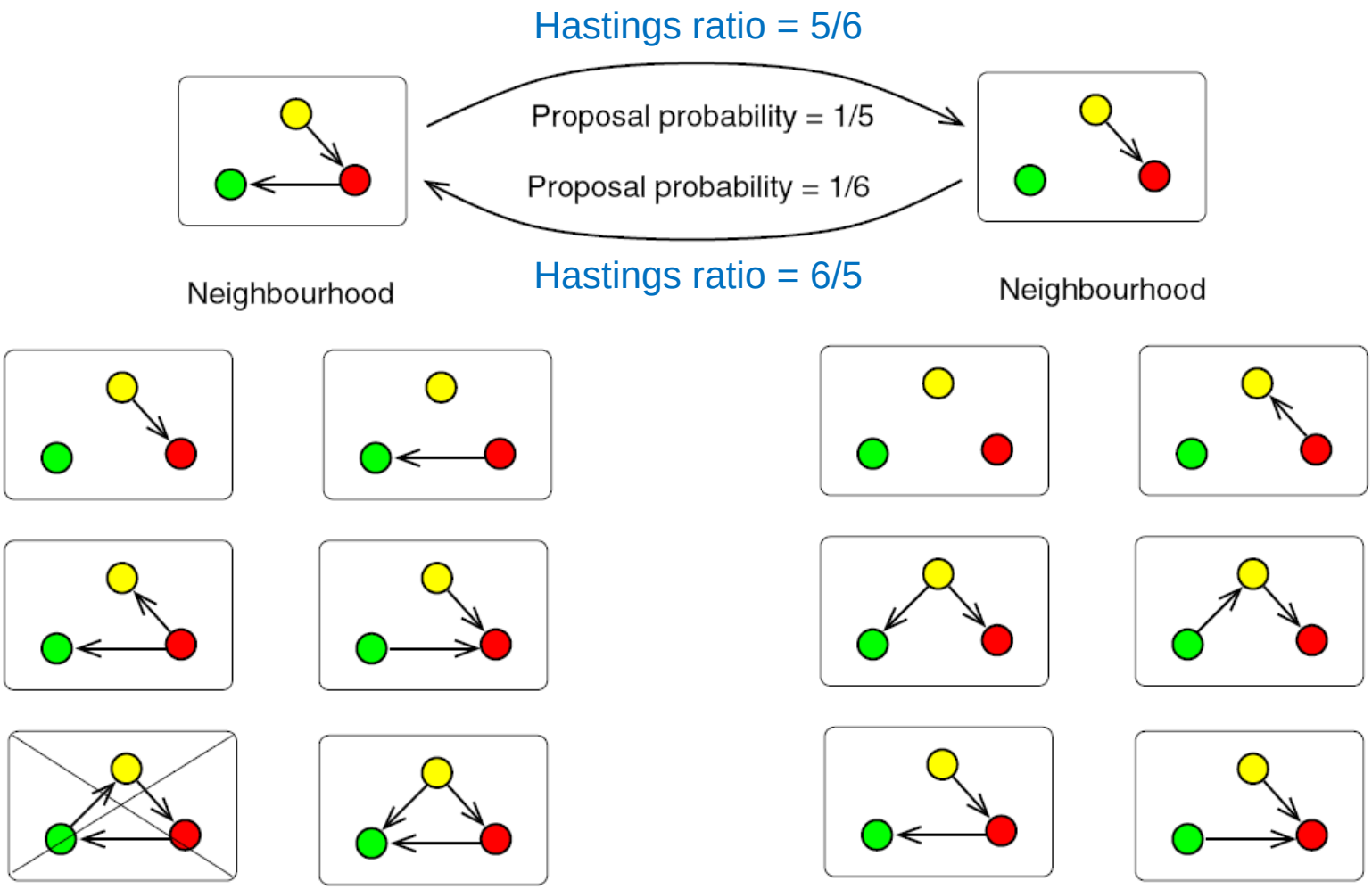


Remove edge

Reverse edge

Add edge

DAG neighborhoods and Hastings ratio



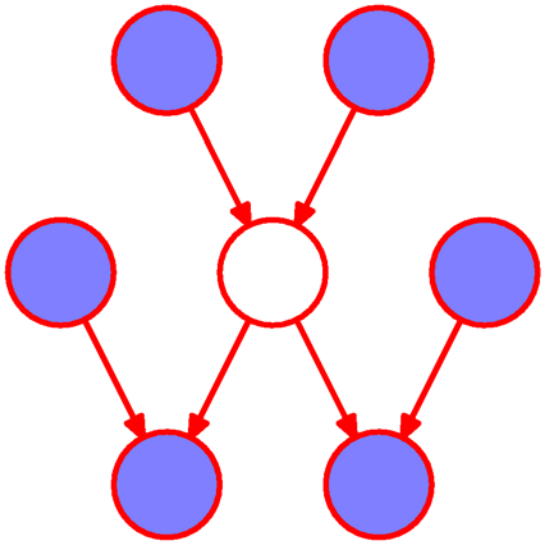
Gibbs sampling

- Sample conditional probabilities of $P(X_1, ..., X_M)$ iteratively:

$$
\begin{aligned}
X_1^{(n+1)} &\sim P\left(X_1 \mid X_2^{(n)}, \ldots, X_M^{(n)}\right) \\
X_2^{(n+1)} &\sim P\left(X_2 \mid X_1^{(n+1)}, X_3^{(n)}, \ldots, X_M^{(n)}\right) \\
&\vdots \\
X_j^{(n+1)} &\sim P\left(X_j \mid X_1^{(n+1)}, \ldots, X_{j-1}^{(n+1)}, X_{j+1}^{(n)}, \ldots, X_M^{(n)}\right) \\
&\vdots \\
X_M^{(n+1)} &\sim P\left(X_M \mid X_1^{(n+1)}, \ldots, X_{M-1}^{(n+1)}\right)
\end{aligned}
$$

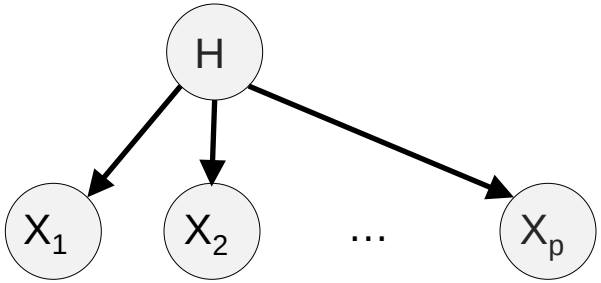
Gibbs sampling for graphical models

- Gibbs sampling is particularly useful, if it is much easier to sample from the conditionals $P(X_k \mid X_{\backslash k})$ than from the joint distribution $P(X_1, ..., X_M)$.
- For graphical models, $P(X_k \mid X_{\backslash k}) = P(X_k \mid X_{MB(k)})$.



Hidden (unobserved) variables

- Observed variables X_j, *hidden (unobserved) variables* H_k
- Example: Clustering (H module/functional pathway, X_j gene expression in functional context)



$$X_i \perp X_j \mid H$$

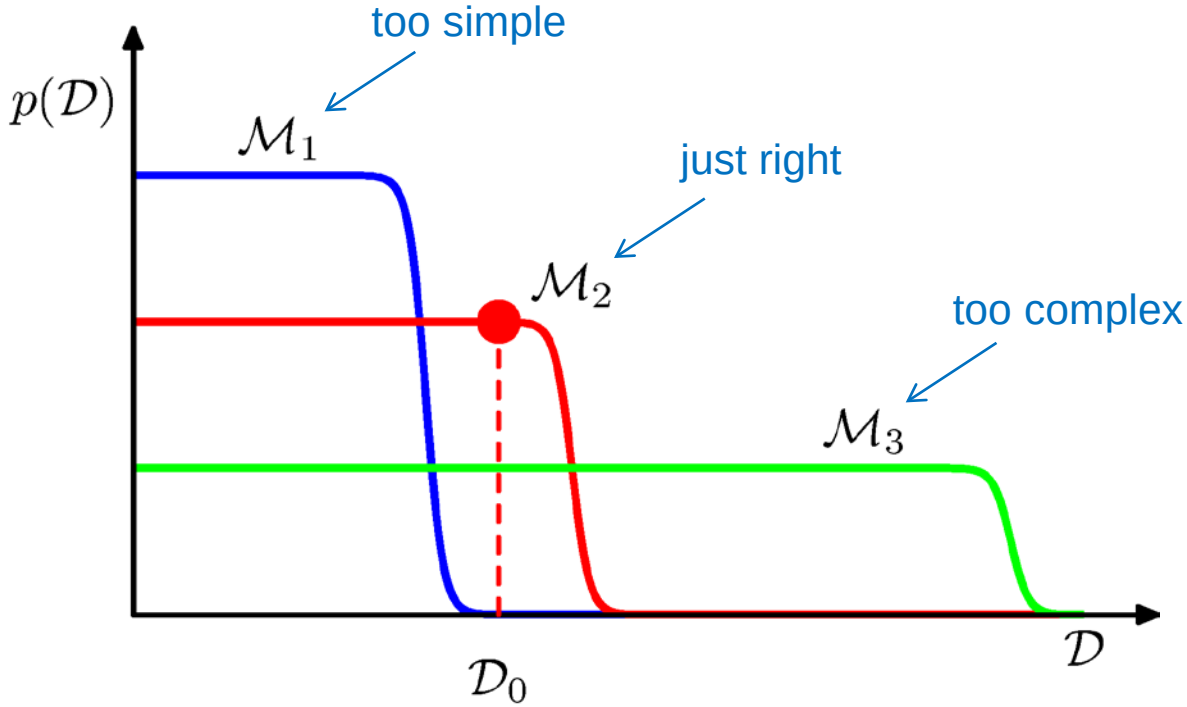
- Marginal distr.: $\Pr(X) = \sum_H \Pr(X_1 \mid H) \cdot \ldots \cdot \Pr(X_p \mid H) \Pr(H)$
- Other examples:
  - Pair hidden Markov model for sequence alignment
  - Phylogenetic trees

EM algorithm: MLE in the presence of hidden variables

$$
\begin{aligned}
\log L(\theta) &= \log \int_H P(X, H \mid \theta)\,\mathrm{d}H \\
&= \log \int_H q(H) \frac{P(X, H \mid \theta)}{q(H)}\,\mathrm{d}H \qquad \text{(for any } q\text{)} \\
&\geq \int_H q(H) \log \frac{P(X, H \mid \theta)}{q(H)}\,\mathrm{d}H \qquad \text{(Jensen's inequality)} \\
&= \mathrm{E}_q[\log P(X, H \mid \theta)] - \mathrm{E}_q[\log q(H)]
\end{aligned}
$$

- The Expectation Maximization (EM) algorithm iteratively maximizes this lower bound,
  - in the E step, w.r.t. q, $q = \Pr(H \mid X, \theta)$
  - in the M step, w.r.t. θ, $\theta = \operatorname{argmax}_\eta L(\eta \mid X, H)$

Model complexity



Data sets

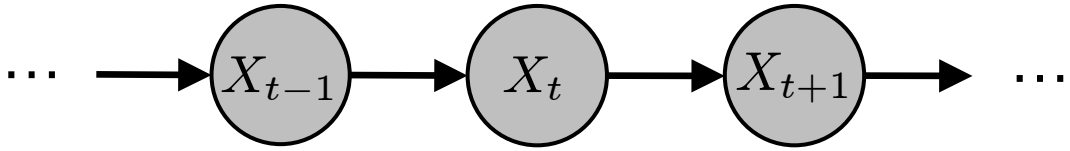
Learning Bayesian networks

	Fully observed data	**Missing data / hidden variables**
Known graph structure	Sample statistics	EM algorithm Gradient ascent MCMC Variational inference
Unknown graph structure	Search-and-score (BIC) PC algorithm MCMC	Structural EM MCMC

Learning Bayesian networks

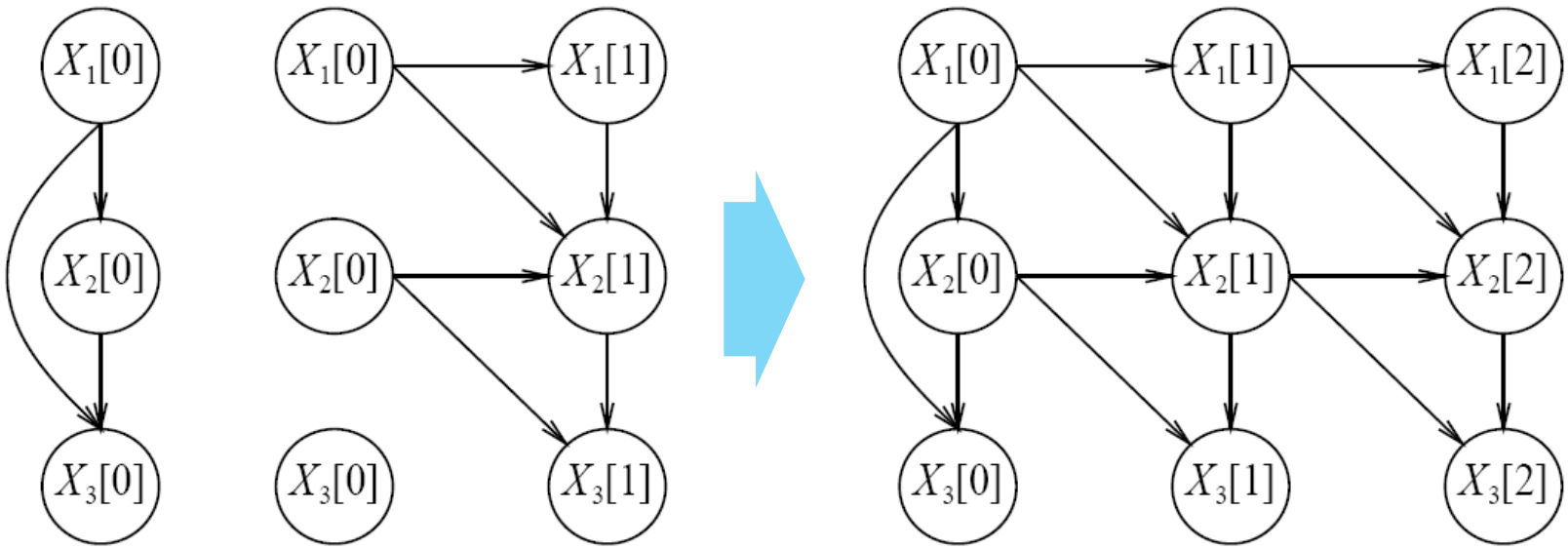
	Fully observed data	**Missing data / hidden variables**
Known graph structure	Sample statistics easy	EM algorithm Gradient ascent MCMC Variational inference doable
Unknown graph structure	Search-and-score (BIC) PC algorithm MCMC doable	Structural EM MCMC hard

Dynamic Bayesian network (DBN)



- A DBN represents random variables evolving over time.
- $X_{t+1} \perp X_{t-1} \mid X_t$ (Markov property)
- The random variables $\{X_t\}$ can be discrete or continuous.
- In general, X_t is multivariate and transitions are modeled by a Bayesian network. Thus, the DBN is an “unrolled BN”.
  - Sparse (factored) representation of states
  - Sparse transition matrices
- There can be hidden variables.

Definition of DBN



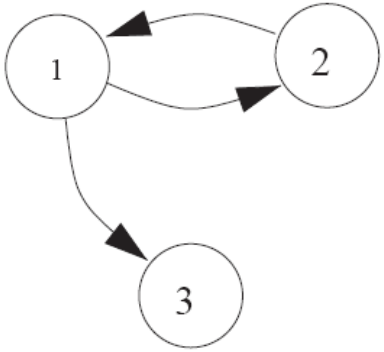
G_0, prior network

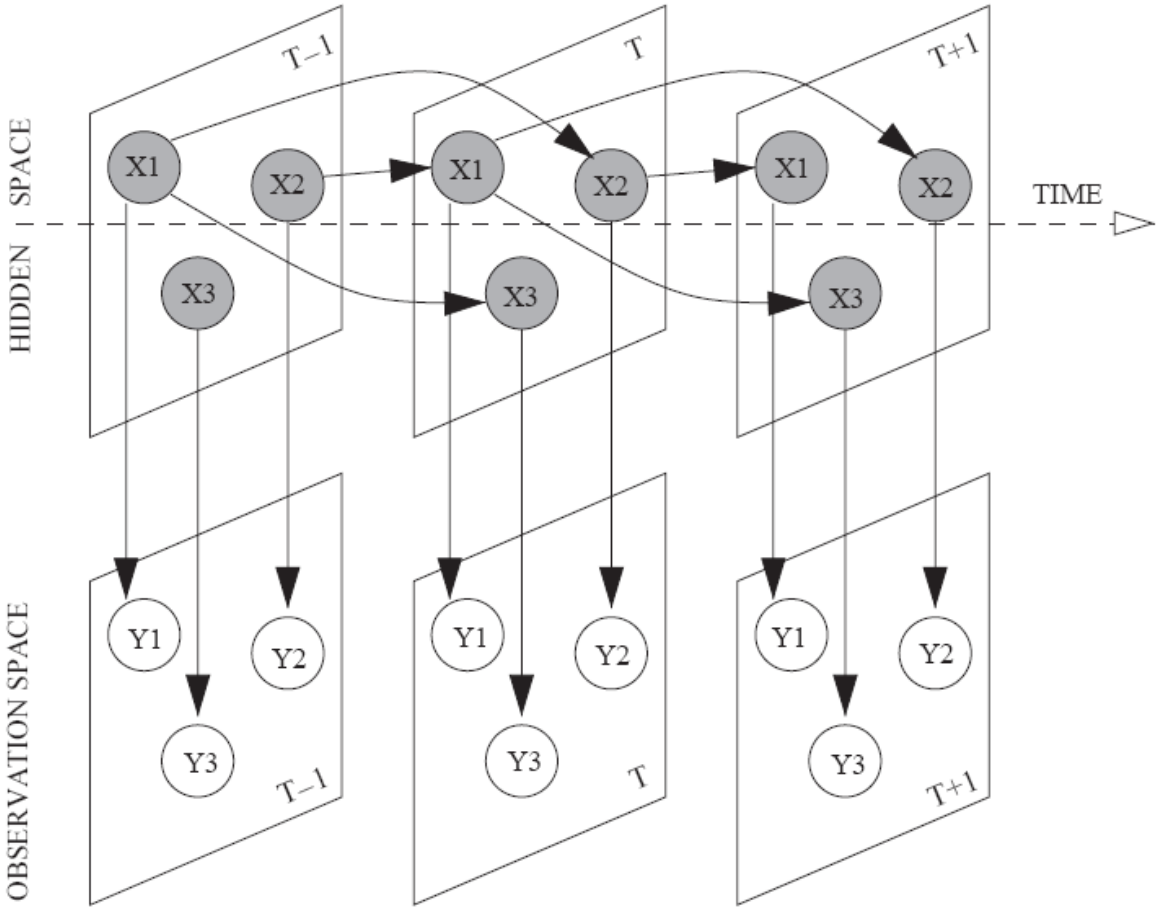
$G_\rightarrow$, transition network

G, DBN (unrolled network)

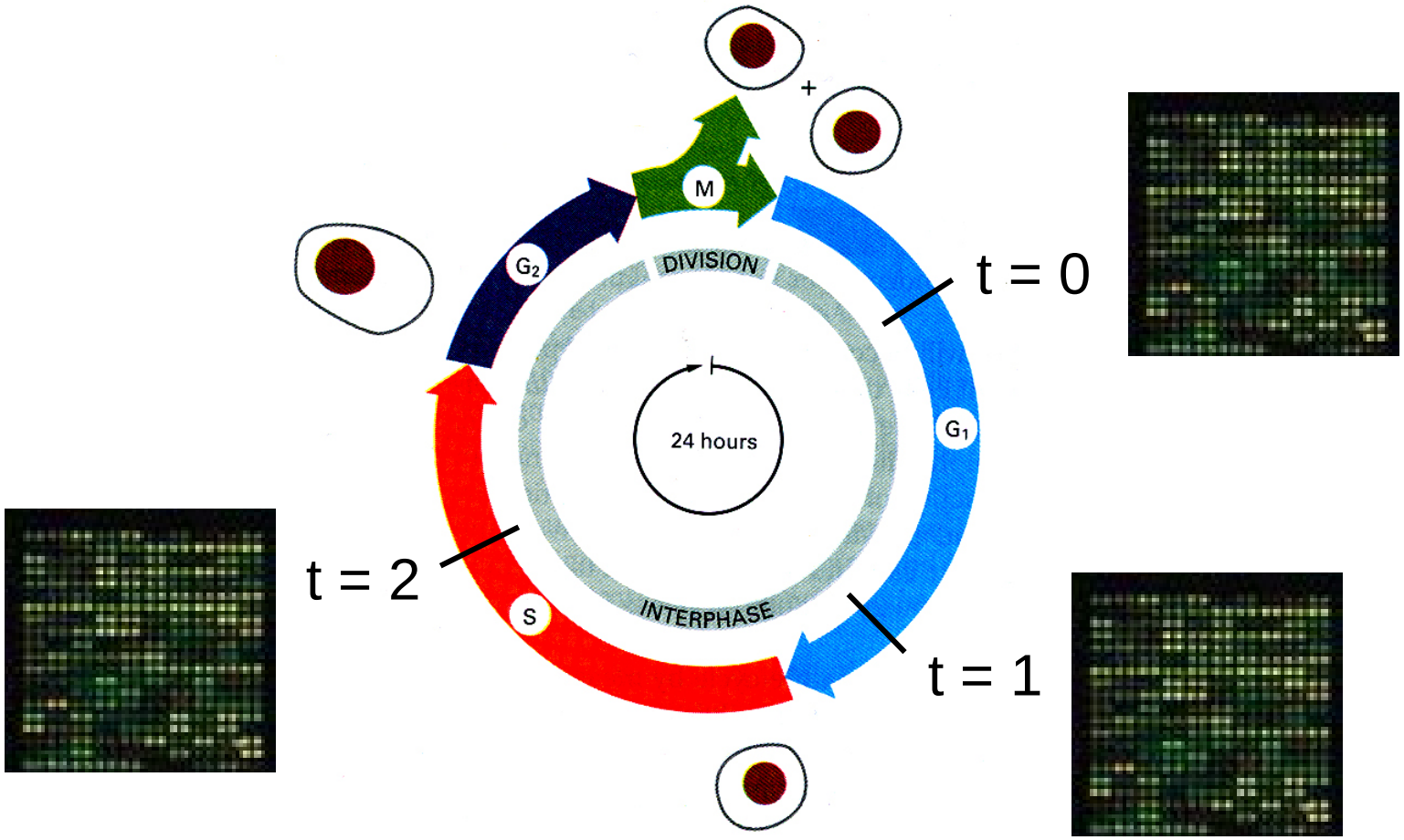
$$P(X[0],\ldots,X[T]) = P_0(X[0]) \prod_{t=0}^{T-1} P_\rightarrow(X[t+1] \mid X[t])$$

The DBN can resolve feedback loops



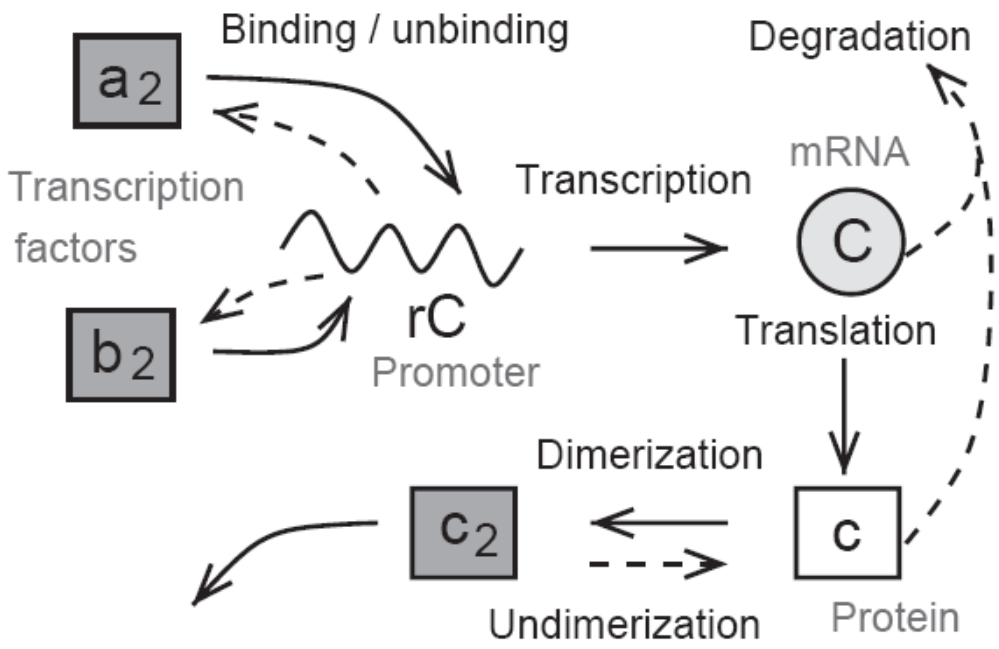


Cell cycle: gene expression time series

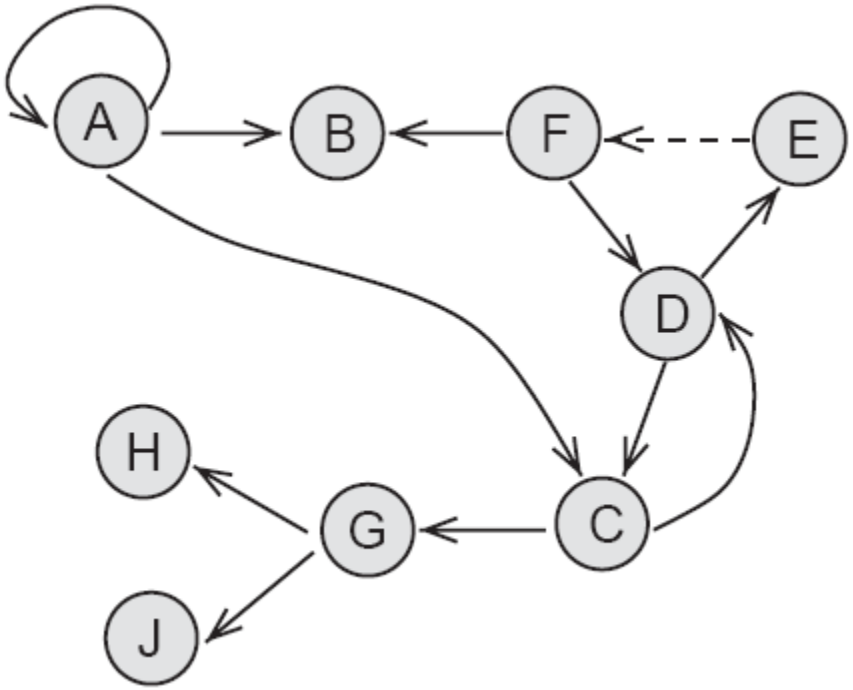


Simulation study

- Elementary processes
  - Transcription factor binding to promoter sequence
  - Transcription
  - Translation
  - Dimerization



Induced mRNA network



ODE model

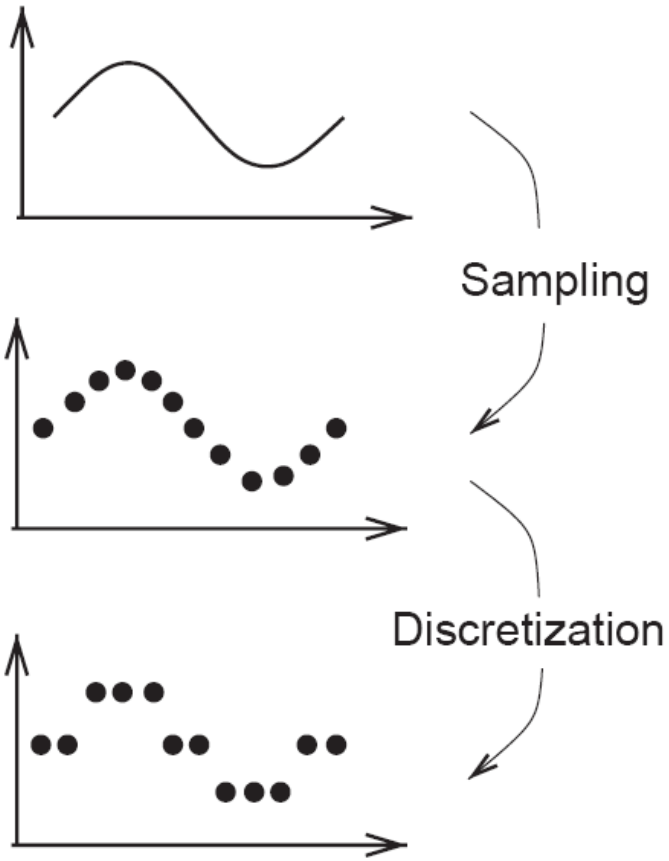
$$\frac{\mathrm{d}}{\mathrm{d}t}[a_2 \cdot rC] = \lambda^{+}_{a_2 \cdot rC}[a_2][rC] - \lambda^{-}_{a_2 \cdot rC}[a_2 \cdot rC],$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[C] = \lambda_{rC}[rC] + \lambda_{a_2 \cdot rC}[a_2 \cdot rC] + \lambda_{b_2 \cdot rC}[b_2 \cdot rC] - \lambda_C[C],$$

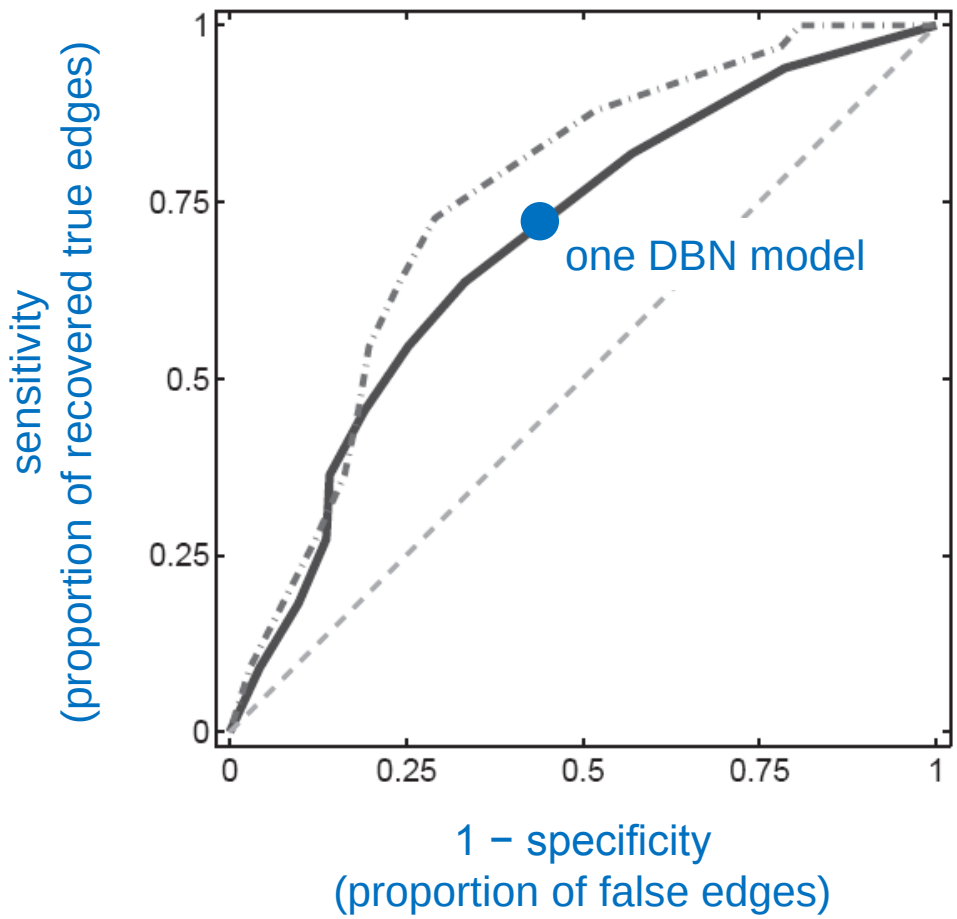
$$\frac{\mathrm{d}}{\mathrm{d}t}[c] = \lambda_{Cc}[C] - \lambda_c[c], \quad \frac{d}{dt}[c_2] = \lambda^{+}_{cc}[c]^2 - \lambda^{-}_{cc}[c_2]$$

Sampling and discretization

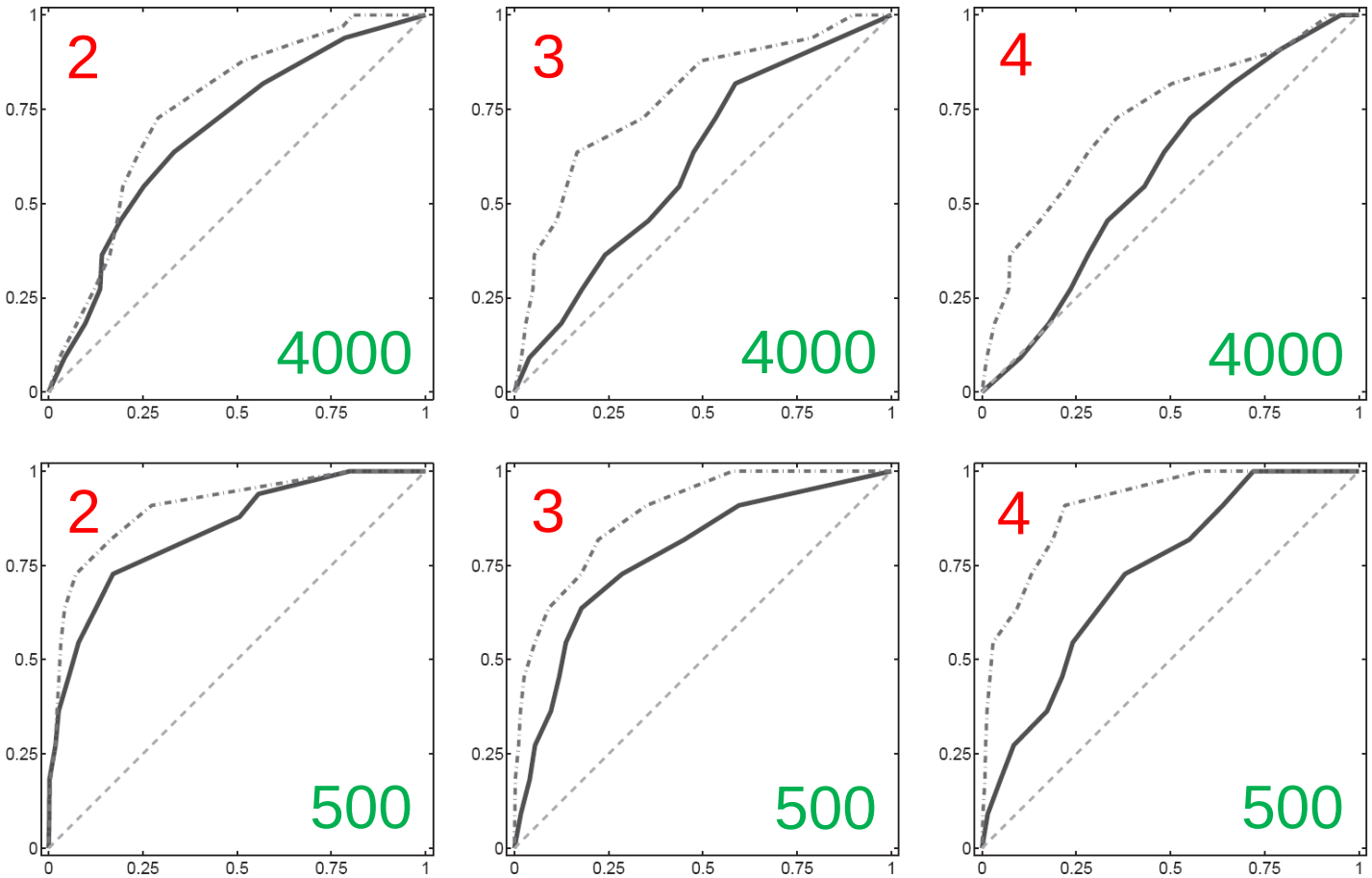
- First experiment:
  - Collect 12 data points over 4000 min after ligand injection
- Second experiment:
  - Collect 12 data points over 500 min after ligand injection
- Use MCMC (Metropolis-Hastings) to sample from P(G | $\mathcal{D}$).
- Different priors restricting the number of incoming edges ("fan-in") are tested.



ROC curve


sensitivity
(proportion of recovered true edges)
1
0.75
0.5
0.25
0
0
0.25
0.5
0.75
1
one DBN model
1 − specificity
(proportion of false edges)

DBN Performance

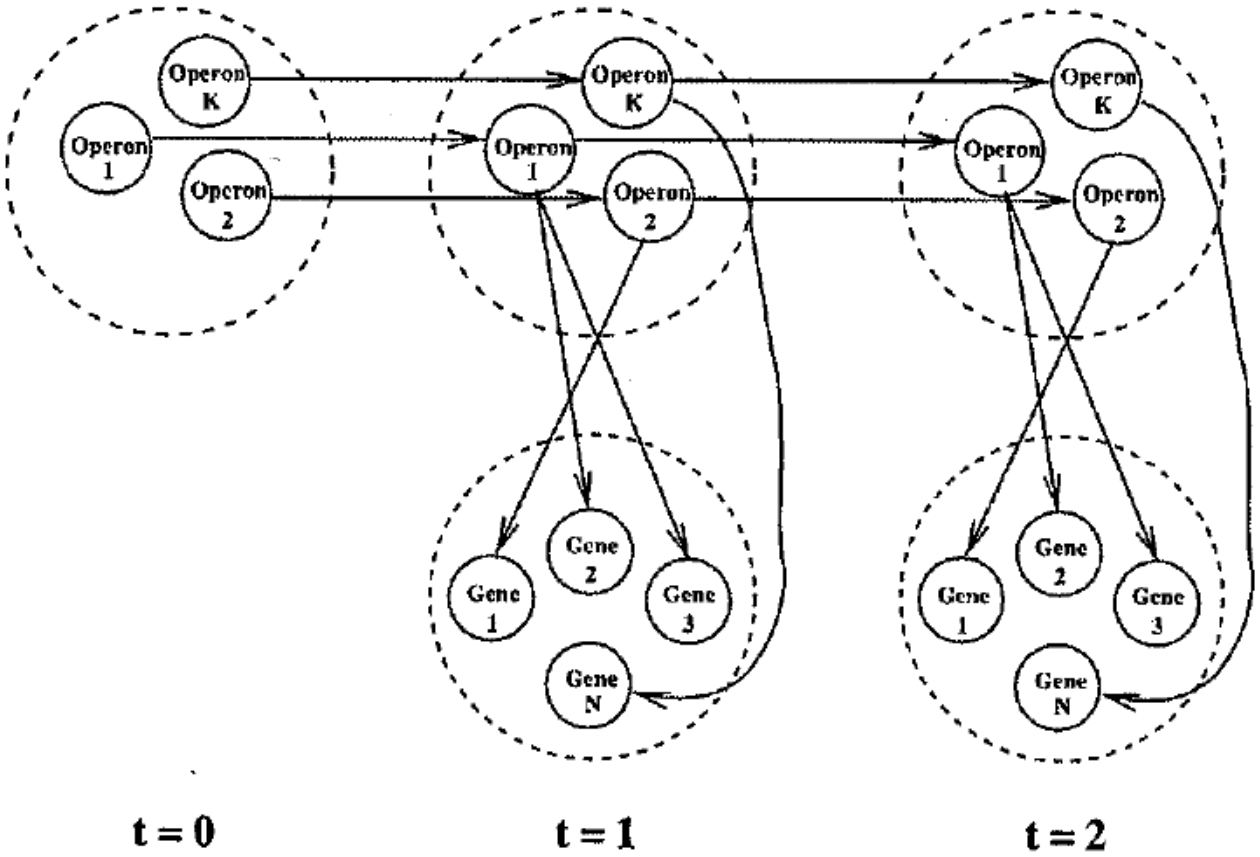


Max. fan-in

Minutes after ligand infection

Additional sequence-based model component

Regulatory pathways in E. coli



References

- Beerenwinkel N, 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
- Books:
  - Bishop CM. Pattern Recognition and Machine Learning.
  - Husmeier D, Dybowski R, Roberts S (eds.). Probabilistic Modeling in Bioinformatics and Medical Informatics.
  - Koller D, Friedman N. Probabilistic Graphical Models.
  - Darwiche A. Modeling and Reasoning with Bayesian Networks.
  - Neapolitan RE. Probabilistic Methods for Bioinformatics (with an Introduction to Bayesian Networks)
- Software
  - Murphy K. Bayes Net Toolbox for MATLAB, http://code.google.com/p/bnt/
  - gR – gRaphical Models in R, http://cran.r-project.org/web/views/gR.html