Learning Causation from Data

Fundamental of Casual Inference and its Applications

Huang Xiao

Chair of IT Security (I20)
Department of Informatics
Technische Universität München

November 16, 2012

Overview

What is Causality?

A definition from Wikipedia

Causality (also referred to as causation) is the relationship between an event (*the cause*) and a second event (*the effect*), where the second event is understood as a consequence of the first.

What is Causality?

A definition from Wikipedia

Causality (also referred to as causation) is the relationship between an event (*the cause*) and a second event (*the effect*), where the second event is understood as a consequence of the first.

An example in real life : Does smoking cause lung cancer?

What is Causality?

A definition from Wikipedia

Causality (also referred to as causation) is the relationship between an event (*the cause*) and a second event (*the effect*), where the second event is understood as a consequence of the first.

An example in real life : Does smoking cause lung cancer?

Yes, it might be!

From Probabilistic View

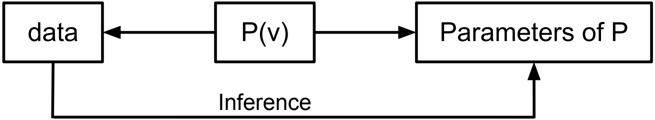
Problem: Does smoking cause lung cancer?

From Probabilistic View

Problem: Does smoking cause lung cancer?

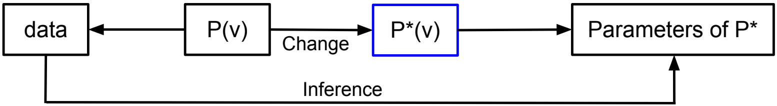
- Smoking does increase the probability of getting lung cancer.

Statistical Inference Overview



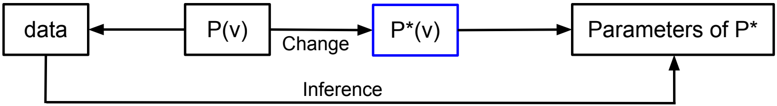
- Approximate an estimate of X given evidence e, namely, $Pr(X \mid e)$. E.g., Regression or Classification problems.
- Rejection of hypothesis, i.e., assert wether samples are from a certain distribution.
- Confidence interval, i.e., construct an interval based on dataset

Causal Inference Overview



- What if P has shifted itself to P^*?

Causal Inference Overview



- What if P has shifted itself to P^*?
- **Key factors:** Causes, Changes, and Invariants .
- Inference of P^* and reasoning of changes.

What makes Causal Inference interesting?

- Human understands the world in terms of causes and effects.
- Empirical science is about establishing causes.
- Causal inference gives a mathematical language for causal statements, and tools to solve causal problems formally.
- Alternative exercising to decision making, reasoning, etc.

Association

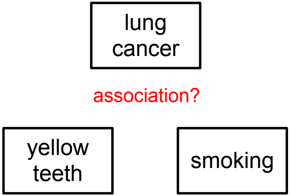
- Now we want to find out what causes lung cancer

Association

- Now we want to find out what causes lung cancer

		Lung cancer	
smoking	yellow teeth	yes	no
yes	yes	100	400
yes	no	100	400
no	yes	1	450
no	no	9	8540

Table: Data observations from 10000 people



Measurements of Association

To find out associations among variables

- Mutual information (Information theory)
- Pearson (linear) correlation
- Spearman's rho (rank correlation)
- Effect size between two variables
- Many others

Observations from Data

Obviously

- *yellow teeth* and *lung cancer* are associated.

Observations from Data

Obviously

- *yellow teeth* and *lung cancer* are associated.

But...

- Bleaching the teeth does not help reduce the probability of getting lung cancer.

Observations from Data

Obviously

- *yellow teeth* and *lung cancer* are associated.

But...

- Bleaching the teeth does not help reduce the probability of getting lung cancer.

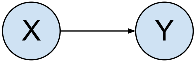
Caution!

Correlation does not imply Causation

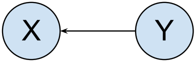
Statistical Implication

Reichenbach's *Common Cause Principle*

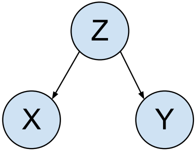
If X and Y are correlated, then either X causes Y or Y causes X or they share a latent common cause Z.



(a) X causes Y



(b) Y causes X

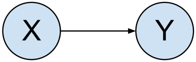


(c) A common latent cause Z

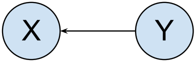
Statistical Implication

Reichenbach's *Common Cause Principle*

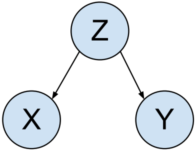
If X and Y are correlated, then either X causes Y or Y causes X or they share a latent common cause Z.



(a) X causes Y



(b) Y causes X



(c) A common latent cause Z

- It links causality with probability

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \ldots, X_n\}$

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \ldots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \ldots, X_n\}$

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \ldots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \ldots, X_n\}$
- Parents of node X_i in $\mathcal{G}$ are its direct causes

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \ldots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \ldots, X_n\}$
- Parents of node X_i in $\mathcal{G}$ are its direct causes
- $X_i = f_i(Parents(X_i), \epsilon_i)$, where $\{\epsilon_1, \ldots, \epsilon_n\}$ are jointly independent noises

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \ldots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \ldots, X_n\}$
- Parents of node X_i in $\mathcal{G}$ are its direct causes
- $X_i = f_i(Parents(X_i), \epsilon_i)$, where $\{\epsilon_1, \ldots, \epsilon_n\}$ are jointly independent noises
- The above entails a joint probability distribution $P(X_1, \ldots, X_n)$

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \ldots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \ldots, X_n\}$
- Parents of node X_i in $\mathcal{G}$ are its direct causes
- $X_i = f_i(Parents(X_i), \epsilon_i)$, where $\{\epsilon_1, \ldots, \epsilon_n\}$ are jointly independent noises
- The above entails a joint probability distribution $P(X_1, \ldots, X_n)$
- Problems are twofold:

Functional Causal Model (pearl et al.)

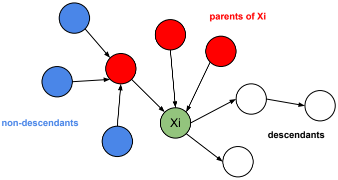
- A set of variables (factors) $\{X_1, \dots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \dots, X_n\}$
- Parents of node X_i in $\mathcal{G}$ are its direct causes
- $X_i = f_i(Parents(X_i), \epsilon_i)$, where $\{\epsilon_1, \dots, \epsilon_n\}$ are jointly independent noises
- The above entails a joint probability distribution $P(X_1, \dots, X_n)$
- Problems are twofold:
  1. How is the P like?

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \dots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \dots, X_n\}$
- Parents of node X_i in $\mathcal{G}$ are its direct causes
- $X_i = f_i(Parents(X_i), \epsilon_i)$, where $\{\epsilon_1, \dots, \epsilon_n\}$ are jointly independent noises
- The above entails a joint probability distribution $P(X_1, \dots, X_n)$
- Problems are twofold:
  1. How is the P like?
  2. Can we recover $\mathcal{G} from P$?

Functional Causal Model (pearl et al.)

- A set of variables (factors) $\{X_1, \ldots, X_n\}$
- Directed acyclic graph $\mathcal{G}$ with vertices $\{X_1, \ldots, X_n\}$
- Parents of node X_i in $\mathcal{G}$ are its direct causes
- $X_i = f_i(Parents(X_i), \epsilon_i)$, where $\{\epsilon_1, \ldots, \epsilon_n\}$ are jointly independent noises
- The above entails a joint probability distribution $P(X_1, \ldots, X_n)$
- Problems are twofold:
  1. How is the P like?
  2. Can we recover $\mathcal{G} from P$?



Functional Causal Model, ctd.

The following are equivalent:

- A functional causal model exists
- Local causal Markov condition: X_i is statistically independent of its non-descendants given X_i's parents
- Global Causal Markov condition: **d-separation** characterize the set of independences over all the observables
- Factorization: $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Parents(X_i))$

Learning causation from Data?

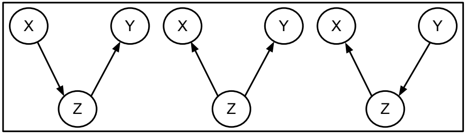
Question

Given observational data, can we infer $\mathcal{G}$?

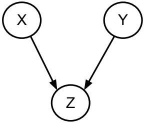
- **Simple answer:** impossible without additional information
- Possible with interventions (outside force, empirical treatment, etc.)
- By conditional independence tests, *Markov equivalence class* containing $\mathcal{G}$ can be learned. But, it fails in simplest 2-nodes case.
- 2-nodes case can be tackled applying residual dependence test. (see Hoyer et al.)

Markov Equivalence Class

Simplest case with three variables



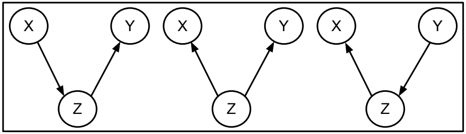
(a) Equivalence



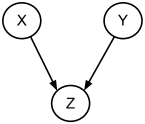
(b) Non-equivalence

Markov Equivalence Class

Simplest case with three variables



(a) Equivalence

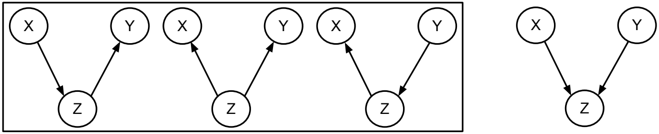


(b) Non-equivalence

- Samples can be explained by all graphs in equivalence class

Markov Equivalence Class

Simplest case with three variables



(a) Equivalence (b) Non-equivalence

- Samples can be explained by all graphs in equivalence class
- For example:

Equivalence class	Non-equivalence class
$Dep(X, Z \mid \emptyset)$	$Dep(X, Z \mid \emptyset)$
$Dep(Y, Z \mid \emptyset)$	$Dep(Y, Z \mid \emptyset)$
$Dep(X, Y \mid \emptyset)$	$Ind(X, Y \mid \emptyset)$
$Ind(X, Y \mid Z)$	$Dep(X, Y \mid Z)$

Overview

Assumptions

- Causal Markov Condition

Assumptions

- Causal Markov Condition
  - Every variable is independent of its non-descendants given its parents

Assumptions

- Causal Markov Condition
  - Every variable is independent of its non-descendants given its parents
  - Factorization: $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Parents(X_i))$

Assumptions

- Causal Markov Condition
  - Every variable is independent of its non-descendants given its parents
  - Factorization: $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Parents(X_i))$
- Faithfulness: causal structure fully determines independences

Assumptions

- Causal Markov Condition
  - Every variable is independent of its non-descendants given its parents
  - Factorization: $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Parents(X_i))$
- Faithfulness: causal structure fully determines independences
- Acyclic: needs to be defined in problem setting

Assumptions

- Causal Markov Condition
  - Every variable is independent of its non-descendants given its parents
  - Factorization: $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Parents(X_i))$
- Faithfulness: causal structure fully determines independences
- Acyclic: needs to be defined in problem setting
- Causal sufficiency

Assumptions

- Causal Markov Condition
  - Every variable is independent of its non-descendants given its parents
  - Factorization: $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Parents(X_i))$
- Faithfulness: causal structure fully determines independences
- Acyclic: needs to be defined in problem setting
- Causal sufficiency
  - Assume no latent common cause

Assumptions

- Causal Markov Condition
  - Every variable is independent of its non-descendants given its parents
  - Factorization: $P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Parents(X_i))$
- Faithfulness: causal structure fully determines independences
- Acyclic: needs to be defined in problem setting
- Causal sufficiency
  - Assume no latent common cause
  - For efficient learning, also for causal interpretation of output

Causal Bayesian Network

Definition

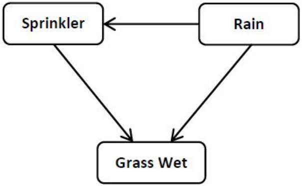
Given a set of variables $X_1, \ldots, X_n$, a Bayesian network is a probabilistic graphical model $B = (\mathcal{G}, \Theta)$, where $\mathcal{G} = (V, E)$ is a directed acyclic graph (DAG) and Θ is the set of the parameters in all conditional probability distributions (CPDs).

A Bayesian network B is said to be causal when do intervention on any subset $X \subseteq V$, i.e., do(X), resulting in a set of interventional distributions P_x, denoted by P_*, and the following three conditions hold:

- P_x is Markov relative to $\mathcal{G}$
- $P_x(v_i) = 1$ for all variables $v_i \in X$
- $P_x(v_i \mid pa_i) = P(v_i | pa_i)$ for all variables $v_i \notin X$

An example

	Sprinkler	
Rain	T	F
T	0.4	0.6
F	0.05	0.95



Rain	
T	F
0.2	0.8

		Grass Wet	
Sprinkler	Rain	T	F
F	F	0.0	1.0
F	T	0.75	0.25
T	F	0.85	0.15
T	T	0.99	0.01

Problem Setting

Goal

Given a dataset $\mathcal{D}$, try to learn the graph $\mathcal{G}$ and the parameters of all conditional probability distribution Θ.

Problem Setting

Goal

Given a dataset $\mathcal{D}$, try to learn the graph $\mathcal{G}$ and the parameters of all conditional probability distribution Θ.

Traditional method

- First step: structure learning
- Second step: parameter estimation conform to the inferred structure

Structure learning

Constraint based
Run conditional independence tests in data and find a DAG faithful to them.

- *Methods:* SGS, PC, TPDA, CPC

Hybrid
Combining both constraint based and score based.

- *Methods:* MMHC, CB, ECOS

Score based
Find a DAG by maximizing the posteriori probability given the data.

- *Methods:* K2, Sparse Candidate, GBPS, BIC/AIC

Parameter Estimation

Given the structure $\mathcal{G}$ learned from last step, factorization will be applied according to local terms governed by parameters θ_i

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid Pa_i, \theta_i)$$

Any parameter estimator will work here, e.g, MLE, MAP

Some problems in BN Learning

- Search space is exponentially large in high dimension
- Too many conditional tests
- Local minimum
- Parametric form needed
- Missing values
- Latent variables
- Hybrid node type

Overview

Copula functions

Definition

Let $U_1, \ldots, U_N$ be real random variables marginally uniformly distributed over $[0, 1]$. A Copula function C is a cumulative joint probability function: $[0, 1]^N \rightarrow [0, 1]$.

$$C(u_1, \ldots, u_N) = P(U_1 \leq u_1, \ldots, U_N \leq u_N)$$

A Copula function C can be viewed as a probability function of points distribution in a N-dimensional unit hypercube.

Sklar's theorem

Copula function is important because of the Sklar's theorem

Theorem (Sklar 1959)

Let $F(x_1, \ldots, x_N)$ be any cumulative multivariate distribution over real-valued random variables, then there exists a copula function C such that

$$F(x_1, \ldots, x_N) = C(F(x_1), \ldots, F(x_N)),$$

where $F(X_i)$ is marginal cumulative density distribution of variable X_i and furthermore if each $F(X_i)$ is continuous then the Copula is unique.

Copula Multivariate Modelling

Advantages

- Total independent free choice of marginal distributions
- Ability of transforming any joint distribution into a specific parametric form
- Decrease the number of paramters to be estimated dramatically
- Non-parametric estimators are allowed on marginals

Copula Multivariate Modelling

Advantages

- Total independent free choice of marginal distributions
- Ability of transforming any joint distribution into a specific parametric form
- Decrease the number of paramters to be estimated dramatically
- Non-parametric estimators are allowed on marginals

Multivariate modelling by Copula functions

1. Finding univariate marginals via either parametric or non-parametric ways

Copula Multivariate Modelling

Advantages

- Total independent free choice of marginal distributions
- Ability of transforming any joint distribution into a specific parametric form
- Decrease the number of paramters to be estimated dramatically
- Non-parametric estimators are allowed on marginals

Multivariate modelling by Copula functions

1. Finding univariate marginals via either parametric or non-parametric ways
2. Defining a Copula function to capture the dependence structure of model

Gaussian Copulas

Gaussian Copula is a widely used Copula function because of its extensive practical importance in many fields and also for computational simplicity. It has the form as follows:

$$C\left(\{F\left(x_i\right)\}\right) = \Phi_\Sigma\left(\phi^{-1}\left(F\left(x_1\right)\right), \ldots, \phi^{-1}\left(F\left(x_n\right)\right)\right)$$

where ϕ is standard normal distribution, Φ_Σ is zero mean normal distribution with correlation matrix Σ.
Other Copulas like Archimedean Copulas, Clayton Copulas, Vine Copula Models are also well studied.

Gaussian Copulas, ctd.

Taking the N-th order derivatives of C, we obtain the Gaussian Copula density function $c\left(\{F\left(x_i\right)\}\right) =$

$$\frac{1}{\sqrt{det\Sigma}} exp \left(-\frac{1}{2} \begin{pmatrix} \phi^{-1}\left(F\left(x_1\right)\right) \\ \vdots \\ \phi^{-1}\left(F\left(x_N\right)\right) \end{pmatrix}^T \left(\Sigma^{-1} - \mathbf{I}\right) \begin{pmatrix} \phi^{-1}\left(F\left(x_1\right)\right) \\ \vdots \\ \phi^{-1}\left(F\left(x_N\right)\right) \end{pmatrix} \right)$$

where $\mathbf{I}$ is the identity matrix. Using Sklar's theorem, the multivariate Gaussian density distribution can be obtained. In a learning scheme, the correlation matrix Σ is the only parameters to be estimated when univariate marginals are known from data observations.

Conditional Copula Density Function

Let x denote a variable and $\mathbf{y} = \{y_1, \ldots, y_k\}$ are the parents of x. And $f(x \mid \mathbf{y})$ is the conditional density function and $f(x)$ denotes the marginal density of x. And there exists a Copula density function $c(F(x), F(y_1), \ldots, F(y_k))$ such that:

$$f(x|\mathbf{y}) = R_c(F(x), F(y_1), \ldots, F(y_k))$$

where R_c is the Copula ratio

$$\begin{aligned} R_c(F(x), F(y_1), \ldots, F(y_k)) &\equiv \frac{c(F(x), F(y_1), \ldots, F(y_k))}{\int c(F(x), F(y_1), \ldots, F(y_k)) f(x) dx} \\ &= \frac{c(F(x), F(y_1), \ldots, F(y_k))}{\frac{\partial^k C(1, F(y_1), \ldots, F(y_k))}{\partial F(y_1) \ldots \partial F(y_k)}} \end{aligned}$$

and R_c is defined to be 1 when $\mathbf{y} = \emptyset$.

Factorization of Copulas

Consider again the factorization of Bayesian network:
$p(X) = \prod_{i=1}^{m} p(x_i \mid \mathbf{Pa_i})$
Copulas can be decomposed in a similar way:

Decomposition of Copulas

Given a directed acyclic graph $\mathcal{G}$ encoding conditional independences over $\mathcal{X}$, the Copula density $c(F(x_1), \ldots, F(x_N))$ can also be decomposed according to $\mathcal{G}$

$$c(F(x_1), \ldots, F(x_N)) = \prod_i R_{c_i}(F(x_i), \{F(\mathbf{Pa_{ik}})\})$$

where c_i is the local Copula ratio on variable x_i

Copula Bayesian Network Model

Definition

A Copula Bayesian Network (CBN) is a triplet $\mathcal{C} = (\mathcal{G}, \Theta_C, \Theta_f)$ encoding the joint density $f_{\mathcal{X}}(x)$. Θ_C is a set for all local Copula densities $c_i(F(x_i), \{F(\mathbf{Pa_{ik}})\})$ and Θ_f is the set of parameters representing the univariate marginals $f(x_i)$. Then $f_{\mathcal{X}}(x)$ can be parameterized as

$$f_{\mathcal{X}}(x) = \prod_i R_{c_i}(F(x_i), \{F(\mathbf{Pa_{ik}})\}) f(x_i)$$

By sharing the global univariate marginals, the hypothesis space on parameters has been largely reduced.

Parameter Estimation

In the case of Gaussian Copula, and we take the MLE method given data observations T. The log-likelihood can be written as:

$$\ell(\theta) = \sum_{t=1}^{|T|} \ln c(F_1(x_1^t; \theta_1), \cdots, F_N(x_N^t; \theta_N), \alpha) + \sum_{t=1}^{|T|} \sum_{n=1}^{N} \ln f_n(x_n^t; \theta_n)$$

where θ_i is the parameters of marginal distribution x_i and α is the set of parameters governing the dependencies.

Structure Learning

For structure learning, we use the partial inverse correlation matrix (PICM) method (constraint based).

Basic idea: simply inverse the estimated covariance matrix Σ in Gaussian Copula, and scale the diagonals as 1.

Structure Learning

For structure learning, we use the partial inverse correlation matrix (PICM) method (constraint based).

Basic idea: simply inverse the estimated covariance matrix Σ in Gaussian Copula, and scale the diagonals as 1.

$$\Sigma^{-1} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdot & s_{n,n} \end{pmatrix}$$

Structure Learning

For structure learning, we use the partial inverse correlation matrix (PICM) method (constraint based).

Basic idea: simply inverse the estimated covariance matrix Σ in Gaussian Copula, and scale the diagonals as 1.

$$\Sigma^{-1} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdot & s_{n,n} \end{pmatrix} \xrightarrow{scale} \begin{pmatrix} 1 & \frac{s_{1,2}}{s_{1,1}*s_{2,2}} & \cdot & \frac{s_{1,n}}{s_{1,1}*s_{n,n}} \\ \frac{s_{2,1}}{s_{2,2}*s_{1,1}} & 1 & \cdot & \frac{s_{2,n}}{s_{2,2}*s_{n,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{n,1}}{s_{n,n}*s_{1,1}} & \frac{s_{n,2}}{s_{n,n}*s_{2,2}} & \cdot & 1 \end{pmatrix}$$

Structure Learning

For structure learning, we use the partial inverse correlation matrix (PICM) method (constraint based).

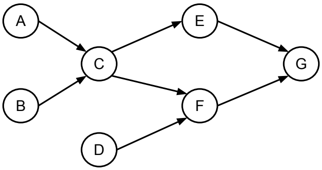
Basic idea: simply inverse the estimated covariance matrix Σ in Gaussian Copula, and scale the diagonals as 1.

$$\Sigma^{-1} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdot & s_{n,n} \end{pmatrix} \xrightarrow{scale} \begin{pmatrix} 1 & \frac{s_{1,2}}{s_{1,1}*s_{2,2}} & \cdot & \frac{s_{1,n}}{s_{1,1}*s_{n,n}} \\ \frac{s_{2,1}}{s_{2,2}*s_{1,1}} & 1 & \cdot & \frac{s_{2,n}}{s_{2,2}*s_{n,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{n,1}}{s_{n,n}*s_{1,1}} & \frac{s_{n,2}}{s_{n,n}*s_{2,2}} & \cdot & 1 \end{pmatrix}$$

If $\frac{s_{i,j}}{\sqrt{s_{i,i}*s_{j,j}}} \leq \sigma(\text{very small}) \Rightarrow X_i \perp\!\!\!\perp X_j \mid \{X_{q \neq i,j}\}$

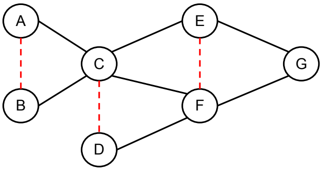
PICM to Moral Graph

A zero-entry in PICM implies no direct edge between two variables, we construct a moral graph accordingly, e.g.,



(a) Original DAG

$\Rightarrow$



(b) Moral graph

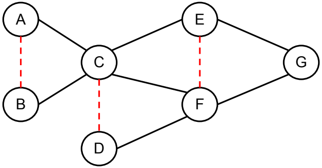
Moral is a term known as the married edge of the parents for a collider (two nodes converge at a common node, i.e., non-equivalence).

Detriangulation of Moral Graph

Note that the additional dependences are brought by colliders (see nodes C, F, and G) and by conditioned on colliders, dependences will disappear, namely, $A \not\perp\!\!\!\perp B \mid C$ but $A \perp\!\!\!\perp B \mid \emptyset$. This motivates us to remove those additional dependences (married edges).

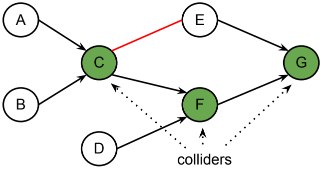
Detriangulation of Moral Graph

Note that the additional dependences are brought by colliders (see nodes C, F, and G) and by conditioned on colliders, dependences will disappear, namely, $A \not\perp\!\!\!\perp B \,|\, C$ but $A \perp\!\!\!\perp B \,|\, \emptyset$. This motivates us to remove those additional dependences (married edges).



(a) Moral graph

$\Rightarrow$



(b) Partially directed graph after removing moral edges

Contraint Propagation (Judea Pearl 2000)

Constraints: Colliders, Acyclicity

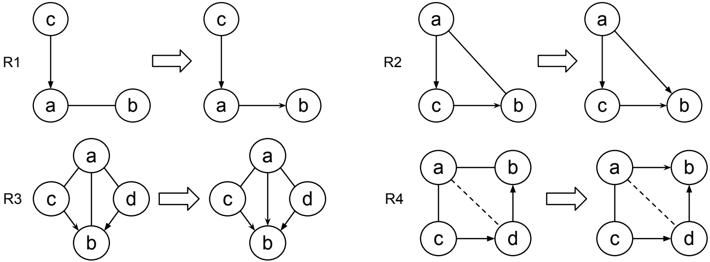
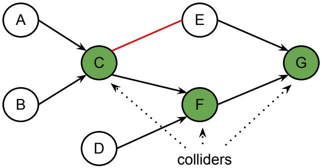


Figure: Rules for completion of orientations

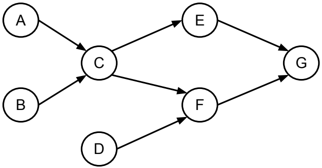
Maximally Oriented Acyclic Graph

Recursively propagate constraints, we obtain a maximally oriented acyclic graph (only equivalence class)



(a) Partially directed graph after removing moral edges

$\Rightarrow$



(b) Maximally oriented acyclic graph

Overview

Structural Hamming Distances

5 synthetic networks of size $5, 7, 10, 20, 50$. (ground truth structures are known.)

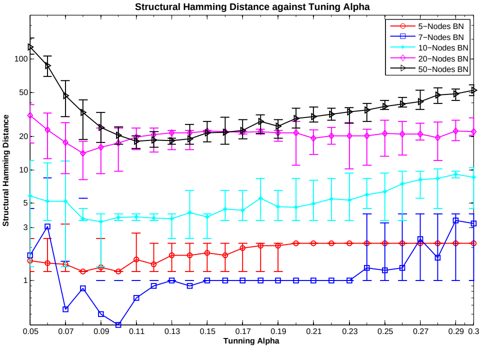


Figure: Structural hamming distance (SHD) against threshold σ

Error rates

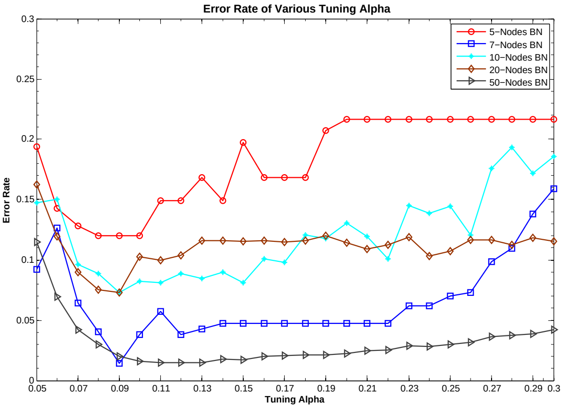
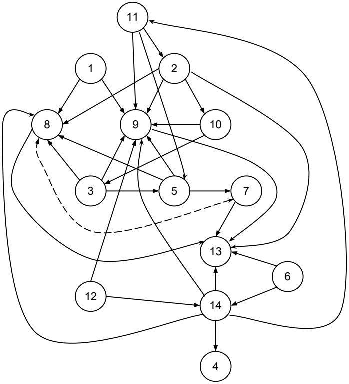


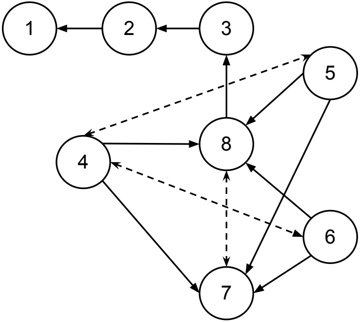
Figure: Error rates in terms of SHD

Boston Housing Price (UCI Repo.)



The factor names
1. crime rate by town
2. percent of residential land
3. proportion of non-retail business
4. Resided by Charles River?
5. nitric oxides concentration
6. average number of rooms
7. proportion of built prior to 1940
8. distances to employment centres
9. accessibility to radial highways
10. property-tax rate
11. pupil-teacher ratio
12. the percent of blacks by town
13. lower status of the population
14. Median value of own homes

Abalone Data Set (UCI Repo.)



The factor names
1. Length (longest shell)
2. Diameter
3. Height (with meat)
4. Whole weight
5. Shucked weight (without shell)
6. Viscera weight (after bleeding)
7. Shell weight
8. Rings (indicating age)

Overview

Wrap up

- Causal Inference is a powerful tool for reasoning and predicting
- Graphical model has a strong representation on causal conditions
- Learning causation from data observation is totally feasible
- Without additional information, only equivalence class can be achieved
- Copula functions enable efficient parameter estimation
- PICM based structure learning is also efficient
- Promising experimental results show the interestingness of causal inference

Overview

References

- J. Pearls, Causality: Models, Reasoning, and Inference *Cambridge University Press*, **2000**.
- Joe Whittaker., Graphical Models in Applied Multivariate Statistics, *John Wiley & Sons*, Ltd, **2008**.
- Roger B. Nelsen, An Introduction to Copulas, *Springer Science+Business Media*, Inc, **2006**.
- Jonas Peters, et al., Identifying Cause and Effect on Discrete Data using Additive Noise Models, *JMLR*, Proceedings Track 9: 597-604, **2010**.