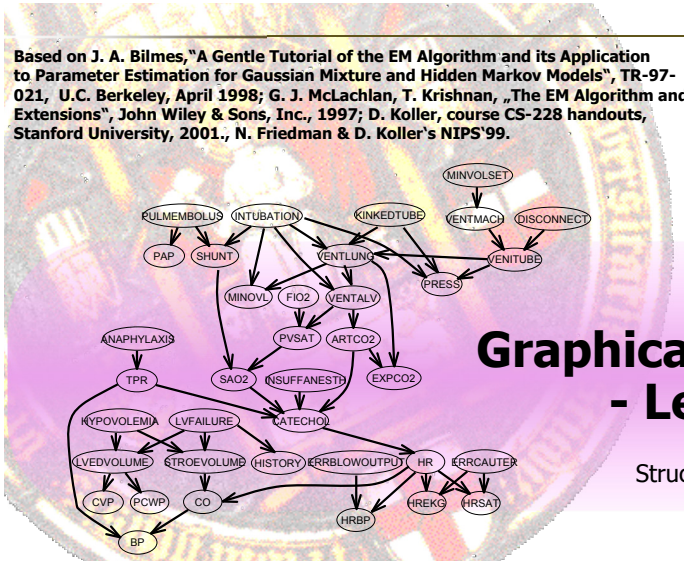
Based on J. A. Bilmes,"A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", TR-97-021, U.C. Berkeley, April 1998; G. J. McLachlan, T. Krishnan, „The EM Algorithm and Extensions", John Wiley & Sons, Inc., 1997; D. Koller, course CS-228 handouts, Stanford University, 2001., N. Friedman & D. Koller's NIPS'99.

Advanced I WS 06/07



Graphical Models - Learning -

Structure Learning

Wolfram Burgard, Luc De Raedt, Kristian Kersting, Bernhard Nebel

Albert-Ludwigs University Freiburg, Germany

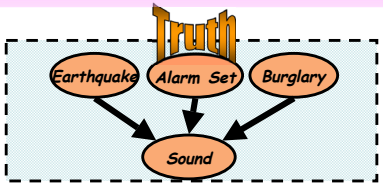
Learning With Bayesian Networks

observed	Fixed structure (A → B)	Fixed variables (A ? B)	Hidden variables (A ? B ? H)
fully	Easiest problem counting	Selection of arcs New domain with no domain expert Data mining	
Partially	Numerical, nonlinear optimization, Multiple calls to BNs, Difficult for large networks	Encompasses to difficult subproblem, „Only" Structural EM is known	Scientific discouvery

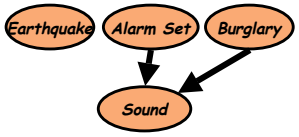
Parameter Estimation

Stucture learing ?

Why Struggle for Accurate Structure?

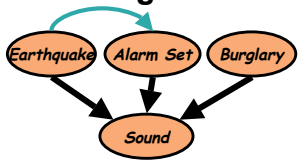


Missing an arc



- Cannot be compensated for by fitting parameters
- Wrong assumptions about domain structure

Adding an arc



- Increases the number of parameters to be estimated
- Wrong assumptions about domain structure

Unknown Structure, (In)complete Data

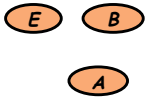
E, B, A
<Y,N,N>
<Y,N,Y>
<N,N,Y>
<N,Y,Y>
.
.
<N,Y,Y>

- Network structure is not specified
  - Learnerr needs to select arcs & estimate parameters
- Data does not contain missing values

E	B	P(A | E,B)
e	b	? ?
e	$\bar{b}$	? ?
$\bar{e}$	b	? ?
$\bar{e}$	$\bar{b}$	? ?

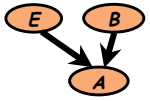
Learning algorithm

E	B	P(A | E,B)
e	b	.9 .1
e	$\bar{b}$	.7 .3
$\bar{e}$	b	.8 .2
$\bar{e}$	$\bar{b}$	.99 .01



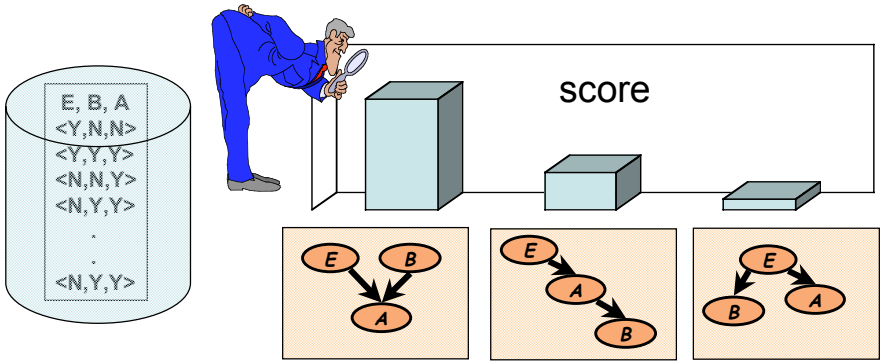
E, B, A
<Y,?,N>
<Y,N,?>
<N,N,Y>
<N,Y,Y>
.
.
<?,Y,Y>

- Network structure is not specified
- Data contains missing values
  - Need to consider assignments to missing values



Score-based Learning

Define scoring function that evaluates how well a structure matches the data



Search for a structure that maximizes the score

Structure Search as Optimization

Input:

- Training data
- Scoring function
- Set of possible structures

Output:

- A network that maximizes the score

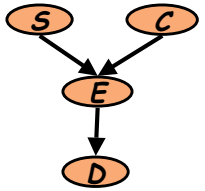
Heuristic Search

- Define a search space:
  - search states are possible structures
  - operators make small changes to structure
- Traverse space looking for high-scoring structures
- Search techniques:
  - Greedy hill-climbing
  - Best first search
  - Simulated Annealing
  - ...

Theorem: Finding maximal scoring structure with at most *k* parents per node is NP-hard for *k > 1*

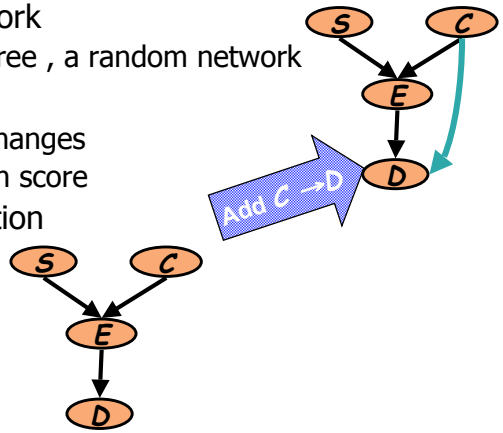
Typically: Local Search

- Start with a given network
  - empty network, best tree , a random network
- At each iteration
  - Evaluate all possible changes
  - Apply change based on score
- Stop when no modification improves score



Typically: Local Search

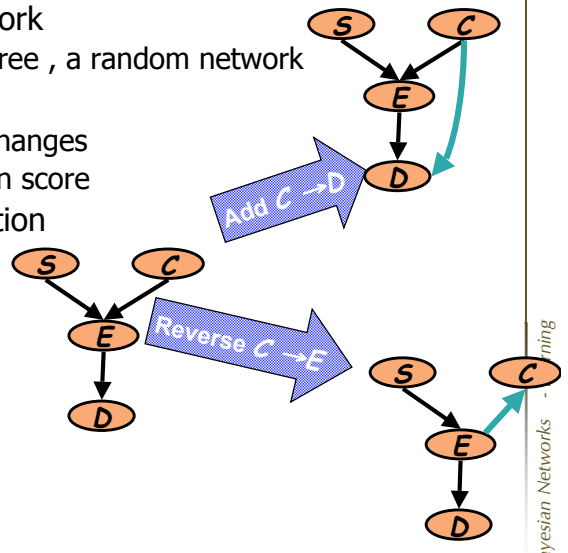
- Start with a given network
  - empty network, best tree , a random network
- At each iteration
  - Evaluate all possible changes
  - Apply change based on score
- Stop when no modification improves score



Advanced I WS 06/07

Typically: Local Search

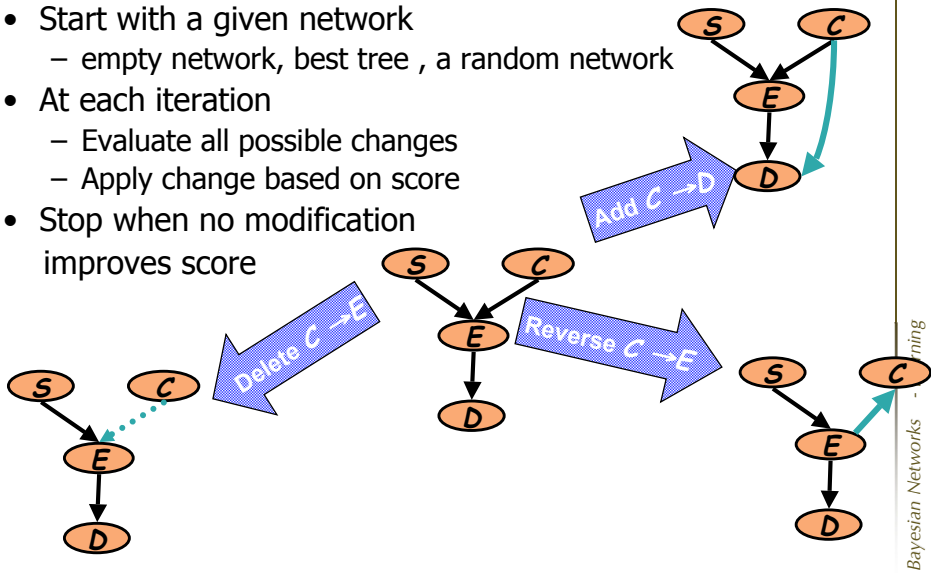
- Start with a given network
  - empty network, best tree , a random network
- At each iteration
  - Evaluate all possible changes
  - Apply change based on score
- Stop when no modification improves score



Advanced I WS 06/07

Typically: Local Search

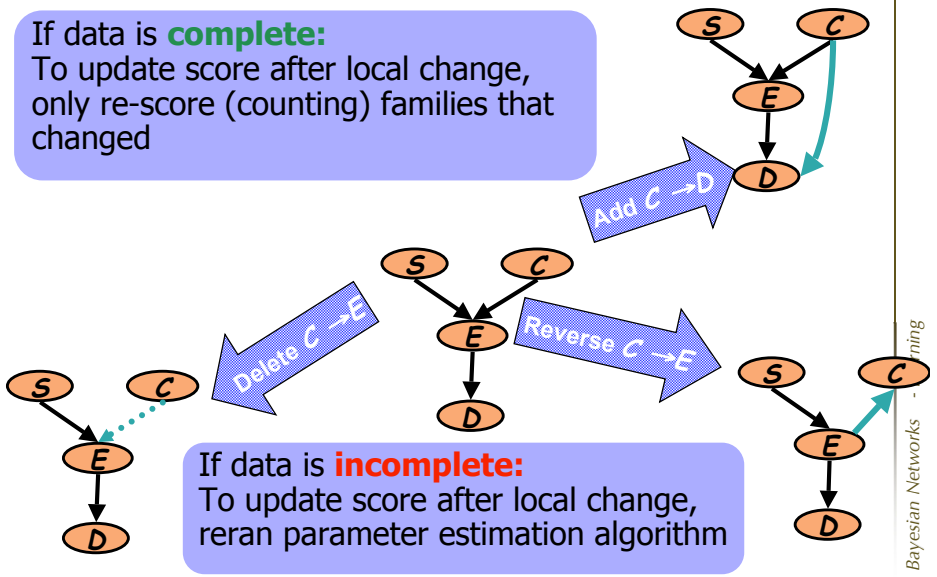
- Start with a given network
  - empty network, best tree , a random network
- At each iteration
  - Evaluate all possible changes
  - Apply change based on score
- Stop when no modification improves score



Advanced I WS 06/07

Typically: Local Search

If data is **complete:**
To update score after local change, only re-score (counting) families that changed



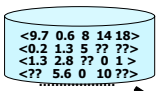
If data is **incomplete:**
To update score after local change, reran parameter estimation algorithm

Local Search in Practice

- Local search can get stuck in:
  - **Local Maxima:**
    - All one-edge changes reduce the score
  - **Plateaux:**
    - Some one-edge changes leave the score unchanged
- Standard heuristics can escape both
  - Random restarts
  - TABU search
  - Simulated annealing

Local Search in Practice

- Using LL as score, adding arcs always helps
  - Max score attained by fully connected network
  - Overfitting: A bad idea…
- Minimum Description Length:
  - Learning ⇔ data compression

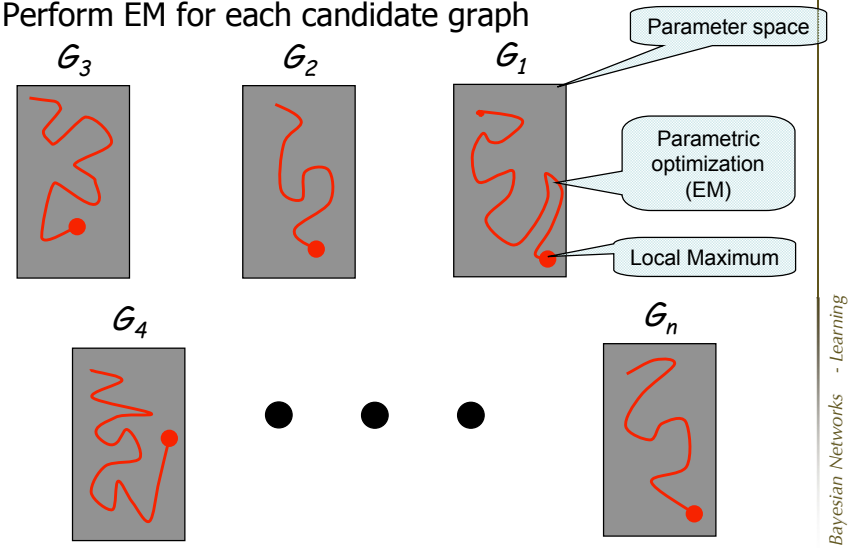


$$MDL(BN \mid D) = \underbrace{-\log P(D \mid \Theta, G)}_{\text{DL(Data|model)}} + \underbrace{\frac{\log N}{2} \mid \Theta \mid}_{\text{DL(Model)}}$$

- Other: BIC (Bayesian Information Criterion), Bayesian score (BDe)

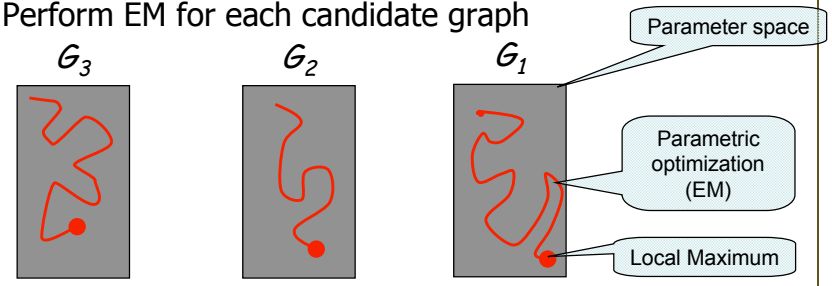
Local Search in Practice

- Perform EM for each candidate graph



Local Search in Practice

- Perform EM for each candidate graph



- Computationally expensive:
  - Parameter optimization via EM — non-trivial
  - Need to perform EM for all candidate structures
  - Spend time even on poor candidates

⇒ In practice, considers only a few candidates

Structural EM
[Friedman et al. 98]

Recall, in complete data we had
- Decomposition ⇒ efficient search

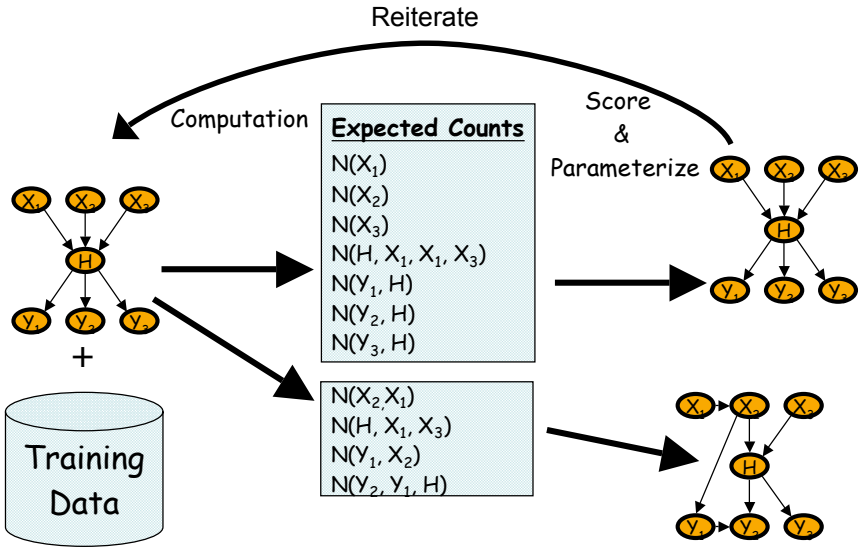
Idea:
- Instead of optimizing the real score…
- Find **decomposable** alternative score
- Such that maximizing new score
  ⇒ improvement in real score

Structural EM
[Friedman et al. 98]

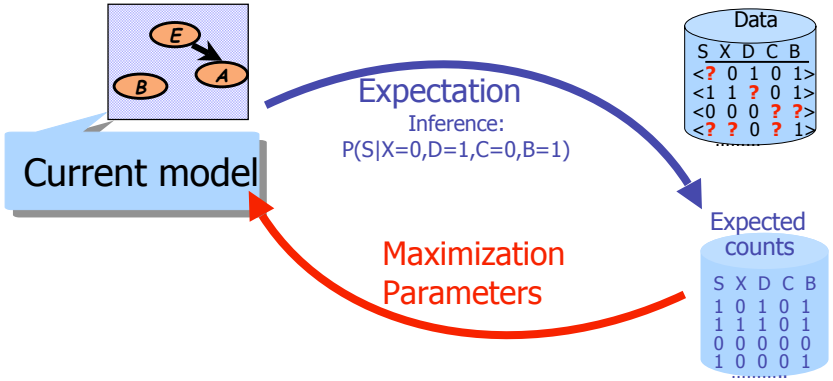
Idea:
- Use current model to help evaluate new structures

Outline:
- Perform search in (Structure, Parameters) space
- At each iteration, use current model for finding either:
  - Better scoring parameters: "parametric" EM step
  or
  - Better scoring structure: "structural" EM step

Structural EM
[Friedman et al. 98]

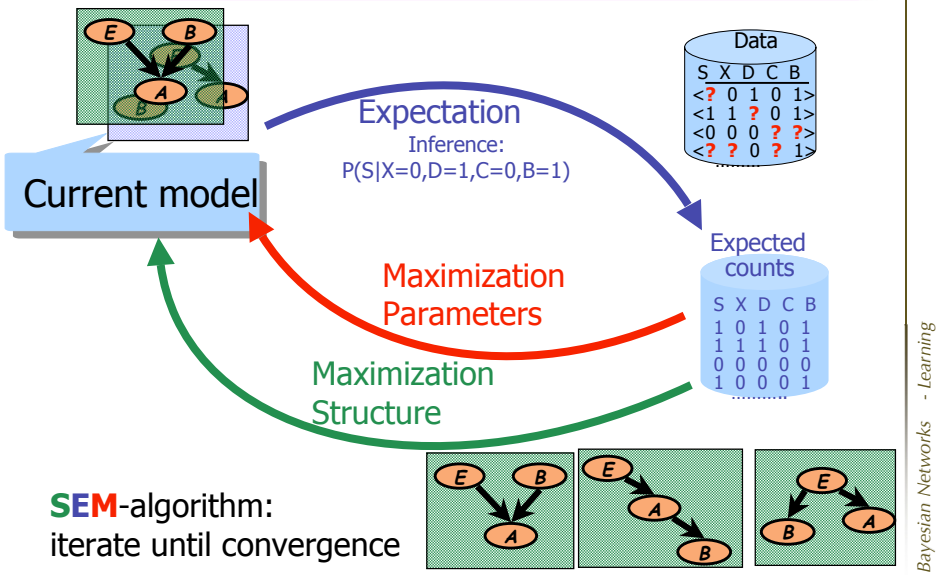


Structure Learning: incomplete data



EM-algorithm:
iterate until convergence

Structure Learning: incomplete data



SEM-algorithm:
iterate until convergence

Structure Learning: Summary

- Expert knowledge + learning from data
- Structure learning involves parameter estimation (e.g. EM)
- Optimization w/ score functions
  - likelihood + complexity penality = MDL
- Local traversing of space of possible structures:
  - add, reverse, delete (single) arcs
- Speed-up: Structural EM
  - Score candidates w.r.t. current best model