Learning Bayesian Networks

Luigi Portinale

University of Piemonte Orientale, Italy

March 7-10, 2017

Introduction

Learning in PGM: issues

- *Parameter Learning*: to learn the quantitative (probabilistic) part of the net
- *Structure Learning*: to learn the qualitative (graph) part of the net
- *Missing Data*: how to deal with data that are missing, especially if there are *latent variables* (i.e., variables that are never observed)

We mainly focus on BN learning

Estimating parameter for binary variables

Suppose θ is the parameter to be estimated: e.g., given a binary variable X with no parent $\theta = \mathcal{P}(X = T)$ or given a variable X with a parent Y, then $\theta = \mathcal{P}(X = T|Y = T)$.

Common assumptions: observed data come from a binomial distribution.

Suppose to observe N successes in N trials:

Maximum Likelihood Estimation

$$\hat{\theta} = \frac{M}{N}$$

Bayesian Estimation (p prior, γ prior's confidence)

$$\hat{\theta} = \frac{\gamma p + M}{\gamma + N}$$

- MLE corresponds to a frequentist view
- BE corresponds to a subjective view
- Confidence γ is also called *equivalent sample size*:
- Prior p can be viewed as the frequency of "hypothetical" m successes over "hypothetical" n trials (before data are observed)
- the confidence is given by $\gamma = n$: the larger is the "hypothetical" sample size, the greater is the confidence in the prior
- it easy to see that $\hat{\theta} = \frac{\gamma p + M}{\gamma + N} = \frac{m + M}{n + N}$
- BE is like MLE *including* the hypothetical samples

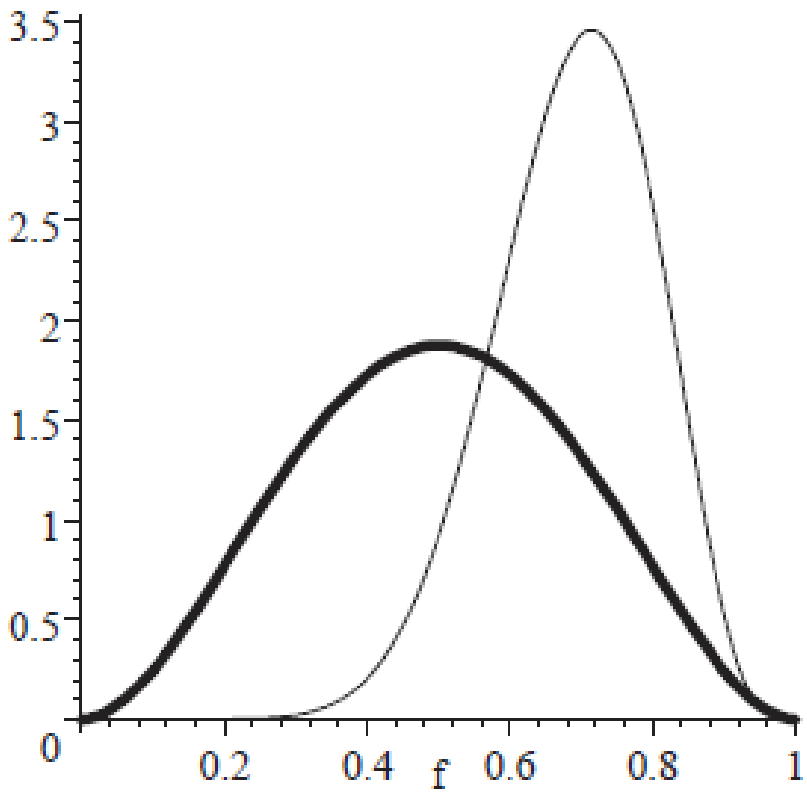
Beta Distribution

$$f_B(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

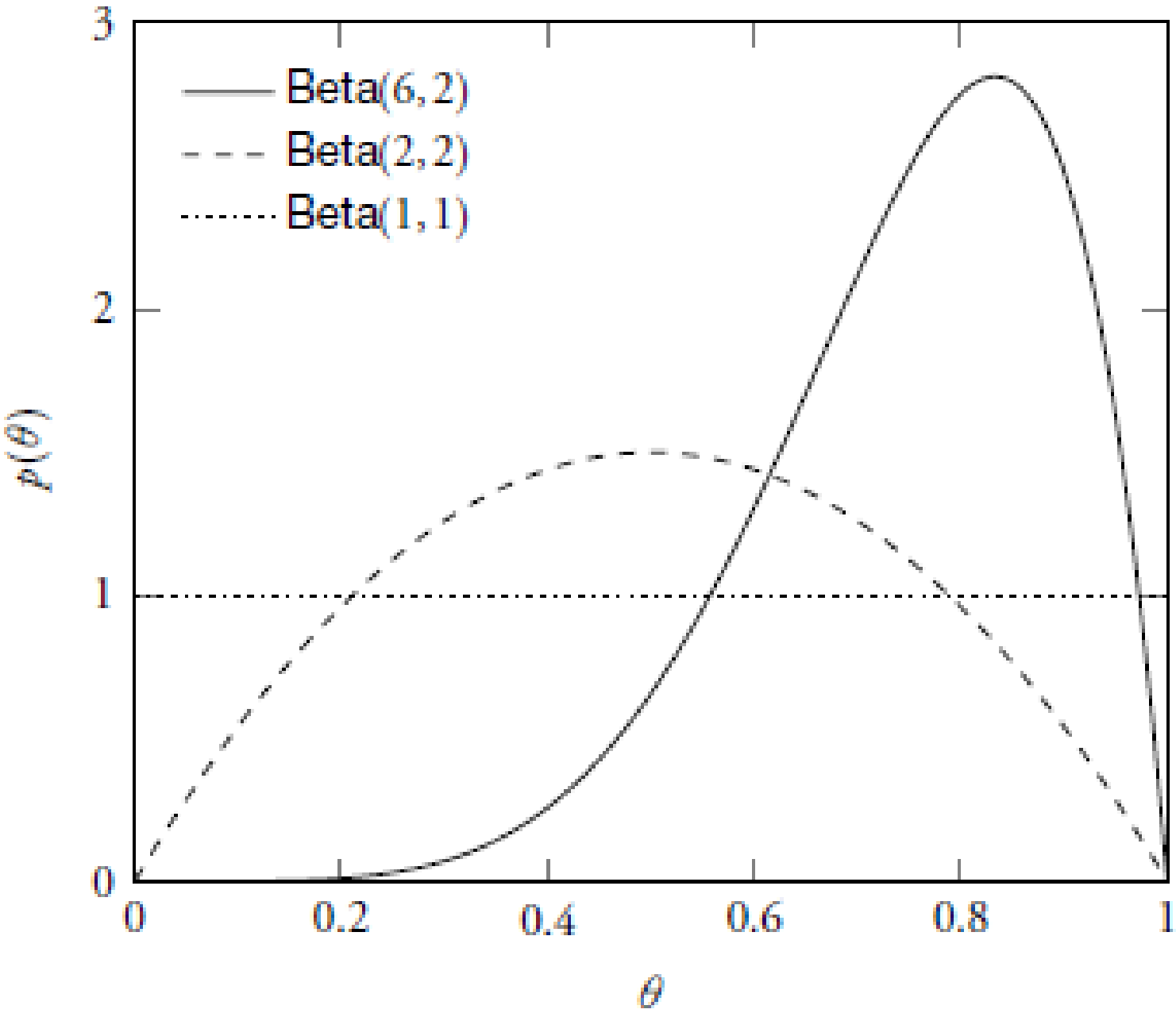
where $\Gamma(n) = (n-1)!$ when n is integer

- can be used to model binary priors when reference distribution is binomial
- $\mathbf{\Theta} = (\Theta_1, \Theta_2) \sim Beta(\alpha, \beta)$ if Θ_1 has density $f_B(\theta, \alpha, \beta)$ and $\Theta_2 = 1 - \Theta 1$
- $E(\Theta_1) = \int_0^1 \theta f_B(\theta; \alpha, \beta) d\theta = \frac{\alpha}{\alpha+\beta}$
- it is a *conjugate prior* for the binomial distribution
- if you have a prior $Beta(\alpha, \beta)$ and actually observe data D with M successes on N trials, then

$$f_B(\theta; \alpha, \beta | D) = f_B(\theta; \alpha + M, \beta + N - M)$$



The thickly plotted density function is $beta(\theta; 3, 3)$ and represents our prior belief concerning the relative frequency of heads. The thinly plotted one is $beta(\theta; 11, 5)$, and represents our posterior belief after we have seen 8 heads in 10 trials.



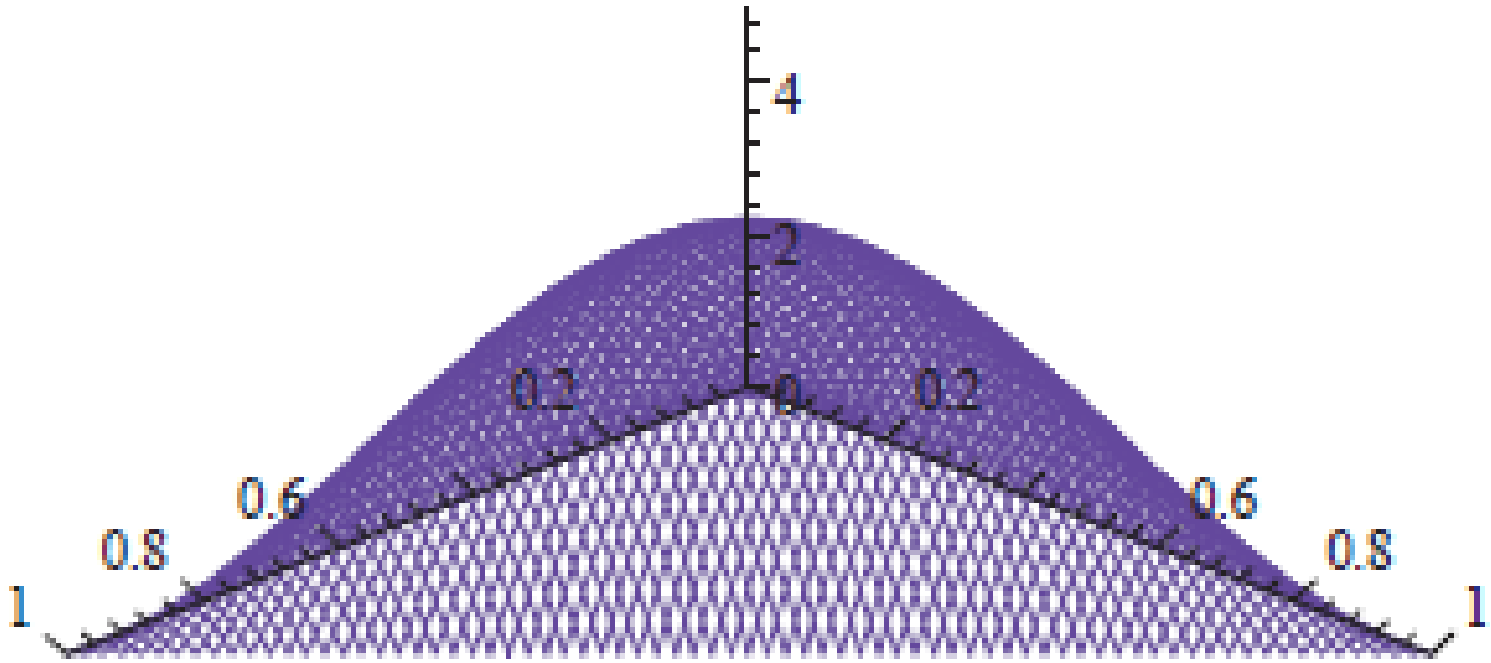
Dirichlet Distribution

$$f_D(\theta_{1\ldots k-1};\alpha_{1\ldots k}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$$

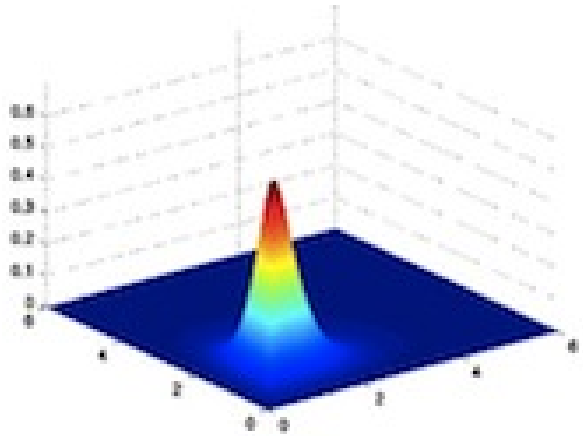
where $\alpha_0 = \sum_{i=1}^{k}\alpha_i$; $\theta_i > 0$ $(i = 1\ldots k)$; $\sum_{i=1}^{k}\theta_i = 1$

- $\boldsymbol{\Theta} = (\Theta_1 \ldots \Theta_k) \sim Dir(\alpha_{1\ldots k})$ if first $k-1$ r.v. have density f_D and $\Theta_k = 1 - \sum_{i=1}^{k-1}\Theta_i$
- used as a prior for multinomial distribution
- $E[\Theta_i] = \frac{\alpha_i}{\alpha_0}$
- it is a conjugate distribution for multinomial data
- given a prior $Dir(\alpha_{1,\ldots k})$ and a multinomial data sample D with M_i occurences of the i-th result over $N = \sum_{i=1}^{k} M_i$ trials, then

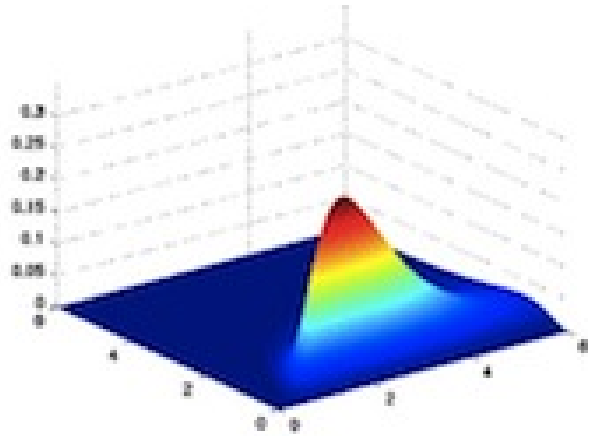
$$f_D(\theta_{1\ldots k-1};\alpha_{1\ldots k}|D) = f_D(\theta_{1\ldots k-1};\alpha_1 + M_1, \ldots \alpha_k + M_k)$$



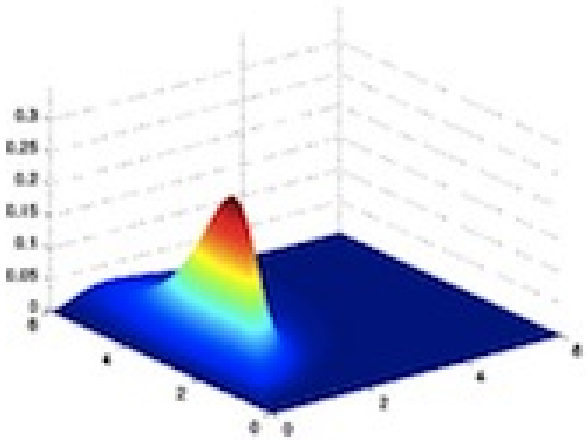
$Dir(\theta_1, \theta_2, \theta_3; 2, 2, 2)$: the density $f_D(\theta_1, \theta_2; 2, 2, 2)$



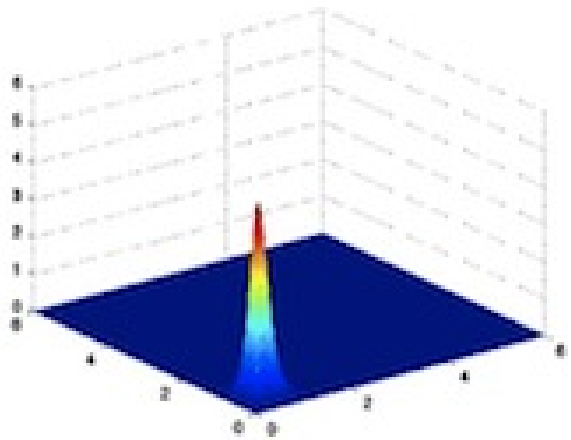
(a) $\alpha_1 = 3.5, \alpha_2 = 3.5, \alpha_3 = 3.5$



(b) $\alpha_1 = 10, \alpha_2 = 3.5, \alpha_3 = 3.5$

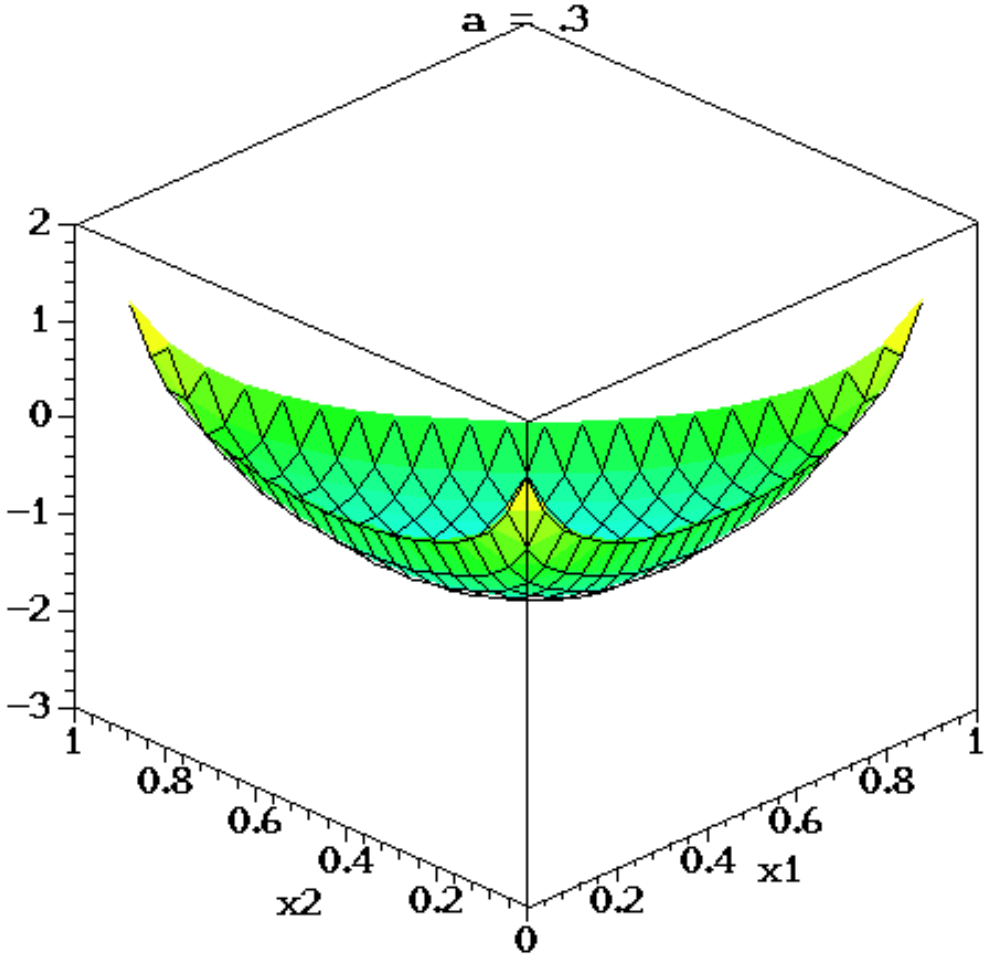


(c) $\alpha_1 = 3.5, \alpha_2 = 10, \alpha_3 = 3.5$



(d) $\alpha_1 = 3.5, \alpha_2 = 3.5, \alpha_3 = 10$

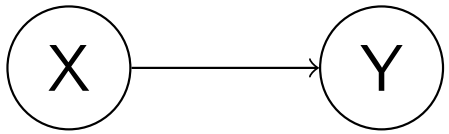
- When all parameters are equal the distrubution is called *symmetric*; it is characterized by parameter α called *concentration parameter*
- $\alpha = 1$ means uniform prior among all the possible outcomes (states of the variable)
- $\alpha > 1$ means dense random variates (i.e., all the values within a single sample are similar to each other)
- $\alpha < 1$ means sparse random variates (i.e., most of the values within a single sample will be close to 0, and the vast majority of the mass will be concentrated in a few of the values)



Illustrating how the log of the density function changes when $k = 3$ as we change the concentration parameter from 0.3 to 2.0.

Example

$P(X = t) = 0.2$



$P(Y = t|X = f) = 0.1$
$P(Y = t|X = t) = 0.7$

Equivalent sample sizes: $\gamma_X = 1000; \gamma_{Y|f} = \gamma_{Y|t} = 500$

$n_1 = \#(X = f, Y = t) = 300$
$n_2 = \#(X = t, Y = t) = 1000$
$n_3 = \#(X = t, Y = f) = 200$
$n_4 = \#(X = f, Y = f) = 500$

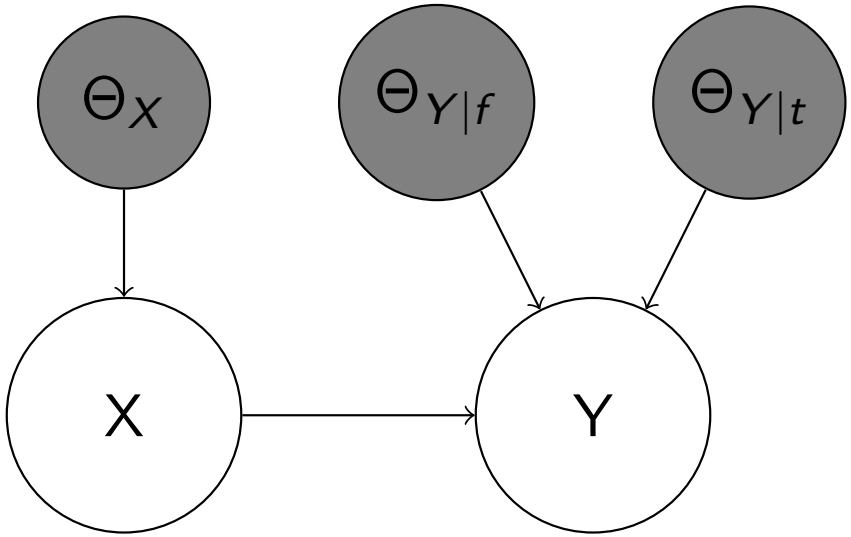
$N \Rightarrow 2000$

$P(X = t|D) = \frac{\gamma_X 0.2 + n_2 + n_3}{\gamma_X + N} = 0.47$

$P(Y = t|X = f, D) = \frac{\gamma_{Y|f} 0.1 + n_1}{\gamma_{Y|f} + n_1 + n_4} = 0.27$

$P(Y = t|X = t, D) = \frac{\gamma_{Y|t} 0.7 + n_2}{\gamma_{Y|t} + n_2 + n_3} = 0.79$

Example



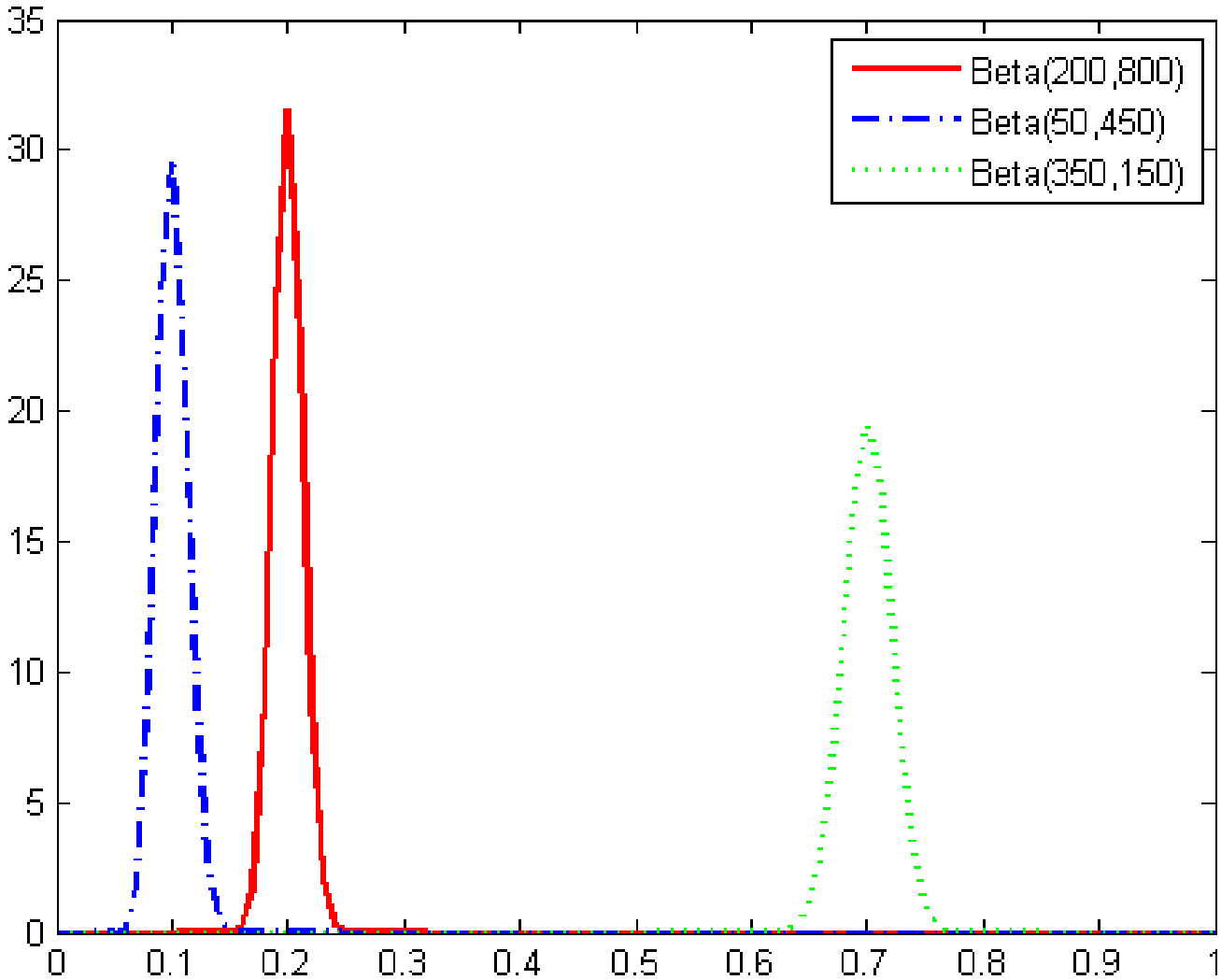
$$\Theta_X \sim Beta(200, 800)$$

since $\alpha = \gamma_X 0.2$ and $\beta = \gamma_X - \alpha$

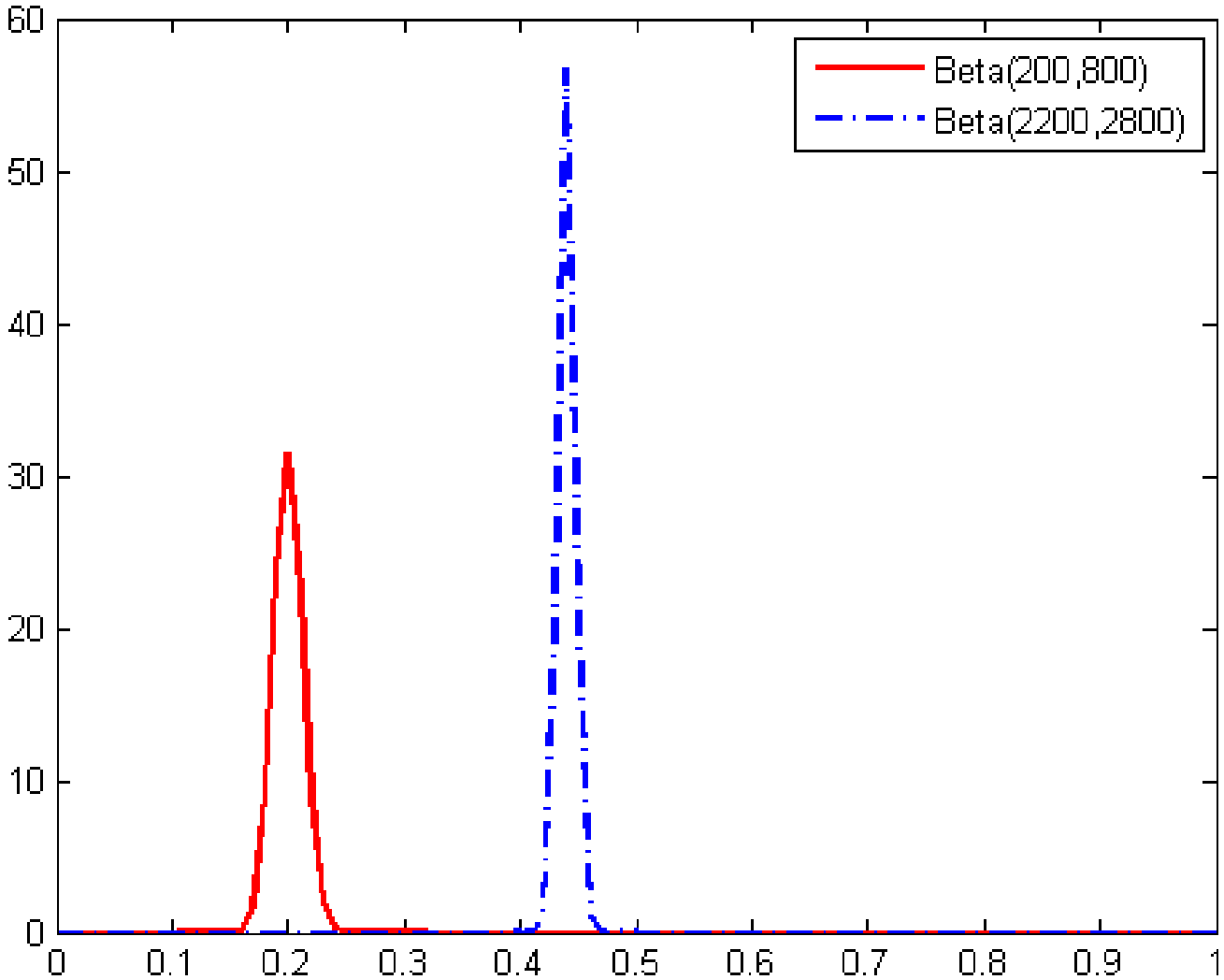
$$\Theta_{Y|f} \sim Beta(50, 450)$$

$$\Theta_{Y|t} \sim Beta(350, 150)$$

Beta distributions from the priors and the given confidence



Prior/Posterior distribution on $X = t$



Known Structure, Complete Data

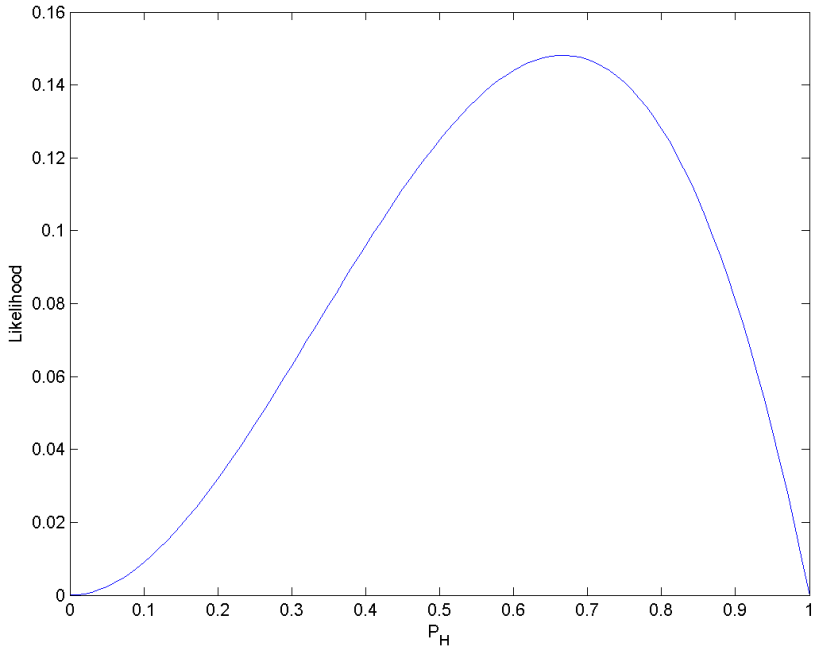
Some notation:

- N: number of BN's random variables
- n: cardinality of dataset (number of cases)
- r_i: cardinality (number of states) of X_i
- q_i: number of configurations of parents of X_i
- D: the dataset (observed sample) with cases D_m $(m = 1 \ldots n)$
- X_i: generic variable with states x_{ik}
- pa_i parent variables of X_i
- π_{ij}: j-th configuration of parents of X_i

- $D = \langle D_1, D_2, \ldots D_n \rangle$ is the dataset
- $D_m = (x_1[m], x_2[m], \ldots x_N[m])$ generic case ($x_i[m]$ observed state of variable X_i in case m)
- $\theta = (\theta_{ijk})$ parameter vector
- $\theta_{ijk} = \mathcal{P}(x_{ik} | \pi_{ij} : \theta)$
- $\theta_{ijl} = 1 - \sum_{k \neq l} \theta_{ijk}$
- $\theta_{ij} = (\theta_{ij1}, \theta_{ij2}, \ldots \theta_{ijr_i})$: parameter vector of variable X_i given the j-th configuration of parents
- $\theta_{i|pa_i} = (\theta_{ij_1}, \theta_{ij_2} \ldots)$ parameter vector of X_i's family.

Parameter Estimation

- Likelihood function: $\mathcal{L}(\theta : D) = \mathcal{P}(D|\theta)$



coin landing heads-up without prior knowledge after observing HHT

- Maximum Likelihood Estimation (MLE): $\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta : D)$
- Bayesian Estimation (BE): $\hat{\theta} = E_{\mathcal{P}}[\theta]$ where $\mathcal{P}(\theta|D) \propto \mathcal{P}(\theta)\mathcal{L}(\theta : D)$
- likelihood function has to be computed

Given a set of cases $D = \{D_m\}$ we assume i.i.d. samples, thus $\mathcal{L}(\theta : D) = \prod_{m=1}^{n} \mathcal{P}(D_m : \theta)$

Because of network factorization:

$$\mathcal{L}(\theta : D) = \prod_{m=1}^{n} \prod_{i=1}^{N} \mathcal{P}(x_i[m]|pa_i[m] : \theta) = \prod_{i=1}^{N} \prod_{m=1}^{n} \mathcal{P}(x_i[m]|pa_i[m] : \theta)$$

If the parameters local to a given family (CPD) are mutually independent (i.e., $\theta_{i|pa_i}$ are independent from $\theta_{i'|pa_{i'}}$) we have that

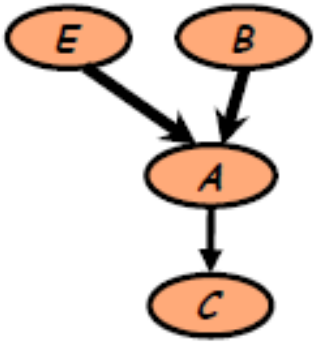
$$\mathcal{L}(\theta : D) = \prod_{i=1}^{N} \mathcal{L}_i(\theta_{i|pa_i} : D)$$

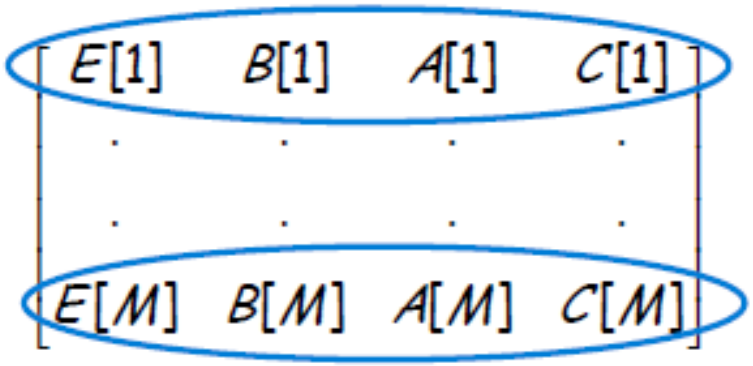
where $\mathcal{L}_i(\theta_{i|pa_i} : D) = \prod_{m=1}^{n} \mathcal{P}(x_i[m]|pa_i[m] : \theta_{i|pa_i})$ is the *conditional likelihood* of X_i

Example

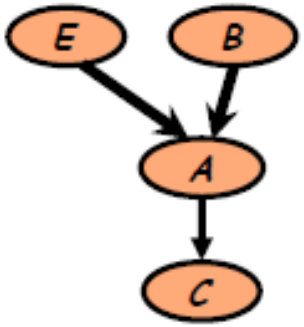
$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta) =$$

$$= \prod P(E[m] : \Theta) P(B[m] : \Theta) P(A[m] \mid B[m], E[m] : \Theta) P(C[m] \mid A[m] : \Theta)$$





Example



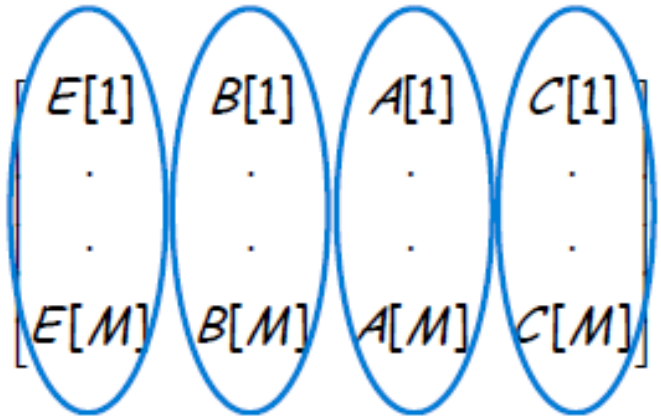
$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta) =$$

$$= \prod_m P(E[m] : \Theta)$$

$$\prod_m P(B[m] : \Theta)$$

$$\prod_m P(A[m] \mid B[m], E[m] : \Theta)$$

$$\prod_m P(C[m] \mid A[m] : \Theta)$$



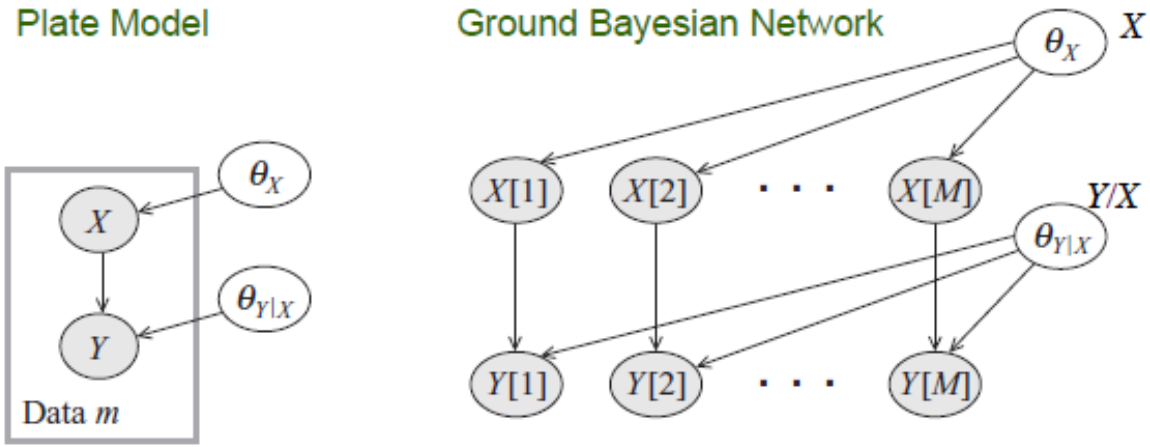
- If global likelihood factorizes into conditional local likelihoods, parameters for each variable can be maximized independently
- if $\hat{\theta}_{i|pa_i}$ are parameters maximazing $\mathcal{L}_i(\theta_{i|pa_i} : D)$ then $\hat{\theta} = (\hat{\theta}_{1|pa_1}, \hat{\theta}_{2|pa_2}, \ldots \hat{\theta}_{N|pa_N})$ maximizes $\mathcal{L}(\theta : D)$
- in case of multinomial distribution, if $M_i[k,j]$ is the number of times $X_i = k$ and pa_i are in configuration j, then we get a further decomposition $\mathcal{L}_i(\theta_{i|pa_i} : D) = \prod_j \prod_k \theta_{ijk}^{M_i[k,j]}$
- it follows that MLE is $\hat{\theta}_{i|pa_i} = (\hat{\theta}_{ij_1}, \hat{\theta}_{ij_2} \ldots \hat{\theta}_{ij_{q_i}})$ where $\hat{\theta}_{ij} = (\hat{\theta}_{ijk_1}, \hat{\theta}_{ijk_2} \ldots \hat{\theta}_{ijk_{r_i}})$ and finally

$$\hat{\theta}_{ijk} = \frac{M_i[k,j]}{\sum_k M_i[k,j]}$$

$\hat{\theta}_{ijk}$ estimates $\mathcal{P}(X_i = k | pa_i = j)$

Bayesian Estimation of BN Parameters

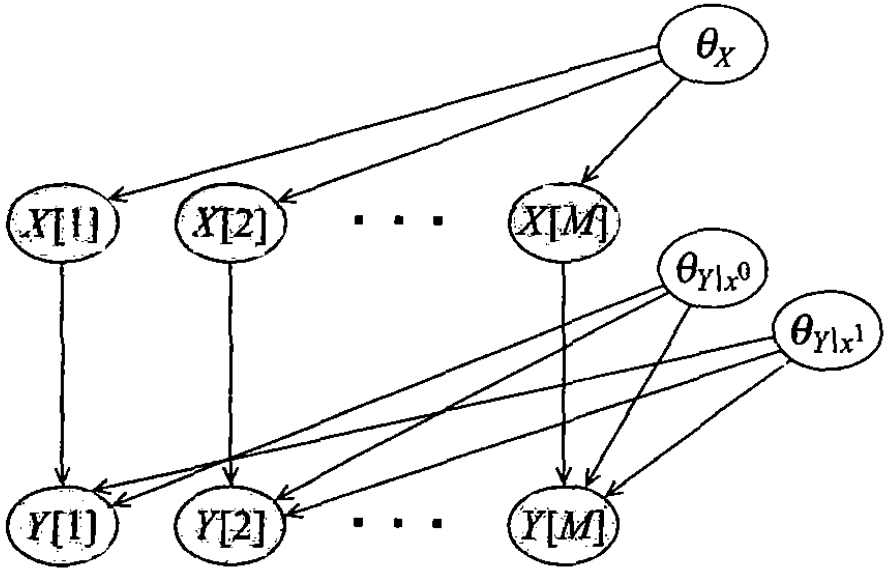
Let $\mathcal{G}$ be a BN structure with parameters $\theta = (\theta_{1|pa_1}, \ldots \theta_{N|pa_N})$, a prior $\mathcal{P}(\theta)$ satisfies the *global parameter independence* if $\mathcal{P}(\theta) = \prod_{i=1}^{N} \mathcal{P}(\theta_{i|pa_i})$



$$\mathcal{P}(\theta|D) = \prod_{i=1}^{N} \mathcal{P}(\theta_{i|pa_i}|D)$$

we can determine the posterior over parameters independently

Let X_i be a variable with parents pa_i, the prior $\mathcal{P}(\theta_{i|pa_i})$ satisfies *local parameter independence* if $\mathcal{P}(\theta_{i|pa_i}) = \prod_{j=1}^{q_i} \mathcal{P}(\theta_{ij})$



Let $\mathcal{G}$ be a BN structure with parameters $\theta = (\theta_{1|pa_1}, \ldots \theta_{N|pa_N})$, if the prior $\mathcal{P}(\theta)$ satisfies global and local parameter independence then

$$\mathcal{P}(\theta|D) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \mathcal{P}(\theta_{ij}|D)$$

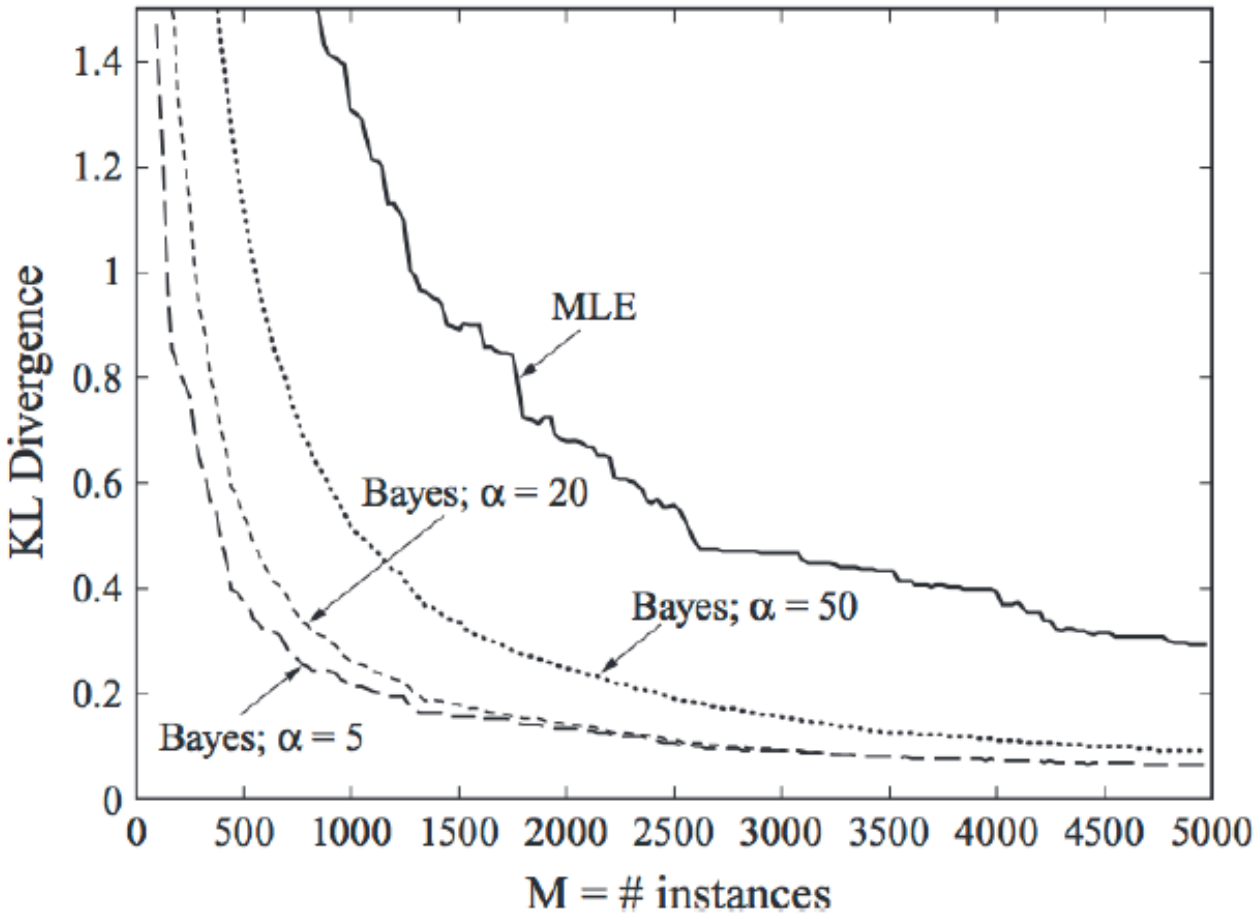
- Let $\theta_{ij} = (\theta_{ijk_1}, \ldots \theta_{ijk_{r_i}}) \sim Dir(\alpha_{ijk_1} \ldots \alpha_{ijk_{r_i}})$ then $(\theta_{ij}|D) \sim Dir(\alpha_{ijk_1} + M_i[k_1, j] \ldots \alpha_{ijk_{r_i}} + M_i[k_{r_i}, j])$
- Estimate

$$\mathcal{P}(X_i = k|pa_i = j) = E[\theta_{ijk}|D] = \frac{\alpha_{ijk} + M_i[k,j]}{\sum_k \alpha_{ijk} + \sum_k M_i[k,j]}$$

- How to set hyper-parameters?
  - *K2 prior*: use a fixed value (say $\alpha_{ijk} = 1$) for all the net's hyper-parameters [Cooper & Herskovitz:93]
  - *BDe prior* (Bayesian Dirichlet equivalent): set an imaginary data set size α and a representation $\mathcal{P}'(X_1 \ldots X_N)$ of the probability of each possible imaginary sample; set then $\alpha_{ijk} = \alpha \mathcal{P}'(X_i = k|pa_i = j)$

- in BDe prior the number of imaginary samples for different choices of parent values is identical
- we can use a prior BN to model $\mathcal{P}'$ (only for the parameters and not necessarily for the structure); we can then compute efficiently $\mathcal{P}'(X_i = k|pa_i = j)$
- it is common to define $\mathcal{P}'$ as a set of independent marginals $\mathcal{P}'(X_i)$ $(i = 1 \ldots N)$
- another choice can be to set $\mathcal{P}'$ to uniform distribution

Case study: Alarm ICU network

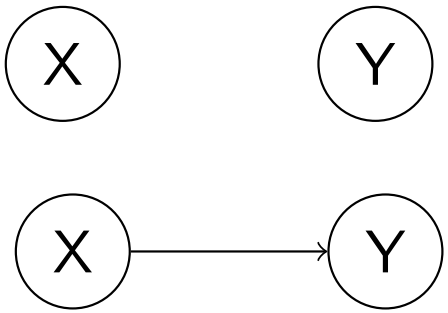


Lowest error with weakest prior ($\alpha = 5$)

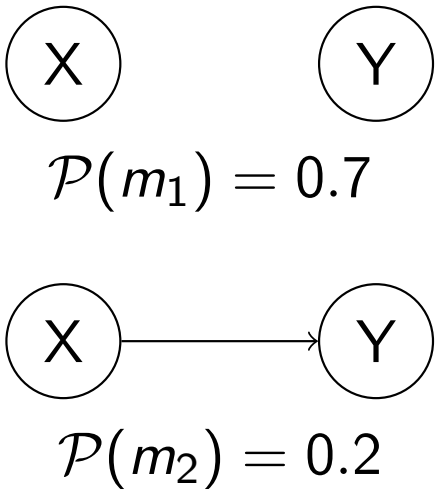
Larger values introduce bias; bias disappears as samples increase

Structure Learning

- Given data which model is *correct*?



- **Bayesian approach:** given data which model is ~~correct~~ *more likely*?



$\xrightarrow{\text{data } D}$

$\mathcal{P}(m_1|D) = 0.1$
$\mathcal{P}(m_2|D) = 0.9$

Bayesian Model Selection: $\tilde{m} = \arg\max_m \mathcal{P}(m|D)$

In the example select model m_2 and use it as the *correct model*

Score-based Search

- main idea for model selection: assign a score to each candidate network, then search for the network with the highest score
- optimization problem
- **Pros**: statistically motivated, takes the structure of conditional probability into account
- **Cons**: computationally hard
- Heuristic search: hill climbing, best-first, simulated annealing

- Natural score measure: the *likelihood function* of $\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle$

$$\max_{\mathcal{G},\theta_{\mathcal{G}}} \mathcal{L}(\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : D) = \max_{\mathcal{G}} [\max_{\theta_{\mathcal{G}}} \mathcal{L}(\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : D)]$$

$$= \max_{\mathcal{G}} [\mathcal{L}(\langle \mathcal{G}, \hat{\theta}_{\mathcal{G}} \rangle : D)]$$

- to find max likelihood pair $\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle$ we search structure $\mathcal{G}$ that achieves highest likelihood *when using MLE parameters* $\hat{\theta}_{\mathcal{G}}$ *for* $\mathcal{G}$
- let $\ell(\hat{\theta}_{\mathcal{G}} : D)$ be the log-likelihood function

$$score_{\mathcal{L}}(\mathcal{G} : D) = \ell(\hat{\theta}_{\mathcal{G}} : D)$$

Information theoretic interpretation

$$H_{\mathcal{P}}(X) = -\sum_{x} \mathcal{P}(x) \log \mathcal{P}(x)$$

entropy of X

$$I_{\mathcal{P}}(X;Y) = \sum_{x,y} \mathcal{P}(x,y) \log \frac{\mathcal{P}(x,y)}{\mathcal{P}(x)\mathcal{P}(y)}$$

Mutual Information between X and Y

- it measures how much information each variable provides about the other
- $I_{\mathcal{P}}(X;Y) \geq 0$
- $I_{\mathcal{P}}(X;Y) = 0$ iff X and Y are independent
- $I_{\mathcal{P}}(X;Y) = H_{\mathcal{P}}(X)$ iff X is totally predictable given Y

It can be proved that:

$$score_{\mathcal{L}}(\mathcal{G} : D) = n \sum_{i=1}^{N} [I_{\mathcal{P}}(X_i; pa_i^{\mathcal{G}}) - H_{\mathcal{P}(X_i)}]$$

where n is the number of samples

- the term $n \sum_{i=1}^{N} H_{\mathcal{P}(X_i)}$ does not depend on structure $\mathcal{G}$
- the score measures the strength of the dependence between variables and their parents
- *good news:* we prefer networks where the parents of each variable are informative about it
- *bad news:* adding arcs always helps
  $I(X; Y) \leq I(X; Y \cup Z)$
  maximal score attained by fully connected network
  such network can overfit data

Avoiding overfitting

- Classical issue in Machine Learning
- *Restricting hypotheses space:* restrict the number of parents and/or the number of parameters
- *Minimum Description Length:* penalize models that are too complex (Occam's razor)
- *Bayesian Methods:* use prior knowledge and/or average over all parameters values

Bayesian Score

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

$$score_B(\mathcal{G} : D) = \log \mathcal{P}(D|\mathcal{G}) + \log \mathcal{P}(\mathcal{G})$$

Problem: computation of the *marginal likelihood*

$$\mathcal{P}(D|\mathcal{G}) = \int_{\theta_\mathcal{G}} \mathcal{P}(D|\theta_\mathcal{G}, \mathcal{G})\mathcal{P}(\theta_\mathcal{G}|\mathcal{G})$$

$\mathcal{P}(D|\theta_\mathcal{G}, \mathcal{G})$ likelihood of data given network $\langle \mathcal{G}, \theta_\mathcal{G} \rangle$

$\mathcal{P}(\theta_\mathcal{G}|\mathcal{G})$ prior distribution over parameters of $\mathcal{G}$

n.b.: MLE returns *maximum* of likelihood function; marginal likelihood returns the *expected value* of the function

Cooper-Herskovitz formula

Given a network $\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle$ with $\mathcal{P}(\theta_{\mathcal{G}}|\mathcal{G})$ satisfying global and local independence; let each parameter $\theta_{ij} \sim Dir(\alpha_{ijk} : (k = 1 \ldots r_i))$ then

$$\mathcal{P}(D|\mathcal{G}) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + M_i[j])} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + M_i[k,j])}{\Gamma(\alpha_{ijk})}$$

where $\alpha_{ij} = \sum_k \alpha_{ijk}$ and $M_i[j] = \sum_k M_i[k,j]$

BIC: Bayesian Information Criterion

We can use an approximation for marginal likelihood exploiting the following theorem

Theorem

If we use Dirichlet priors for all parameters of the network, then when $n \rightarrow \infty$ we have that

$$\log \mathcal{P}(D|\mathcal{G}) = \ell(\hat{\theta}_{\mathcal{G}} : D) - \frac{\log n}{2} Dim[\mathcal{G}] + O(1)$$

where $Dim[\mathcal{G}]$ is the number of independent parameters in $\mathcal{G}$.

$$score_{BIC}(\mathcal{G} : D) = \ell(\hat{\theta}_{\mathcal{G}} : D) - \frac{\log n}{2} Dim[\mathcal{G}]$$

It exploit MLE, by penalizing too complex models (MDL)

- In general Bayesian scores are biased towards simpler structure, by recognizing that more complex structure is necessary as more data are available (trade-off fit to data with model complexity)
- if variables are independent, small fluctuations in the data (sampling noise) are unlikely to cause preference for more complex structures
- in case of BIC it is evident that, the stronger is dependence from parents, the higher is the score; the more complex is the network, the lower is the score
- the data likelihood term grows linearly with n, while complexity grows logarithmically (the larger n is, the more emphasis is given to fit the data)

Score consistency

Suppose $\mathcal{G}^*$ is a *perfect map* for a distribution $\mathcal{P}$. A scoring function is *consistent* if the following properties hold as $n \rightarrow \infty$, with probability that approaches 1 (over all possible dataset D):

- $\mathcal{G}^*$ will maximize the score
- all strcture $\mathcal{G}$ that are not I-equivalent to $\mathcal{G}^*$ will have strictly lower score

Theorem

Bayesian score and BIC are consistent

Asymptotically, these scores prefer a structure that exactly fits the dependencies in the data.

Structure priors

- structure prior does not grow with n and do not play a role asymptotically (unless it rules out some structures with 0 prob)
- usual choices: uniform priors or *edge penalty* where $\mathcal{P} \propto c^{|\mathcal{E}|}$ ($c < 1$ and E edges of $\mathcal{G}$)
- no need to worry about the actual number of prior networks, since it suffices a value proportional to the actual prior (e.g. $\mathcal{P}(\mathcal{G}) = 1$ when uniform)
- *structure modularity*:

$$\mathcal{P}(\mathcal{G}) \propto \prod_{i=1}^{N} \mathcal{P}(pa_i = pa_i^{\mathcal{G}})$$

$\mathcal{P}(pa_i = pa_i^{\mathcal{G}})$ prior to choosing the set of parents of X_i

Score Decomposability

- a structure score is *decomposable* iff

$$score(\mathcal{G} : D) = \sum_{i=1}^{N} Fscore(X_i|pa_i : D)$$

  where $Fscore(X_i|pa_i : D)$ measures how well pa_i serves as parents of X_i in D
- e.g., likelihood score is decomposable since $Fscore_{\mathcal{L}}(X_i|pa_i : D) = n(I_{\mathcal{P}}(X_i; pa_i) - H_{\mathcal{P}}(X_i))$
- with a decomposable score, a local change in structure does not change the score of other parts (easier search)
- under which condition is Bayesian score decomposable?

Decomposability of Bayesian Score

- Let $\mathcal{P}(\theta_{\mathcal{G}}|\mathcal{G})$ be a set of parameter priors with global independence. They satisfy *parameter modularity* if for each $\mathcal{G}, \mathcal{G}'$ such that $pa_i^{\mathcal{G}} = pa_i^{\mathcal{G}'}$, then $\mathcal{P}(\theta_{X_i|pa_i^{\mathcal{G}}}|\mathcal{G}) = \mathcal{P}(\theta_{X_i|pa_i^{\mathcal{G}'}}|\mathcal{G}')$
- above property states that prior over X_i depends only on local structure

Theorem

Let $\mathcal{G}$ be a network structure, $\mathcal{P}(\mathcal{G})$ be a structure prior with structure modularity, and $\mathcal{P}(\theta_{\mathcal{G}}|\mathcal{G})$ be a parameter prior with global independence and parameter modularity, then the Bayesian score is decomposable

Parameter Priors

- impossible to elicit parameter priors for each possible network (superexponential)
- *K2 prior*: choose $Dir(\alpha, \ldots \alpha)$ for every parameter, where α is a fixed constant
- *BDe prior*: choose an equivalent sample size α and set $\alpha_{ijk} = \alpha \mathcal{P}'(X_i = k | pa_i = j)$
- BDe prior allows also to satisfy the important property of score equivalence

Score equivalence

Let $score(\mathcal{G} : D)$ be a score. It satisfies *score equivalence* if for all I-equivalent networks $\mathcal{G}, \mathcal{G}'$, we have that $score(\mathcal{G} : D) = score(\mathcal{G}' : D)$ for all dataset D

Theorem

The likelihood and the BIC scores satisfy score equivalence

What about bayesian score?

Theorem

Let $\mathcal{P}(\mathcal{G})$ be a structure score assigning equal prior to I-equivalent networks; let $\mathcal{P}(\theta_{\mathcal{G}}|\mathcal{G})$ be a Dirichlet parameter prior with global and local independence. The Bayesian score with this prior satisfies score equivalence iff the prior is a BDe prior for some α and $\mathcal{P}'$.

It follows that if we use Dirichlet priors and want decomposition, then to satisfy score equivalence we MUST use BDe prior

Structure Learning as Search

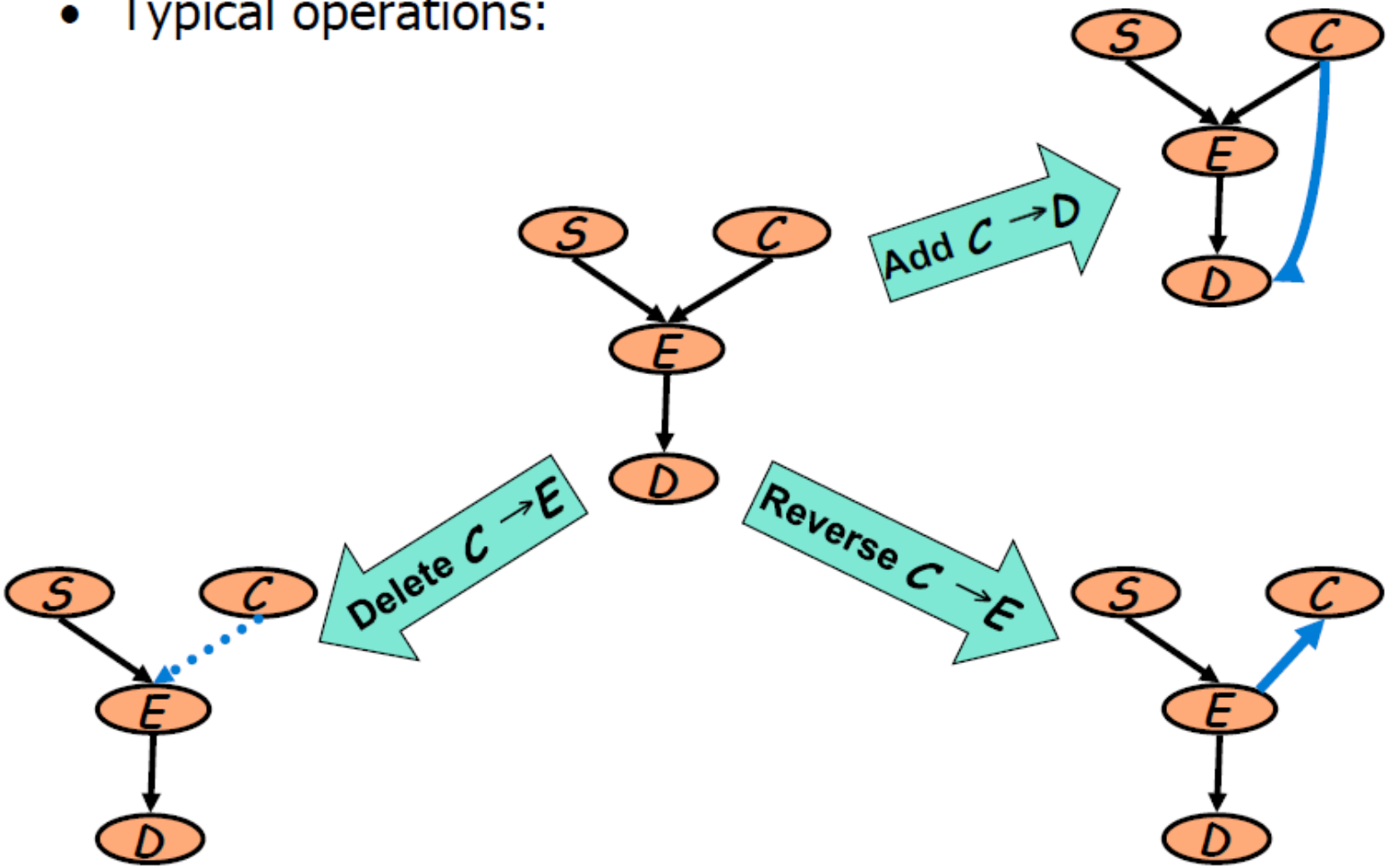
Input: training set, scoring function, priors (if needed), set of possible structures

Output: a networks (or networks) maximising the score

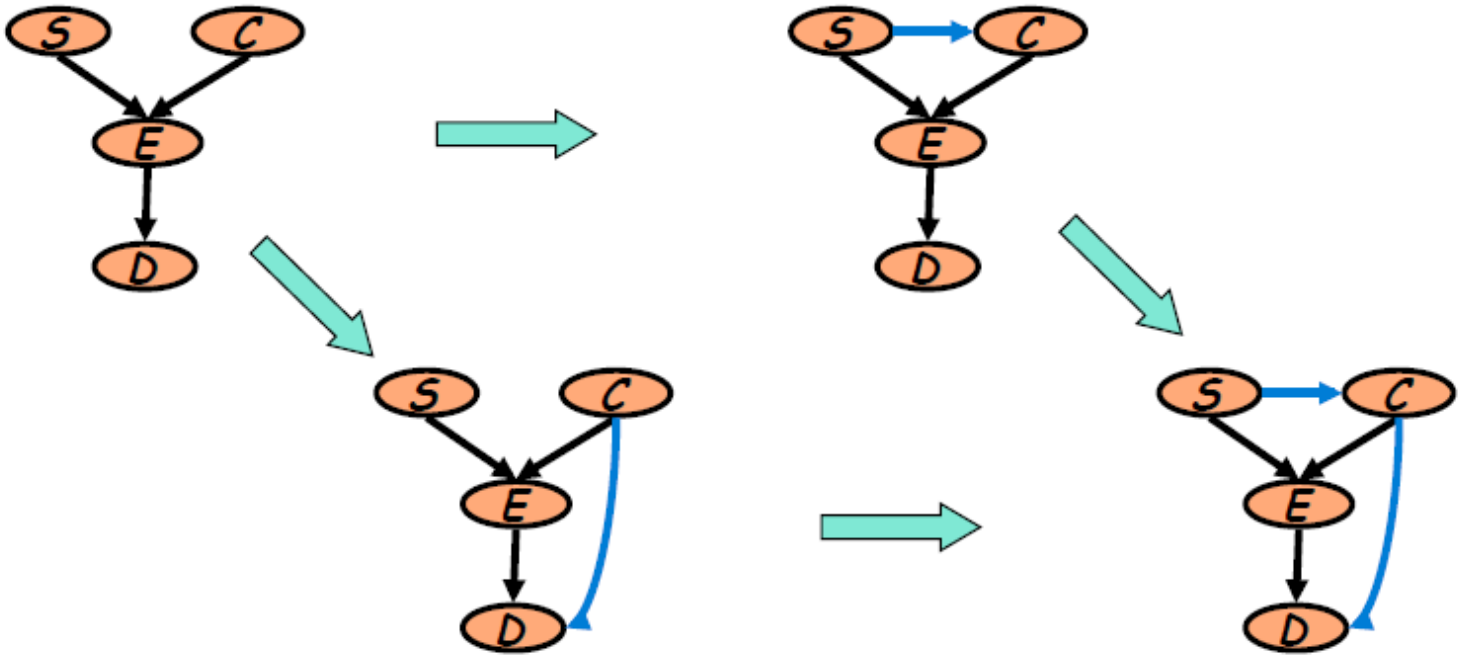
Key property: *decomposability*: the score is a sum of local terms

Optimization problem to be solved by heuristic search (hill climbing, simulated annealing, genetic algorithms, best first, etc. . .

- Typical operations:



Exploiting Decomposability in Local Search



- **Caching:** To update the score after a local change, we only need to re-score the families that were changed in the last move

Hill Climbing

- start with an initial network (random network, empty network, ...)
- at each iteration
  - evaluate all possible changes
  - apply change resulting to best increase in score
  - reiterate
- stop when no change improves the score

Better results obtained with *random restarting*

Each step requires evaluating approximately N new changes