Computational Cognitive Modeling
Probabilistic graphical models

Brenden Lake & Todd Gureckis

email address for instructors:
instructors-ccm-spring2020@nyucll.org

course website:
https://brendenlake.github.io/CCM-site/
Probabilistic graphical models

In this section, we will cover:

- The basic technical concepts behind probabilistic graphical models and how to work with them.
- Applications in computational cognitive modeling, including problems in classification, causal learning, and structure discovery.
Bayesian networks ("Bayes net")

- Bayesian network: a directed graph that represents dependencies between random variables, giving a concise specification of a joint probability distribution.

- In a well-constructed network, an arrow indicates that two variables have a path of direct (causal) influence.

- Bayesian networks must be directed, acyclic graphs (DAGs), meaning that they have no cycles.

Factorization of the joint distribution:

\[
P(x_1, x_2, x_3) = P(x_1)P(x_2)P(x_3|x_1, x_2)
\]
Bayesian networks

\[ P(x_1, \ldots, x_7) = P(x_1)P(x_2)P(x_3)P(x_4|x_1, x_2, x_3) \]
\[ P(x_5|x_1, x_3)P(x_6|x_4)P(x_7|x_4, x_5) \]

General formula for factorizing the joint distribution over a Bayes net:

\[ P(X) = \prod_{i=1}^{K} P(x_i|\text{Parents}(x_i)) \]

Slide credit: Christopher Bishop
An example: the alarm network

\[ P(B, E, A, J, M) = P(B)P(E)P(A|B, E)P(J|A)P(M|A) \]

(particular version from Russell and Norvig)
Evaluating the joint probability of data

We use the decomposed joint distribution to evaluate the probability of a setting of all of the variables.

\[ P(B, E, A, J, M) = P(B)P(E)P(A|B, E)P(J|A)P(M|A) \]

What is the probability that there is no burglary or earthquake, and yet the alarm rings and both John and Mary call?

\[ P(B = 0, E = 0, A = 1, J = 1, M = 1) \]
\[ = P(B = 0)P(E = 0)P(A = 1|B = 0, E = 0)P(J = 1|A = 1)P(M = 1|A = 1) \]
\[ = 0.999 \times 0.998 \times 0.001 \times 0.9 \times 0.7 = 0.00063 \]
Causal mechanisms are important in everyday categorization and reasoning. Is useful to think of the function $f$ (the representation for a specific category) as a Bayesian network?
A Causal-Model Theory of Conceptual Representation and Categorization

Bob Rehder
New York University

This article presents a theory of categorization that accounts for the effects of causal knowledge that relates the features of categories. According to causal-model theory, people explicitly represent the probabilistic causal mechanisms that link category features and classify objects by evaluating whether they were likely to have been generated by those mechanisms. In 3 experiments, participants were taught causal knowledge that related the features of a novel category. Causal-model theory provided a good quantitative account of the effect of this knowledge on the importance of both individual features and interfeature correlations to classification. By enabling precise model fits and interpretable parameter estimates, causal-model theory helps place the theory-based approach to conceptual representation on equal footing with the well-known similarity-based approaches.

For the last several decades, research on the topic of categorization has focused on the problem of learning new categories via examples of category members, that is, from empirical observations. The result has been a host of categorization models that are based on representational ideas such as central prototypes, stored exemplars, and variabilized rules, and on processing principles such as similarity, that have considerable explanatory power and experimental support. More recently, the influence of the prior “theoretical” knowledge that learners often contribute to their representations of categories has also been a topic of study (Carey, 1985; Keil, 1989; Murphy & Medin, 1985; Schank, Collins, & Hunter, 1986). For example, people not only know that birds have wings and that they can fly and build nests in trees, but also that birds build nests in trees because they can fly, and fly because they have wings. Many people even believe that morphological features of birds such as wings are ultimately caused by the kind of DNA that birds possess. However, in comparison with the development of models accounting for the effects of empirical observations, there has been relatively little development of formal models to account for the effects of such prior knowledge (although see Heit, 1994; Heit & Bott, 2000; Pazzani, 1991; Rehder & Murphy, in press; Sloman, Love, & Ahn, 1998).

The purpose of this article is to present a theory of categorization that accounts for the effects of causal knowledge that relates the features of categories. According to causal-model theory, people explicitly represent the probabilistic causal mechanisms that link category features and classify objects by evaluating whether they were likely to have been generated by those mechanisms. In 3 experiments, participants were taught causal knowledge that related the features of a novel category. Causal-model theory provided a good quantitative account of the effect of this knowledge on the importance of both individual features and interfeature correlations to classification. By enabling precise model fits and interpretable parameter estimates, causal-model theory helps place the theory-based approach to conceptual representation on equal footing with the well-known similarity-based approaches.

Further, according to this theory, people use causal models to determine a new object's category membership.

In this article, causal-model theory is applied to two outstanding problems in the domain of categorization research. The first problem concerns determining the importance, or weight, that individual features have on establishing category membership. Since the popularization of the notion of probabilistic categories in the 1970s, it has usually been assumed that features of a category vary regarding their influence on category membership (Hampton, 1979; Rosch, 1973; Rosch & Mervis, 1975; Smith & Medin, 1981). Indeed, formal models of categorization have formalized the manner in which a feature’s weight is influenced by its perceptual saliency (Lamberts, 1995, 1998) and by the frequency with which it appears in category members and nonmembers (Nosofsky, 1986; Reed, 1972; Rosch & Mervis, 1975; Shepard, Hovland, & Jenkins, 1961). However, these models do not account for the fact that feature weights are also determined by a categorizer’s domain theories. For instance, Medin and Shoben (1988) have found that straight bananas are rated as better members of the category bananas than straight boomerangs are of the category boomerangs, a result they attribute to the default feature assignment that boomerangs have of being curved.
Artificial categorization task

Task: Learn and make predictions about a new category, e.g., “Lake Victoria Shrimp”

Four binary features

F₁: High amounts of ACh neurotransmitter.
F₂: Long-lasting flight response.
F₃: Accelerated sleep cycle.
F₄: High body weight.

Base rate information: 75% of Lake Victoria Shrimp have each feature, e.g., 75% have feature F₄
Artificial categorization task
participants assigned to one of two conditions

chain (causal framing) condition

\[ F_1 \overset{m}{\rightarrow} F_2 \overset{m}{\rightarrow} F_3 \overset{m}{\rightarrow} F_4 \]

free parameters \( c, b \) (background mechanism) and \( m \) (causal strength)

**F1** → **F2**
A high quantity of the ACh neurotransmitter causes a long-lasting flight response. The duration of the electrical signal to the muscles is longer because of the excess amount of neurotransmitter.

**F2** → **F3**
A long-lasting flight response causes an accelerated sleep cycle. The long-lasting flight response causes the muscles to be fatigued, and this fatigue triggers the shrimp’s sleep center.

**F3** → **F4**
An accelerated sleep cycle causes a high body weight. Shrimp habitually feed after waking, and shrimp on an accelerated sleep cycle wake three times a day instead of once.

control condition (no causal framing)

\[ F_1 \quad F_2 \quad F_3 \quad F_4 \]

exactly the same instructions, but without causal information between features

\( F_1: \) High amounts of ACh neurotransmitter.
\( F_2: \) Long-lasting flight response.
\( F_3: \) Accelerated sleep cycle.
\( F_4: \) High body weight.
Artificial categorization task: Results

Conclusions: Causal/structural information influences people’s categorization decisions, in a way predicted by a Bayesian network model.

**Test judgments:** is $F$ a Lake Victoria Shrimp?

In causal condition, compute judgement as:

$$P(F_1, F_2, F_3, F_4) = P(F_1)P(F_2|F_1)P(F_3|F_2)P(F_4|F_3)$$

Key idea: categorization decision is computing joint probability under a Bayes net model of that category.
Causal structure matters in categorization judgments

Further work from Rehder and colleagues have studied categories with these alternative Bayes net structures…
(e.g., Rehder and Hastie, 2001)
How do Bayesian networks relate to other Bayesian models used in cognitive modeling?

\[
P(x_1, \ldots, x_7) = P(x_1)P(x_2)P(x_3)P(x_4|x_1, x_2, x_3) \\
P(x_5|x_1, x_3)P(x_6|x_4)P(x_7|x_4, x_5)
\]

General formula for factorizing the joint distribution

\[
P(X) = \prod_{i=1}^{K} P(x_i | \text{Parents}(x_i))
\]

The number game

Perceptual magnet effect

The speaker makes an intended sound production \(T\). Noise in the air perturbs \(T\) into \(S\). The listener calculates the posterior \(P(T|S)\).
Connection with simple Bayesian models

Many of the Bayesian models developed for cognitive modeling can be interpreted as two node Bayesian networks, with a complex (potentially very complex) conditional probability table (aka likelihood function).

Diagnosis example

Data ($D$): John is coughing

Hypotheses:

\[ h_1 = \text{John has a cold} \]
\[ h_2 = \text{John has emphysema} \]
\[ h_3 = \text{John has a stomach flu} \]

The number game

16 8 2 64

Perceptual magnet effect

$P(d, h) = P(d|h)P(h)$
Review from the number game: Probabilistic inference is very flexible!

$$E[\phi(h)|D] \approx \frac{1}{M} \sum_{m} \phi(h^{(m)})$$

If we can compute the posterior, or draw samples from the posterior, we can automatically reason about a huge range of questions $\phi(\cdot)$.

**Examples of reusing the sample for new queries**

- Is 64 a member of the set? *(probability is 0.73)*

- Are both 36 and 64 members of the set? *(0.36)*

- Is there a member of the set greater than or equal to 80? *(0.27)*

- If we sample a new number from the hypothesis, what is the chance it will be 64? *(0.16)*

- If we sample a new number from the hypothesis, what is the chance it will be 80? *(0.02)*

The type of flexibility in reasoning is natural in Bayesian models, but it is difficult to capture in neural networks trained with supervised learning, or model-free reinforcement learning.

Inference flexibility is not specific to rejection sampling, but to Bayesian models in general.
Probabilistic inference
as a generalization of Bayes’ rule to arbitrary queries in a probabilistic model

General formula for probabilistic inference

\[ P(X|e) \propto \sum_y P(X, e, y) \]

\[ P(X|e) = \frac{\sum_y P(X, e, y)}{\sum_{y, X'} P(X', e, y)} \]

\( X = \) query variables
\( e = \) evidence variables
\( Y = \) hidden variables

Example with the alarm network:

Probability of a burglary given that Mary calls:

\[ P(B = 1|M = 1) = 0.056 \]

\( X = \{B\} \)
\( e = \{M\} \)
\( Y = \{E, A, J\} \)
Flexible reasoning through probabilistic inference

as a generalization of Bayes’ rule to arbitrary queries in a probabilistic model

**Marginal distributions**

\[
P(B = 1) = 0.001 \\
P(A = 1) = 0.003 \\
P(M = 1) = 0.01
\]

Is there a burglary, given John and Mary both call?

\[
P(B = 1|J = 1, M = 1) = 0.284
\]

Is there an earthquake, given John and Mary both call?

\[
P(E = 1|J = 1, M = 1) = 0.176
\]

Algorithms for inference:

- exact enumeration (equation from previous slide)
- rejection sampling
- importance sampling
- MCMC
- etc.

**Does Mary call, given a burglary?**

\[
P(M = 1|B = 1) = 0.66
\]

**Does Mary call, given a burglary and earthquake?**

\[
P(M = 1|B = 1, E = 1) = 0.67
\]

What is the chance there is a burglary with an alarm and Mary calls, assuming no earthquake?

\[
P(A = 1, B = 1, M = 1|E = 0) = 0.0006
\]
“Explaining away” with Bayes nets

Where one cause can explain away the need for a second one…

If alarm rings, then an earthquake becomes a real possibility:

$$P(E = 1|A = 1) = 0.23$$

Unless we know \textit{there was a burglary}, in which case we can “explain away” the indication for an earthquake.

$$P(E = 1|A = 1, B = 1) = 0.002$$
Probabilistic inference in practice

Pathfinder project for medical diagnosis (Heckerman, Horvitz, & colleagues; late 1980s)

- Commercial system for diagnosing lymph-node pathology
- Probabilistic inference used to compute $P(\text{Disease}|\text{Symptoms})$
- CPTs determined by expert knowledge from pathologists
Example of probabilistic inference: Interactive activation model

**recurrent neural network**

- name
- age
- occupation
- marital status
- gang
- hidden / instance

**Bayesian network alternative**

- hidden / instance
- noisy links

- Retrieval by name \[ P(X|\text{name} = \text{Ken}) \]
- Content addressability \[ P(\text{name}|\text{age} = \text{30s}, \text{gang} = \text{Sharks}) \]
- Spontaneous generalization \[ P(\text{age}|\text{gang} = \text{Sharks}) \]
Conditional independence

\( x_1 \) is independent of \( x_2 \) given \( x_3 \)

\[
P(x_1 | x_2, x_3) = P(x_1 | x_3)
\]

Equivalently

\[
P(x_1, x_2 | x_3) = P(x_1 | x_2, x_3)P(x_2 | x_3) \quad \text{(product rule)}
\]

\[
= P(x_1 | x_3)P(x_2 | x_3)
\]

Written as

\[
\begin{align*}
  x_1 & \perp \! \! \! \! \! \perp x_2 \mid x_3
\end{align*}
\]

Slide credit: Christopher Bishop
Conditional independence: Example

\[ P(J, M | A) = \frac{P(J, A, M)}{P(A)} \]  

(def. condition prob)

\[ = \frac{P(J | A)P(M | A)P(A)}{P(A)} \]

\[ = P(J | A)P(M | A) \]

\[ J \perp M | A \]
Conditional independence: Example

\[ P(J, A, M) = P(J|A)P(M|A)P(A) \]

\[ P(J, M) = \sum_A P(J|A)P(M|A)P(A) \]

\[ J \perp M \]
Conditional independence: More examples

- Burglary → Alarm → JohnCalls
  - $B \perp J$

- Burglary → Alarm → JohnCalls
  - $B \perp J \mid A$

- Burglary → Earthquake → Alarm
  - $B \perp E$

- Burglary → Earthquake → Alarm
  - $B \perp E \mid A$

“explaining away” case
General statement on conditional independence

A variable is conditionally independent of its non-descendants given its parents.
Significance of Bayes nets and conditional independence

- We can read conditional independence properties directly off the graph structure, rather than having to derive them analytically (as we did with simple examples of conditional independence).

- We can exploit the conditional independence properties for efficient probabilistic inference / Bayesian reasoning (using exact inference, MCMC, etc.)
Learning Bayesian networks: Parameter learning

Known structure (e.g., consulting experts, prior knowledge), but unknown parameters

\[ P(B, E, A, J, M) = P(B)P(E)P(A|B, E)P(J|A)P(M|A) \]

maximum likelihood parameter learning:

\[ \arg \max_\theta \sum_i \log P(D^{(i)}|\theta; S) \]

straightforward solution: we can fit CPTs independently, and each CPT is very intuitive (simply count the relevant occurrences of a variable given its parents)
Learning Bayesian networks: Structure learning

Unknown structure, unknown parameters

$$\arg \max_{\theta, S} \sum_i \log P(D^{(i)}|\theta, S) - \text{cost}(S)$$

- Structure learning is much more difficult computationally than parameter learning.
- The objective function includes some type of regularization to favor simple graphs (BIC, AIC, etc.).
- Finding the optimal graph structure $S$ often involves a huge combinatorial search problem over structures, and we need to be careful not to introduce cycles.
- We usually have to resort to heuristic search methods (such as greedy proposal for adding, removing, or switching the direction of edges).
- Data can include both observations and (optionally) interventions.

**example proposal to add an edge**
Learning Bayesian networks: Structure learning

We can also search over “node orders”, where network structure is determined by ordering.

These networks can represent the same probability distribution as the original alarm network, but are much clumsier and require more parameters (node order is indicated by height on the slide).

If we get the causal structure wrong, we can easily overfit when learning the parameters and make bad inferences.
A Theory of Causal Learning in Children: Causal Maps and Bayes Nets

Alison Gopnik
University of California, Berkeley

Clark Glymour
Carnegie Mellon University and
Institute for Human and Machine Cognition

David M. Sobel
Brown University

Laura E. Schulz and Tamar Kushnir
University of California, Berkeley

The authors outline a cognitive and computational account of causal learning in children. They propose that children use specialized cognitive systems that allow them to recover an accurate “causal map” of the world: an abstract, coherent, learned representation of the causal relations among events. This kind of knowledge can be perspicuously understood in terms of the formalism of directed graphical causal models, or Bayes nets. Children’s causal learning and inference may involve computations similar to those for learning causal Bayes nets and for predicting with them. Experimental results suggest that 2- to 4-year-old children construct new causal maps and that their learning is consistent with the Bayes net formalism.

The input that reaches children from the world is concrete, particular, and limited. Yet, adults have abstract, coherent, and largely veridical representations of the world. The great epistemological question of cognitive development is how human beings get from one place to the other: How do children learn so much about the world so quickly and effortlessly? In the past 30 years, cognitive developmentalists have demonstrated that there are systematic changes in children’s knowledge of the world. However, psychologists know much less about the representations that underlie that knowledge and the learning mechanisms that underlie changes in that knowledge.

In this article, we outline one type of representation and several related types of learning mechanisms that may play a particularly important role in cognitive development. The representations are of the causal structure of the world, and the learning mechanisms involve a particularly powerful type of causal inference. Causal knowledge is important for several reasons. Knowing about causal structure permits us to make wide-ranging predictions about future events. Even more important, knowing about causal structure allows us to intervene in the world to bring about new events—often events that are far removed from the interventions themselves.
The results of these experiments rule out many possible hypotheses about children’s causal learning. Because children did not activate the detector themselves, they could not have solved these tasks through operant conditioning or through trial-and-error learning. The blickets and nonblickets were perceptually indistinguishable, and both blocks were in contact with the detector, so children could not have solved the tasks through their substantive prior knowledge about everyday physics.

The “make it stop” condition in this experiment also showed that children’s inferences went beyond classical conditioning, simple association, or simple imitative learning. Children not only associated the word and the effect, they combined their prior causal knowledge and the new causal knowledge they inferred from the dependencies to create a brand-new intervention that they had never witnessed before. As we mentioned above, this kind of novel intervention is the hallmark of a causal map. It is interesting that there is, to our knowledge, no equivalent of this result in the vast animal conditioning literature, although such an experiment would be easy to design. Would Pavlov’s dogs, for example, intervene to silence a bell that led to shock, if they had simply experienced an association between the bell and the shock but had never intervened in this way before?

In all these respects, children seemed to have learned a new causal map. Moreover, this experiment showed that children were not using simple frequencies to determine the causal structure of this map but were using more complex patterns of conditional dependence. However, this experiment was consistent with all four learning models we described above, including the causal interpretation of the RW model.

Inference from indirect evidence: Backward blocking.

In the next study we wanted to see whether children’s reasoning would extend to even more complex types of conditional dependence and, in particular, if children would reason in ways that went beyond causal RW. There are a number of experimental results that argue against the RW model for adult human causal learning. One such phenomenon is “backward blocking” (Shanks, 1985; Shanks & Dickinson, 1987; Wasserman & Berglan, 1998). In backward blocking, learners decide whether an object causes an effect by using information from trials in which that object never appears.

Sobel and colleagues (Sobel, Tenenbaum, & Gopnik, in press) have demonstrated backward blocking empirically in young children. In one experiment (Sobel et al., in press, Experiment 2), 3- and 4-year-olds were introduced to the blicket detector in the same manner as in the Gopnik et al. (2001) experiments. They were told that some blocks were blickets and that blickets make the machine go. In a pretest, children saw that some blocks, but not others, made the machine go, and the active objects were labeled as blickets. Then children were shown two new blocks (A and B). In one condition, the control, inference condition, A and B were placed on the detector together twice, and the detector responded both times. Then children observed that Object A did not activate the detector by itself. In the other condition, the backward blocking condition, children saw that two new blocks, A and B, activated the detector together twice. Then they observed that Block A did activate the detector by itself. In both conditions, children were then asked whether each block was a blicket and were asked to make the machine go (see Figure 13).

Results: 75% of 3-4 year olds remove Object A from the detector (Gopnik et al., 2001)
3-4 year old children categorized **Object B as a blicket only 31% of the time** in backward blocking condition, but 100% of the time in control (Sobel et al., 2004)
Four hypotheses for Bayesian structure learning
(prior favors fewer edges)

- A → D
- B → D
- A
- B

Neural network representation

Neural nets only account for decrease to baseline with particular input encoding schemes, where it comes naturally from structure learning.

Adult vs. model judgements

Key effect: reduction to baseline

(Tenenbaum & Griffiths, 2003)
The discovery of structural form
Charles Kemp* and Joshua B. Tenenbaum†

*Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213; and †Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 30, 2008 (received for review March 17, 2008)

Algorithms for finding structure in data have become increasingly important both as tools for scientific data analysis and as models of human learning, yet they suffer from a critical limitation. Scientists discover qualitatively new forms of structure in observed data: For instance, Linnaeus recognized the hierarchical organization of biological species, and Mendeleev recognized the periodic structure of the chemical elements. Analogous insights play a pivotal role in cognitive development: Children discover that object category labels can be organized into hierarchies, friendship networks are organized into cliques, and comparative relations (e.g., “bigger than” or “better than”) respect a transitive order. Standard algorithms, however, can only learn structures of a single form that must be specified in advance: For instance, algorithms for hierarchical clustering create tree structures, whereas algorithms for dimensionality-reduction create low-dimensional spaces. Here, we present a computational model that learns structures of many different forms and that discovers which form is best for a given dataset. The model makes probabilistic inferences over a space of graph grammars representing trees, linear orders, multidimensional spaces, rings, dominance hierarchies, cliques, and other forms and successfully discovers the underlying structure of a variety of physical, biological, and social domains. Our approach brings structure learning methods closer to human abilities and may lead to a deeper computational understanding of cognitive developmental challenges.

Humans are adept at making inferences that take them beyond the limits of their direct experience. Even young children can learn the meaning of a novel word from a single labeled example (Heibbeck & Markman, 1987), predict the trajectory of a moving object when it passes behind an occluder (Spelke, 1990), and choose a gait that allows them to walk over terrain they have never before encountered. Inferences like these may differ in many respects, but common to them all is the need to go beyond the information given (Bruner, 1973).

This article describes a formal approach to inductive inference that should apply to many different problems, but we focus on the problem of property induction (Sloman & Lagnado, 2005). In particular, we consider cases where one or more categories in a domain are observed to have a novel property and the inductive task is to predict how the property is distributed over the remaining categories in the domain. For instance, given that bears have sesamoid bones, which species is more likely to share this property: moose or salmon (Osherson, Smith, Wilkie, Lopez, & Shafir, 1986).
Review: A neural network model of semantic cognition

- Network is trained to answer queries involving an item (e.g., “Canary”) and a relation (e.g., “CAN”), outputting all attributes that are true of the item/relation pair (e.g., “grow, move, fly, sing”)

- Trained with stochastic gradient descent, as we learned about in this lecture

- The model helps us to understand the broad-to-specific pattern of differentiation in children’s cognitive development

- It also helps us to understand the specific-to-general deterioration in semantic dementia
Alternative: Property induction as probabilistic inference in a probabilistic graphical model

Question: “Given that cows and seals have T9 hormones, how likely is it that horses do?”

property induction as probabilistic inference:

\[ P(f_Y = 1 | f_X = 1) \]

\[ f : \text{T9 hormones} \]
\[ Y = \{\text{horses}\} \]
\[ X = \{\text{cows, seals}\} \]

Bayesian modeling roadmap

Biological reasoning about animals
A tree fits better than a 2D space

Cows have property P.
Elephants have property P.
Horses have property P.

Gorillas have property P.
Mice have property P.
Seals have property P.

All mammals have property P.

Correlation of human participant judgments with model judgments

Evaluates across a range of difference premise/conclusion combinations.

slide credit: Josh Tenenbaum
Spatial reasoning about cities
A 2D space fits better than a tree

“Given that a certain kind of native American artifact has been found in sites near city X, how likely is the same artifact to be found near city Y?”

slide credit: Josh Tenenbaum
Learning structural forms
How do we know what the right form is?

People can discover intuitive structural forms:

Famous examples in Science

Linnaeus
Kingdom Animalia
Phylum Chordata
Class Mammalia
Order Primates
Family Hominidae
Genus Homo
Species *Homo sapiens*

Darwin

Mendeleev

Examples from childhood
• e.g., days of the week form a cycle, social networks are cliques, comparative relations are transitive, names can be organized in taxonomies

slide credit: Josh Tenenbaum
Learning structural forms
Kemp & Tenenbaum (2008). The discovery of structural form. *PNAS.*
Bayesian structural forms model
Kemp & Tenenbaum (2008)

Form $F$
- Partition
- Chain
- Order
- Ring
- Hierarchy
- Tree
- Grid
- Cylinder

Structure $S$
- gorilla
- dolphin
- seal
- chimp
- mouse
- squirrel
- elephant
- rhino

Data $f^{(k)}$
- Horse
- Cow
- Chimp
- Gorilla
- Mouse
- Squirrel
- Dolphin
- Seal
- Rhino
- Elephant

Key
- observed variable (object)
- latent variable (Markov random field)
Very brief intro. to undirected graphical models

- Also known as Markov Random Fields or Markov Networks
- Bayesian networks are better suited for representing causal processes, while Markov networks are better suited for capturing soft constraints between variables

In the structural forms model, structure is operationalized as a Gaussian Markov Random Field, which enforces smoothness over the graph:

\[
P(f^{(k)}|S) \propto \exp \left( - \frac{1}{4} \sum_{i,j} s_{ij} (f_i^{(k)} - f_j^{(k)})^2 \right)
\]

Feature \( f^{(k)} \)

- **on**
- **off**

**High probability**
- gorilla
- dolphin
- seal
- chimp
- mouse
- squirrel
- elephant
- rhino

**Low probability**
- gorilla
- dolphin
- seal
- chimp
- mouse
- squirrel
- rhino
- elephant

*essential regularization term not shown, see paper for details
Results: Bayesian structural forms
Results: Bayesian structural forms
The Emergence of Organizing Structure in Conceptual Representation

Brenden M. Lake, a,b Neil D. Lawrence, c,† Joshua B. Tenenbaum d,e

a Center for Data Science, New York University
b Department of Psychology, New York University
c Department of Computer Science, University of Sheffield
d Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
e Center for Brains, Minds and Machines

Received 28 November 2016; received in revised form 20 September 2017; accepted 6 November 2017

Abstract
Both scientists and children make important structural discoveries, yet their computational underpinnings are not well understood. Structure discovery has previously been formalized as probabilistic inference about the right structural form—where form could be a tree, ring, chain, grid, etc. (Kemp & Tenenbaum, 2008). Although this approach can learn intuitive organizations, including a tree for animals and a ring for the color circle, it assumes a strong inductive bias that considers only these particular forms, and each form is explicitly provided as initial knowledge. Here we introduce a new computational model of how organizing structure can be discovered, utilizing a broad hypothesis space with a preference for sparse connectivity. Given that the inductive bias is more general, the model’s initial knowledge shows little qualitative resemblance to some of the discoveries it supports. As a consequence, the model can also learn complex structures for domains that lack intuitive description, as well as predict human property induction judgments without explicit structural forms. By allowing form to emerge from sparsity, our approach clarifies how both the richness and flexibility of human conceptual organization can coexist.

Keywords: Structure discovery; Unsupervised learning; Bayesian modeling; Sparsity

1. Introduction
Structural discoveries play an important role in science and cognitive development (Carey, 2009; Kuhn, 1962). In biology, Linnaeus realized that living things were best organized as a tree, displacing the “great chain of being” used for centuries before.
**Structural sparsity model** (Lake et al., 2018)

(more closely akin to traditional graphical model structure learning in machine learning)

Find structure $S$ that maximizes the objective function:

$$\arg\max_{S} p(S|f^{(1)}, \ldots, f^{(k)}) \propto \prod_{i=1}^{m} p(f^{(i)}|S)p(S)$$

Likelihood favors fit to the data

Prior favors sparse graphs (fewest possible edges)
Learning complex structural organizations
Table 3: Correlations between how people and the structural sparsity model judge inductive strength, for several tasks concerning mammals and cities. Rows of the table indicate different values of the sparsity parameter $\alpha$. 

<table>
<thead>
<tr>
<th>Sparsity</th>
<th>Marshall</th>
<th>Brennan</th>
<th>White</th>
<th>Blackmun</th>
<th>Rehnquist</th>
<th>O'Connor</th>
<th>Scalia</th>
<th>Stevens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.94</td>
<td>0.87</td>
<td>0.70</td>
<td>0.59</td>
<td>0.71</td>
<td>0.94</td>
<td>0.86</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>0.95</td>
<td>0.84</td>
<td>0.69</td>
<td>0.51</td>
<td>0.68</td>
<td>0.93</td>
<td>0.84</td>
<td>0.69</td>
<td>0.51</td>
</tr>
<tr>
<td>0.91</td>
<td>0.89</td>
<td>0.91</td>
<td>0.51</td>
<td>0.68</td>
<td>0.92</td>
<td>0.89</td>
<td>0.91</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Crisp structural discoveries
Accounting for inductive judgments without special purpose structural forms

Participants ranked each set of premises, followed by the conclusion category labeled on the right. The dots on the left side of the figure represent each set of premises, followed by the conclusion category labeled on the right. Each dot represents a different set of premises, with the conclusion category labeled on the right. The dots are ordered from left to right according to increasing strength of the argument. The figure shows the results of an experiment comparing models and human responses regarding property induction for mammals and cities.

- **Biological reasoning**
  - Sparse: r = 0.89, r = 0.91
  - Tree: r = 0.85, r = 0.9
  - Spatial: r = 0.79, r = 0.74

- **Spatial reasoning**
  - Sparse: r = 0.7, r = 0.6
  - Tree: r = 0.48, r = 0.35
  - Spatial: r = 0.74, r = 0.82

The figure includes a tree diagram showing the pairwise paths between the 10 mammals, with the conclusion category labeled on the right. The arguments about mammals mention only the 10 mammals shown in (A), although predictions were made by using the full structures learned for 50 mammals. Here only the subtrees (or space) that compared for different types of structures, including those learned with structural sparsity (i), trees (ii), and 2D spaces (iii).

Dolphins and seals have this property, therefore horses do. The argument strength for the models varies across participants (A), Minneapolis, Houston, or all cities. Argument strength for the models varies across participants (A), Minneapolis, Houston, or all cities. Argument strength for the models varies across participants (A), Minneapolis, Houston, or all cities. Argument strength for the models varies across participants (A), Minneapolis, Houston, or all cities.

- **Data sets for property induction**
  - The property induction data concerning mammals, including the Osherson horse and Osherson seal, was used in the experiment. Participants were shown arguments of the form “Cows and chimps require biotin for hemoglobin synthesis.” The Osherson horse argument was “Horses require biotin for hemoglobin synthesis.” Therefore, horses require biotin for hemoglobin synthesis.

- **Matrix synthesis**
  - The biological reasoning model vs. human judgments for 10 species: horse, cow, chimp, gorilla, mouse, squirrel, dolphin, seal, and rhino. Participants ranked each set of premises, followed by the conclusion category labeled on the right. Each dot represents a different set of premises, with the conclusion category labeled on the right. The dots are ordered from left to right according to increasing strength of the argument. The results show that the biological reasoning model performs well in predicting human judgments.
Conclusions: probabilistic graphical models

- Probabilistic graphical models are a powerful paradigm in machine learning, and they have been applied in computational cognitive modeling to problems in classification, causal learning, and structure discovery.

- Especially well-suited for modeling data where the causal process is transparent.

- Probabilistic inference, using a model of the world, helps us to understand the productivity of human thought and reasoning.
Excellent textbook for deeper reading