Graphical Models and the PC Algorithm

Ewan Donnachie

14 July 2006

Contents

The Problem with Causality

"Causality is the centerpiece of the universe" [1]

"The central aim of many studies . . . is the elucidation of cause-effect relationships between variables or events" [2]

- Criticism of statistical science: focus on probabilistic and statistical inference at the expense of causational enquiry

[1] Causality - Wikipedia, the free encyclopedia
[2] Preface to *Pearl (2000)*

Outline

Conditional Independence

Definition (Conditional Independence)

The random variables X and Y are said to be conditionally independent given the value of a third random variable Z, if $f(X|Y,Z) = f(X|Z)$.

- Write: $X \perp\!\!\!\perp Y \mid Z$
- Intuitively, *if* Z is known, Y adds no information about the value of X.
- The difference between independence and conditional independence is demonstrated by the Yule-Simpson Paradox.

Yule-Simpson Paradox

Let n_{ij}, N_{ij}, $i \in \{1,2\}$ and $j \in \{A,B\}$, be integers. Then it is possible that:

$$\frac{n_{1A}}{N_{1A}} < \frac{n_{1B}}{N_{1B}}$$

and

$$\frac{n_{2A}}{N_{2A}} < \frac{n_{2B}}{N_{2B}}$$

but

$$\frac{n_{1A}+n_{2A}}{N_{1A}+N_{2A}} > \frac{n_{1B}+n_{2B}}{N_{1B}N_{2B}}$$

Applying this to the calculation of conditional probabilities leads to the Yule-Simpson paradox, credited to George Udny Yule (1903) and popularised by E.H. Simpson (1951).

Example: The Berkeley sex-bias case

The University of California, Berkeley, were sued for bias against women applying to grad school:

- In the university as a whole, men were more likely to be admitted to a course than women
- Examining individual departments (conditioning on the departments), there was no significant bias against women—in fact, most departments showed a slight bias against men
  Explanation:
  - women tended to apply for courses with low admission rates
  - men tended to apply for courses with high admission rates

Outline

Graphical Models

Nodes: The vertices ($i \in V$) of the graph
(Nodes and vertices used interchangeably)
Edges: Connections ($(i,j) \in E$) between vertices
Path: A route along (directed) edges from one node to another
(e.g. $i \to j \to k \to l$)

Definition (Graphical Model)

A graphical model G is a system of nodes and connecting edges: $G = (V, E)$

Why Graphical Models?

The role of graphs in probabilistic and statistical modeling is threefold:

1. to provide convenient means of expressing substantive assumptions;
2. to facilitate economical representation of joint probability functions; and
3. to facilitate efficient inferences from observations.

Conditional Independence Graph

Definition (Conditional Independence Graph)

The conditional independence graph of X is the **undirected** graph $G = (V, E)$ where $V = \{1, 2, \dots v\}$ and (i,j) is not in the edge set E iff $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$.

More informally:

- Start with the complete graph, where each node is connected to all other nodes
- Remove the edge between X_i and X_j if

$$X_i \perp\!\!\!\perp X_j \mid rest$$

N.B.: The conditional dependencies do not represent causal or directed relationships between variables.

The Pairwise Markov Property

A graph has the pairwise Markov property if, for all non-adjacent (not directly connected) vertices i and j,

$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$$

- Undirected conditional independence graphs are formed using this definition

Therefore, if X_i and X_j are non-adjacent vertices:

- they are independent conditional on the remaining nodes
- X_j is irrelevant for the prediction of X_i, and vice-versa
- *Separation Theorem*: $X_i \perp\!\!\!\perp X_j \mid rest \Rightarrow X_i \perp\!\!\!\perp X_j \mid X_a$, where X_a are the vertices separating X_i and X_j.

The Local Markov Property

A graph has the local Markov property if, for every vertix i, with boundary $a = bd(i)$ and b the set of remaining verties,

$$X_i \perp\!\!\!\perp X_b \mid X_a$$

More informally, if:

$$X_i \perp\!\!\!\perp rest \mid boundary$$

- Closely related to prediction—conditioned only on adjacent variables

The Global Markov Property

Let a, b and c be disjoint subsets of V. Then, a graph has the global Markov property if, whenever b and c are separated by a in the graph, then:

$$X_b \perp\!\!\!\perp X_c \mid X_a$$

- Global in the sense that the subsets are potentially arbitrary

Equivalence of Markov Properties

The three Markov properties: pairwise Markov, local Markov and global Markov, are equivalent.

- As the boundary set is always a separating set, global Markov $\Longrightarrow$ local Markov
- Local Markov $\Longrightarrow$ pairwise Markov
- By separation theorem, pairwise Markov $\Longrightarrow$ global Markov

Outline

Directed Acyclic Graphs

Definition (Directed Acyclic Graph)

A graph $G = (V, E)$ is called a directed acyclic graph if all edges are directed and there are no cycles (i.e. it is impossible to return to any point).

- $X \to Y \quad \Longrightarrow \quad X$ "causes" Y
- Various theorems—and background information—can be used to identify which conditional dependencies are causal in nature.
- Independent variables (i.e. no directed edge) may be dependent conditional on the remaining variables (Berkson's Paradox)

Types of Connection and d–separation

1. **Serial Connection**
   A series of nodes: $i \to j \to k$
2. **Diverging Connection**
   One node leads to several: $j \leftarrow i \to k$
3. **Converging Connection**
   Several nodes lead to one path: $j \to i \leftarrow k$

Definition (d–separation)

A set Z is said to d–separate (directionally separate) X from Y iff Y blocks every path from a node in X to a node in Y

Properties of a DAG

Definition (Faithfulness)

A distribution P is faithful to a DAG D if the all conditional independence relations for P can be derived from d–separation.

- Faithful graphs can be estimated using conditional independence relations
- Direction means that the graph is conditioned only on previous nodes
- Directed independence graphs are therefore based on the local and not pairwise Markov property

Definition (Skeleton of a DAG)

The graph generated by replacing all directed edges of a DAG with undirected edges is called a skeleton.

Outline

Estimating DAG Structures

Suppose we have a multivariate data sample and assume:

- p variables and sample size n
- $X \sim N_p(\mu, \Sigma)$
- This multivariate normal distribution is faithful
- The underlying graph is sparse (i.e. not too many edges)

Then, the structure of a DAG can be recovered using conditional independence relations.

Pairwise vs. Local Markov Property

Estimate the skeleton using the pairwise Markov, not the local Markov property:

- For any given vertex, there are 2^{p-1} ways of partitioning the remaining vertices into "boundary" and "rest" groups
- If p is large (or $p > n$), this is both computationally and statistically infeasible
- In contrast, the pairwise property has only $(k-1)$ ways of partitioning the remaining vertices

Conditional Independence

Definition (Partial Correlation)

For $i \neq j \in 1, \ldots, p$, $k \in rest$, let $\rho_{i,j|\mathbf{k}}$ be the partial correlation between X_i and X_j given X_r; $r \in \mathbf{k}$.

- As the distribution is multivariate normal,
  $X_i \perp\!\!\!\perp X_j \mid X_r \quad \Leftrightarrow \quad \rho_{i,j|\mathbf{k}}$
- A test for conditional independence is therefore a test for partial correlation between the variables
- The partial correlations can be estimated, for example, via regression

Test for Conditional Independence

Definition (Fisher's Z-Transform)

Let:

$$Z(i,j|\mathbf{k}) = \frac{1}{2}\left(\frac{1+\hat{\rho}_{i,j|\mathbf{k}}}{1-\hat{\rho}_{i,j|\mathbf{k}}}\right)$$

Then:

$$\sqrt{n-|\mathbf{k}|-3}\,|Z(i,j|\mathbf{k})| \sim N(0,1)$$

- Test for independence using classical test at significance level α Kalisch and Bühlmann show that the choice of α is not too important
- Various other tests are available, using different approaches and for different distributions

Outline

1. Introduction
   - Conditional Independence
   - Graphical Models
   - Directed Acyclic Graphs
2. Estimating DAG Structures
   - General Approach
   - The PC Algorithm
   - Example

The PC Algorithm

Start with the complete undirected graph, $\tilde{C}$ with vertices $V = X_1, \dots, X_p$. Then:

1. Set $\ell = -1$ and $C = \tilde{C}$
2. Increase ℓ by one. For all pairs of adjacent nodes:
   - Check for conditional independence
   - Remove edge (X_i, X_j) if $X_i \perp\!\!\!\perp X_j \mid rest$
3. Repeat step 2 until $\ell = m$ *or* until each node has fewer than $\ell - 1$ neighbours

Stopping level m

Let $m_reach \in \max \ell, m$ denote the stopping level of the algorithm and q be the maximum number of neighbours. It can be shown that:

1. The PC Algorithm constructs the true skeleton of the DAG
2. The stopping level is $m_reach \in q-1, 1$

Consistency of the PC Algorithm I

Let G be a DAG with probability distribution P. The following assumptions are made:

1. The distribution is multivariate normal and is faithful w.r.t. G
2. The dimension is $p_n = O(n^a),\ a \geq \infty$
   $\rightarrow$ high dimensionality
3. The maximum number of neighbours, $q_n = O(n^{1-b}),\ 0 < b \leq 1$
   $\rightarrow$ the graph is sparse
4. The partial correlations (absolute values) are bounded from above and below:
   $\rightarrow$ a regularity condition

Consistency of the PC Algorithm II

Denote by G_{skel} the true skeleton of a DAG G, and let the estimate from the PC Algorithm be $\hat{G}_{skel}$.
Then, under the above assumptions, it can be shown that, for some $C \geq 0$:

$$P(\hat{G}_{skel} = G_{skel}) = 1 - O(\exp\left(-Cn^{1-2d}\right)) \ \rightarrow 1,\ n \rightarrow \infty$$

Additionally, the stopping level is data dependent,

Outline

1. Introduction
   - Conditional Independence
   - Graphical Models
   - Directed Acyclic Graphs
2. Estimating DAG Structures
   - General Approach
   - The PC Algorithm
   - Example

Example using Simulated Data

Construct an adjacency matrix describing the conditional independence relations contained in a randomly generated graph of dimension p.

- Begin with a matrix of zeroes (i.e. no edges)
- Independent realisations of a Bernoulli random variable with parameter s determine which edges are connected. Call s the sparseness of the model
- For the edges in the graph (ones in the adjacency matrix), independent realisations of a Uniform$[0.1, 1]$ distribution are used to model the partial correlations

Then, $X_1 = \epsilon_1 \sim N(0, 1)$, and the remaining nodes are calculated recursively as follows:

$$X_i = \sum_{k-1}^{i-1} A_{ik} X_k + \epsilon_i \quad i = 2, \dots, p$$

Summary Statistics

The PC algorithm is to be compared with two alternative methods:

- Greedy Equivalent Search (GES)
- Maximum Weight Spanning Trees (MWST)

The following statistics allow their characteristics to be compared:

TDR True discovery rate, the proportion of edges in the estimated model that are edges in the true model

FPR False positive rate, the proportion of edges in the estimated model that have been falsely identified

TPR True positive rate, the proportion of true edges that have been identified by the model

Results

Method	ave[TPR]	ave[FPR]	ave[TDR]
PC	0.57 (0.06)	0.02 (0.01)	0.91 (0.05)
GES	0.85 (0.05)	0.13 (0.04)	0.71 (0.07)
MWST	0.66 (0.07)	0.06 (0.01)	0.78 (0.06)

The PC algorithm:

- achieves much higher TDR than GES or MWST
- identifies a lower proportion of the true nodes, but also has fewer false positives

Bibliography

- Kalisch, M and Bühlmann, B (2006). *Estimating high-dimensional directed acyclic graphs with the PC-Algorithm.*
- Pearl, J (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chicester.
- Lauritzen, S (2005). *Graphical Models and Inference.* Lecture notes from a course given at Oxford University.