Probabilistic Graphical Models

Structure learning in Bayesian networks

Siamak Ravanbakhsh

Fall 2019

Learning objectives

- why structure learning is hard?
- two approaches to structure learning
  - constraint-based methods
  - score based methods
- MLE vs Bayesian score

Structure learning in BayesNets

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets

Structure learning in BayesNets

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets
- search over the combinatorial space, maximizing a **score**

$2^{\mathcal{O}(n^2)}$

Structure learning in BayesNets

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets
- search over the combinatorial space, maximizing a **score**

  $2^{\mathcal{O}(n^2)}$
- Bayesian model averaging
  - integrate over all possible structures

Structure learning in BayesNets

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets
- search over the combinatorial space, maximizing a **score**
- Bayesian model averaging
  - integrate over all possible structures

Structure learning in BayesNets

Identifiable up to I-equivalence

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets

a DAG with the same set of conditional independencies (CI) $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p_{\mathcal{D}})$

Structure learning in BayesNets

Identifiable up to I-equivalence

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets

Perfect MAP

a DAG with the same set of conditional independencies (CI) $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p_{\mathcal{D}})$

Structure learning in BayesNets

Identifiable up to I-equivalence

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets

Perfect MAP

a DAG with the same set of conditional independencies (CI) $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p_{\mathcal{D}})$

hypothesis testing

Structure learning in BayesNets

Identifiable up to I-equivalence

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets

a DAG with the same set of conditional independencies (CI)

Perfect MAP

$\mathcal{I}(\mathcal{G}) = \mathcal{I}(p_{\mathcal{D}})$

↓

hypothesis testing

$X \perp Y \mid \mathbf{Z}?$

Structure learning in BayesNets

Identifiable up to I-equivalence

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets

Perfect MAP

a DAG with the same set of conditional independencies (CI) $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p_{\mathcal{D}})$

first attempt: a DAG that is **I-map** for $p_{\mathcal{D}}$ $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(p_{\mathcal{D}})$

hypothesis testing

$X \perp Y \mid \mathbf{Z}?$

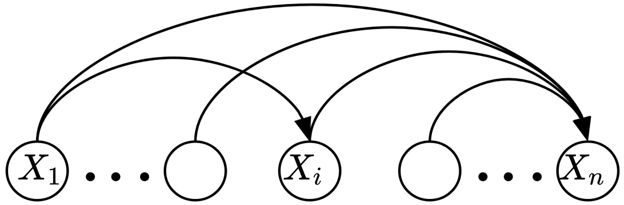
minimal I-map from CI test

a DAG where removing an edge violates I-map property

input: IC test oracle; an ordering $X_1, \ldots, X_n$

output: a minimal I-map G

for i=1...n



- find minimal $\mathbf{U} \subseteq \{X_1, \ldots, X_{i-1}\}$ s.t. $(X_i \perp X_1, \ldots, X_{i-1} - \mathbf{U} \mid \mathbf{U})$
- set $Pa_{X_i} \leftarrow \mathbf{U}$

$X_i \perp NonDesc_{X_i} \mid Pa_{X_i}$

minimal I-map from CI test

☹ **Problems:**

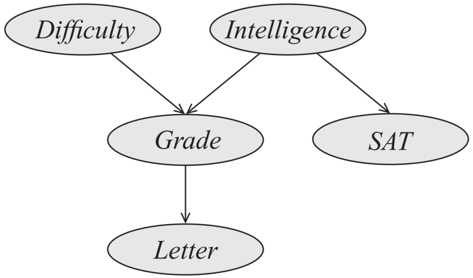
- CI tests involve many variables
- number of CI tests is exponential
- a minimal I-MAP may be far from a P-MAP

minimal I-map from CI test

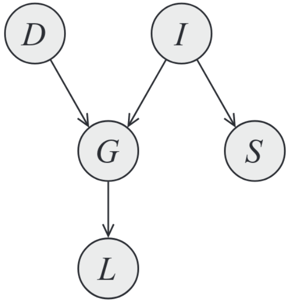
Problems:

- CI tests involve many variables
- number of CI tests is exponential
- a minimal I-MAP may be far from a P-MAP

Example:

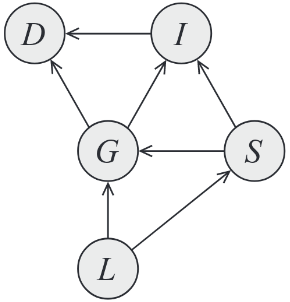


different orderings give different graphs

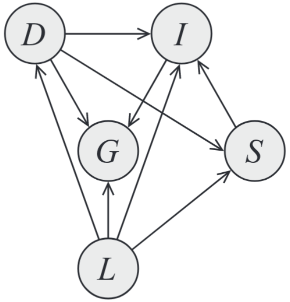


D,I,S,G,L

(a topological ordering)



L,S,G,I,D



L,D,S,I,G

Structure learning in BayesNets

Identifiable up to I-equivalence

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets

a DAG with the same set of conditional independencies (CI)

first attempt: a DAG that is **I-map** for $p_{\mathcal{D}}$ $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(p_{\mathcal{D}})$

second attempt: a DAG that is **P-map** for $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p_{\mathcal{D}})$

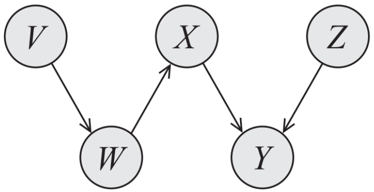
can we find a perfect MAP with fewer IC tests
involving fewer variables?

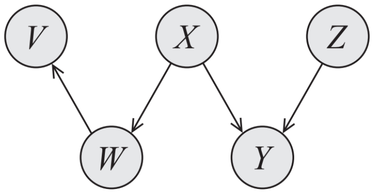
Perfect map from CI test

only up to **I-equivalence**

the same set of CIs

- same skeleton
- same immoralities



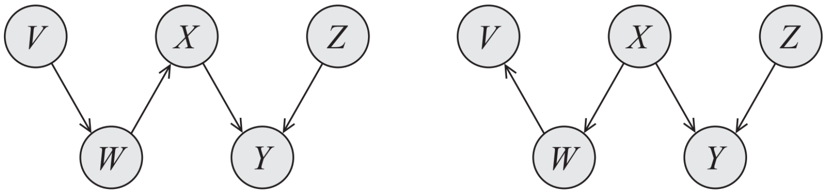


Perfect map from CI test

only up to **I-equivalence**

the same set of CIs

- same skeleton
- same immoralities



procedure:

1. find the undirected skeleton using CI tests
2. identify immoralities in the undirected graph

Perfect map from CI test

1. finding the undirected skeleton

observation: if X and Y are not adjacent then $X \perp Y \mid Pa_X$ OR $X \perp Y \mid Pa_Y$

Perfect map from CI test

1. finding the undirected skeleton

observation: if X and Y are not adjacent then $X \perp Y \mid Pa_X$ OR $X \perp Y \mid Pa_Y$

assumption: max number of parents d

Perfect map from CI test

1. finding the undirected skeleton

observation: if X and Y are not adjacent then $X \perp Y \mid Pa_X$ OR $X \perp Y \mid Pa_Y$

assumption: max number of parents d

idea: search over all subsets of size d, and check CI above

Perfect map from CI test

1. finding the undirected skeleton

observation: if X and Y are not adjacent then $X \perp Y \mid Pa_X$ OR $X \perp Y \mid Pa_Y$

assumption: max number of parents d

idea: search over all subsets of size d, and check CI above

input: CI oracle; bound on #parents d

output: undirected skeleton

initialize **H** as a complete *undirected* graph

for all pairs X_i, X_j

for all subsets **U** of size $\leq d$ (within current neighbors of X_i, X_j)

If $X_i \perp X_j \mid \mathbf{U}$ then remove $X_i - X_j$ from **H**

return **H**

Perfect map from CI test

1. finding the undirected skeleton

observation: if X and Y are not adjacent then $X \perp Y \mid Pa_X$ OR $X \perp Y \mid Pa_Y$

assumption: max number of parents d

idea: search over all subsets of size d, and check CI above

input: CI oracle; bound on #parents d

output: undirected skeleton

initialize **H** as a complete *undirected* graph

for all pairs X_i, X_j

for all subsets **U** of size $\leq d$ (within current neighbors of X_i, X_j)

If $X_i \perp X_j \mid \mathbf{U}$ then remove $X_i - X_j$ from **H**

return **H**

$\mathcal{O}(n^{d+2})$

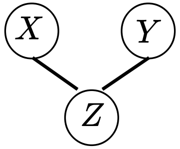
$= \mathcal{O}((n^2) \times \mathcal{O}((n-2)^d)$

Perfect map from CI test

2. finding the immoralities

potential immorality

$X - Z, Y - Z \in \mathcal{H}, X - Y \notin \mathcal{H}$

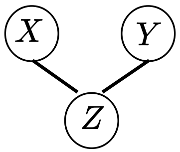


Perfect map from CI test

2. finding the immoralities

potential immorality

$X - Z, Y - Z \in \mathcal{H}, X - Y \notin \mathcal{H}$

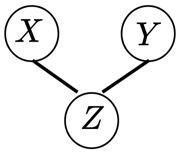


Perfect map from CI test

2. finding the immoralities

potential immorality

$X - Z, Y - Z \in \mathcal{H}, X - Y \notin \mathcal{H}$



not immorality only if

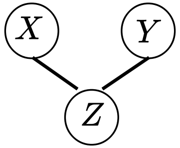
$X_i \perp X_j \mid \mathbf{U} \Rightarrow Z \in \mathbf{U}$

Perfect map from CI test

2. finding the immoralities

potential immorality

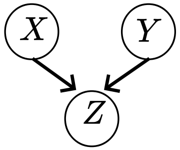
$X - Z, Y - Z \in \mathcal{H}, X - Y \notin \mathcal{H}$



not immorality only if

$X_i \perp X_j \mid \mathbf{U} \Rightarrow Z \in \mathbf{U}$

- save the **U** when removing X-Y
- see if Z in **U?**
  - if no, then we have immorality



input: CI oracle; bound on #parents d

output: undirected skeleton

initialize **H** as a complete *undirected* graph

for all pairs X_i, X_j

for all subsets **U** of size $\leq d$ (within current neighbors of X_i, X_j)

If $X_i \perp X_j \mid \mathbf{U}$ then remove $X_i - X_j$ from **H**

return **H**

Perfect map from CI test

3. propagate the constraints

at this point: a mix of directed and undirected edges

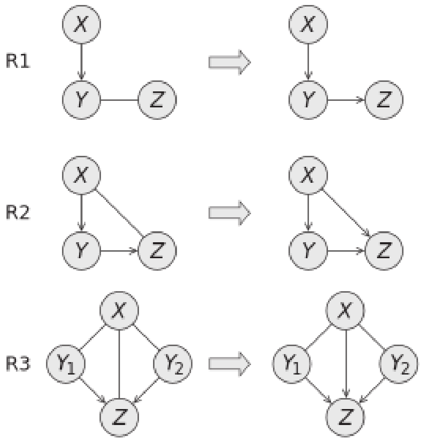
Perfect map from CI test

3. propagate the constraints

at this point: a mix of directed and undirected edges

add directions using the following rules (needed to preserve immoralities / DAG structure)

until convergence



for exact CI tests, this guarantees the exact I-equivalence family

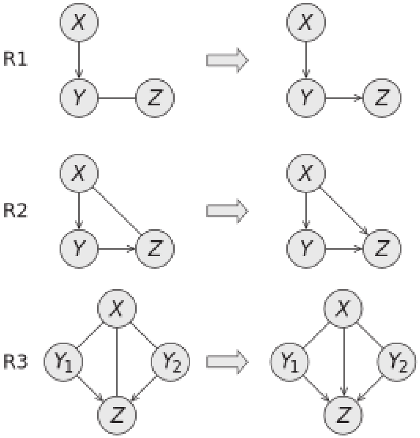
Perfect map from CI test

3. propagate the constraints

at this point: a mix of directed and undirected edges

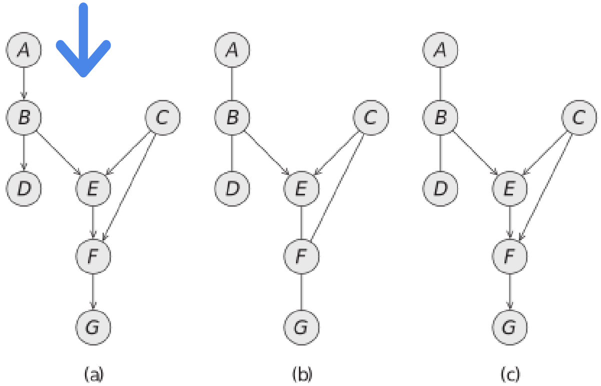
add directions using the following rules (needed to preserve immoralities / DAG structure)

until convergence



Example

Ground truth DAG



for exact CI tests, this guarantees the exact I-equivalence family

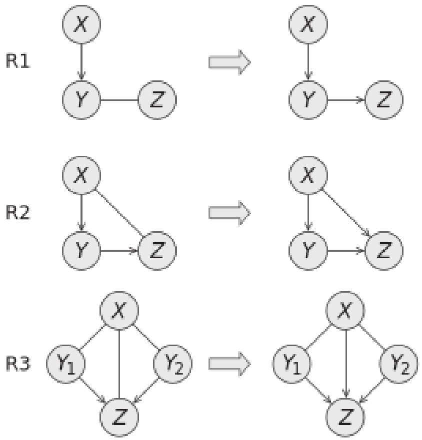
Perfect map from CI test

3. propagate the constraints

at this point: a mix of directed and undirected edges

add directions using the following rules (needed to preserve immoralities / DAG structure)

until convergence

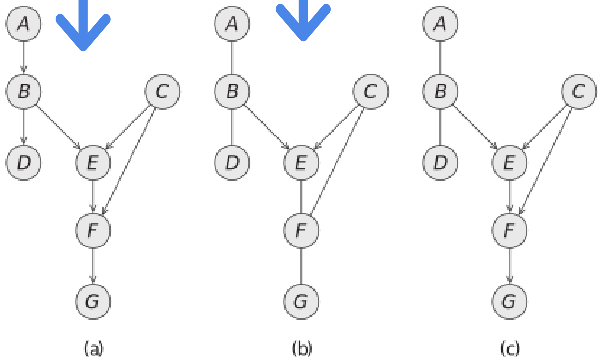


Example

Ground truth DAG

undirected skeleton

+immoralities



for exact CI tests, this guarantees the exact I-equivalence family

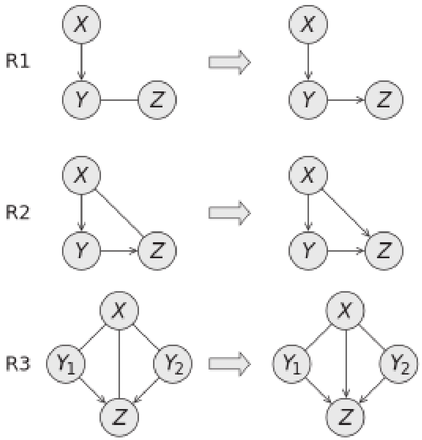
Perfect map from CI test

3. propagate the constraints

at this point: a mix of directed and undirected edges

add directions using the following rules (needed to preserve immoralities / DAG structure)

until convergence



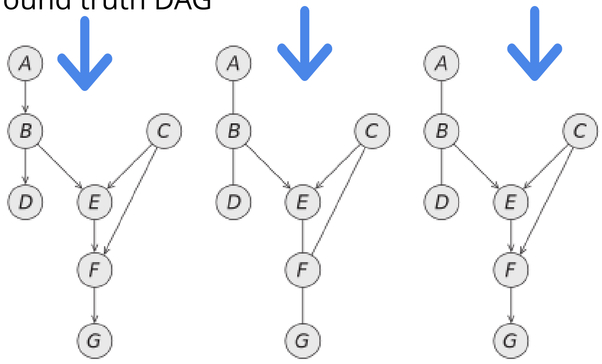
Example

Ground truth DAG

undirected skeleton

+immoralities

using rules R1,R2,R3



(a) (b) (c)

for exact CI tests, this guarantees the exact I-equivalence family

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

measure the deviance of $p_{\mathcal{D}}(X \mid Z) p_{\mathcal{D}}(Y|Z)$ from $p_{\mathcal{D}}(X, Y|Z)$

- conditional mututal information

$$d_I(\mathcal{D}) = \mathbb{E}_Z[\mathbf{D}(p_{\mathcal{D}}(X, Y|Z) || p_{\mathcal{D}}(X|Z) p_{\mathcal{D}}(Y|Z))]$$

- χ^2 statistics

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

measure the deviance of $p_{\mathcal{D}}(X \mid Z) p_{\mathcal{D}}(Y|Z)$ from $p_{\mathcal{D}}(X, Y|Z)$

- conditional mututal information

$$d_I(\mathcal{D}) = \mathbb{E}_Z[\mathbf{D}(p_{\mathcal{D}}(X, Y|Z) || p_{\mathcal{D}}(X|Z) p_{\mathcal{D}}(Y|Z))]$$

- χ^2 statistics

$$d_{\chi^2}(\mathcal{D}) = |\mathcal{D}| \sum_{x,y,z} \frac{(p_{\mathcal{D}}(x,y,z) - p_{\mathcal{D}}(z) p_{\mathcal{D}}(x|z) p_{\mathcal{D}}(y|z))^2}{p_{\mathcal{D}}(z) p_{\mathcal{D}}(x|z) p_{\mathcal{D}}(y|z)}$$

using frequencies in the dataset

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

measure the deviance of $p_{\mathcal{D}}(X \mid Z) p_{\mathcal{D}}(Y|Z)$ from $p_{\mathcal{D}}(X, Y|Z)$

- conditional mututal information

$$d_I(\mathcal{D}) = \mathbb{E}_Z[\mathbf{D}(p_{\mathcal{D}}(X, Y|Z) || p_{\mathcal{D}}(X|Z) p_{\mathcal{D}}(Y|Z))]$$

- χ^2 statistics

$$d_{\chi^2}(\mathcal{D}) = |\mathcal{D}| \sum_{x,y,z} \frac{(p_{\mathcal{D}}(x,y,z) - p_{\mathcal{D}}(z) p_{\mathcal{D}}(x|z) p_{\mathcal{D}}(y|z))^2}{p_{\mathcal{D}}(z) p_{\mathcal{D}}(x|z) p_{\mathcal{D}}(y|z)}$$

using frequencies in the dataset

large deviance rejects the null hypothesis (of conditional independence)

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

measure the deviance of $p_{\mathcal{D}}(X \mid Z)p_{\mathcal{D}}(Y|Z)$ from $p_{\mathcal{D}}(X, Y|Z)$

- conditional mututal information

$$d_I(\mathcal{D}) = \mathbb{E}_Z[\mathbf{D}(p_{\mathcal{D}}(X, Y|Z)||p_{\mathcal{D}}(X|Z)p_{\mathcal{D}}(Y|Z))]$$

- χ^2 statistics

$$d_{\chi^2}(\mathcal{D}) = |\mathcal{D}| \sum_{x,y,z} \frac{(p_{\mathcal{D}}(x,y,z) - p_{\mathcal{D}}(z)p_{\mathcal{D}}(x|z)p_{\mathcal{D}}(y|z))^2}{p_{\mathcal{D}}(z)p_{\mathcal{D}}(x|z)p_{\mathcal{D}}(y|z)}$$

using frequencies in the dataset

large deviance rejects the null hypothesis (of conditional independence)

↓

pick a threshold $d(\mathcal{D}) > t$

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

measure the deviance of $p_{\mathcal{D}}(X \mid Z) p_{\mathcal{D}}(Y|Z)$ from $p_{\mathcal{D}}(X, Y|Z)$

- conditional mututal information

$$d_I(\mathcal{D}) = \mathbb{E}_Z[\mathbf{D}(p_{\mathcal{D}}(X, Y|Z) || p_{\mathcal{D}}(X|Z) p_{\mathcal{D}}(Y|Z))]$$

- χ^2 statistics

$$d_{\chi^2}(\mathcal{D}) = |\mathcal{D}| \sum_{x,y,z} \frac{(p_{\mathcal{D}}(x,y,z) - p_{\mathcal{D}}(z) p_{\mathcal{D}}(x|z) p_{\mathcal{D}}(y|z))^2}{p_{\mathcal{D}}(z) p_{\mathcal{D}}(x|z) p_{\mathcal{D}}(y|z)}$$

using frequencies in the dataset

large deviance rejects the null hypothesis (of conditional independence)

↓

pick a threshold $d(\mathcal{D}) > t$

p-value is the probability of false rejection $\text{pvalue}(t) = P(\{\mathcal{D} : d(\mathcal{D}) > t\} \mid X \perp Y \mid Z)$

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

large deviance rejects the null hypothesis (of conditional independence)

↓

pick a threshold $d(\mathcal{D}) > t$

p-value is the probability of false rejection $\mathrm{pvalue}(t) = P(\{\mathcal{D} : d(\mathcal{D}) > t\} \mid X \perp Y \mid Z)$

↓

over all possible datasets ☹

conditional independence (CI) test

how to decide $X \perp Y \mid Z$ from the dataset $\mathcal{D}$

large deviance rejects the null hypothesis (of conditional independence)

↓

pick a threshold $d(\mathcal{D}) > t$

p-value is the probability of false rejection $\text{pvalue}(t) = P(\{\mathcal{D} : d(\mathcal{D}) > t\} \mid X \perp Y \mid Z)$

↓

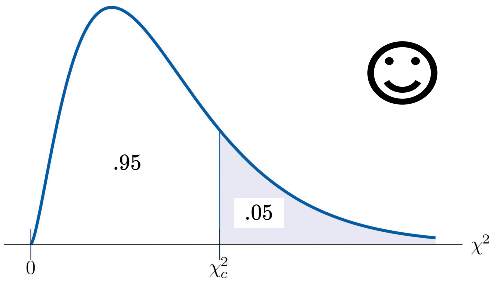
over all possible datasets ☹

it is possible to derive the distribution of deviance measures

- e.g., χ^2 distribution

reject a hypothesis (CI) for small p-values (.05)

☺



Structure learning in BayesNets

family of methods

- constraint-based methods
  - estimate cond. independencies from the data
  - find compatible BayesNets
- search over the combinatorial space, maximizing a **score**
- Bayesian model averaging
  - integrate over all possible structures

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y)$$

↓

conditional entropy $\sum_x p(x)H(p(y|x))$

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

symmetric $= I(Y,X)$

conditional entropy $\sum_x p(x)H(p(y|x))$

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

symmetric $= I(Y,X)$

conditional entropy $\sum_x p(x)H(p(y|x))$

$$I(X,Y) = \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$

Mutual information

how much information does X encode about Y?

reduction in the uncertainty of X after observing Y

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

symmetric $= I(Y,X)$

conditional entropy $\sum_x p(x)H(p(y|x))$

$$I(X,Y) = \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$

$$= D_{KL}(p(x,y)\|p(x)p(y))$$ positive

MLE in Bayes-nets mutual information form

log-likelihood $\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

MLE in Bayes-nets mutual information form

log-likelihood

$$\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

$$= \sum_i \sum_{(x_i, Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

MLE in Bayes-nets mutual information form

log-likelihood

$$\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i};\theta_{i|Pa_i})$$

$$= \sum_i \sum_{(x_i, Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i};\theta_{i|Pa_i})$$

using the empirical distribution

$$= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i};\theta_{i|Pa_i})$$

MLE in Bayes-nets mutual information form

log-likelihood $\ell(\mathcal{D};\theta) = \sum_{x\in\mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

$$= \sum_i \sum_{(x_i, Pa_{x_i})\in\mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$$

using the empirical distribution $= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

use MLE estimate $\ell(\mathcal{D}, \theta^*) = N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \log p_{\mathcal{D}}(x_i \mid Pa_{x_i})$

MLE in Bayes-nets mutual information form

log-likelihood $\ell(\mathcal{D};\theta) = \sum_{x \in \mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

$= \sum_i \sum_{(x_i, Pa_{x_i}) \in \mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

using the empirical distribution $= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

use MLE estimate $\ell(\mathcal{D}, \theta^*) = N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \log p_{\mathcal{D}}(x_i \mid Pa_{x_i})$

$= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \left(\log \frac{p_{\mathcal{D}}(x_i, Pa_{x_i})}{p_{\mathcal{D}}(x_i) p_{\mathcal{D}}(Pa_{x_i})} + \log p_{\mathcal{D}}(x_i) \right)$

MLE in Bayes-nets mutual information form

log-likelihood $\ell(\mathcal{D};\theta) = \sum_{x \in \mathcal{D}} \sum_i \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

$= \sum_i \sum_{(x_i, Pa_{x_i}) \in \mathcal{D}} \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

using the empirical distribution $= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x, Pa_{x_i}) \log p(x_i \mid Pa_{x_i}; \theta_{i|Pa_i})$

use MLE estimate $\ell(\mathcal{D}, \theta^*) = N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \log p_{\mathcal{D}}(x_i \mid Pa_{x_i})$

$= N \sum_i \sum_{x_i, Pa_{x_i}} p_{\mathcal{D}}(x_i, Pa_{x_i}) \left(\log \frac{p_{\mathcal{D}}(x_i, Pa_{x_i})}{p_{\mathcal{D}}(x_i) p_{\mathcal{D}}(Pa_{x_i})} + \log p_{\mathcal{D}}(x_i) \right)$

using the definition of mutual information $= N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

Optimal solution for trees

likelihood score $\ell(\mathcal{D}, \theta^*) = N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

Optimal solution for trees

likelihood score $\ell(\mathcal{D}, \theta^*) = N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

does not depend on structure

Optimal solution for trees

likelihood score $\ell(\mathcal{D}, \theta^*) = N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

does not depend on structure

$I_{\mathcal{D}}(X_i, X_j)$

Optimal solution for trees

likelihood score $\ell(\mathcal{D}, \theta^*) = N \sum_i I_{\mathcal{D}}(X_i, Pa_{X_i}) - H_{\mathcal{D}}(X_i)$

does not depend on structure

$I_{\mathcal{D}}(X_i, X_j)$

structure learning algorithms use mutual information in the structure search:

- **Chow-Liu algorithm**: find the max-spanning **tree:**
  - edge-weights = mutual information
  - add direction to edges later $I_{\mathcal{D}}(X_j, X_i) = I_{\mathcal{D}}(X_i, X_j)$
    - make sure each node has at most one parent (i.e., no v-structure)

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G})$$

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \xrightarrow{\log} \text{score}_B(\mathcal{G}, \mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}) + \log P(\mathcal{G})$$

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \xrightarrow{\log} \text{score}_B(\mathcal{G},\mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}) + \log P(\mathcal{G})$$

$$\int_{\theta \in \Theta_{\mathcal{G}}} P(\mathcal{D}|\theta,\mathcal{G})P(\theta \mid \mathcal{G})\mathrm{d}\theta$$

marginal likelihood for a structure $\mathcal{G}$

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \xrightarrow{\log} \text{score}_B(\mathcal{G},\mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}) + \log P(\mathcal{G})$$

$$\int_{\theta\in\Theta_{\mathcal{G}}} P(\mathcal{D}|\theta,\mathcal{G})P(\theta \mid \mathcal{G})\mathrm{d}\theta$$

marginal likelihood for a structure $\mathcal{G}$

assuming local and global parameter independence

factorizes to the marginal likelihood of each node

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G})$ $\xrightarrow{\log}$ $\text{score}_B(\mathcal{G},\mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}) + \log P(\mathcal{G})$

$\int_{\theta \in \Theta_{\mathcal{G}}} P(\mathcal{D}|\theta,\mathcal{G})P(\theta \mid \mathcal{G})\mathrm{d}\theta$ **marginal likelihood** for a structure $\mathcal{G}$

assuming local and global parameter independence

factorizes to the marginal likelihood of each node

for Dirichlet-multinomial has closed form

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \xrightarrow{\log} \text{score}_B(\mathcal{G},\mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}) + \log P(\mathcal{G})$$

$$\int_{\theta \in \Theta_{\mathcal{G}}} P(\mathcal{D}|\theta,\mathcal{G})P(\theta \mid \mathcal{G})\mathrm{d}\theta$$

marginal likelihood for a structure $\mathcal{G}$

assuming local and global parameter independence

factorizes to the marginal likelihood of each node

for Dirichlet-multinomial has closed form

for large sample size

any *exp-family* member

Bayesian Information Criterion (**BIC**) $\text{score}_B(\mathcal{G},\mathcal{D}) \approx \ell(\mathcal{D},\theta^*{}_{\mathcal{G}}) - \frac{1}{2}\log(|\mathcal{D}|)K$

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \xrightarrow{\log} \text{score}_B(\mathcal{G},\mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}) + \log P(\mathcal{G})$$

$$\int_{\theta \in \Theta_{\mathcal{G}}} P(\mathcal{D}|\theta,\mathcal{G})P(\theta \mid \mathcal{G})\mathrm{d}\theta$$

marginal likelihood for a structure $\mathcal{G}$

assuming local and global parameter independence

factorizes to the marginal likelihood of each node

for Dirichlet-multinomial has closed form

for large sample size

any *exp-family* member

#parameters

Bayesian Information Criterion (**BIC**) $\text{score}_B(\mathcal{G},\mathcal{D}) \approx \ell(\mathcal{D},\theta^*{}_{\mathcal{G}}) - \frac{1}{2}\log(|\mathcal{D}|)K$

Bayesian Score for BayesNets

Bayesian about both structure $\mathcal{G}$ and parameters θ

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \xrightarrow{\log} \text{score}_B(\mathcal{G},\mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}) + \log P(\mathcal{G})$$

$$\int_{\theta \in \Theta_{\mathcal{G}}} P(\mathcal{D}|\theta,\mathcal{G})P(\theta \mid \mathcal{G})\mathrm{d}\theta$$

marginal likelihood for a structure $\mathcal{G}$

assuming local and global parameter independence

factorizes to the marginal likelihood of each node

for Dirichlet-multinomial has closed form

for large sample size

any *exp-family* member

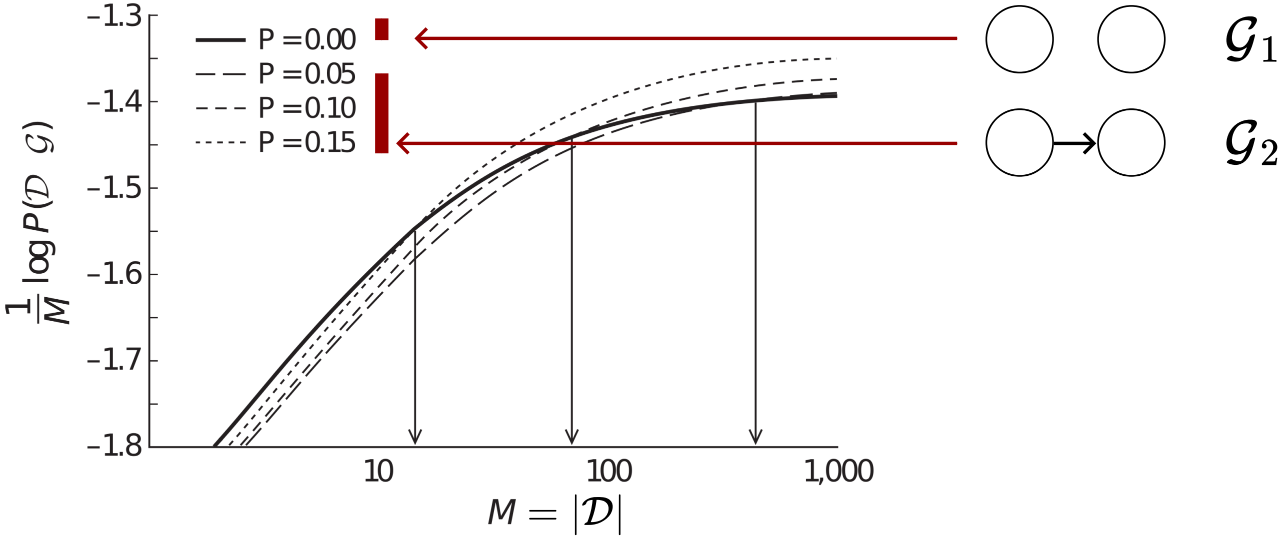
#parameters

Bayesian Information Criterion (**BIC**) $\text{score}_B(\mathcal{G},\mathcal{D}) \approx \ell(\mathcal{D},\theta^*{}_{\mathcal{G}}) - \frac{1}{2}\log(|\mathcal{D}|)K$

Akaike Information Criterion (**AIC**) $\ell(\mathcal{D},\theta^*{}_{\mathcal{G}}) - \frac{1}{2}K$

Bayesian Score for BayesNets

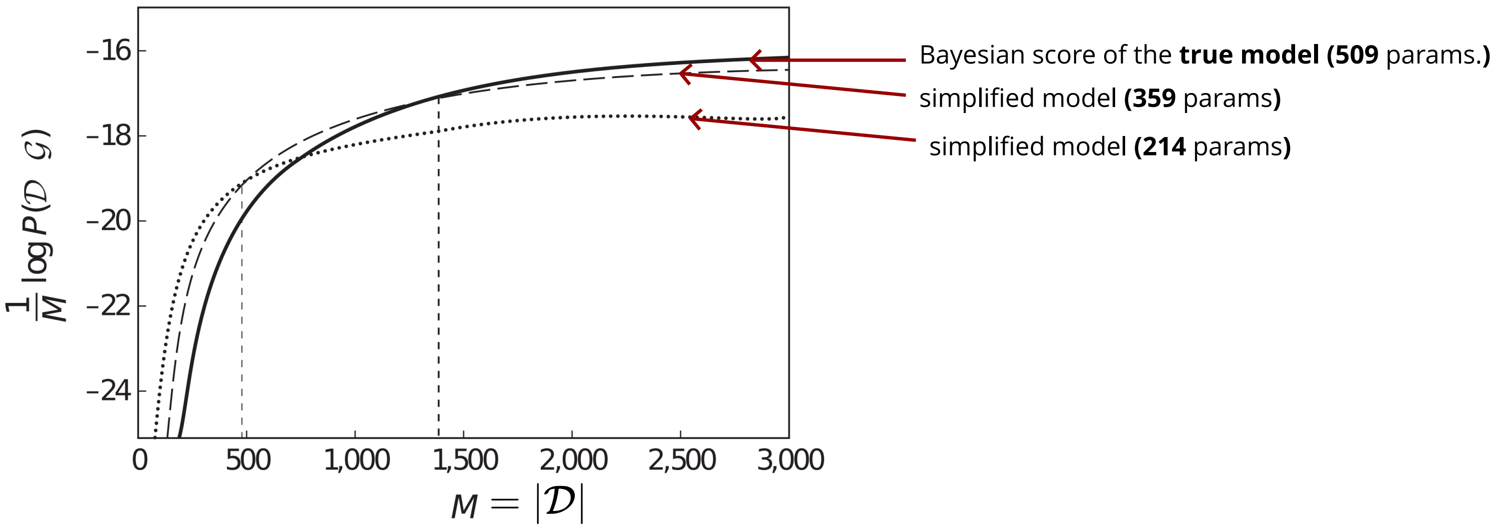
Example The Bayesian score is biased towards simpler structures



Bayesian Score for BayesNets

Example The Bayesian score is biased towards simpler structures

data sampled from ICU alarm Bayesnet



Structure search

$\arg\max_{\mathcal{G}} \text{Score}(\mathcal{D}, \mathcal{G})$ is NP-hard

use heuristic search algorithms (discussed for **MAP inference**)

Structure search

$\arg\max_{\mathcal{G}} \text{Score}(\mathcal{D}, \mathcal{G})$ is NP-hard

use heuristic search algorithms (discussed for **MAP inference**)

local search using:

- edge addition
- edge deletion
- edge reversal

Structure search

$\arg\max_{\mathcal{G}} \text{Score}(\mathcal{D}, \mathcal{G})$ is NP-hard

use heuristic search algorithms (discussed for **MAP inference**)

local search using:

- edge addition
- edge deletion
- edge reversal

$\mathcal{O}(N^2)$ possible moves

Structure search

$\arg\max_{\mathcal{G}} \text{Score}(\mathcal{D}, \mathcal{G})$ is NP-hard

use heuristic search algorithms (discussed for **MAP inference**)

local search using:

- edge addition
- edge deletion
- edge reversal

$\mathcal{O}(N^2)$ possible moves

↓

- collect sufficient statistics (frequencies)
- estimate the score

Structure search

$\arg\max_{\mathcal{G}} \text{Score}(\mathcal{D}, \mathcal{G})$ is NP-hard

use heuristic search algorithms (discussed for **MAP inference**)

local search using:
- edge addition
- edge deletion
- edge reversal

$\mathcal{O}(N^2)$ possible moves

↓

- collect sufficient statistics (frequencies)
- estimate the score

↓

use the **decomposition** of the score

Structure search

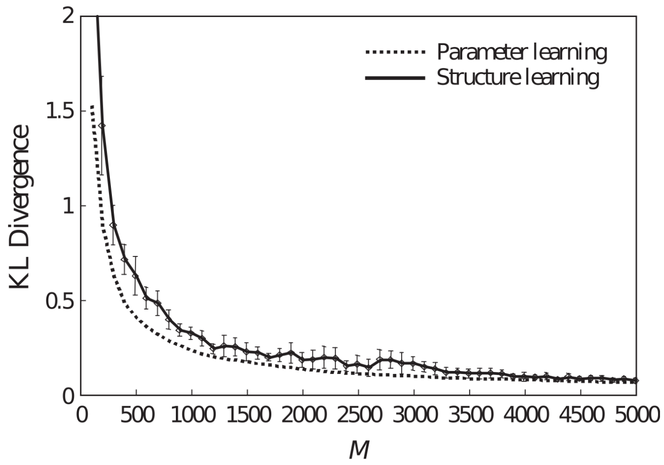
$\arg\max_{\mathcal{G}} \text{Score}(\mathcal{D}, \mathcal{G})$ is NP-hard

use heuristic search algorithms (discussed for **MAP inference**)

local search using:

- edge addition
- edge deletion
- edge reversal

example ICU-alarm network



$\mathcal{O}(N^2)$ possible moves

↓

- collect sufficient statistics (frequencies)
- estimate the score

↓

use the **decomposition** of the score

Summary

Structure learning is NP-hard

Make assumptions to simplify:

Summary

Structure learning is NP-hard

Make assumptions to simplify:

- constraint-based methods:
  - limit the max number of parents
  - rely on CI tests
  - identifies the *I-equivalence class*

Summary

Structure learning is NP-hard

Make assumptions to simplify:

- constraint-based methods:
  - limit the max number of parents
  - rely on CI tests
  - identifies the *I-equivalence class*
- score based methods:
  - tree structure
  - use a Bayesian score + heuristic search
  - finds a *locally optimal* structure