Causal Discovery

"Proceeding" in Advanced Methods in Applied Statistics

Jakob S. Harteg (wmc573) | March 8, 2023 | University of Copenhagen

Abstract

In this write up, I present a short introduction to the PC-algorithm, a classic algorithm for causal discovery, and give a quick review of an application in Earth system science by Runge et al.[1].

Introduction

Understanding the causes behind the phenomena we observe in nature has always been at the center of scientific pursuit. Standard methods for discovering causality rely on experiments in which we control for certain variables and observe the outcome on others. In many cases, however, performing experiments can be either unethical, impractical or impossible. Luckily, the steady increase in computer power and amount of observational data, now opens up for novel data-driven methods for causal discovery that surpass common correlation analysis techniques.

Graphs as causal networks In mathematics, a graph is a set of points, called vertices or nodes, that are connected by lines, called edges. If all edges of a graph are directed, i.e. show an arrow in one direction, the graph is called directed. If the graph contains no cycles, the graph is called acyclic. Graphs that possess these two characteristics are called Acyclic Directed Graphs, or DAGs for short (see Figure 1).

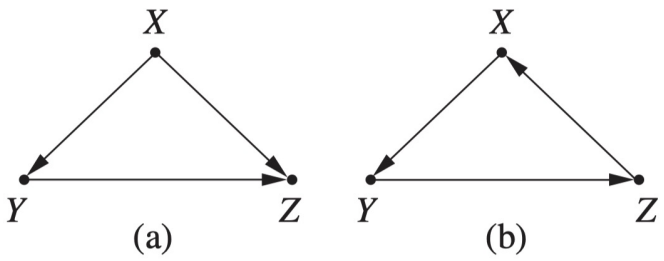

Figure 1: *a: acyclic directed graph, b: cyclic directed graph (from [2]).*

DAGs allow us to illustrate causal connections by using nodes to represent relevant variables and linking them with directed edges that indicate which variables directly affect each other. And with a few assumptions, it is possible to infer the DAG underlying a given data set and thus discover the causal structure of an observed system. There are several approaches to causal discovery and I will describe but one in this write-up: the PC-algorithm (named after its authors Peter and Clark)[3].

PC algorithm

To understand the PC algorithm we must first understand how nodes in a DAG can be either dependent, marginally independent and conditionally independent.

Independence

The undirected graph $X - Y$ signifies that X and Y are dependent, while the directed graph $X \to Y$ signifies that X causes Y. The graph $X \to Y \to Z$ implies that X causes Y which causes Z, but X does not cause Z. This can be demonstrated by *conditioning* on Y, i.e. when we hold Y fixed, X and Z become independent as the link or *path* between them has been *blocked*. We say that X and Y are conditionally independent on Z, and write $X \perp Y \mid Z$.

Figure 2a pictures a *fork*, where a variable Z cause two variables X and Y. Given such a structure, X and Y are clearly dependent, but conditioning on Z *blocks* the *path* between them, making them independent; thus $X \perp Y \mid Z$.

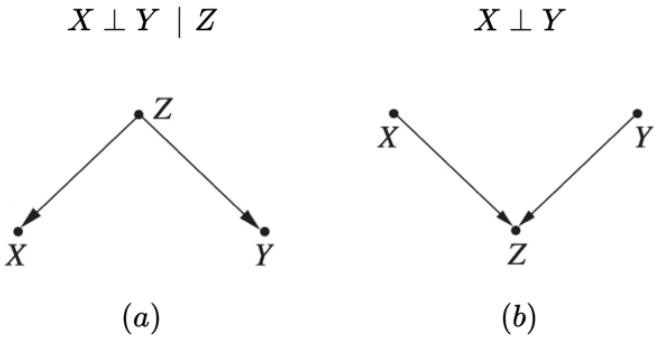

Figure 2: *(a) shows a fork and (b) a collider (adapted from [2]).*

Figure 2b pictures a *collider*. Here two variables, X and Y, join to determine the value of Z, and, as such, X and Y are independent; $X \perp Y$. But if we condition on Z, then X and Y become dependent. This is clear, since, as Z only depends on X and Y and we hold the value of Z fixed, then any change in X must be accompanied by a change in Y and vice versa. Conditioning on Z thus *unblocks* the *path*. These examples generalise into the concept of *d-separation*:

Definition 1 (d-separation, def. 1.2.3 in [4]): A path p is blocked by a set of nodes $\mathbf{Z}$ if and only if

1. p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in $\mathbf{Z}$, or
2. p contains a collider $A \to B \leftarrow C$ such that neither B nor any descendant of B is in $\mathbf{Z}$.

If $\mathbf{Z}$ blocks every path between X and Y then X and Y are d-separated conditional on $\mathbf{Z}$.

In other words, and this is key, if we are unable to find a set of nodes $\mathbf{Z}$ to condition on that blocks every path between

nodes X and Y, then X and Y must have a direct connection. If so, we say that X and Y are adjacent. With the tool of d-separation, it is thus possible to determine using conditional independence tests whether two nodes are causally connected.

The different kinds and complexity levels of conditional independence tests are many. For now, you can simply imagine that if a linear regression of X on Y given $\mathbf{Z}$ gives a slope of zero, then X and Y are conditionally independent on $\mathbf{Z}$.

Assumptions

The inference of causality from d-separation as explained depends on three assumptions.

Assumption 1 (Causal Markov Condition): Every node in a graph is conditionally independent of its non-descendants given its parents[3]. This means that every node in the graph is only directly influenced by its own parents and essentially says that if variables are d-separated, then they are conditionally independent.

Assumption 2 (Faithfulness Condition): The faithfulness condition states, in simplified terms, that independencies must arise from structure and not by coincidence[5]. Two variables may for instance appear independent if they are connected by two opposite but equal effects that cancel each other out.

Assumption 3 (Causal sufficiency): The set of observed variables is causally sufficient[3]. The assumption of Causal Sufficiency is satisfied if we have measured all the common causes of the measured variables[5].

Outline of the PC algorithm

The PC algorithm can be split into three steps: learning the skeleton, identifying colliders, and orienting edges. Each step is described in detail below, while a more rigorous description is given in[3].

1. We start with a complete, undirected graph (i.e. a graph with undirected edges between all nodes) and recursively removes edges based on conditional independence tests: an edge between X and Y is removed, if we can find a set $\mathbf{Z}$ such that $X \perp Y \mid Z$. The set $\mathbf{Z}$ must not contain X or Y, but need only be a subset of adjacent nodes to X and Y. The independence tests are ordered by levels, starting with testing every pair of nodes on the empty set, followed by sets of increasing size until the level exceeds the number of adjacent nodes.
2. Now we identify colliders by checking for each triplet of the form $A - C - B$, if C was in the conditional set for A and B. If not, the triplet must be a collider (cf. the definition of d-separation) and we orient the triplet $A \rightarrow B \leftarrow C$.
3. Since all colliders have now been identified, we can orient each remaining triplet of the form $A \rightarrow B - C$ as $A \rightarrow B \rightarrow C$.

The PC algorithm is not perfect though. It will often not be able to orient all edges, as some conditional independencies can give rise to several graphs. $X \perp Y \mid Z$ is for instance compatible with $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$ and $X \leftarrow Y \rightarrow Z$. Conditional independence tests themselves can be hard[6], the runtime is potentially exponential to the number of nodes[7], the causal sufficiency assumption can be difficult to justify, and the algorithm is variable order-dependent, meaning that it can yield different results depending on the ordering of the conditional independence tests[8]. Various extensions to the PC algorithm have since been developed to deal with these issues.

Review

One of many fields in which of causal discovery algorithms can be useful is Earth system science. In [1], Jakob Runge et al. demonstrate an extended version of the PC-algorithm (called PCMCI) that allows for identifying time lagged causal dependencies in high dimensional, nonlinear time series. PCMCI first applies the PC-algorithm to identify possible causal links at a given time, and then conducts *momentary conditional independence* (MCI) tests between that time and a number of earlier time steps to establish time-lagged causal dependencies.

Figure 3[h] reproduces the results of an example from [1] that compares PCMCI and a pure lagged correlation analysis (Corr) for identifying the Walker circulation, a well understood atmospheric circulation pattern in the tropical Pacific Ocean: Warm surface air temperature anomalies in the East Pacific (EPAC) are carried westward by trade winds across the Central Pacific (CPAC)[1]. Then, the moist air rises over the West Pacific (WPAC), and the circulation is closed by the cool and dry air sinking eastward across the entire tropical Pacific[1]. The CPAC region also links temperature anomalies to the tropical Atlantic (ATL) via an atmospheric bridge[1]. Corr results in a completely connected graph, while the PCMCI correctly identifies the underlying causal structure, including the link from EPAC $\rightarrow$ WPAC being mediated through CPAC, as well as the one way influence of CPAC on ATL.

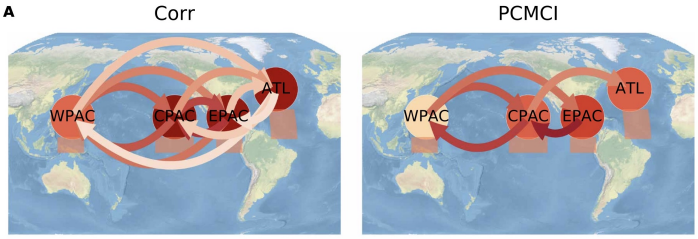

Figure 3: *Correlation compared to PCMCI for identifying the Walker circulation (from [1])*

Since it is typically difficult to justify causal sufficiency, especially in large, complex systems, direct links must always be viewed in the light of possible unknown mediators[1]. However, the absence of a direct link can indeed be interpreted as the absence of a direct causal relationship[1].

Conclusion The original PC-algorithm provides an intuitive introduction to causal discovery and lies at the foundation of many contemporary methods for inferring causality from data, a pursuit that seems very promising for improving our understanding of complex systems, especially given the ongoing growth in observational data.

References

[1] Jakob Runge et al. "Detecting and quantifying causal associations in large nonlinear time series datasets". In: *Science advances* 5.11 (2019), eaau4996.

[2] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[3] Peter Spirtes et al. *Causation, prediction, and search*. MIT press, 2000.

[4] Judea Pearl et al. "Models, reasoning and inference". In: *Cambridge, UK: CambridgeUniversityPress* 19.2 (2000).

[5] Richard Scheines. "An introduction to causal inference". In: (1997).

[6] Rajen D Shah and Jonas Peters. "The hardness of conditional independence testing and the generalised covariance measure". In: (2020).

[7] Thuc Duy Le et al. "A fast PC algorithm for high dimensional causal discovery with multi-core PCs". In: *IEEE/ACM transactions on computational biology and bioinformatics* 16.5 (2016), pp. 1483–1495.

[8] Diego Colombo and Marloes H Maathuis. "A modification of the PC algorithm yielding order-independent skeletons". In: *arXiv preprint arXiv:1211.3295* (2012).