Learning Causality

Sargur N. Srihari

University at Buffalo, The State University of New York

USA

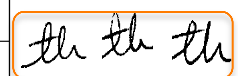
Plan of Discussion

- Bayesian Networks
- Causal Models
- Learning Causal Models

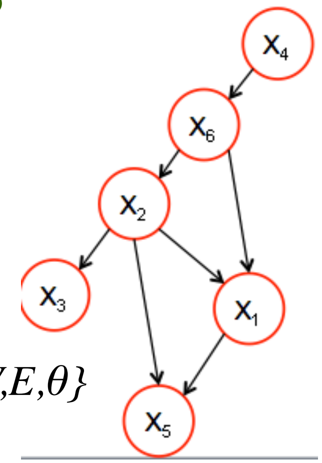
BN and Complexity of Prob Distributions

- For n variables with k states for each variable
  - Full Distribution requires k^n-1 parameters
    - with n=6, k=5 need $15{,}624$ parameters

R = Height Relationship of t to h	L = Shape of Loop of h	A = Shape of Arch of h	C = Height of Cross on t staff	B = Baseline of h	S = Shape of t
r^0= t shorter than h	l^0 = retraced	a^0 = rounded arch	c^0 = upper half of staff	b^0 = slanting upward	s^0 = tented
r^1 = t even with h	l^1 = curved right side and straight left side	a^1 = pointed	c^1 = lower half of staff	b^1 = slanting downward	s^1 = single stroke
r^2 = t taller than h	l^2 = curved left side and straight right side	a^2 =no set pattern	c^2 = above staff	b^2 = baseline even	s^2 = looped
r^3 = no set pattern	l^3 = both sides curved		c^3 = no fixed pattern	b^3 = no set pattern	s^3 = closed
	l^4 = no fixed pattern				s^4 = mixture of shapes



G={V,E,θ}



- BN Provides a factorization of joint distribution
    - Nodes are variables, edges are influences

$$P(\mathbf{x}) = P(x_4)P(x_6 \mid x_4)P(x_2 \mid x_6)P(x_3 \mid x_2)P(x_1 \mid x_2, x_6)P(x_5 \mid x_1, x_2)$$

θ: $4+(3*24)+(2*125)=326$ parameters

Organized as Six CPTs, e.g.

$P(X_5|X_1,X_2)$

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0, X_2 = 0$	0.50	0	0	0.50
$X_1 = 0, X_2 = 1$	0	1.00	0	0
$X_1 = 0, X_2 = 2$	0.18	0.36	0.27	0.18
$X_1 = 0, X_2 = 3$	0.27	0.40	0.30	0.03
$X_1 = 0, X_2 = 4$	0.22	0.45	0.28	0.05
$X_1 = 1, X_2 = 0$	0.43	0	0.28	0.29
$X_1 = 1, X_2 = 1$	NaN	NaN	NaN	NaN
$X_1 = 1, X_2 = 2$	0.39	0.06	0.33	0.22
$X_1 = 1, X_2 = 3$	0.33	0.17	0.33	0.17
$X_1 = 1, X_2 = 4$	0.42	0.11	0.29	0.18
......	......	......	......	......

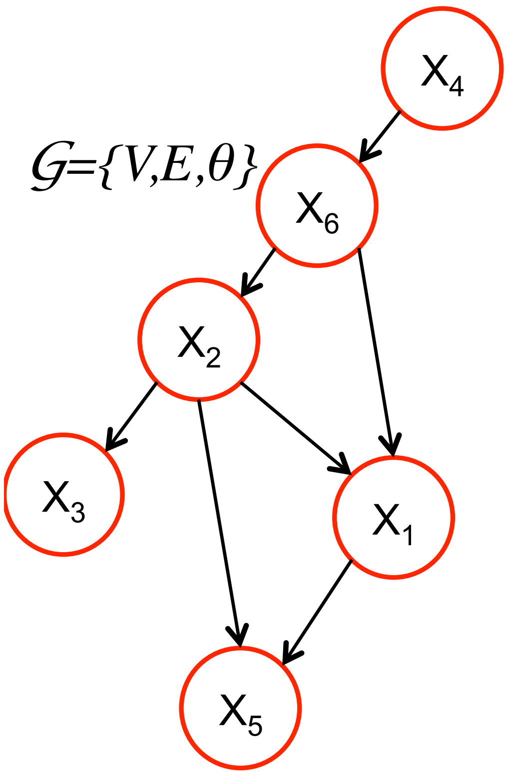
Learning Problems

- Parameters
  - When structure is specified by an expert
    - Experts cannot usually specify parameters
  - Data sets can change over time
  - Need to learn parameters when structure is learnt

- Structure
  - Cannot be easily determined by experts
  - Variables and Structure may change with new data sets

Learning Parameters of BN

- Parameters define local interactions
- Straight-forward since local CPDs



$\mathcal{G}=\{V,E,\theta\}$

Max Likelihood Estimate

$P(x_5|x_1,x_2)$

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0, X_2 = 0$	0.50	0	0	0.50
$X_1 = 0, X_2 = 1$	0	1.00	0	0
$X_1 = 0, X_2 = 2$	0.18	0.36	0.27	0.18
$X_1 = 0, X_2 = 3$	0.27	0.40	0.30	0.03
$X_1 = 0, X_2 = 4$	0.22	0.45	0.28	0.05
$X_1 = 1, X_2 = 0$	0.43	0	0.28	0.29
$X_1 = 1, X_2 = 1$	NaN	NaN	NaN	NaN
$X_1 = 1, X_2 = 2$	0.39	0.06	0.33	0.22
$X_1 = 1, X_2 = 3$	0.33	0.17	0.33	0.17
$X_1 = 1, X_2 = 4$	0.42	0.11	0.29	0.18
……	……	……	……	……

Bayesian Estimate

with Dirichlet Prior

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0, X_2 = 0$	0.29	0.14	0.29	0.29
$X_1 = 0, X_2 = 1$	0.25	0.25	0.25	0.25
$X_1 = 0, X_2 = 2$	0.25	0.38	0.25	0.12
$X_1 = 0, X_2 = 3$	0.22	0.41	0.31	0.06
$X_1 = 0, X_2 = 4$	0.16	0.52	0.25	0.07
$X_1 = 1, X_2 = 0$	0.29	0.14	0.29	0.29
$X_1 = 1, X_2 = 1$	0.25	0.25	0.25	0.25
$X_1 = 1, X_2 = 2$	0.37	0.05	0.47	0.11
$X_1 = 1, X_2 = 3$	0.33	0.22	0.33	0.11
$X_1 = 1, X_2 = 4$	0.38	0.13	0.29	0.20
……	……	……	……	……

Dirichlet Prior

Prior $\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ $\alpha_1 = \ldots = \alpha_k = 1$

Likelihood $O = \{o_1, \ldots, o_k\} \sim \mathrm{Multinomial}(\theta_1, .., \theta_k)$

Posterior $\boldsymbol{\theta}|O \sim \mathrm{Dirichlet}(\alpha'_1, \ldots, \alpha'_k)$

$$\alpha'_i = \alpha_i + o_i, \text{for } i = 1, \ldots, k$$

BN Structure Learning

1. Local: Deviance from Independence Tests

$$d_{\chi^2}(\mathcal{D}) = \sum_{x_i, x_j} \frac{\left(M[x_i, x_j] - M \cdot \hat{P}(x_i) \cdot \hat{P}(x_j)\right)^2}{M \cdot \hat{P}(x_i) \cdot \hat{P}(x_j)} \qquad d_{\mathcal{I}}(\mathcal{D}) = \frac{1}{M} \sum_{x_i, x_j} M[x_i, x_j] \log \frac{M[x_i, x_j]}{M[x_i] M[x_j]}$$

   1. Rule for accepting/rejecting hypothesis of independence

$$R_{d,t}(\mathcal{D}) = \begin{cases} \text{Accept} & d(\mathcal{D}) \leq t \\ \text{Reject} & d(\mathcal{D}) > t \end{cases}$$

False Rejection probability due to choice of t is its p-value

2. Global: Structure Scoring
   – Goodness of Network

Independence Tests

1. For variables x_i, x_j in data set $\mathcal{D}$ of M samples
   1. Pearson's Chi-squared ($\mathcal{X}^2$) statistic

$$d_{\chi^2}(\mathcal{D}) = \sum_{x_i, x_j} \frac{\left(M[x_i, x_j] - M \cdot \hat{P}(x_i) \cdot \hat{P}(x_j)\right)^2}{M \cdot \hat{P}(x_i) \cdot \hat{P}(x_j)}$$

Sum over all values of x_i and x_j

      - Independence → $d_{\mathcal{X}}(\mathcal{D})=0$, larger value when Joint $M[x,y]$ and expected counts (under independence assumption) differ

   2. Mutual Information (K-L divergence) between joint and product of marginals

$$d_{\mathcal{I}}(\mathcal{D}) = \frac{1}{M} \sum_{x_i, x_j} M[x_i, x_j] \log \frac{M[x_i, x_j]}{M[x_i] M[x_j]}$$

      - Independence → $d_{\mathcal{I}}(\mathcal{D})=0$, otherwise a positive value

- 2. Decision rule

$$R_{d,t}(\mathcal{D}) = \begin{cases} \text{Accept} & d(\mathcal{D}) \le t \\ \text{Reject} & d(\mathcal{D}) > t \end{cases}$$

False Rejection probability due to choice of t is its p-value

Structure Scoring

1. Log-likelihood Score for $\mathcal{G}$ with n variables

$$score_L(\mathcal{G} : \mathcal{D}) = \sum_{\mathcal{D}} \sum_{i=1}^{n} \log \hat{P}(x_i \mid pax_i)$$

Sum over all data and variables x_i

2. Bayesian Score

$$score_B(\mathcal{G} : \mathcal{D}) = \log p(\mathcal{D} \mid \mathcal{G}) + \log p(\mathcal{G})$$

3. Bayes Information Criterion
   - With Dirichlet prior over graphs

$$score_{BIC}(\mathcal{G} : D) = l(\hat{\theta}_G : D) - \frac{\log M}{2} Dim(\mathcal{G})$$

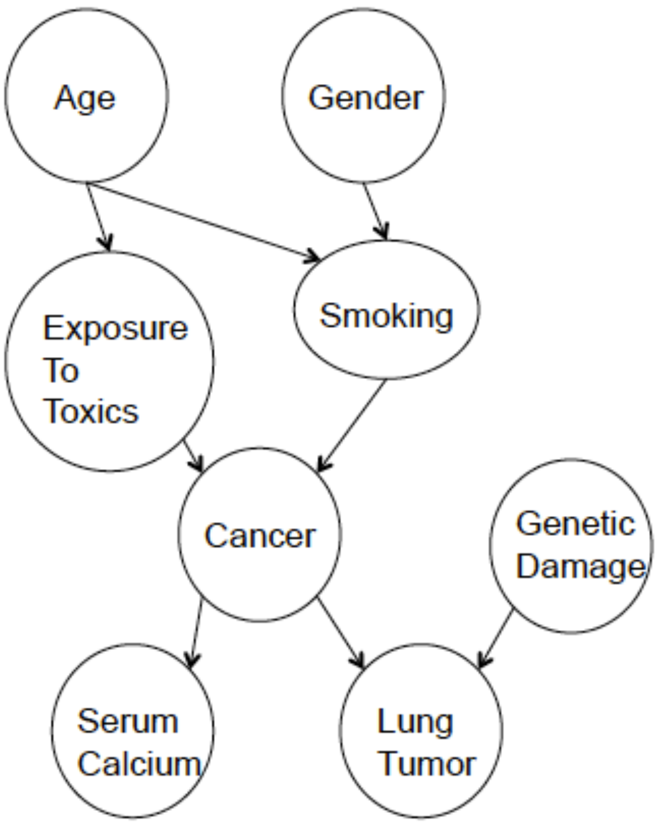
BN Structure Learning Algorithms

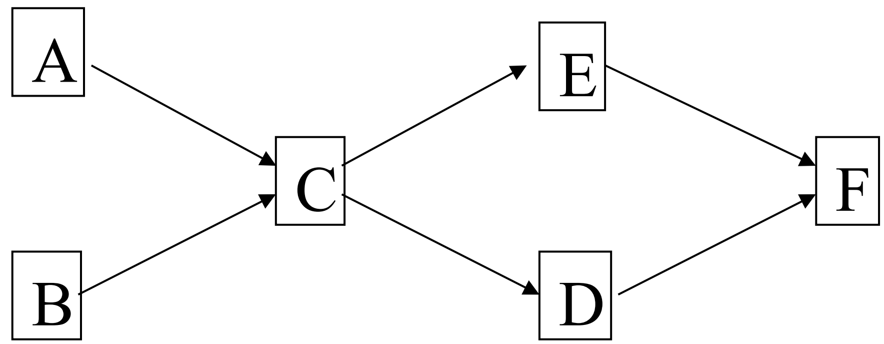
- Constraint-based
  - Find structure that best explains determined dependencies
  - Sensitive to errors in testing individual dependencies
    - Koller and Friedman, 2009
- Score-based
  - Search the space of networks to find high-scoring structure
  - Since space is super-exponential, need heuristics
    - K2 algorithm((Cooper & Herskovits, 1992)
    - Optimized Branch and Bound (deCampos, Zheng and Ji, 2009)
- Bayesian Model Averaging
  - Prediction over all structures
  - May not have closed form, Limitation of X^2
    - Peters, Danzing and Scholkopf, 2011

Causal Models

- Causality:
  - Relation between an event (the *cause*) and a second event (the *effect*), where the second is understood to be a consequence of the first
  - Examples
    - Rain causes mud, Smoking causes cancer, Altitude lowers temperature

Causal BNs





- A and B are *causally independent*;
- C, D, E, and F are *causally dependent* on A and B;
- A and B are *direct* causes of C;
- A and B are *indirect* causes of D, E and F;
- If C is prevented from changing with A and B, then A and B will no longer cause changes in D, E and F

BN is not necessarily Causal

- BN is only an efficient representation of a joint distribution in terms of conditional distributions
- Several different BNs can represent the same distribution– equivalence class

Causality in Philosophy

- Dream of philosophers
  - Democritus 460-370BC, father of modern science
    - “I would rather discover one causal law than gain the kingdom of Persia”
- Indian philosophy
  - Karma in Sanatana Dharma
    - A person’s actions causes certain effects in current and future life either positively or negatively

Causality in Medicine

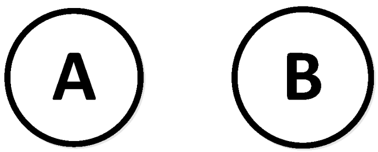
- Medical treatment
  - Possible effects of a medicine
    - Right treatment saves lives
- Vitamin D and Arthritis
  - Correlation versus Causation
  - Need for Randomized Correlation Test

Determining Causal Relationships

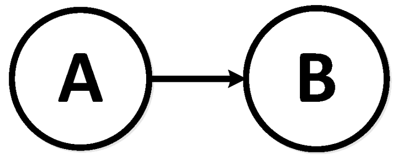
- Best way: Randomized controlled experiments
  - May be too difficult or immoral to perform
- Want to rely on “observational” data to infer causal relationships
- Current statistical methods are good at determining correlation
  - Correlation hints at causality

Relationships between events

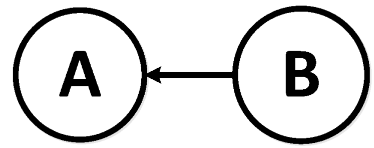
A | B



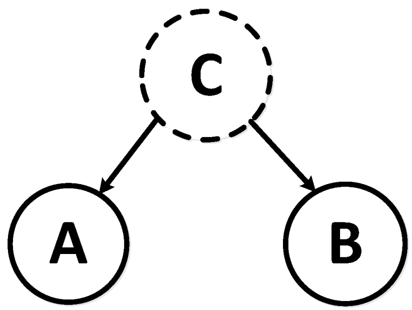
A Causes B



B Causes A



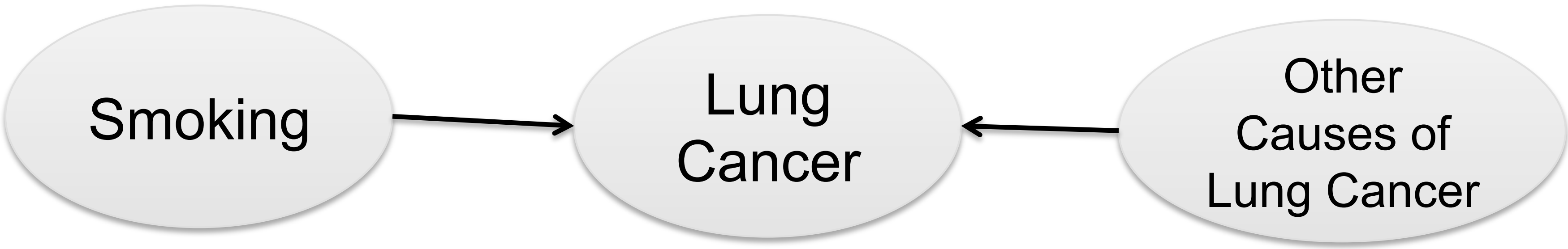
Common Causes for A and B, which do not cause each other



A = ice cream sales
B = drowning deaths
C = hot summer

Correlation is a broader concept than causation

Examples of Causal Model

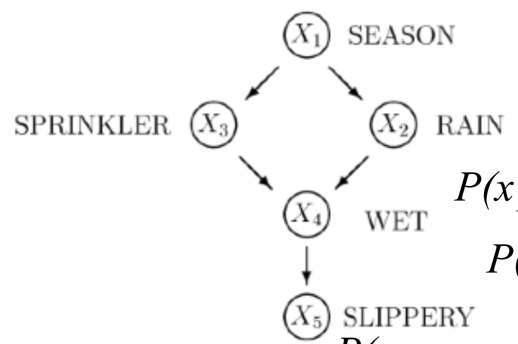


- Statement `Smoking causes cancer' implies an asymmetric relationship:
  - Smoking leads to lung cancer, but
  - Lung cancer will not cause smoking
- Arrow indicates such causal relationship
- No arrow between smoking and `Other causes of lung cancer'
  - Means: no direct causal relationship between them

Probabilistic Causality

- Deterministic causation
  - If *A* causes *B*, then *A* must *always* be followed by *B*.
    - War does not cause deaths, nor does smoking cause cancer.
- Probabilistic causation
  - *A* probabilistically causes *B* if *A*'s occurrence increases the probability of *B*
    - Smoking causes cancer
  - Reflects either imperfect knowledge of a deterministic system or system under study is inherently probabilistic, such as quantum mechanics

Inference with Causal Networks



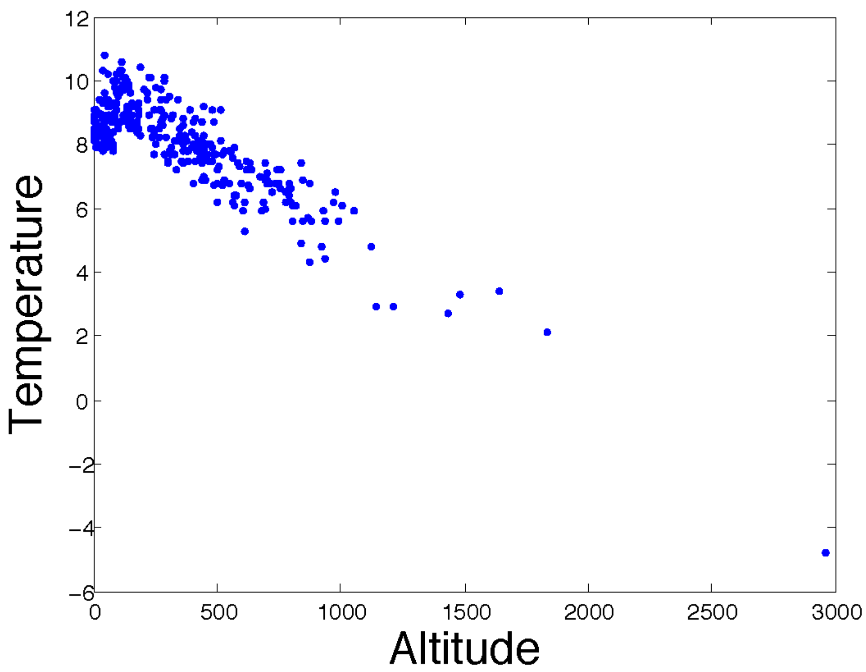
$P(x_1,x_2,x_3,x_4)=P(x_1)P(x_2|x_1)$
$P(x_3|x_1)P(x_4|x_2,x_3)P(x_5|x_4)$

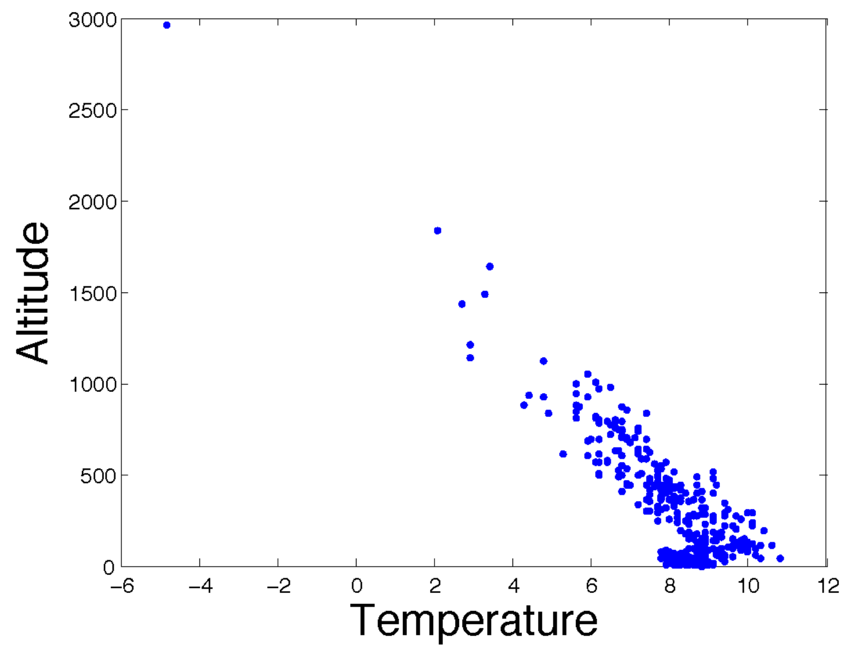
$P(x_1,x_2,x_3,x_4|do(X_3=x_3))=P(x_1)P(x_2|x_1)$
$P(x_4|x_2,X_3=on)P(x_5|x_4)$

1. Probabilistic Queries
   – Similar to other PGMs
2. Intervention Queries
   – Ideal with no other effect
      - If patient takes this medication what are chances of getting well $P(H|do(M=m^1))$
         – Where H=Health. Which is different from $P(H|m^1)$
            » Patients taking meds on their own are healthier
3. Contra-factual Queries
      - Would the accident have happened if driver was not drunk?

CAUSAL STRUCTURE LEARNING

Statistical Modeling of Cause-Effect





Data: National Climate Data Center (546 stations)

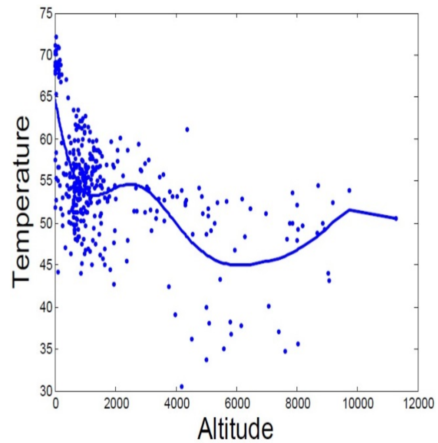
Additive Noise Model

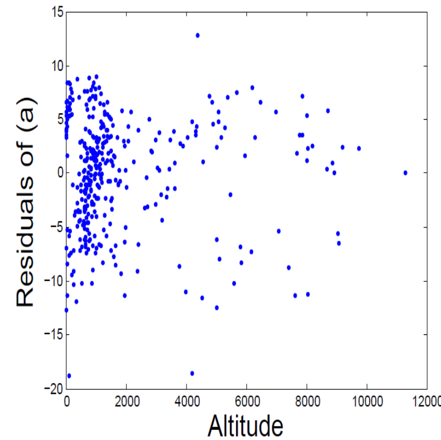
1. Test if variables x and y are independent
2. If not test if $y = f(x) + \varepsilon$ is consistent with data
   - Where f is obtained by regression
   - If residuals $\varepsilon = y - f(x)$ are independent of x then accept $y = f(x) + \varepsilon$. If not reject it.
3. Similarly test for $x = g(y) + \varepsilon$
4. If both accepted/rejected then need more complex relationship

Additive Noise Model: Example

Forward Model

$$y = f(x) + \varepsilon$$



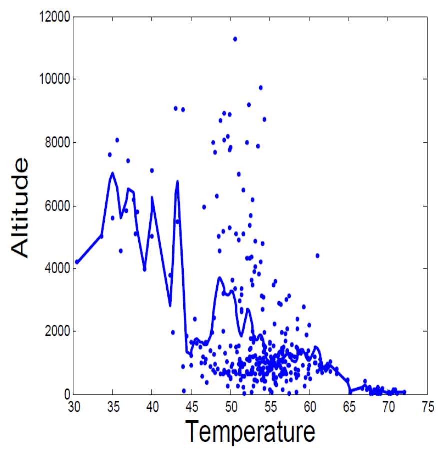


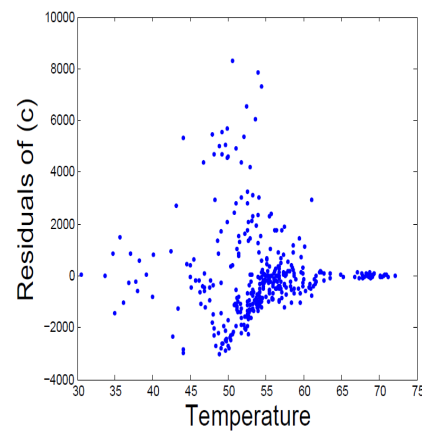
Residuals More Independent

$p=0.0026$

Backward Model

$$x = g(y) + \varepsilon$$





Residuals Less Independent

$p=5 \times 10^{-12}$

Admit altitude → temperature

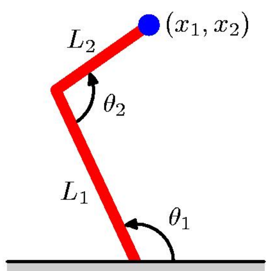
Justifying additive noise model

- Algorithmic Information Theory
  - Also called Kolmogorov Complexity
- True causal description has a shorter description

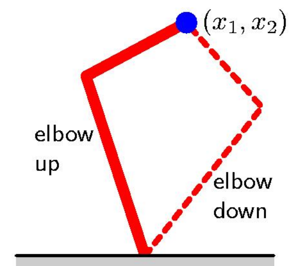
Forward and Inverse Problems

- Kinematics of a robot arm

Forward problem:
Find end effector position given joint angles
Has a unique solution



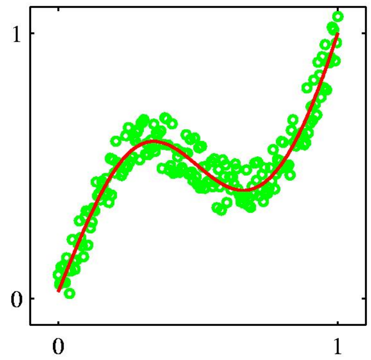
Inverse kinematics: two solutions:
Elbow-up and elbow-down



- Forward problems correspond to causality in a physical system
  have a unique solution
  e.g., symptoms caused by disease
- If forward problem is a many-to-one mapping, inverse has multiple solutions

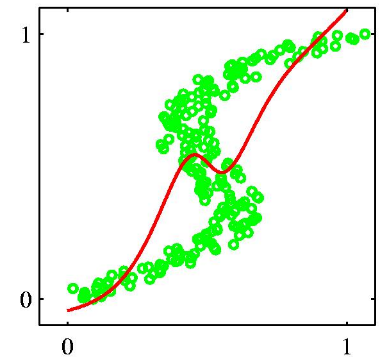
Regression Problems

Forward problem
data set



Red curve is result of
fitting a two-layer
neural network
by minimizing
sum-of-squared
error

Corresponding inverse
problem by reversing
x and t



Very poor fit
to data:
GMMs used here

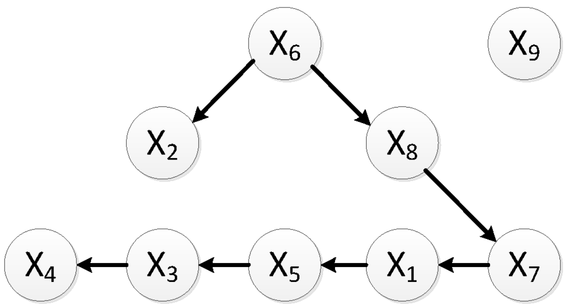
A Causal BN Structure Learning Algorithm*

- Construct PDAG by removing edges from complete undirected graph
- Use X^2 test to sort dependencies
- Orient most dependent edge using additive noise model
- Apply causal forward propagation to orient other undirected edges
- Repeat until all edges are oriented

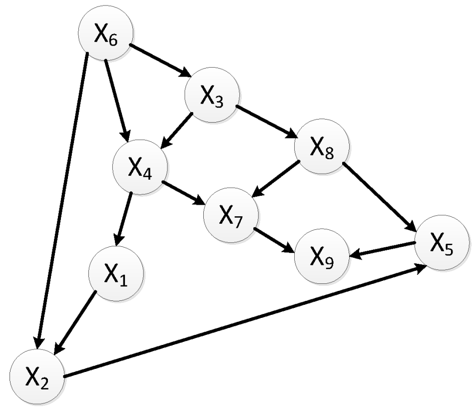
* Zhen and Srihari 2014

Comparison of Algorithms

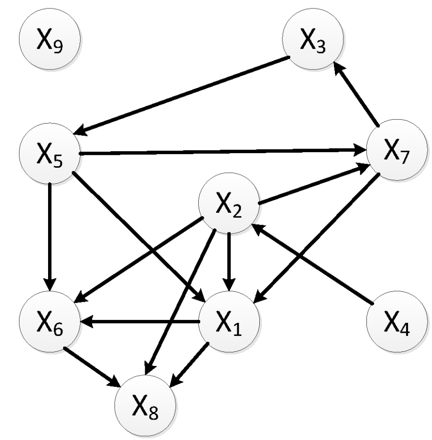
Greedy



B & B



Causal



Data			**Algorithm**			
Data Set	**Type of Data**	**No. of Vars.**	**Ind.Vars. (No Edges)**	**Greedy Algo. [11]**	**B & B Algo. [10]**	**Causal Algo.**
Set 1	Cursive	9	25994	25329	25642	25228
Set 1	Handprint	9	8059	7898	7301	7094
Set 2	Cursive	12	5316	5142	5139	5008
Set 2	Handprint	12	7825	7004	6976	6956

Some Relevant Papers

- Greedy Algorithm
  - M. Puri, S. N. Srihari, Y. Tang, "Bayesian Network Structure Learning and Inference Methods for Handwriting," *ICDAR* 2013
- Causal Algorithm
  - Zhen and Srihari, Learning Causal Networks, 2014
- PGMs:
  - Srihari, Probabilistic Graphical Models, *Encyclopedia of Social Networks*, Springer 2014
- BN Inference:
  - M. Puri, S. N. Srihari, L. Hanson, "Probabilistic Modeling of Children's Handwriting," *DRR* 2014