High-Dimensional Structure Learning of Graphical Models: Trees, Latent Trees & Beyond

Anima Anandkumar

Electrical Engineering and Computer Science
U.C. Irvine

Joint work with Myung Jin Choi, Vincent Tan, and Alan Willsky.

UIUC Seminar

Graphical Models: Motivation

Example: Contextual Object Recognition

SKY

ROAD AREA

TREE CAR CAR PEOPLE ROAD

TRAFFIC LIGHT CROSSWALK

Graphical Models: Motivation

Example: Contextual Object Recognition

SKY

ROAD AREA

TREE CAR CAR PEOPLE ROAD

TRAFFIC LIGHT CROSSWALK

Robust Recognition using Context

- Multivariate distribution over set of known object categories
- Co-occurrence probabilities for different objects to occur in the same image

Graphical Models: Motivation

Example: Contextual Object Recognition

SKY

FLOOR ROAD AREA

WALL TREE CAR CAR PEOPLE ROAD

TRAFFIC LIGHT CROSSWALK

Robust Recognition using Context

- Multivariate distribution over set of known object categories
- Co-occurrence probabilities for different objects to occur in the same image

Motivation Using Contextual Object Recognition

SKY

FLOOR ROAD AREA

WALL TREE CAR CAR PEOPLE ROAD

TRAFFIC LIGHT CROSSWALK

M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting Hierarchical Context on a Large Database of Object Categories", to appear at 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Motivation Using Contextual Object Recognition

SKY

FLOOR ROAD AREA

WALL TREE CAR CAR PEOPLE ROAD

TRAFFIC LIGHT CROSSWALK

Challenges in Using Context :Curse of Dimensionality

- Many training images for learning and complexity inference for testing
- SUN09 dataset with ~ 100 object categories, ~ 4000 training images.
- Require learning $\sim 2^{100}$ co-occurrence probability table
- Object recognition in test images: search over probability table

M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting Hierarchical Context on a Large Database of Object Categories", to appear at 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Motivation Using Contextual Object Recognition

SKY

FLOOR ROAD AREA

WALL TREE CAR CAR PEOPLE ROAD

TRAFFIC LIGHT CROSSWALK

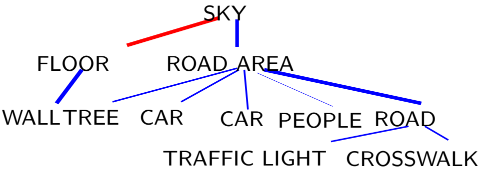
Challenges in Using Context :Curse of Dimensionality

- Many training images for learning and complexity inference for testing
- SUN09 dataset with ~ 100 object categories, ~ 4000 training images.
- Require learning $\sim 2^{100}$ co-occurrence probability table
- Object recognition in test images: search over probability table

Succinct representation of contextual image information as a graphical model

M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting Hierarchical Context on a Large Database of Object Categories", to appear at 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Motivation Using Contextual Object Recognition



Challenges in Using Context :Curse of Dimensionality

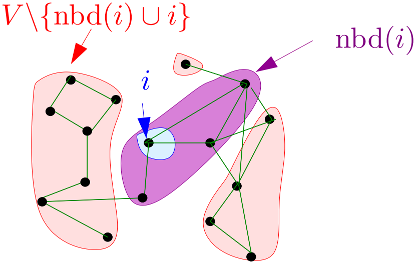
- Many training images for learning and complexity inference for testing
- SUN09 dataset with ~ 100 object categories, ~ 4000 training images.
- Require learning $\sim 2^{100}$ co-occurrence probability table
- Object recognition in test images: search over probability table

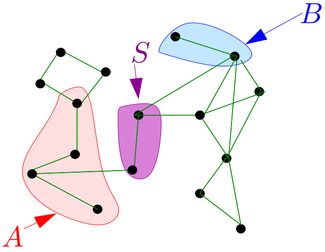
Succinct representation of contextual image information as a graphical model

M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting Hierarchical Context on a Large Database of Object Categories", to appear at 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Graphical Models: Introduction

- Graph structure $G = (V, E)$ in the multivariate distribution of random variables, with $V = \{1, \ldots, m\}$.
- Nodes $i \in V$ correspond to random variable X_i.
- Edges E correspond to conditional independence relationships.





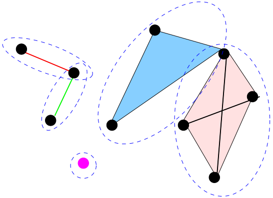
$X_i \perp\!\!\!\perp \mathbf{X}_{V \setminus \{\mathrm{nbd}(i) \cup i\}} | \mathbf{X}_{\mathrm{nbd}(i)}$

$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$

From Conditional Independence to Gibbs Distribution

Hammersley-Clifford Theorem'71

Let P be joint pmf of model with graph $G = (V, E)$,



$$P(\mathbf{x}) = \frac{1}{Z} \exp[\sum_{c \in \mathcal{C}} \Psi_c(\mathbf{x}_c)].$$

where $\mathcal{C}$ is the set of maximal cliques.

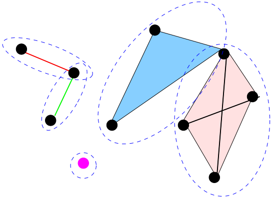
From Conditional Independence to Gibbs Distribution

Hammersley-Clifford Theorem'71

Let P be joint pmf of model with graph $G = (V, E)$,

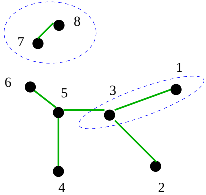
$$P(\mathbf{x}) = \frac{1}{Z} \exp[\sum_{c \in \mathcal{C}} \Psi_c(\mathbf{x}_c)].$$

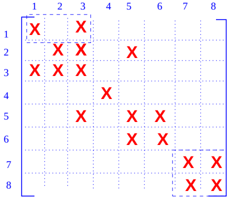
where $\mathcal{C}$ is the set of maximal cliques.



Gaussian Graphical Models

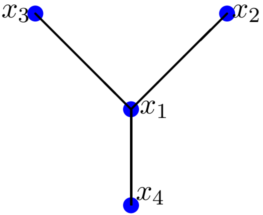
Dependency Graph





Inverse of Covariance Matrix

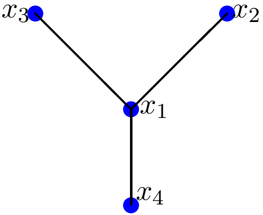
Tree Structured Graphical Models



$$P(\mathbf{x}) = \prod_{i \in V} P_i(x_i) \prod_{(i,j) \in E} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)}$$

$$= P_1(x_1) \frac{P_{1,2}(x_1, x_2)}{P_1(x_1)} \frac{P_{1,3}(x_1, x_3)}{P_1(x_1)} \frac{P_{1,4}(x_1, x_4)}{P_1(x_1)}.$$

Tree Structured Graphical Models



$$P(\mathbf{x}) = \prod_{i \in V} P_i(x_i) \prod_{(i,j) \in E} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)}$$

$$= P_1(x_1) \frac{P_{1,2}(x_1, x_2)}{P_1(x_1)} \frac{P_{1,3}(x_1, x_3)}{P_1(x_1)} \frac{P_{1,4}(x_1, x_4)}{P_1(x_1)}.$$

Tree Graphical Models: Tractable Learning & Inference

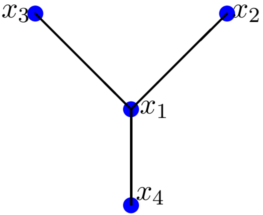
- Maximum likelihood learning of tree structure is tractable

  Chow-Liu Algorithm (1968)

- Inference on tree models is tractable

  Belief Propagation

Tree Structured Graphical Models



$$P(\mathbf{x}) = \prod_{i \in V} P_i(x_i) \prod_{(i,j) \in E} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)}$$

$$= P_1(x_1) \frac{P_{1,2}(x_1, x_2)}{P_1(x_1)} \frac{P_{1,3}(x_1, x_3)}{P_1(x_1)} \frac{P_{1,4}(x_1, x_4)}{P_1(x_1)}.$$

Tree Graphical Models: Tractable Learning & Inference

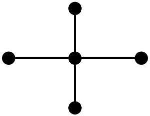
- Maximum likelihood learning of tree structure is tractable
  - Chow-Liu Algorithm (1968)
- Inference on tree models is tractable
  - Belief Propagation

What other classes of graphical models are tractable for learning and inference?

Graphical Models: Trees & Beyond

Analysis of Tree Structure Learning: Extremal Trees for Learning

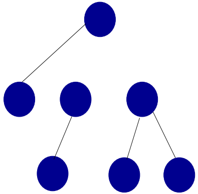
Star



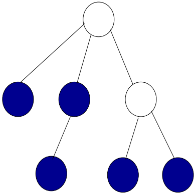
Chain

Structure Learning in Graphical Models Beyond Trees

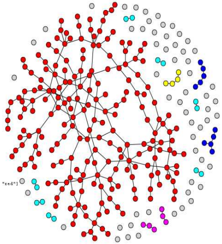
Forests



Latent Trees



Random Graphs



High Dimensional Learning of Graphical Models

- Given n i.i.d. samples $\mathbf{x}^n$ from model P with structure G
- Information about model class, e.g., trees, forests, latent trees etc.
- Output estimated structure $\widehat{G}$ and model $\widehat{P}$

Structural Consistency

$$\lim_{n\to\infty} \Pr(\{\mathbf{x}^n : \widehat{G}^n \neq G\}) = 0.$$

Sample Complexity: High Dimensional Regime

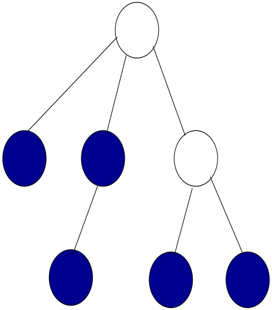
- m is number of observed nodes in the graphical model.
- m can be large compared to n
- When $n > f(m;\delta)$, $P_{err}(n) < \delta$, for every $\delta > 0$, then sample complexity is $\Omega(f(m))$

Structure Learning Algorithms with Low Sample Complexity

Outline

Result 1: Learning Latent Tree Models

- Latent tree model is a tree model on $W := V \cup H$
- Visible Nodes V, Hidden Nodes H.



Latent Tree Reconstruction

- Given n IID samples from node set V, estimate latent tree model
- No knowledge on number of hidden variables

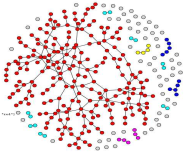
Result 1: Learning Latent Trees Contd.,

Reconstruction of general latent tree models from samples

- Propose two novel algorithms under unified approach for Gaussian and discrete models
- Provide theoretical guarantees: consistency, computational and sample complexities
  - Structural and risk consistency for any minimal latent tree
  - Sample complexity of $\Omega(\log m)$ for m observed nodes when effective depth is constant
  - Low computational complexity
- Experimental results demonstrate efficiency of methods

M.J. Choi, V. Tan, A. Anandkumar & A. Willsky, "Learning Latent Tree Graphical Models," Submitted to *J. of Machine Learning Research*, available on Arxiv.

Result 2: Learning Random Graphs



- Binary discrete (Ising) model on Erdős-Rényi random graphs $G_m \sim \mathcal{G}(m, c/m)$
- n samples available at nodes to estimate structure

Challenges

- Random graphs have many large degrees nodes
- Previous algorithms cannot guarantee consistent estimation

Intuitions

- Random graphs are locally tree-like
- Correlation decay: Effect of faraway nodes negligible, model behaves locally as a tree distribution

---

A. Anandkumar, V. Tan, A. S. Willsky "High Dimensional Structure Learning of Ising Models on Sparse Random Graphs," preprint on webpage.

Result 2: Learning Random Graphs Contd.,

- Propose two local algorithms
- Analyze structure learning performance under correlation decay

Conditional Mutual Information Thresholding

- Consistent structure learning under correlation decay
- Require number of samples $n = \omega(\log m)$

Correlation Thresholding

- Finite edit distance under correlation decay
- Consistent structure reconstruction under additional conditions
- Require number of samples $n = \Omega(\log m)$

Lower bound on sample complexity

Require $n = \Omega(c \log m)$ samples to estimate random graphical structures

Related Work in Structure Learning

Efficient Algorithms for Structure Learning

- ML for trees: Max. weight spanning tree with mutual information weights (Chow & Liu 68)
- Causal dependence trees: directed mutual information (Quinn, Coleman & Kiyavash '10)
- Tree augmented models: (Santhanam, Dingel, & Milenkovic, '09)
- Convex relaxation methods: ℓ_1 regularization

  Gaussian Graphical Models (Meinshausen and Buehlmann 2006)
  Logistic regression for Ising models (Ravikumar et. al. 10)
- Brute-force conditional independence test for bounded degree graphs (Bresler et. al. '09)
- Greedy modification for large-girth graphs under correlation decay (Netrapalli et. al. '10)
- Learning thin junction trees through conditional mutual information tests (Chechetka et. al. '07)

Related Work Contd.

Lower Bounds on Sample Complexity

- Information-theoretic bounds for bounded degree graphs (Santhanam & Wainwright '08, Wang et. al. '10)
- Strong converse bounds for bounded degree graphs (Mitliagkas & Vishwanath '10)

Latent Graphical Models

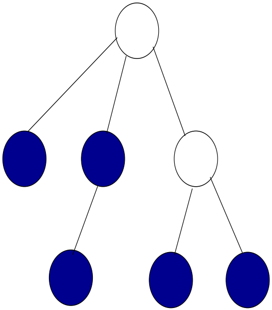
- Neighborhood joining: Fast implementation but large sample complexity (Saitou & Nei '87)
- Quartet methods: Local tests but non-trivial merging (Erdos et. al 99, Attenson 99, Daskalakis et al. 06)
- Expectation Maximization: Greedy local structural search (Kemp & Tenenbaum 08, Zhang & Kocka 04, Elidan & Friedman 05)
- Convex Methods: Sparse observed graph and small number of hidden variables (Chandrasekaran et. al. '10)

Outline

Outline

Latent Tree Model

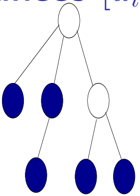
- Visible Nodes V, Hidden Nodes H and $W := V \cup H$
- $T = (W, E)$ is a tree on W



Latent Tree Reconstruction

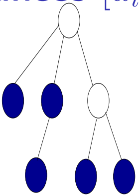
Given n IID samples from node set V, reconstruct latent tree model

Information Distances $[d_{i,j}]$ on Tree Models



Gaussian Model: $\mathbf{X}_W \sim N(\mathbf{0}, \mathbf{\Sigma})$, $d_{ij} := -\log|\rho_{ij}|$, $\rho_{ij} := \frac{\mathbf{\Sigma}_{ij}}{\sqrt{\mathbf{\Sigma}_{ii}\mathbf{\Sigma}_{jj}}}$

Information Distances $[d_{i,j}]$ on Tree Models

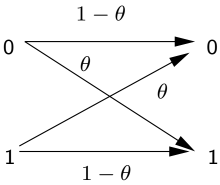


Gaussian Model: $\mathbf{X}_W \sim N(\mathbf{0}, \mathbf{\Sigma})$, $d_{ij} := -\log|\rho_{ij}|$, $\rho_{ij} := \frac{\mathbf{\Sigma}_{ij}}{\sqrt{\mathbf{\Sigma}_{ii}\mathbf{\Sigma}_{jj}}}$

Discrete Symmetric Model

- $X_i \in \{1, 2, \ldots, K\}$ and for $\theta_{ij} \in (0, 1/K)$,

$$P(x_i|x_j) = \begin{cases} 1 - (K-1)\theta_{ij} & \text{if } x_i = x_j \\ \theta_{ij}, & \text{o.w.} \end{cases}$$



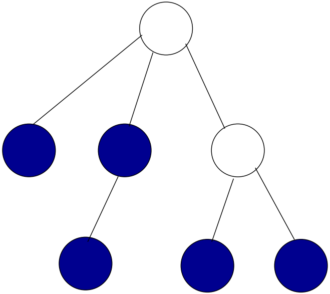
- node marginal is uniform
- Distance is $d_{i,j} := -\log(1 - K\theta_{ij})$.

Markov property on information distances

Markov Property on Trees: $[d_{i,j}]$ is a tree metric

$$d_{k,l} = \sum_{(i,j) \in \mathrm{Path}(k,l;E)} d_{i,j},$$

where $\mathrm{Path}(k,l;E)$ is the path from k to l along edges E of tree.

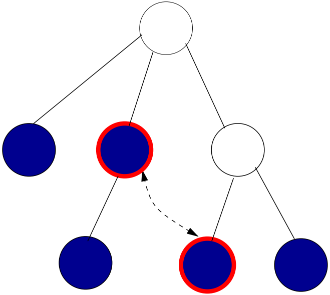


Markov property on information distances

Markov Property on Trees: $[d_{i,j}]$ is a tree metric

$$d_{k,l} = \sum_{(i,j) \in \mathrm{Path}(k,l;E)} d_{i,j},$$

where $\mathrm{Path}(k,l;E)$ is the path from k to l along edges E of tree.

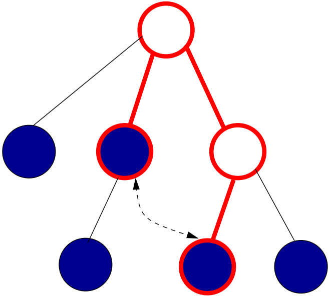


Markov property on information distances

Markov Property on Trees: $[d_{i,j}]$ is a tree metric

$$d_{k,l} = \sum_{(i,j)\in \mathrm{Path}(k,l;E)} d_{i,j},$$

where $\mathrm{Path}(k,l;E)$ is the path from k to l along edges E of tree.



Minimal Tree Extensions

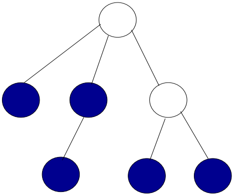
Minimal Tree Extension (Pearl 88)

Tree with least hidden variables explaining observed statistics

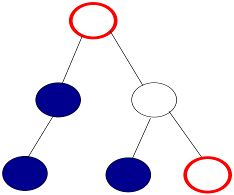
Conditions for Minimality

- Each hidden variable has at least three neighbors: Leaves are visible
- No two variables are perfectly dependent or independent:

$$0 < l \leq d_{i,j} \leq u < \infty, \quad \forall\ (i,j) \in E.$$



Minimal



Non-minimal

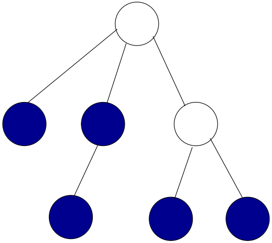
Outline

Siblings Test Based on Information Distances

Exact Statistics: Distances $[d_{i,j}]$

Let $\Phi_{ijk} := d_{i,k} - d_{j,k}$.

- $-d_{i,j} < \Phi_{ijk} = \Phi_{ijk'} < d_{i,j}\ \forall\, k, k' \neq i, j,\ \iff\ i, j$ leaves with common parent
- $\Phi_{ijk} = d_{i,j},\ \forall\, k \neq i, j,\ \iff\ i$ is a leaf and j is its parent.



Sample Statistics: ML Estimates $[\hat{d}_{i,j}]$

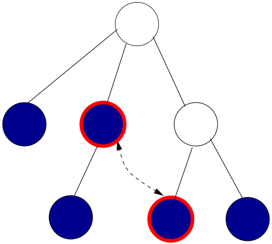
Use only short distances: $d_{i,k}, d_{j,k} < \tau$, Relax equality relationships

Siblings Test Based on Information Distances

Exact Statistics: Distances $[d_{i,j}]$

Let $\Phi_{ijk} := d_{i,k} - d_{j,k}$.

- $-d_{i,j} < \Phi_{ijk} = \Phi_{ijk'} < d_{i,j} \; \forall \, k, k' \neq i, j, \iff i, j$ leaves with common parent
- $\Phi_{ijk} = d_{i,j}, \; \forall \, k \neq i, j, \iff i$ is a leaf and j is its parent.



Sample Statistics: ML Estimates $[\hat{d}_{i,j}]$

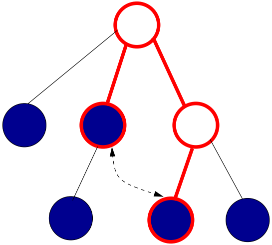
Use only short distances: $d_{i,k}, d_{j,k} < \tau$, Relax equality relationships

Siblings Test Based on Information Distances

Exact Statistics: Distances $[d_{i,j}]$

Let $\Phi_{ijk} := d_{i,k} - d_{j,k}$.

- $-d_{i,j} < \Phi_{ijk} = \Phi_{ijk'} < d_{i,j} \; \forall\, k, k' \neq i, j,$ $\iff$ i, j leaves with common parent
- $\Phi_{ijk} = d_{i,j}, \; \forall\, k \neq i, j,$ $\iff$ i is a leaf and j is its parent.



Sample Statistics: ML Estimates $[\hat{d}_{i,j}]$

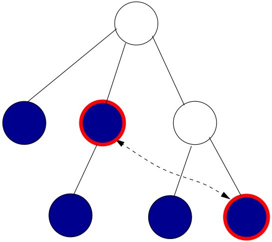
Use only short distances: $d_{i,k}, d_{j,k} < \tau$, Relax equality relationships

Siblings Test Based on Information Distances

Exact Statistics: Distances $[d_{i,j}]$

Let $\Phi_{ijk} := d_{i,k} - d_{j,k}$.

- $-d_{i,j} < \Phi_{ijk} = \Phi_{ijk'} < d_{i,j}\ \forall\, k, k' \neq i, j, \iff i, j$ leaves with common parent
- $\Phi_{ijk} = d_{i,j},\ \forall\, k \neq i, j, \iff i$ is a leaf and j is its parent.



Sample Statistics: ML Estimates $[\hat{d}_{i,j}]$

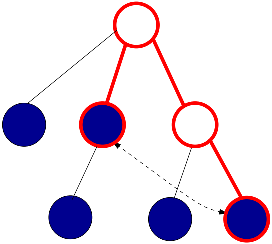
Use only short distances: $d_{i,k}, d_{j,k} < \tau$, Relax equality relationships

Siblings Test Based on Information Distances

Exact Statistics: Distances $[d_{i,j}]$

Let $\Phi_{ijk} := d_{i,k} - d_{j,k}$.

- $-d_{i,j} < \Phi_{ijk} = \Phi_{ijk'} < d_{i,j}\ \forall\, k, k' \neq i, j,$ $\iff$ i, j leaves with common parent
- $\Phi_{ijk} = d_{i,j},\ \forall\, k \neq i, j,$ $\iff$ i is a leaf and j is its parent.



Sample Statistics: ML Estimates $[\hat{d}_{i,j}]$

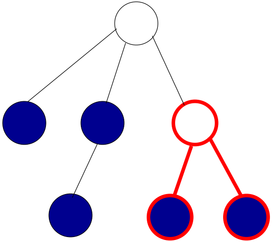
Use only short distances: $d_{i,k}, d_{j,k} < \tau$, Relax equality relationships

Siblings Test Based on Information Distances

Exact Statistics: Distances $[d_{i,j}]$

Let $\Phi_{ijk} := d_{i,k} - d_{j,k}$.

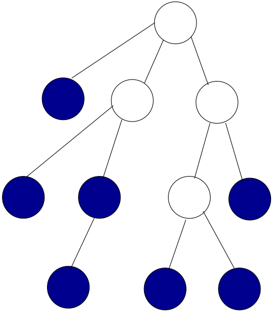
- $-d_{i,j} < \Phi_{ijk} = \Phi_{ijk'} < d_{i,j} \ \forall\, k, k' \neq i, j,$ $\iff$ i, j leaves with common parent
- $\Phi_{ijk} = d_{i,j}, \ \forall\, k \neq i, j,$ $\iff$ i is a leaf and j is its parent.



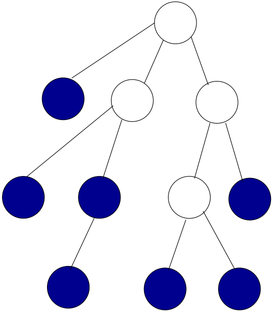
Sample Statistics: ML Estimates $[\hat{d}_{i,j}]$

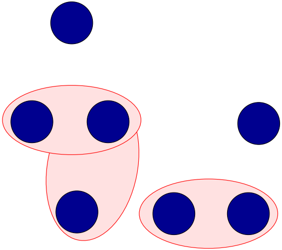
Use only short distances: $d_{i,k}, d_{j,k} < \tau$, Relax equality relationships

Recursive Grouping: Example and Guarantees

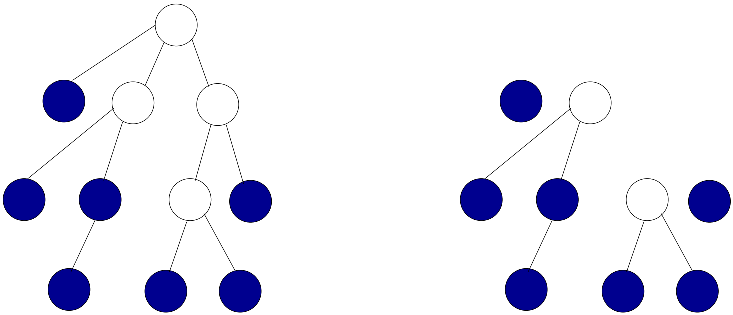


Recursive Grouping: Example and Guarantees

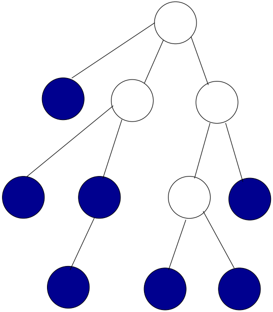


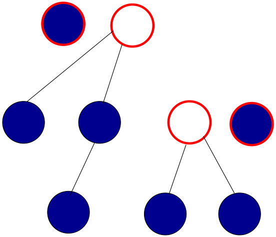


Recursive Grouping: Example and Guarantees

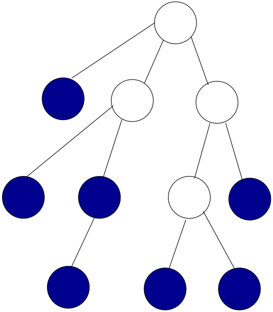


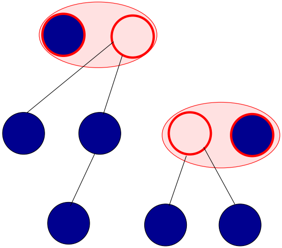
Recursive Grouping: Example and Guarantees



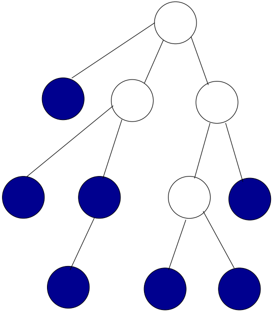


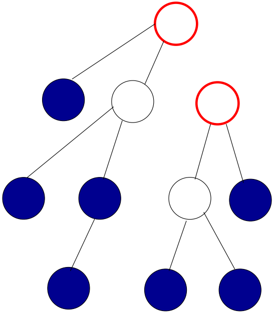
Recursive Grouping: Example and Guarantees



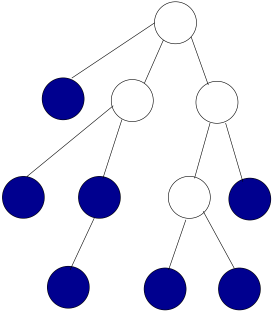


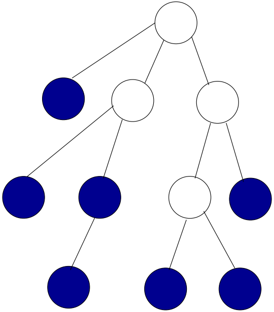
Recursive Grouping: Example and Guarantees



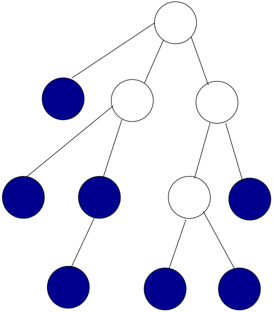


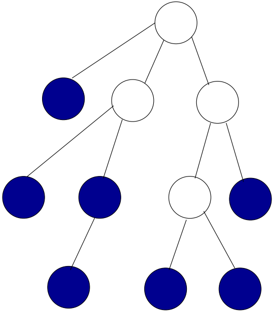
Recursive Grouping: Example and Guarantees





Recursive Grouping: Example and Guarantees





Guarantees

- Structural and estimation consistency for all minimal latent trees
- Sample complexity of $\Omega(\log m)$ for m observed nodes when effective depth is fixed
- Computational complexity of $O(\mathrm{diam}(\hat{T})m^3)$.

Outline

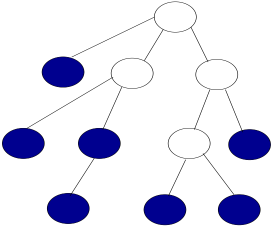
Overview of Chow-Liu Based Grouping

Shortcomings of Recursive Grouping

- Computationally intensive: check all observed node pairs as siblings
- Sibling test: local test. Error prone

Pre-processing to improve efficiency and accuracy

Build a Chow-Liu tree, rule out many pairs of observed nodes as siblings



Reconstruct Latent Tree by Transforming Chow-Liu Tree

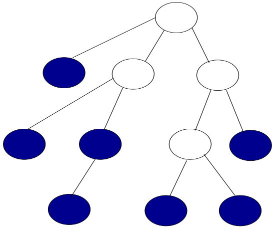
Overview of Chow-Liu Based Grouping

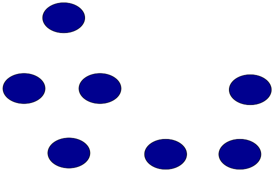
Shortcomings of Recursive Grouping

- Computationally intensive: check all observed node pairs as siblings
- Sibling test: local test. Error prone

Pre-processing to improve efficiency and accuracy

Build a Chow-Liu tree, rule out many pairs of observed nodes as siblings





Reconstruct Latent Tree by Transforming Chow-Liu Tree

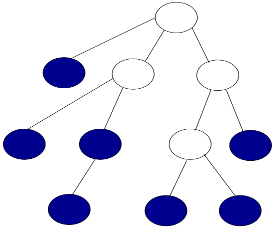
Overview of Chow-Liu Based Grouping

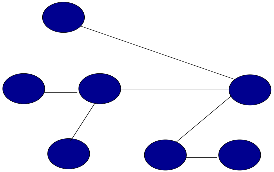
Shortcomings of Recursive Grouping

- Computationally intensive: check all observed node pairs as siblings
- Sibling test: local test. Error prone

Pre-processing to improve efficiency and accuracy

Build a Chow-Liu tree, rule out many pairs of observed nodes as siblings





Reconstruct Latent Tree by Transforming Chow-Liu Tree

Chow-Liu Tree on Observed Nodes

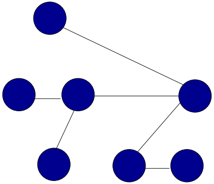
Chow-Liu tree: ML tree over observed nodes V

- $\widehat{P}_{\mathrm{CL}}$: Tree distribution closest (in KL-divergence) to the empirical distribution

$$\widehat{P}_{\mathrm{CL}} := \operatorname*{argmin}_{Q \in \mathsf{Tree}} D(\widehat{P} \,||\, Q).$$

- Chow-Liu algorithm: $\widehat{T}_{\mathrm{CL}} = \operatorname*{argmax}_{T=(V,E)\in\mathcal{T}} \sum_{e\in E} I(\widehat{P}_e)$
- Chow-Liu tree in terms of distance estimates

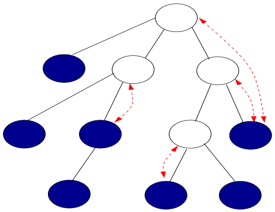
$$\widehat{T}_{\mathrm{CL}} = \mathrm{MST}(V; \widehat{\mathbf{d}}) := \operatorname*{argmin}_{T=(V,E)\in\mathcal{T}} \sum_{e\in E} \hat{d}_e.$$

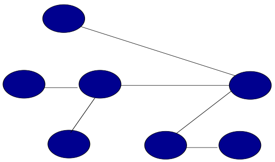


Relating Chow-Liu Tree with Latent Tree

Surrogate $\mathrm{Sg}(i)$ for node i: visible node with strongest correlation

$$\mathrm{Sg}(i; T_p, V) := \operatorname*{argmin}_{j \in V} d_{i,j}$$





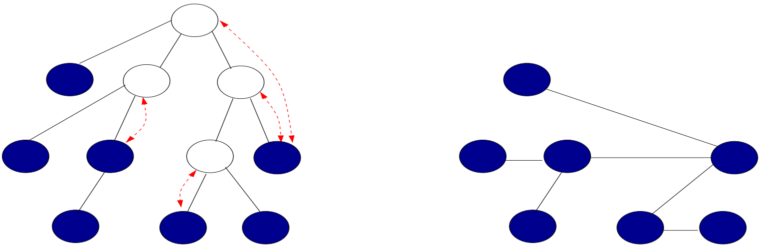
Properties of Chow-Liu Tree and Surrogacy

Neighborhood Preservation: for $i, j \in W$ with $\mathrm{Sg}(i) \neq \mathrm{Sg}(j)$,

$$(i, j) \in E_p \Rightarrow (\mathrm{Sg}(i), \mathrm{Sg}(j)) \in \mathrm{MST}(V; \mathbf{d}).$$

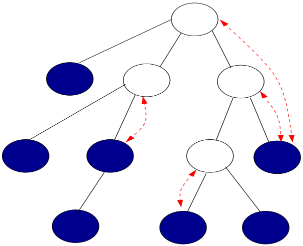
Chow-Liu Grouping for General Latent Trees

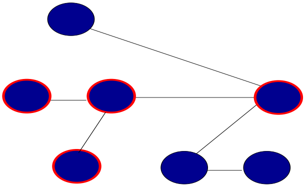
- Initialize tree estimate as Chow-Liu tree: $\widehat{T} \leftarrow \widehat{T}_{\mathrm{CL}}$
- Pick an internal node i in Chow-Liu tree $\widehat{T}_{\mathrm{CL}}$ not visited before, Recursive grouping over closed neighborhood $\mathrm{nbd}[i; \widehat{T}]$
- In $\widehat{T}$, replace subtree over $\mathrm{nbd}[i; \widehat{T}]$ with output of recursive grouping



Chow-Liu Grouping for General Latent Trees

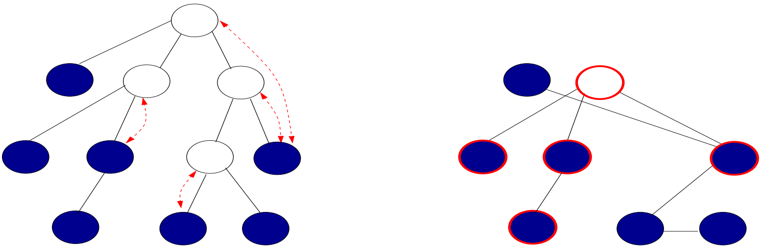
- Initialize tree estimate as Chow-Liu tree: $\widehat{T} \leftarrow \widehat{T}_{\mathrm{CL}}$
- Pick an internal node i in Chow-Liu tree $\widehat{T}_{\mathrm{CL}}$ not visited before, Recursive grouping over closed neighborhood $\mathrm{nbd}[i;\widehat{T}]$
- In $\widehat{T}$, replace subtree over $\mathrm{nbd}[i;\widehat{T}]$ with output of recursive grouping





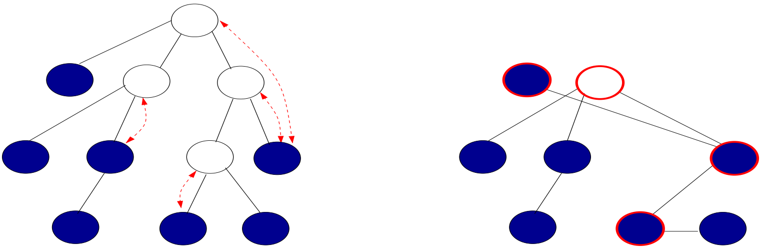
Chow-Liu Grouping for General Latent Trees

- Initialize tree estimate as Chow-Liu tree: $\widehat{T} \leftarrow \widehat{T}_{\mathrm{CL}}$
- Pick an internal node i in Chow-Liu tree $\widehat{T}_{\mathrm{CL}}$ not visited before, Recursive grouping over closed neighborhood $\mathrm{nbd}[i; \widehat{T}]$
- In $\widehat{T}$, replace subtree over $\mathrm{nbd}[i; \widehat{T}]$ with output of recursive grouping



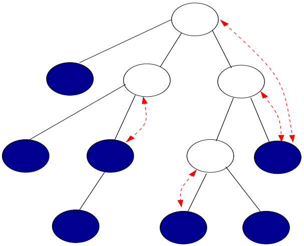
Chow-Liu Grouping for General Latent Trees

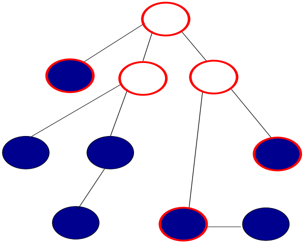
- Initialize tree estimate as Chow-Liu tree: $\widehat{T} \leftarrow \widehat{T}_{\mathrm{CL}}$
- Pick an internal node i in Chow-Liu tree $\widehat{T}_{\mathrm{CL}}$ not visited before, Recursive grouping over closed neighborhood $\mathrm{nbd}[i; \widehat{T}]$
- In $\widehat{T}$, replace subtree over $\mathrm{nbd}[i; \widehat{T}]$ with output of recursive grouping



Chow-Liu Grouping for General Latent Trees

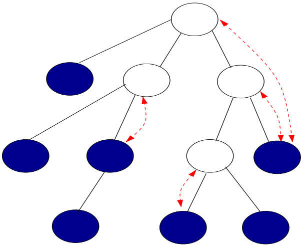
- Initialize tree estimate as Chow-Liu tree: $\widehat{T} \leftarrow \widehat{T}_{\mathrm{CL}}$
- Pick an internal node i in Chow-Liu tree $\widehat{T}_{\mathrm{CL}}$ not visited before, Recursive grouping over closed neighborhood $\mathrm{nbd}[i;\widehat{T}]$
- In $\widehat{T}$, replace subtree over $\mathrm{nbd}[i;\widehat{T}]$ with output of recursive grouping

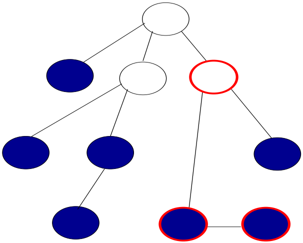




Chow-Liu Grouping for General Latent Trees

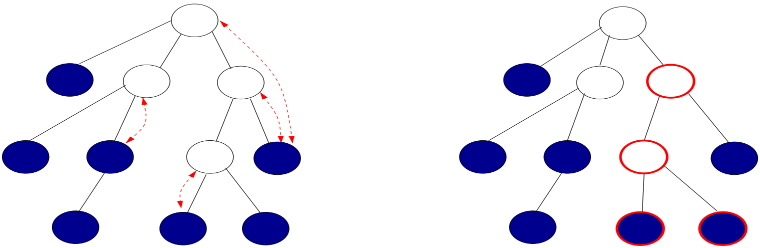
- Initialize tree estimate as Chow-Liu tree: $\widehat{T} \leftarrow \widehat{T}_{\mathrm{CL}}$
- Pick an internal node i in Chow-Liu tree $\widehat{T}_{\mathrm{CL}}$ not visited before, Recursive grouping over closed neighborhood $\mathrm{nbd}[i; \widehat{T}]$
- In $\widehat{T}$, replace subtree over $\mathrm{nbd}[i; \widehat{T}]$ with output of recursive grouping





Chow-Liu Grouping for General Latent Trees

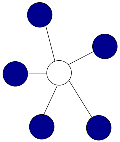
- Initialize tree estimate as Chow-Liu tree: $\widehat{T} \leftarrow \widehat{T}_{\mathrm{CL}}$
- Pick an internal node i in Chow-Liu tree $\widehat{T}_{\mathrm{CL}}$ not visited before, Recursive grouping over closed neighborhood $\mathrm{nbd}[i;\widehat{T}]$
- In $\widehat{T}$, replace subtree over $\mathrm{nbd}[i;\widehat{T}]$ with output of recursive grouping



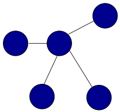
Guarantees for Chow-Liu Grouping

- Structural and estimation consistency for all minimal latent trees
- Sample complexity of $\Omega(\log m)$ for m observed nodes when effective depth is constant
- Computational complexity of $O(m^2 \log m + \text{(No. of internal nodes in CL-tree)} \times \text{(Max. Deg)}^3$.

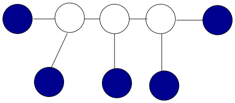
Star: Latent Tree



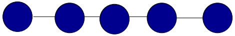
Chow-Liu Tree



HMM: Latent Tree



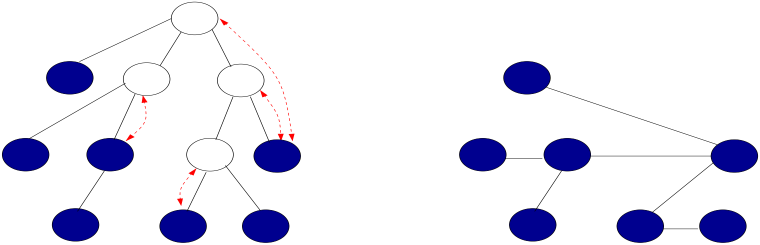
Chow-Liu Tree



Regularized Chow-Liu Grouping

- Chow-Liu pre-processing step provides a natural means to tradeoff accurate model fitting with model complexity
- Can stop at any stage: tree with fewer no. of hidden variables
- Relevant for real data: stopping rule through Bayesian information criterion (BIC) score

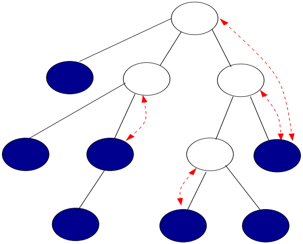
$$\text{BIC}(\hat{T}) := \log P(\mathbf{x}^n; \hat{T}) - C|H(\hat{T})| \log n.$$

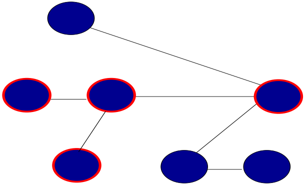


Regularized Chow-Liu Grouping

- Chow-Liu pre-processing step provides a natural means to tradeoff accurate model fitting with model complexity
- Can stop at any stage: tree with fewer no. of hidden variables
- Relevant for real data: stopping rule through Bayesian information criterion (BIC) score

$$\text{BIC}(\hat{T}) := \log P(\mathbf{x}^n; \hat{T}) - C|H(\hat{T})| \log n.$$

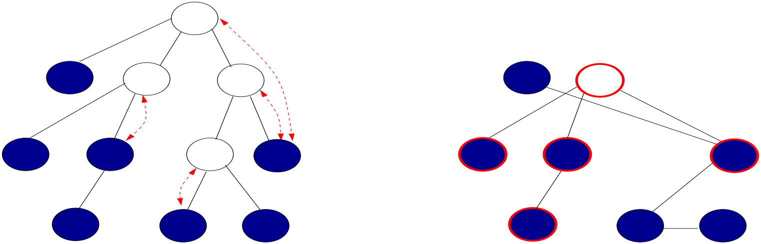




Regularized Chow-Liu Grouping

- Chow-Liu pre-processing step provides a natural means to tradeoff accurate model fitting with model complexity
- Can stop at any stage: tree with fewer no. of hidden variables
- Relevant for real data: stopping rule through Bayesian information criterion (BIC) score

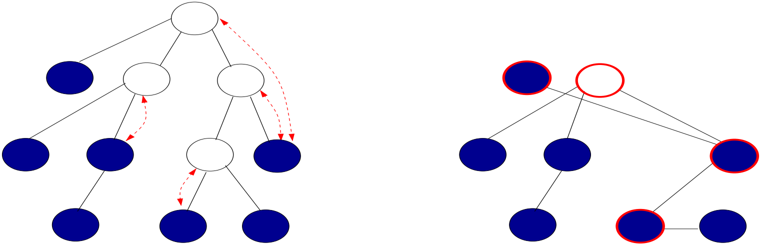
$$\text{BIC}(\hat{T}) := \log P(\mathbf{x}^n; \hat{T}) - C|H(\hat{T})| \log n.$$



Regularized Chow-Liu Grouping

- Chow-Liu pre-processing step provides a natural means to tradeoff accurate model fitting with model complexity
- Can stop at any stage: tree with fewer no. of hidden variables
- Relevant for real data: stopping rule through Bayesian information criterion (BIC) score

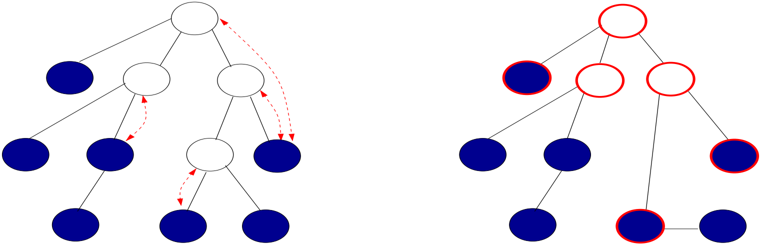
$$\text{BIC}(\hat{T}) := \log P(\mathbf{x}^n; \hat{T}) - C|H(\hat{T})| \log n.$$



Regularized Chow-Liu Grouping

- Chow-Liu pre-processing step provides a natural means to tradeoff accurate model fitting with model complexity
- Can stop at any stage: tree with fewer no. of hidden variables
- Relevant for real data: stopping rule through Bayesian information criterion (BIC) score

$$\text{BIC}(\hat{T}) := \log P(\mathbf{x}^n; \hat{T}) - C|H(\hat{T})| \log n.$$



Outline

Results on Data sets

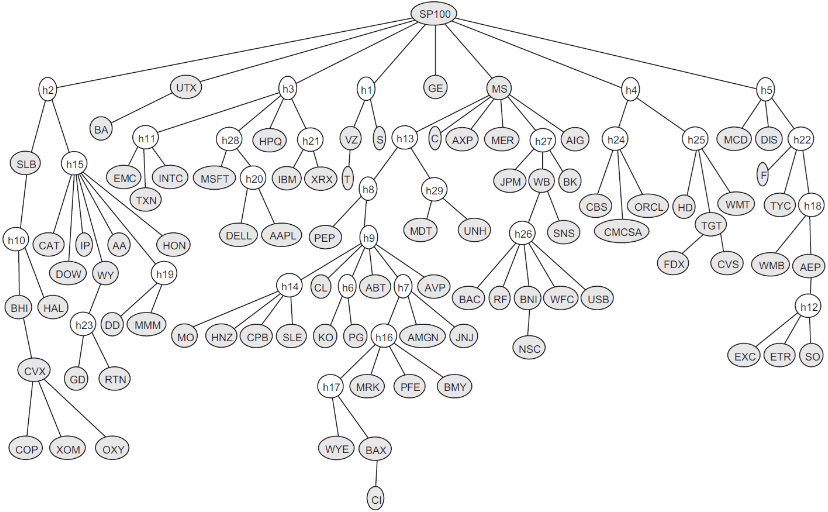
S & P 100 Stock Data

- Monthly returns of 84 companies in S&P 100.
- Samples from 1990 to 2007.
- Latent tree learned using CLNJ.

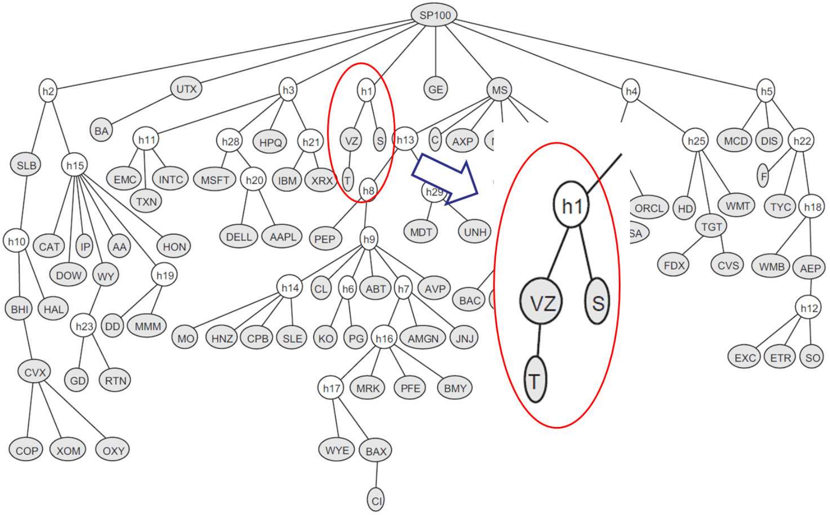
20 Newsgroups with 100 words

- 16,242 binary samples of 100 words
- Latent tree learned using regCLRG.

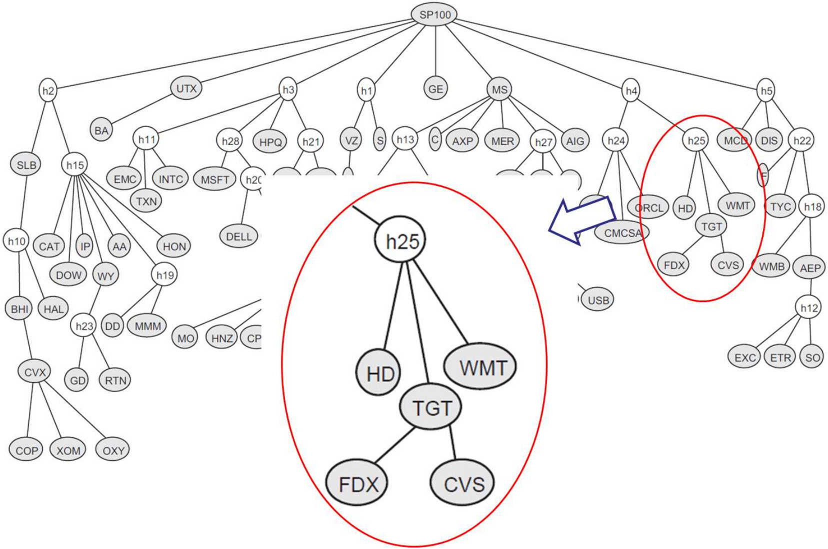
S & P Monthly Returns



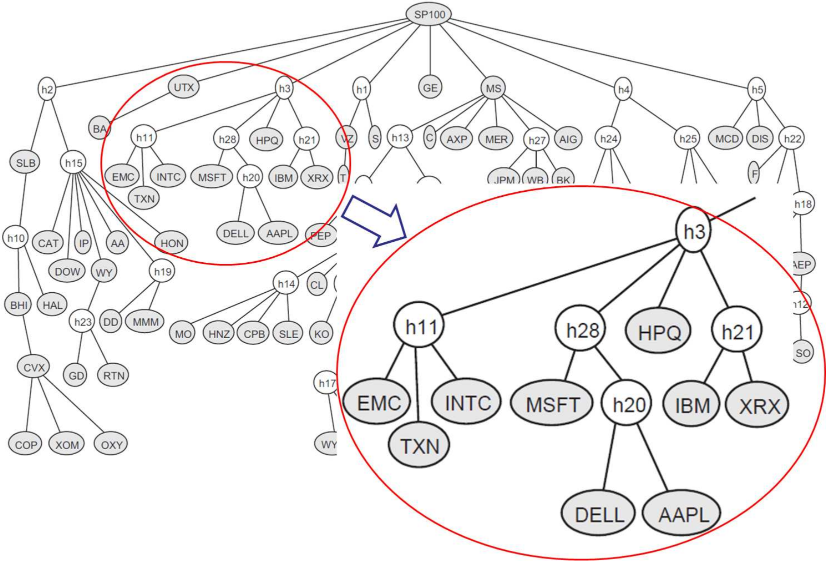
S & P Monthly Returns



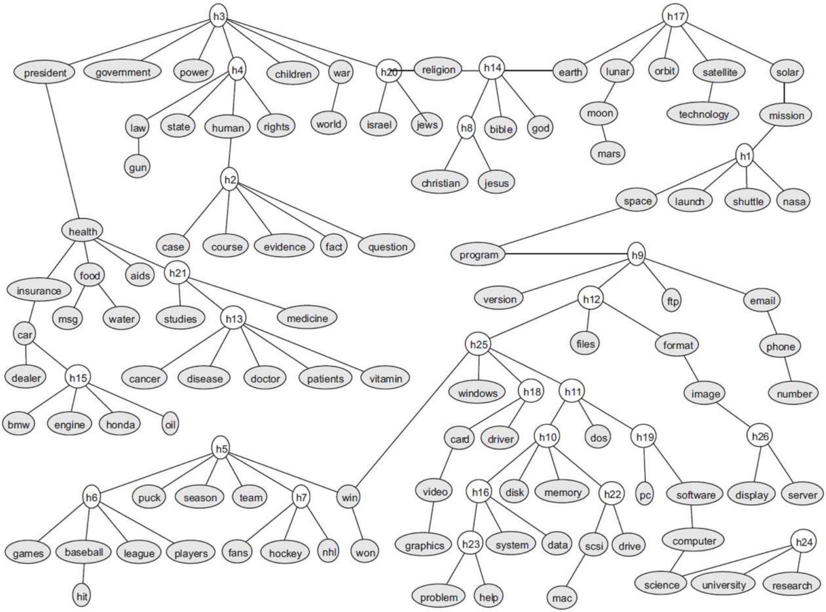
S & P Monthly Returns



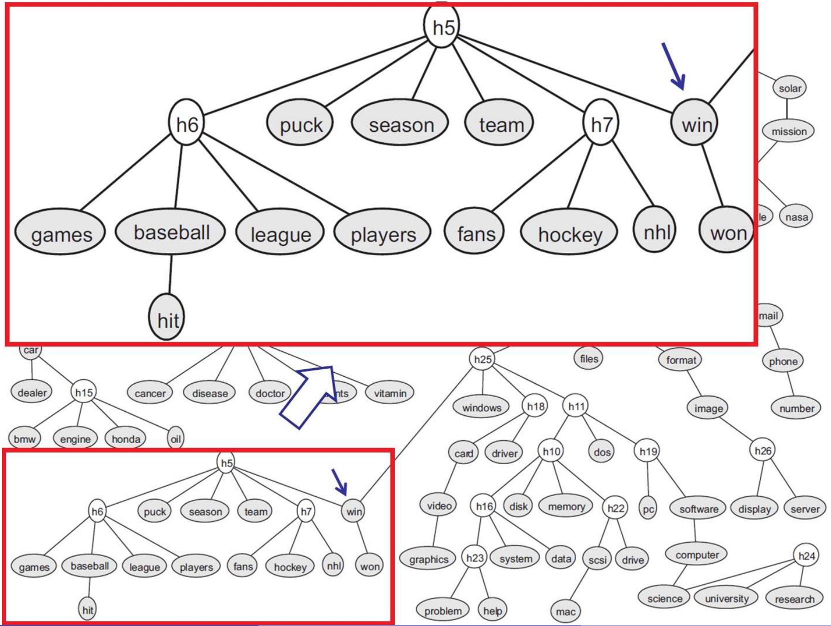
S & P Monthly Returns



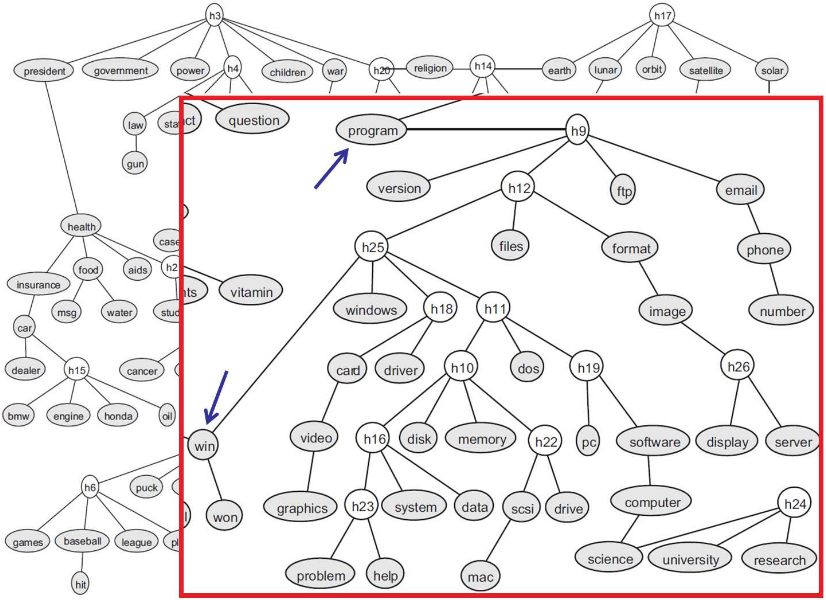
Newsgroup Data



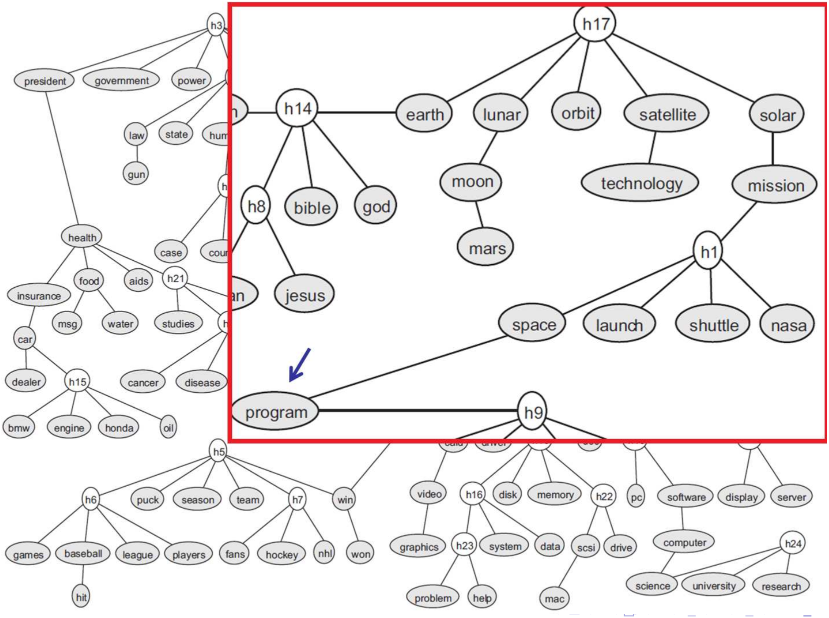
Newsgroup Data



Newsgroup Data

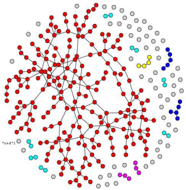


Newsgroup Data



Outline

Setup: Ising Models on Random Graphs



- n samples available at nodes to estimate structure
- Erdős-Rényi random graphs $G_m \sim \mathcal{G}(m, c/m)$: each edge has probability c/m
- Ising Models (Binary Pairwise Model)

$$P(\mathbf{x}) = \frac{1}{Z} \exp[\sum_{(i,j) \in G} J_{i,j} x_i x_j]$$

- For $(i,j) \in G_n$, $0 < J_{\min} \leq J_{i,j} \leq J_{\max} < \infty$

Two Algorithms for Structure Learning

Correlation Thresholding (CT)

- Empirical Correlations from Samples: $\widehat{C}^n(i,j) := \frac{1}{n}\sum_{k=1}^{n} x_i(k)x_j(k)$
- $(i,j) \in \widehat{G}$ if $\widehat{C}^n(i,j) > \delta(J_{\min}, J_{\max})$.

Two Algorithms for Structure Learning

Correlation Thresholding (CT)

- Empirical Correlations from Samples: $\widehat{C}^n(i,j) := \frac{1}{n}\sum_{k=1}^{n} x_i(k)x_j(k)$
- $(i,j) \in \widehat{G}$ if $\widehat{C}^n(i,j) > \delta(J_{\min}, J_{\max})$.

Conditional Mutual Information Thresholding (CMIT)

- Empirical Mutual Information from samples

$$\widehat{I}^n(X;Y) := \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \widehat{P}^n(x,y) \log \frac{\widehat{P}^n(x,y)}{\widehat{P}^n(x)\widehat{P}^n(y)},$$

  where $\widehat{P}^n$ is the type or the empirical distribution
- $(i,j) \in \widehat{G}$ if $\min\limits_{\substack{S\subset V\setminus\{i,j\} \\ |S|\leq 3}} \widehat{I}(X_i;X_j|\mathbf{X}_S) > \xi_{n,m}$
- Threshold $\xi_{n,m}$: depends on no. of samples n and no. of nodes m: parameter free

Results on Conditional Mutual Information Thresholding

- Ising model on the random graph $G_m = (V_m, E_m) \sim \mathcal{G}(m, \frac{c}{m})$
- No. of samples $n > M g_m \log m$, with $\lim_{m\to\infty} g_m = \infty$.
- Correlation decay: $c \tanh J_{\max} < 1$.

Structural Consistency of CMIT

CMIT is consistent for a.e. graph G_m

$$\lim_{\substack{m,n\to\infty \\ n > M g_m \log m}} \mathbb{P}\left[\text{CMIT}\left(\{\mathbf{x}^n\}; \xi_{n,m}\right) \neq G_m\right] = 0.$$

Results on Correlation Thresholding

- Number of samples $n > M \log m$
- Correlation decay: $c \tanh J_{\max} < 1$.

Edit Distance Guarantee for CT

Finite edit distance for a.e. graph

$$\lim_{\substack{m,n\to\infty \\ n>M\log m}} \mathbb{P}\left[\left|\mathsf{CT}(\{\widehat{C}^n_{i,j}\};\delta)\triangle G_m\right| > \omega(1)\right] = 0,$$

- Assume homogeneity: $2\tanh^2 J_{\max} < \tanh J_{\min}$

Structural Consistency for CT

CT is consistent for a.e. G_m

$$\lim_{\substack{m,n\to\infty \\ n>M\log m}} \mathbb{P}\left[\mathsf{CT}(\{\widehat{C}^n_{i,j}\};\delta) \neq G_m\right] = 0$$

Lower Bound on Sample Complexity

- Proposed algorithms with performance guarantees and sample complexities
- Converse result: lower bound on sample complexity below which any algorithm fails
- $G_m \sim \mathcal{G}(m, c/m)$ for any $c \leq 0.5m$: not required to be sparse

Converse Result

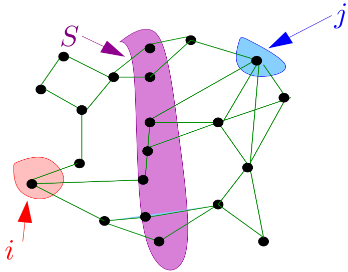
If $n \leq \epsilon c \log m$ for sufficiently small $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\widehat{G}_m \neq G_m) = 1.$$

$\Omega(c \log m)$ samples needed for random graph structure estimation.

Proof Ideas: Conditional Mutual Information

Separators in Graphical Models



$$X_i \perp\!\!\!\perp X_j | \mathbf{X}_S \iff I(X_i; X_j | \mathbf{X}_S) = 0$$

Challenges

- Structure learning through conditional mutual information testing
- Large separator sets in general graphs

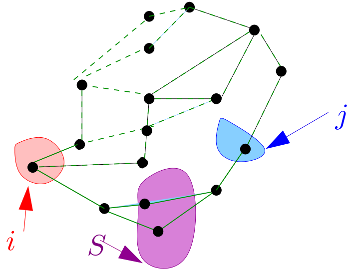
Proof Ideas Contd.

Approximate Separator Sets

Subset of separator on short paths.

Bound on Approx. Separator Set

- In random graphs, size of separator is at most two asymptotically
- Short cycles do not overlap in random graphs



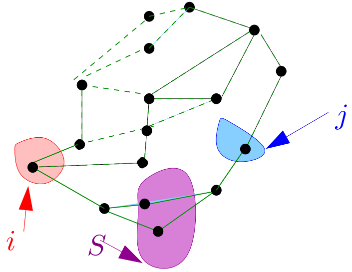
Proof Ideas Contd.

Approximate Separator Sets

Subset of separator on short paths.

Bound on Approx. Separator Set

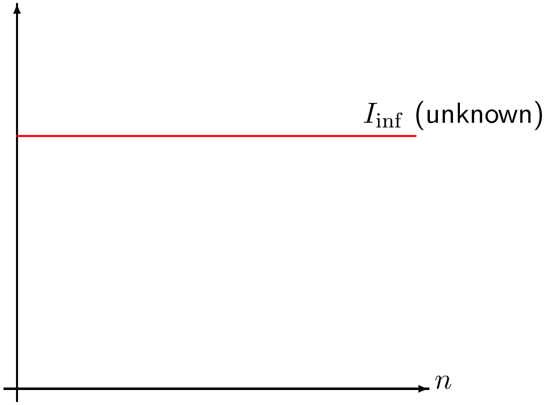
- In random graphs, size of separator is at most two asymptotically
- Short cycles do not overlap in random graphs



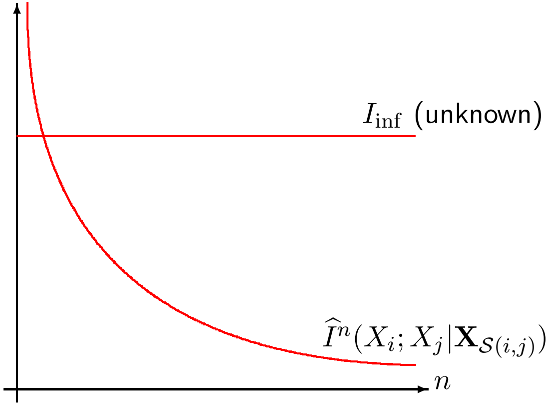
Decay of Conditional Mutual Information

- Under correlation decay, short paths contain most of the information
- $I(X_i; X_j | \mathbf{X}_S)$ decays as the graph size grows

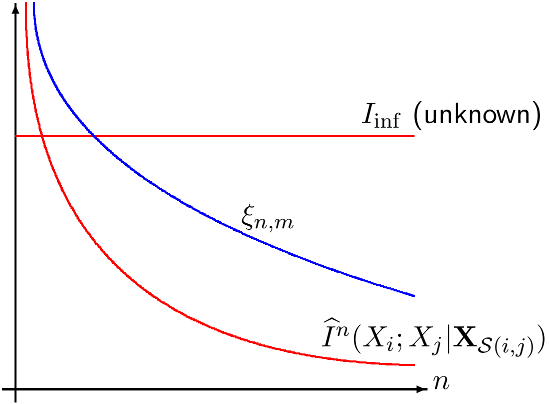
Proof Ideas: Choice of Threshold $\xi_{n,m}$



Proof Ideas: Choice of Threshold $\xi_{n,m}$



Proof Ideas: Choice of Threshold $\xi_{n,m}$

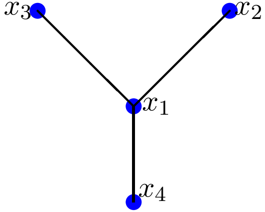


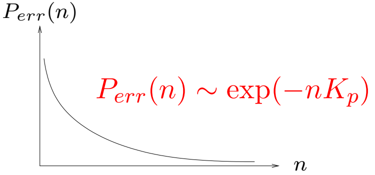
- Threshold $\xi_{n,m}$ depends both on the graph size m and number of samples n
- Asymptotically, $\xi_{n,m}$ distinguishes edges and non-edges.

Outline

Outline

Result 1: Error Exponent Tree Structure Learning





Error Exponent

Rate of exponential decay of prob. that estimated tree $\neq$ true tree.

Results for discrete tree models

- Error exponent as optimization of error rates for local events
- In very-noisy regime, error exponent $\approx$ SNR for learning.

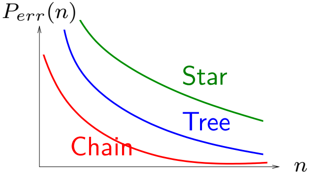
V. Tan, A. Anandkumar, L. Tong, A. Willsky "A Large-Deviation Analysis of the Maximum-Likelihood Learning of Markov Tree Structures," *submitted to IEEE Tran. on Information Theory*, on Arxiv.

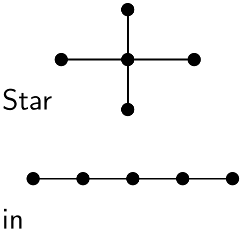
Result 1: Error Exponent for Tree Learning Contd.,

Extremal Tree Structures for Learning

For Gaussian distribution in very noisy learning regime

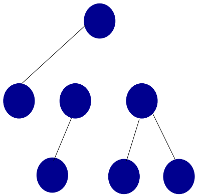
- Star graphs are hardest to learn, Markov chains are easiest to learn.
- Error exponent increases with tree diameter.
- Keeping the correlations on edges fixed.





V. Tan, A. Anandkumar, A. Willsky "Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures," *IEEE Tran. on Signal Proc.*, Vol. 58, No. 5, May 2010, pp. 2701-2714.

Result 2: Learning High-Dimensional Forests



Setup

- High dimensional regime: both number of samples n and number of nodes m grow.
- Goal: learn forest distributions.

Intuitions

- Learn tree models and remove “weak” edges to prevent overfitting
- Challenge in edge thresholding: finite samples results in noisy edge strengths
- Regularized Threshold: as a function of number of samples n

Result 2: Learning High-Dimensional Forests Contd.,

- Propose CLThres, a thresholding algorithm: Chow-Liu Algorithm + Threshold

Result 2: Learning High-Dimensional Forests Contd.,

- Propose CLThres, a thresholding algorithm: Chow-Liu Algorithm + Threshold
- Prove error rates: exponential decay of underestimation and super-polynomial decay of overestimation errors for fixed-size models

Result 2: Learning High-Dimensional Forests Contd.,

- Propose CLThres, a thresholding algorithm: Chow-Liu Algorithm + Threshold
- Prove error rates: exponential decay of underestimation and super-polynomial decay of overestimation errors for fixed-size models
- Prove achievable scaling laws on (n, m, k) for consistent recovery in high-dimensions.

$$n > \max(C_1 \log^{1+\delta}(d-k), C_2 \log d), \qquad \forall \delta > 0$$

is achievable, where n: no. of samples, m: no. of nodes, k: no. of edges.

Result 2: Learning High-Dimensional Forests Contd.,

- Propose CLThres, a thresholding algorithm: Chow-Liu Algorithm + Threshold
- Prove error rates: exponential decay of underestimation and super-polynomial decay of overestimation errors for fixed-size models
- Prove achievable scaling laws on (n, m, k) for consistent recovery in high-dimensions.

$$n > \max(C_1 \log^{1+\delta}(d-k), C_2 \log d), \qquad \forall \delta > 0$$

is achievable, where n: no. of samples, m: no. of nodes, k: no. of edges.

Consistent estimation of forests is even when m grows polynomially in n

---

V. Tan, A. Anandkumar, A. Willsky "Learning High-Dimensional Markov Forest Distributions: Analysis of Error Rates", Submitted to *J. of Machine Learning Research*, available on Arxiv.

Outline

Conclusion

Learning Latent Tree Models

- Proposed two novel algorithms under unified approach for Gaussian and discrete latent tree models
- Consistency, computational and sample complexities

  Structural and estimation consistency for any minimal latent tree
  Sample complexity of $\Omega(\log m)$ for m observed nodes for fixed depth
  Low computational complexity

Learning Random Graphical Models

- Proposed two local algorithms
- Provided guarantees under correlation decay
- Efficient structure learning

---

http://newport.eecs.uci.edu/anandkumar