In Proceedings of the Sixth International Conference on Cognitive Modeling, 342-343.
Mahwah, NJ: Lawrence Earlbaum.

Constraint-Based Human Causal Learning

David Danks (ddanks@cmu.edu)
Department of Philosophy, Carnegie Mellon University; & Institute for Human & Machine Cognition
135 Baker Hall, Pittsburgh, PA 15213 USA

Theories of Human Causal Learning

Much of human cognition and activity depends on causal beliefs and reasoning. In psychological research on human causal learning and inference, we usually suppose that we have a set of binary potential causes, C_1, …, C_n, and a known binary effect, E, all typically *present-absent* values of a property or event. The differentiation into potential causes and effect is made on the basis of external factors, including prior knowledge or temporal information.

Given these variables, people are then asked to infer the existence and strength of causal relationships between the C_i's and E from observed data in one of several formats (serially, as a list, or in a summary). The standard measure of people's causal beliefs is a rating of some proxy for causal influence, where a zero rating indicates no causal relationship. The exact probe question varies between experiments, and has been found to significantly impact participants' ratings (e.g., Collins & Shanks, under review).

A variety of theories have been proposed to explain people's causal inferences in this type of highly limited scenario (see Danks, forthcoming, for a theoretical overview and synthesis). One general view for which there is a growing body of evidence is that people's causal beliefs and learning are well-modeled as though they are learning a causal Bayesian network (CBN, henceforth).

Causal Learning with Bayesian Networks

CBNs have proven to be a powerful framework for representing and learning causal structure from observational, experimental, and mixed data (e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 2000). At a general level, a CBN contains two distinct, related components: a directed acyclic graph (DAG) that represents qualitative causal relationships ($X \rightarrow Y$ means that X is a direct cause of Y); and quantitative information about the strengths of the various causal connections (e.g., a joint probability distribution; a set of a linear equations; and so on). These components are connected through the Markov and Faithfulness/Stability assumptions, which constrain the ways in which causal relationships manifest themselves in observational and experimental data. These assumptions are domain-general, and themselves testable.

There are essentially two different strategies for learning a CBN from data: (i) score-based or Bayesian approaches; or (ii) constraint-based (C-B) approaches. In the former approach, we search either heuristically or exhaustively for the CBN that maximizes P(CBN | observed data). We focus here on the latter approach, in which we determine the set of DAGs that could possibly have produced the observed independencies and associations (given Markov and Faithfulness). C-B algorithms thus take a set of independence and association judgments as input, and output an equivalence class of DAGs, all of which make identical predictions about the observed data. For small numbers of variables, the equivalence class will frequently *not* be a singleton, and so we will have a set of possibilities that cannot be distinguished given the data.

As an example of a C-B algorithm, suppose that we have data on X, Y, and Z. There are six different independencies (conditional and unconditional) that may or may not hold for these three variables. Suppose that some process yields the following statistical judgments about the variables: the only independence (of the six possibilities) is that X and Z are unconditionally independent. If we further suppose that there are no unobserved common causes (latents) of these variables, then there is exactly one DAG that could have produced these data: $X \rightarrow Y \leftarrow Z$. (If we drop the "no latents" assumption, then there might be unobserved common cause(s) of X and Y, or Z and Y, in addition to or in place of the $X \rightarrow Y$, $Z \rightarrow Y$, edge.)

Both types of approaches have been used for rational analyses of causal learning and categorization. Examples using Bayesian approaches include Griffiths, Baraff, & Tenenbaum (2004); and Tenenbaum & Griffiths (2001), (2003). Rational analyses with C-B algorithms include Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks (2004); and Kushnir, Gopnik, Schulz, & Danks (2003).

Constraint-Based Human Causal Learning

A range of evidence suggests that human causal learning is best-modeled by CBN structure learning, including: learning from manipulations (as opposed to just observations); learning when the variables are not differentiated into causes and effect; and differences in predictive and diagnostic learning. All of these phenomena can be explained both by C-B and Bayesian approaches to CBN structure learning, and have been modeled elsewhere.

Another important piece of evidence for the CBN theory is that people are seemingly able to use domain-specific knowledge to draw causal conclusions based on small sample sizes, and CBN learning algorithms are the only ones currently on offer that have the flexibility to model this adaptive behavior (Griffiths, et al., 2004; Tenenbaum & Griffiths, 2003).

Griffiths, et al. (2004) have drawn a further conclusion based on inferences from small samples: C-B algorithms cannot model this phenomenon—and so are incorrect—because they do not incorporate domain-specific knowledge

about the manner or functional form for the operation of possible causal mechanisms. This latter claim about C-B algorithms is based on standard machine learning implementations that use general statistical tests (e.g., χ^2 or correlation) for independence and association judgments.

Griffiths, et al. (2004) do not consider a psychological C-B learning algorithm—as opposed to a rational analysis—because one has not previously been proposed. The most straightforward such algorithm would be to estimate all of the possible independencies (conditional and unconditional) using some method, and then derive the equivalence class of DAGs from these estimates using a "standard" C-B learning algorithm (e.g., the PC algorithm of Spirtes, et al., 2000).

C-B algorithms are not computationally intensive once the independence judgments are provided, and many of their steps are intuitively sensible (e.g., if X and Y are un/conditionally independent, then conclude that there is no direct causal connection). One possible concern with the psychological plausibility of this theory is that the number of possible independencies grows exponentially with the number of variables. However, there is little evidence that people actually can learn causal structures for large numbers of variables without substantial prior knowledge constraints, and so this worry is irrelevant for psychological modeling.

This proposed theory has left relatively open the question of the source of the independence and association judgments. The C-B learning framework simply requires association/independence judgments; machine learning implementations have used general statistical tests, but the framework itself does not require that choice. So a C-B algorithm could, for example, estimate the probability of association using Bayesian statistics, perhaps even using a similar method as that advocated by Griffiths, et al. (2004).

In fact, given information about the particular (type of) causal mechanism, we can use even simpler theories than full Bayesian statistics. For example, if we know or believe that the causal influences operate as though they have causal powers, as proposed by Cheng (1997), then we can use the dynamical theory of Danks, Griffiths, and Tenenbaum (2003) to estimate conditional association in an online manner. A simple transformation of that equation leads to dynamical computation of *un*conditional association, assuming causal influences combine as causal powers. And there are similar dynamical equations for a variety of other underlying "mechanisms," at least when those mechanisms are equated with functional form (see Danks, forthcoming, for an overview of some of the dynamical equations). C-B algorithms can thus provide a psychological theory—and not just rational analysis—of domain-specific causal learning phenomena by incorporating the domain knowledge into the particular statistical tests used to make estimates of independence and association.

Moreover, other proposed psychological theories of causal learning can be modeled as special cases of this psychologically plausible theory of C-B structure learning: the other theories correspond to cases in which people estimate only a subset of the possible independencies. For example, suppose prior knowledge leads us to believe that only the independencies of the C_i's and E conditional on the other C_i's are potentially relevant. (There are a variety of plausible situations in which we might believe this.) If we also believe that causal influences operate like causal powers, we will use the dynamical method of Danks, et al. (2003). The resulting theory is exactly the same as Cheng's (1997) power PC theory. That is, we can model the power PC theory as C-B learning of CBN structure, but where people only estimate a subset of the possibly relevant independencies. A similar story can be told for the other non-CBN psychological theories. Hence, this version of C-B learning explains not only the data supporting CBN structure learning, but also a range of the data that seemingly support non-CBN theories.

Acknowledgments

This research was supported in part by NASA grants NCC2-1399 and NCC2-1377.

References

Cheng, P.W. (1997). From covariation to causation: a causal power theory. *Psychological Review, 104*, 367-405.

Collins, D. J. & Shanks, D. R. (Under review). Conformity to the power PC theory of causal induction depends on type of probe question. *Memory and Cognition*.

Danks, D. (Forthcoming). Causal learning from observations and manipulations. In M. Lovett & P. Shah (Eds.), *Thinking with data*. Erlbaum.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. *Advances in Neural Information Processing Systems 15* (pp. 67-74). Cambridge, MA.: The MIT Press.

Griffiths, T. L., Baraff, E. R., & Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review, 111*, 3-32.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). Inferring hidden causes. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 699-703). Boston: Cognitive Science Society.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, prediction, and search*. 2nd edition. Cambridge, MA: The MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. *Advances in Neural Information Processing 13* (pp. 59-65). Cambridge, MA: The MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems 15* (pp. 35-42). Cambridge, MA: The MIT Press.