Learning Causal Schemata

Charles Kemp, Noah D. Goodman & Joshua B. Tenenbaum
{ckemp, ndg, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Causal inferences about sparsely observed objects are often supported by causal schemata, or systems of abstract causal knowledge. We present a hierarchical Bayesian framework that learns simple causal schemata given only raw data as input. Given a set of objects and observations of causal events involving some of these objects, our framework simultaneously discovers the *causal type* of each object, the causal powers of these types, the characteristic features of these types, and the characteristic interactions between these types. Previous behavioral studies confirm that humans are able to discover causal schemata, and we show that our framework accounts for data collected by Lien and Cheng and Shanks and Darby.

Keywords: causal induction; intuitive theories; hierarchical Bayesian models; categorization

Introduction

People often make accurate causal inferences based on very sparse data. Imagine, for instance, that you are travelling in the tropics, and on your very first morning you take an anti-malarial pill and wash it down with guava juice. Soon afterward you develop a headache and wonder what might have caused it. Suppose that you have very little direct information about the two potential causes—you have never before tasted guava juice or taken anti-malarial pills. Even so, you will probably correctly attribute your headache to the pill rather than the juice.

Accurate inferences from sparse data often rely on the top-down influence of abstract knowledge. Even if you have never come across anti-malarial pills or guava juice, you probably know about the causal powers of pills in general and juices in general—in particular, you know that pills tend to cause headaches but that juices do not. Abstract causal beliefs of this sort are sometimes called causal schemata [6] or intuitive theories.

Two fundamental questions can be asked about causal schemata: how do these schemata support top-down inferences given relatively sparse data, and how are these schemata acquired? This paper develops a hierarchical Bayesian framework that provides a unified approach to both questions. Griffiths [4] has previously shown that hierarchical Bayesian models help to explain how top-down inferences can be guided by causal schemata. Here we focus on the acquisition question, and show that hierarchical Bayesian models help to explain how causal schemata can be acquired by bottom-up learning.

Top-down and bottom-up approaches to causal learning are sometimes seen as competitors. The top-down approach [6, 14] emphasizes inferences that are based on knowledge about causal powers, and the bottom-up approach emphasizes statistical inferences that are based on patterns of covariation. As Cheng [2] and others have argued, these perspectives are best regarded as complementary: top-down knowledge about causal power plays a role in many inferences, and bottom-up statistical learning can help to explain how this knowledge is acquired. The apparent conflict between these perspectives may have developed in part because there is no well-established framework that accommodates them both. Kelley, for example, argued for both top-down [6] and bottom-up approaches [7] to causal reasoning, but did not develop a single theoretical framework that properly unified his two proposals. We will argue that a hierarchical Bayesian approach provides this missing theoretical framework, and will develop a model that shows how schemata support causal reasoning and how schemata can be acquired by statistical learning.

Part of our task is to formalize the notion of a causal schema. Suppose that we are interested in a set of objects—for example, a set of pills. This paper works with schemata that assign each object to a causal type, and specify the causal powers and features of each type. Our pills, for instance, may represent four causal types—pills of type A cause headaches, pills of type B relieve headaches, and pills of types C and D neither cause nor relieve headaches. A causal schema may also specify how causal types interact. For instance, a C-pill and a D-pill may cause a headache when taken together, even though neither pill causes a headache on its own.

The first section of this paper considers the well-studied problem (Fig. 1a) of learning a causal model that captures the relationship between a single object (e.g. a pill) and an effect (e.g. a headache). Causal models for several objects can be learned independently, but this approach ignores any information that might be shared across objects: for instance, two blood-pressure medications are likely to have similar side effects, enabling us to predict that a new blood-pressure medication will cause headaches if several others already have. The second section introduces causal schemata that group the objects into types, and specify the likely causal powers of the objects belonging to each type (Fig. 1b). We show how these schemata can be acquired in settings where learners must learn a schema at the same time as they are learning causal models for many different objects.

By tracking the characteristic features of causal types, learners can often make strong predictions about a novel object before it is observed to participate in any causal interactions. For instance, predictions about a pill with a given color, size, shape and imprint can be based on the effects produced by previous pills which shared these features. The third sec-

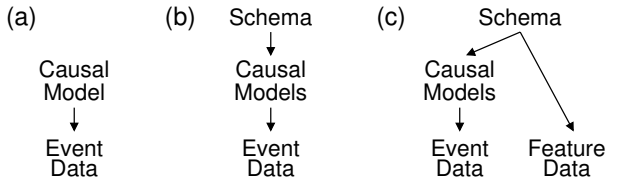

Figure 1: (a) A generative framework for discovering the causal powers of a single object. (b) A generative framework for learning a schema that guides inferences about multiple objects. The schema organizes the objects into causal types, and specifies the causal powers of each type. (c) A generative framework for learning a schema that includes information about the characteristic features of each type. Concrete examples of each framework are shown in Figs. 2b, 2c, and 3.

tion extends the notion of a causal schema by including information about the characteristic features of each causal type (Fig. 1c). Although we begin with cases where at most one object is present at any time, the final section considers cases where multiple objects may be present. We extend the notion of a schema one more time by allowing interactions between different types (for instance, pills of type C may interfere with pills of type D), and we show how these characteristic interactions can be learned.

Learning a single causal model

Suppose that we are interested in the relationship between a single object o and an effect e, and that we have observed a collection of event data V. Each observation in V represents a trial where the object o was either present or absent and the effect e either was or was not observed. For instance, if object o is a pill and effect e is a headache, each trial might indicate whether or not a patient took a pill on a given day, and whether or not she subsequently experienced a headache. To simplify our notation, o will refer both to the pill and to the event of the patient swallowing the pill.

We assume that the outcome of each trial is generated from a causal model M that captures the causal relationship between o and e (Figs. 1a and 2b). Having observed the event data V, our beliefs about the causal model can be summarized by the posterior distribution

$$P(M|V) \propto P(V|M)P(M). \quad (1)$$

We build on the approach of Griffiths and Tenenbaum [5] and parameterize the causal model M using four causal variables (Fig. 2a and 2b). Let a indicate whether there is an arrow joining o and e, and let g indicate the polarity of this causal relationship ($g = 1$ if o is a generative cause and $g = 0$ if o is a preventive cause). Suppose that s is the strength of the relationship between o and e.[1] To capture the possibility that e will be present even though o is absent, we assume that a generative background cause of strength b is always present We specify the distribution $P(e|o)$ by assuming that generative and preventive causes combine according to a network of noisy-OR and noisy-AND-NOT gates.

Now that we have parameterized model M in terms of the triple (a, g, s) and the background strength b, we can rewrite Equation 1 as

$$p(a, g, s, b|V) \propto P(V|a, g, s, b)P(a)P(g)p(s)p(b). \quad (2)$$

To complete our framework we must place prior distributions on the four causal variables. We use uniform priors on the two binary variables (a and g), and assume that the two continuous variables (s and b) represent the logistic transformations of Gaussian variables drawn from conjugate priors.[2] We set the hyperparameters of these conjugate priors to encourage b to be small and s to be large.

Learning multiple causal models

Suppose now that we are interested in a set of objects $\{o_i\}$ and a single effect e. We begin with the case where at most one object is present at any time: for example, suppose that our patient takes many different pills, but at most one per day. Instead of learning a single causal model our goal is to learn a set $\{M_i\}$ of causal models, one for each pill (Figs. 1b and 2c). There is now a triple (a_i, g_i, s_i) describing the causal model for each pill o_i, and we collect these variables into three vectors, $\boldsymbol{a}$, $\boldsymbol{g}$ and $\boldsymbol{s}$. Let Ψ be the tuple $(\boldsymbol{a}, \boldsymbol{g}, \boldsymbol{s}, b)$ which includes all the parameters of the causal models. As for the single object case, we assume that a generative background cause of strength b is always present.

Instead of learning each causal model separately, we introduce the notion of a schema. A schema specifies a grouping of the objects into causal types, and indicates the causal powers of each of these types. The schema in Fig. 2c indicates that there are two causal types: objects of type t_1 tend to prevent the effect, and objects of type t_2 tend to cause the effect. Formally, let z_i indicate the type of o_i, and let $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{g}}$, and $\bar{\boldsymbol{s}}$ be schema-level analogues of $\boldsymbol{a}$, $\boldsymbol{g}$, and $\boldsymbol{s}$: $\bar{a}(t)$ is the probability that any given object of type t will be causally related to the effect, and $\bar{g}(t)$ and $\bar{s}(t)$ are the expected polarity and causal strength for objects of type t. Even though $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{g}}$ are vectors of probabilities, Fig. 2c simplifies by showing each $\bar{a}(t)$ and $\bar{g}(t)$ as a binary variable.

To generate a causal model for each object, we assume that each arrow variable a_i is generated by tossing a coin with weight $\bar{a}(z_i)$, that each polarity g_i is generated by tossing a coin with weight $\bar{g}(z_i)$, and that each strength s_i is drawn from the logistic transform of a Gaussian distribution with mean $\bar{s}(z_i)$ and variance $\bar{\sigma}(z_i)$. Let $\bar{\Psi}$ be the tuple

[1] To simplify the later development of our model, we assume that g and s are defined even if $a = 0$ and there is no causal relationship between o and e. When $a = 0$, g and s can be interpreted as the polarity and strength that the causal relationship between o and e might have had if this relationship actually existed.

[2] For instance, logit(s) is drawn from a Gaussian with mean μ and variance σ^2. σ^2 is drawn from an inverse gamma distribution with shape parameter a and scale parameter b, and μ is drawn from a Gaussian with mean m and variance $v\sigma^2$. We set $v = 10$, $a = 2$ and $b = 0.3$ for all continuous variables. For strength variables, we set $m = 1$, and for the background variable we set $m = -1$.

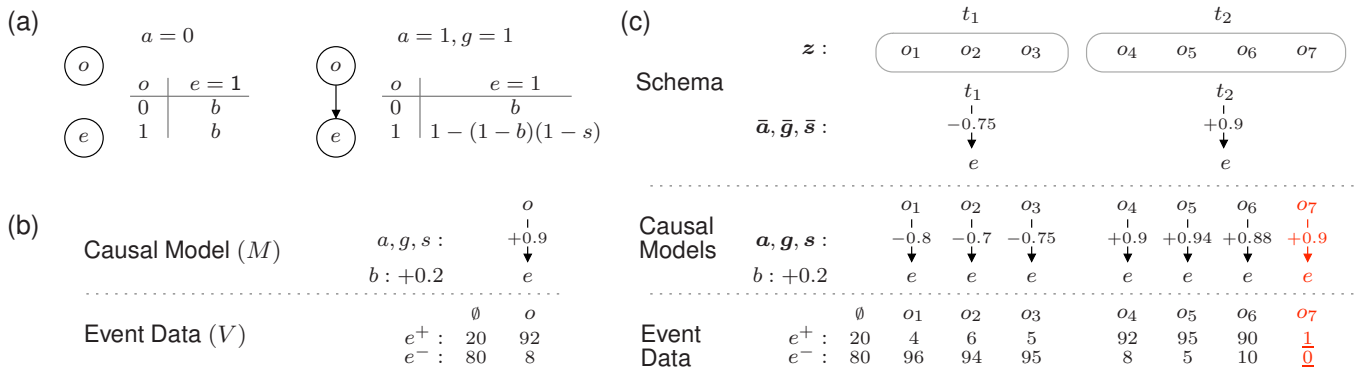

Figure 2: (a) Causal graphical models which capture two possible relationships between an object o and an effect e. a indicates whether there is a causal relationship between o and e, g indicates whether this relationship is generative or preventive, and s indicates the strength of this relationship. A generative background cause of strength b is always present. A third possible model ($a = 1, g = 0$) is not shown. (b) Learning a causal model M from event data V (see Fig. 1a). The event data specify the number of times the effect was (e^+) and was not (e^-) observed when o was absent and when o was present. The model shown has $a = 1$, $g = 1$, $s = 0.9$ and $b = 0.2$, and is an instance of the second model in (a). (c) Learning a schema and a set of causal models (see Fig. 1b). $\boldsymbol{z}$ specifies a set of causal types, where objects belonging to the same type have similar causal powers, and $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{g}}$, and $\bar{\boldsymbol{s}}$ specify the causal powers of each type. Note that the schema supports inferences about an object (o_7) that is very sparsely observed.

$(\bar{\boldsymbol{a}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{\sigma}})$. To complete the model, we specify prior distributions on $\boldsymbol{z}$ and $\bar{\Psi}$. We use a prior $P(\boldsymbol{z})$ that assigns some probability mass to all possible partitions but favors partitions that use a small number of types.[3]

Having defined a generative model, we can use it to learn the type assignments $\boldsymbol{z}$, the schema parameters $\bar{\Psi}$ and the parameters Ψ of the causal models that are most probable given the event data V we have observed:

$$p(\boldsymbol{z}, \bar{\Psi}, \Psi | V) \propto P(V|\Psi)P(\Psi|\bar{\Psi}, \boldsymbol{z})p(\bar{\Psi}|\boldsymbol{z})P(\boldsymbol{z}). \quad (3)$$

Fig. 2c shows how a schema and a set of causal models (top two rows) can be simultaneously learned from the event data V in the bottom row. All of the variables in the figure have been set to values with high posterior probability according to Equation 3: for instance, the partition $\boldsymbol{z}$ shown is the $\boldsymbol{z}$ with maximum posterior probability. Note that learning a schema supports confident inferences about object o_7, which is very sparsely observed (see the underlined entries in Fig. 2c). On its own, a single trial might not be very informative about the causal powers of a novel object, but experience with previous objects allows our model to predict that o_7 will produce the effect as regularly as the other members of type t_2.

To compute the predictions of our model we implemented a Markov chain Monte Carlo scheme that samples from the posterior distribution in Equation 3. Our implementation, however, is not intended as a process model, and our primary contribution is the computational theory summarized by Equation 3.

[3]We use a Chinese Restaurant Process prior on $P(\boldsymbol{z})$, and set the concentration parameter to 1. The entries in $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{g}}$ are independently drawn from a Beta$(0.1, 0.1)$ distribution, and the means and variances in $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{\sigma}}$ are drawn from the conjugate prior already described.

Learning causal types given feature data

Imagine that you are allergic to nuts, and that one day you discover a small white sphere in your breakfast cereal—a macadamia nut, although you do not know it. To discover the causal powers of this novel object you could collect some causal data—you could eat it and wait to see what happens. Probably, however, you will observe the features of the object (its color, shape and texture) and decide to avoid it since it is similar to other allergy-producing foods you have encountered.

Our formal framework naturally handles the idea that instances of a given causal type tend to have similar features (Figs. 1c and 3). Suppose that we have a matrix F which captures many features of the pills in our study, including their sizes, shapes, colors, and imprints. We assume that objects belonging to the same type have similar features. For instance, the schema in Fig. 3 specifies that objects of type t_1 tend to have feature f_7 but not f_8. Formally, let the schema parameters $\bar{\Psi}$ include a matrix $\bar{F}$, where $\bar{f}_j(t)$ specifies the expected value of feature f_j within causal type t.[4] Building on previous models of categorization [1], we assume that the value of f_j for object o_i is generated by tossing a coin with bias $\bar{f}_j(z_i)$. Our goal is now to use the features F along with the event data V to learn a schema and a set of causal models:

$$p(\boldsymbol{z}, \bar{\Psi}, \Psi | V, F) \propto P(V|\Psi)P(F|\bar{\Psi}, \boldsymbol{z})p(\Psi|\bar{\Psi}, \boldsymbol{z})p(\bar{\Psi}|\boldsymbol{z})P(\boldsymbol{z}).$$

There are many previous models for discovering categories of objects with similar features [1], and feature-based categorization is sometimes pitted against causal categorization [3]. Our framework works with the idea that real-world categories are often distinguished both by their characteristic features

[4]The prior on $\bar{F}$ assumes that all entries in this matrix are independent draws from a Beta$(0.5, 0.5)$ distribution.

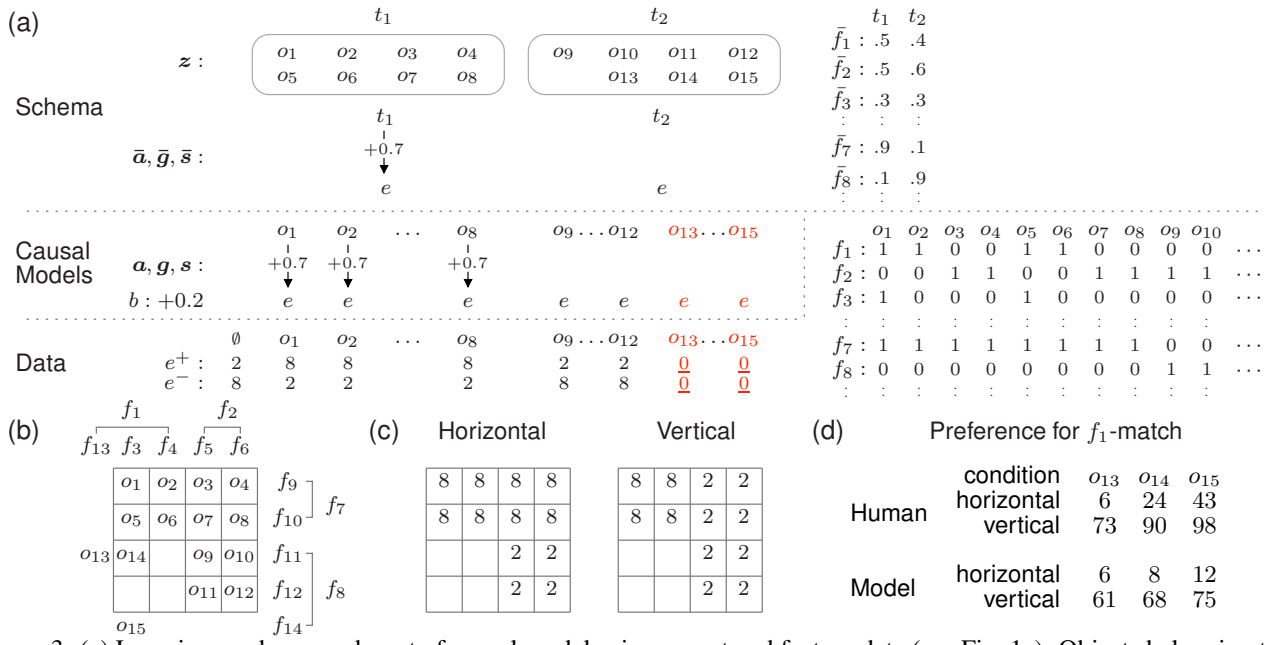

Figure 3: (a) Learning a schema and a set of causal models given event and feature data (see Fig. 1c). Objects belonging to the same type have similar causal powers and similar features, and $\bar{f}_i$ specifies the expected value of feature f_i within each type. The event and feature data shown are for the horizontal condition of the Lien and Cheng experiment. (b) A summary of the feature matrix shown in (a). Feature f_7 is shared by all and only the first eight objects, and f_9 is shared only by the first four objects. (c) Event data for two conditions. 10 trials were shown for each of the first 12 objects. In the horizontal condition, each object with feature f_7 produces the effect on 8 out of 10 trials. In the vertical condition, only objects with f_1 regularly produce the effect. (d) Predictions for the sorting task of Lien and Cheng [9]. The first two rows show the percentage of subjects who grouped a novel object (o_{13}, o_{14} or o_{15}) with the f_1-match (o_1) rather than the f_8-match (o_{10}). Only subjects in the vertical condition tend to sort according to f_1. The model predictions represent the relative probability that each novel object belongs to the same causal type as the f_1-match.

and their characteristic causal interactions. More often than not, one kind of information will support the categories indicated by the other, but there will also be cases where the causal data and the feature data conflict. In cases like this, our model may discover the feature-based categories, the causal categories, or some combination of both—the categories preferred will depend on the relative weights of the statistical information present in the two kinds of data.

Behavioral data

Lien and Cheng [9] ran several experiments that explore how perceptual features and causal observations can both inform causal judgments. Our framework can handle all of their tasks, but we focus here on a simplified version of their first task. The effect of interest is whether a certain kind of plant blooms, and the potential causes are 15 chemicals (objects o_1 through o_{15}). Fig. 3b shows that the features of these objects (f_1 through f_{14}) support two systems of categorization. The first is based on color: each object has a cool color (f_7) or a warm color (f_8), and the warm-colored objects are either yellow (f_{11}), red (f_{12}) or orange (f_{14}). Similarly, each object has an irregular shape (f_1) or a regular shape (f_2) and there are three kinds of irregular shapes (f_{13}, f_3 and f_4).

We show our model 10 trials for each of the first 12 objects, and Fig. 3c summarizes the results of these trials. In the *horizontal* condition, each object with a cool color (f_7) causes blooming on 8 out of 10 occasions, and the remaining objects lead to blooming less often. In the *vertical* condition, objects with irregular shapes (f_1) are the only ones that tend to cause blooming. In both conditions, the model is shown that blooming occurs on 2 out of 10 trials when no chemicals are applied.

We test our model by requiring it to reason about three objects (o_{13}, o_{14} and o_{15}) for which no trials were observed (see the underlined entries in Fig. 3a). Object o_{13} has a novel shape, o_{15} has a novel color, and o_{14} is a novel combination of a known shape and known color. Each novel object was presented as part of a trio that also included o_1 and o_{10}, and we computed whether the model preferred to group each novel object with the shape match (o_1) or the color match (o_{10}).[5] In the horizontal condition, the model prefers to sort each trio according to color (f_8), but in the vertical condition the model sorts each trio according to shape (f_1) (see Fig. 3d). Note that the feature data and the causal data must be combined to produce this result: a model that relied on the features alone would predict no difference between the two conditions, and

[5] We implemented this sorting task by computing the posterior distribution $p(\boldsymbol{z}|V, F)$, and comparing the probability that the novel object and its color match belong to the same causal type with the probability that the novel object is grouped with the shape match.

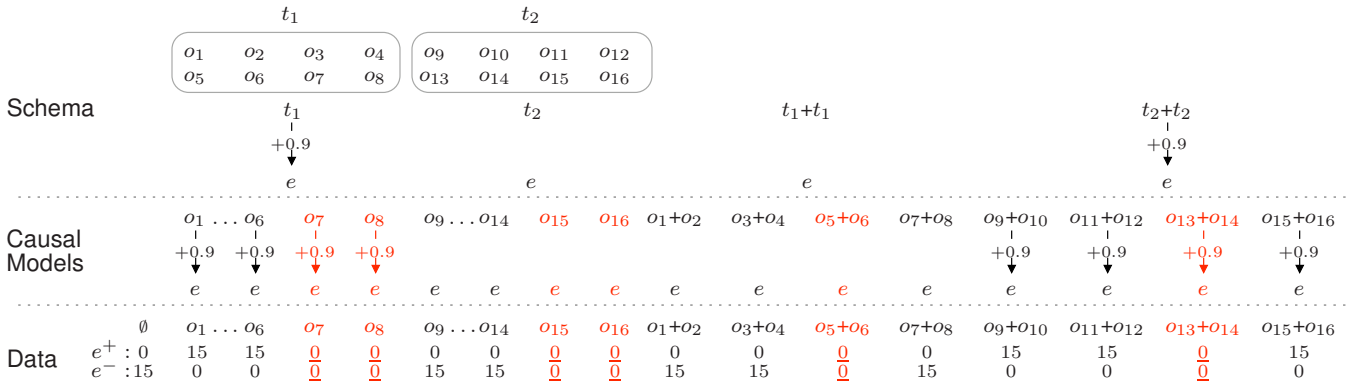

Figure 4: Learning about interactions between objects. The schema specifies the causal powers of each type and of each combination of types (the combination t_1+t_2) is not shown. The collection of causal models includes a model for each combination of objects. The event data are inspired by the experiment of Shanks and Darby [13]. The model groups the objects into two types: objects belonging to type t_1 cause the effect on their own but not when paired with each other, and objects belonging to the type t_2 cause the effect only when paired with each other.

a model that used only the causal data would be unable to make useful predictions about the three novel objects. Since we have modeled a simplified version of the Lien and Cheng task, the quantitative predictions of our model are not directly comparable to their results, but Fig. 3d shows that our model captures the main qualitative patterns in their data.[6]

Discovering interactions between causal types

So far we have considered problems where at most one object o_i can be present at a time. Suppose now that multiple objects can be present on any trial. For instance, consider the problem of discovering which drugs produce a certain allergy—two drugs which are innocuous on their own may produce the allergy when combined. Our goal is to discover a schema and a set of causal models that allow us to predict whether any given combination of drugs is likely to produce an allergic reaction. Formally, we would like to learn a causal model M for each possible combination of objects.

We assume that each combination of objects corresponds to a conjunctive cause that may be generative or preventive, and extend Ψ to include an arrow a, a polarity g and a strength s for each combination of objects. We extend the schema in a similar fashion, and include schema parameters $\bar{a}$, $\bar{g}$, $\bar{s}$ and $\bar{\sigma}$ for each combination of causal types. The causal model parameters for sets of objects are generated, as before, from the schema parameters for the corresponding set of types. For instance, Fig. 4 shows how the causal model for o_{13}+o_{14} is generated from the schema-level knowledge that pairs of objects drawn from type t_2 tend, in combination, to generate the effect with strength 0.9. As before, we assume that a generative background cause of strength b is always present.

There are several possible strategies for handling conjunctive causes and our approach makes several simplifying assumptions. For instance, we assume that the causal power of a conjunction of objects is independent of the causal powers that correspond to any subset of these objects. To accurately capture human intuitions, it will be necessary to relax our simplifying assumptions, and to combine our framework with a sophisticated approach to conjunctive causality [11]. Here, however, we have aimed to provide the simplest possible example of how our framework can discover interactions between causal types.

Behavioral data

Shanks and Darby [13] ran an experiment which suggests that humans can acquire abstract knowledge about interactions between causal types. These authors used a task where the potential causes were foods, and the effect of interest was an allergic reaction. The data observed by participants in their second experiment are shown in Fig. 4.[7] When supplied with these data, our model discovers two causal types: foods of type t_1 (o_1 through o_8) produce the allergy on their own, but foods of type t_2 (o_9 through o_{16}) do not. The model also discovers that two foods of type t_2 will produce the allergy when eaten together, but two foods of type t_1 will not (Fig. 4).

Shanks and Darby were primarily interested in predictions about cases which had never been observed in training—the cases underlined in Fig. 4. Their participants can be divided into two groups according to their scores when tested on the training data. Learners in the high group (learners who scored well on the test) tended to make the same predictions as our model: for instance, they tended to predict that o_7 and o_8 produce the allergy when eaten in isolation, that o_{15} and o_{16} do not, that the combination of o_{13} and o_{14} produces the allergy, and that the combination of o_5 and o_6 does not. Learners in the low group tended to make the opposite predictions: for instance, they tended to predict that o_7 and o_8 do not produce the allergy when eaten in isolation. Since our compu-

[6] Lien and Cheng report that a handful of subjects did not group the novel objects with either the shape match or the color match. These subjects were dropped before computing the percentages in Fig. 3d.

[7] Different subjects saw different amounts of training data, but we overlook this detail.

tational framework does not suffer from memory limitations or lapses of attention, it is not surprising that it accounts only for the predictions of learners who absorbed the information provided during training.

Discussion

We described a hierarchical Bayesian framework (Fig. 1c) for learning causal schemata. Our hierarchical framework supports several kinds of inferences. We focused on bottom-up learning and showed that the model helps to explain how a causal schema and a set of specific causal models can be simultaneously learned given event data and feature data. If the causal schema is known in advance, then the framework serves as a computational theory of top-down causal learning, and explains how inferences about a set of causal models can simultaneously draw on low-level event data and top-down knowledge.

Our work exploits the fact that probabilistic approaches are modular and can be composed to build integrated models of inductive reasoning. The model in Fig. 1c can be created by combining three models: probabilistic causal models [12] specify how the event data are generated given a set of causal models, the infinite relational model [8] specifies how the causal models are generated, and Anderson's rational approach to categorization [1] specifies how the features are generated. Since all three models work with probabilities it is straightforward to combine them to create a single integrated framework for causal reasoning.

We showed that our framework helps to explain some aspects of the data collected by Lien and Cheng [9] and Shanks and Darby [13], and it also accounts for several other results in the literature. Waldmann and Hagmayer [16] showed that a known set of categories can influence future causal learning, and our approach predicts a similar result if we fix the causal types z then use our framework to discover a set of causal models given event data. Our framework can also model experiments carried out using the blicket detector [3] or causal blocks world [15] paradigms. Many aspects of these experiments have been previously modeled, but our framework captures phenomena that are not addressed by most existing models. For instance, our model suggests why two identical looking blocks might both be categorized as blickets even though a handful of observations suggest that they have different effects on a blicket detector [3].

Several extensions of our approach may be worth exploring. We restricted ourselves to problems where the distinction between a set of potential causes and a set of effects [8] is known in advance, but in some cases this distinction might need to be learned [10]. A second limitation is that we focused on cases where feature data and contingency data represent the only input to our model. Human learners are sometimes directly supplied with abstract causal knowledge—for example, a science student might be told that "pineapple juice is an acid, and acids turn litmus paper red." Statements like these correspond to fragments of a causal schema, and future experiments should explore how schemata are learned when parts of these schemata are directly supplied.

More often than not, competing accounts of a given phenomenon both capture some element of the truth. Where possible, cases like these should be handled by building unified accounts that subsume the two competing views. We have developed a hierarchical Bayesian model that attempts to unify top-down and bottom-up approaches to causal reasoning. Similar conflicts between top-down and bottom-up approaches are found in other areas of cognitive science, and the hierarchical Bayesian approach may be useful for resolving these conflicts wherever they occur.

Acknowledgments Supported by the William Asbjornsen Albert memorial fellowship (CK), the James S. McDonnell Foundation Causal Learning Collaborative Initiative (NDG, JBT) and the Paul E. Newton chair (JBT).

[8] This paper has focused on problems where there is a single effect, but our approach also handles problems with multiple effects.

References

[1] Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

[2] Cheng, P. W. (1993). Separating causal laws from casual facts: Pressing the limits of statistical relevance. In *The psychology of learning and motivation*, volume 30, pages 215–264. Academic Press, San Diego.

[3] Gopnik, A. and Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, 71:1205–1222.

[4] Griffiths, T. L. (2005). *Causes, coincidences, and theories*. PhD thesis, Stanford University.

[5] Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51:354–384.

[6] Kelley, H. H. (1972). Causal schemata and the attribution process. In Jones, E. E., Kanouse, D. E., Kelley, H. H., Nisbett, R. S., Valins, S., and Weiner, B., editors, *Attribution: Perceiving the causes of behavior*, pages 151–174, Morristown, NJ. General Learning Press.

[7] Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28:107–128.

[8] Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI 06*.

[9] Lien, Y. and Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, 40:87–137.

[10] Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., and Griffiths, T. L. (2006). Structured priors for structure learning. In *UAI 06*.

[11] Novick, L. R. and Cheng, P. W. (2004). Assessing interactive causal inference. *Psychological Review*, 111:455–485.

[12] Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press, Cambridge, UK.

[13] Shanks, D. R. and Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 24(4):405–415.

[14] Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47(1):1–51.

[15] Tenenbaum, J. B. and Niyogi, S. (2003). Learning causal laws. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*.

[16] Waldmann, M. R. and Hagmayer, Y. (2006). Categories and causality: the neglected direction. *Cognitive Psychology*, 53:27–58.