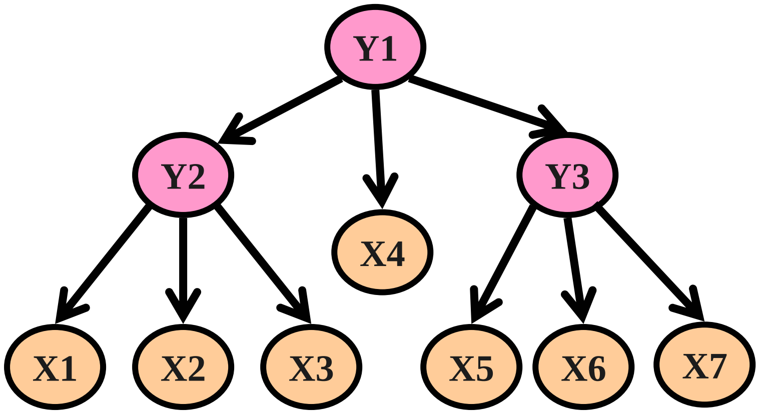
Efficient Model Evaluation in the Search-Based Approach to Latent Structure Discovery

Tao Chen, Nevin L. Zhang and Yi Wang

Department of Computer Science & Engineering
The Hong Kong University of Science & Technology

Latent Tree Models (LTMs)

- Bayesian networks with
  - Rooted tree structure
  - Discrete random variables
  - Leaves observed (manifest variables)
  - Internal nodes latent (latent variables)
- Denoted by (m, θ)
  - m is the model structure
  - θ is the model parameters
- Also known as hierarchical latent class (HLC) models, (Zhang 2004)



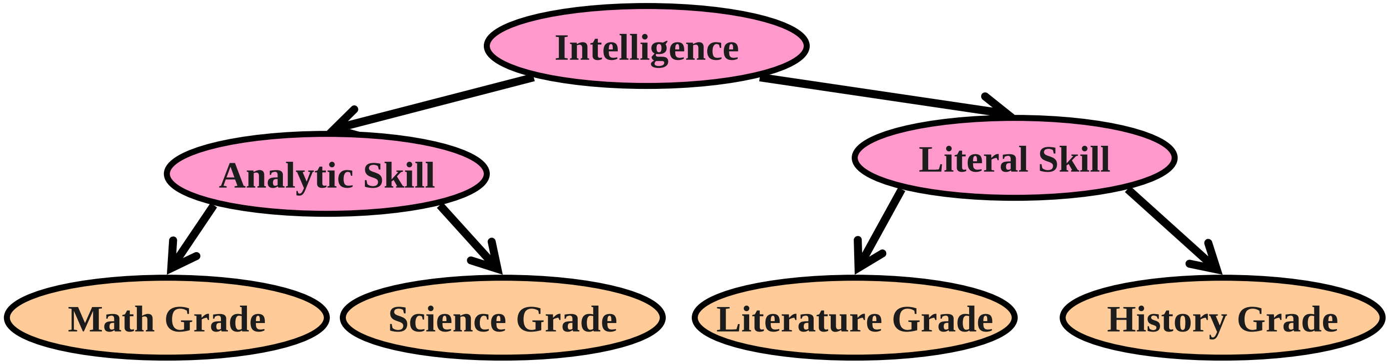
P(Y1),

P(Y2|Y1),

P(X1|Y2), P(X2|Y2), …

Example

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



Learning Latent Tree Models

Search-Based method

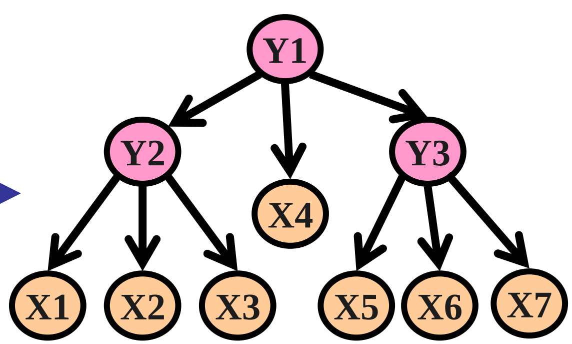
- maximizing the BIC score:

$$BIC(m|D) = \max_{\theta} \log P(D|m, \theta) - d(m) \log N/2$$

Maximized loglikelihood

Penalty

X1	X2	…	X6	X7
1	0	…	1	1
1	1	…	0	0
0	1	…	0	1
…	…	…	…	…



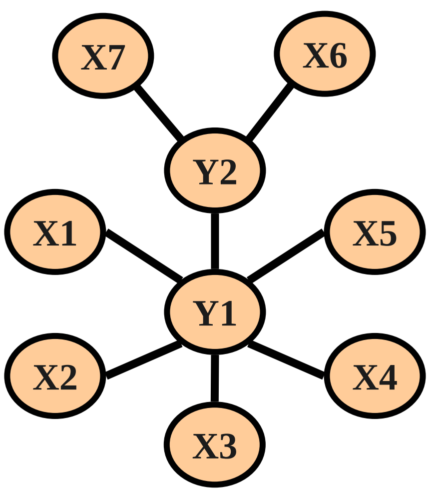
- Number of latent variables
- Cardinality (i.e. number of states) of each latent variable
- Model Structure
- Conditional probability distributions

Outline

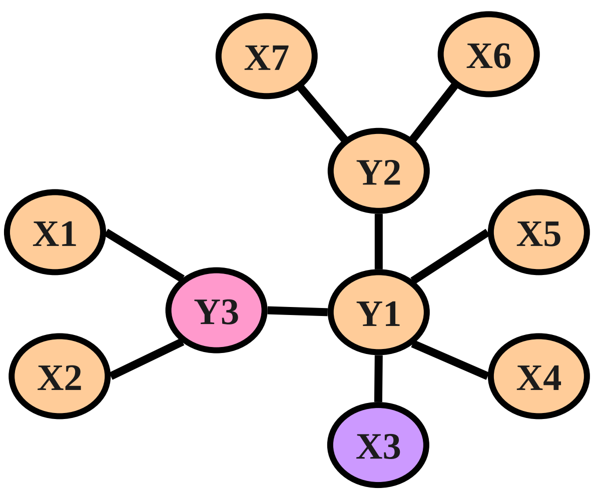
- EAST Search
- Efficient Model Evaluation
- Experiment Results and Explanations
- Conclusions

Search Operators

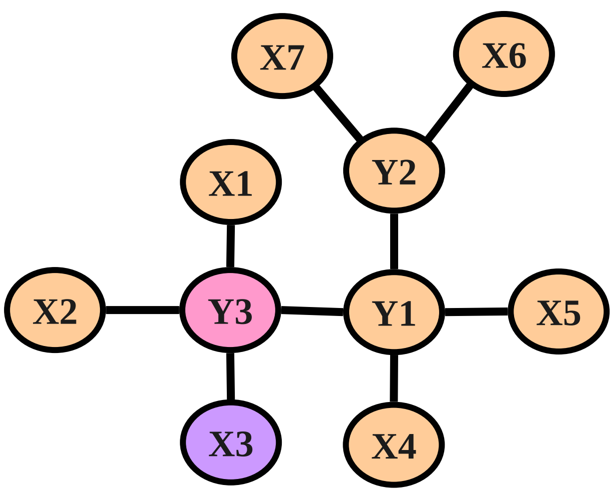
- Expansion operators:
  - Node introduction (NI): $m_1 \Rightarrow m_2$; |Y3| = |Y1|
  - State introduction (SI): add a new state to a latent variable
- Adjustment operator: node relocation (NR), $m_2 \Rightarrow m_3$
- Simplification operators: node deletion (ND), state deletion (SD)



(a) m_1



(a) m_2



(a) m_3

Naïve Search

- At each step:
  - Construct all possible candidate models by applying the search operators to the current model.
  - Evaluate them one by one (BIC)
  - Pick the best one

- Complexity:
  - SI: $O(l)$ l: the number of latent variables in the current model
  - SD: $O(l)$
  - NR: $O(l(l+n))$ n: the number of manifest variables (current)
  - NI: $O(l\, r(r-1)/2)$ r: the maximum number of neighbors (current)
  - ND: $O(l\, r)$

  - **Total** : **T =** $O(l(2 + r/2 + r^2/2 + l + n))$

Reducing Number of Candidate Models

- Reduce number of operators used at each step
- How?

  $\text{BIC}(m|D) = \max_{\theta} \log P(D|m, \theta) - d(m) \log N/2$

- Three phases:
  - <u>Expansion Phase:</u> $O(l\,(1 - r/2 + r^2/2)) < T$
    - Search with expansion operators NI and SI
    - Improve the maximized likelihood term of BIC
  - <u>Simplification Phase:</u> $O(l\,(1+r)) < T$
    - Search with simplification operators ND and SD, separately
    - Reduce penalty term
  - <u>Adjustment Phase:</u> $O(l\,(l+n)) < T$
    - Search with adjustment operators NR
    - Restructure

EAST Search

- Start with a simple initial model
- Repeat until model score ceases to improve
  1. Expansion Phase (NI, SI)
  2. Adjustment Phase (NR)
  3. Simplification Phase (ND, SD)

- EAST: **E**xpansion, **A**djustment, **S**implification until **T**ermination

Outline

- EAST Search
- Efficient Model Evaluation
- Experiment Results and Explanations
- Conclusions

The Complexity of Model Evaluation

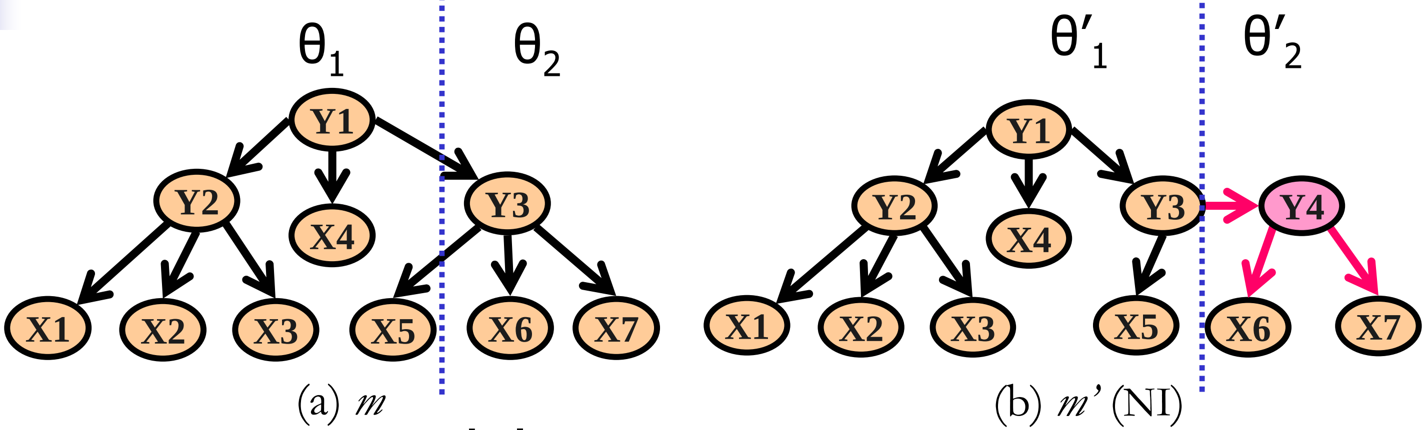
- Compute likelihood term $\max_\theta \log P(D|m, \theta)$ in BIC
- EM algorithm necessary because of latent variables
- EM is an iterative algorithm
  - At each iteration, do inference for every data case

$l = 30$ the number of latent variables in the current model
$n = 70$ the number of manifest variables in the current model

- The complexity of EM algorithm has THREE factors
  1. #of iterations: M = 100
  2. Sample size: N = 10,000
  3. Complexity of inference for one data case is the model size: $O(l + n)$
- Evaluating a candidate model: $O(MN(l + n)) \rightarrow 10^8$

- How to reduce the complexity:
  - Restricted Likelihood (RL) Method
  - Data Completion (DC) Method

Restricted Likelihood: Parameter Composition



- *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', θ_2')

old new

Restricted Likelihood

- Know optimal parameter values for m: (θ_1^*, θ_2^*);
- maximum restricted likelihood:
  - Freezing $\theta_1' = \theta_1^*$ and Varying θ_2'
  - Likelihood $\approx$ Restricted Likelihood

$$\max_{\theta 2'} \log P(D|m', \theta_1^*, \theta_2') \approx \max_{(\theta 1', \theta 2')} \log P(D|m', \theta_1', \theta_2')$$

- **RL based evaluation**: likelihood→restricted likelihood

$$\text{BIC\_RL}(m'|D) = \max_{\theta 2'} \log P(D|m', \theta_1^*, \theta_2') - d(m') \log N/2$$

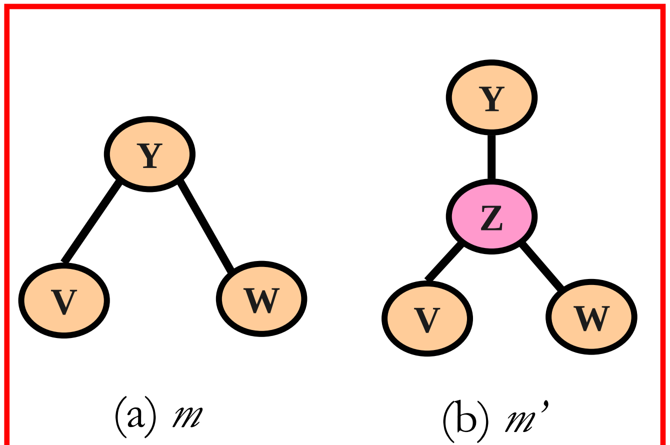
- How the complexity is reduced? (sample size N = 10,000)
  1. Need less iterations before convergence: M′ = 10
  2. Inference is restricted to new parameters: model size = $O(1)$

$$M'N\ O(1) \rightarrow 10^5$$

Data Completion

- Complete data D using (m, θ*) → $\bar{D}$
- Use $\bar{D}$ to evaluate candidate models

NI example



- Null Hypothesis:
  - V and W are conditionally independent given Y
- G-squared Statistic from $\bar{D}$

$$2N \sum_{Y,V,W} \bar{P}(Y,V,W) \log \frac{\bar{P}(V,W|Y)}{\bar{P}(V|Y)\bar{P}(W|Y)}$$

- Model Selection

- How the complexity is reduced? (sample size N = 10,000)
  - No iterations any more
  - Linear in sample size

O(N) → 10^4 (RL: 10^5)

Outline

- EAST Search
- Efficient Model Evaluation
- Experiment Results and Explanations
- Conclusions

RL vs. DC: Data Analysis

- **Two Algorithms: EAST-RL and EAST-DC**
- **Date sets:**
  - Synthetic data

	1k	5k	10k
m_7	$\mathcal{D}_7(1k)$	$\mathcal{D}_7(5k)$	$\mathcal{D}_7(10k)$
m_{12}	$\mathcal{D}_{12}(1k)$	$\mathcal{D}_{12}(5k)$	$\mathcal{D}_{12}(10k)$
m_{18}	$\mathcal{D}_{18}(1k)$	$\mathcal{D}_{18}(5k)$	$\mathcal{D}_{18}(10k)$

  - Real-world data

	num vars	num states per var	sample size	
			train	test
ICAC	31	3.5	1200	301
KIDNEY	35	4.0	2000	600
COIL	42	2.7	5822	4000
DEPRESSION	100	2	500	104

- **Quality measure:**
  - Synthetic: empirical KL divergence (approximate); 10 runs
  - Real-world: logarithmic score on testing data (prediction); 5 runs

RL vs. DC: Efficiency

- Synthetic data:

time	D_7(1k)	D_7(5k)	D_7(10k)	D_{12}(1k)	D_{12}(5k)	D_{12}(10k)	D_{18}(1k)	D_{18}(5k)	D_{18}(10k)
RL	.7	7.1	8.3	17.2	1.4	2.6	.7	6.0	18.4
DC	.6	5.8	8.4	6.6	0.7	1.4	.6	3.9	8.2
RL/DC	1.1	1.2	1.0	2.6	2.0	1.9	1.2	1.5	2.2

- Real-world data:

time	ICAC	KID.	COIL	DEP.
RL	0.22	1.00	2.31	3.58
DC	0.09	0.27	0.68	0.58
RL/DC	2.4	3.7	3.4	6.2

RL vs. DC: Model Quality

- Synthetic data:
  - 12 and 18 variables : EAST_RL beats EAST_DC
  - 7 variables : identical models

emp-KL	D_{12}(1k)	D_{12}(5k)	D_{12}(10k)	D_{18}(1k)	D_{18}(5k)	D_{18}(10k)
RL	.0999	.0311	.0032	.1865	.0148	.0047
DC	.1659	.0590	.0051	.2171	.0371	.0113
DC/RL	1.7	1.9	1.6	1.2	2.5	2.4

- Real-world data: EAST_RL beats EAST_DC

logScore	ICAC	KID.	COIL	DEP.
RL	-6172	-16761	-34121	-4220
DC	-6231	-17236	-35025	-4392
Ratio	0.6%	2.8%	2.6%	3.9%

Theoretical Relationships

- **Objective function: BIC functions**
  - Resort to RL and DC due to hardness
  - How RL and DC are related to BIC?

- Proposition 1 (RL and BIC) : For any candidate model m' obtained from the current model m,

  RL functions $\leqslant$ BIC functions.

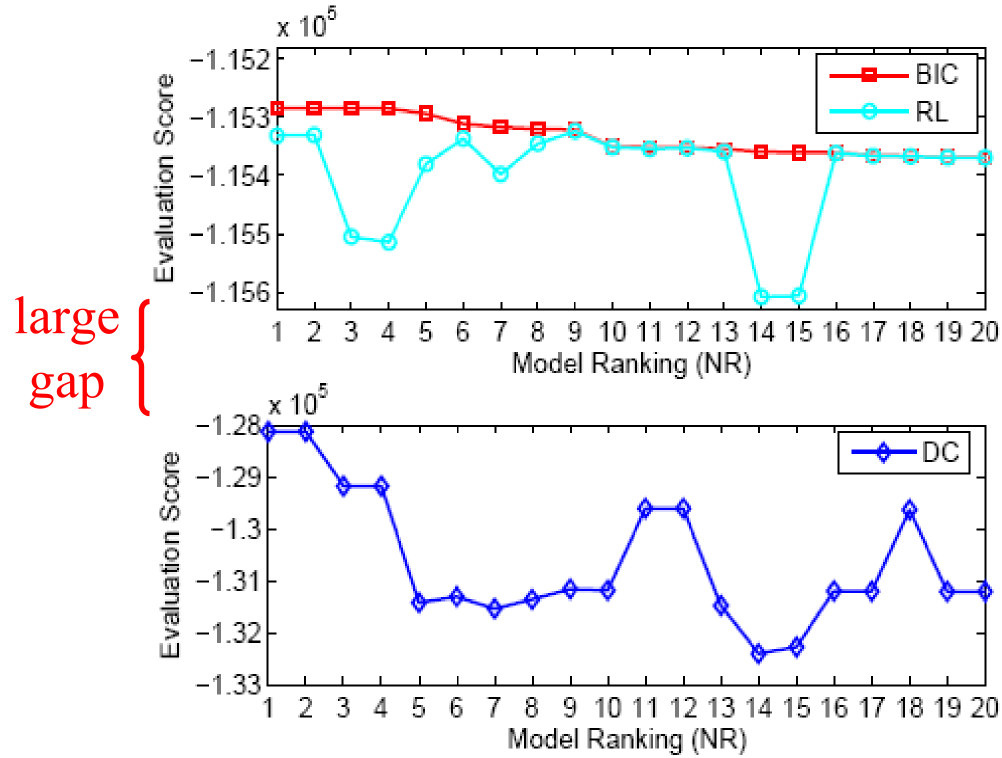
- Proposition 2 (DC and BIC): For any candidate model m' obtained from the current model m using the NR, ND or SD operator,

  DC functions (NR, ND and SD) $\leqslant$ BIC functions (NR, ND and SD)

No clear relations between DC and BIC functions in the case of SI and NI operators.

Comparison of Function Values

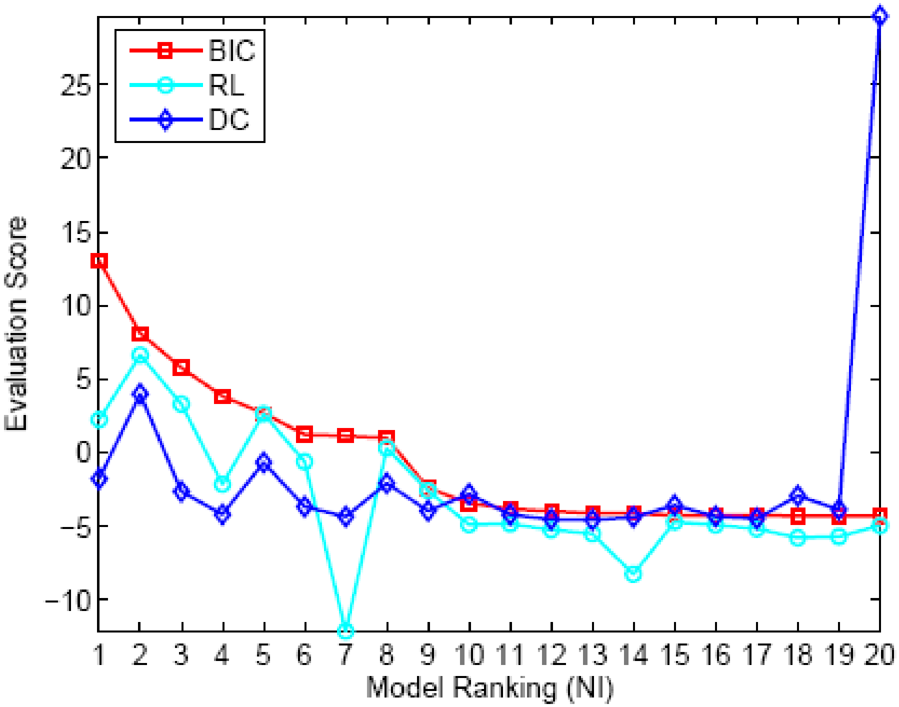
- **RL functions**
  - Tight lower bound BIC
- **DC functions**
  - Lower bound BIC
  - Far away from BIC



- **Similar stories on ND, SD.**

Comparison of Function Values

- **RL functions:**
  - Lower bound
  - Tight in most cases
  - Good ranking
- **DC functions:**
  - Not lower bound
  - Bad ranking



Comparison of Model Selection

# steps	$\mathcal{D}_7(1k)$	$\mathcal{D}_7(5k)$	$\mathcal{D}_7(10k)$	$\mathcal{D}_{12}(1k)$	$\mathcal{D}_{12}(5k)$	$\mathcal{D}_{12}(10k)$	$\mathcal{D}_{18}(1k)$	$\mathcal{D}_{18}(5k)$	$\mathcal{D}_{18}(10k)$
RL > DC	0	0	0	5	5	5	4	7	4
RL = DC	0	2	2	7	14	15	16	25	28
RL < DC	0	0	0	0	0	1	0	3	2
Total	0	2	2	12	19	21	20	35	34

- **D_7(1k), D_7(5k), D_7(10k)**
  - RL and DC picked the same models
- **The other 6 data sets**
  - Most steps : the same models
  - Quite a number of steps : RL picked better models.

Performance Difference Explained

- EAST_RL uses RL functions in model evaluation
- EAST_DC uses DC functions in model evaluation

- RL functions are more closely related to BIC functions than DC functions
  - Theoretically
  - Empirically

- Model Selection
  - RL picks better models than DC during search

- EAST_RL finds better models than EAST_DC

Conclusions

- EAST Search
- Efficient Model Evaluation
  - RL: find better models
  - DC: more efficient
- Deeper understanding →
  new search-based algorithms (future work)

Thank you!