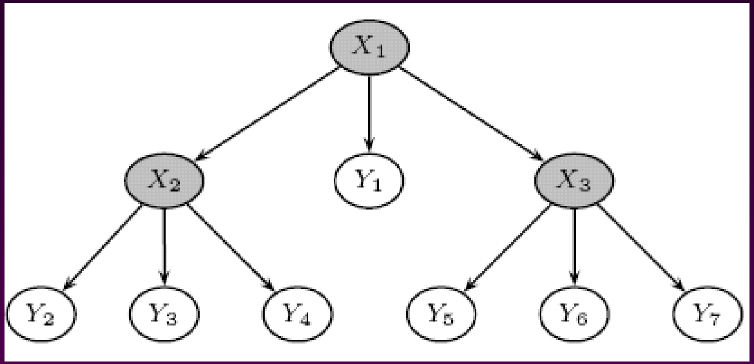
Learning Latent Tree Models

Nevin L. Zhang

Department of Computer Science & Engineering

The Hong Kong University of Science & Technology

Latent Tree Models (LTM)



- Bayesian networks with
  - Rooted tree structure
  - Discrete random variables
  - Leaves observed (manifest variables)
  - Internal nodes latent (latent variables)

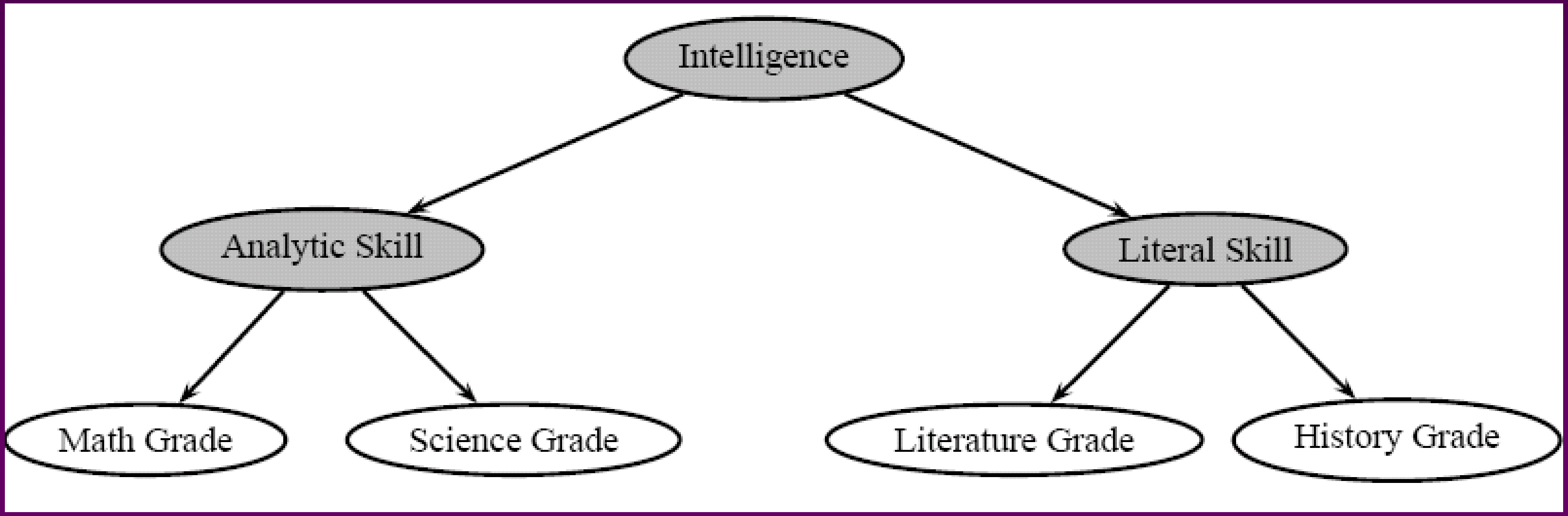
$P(X_2|X_1)$

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.9	0.1
$X_1 = 1$	0.1	0.9

- Also known as hierarchical latent class (HLC) models, HLC models

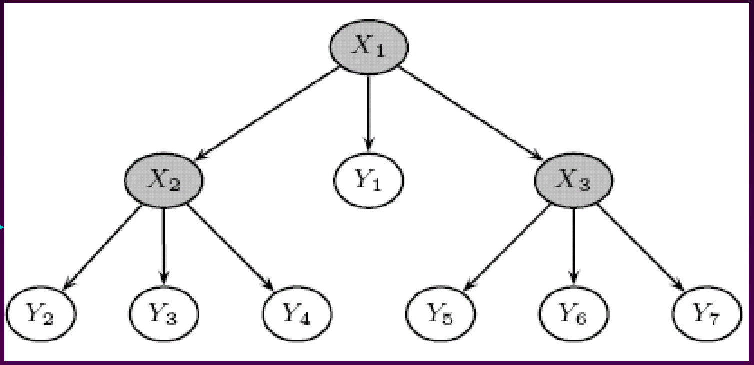
Example

- Manifest variables
  - Math Grade, Science Grade, Literature Grade, History Grade
- Latent variables
  - Analytic Skill, Literal Skill, Intelligence



Learning Latent Tree Models

Y1	Y2	…	Y6	Y7
1	0	…	1	1
1	1	…	0	0
0	1	…	0	1
…	…	…	…	…



Determine

- Number of latent variables
- Cardinality of each latent variable
- Model Structure
- Conditional probability distributions

Outline

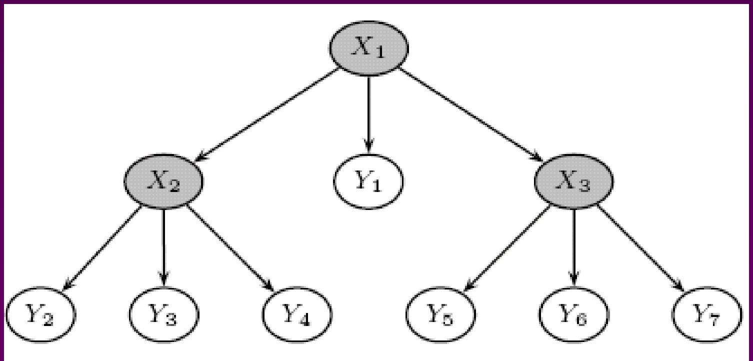
- Problem Statement
- Why Interesting?
- Technical issues
  - Properties of Latent Tree Models
  - Model Selection
  - Model Optimization
- Conclusions

Why Latent Tree Models Interesting?

- Probabilistic modeling
- Latent structure discovery
- Cluster Analysis
- Traditional Chinese Medicine

LTM and Probabilistic Modeling

- Pearl 1988: LTMs
  - Are computationally very simple to work with.
  - Can represent complex relationships among manifest variables.



LTM and Probabilistic Modeling

- New approximate inference algorithm for BN
  - Dense BN with variables Y1, Y2, …, Yn
  - Sample from the BN a data set on Y1, Y2, …, Yn
  - Learn an LTM with manifest variables Y1, Y2, …, Yn and some latent variables
  - Use the LTM to make inference among Y1, Y2, …, Yn
  - Empirical comparison with Loopy Propagation
    - More accurate
    - Much lower online complexity

LTM and Probabilistic Modeling

- New approach for density estimation

Bayes rule: $P(C|A_1, A_2, \ldots, A_m) \propto P(C)P(A_1, A_2, \ldots, A_m|C)$

Density estimation: $P(A_1, A_2, \ldots, A_m|C)$

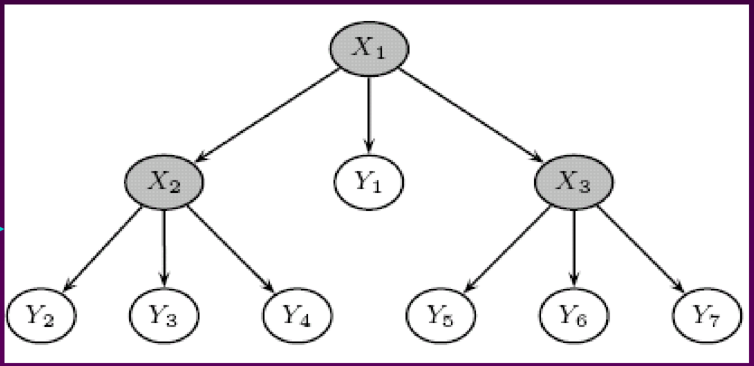
A new method: Learn an LTM for $P(A_1, A_2, \ldots, A_m|C)$

Intuition: attributes influenced by latent factors besides C.

Latent Structure Discovery

- Learning LTM is to discover latent structures

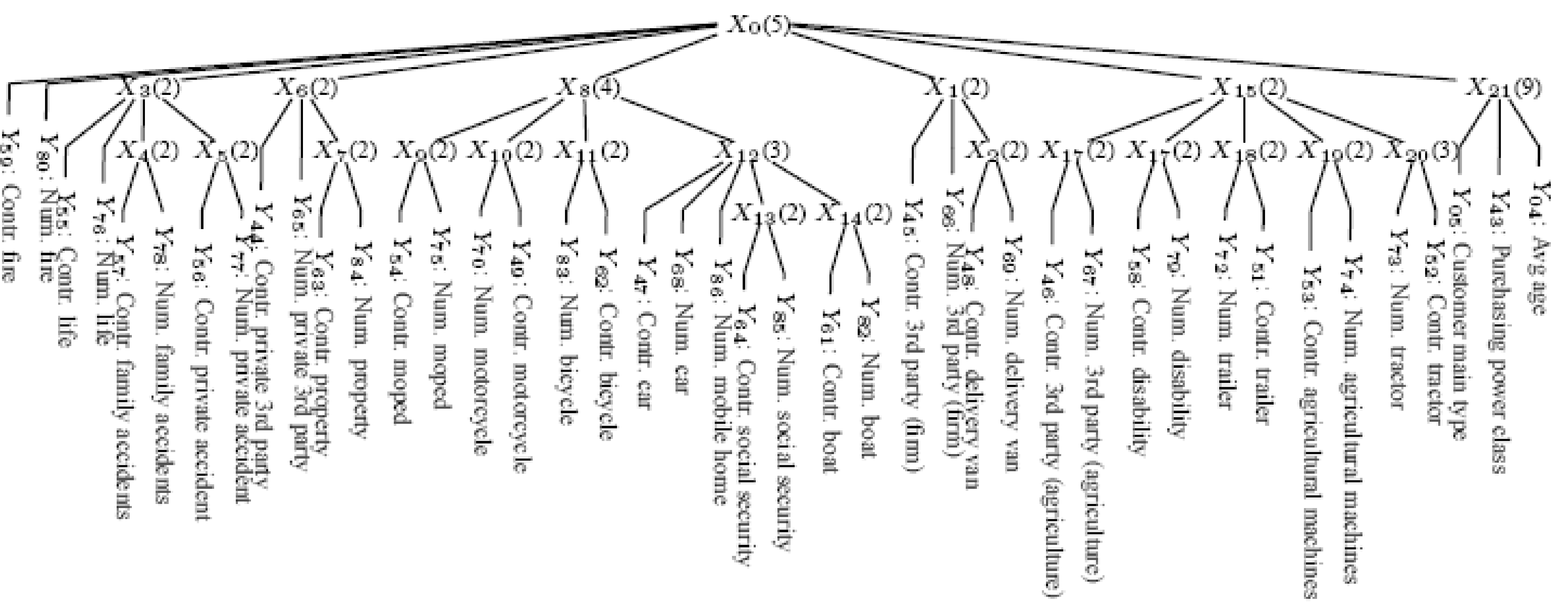
Y1	Y2	…	Y6	Y7
1	0	…	1	1
1	1	…	0	0
0	1	…	0	1
…	…	…	…	…



- Can interesting latent structures be discovered?

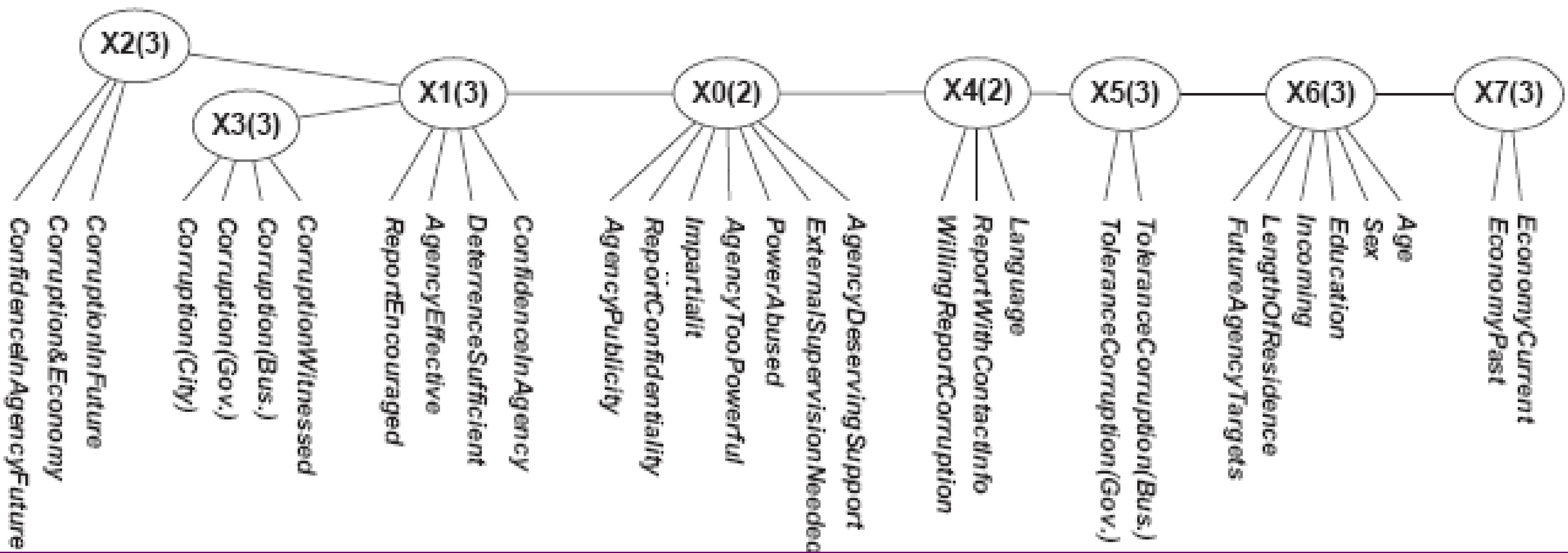
Latent Structure Discovery

- Results on the CoiL Challenge 2000 data set
- Customer records of a Holland Insurance Company
- 42 manifest variables, 5822 records



Latent Structure Discovery

- Hong Kong ICAC survey data
- 31 manifest variables, 12000 records

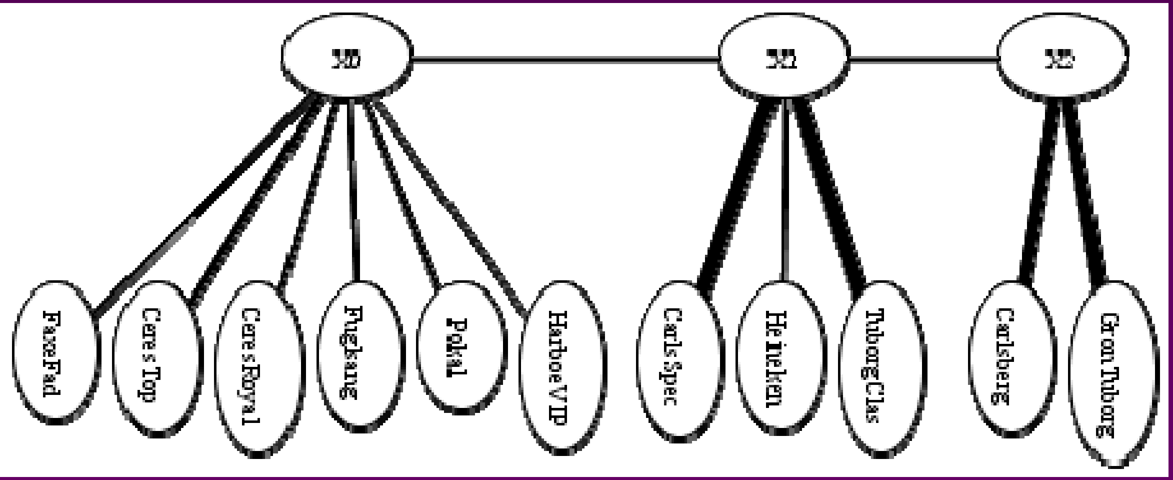


Latent Structure Discovery

- Danish Beer data
- 783 samples
- States of Manifest variables

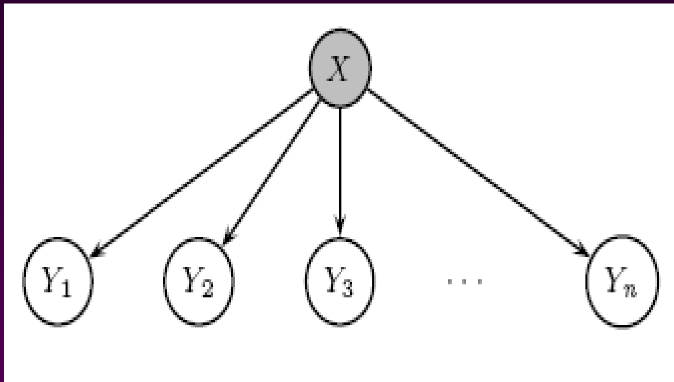
1. Never heard of; 2. heard but not tasted;

3. tasted but don't drink regularly; 4. drink regularly

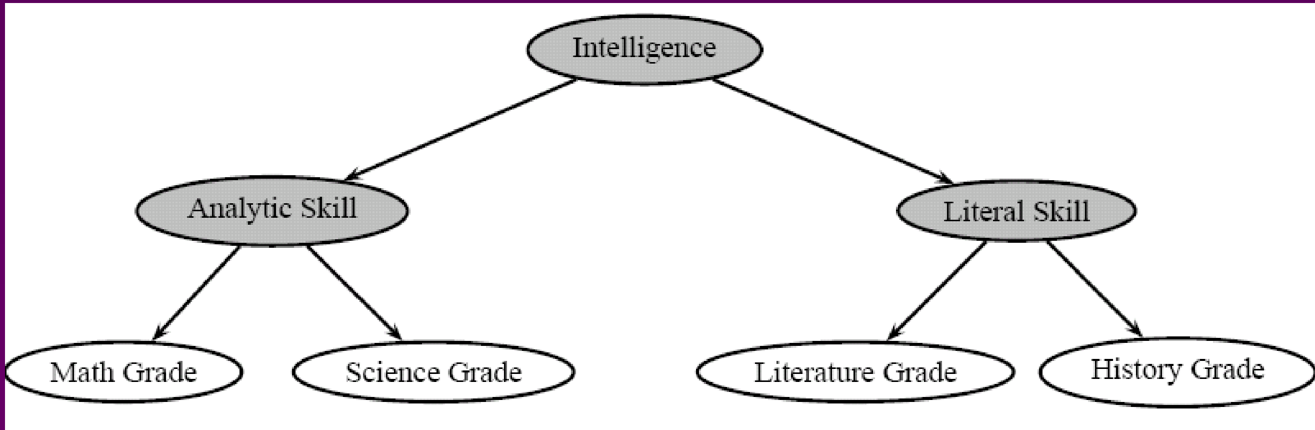


Cluster Analysis

- Latent class model (LCM) for cluster analysis:
  - Each state of X represents a cluster



- LTM generalizes LCM
  - Relaxes strong constraint of LCM
  - Multidimensional clustering



Traditional Chinese Medicine (TCM)

- TCM statement:
  - Yang deficiency (阳虚): intolerance to cold (畏寒), cold limbs (肢冷), cold lumbus and back (腰背冷), and so on ….
  - Regarded by many as not scientific, even groundless.

- Two aspects to the meaning
  1. Claim: There exists a class of patients, who characteristically have the cold symptoms . The cold symptoms co-occur in a group of people,
  2. Explanation offered: Due to deficiency of Yang. It fails to warm the body

- What to do?
  - Previous work focused on 2.
  - New idea: Do data analysis for 1

Objectivity of the Claimed Pattern

- TCM Claim: there exits a class of patients, in whom symptoms such as ‘intolerance to cold’, ‘cold limbs’, ‘cold lumbus and back’, and so on co-occur at the same time

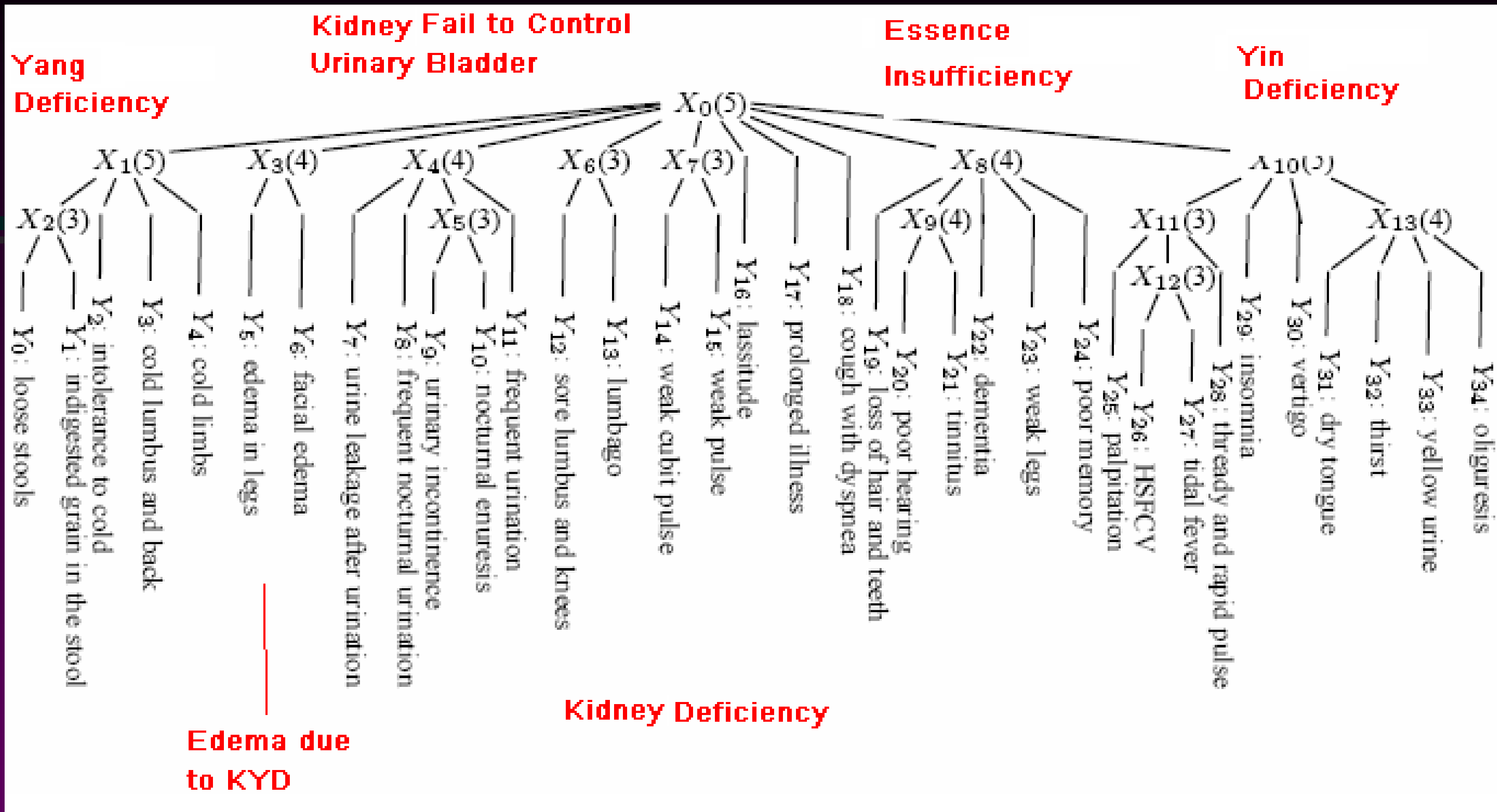
- How to prove or disapprove that such claimed TCM classes exist in the world?
  - Systematically collect data about symptoms of patients.
  - Perform cluster analysis, obtain natural clusters of patients
  - If the natural clusters corresponds to the TCM classes, then YES.
    1. Existence of TCM classes validated
    2. Descriptions of TCM classes refined and systematically expanded
    3. Establish a statistical foundation for TCM

Why Latent Tree Models?

- TCM uses multiple interrelated latent concepts to explain co-occurrence of symptoms
  - Yang deficiency (肾阳虚) , Yin deficiency (肾阴虚): , Essence insufficiency (肾精亏虚) , …
- Need latent structure models
  - With multiple interrelated latent variables..
- Latent Tree Models are the simplest such models

Empirical Results

- Can we find the claimed TCM classes using latent tree models?
  - We collected a data set about kidney deficiency (肾虚)
  - 35 symptom variables, 2600 records



- Y0-Y34: manifest variables from data
- X0-X13: latent variables introduced by data analysis
- Structure interesting, supports TCM's theories about various symptoms.

Latent Clusters

- X1:
  - 5 states: s0, s1, s2, s3, s4
  - Samples grouped into 5 clusters
- Cluster X1=s4

{sample | P(X1=s4|sample) > 0.95} ➔

X1=s4

Y2	Y3	Y4	# samples
3	3	3	8
3	2	3	4
3	2	2	8
2	3	3	4
3	2	1	1
3	3	2	2
2	2	2	30

....

Cold symptoms co-occur in samples

- Class implicitly claimed by TCM found!
- Description of class refined
  - By Math vs by words

Other TCM Data Sets

- From Beijing U of TCM, 973 project
  - Depression
  - Hepatitis B
  - Chronic Renal Failure

- China Academy of TCM
  - Subhealth
  - Type 2 Diabetes

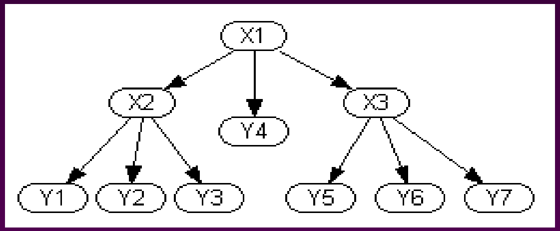
- In all cases, distribution patterns implicitly claimed in TCM theory
  - Validated
  - Quantified and refined

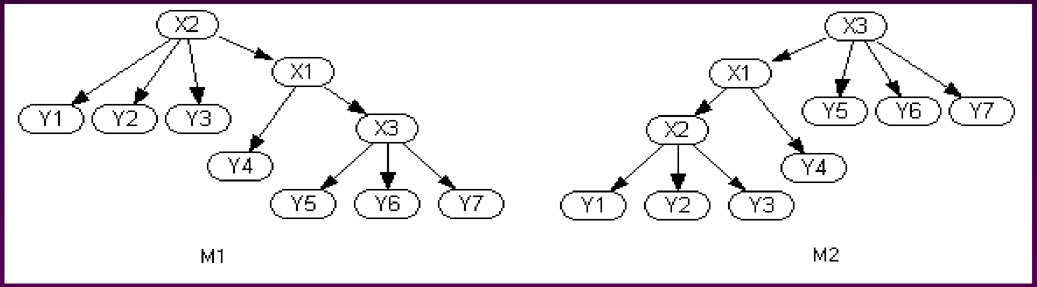
Outline

- Problem Statement
- Why interesting
- <u>Technical issues</u>
  - Properties of Latent Tree Models
  - Model Selection
  - Model Optimization
- Conclusions

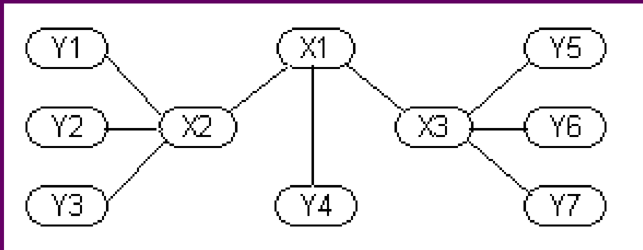
Root Walking and Model Equivalence

- M1: root walks to X2; M2: root walks to X3





- Root walking leads to equivalent models
- Implications:
  - Cannot determine edge orientation from data
  - Can only learn unrooted models



Regularity

- Regular latent tree models: For any latent node Z with neighbors X1, X2, …, Xk

$$|Z| \leq \frac{\prod_{i=1}^{k} |X_i|}{\max_{i=1}^{k} |X_i|},$$

- Can focus on regular models only
  - Irregular models can be made regular
  - Regularized models better than irregular models
- The set of all such models is finite.

Model Selection

- Bayesian score: posterior probability P(m|D)
  - $P(m|D) = P(m) \int P(D|m, \theta)\, d\theta \,/\, P(D)$
- BIC Score: large sample approximation

  $BIC(m|D) = \log P(D|m, \theta^*) - d \log N/2$

- BICe Score:

  $BICe(m|D) = \log P(D|m, \theta^*) - d_e \log N/2$

  effective dimension d_e.

  - Effective dimensions are difficult to compute
  - BICe not realistic

Model Selection

- Other Choices
  - Cheeseman-Stutz (CS): impact of approximation error in BIC reduced
  - AIC
  - Holdout likelihood
  - (Cross validation: too expensive)

- Simulation studies indicate that
  - BIC and CS result in good models
  - AIC and holdout likelihood do not
- Therefore, we chose work with BIC.

Model Optimization

- Search-based algorithm
  - Start with an initial model
  - At each step:
    - Construct all possible candidate models
    - Evaluate them one by one
    - Pick the best one

- Difficult
  - Too many candidate models
  - Too expensive to run EM on all of them

Model Optimization

- Double hill climbing (DHC), 2002
  - 7 manifest variables.
- Single hill climbing (SHC), 2004
  - 12 manifest variables
- Heuristic SHC (HSHC), 2004
  - 50 manifest variables
- EAST, 2007
  - As efficient as HSHC, and more principled
  - 100+ manifest variables

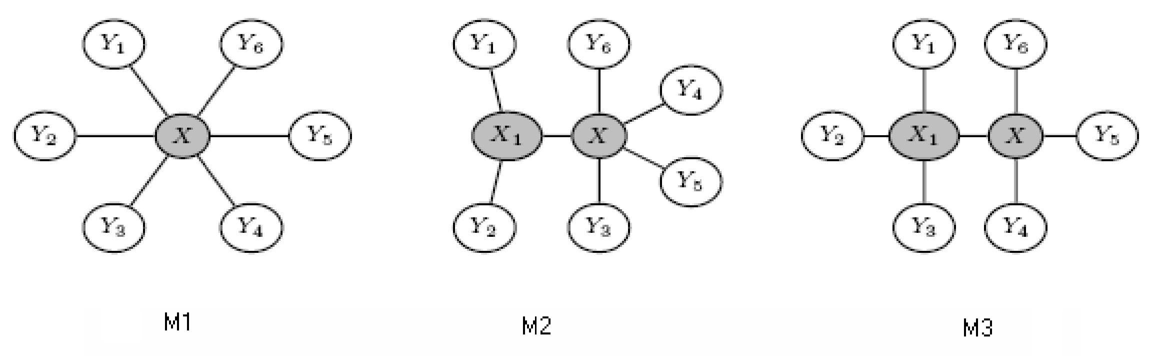
- Heuristic Method (for approximate inference)

The EAST Algorithm

- Search-based algorithm.
- EAST: Expansion, Adjustment, Simplification until Termination

5 Search Operators

- Expansion operators:
  - Node introduction (NI): M1 => M2; |X1| = |X|
    - Constraint: To mediate a latent node and only two of its neighbors
  - State introduction (SI): adds a new state to a latent variable
- Adjustment operator: node relocation (NR), M2 => M3
- Simplification operators: node deletion (ND), state deletion (SD)



Naïve Search

- Start with an initial model
- At each step:
  - Construct all possible candidate models
  - Evaluate them one by one
  - Pick the best one

- Inefficient
  - Too many candidate models
  - Too expensive to run EM on all of them
    - Structural EM assumes fixed set of variables.
    - Does not work here

      Latent variables in models by NI, SI, SD differ from those in current model

Reducing Number of Candidate Models

- Not to use ALL the operators at once.

- How?
  - BIC: $BIC(m|D) = \log P(D|m, \theta^*) - d \log N/2$
  - Improve the two terms alternately

  - SD and ND reduce the penalty term.
  - Which operators to improve the likelihood term?

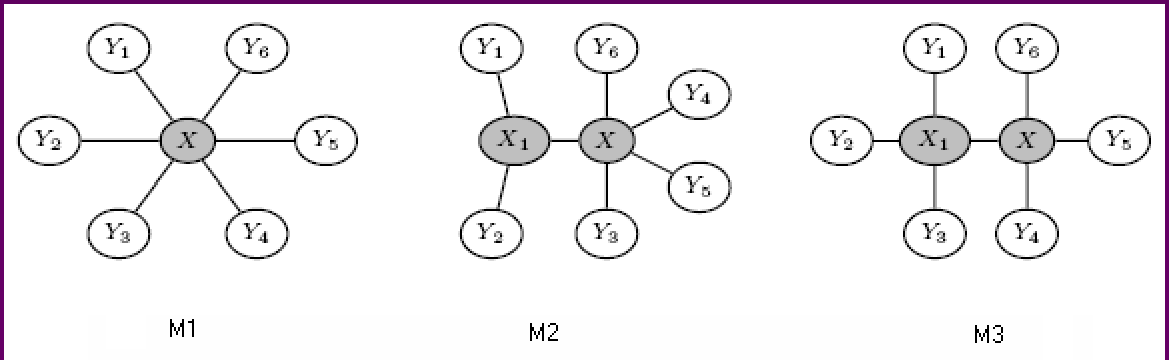
Improve Likelihood Term

- Let be m' obtained from m using NI or SI

  $\log P(D|m', \theta'^*) \geq \log P(D|m, \theta^*)$

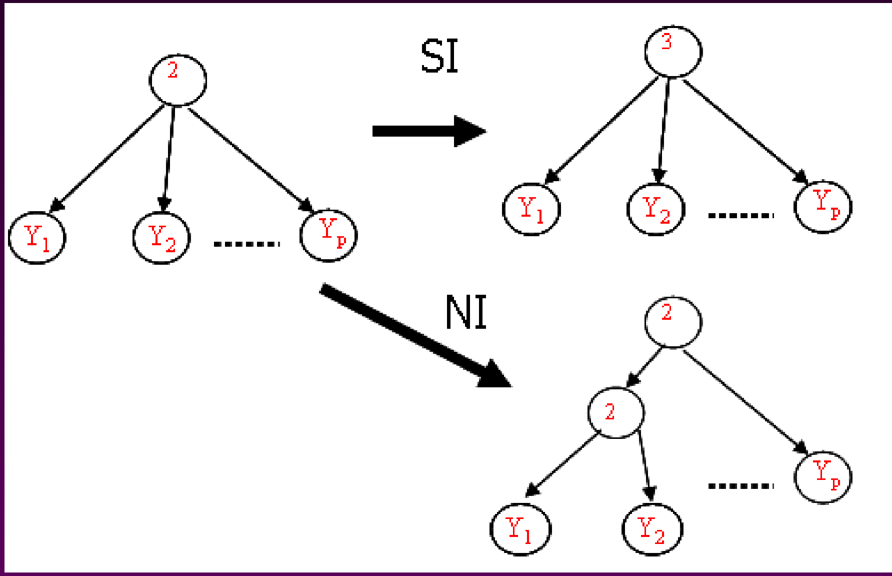
  NI and SI improves the likelihood term

- Follow each NI operation with NR operations.
  - Overcome constraint by NI and allow transition from M1 to M3



Choosing between Models by SI and NI

- Operation Granularity
  - p = 100
  - SI: 101 additional parameters
  - NI: 2 additional parameters
  - Compare shovels with bulldozer
  - SI always preferred initially



- Cost-effectiveness principle
  - Select candidate model with highest improvement ratio

$$IR(m', m|D) = \frac{BIC(m'|D) - BIC(m|D)}{d(m') - d(m)}.$$

The EAST Algorithm

1. Start with a simple initial model
2. Repeat until model score ceases to improve

   EXPANSION: Search with NI, SI

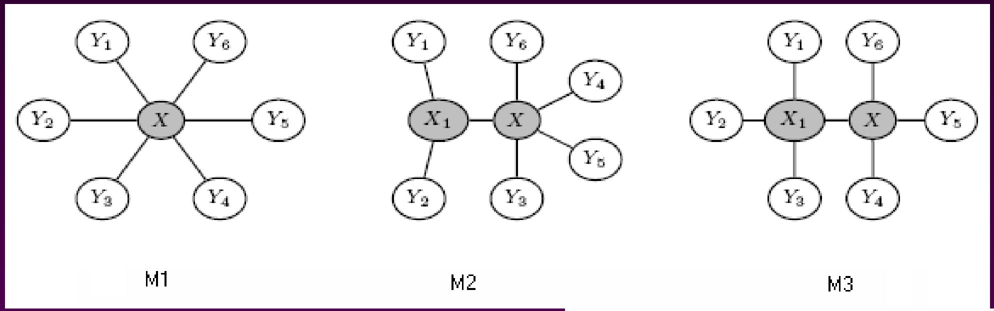
   ADJUSTMENT: Follow each NI operation with NR operations.

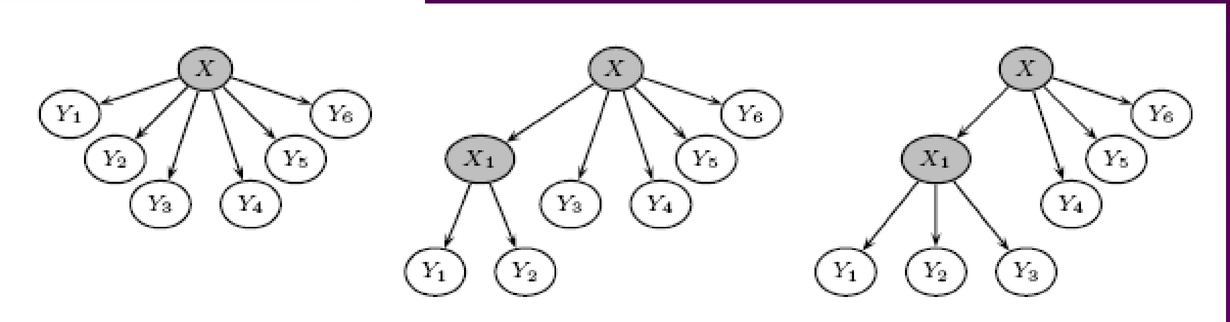
   SIMPLIFICATION: Search with ND, SD

EAST: Expansion, Adjustment, Simplification until Termination

Parameter Sharing

- Internal representation of unrooted model: rooted model





- m: current model;
- m': candidate model generated by applying a search operator on m.
- The two models share many parameters
  - m: (θ 1, θ 2); m': (θ 1, λ 2);

Avoiding EM

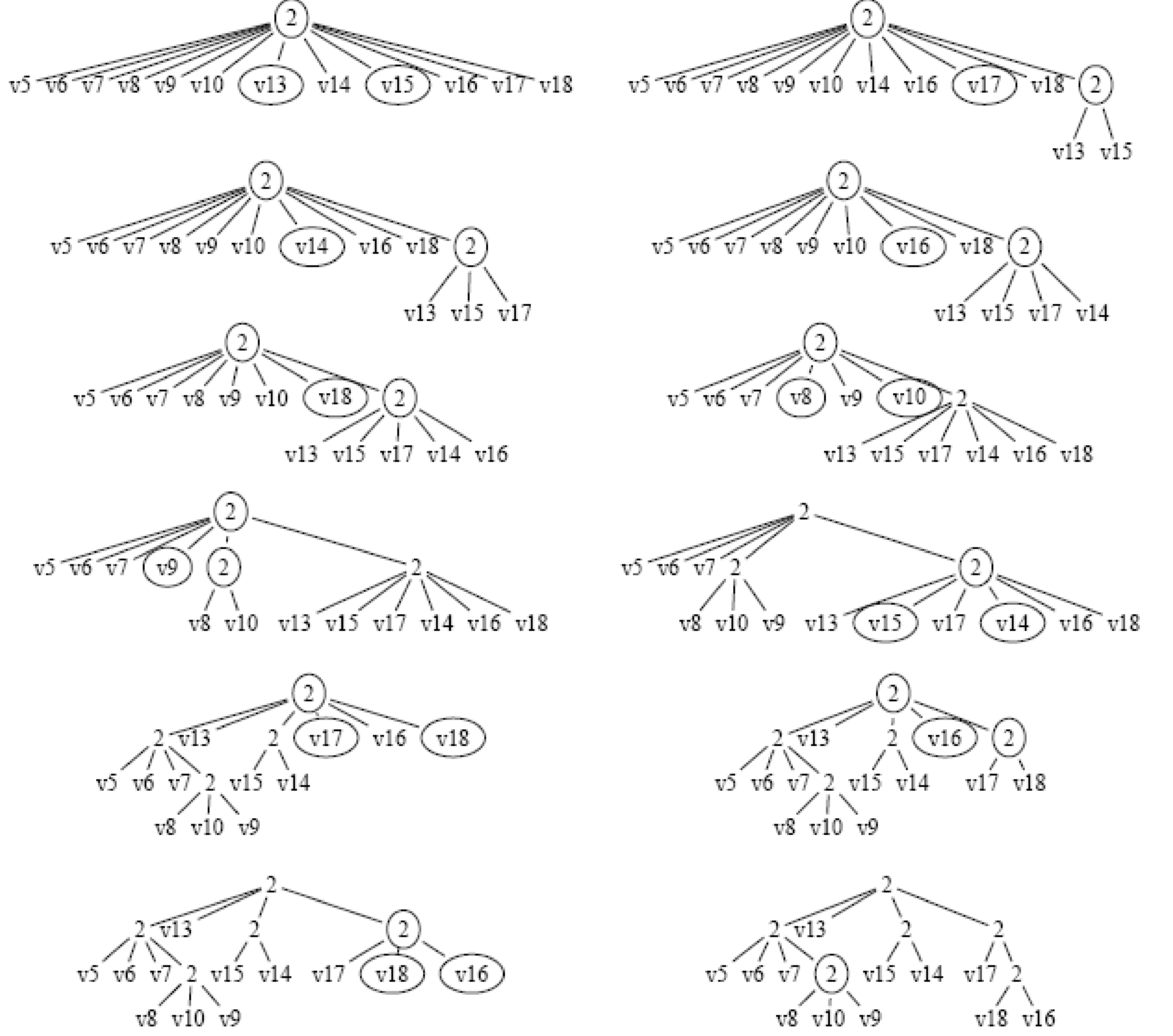
- Run EM to estimate parameters for current model m
  - m: (θ^*_1, θ^*_2);

- Estimate parameters for candidate model m' as follows
  - m': (θ^*_1, λ^*_2);
  - where λ^*_2 is the local MLE

$$\lambda^*_2 = \arg\max_{\lambda_2} \log P(D|m', \theta^*_1, \lambda_2)$$

- Local MLE can be computed efficiently using local EM.

Illustration of the search process



Conclusions

- Latent tree models, and latent structure models in general, offer framework for
  - Probabilistic modeling
    - Approximate reasoning, latent variable in classification
  - Latent structure discovery
  - Multidimensional clustering.
  - Can play a fundamental role in modernizing TCM
  - Can be useful in many other areas
    - such as marketing, survey studies, ….

- We have only scratched the surface. A lot of interesting research work yet to be done.


INTRODUCTION TO
BAYESIAN NETWORKS
贝叶斯网引论
张连文
郭海鹏
著
科学出版社
www.sciencep.com