Modeling active learning decisions during causal learning

Anna Coenen
Department of Psychology
New York University
New York, NY 10003
anna.coenen@nyu.edu

Bob Rehder
Department of Psychology
New York University
New York, NY 10003
bob.rehder@nyu.edu

Todd M. Gureckis
Department of Psychology
New York University
New York, NY 10003
todd.gureckis@nyu.edu

Abstract

An important type of decision making concerns how people choose to gather information which reduces their uncertainty about the world. For example, when learning about a novel piece of technology, like a smartphone, people often actively intervene on various aspects in order to better understand the function of the system. Interventions allow us to tell apart causal structures that are indistinguishable through observation, but only if the right variables are intervened on. Normative models of decision making developed in the machine learning literature specify a process of comparing hypotheses to identify those interventions that will allow a learner to distinguish between them. An experiment that asked subjects to decide between two causal hypotheses found that while they often chose useful interventions, they frequently perform interventions whose expected effects were typical of one causal structure but that did not always allow the two structures to be distinguished. We interpret this tendency as a type of positive-test-strategy with a preference for outcomes that are representative of a single causal structure.

Keywords: active learning; causal learning; interventions; information search

Introduction

To learn about causal relationships in the world, we often cannot rely on passive observation (i.e., unsupervised learning) alone. In order to understand why certain variables covary, we need the ability to actively *change* them and observe the effects of these changes. *Active interventions* are thus a crucial instrument for learning what causal structures underlie patterns of covariation in the world. There exists considerable evidence in psychology that people understand how causal systems behave in response to interventions (Waldmann & Hagmayer, 2005) and that they can use the information obtained from interventions to improve their inferences (Lagnado & Sloman, 2006).

It is still an open question, however, what strategies people use to plan their interventions with the goal of learning, that is how they decide which information would be useful for learning how a causal system works. A medical researcher, for example, needs to decide which of a patient's symptoms to treat in order to find out what illness may have caused their particular pattern of symptoms. Similarly, a scientist has to choose experimental manipulations that will tell apart different scientific hypotheses.

Here, we will examine *two* broad categories of models that can be used to explain people's decision-making processes during intervention-based causal learning. Then, in a behavioral experiment with human participants, we evaluate which class of models provides the best account of our observed choice data. The following section will give a short overview of these two key modeling approaches we have explored.

Comparative strategies

One strategy that might underlie people's causal intervention decisions is based on a rational analysis of the structure learning task. According to this rational perspective, people should choose interventions that will be useful for distinguishing alternative hypotheses. There exists a large group of optimal models, or sampling norms, that have been proposed as methods for achieving this goal (Nelson, 2005). These models share the assumption that people anticipate possible outcomes of their search behavior (i.e., of their interventions), and evaluate how useful these outcomes will be for differentiating hypotheses. Importantly, they all rely on a process of *comparison*, because they only value information that can help tell apart different hypotheses.

One sampling norm that captures the goal of causal structure learning particularly well is the *Information Gain* (IG) model of hypothesis testing. The model values observations based on their potential to reduce a learner's uncertainty about which out of a number of possible hypotheses is might underlie some observed data. It was first applied to causal interventions in the machine-learning literature (Murphy, 2001; Tong & Koller, 2001). However, it has also been proposed as a mechanism that guides people's intervention choices (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003).

Non-comparative strategies

In contrast to such comparative models of intervention choice, there also exists a literature within the psychology literature which shows that people seek information that only pertains to *one specific* hypothesis at a time. For example, it has been shown in rule-learning tasks that behavior often follows a positive-test-strategy (PTS) or positivity bias. This bias is a preference for seeking affirming information given a currently held hypothesis (e.g., Klayman & Ha, 1989), rather than testing whether the rule does not hold for counterexamples.

In the causal domain, PTS could manifest in a preference to intervene on variables (*nodes* in a causal graph), with high *centrality* (e.g., Ahn, Kim, Lassaline, & Dennis, 2000) within one candidate causal structure, irrespective of other hypotheses. Nodes are central if they have a large number of direct or indirect descendant links which could be activated through an intervention and thus count as positive evidence for a given structure. This metric can be completely at odds with a comparative strategy such as IG, because the outcomes of interventions based on this strategy may not be at all helpful for distinguishing one hypothesis from its alternatives.

Goals of this study

The aim of our study is to evaluate the degree to which people engage in comparative or non-comparative search behavior during causal structure learning. To answer this question, we conducted a simple intervention experiment that was set up in a slightly biased way to facilitate the use of a comparative strategy.

Methods

In this experiment, participants were repeatedly asked to make interventions on three-node causal systems to distinguish between two causal hypotheses.

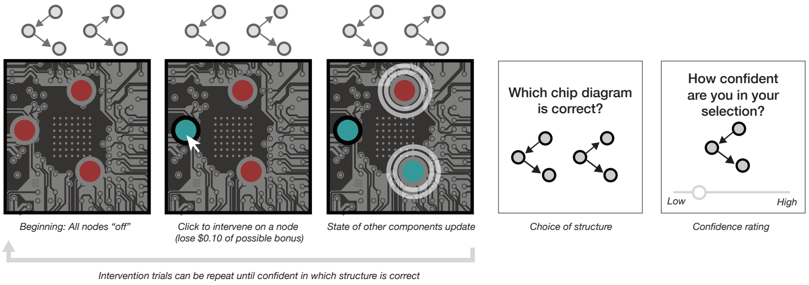
Which chip diagram is correct?
How confident are you in your selection?
Low
High
Beginning: All nodes "off"
Click to intervene on a node (lose $0.10 of possible bonus)
State of other components update
Choice of structure
Confidence rating
Intervention trials can be repeat until confident in which structure is correct

Figure 1: Intervention phase of the experiment which was repeated for each of the 27 structure comparisons. The true underlying causal graph was selected randomly. Participants could make as many interventions as they wished, but lost $0.10 of a potential bonus payment with each intervention.

Participants. We recruited one hundred and five participants (51 women and 55 men) aged 18 to 64 (M = 34.3 years, SD = 12) via Amazon Mechanical Turk. All participants were paid $2 for participation with the option of earning another $1 bonus for their performance in the task (bonus structure is explained below).

Stimuli and Materials. All possible three-node structures with one or two links were used in the experiment. They were exhaustively paired with each other to yield 27 unique structure pairs, which acted as hypotheses. All links had causal strengths of 0.8 and there were no background causes that could turn on nodes without any causal impact from another node or an intervention from the outside. In the experiment, causal graphs were described as computer chips with multiple components (nodes), which could either be on or off as indicated by their color (red or green). Hypotheses (pairs of causal graphs) were illustrated by arrow diagrams that show their respective causal links. During the task, the order of the nodes was randomized on the screen so that each node could appear in one of five different locations.

Procedure. w Participants played a game which had them imagine they were working in a computer chip factory in which an accident had caused some of the chips to be mixed up. They were instructed to help identify the types of individual chips by testing them through interventions. After an extensive instruction phase, participants tested 27 chips corresponding to all 27 causal structure comparisons. They were told that each chip could be described by one of two different chip types (hypotheses), which were presented to them with arrow diagrams. The diagrams remained at the top of the screen the entire time that a chip was tested to facilitate comparison between them. For each chip comparison, one of the hypotheses was randomly selected to be the true underlying structure of the test chip.

Figure 1 illustrates the intervention phase of the experiment. Interventions could be made by clicking on one of the nodes, which could then activate other nodes on the chip. Activated nodes changed their color (from red to green). Participants could make as many interventions as they wished, and were allowed to proceed any time when they felt they had figured out the chip type. They then indicated which of the two hypotheses was most likely given the results of their interventions.

Participants could receive a bonus of up to $1 based on one randomly chosen comparison at the end of the experiment. The bonus was only paid if they chose the correct structure at the end of that particular comparison, and it was further reduced by $0.10 for every intervention they had made. Thus, participants were incentivized to respond accurately and to use a minimal number of interventions.

Results

Model comparison.

To examine the degree to which participants rely on a comparative strategy when choosing their interventions, we calculated the expected information gain for every intervention in all 27 structure comparisons. We fit these predictions of the IG model to participants' choices using a probabilistic choice rule with a temperature scaling parameter that was estimated for each individual participant. We found that IG predicted choices well on some problem types but also considerably deviated from them on others. To make sure that these deviations were not due to just random variation, we compared bootstrapped samples from the choice data to samples from the model's posterior, separately for each problem type (plots are not shown in the interest of space). This gave us an indication of the expected uncertainty around our measurement of people's preferences, as well as the expected distribution of choices that a population of IG users would produce. Even after accounting for uncertainty in this way, model predictions from IG still deviated from the empirical data because the two sets of samples overlapped only barely or not at all on certain problem types. We conducted the same analysis using the PTS model and found similar results (good fit on some, but not all problem types).

Next, we investigated whether a propensity for non-comparative hypothesis testing, like PTS, could explain why IG did not match people's choices in some problems. To do so, we derived a measure of agreement between the two models, by calculating the rank correlation of their predictions for the preference over the three nodes in a given problem type. Figure 2 shows how this measure of model agreement relates to the goodness of fit of the IG model, in each problem type. Indeed, we find that the IG model had a lower likelihood in precisely those problem types in which its predictions conflicted with the PTS model. In addition to the bootstrapping analyses, this provides another reason to believe that deviations from IG on some problem types are not just due to random variation in the data. Instead, the model might particularly struggle on problems where other aspects of the task, like non-comparative considerations, enter people's decision process.

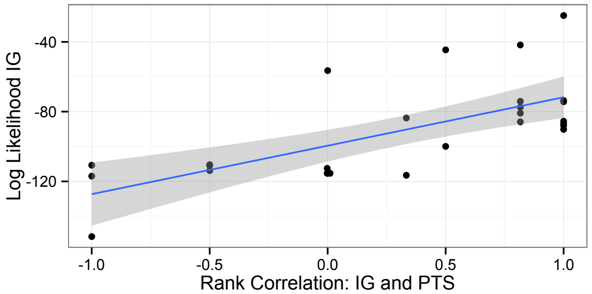

Figure 2: Log likelihood of IG model and agreement of IG and PTS (kendall's tau rank correlation), by problem type.

Finally, we fit a combined model that took a weighted combination of IG and PTS scores before applying the probabilistic choice rule. Again, weights were estimated separately for each participant. When comparing posterior samples of this combined model to bootstrapped samples of the data, we found that it made credible predictions on all 27 structure comparisons.

Reaction times.

If, as the combined model suggests, participants are influenced by both comparative and non-comparative aspects of the task, we expected that it should be particularly difficult to choose an intervention when IG and PTS make divergent predictions about which node to choose. We therefore looked at the time it took participants to make an intervention, separately for each problem type and again depending on the agreement between IG and PTS. As Figure 3 shows, people did take significantly longer to choose an intervention in problems with low model agreement, $r(25) = -0.58, p < 0.005$. This finding confirms that comparative and non-comparative components may both play a role in people's intervention decisions and, when in conflict, can make certain problems more difficult than others.

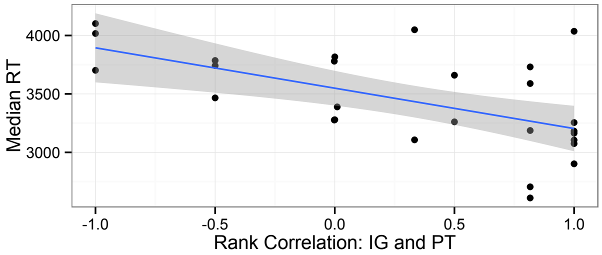

Figure 3: Median response time before making an intervention and agreement of IG and PTS (kendall's tau rank correlation), by problem type.

Individual differences

Using the combined model, we found considerable variation in the relative weights that participants place on the two strategies examined here (IG and PTS). Thus, we were interested in finding out if the individual tendency to use either IG or PTS manifested in other aspects of participants' behavior in the task, besides their intervention choices. To do so, we considered the difference in log likelihood of the separate IG and PTS models for each participant as a proxy for their tendency of making comparison-based interventions. We considered three independent variables in relation to this measure:

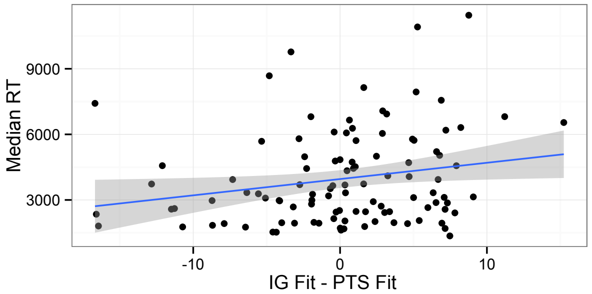

Figure 4: Response time and difference in model fit between IG and PTS, by participant.

First, we predicted that participants who are more prone to using IG, which is a computationally more intensive strategy than PTS, would take longer to decide which intervention to make. As predicted, we find that participants whose behavior is better accounted for by the IG model compared to PTS take significantly longer to choose interventions, $r(103)$ = 0.21, p = 0.03.

We also expected that IG users would be more likely to choose the correct causal structure at the end of the intervention phase. This is plausible because using IG leads to outcomes that will allow the learner to actually discriminate between graphs. It is also possible that if a non-comparative strategy is used, learners are more likely to falsely rely on outcomes that appear to provide evidence for one of the graphs, but in fact do not exclude the possibility that the alternative is true. As expected, we found a positive relationship between the degree to which participants' choices were better fit by the IG model and their average accuracy across all comparisons, $r(103) = 0.28,\ p < 0.01$, as shown in Figure 5.

Finally, we also expected comparative hypothesis testers to need fewer interventions overall before deciding which structure is correct. Again, one reason for this is that they should have received better data on average to help them actually discriminate the two graphs. Another reason is that positive testers might be tempted to want to recreate all positive effects of one of the structures and thus require more interventions to achieve this goal. As figure 6 shows, individuals better fit by IG made fewer interventions than participants who relied more heavily on the non-comparative strategy, $r(103) = -0.35,\ p < 0.001$.

In sum, the combined model of IG and PTS not only provides a better fit to people's choices, but it also has some interesting behavioral implications that we could observe in our data.

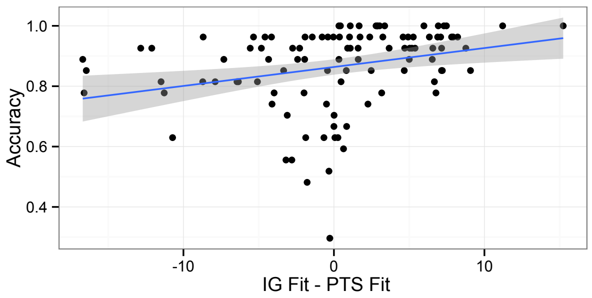

Figure 5: Accuracy and difference in model fit between IG and PTS, by participant.

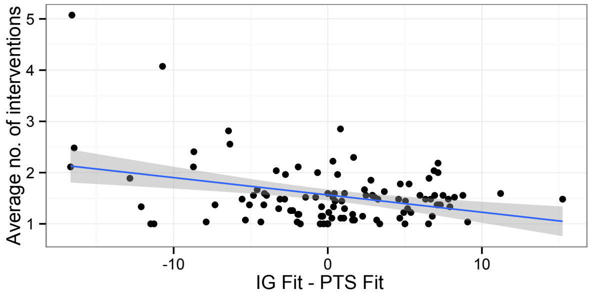

Figure 6: Number of interventions and difference in model fit between IG and PTS, by participant.

Discussion

In contrast to predictions of the rational approach to causal information search, we find that people's intervention choices not always aim at differentiating causal hypotheses. Instead, participants' choices in a simple causal intervention task were best accounted for by a model that also included preferences based on graph-specific, non-comparative features of a given problem. Specifically, participants preferred intervening on causal nodes that had the potential to trigger a large proportion of all the effects associated with *one* of the hypothesized graphs. We interpret this preference as a type of positive-test-strategy, which favors seeking information that will lead to positive outcomes that should be expected if a given graph was true. This finding is at odds with a rational model that is purely based on seeking interventions that lead to surprising outcomes, like the IG model. In reality, it looks like people's decisions are guided by both comparative and non-comparative strategies during intervention-based causal structure learning.

Going forward, we are interested in testing whether people's reliance on non-comparative strategies can be influenced by the task environment. In our current experiment, using a non-comparative strategy still led to outcomes that would enable participants to make correct graph choices, most of the time. However, if graph comparisons were designed so that non-comparative strategies would not aid learning at all, it is possible that participants would switch to a more comparison driven approach.

References

Ahn, W.-k., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*(4), 361–416.

Klayman, J., & Ha, Y.-w. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451.

Murphy, K. P. (2001). Active learning of causal bayes net structure. *Technical Report. Department of Computer Science, U.C. Berkeley.*.

Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4), 979.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive science*, *27*(3), 453–489.

Tong, S., & Koller, D. (2001). Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence* (Vol. 17, pp. 863–869).

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216.