Causal Modeling and Causal Discovery

Causal Modeling and Causal Discovery
Tutorial
Causality Study Fortnight
University of Kent

Kevin Korb
Clayton School of IT
Monash University

kbkorb@gmail.com

Path Modeling
Reichenbach's Common Cause Principle
Bayesian networks
Causal Bayesian networks
Causal discovery algorithms
Knowledge engineering
References

Contents

Path Modeling
Reichenbach's Common Cause Principle
Bayesian networks
Causal Bayesian networks
Causal discovery algorithms
Knowledge engineering
References

Causal Graphical Models

- First systematic use of graphs for reasoning

  Wigmore (1913) charts for legal reasoning

- First systematic use of specifically causal graphs

  Sewall Wright (1921) for analysing population genetics

- Simon-Blalock method for parameterization
- Structural equation models
- Algorithms for Bayesian network modeling

  Pearl (1988)

Path Modeling
Reichenbach's Common Cause Principle
Bayesian networks
Causal Bayesian networks
Causal discovery algorithms
Knowledge engineering
References

Probability and Causality

Since probabilistic causality theory (Suppes, 1970; Salmon, 1984; etc.)

> and esp since the rise of a graphical technology that "embodies" this theory — Bayesian networks

there's been a growing consensus that the two are intimate, that

- causal structure generates probabilistic structure
- from probabilistic structure we can infer causal structure

Probabilistic Dependence

Definition (Independence)

$$X \perp\!\!\!\perp Y \text{ if and only if (iff) } P(X|Y) = P(X)$$

Example (Gambling)

E.g., Two tosses of a die are **independent**.
But, famously, two aces being drawn from a deck are not.

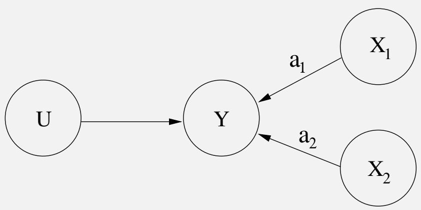
And, of course, X and Y are dependent ($X \not\perp\!\!\!\perp Y$) iff they are not independent.

Example

Rainbow and *Rain* are **dependent**.

Expectation and Correlation

Path Models



$$Y = a_1 X_1 + a_2 X_2 + U \tag{1}$$

In graphical representations U is normally omitted; in SEM models U is normally present.

Path Models

The usual assumptions:

- U is distributed as a Gaussian $N(\mu, \sigma^2)$
- U and X_i are uncorrelated (there is no hidden common cause of both X_i and Y)
- For *path models* all variables are standardized to be unit normal – $N(0, 1)$. E.g.,

$$X_1 = \frac{M_1 - \mu_1}{\sigma_1} \tag{2}$$

  where M_1 is the original (measured) variable
- In path models arcs are causal
- In SEMs "=" is strictly misleading, just like in FORTRAN: it is the causal arc $\leftarrow$ in disguise

Path Models

Theorem (Explained Variation)

Path coefficients are equal to the square root of the variation in the child variable attributable to the parent.

I.e., path coefficients squared are the amount of variation explained by the individual parents.

$$Var(X_j) = 1 = \sum_i p_{ji}^2$$

- As a consequence of standarization
- Requires a residual term U with p_{ju}

Path Model Example

Education Spending

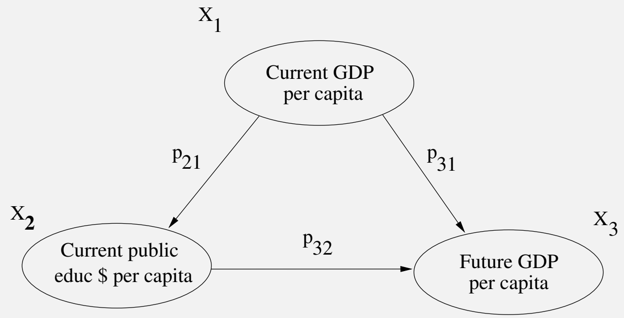


Figure: OECD Public Education Spending Model (Korb et al, 1997)

$$\begin{aligned} X_2 &= p_{21} X_1 \\ X_3 &= p_{32} X_2 + p_{31} X_1 \end{aligned}$$

Correlation

The linear dependency between two continuous variables:

Definition (Correlation)

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Correlation Example

Education Spending

r	1	2	3
1	1		
2	0.8212	1	
3	0.5816	0.6697	1

r = sample estimate of ρ

Wright's Decomposition Rule

Wright developed graphical rules for relating (observed) correlations with path coefficients.

> *Fundamental idea: correlation results from active causal influence along paths between variables.*

Definition (Active Path)

Φ_k is an **active path** between X_i and X_j iff it is an undirected path connecting X_i and X_j s.t. it does not go against the direction of an arc *after* having gone forward.

Wright's Decomposition Rule

This can be thought of as 3 rules in 1 for defining paths supporting causal influence:

1. Directed chains support causal influence
2. Common ancestors support causal influence between descendants
3. Common descendants don't support causal influence between ancestors

(This prefigures Pearl's d-separation rules.)

Wright's Decomposition Rule

To assess the strength of causal influence along an active path:

Definition (Valuation)

The valuation of a path is

$$v(\Phi_k) = \prod_{lm} p_{lm} \quad \text{for all } X_m \rightarrow X_l \in \Phi_k$$

Wright's Decomposition Rule

Definition (Wright's Decomposition Rule)

The correlation r_{ij} between variables X_i and X_j, where X_i is an ancestor of X_j, can be rewritten as:

$$r_{ij} = \sum_k v(\Phi_k) \tag{3}$$

where Φ_k is an active path between X_i and X_j and $v(\cdot)$ is a valuation of that path.

Wright's Decomposition Rule

Example

We can use Wright's Rule to generate equations for each X_i in the Education Spending model:

Example (r_{12})

$$
\begin{aligned}
r_{12} &= \sum_k v(\Phi_k) \\
&= v(X_1 \to X_2) \\
&= p_{21}
\end{aligned}
$$

Example (r_{13})

$$
\begin{aligned}
r_{13} &= v(X_1 \to X_3) + v(X_1 \to X_2 \to X_3) \\
&= p_{31} + p_{21}p_{32}
\end{aligned}
$$

Wright's Decomposition Rule

Example

Example (r_{23})

$$
\begin{aligned}
r_{23} &= v(X_2 \to X_3) + v(X_2 \leftarrow X_1 \to X_3) \\
&= p_{32} + p_{21}p_{31}
\end{aligned}
$$

I.e.,

$$
\begin{aligned}
r_{12} &= p_{21} \\
r_{13} &= p_{31} + p_{21}p_{32} \\
r_{23} &= p_{32} + p_{21}p_{31}
\end{aligned}
$$

Wright's Decomposition Rule

Example

Solving for the p_{ij}:

$$
\begin{aligned}
p_{21} &= r_{12} \\
p_{31} &= \frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2} \\
p_{32} &= \frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}
\end{aligned}
$$

So,

$$
\begin{aligned}
p_{21} &= 0.8212 \\
p_{32} &= 0.5899 \\
p_{31} &= 0.0972
\end{aligned}
$$

Fitted Education Path Model

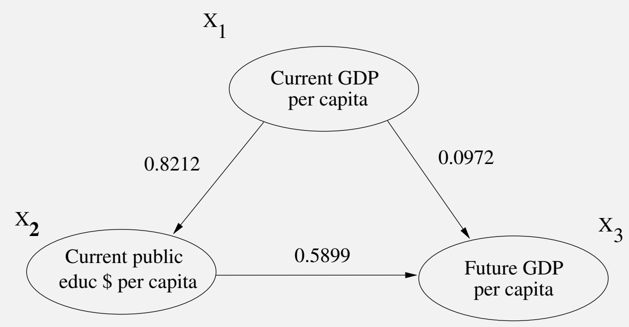


Figure: OECD Education Path Model

Non-standardized Education Path Model

We can invert the process of standardization, yielding:

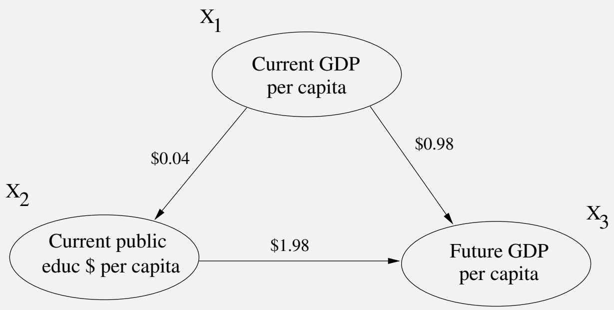


Figure: Non-standardized Education Path Model

This tells a story neo-conservatives don't want to hear!

Path Coefficients

Path coefficients are

- measures of direct linear causal power between two variables
- or, just standardized regression coefficients:

$$p_{ij} = b_{ij}\left(\frac{\sigma_j}{\sigma_i}\right)$$

I.e., we can do the same things with repeated OLS regression models.

OLS regression, however, is more complex and loses the causal story. . .

Summary

- Path models are simple and intuitive representations for linear causal models
  - graphical
  - straightforward parameterization of recursive models from sample correlations
  - non-recursive models represent unknown causal structure

Bayesian Networks

Definition (Bayesian Network)

A graph in which the following holds:

1. A set of random variables makes up the nodes in the network.
2. A set of directed links or arrows connects pairs of nodes.
3. Each node has a conditional probability table that *quantifies* the effects the parents have on the node.
4. It is a directed, acyclic graph (DAG), i.e. no directed cycles.

Pearl's Alarm Example

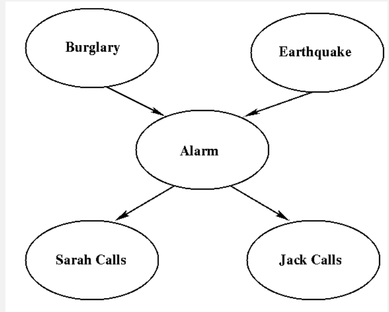
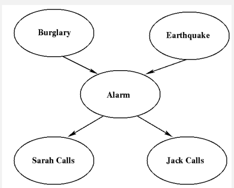


Figure: Pearl's Alarm Example

Factorization

Any joint probability distribution can be factorized using any total order. E.g.,

$$
\begin{aligned}
&P(B, E, A, S, J) \\
&\quad = \frac{P(B, E, A, S, J)}{P(J)} P(J) \\
&\quad = P(B, E, A, S|J) P(J) \\
&\quad = \ldots \\
&\quad = P(B|E, A, S, J) P(E|A, S, J) P(A|S, J) P(S|J) P(J)
\end{aligned}
$$

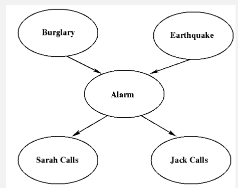


Factorization

The advantage of graphical models is that we have a grahical criterion for systematically simplifying this computation, yielding:

$$P(B, E, A, S, J) = P(S|A) P(J|A) P(A|B, E) P(B) P(E)$$

NB: Note that the order is no longer arbitrary!



The Markov condition

In order to justify the simplification, we will have to invoke (and justify) the Markov condition:

Definition (Markov Condition)
There are no direct dependencies in the system being modeled which are not explicitly shown via arcs.

Equivalently,

Definition (Markov Condition)
Every variable is independent of its non-descendants given a known state for its parents.

The Markov condition

The Markov condition is not automatically true; you have to *make* it true.

> *When it's false, there's a missing arc somewhere. The model is wrong, so go find the right model.*

Inference

Given the above, a large variety of "efficient" algorithms are available for probabilistic inference — i.e., Bayesian inference conditioning upon observations

- exact
- or approximate (complex nets)

Efficiency depends upon network complexity (esp arc density)

- worst case exponential (NP-hard; Cooper, 1990)

D-Separation

Standard BN Semantics

Standard Semantics

A representation of the joint probability distribution.

- Compactness is desirable, but not definitive
- Fully connected networks can represent *any* probability distribution, just with more difficulty (and less speed)

Compactness and Node Ordering

Compactness of BN depends upon how the net is constructued, in particular upon the underling node order

- When constructing a BN, it's best to add nodes in their natural causal order, root causes through to leaves.
- Other orderings tend to produce denser networks

Causal Ordering

Why does variable order affect network density?

Because

- Using the causal order allows direct representation of conditional independencies
- Violating causal order requires new arcs to re-establish conditional independencies

Chickering's Rule

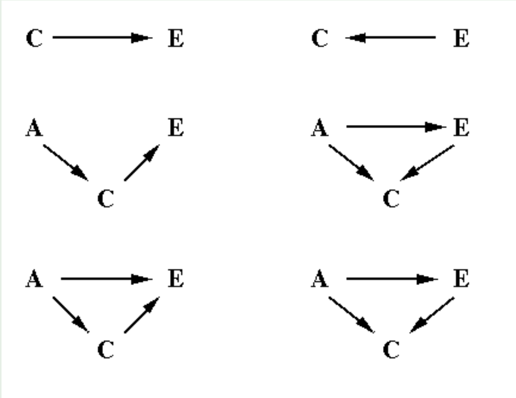
Chickering's Rule (1995)

$C \longrightarrow E$ may be changed to $C \longleftarrow E$ without loss of representational power so long as:

> *if any uncovered collision is introduced or eliminated, then a covering arc is added (e.g., if $A \longrightarrow C \longrightarrow E$ then A and E must be directly connected).*

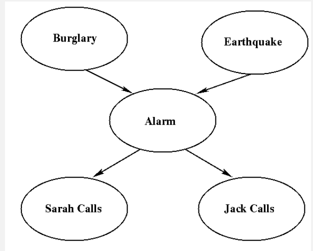
Chickering's Rule

Examples

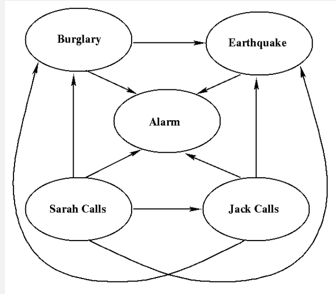


Causal Ordering

Using $\langle B, E, A, S, J \rangle$



Using $\langle S, J, B, E, A \rangle$



Causal Ordering

Dynamic Bayesian networks

Decision networks

Probabilistic Causality

It has been clear for a long time that probability and causality are intimate relations.

In recent decades the theory or probabilistic causality has converged on Bayesian networks, yielding a *causal interpretation of Bayesian networks:*

- Each arc implies a connecting causal process
- Each arc implies a difference-making connection

Two principles:

1. From causal connections arise probabilistic dependence
2. So, from dependencies observed we can infer causal relations (where there's smoke there's fire)

Manipulation and causality

The causal Markov condition

Causal power theory

Constraint-based discovery

Metric discovery

Parameterization

Problems:

Markov equivalence

Unfaithful networks

Latent variables

Variable identification

Evaluation of discovery algorithms

Elicitation of structure and probabilities

Combining expert priors with discovery

Validation

References

D.M. Chickering (1995). "A Tranformational Characterization of Equivalent Bayesian Network Structures," *UAI* (pp. 87-98). Morgan Kaufmann.

G.F. Cooper (1990). The computational complexity of probabilistic inference using belief networks. *Artificial Intelligence, 42,* 393-405.

Korb, K.B., Kopp, C. and Allison, L. (1997). *A statement on higher education policy in Australia*. Technical Report 97/318, Dept Computer Science, Monash University.

Wigmore, J. H. (1913). The problem of proof. *Illinois Law Journal 8,* 77-103.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557-585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.