Cognition 179 (2018) 266–297

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Successful structure learning from observational data

Anselm Rothe[a,*], Ben Deverett[b], Ralf Mayrhofer[c], Charles Kemp[d,1]

[a] Department of Psychology, New York University, NY 10003, United States
[b] Department of Molecular Biology and Princeton Neuroscience Institute, Princeton University, NJ 08544, United States
[c] Department of Psychology, University of Göttingen, Germany
[d] Department of Psychology, Carnegie Mellon University, PA 15213, United States

ARTICLE INFO

Keywords:
Causal structure learning
Causal reasoning
Bayesian modeling

ABSTRACT

Previous work suggests that humans find it difficult to learn the structure of causal systems given observational data alone. We identify two conditions that enable successful structure learning from observational data: people succeed if the underlying causal system is deterministic, and if each pattern of observations has a single root cause. In four experiments, we show that either condition alone is sufficient to enable high levels of performance, but that performance is poor if neither condition applies. A fifth experiment suggests that neither determinism nor root sparsity takes priority over the other. Our data are broadly consistent with a Bayesian model that embodies a preference for structures that make the observed data not only possible but probable.

1. Introduction

Causal networks have been widely used as models of the mental representations that support causal reasoning. For example, an engineer's knowledge of the local electricity system may take the form of a network in which the nodes represent power stations and the links in the network represent connections between stations. Causal networks of this kind may be learned in several ways. For example, an intervention at station A that also affects station B provides evidence for a directed link between A and B. Networks can also be learned via instruction: for example, a senior colleague might tell the engineer that A sends power to B. Here, however, we focus on whether and how causal networks can be learned from observational data. For example, the engineer might observe that A and B both have voltage spikes on some occasions, that B alone has voltage spikes on others, but that A is never the only station with voltage spikes (Fig. 1). Based on these observations alone, the engineer might infer that A sends power to B.

The problem in Fig. 1 is an instance of *structure* learning because it requires a choice between two distinct graph structures: one in which A sends a link to B and the other in which B sends a link to A. Structure learning can be distinguished from *parameter learning* problems that require inferences about the properties of links in a known causal structure (Danks, 2014; Jacobs & Kruschke, 2011). For example, an engineer who knows that station A sends a link to station B might need to learn about the fidelity with which signals at A are transmitted to B. Causal parameter learning is often studied experimentally using paradigms in which a focal effect is clearly distinguished from a set of potential causes, and the learning problem is to infer the strength of the relationship between each candidate cause and the effect (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Sloman, 2005). Here, however, we focus on structure learning problems in which the variables are not presorted into potential causes and effects.

A consensus has emerged that people find causal structure learning to be difficult or impossible given observational data alone. For example, Fernbach and Sloman (2009) cite results obtained by Steyvers, Tenenbaum, Wagenmakers, and Blum (2003), Lagnado and Sloman (2004), and White (2006) to support their claim that "observation of covariation is insufficient for most participants to recover causal structure" (p. 680). Here we challenge this consensus by identifying two conditions that enable successful structure learning from observational data alone. The first condition is causal determinism, and is satisfied if each variable is a deterministic function of its direct causes. The second condition is root sparsity, and is satisfied if each observation is the outcome of a single root cause. Both conditions simplify the structure-learning problem by reducing the number of possible explanations for a given set of observations.

Determinism and root sparsity have both previously been discussed in the literature on causal reasoning. Several lines of research suggest that people tend to assume that causes are deterministic or near-deterministic (Frosch & Johnson-Laird, 2011; Lu et al., 2008; Schulz & Sommerville, 2006; Yeung & Griffiths, 2015), and this assumption has informed previous studies of structure learning (Mayrhofer &

[*] Corresponding author at: Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, United States.
E-mail address: anselm@nyu.edu (A. Rothe).
[1] Present address: School of Psychological Sciences, University of Melbourne, Australia.

https://doi.org/10.1016/j.cognition.2018.06.003
Received 24 March 2017; Received in revised form 31 May 2018; Accepted 4 June 2018
Available online 02 July 2018

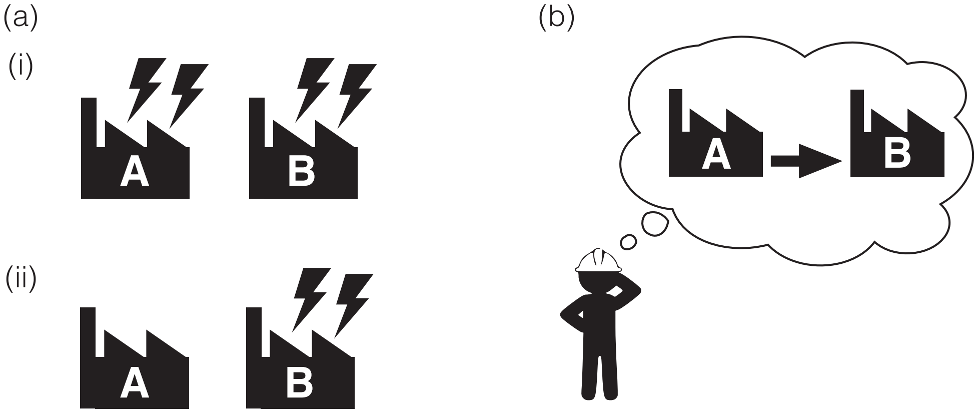

Fig. 1. Learning the causal structure of a power network given observations alone. (a) When voltage spikes are observed, either (i) stations A and B both have voltage spikes or (ii) B alone has voltage spikes. (b) These observations support the inference that station A sends power to station B.

Waldmann, 2011; Mayrhofer & Waldmann, 2016; White, 2006). Our work is related most closely to a previous study by White (2006), who asked participants to learn the structure of deterministic causal systems from observational data alone. White's task proved to be difficult, and performance was poor even when White gave his participants explicit instructions about how to infer causal structure from observational data. In contrast, we find that our participants are reliably able to infer the structure of deterministic causal systems.

Although "root sparsity" is our own coinage, this term is related to a cluster of existing ideas. Some work on causal attribution suggests that people tend to prefer explanations that invoke a single root cause (Chi, Roscoe, Slotta, Roy, & Chase, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017), although Zemla, Sloman, Bechlivanidis, and Lagnado (2017) report the opposite finding. Many studies of causal parameter learning consider cases in which there are two potential causes of an effect: a focal cause and a background cause. In this setting learners seem to expect that exactly one of these potential causes is strong (Lu et al., 2008). Mayrhofer and Waldmann (2015) explore a related idea in their work on prior expectations in structure learning. One of the priors that they consider captures the idea that an effect has a single cause. The notion of root sparsity is also consistent with studies of structure learning that focus on the role of interventions. Several researchers in this literature suggest that people tend to succeed only when interventions are not accompanied by spurious changes. If this condition holds then all changes observed following an intervention can be traced back to a single root cause – that is, to the intervention (Fernbach & Sloman, 2009; Lagnado & Sloman, 2004). Rottman and Keil (2012) show that the same condition supports structure learning from observational data if the temporal sequence of the observations is known.

Our primary goal is to explore the extent to which determinism and root sparsity allow people to succeed at structure learning. We find that people perform well when determinism and root sparsity both apply, and that either condition alone is sufficient to produce high levels of performance. To help us understand our participants' inferences, we compare these inferences to the predictions of several computational models. We initially focus on a model that we refer to as the Bayesian structure learner, or the BSL for short. The BSL serves as a normative benchmark that helps to evaluate the extent to which people succeed at structure learning. Previous discussions of structure learning have also considered Bayesian benchmarks, but Fernbach and Sloman (2009) suggest that there is "little reason to treat them as descriptively correct" (p. 681). In our setting, however, we find that people's inferences align closely with the predictions of our Bayesian model in many cases.

The BSL model contrasts with previous statistical accounts of structure learning that are sensitive to patterns of conditional independence between variables (Pearl, 2000; Spirtes, Glymour, & Scheines, 2001). Like several previous authors (Fernbach & Sloman, 2009; Mayrhofer & Waldmann, 2011), we believe that models that track patterns of conditional independence are often too powerful to capture inferences made by resource-bounded human learners. The BSL model uses statistical inference in a different way, and relies on a computation that assesses how much of a coincidence the available data would be with respect to different possible structures. It is therefore possible that people rely on a similar kind of statistical computation when approaching structure learning problems.

2. Four classes of causal networks

The causal systems that we consider are simple activation networks. Each network can be represented as a graph which may include cycles. Each node in the graph can be active or inactive, and the edges in the graph transmit activation from one node to another.

This paper will consider four qualitatively different classes of causal networks that are summarized in Table 1. The causal links in a network may be deterministic (D) or probabilistic (P), and root causes may be sparse (S) or non-sparse (N), producing a total of four possibilities that we refer to as classes DS, DN, PS, and PN.

Fig. 2a shows an example of activation spreading over a network from class DS. At stage i, node A activates spontaneously. At stage ii, node A has activated nodes B and C. At stage iii, node B has activated node D, and the network has reached a stable end state. The links in the network are deterministic, which means that they always succeed in transferring activation from one node to another. Root causes are sparse, which means that at most one node activates spontaneously per trial. As a result, each end state is the consequence of a single root cause. For example, the end state in Fig. 2a.iii is the consequence of the initial activation of node A.

Fig. 2b shows a network for which root causes are non-sparse. At

Table 1
Four classes of causal networks. For each class, the number of possible causal histories for a network with n nodes and l links is shown.

Causal strength	Number of root causes	
	1	≥ 1
Deterministic	Class DS (deterministic and sparse) Experiment 1 n	Class DN (deterministic and non-sparse) Experiment 2 2^n
Probabilistic	Class PS (probabilistic and sparse) Experiment 3 $n(2^l-1)$	Class PN (probabilistic and non-sparse) Experiment 4 $2^n(2^l-1)$