Discovering Structure by Learning Sparse Graphs

Brenden M. Lake and **Joshua B. Tenenbaum**
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
{brenden, jbt}@mit.edu

Abstract

Systems of concepts such as colors, animals, cities, and artifacts are richly structured, and people discover the structure of these domains throughout a lifetime of experience. Discovering structure can be formalized as probabilistic inference about the organization of entities, and previous work has operationalized learning as selection amongst specific candidate hypotheses such as rings, trees, chains, grids, etc. defined by graph grammars (Kemp & Tenenbaum, 2008). While this model makes discrete choices from a limited set, humans appear to entertain an unlimited range of hypotheses, many without an obvious grammatical description. In this paper, we approach structure discovery as optimization in a continuous space of all possible structures, while encouraging structures to be sparsely connected. When reasoning about animals and cities, the sparse model achieves performance equivalent to more structured approaches. We also explore a large domain of 1000 concepts with broad semantic coverage and no simple structure.

Keywords: structure discovery, semantic cognition, unsupervised learning, inductive reasoning, sparse representation

The act of learning is not just memorizing a list of facts; instead people seem to learn specific organizing structures for different classes of entities. The color circle captures the structure of pure-wavelength hues, a tree captures the biological structure of mammals, and a 2D space captures the geographical structure of cities (Fig. 1a, 1c, 5a). How does the mind discover which type of structure fits which domain?

Discovering structure can be understood computationally as probabilistic inference about the organization of entities. Past work has tackled this problem by considering rings, trees, chains, grids, etc. as mutually exclusive hypotheses called *structural forms* (Kemp & Tenenbaum, 2008). Forms are defined by grammatical constraints on the connections between entities; for example the ring form constrains each color to have two neighbors (Fig. 1a). After considering all of the candidate forms, the structural forms model selects the best fitting form and instance of that form. This can be a powerful approach; the model selects a ring for colors, a tree for mammals, and a globe-like structure for world cities. These structures can then predict human inductive reasoning about novel properties of objects (Kemp & Tenenbaum, 2009).

Despite its power, the structural forms approach is not clearly appropriate when structures stray from the predefined forms, and such exceptions are common in real world domains. While the genetic similarity of animals is captured by an evolutionary tree,[1] everyday reasoning about animals draws on factors that span divergent branches, including shared habitat, role as predator versus prey, and size. While these factors cannot be perfectly explained by a single tree, other domains are interestingly structured and are even further removed from a clean form, such as artifacts and social networks. Since humans learn and reason about all of these domains, they must entertain structural hypotheses without obvious grammatical descriptions.

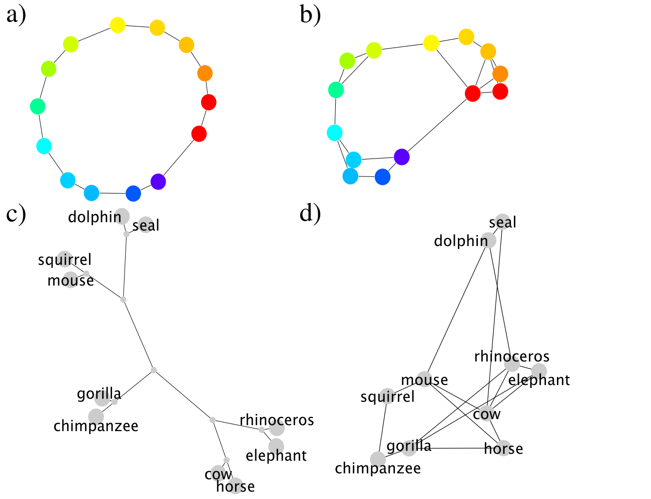

Figure 1: Structure learned by the structural forms model for colors (a) and mammals (c), compared to the sparse model (b, d). Shorter edges correspond to stronger connections. Graphs in this paper, except cities, were drawn with Cytoscape.

These considerations have motivated models without an explicit representation of structure. Rogers and McClelland (2004) demonstrated how structure can emerge in a connectionist network mapping animals (like canary) and relations (can) to output attributes that a canary can do (grow, move, fly, and sing). Without being constrained to follow a tree, their network learns a distributed representation that approximates a tree. But Kemp and Tenenbaum (2009) suggest some advantages of explicit representation: for incorporating observations that have direct structural implications ("Indiana is next to Illinois") and for learning higher-level knowledge (a tree helps learn the word "primate", Fig 1c). It also remains to be seen if this model can predict human inductive inferences about animal properties, as past researchers have found this difficult (Kemp, Perfors, & Tenenbaum, 2004).

Here, we present an approach to structure discovery that incorporates some of the best features of previous probabilistic and connectionist models. Rather than selecting between discrete structural hypotheses defined by grammars, the model

[1] Even this structure has exceptions; for example, Rivera and Lake (2004) provide evidence that at the deepest levels "the tree of life is actually a ring of life" where genomes fused.

learns structure in an unrestricted space of all possible graphs. In order to achieve good inductive generalization, there must be a method for promoting simple graphs. While Kemp and Tenenbaum (2008) used grammars, here we use sparsity, meaning only a small number of edges are active. This structural freedom can approximate cleaner structural forms, such as the ring-like graph for colors in Fig. 1b learned from similarity data printed in Shepard (1980), and on other datasets it deviates, such as mammals (Fig. 1d). Often these deviations capture additional information; while the tree suggests squirrels and mice are equidistant from chimps, the sparse structure suggests squirrels and chimps share additional similarity, like their association with trees.

The sparse model achieves performance equivalent to more structured approaches when predicting human inductive judgements. We show this for biological properties of animals and geographical properties of cities (Kemp & Tenenbaum, 2009). Due to the model's computational efficiency, it can learn on datasets too large for most previous approaches. We demonstrate learning a structure for 1000 concepts with broad semantic coverage, resembling classical proposals for semantic networks (Collins & Loftus, 1975).

The Sparse Model

In the structural forms and sparse models, a structure defines how objects covary with regard to their features. Objects are nodes in a weighted graph, where the strength of connectivity between two objects is related to the strength of covariation with regard to their features. The weights of the graph, denoted as the symmetric matrix W, are learned from data by optimizing an objective function that trades off the fit to the data with the sparsity of the graph.

The data D is an n x m matrix with n objects and m features. The columns of D, denoted as features $\{f^{(1)},...,f^{(m)}\}$, are assumed to be independent and identically distributed draws from $p(f^{(k)}|W)$. If the graph structure fits the data well, features should vary smoothly across the graph. For example, if two objects i and j are connected by a large weight w_{ij} (like seal and dolphin), they often share similar property values ("is active" or "lives in water"). As a result of sparsity, most objects are not directly connected in the learned graph ($w_{ij} = 0$, like dolphin and chimp), meaning they are conditionally independent when all the other objects are observed.

Formally, the undirected graph W defines a Gaussian distribution $p(f^{(k)}|W)$, known as a Gaussian Markov Random Field (GMRF), where the n objects are the n-dimensions of the Gaussian. Learning GMRFs with sparse connectivity has a long history (Dempster, 1972), and recent work has formulated this as a convex optimization problem that can be solved very efficiently, in $O(n^3)$, for the globally optimal structure (e.g., Duchi, Gould, & Koller, 2008). Following Kemp and Tenenbaum (2008), we assume people learn a single set of parameters that fits the observed data well. Thus, we find the maximum *a posteriori* (MAP) estimate of the parameters $\underset{W}{\operatorname{argmax}} \log p(W|D) = \underset{W}{\operatorname{argmax}} \log p(W) + \sum_{i=1}^{m} \log p(f^{(i)}|W)$.

Generative model of features. Following the formulation in Zhu, Lafferty, and Ghahramani (2003), a particular property vector $f^{(k)}$, observed for all n objects $f^{(k)} = (f_1^{(k)},...,f_n^{(k)})$, is modeled as

$$p(f^{(k)}|W) \propto \exp(-\frac{1}{4}\sum_{i,j} w_{ij}(f_i^{(k)} - f_j^{(k)})^2 - \frac{1}{2\sigma^2} f^{(k)T} f^{(k)}).$$

This defines a notion of feature smoothness, and it is equivalent to the n-dimensional Gaussian distribution

$$p(f^{(k)}|W) \sim N(0, \tilde{\Delta}^{-1}),$$

where $\tilde{\Delta} = Q - W + I/\sigma^2$ is the precision (inverse covariance) matrix, $Q = \text{diag}(q_i)$ is a diagonal matrix with entries $q_i = \sum_j w_{ij}$, and I is the identity matrix. We also restrict $w_{ij} \geq 0$, so the model represents only positive correlations. The model assumes the feature mean is zero, and raw data is scaled such that the mean value in D is zero and the maximum value in covariance $\frac{1}{m}DD^T$ is one. The parameter σ^2 can be thought of as the *a priori* feature variance (Zhu et al., 2003), and we choose the value that maximizes the objective function.

Sparsity penalty. To complete the model, we need a prior distribution on graph structures, $p(W)$. To learn a simple graph representation with a minimal number of edges, we assume each weight $p(w_{ij})$ is independently drawn from a distribution $p(w_{ij}) \sim \text{Exponential}(\beta)$, meaning

$$p(W) = \prod_{1 \leq i < j \leq n} \beta e^{-\beta w_{ij}}.$$

This prior encourages small weights, and in practice it produces sparse graph structures by forcing most weights to zero.

Structure Learning. Finding $\underset{W}{\operatorname{argmax}} \log p(W|D)$ is equivalent to the following convex optimization problem:

$$\begin{aligned} &\underset{\tilde{\Delta}\succ 0, W, \sigma^2}{\text{maximize}} \log|\tilde{\Delta}| - \text{trace}(\tilde{\Delta}\frac{1}{m}DD^T) - \frac{\beta}{m}||W||_1 \\ &\text{subject to} \\ &\qquad \tilde{\Delta} = \text{diag}(\sum_j w_{ij}) - W + I/\sigma^2 \\ &\qquad w_{ii} = 0,\ i = 1,\dots,n \\ &\qquad w_{ij} \geq 0,\ i = 1,\dots,n; j = 1,\dots,n \\ &\qquad \sigma^2 > 0. \end{aligned}$$

The first term in the objective, $\log|\tilde{\Delta}| - \text{trace}(\tilde{\Delta}\frac{1}{m}DD^T)$, is proportional to the log-likelihood from Kemp and Tenenbaum (2008) after dropping unnecessary constants, and $\frac{\beta}{m}||W||_1$, where $||W||_1 = \sum_{i=1,j=1}^{n} |w_{ij}|$, comes from the log-prior. $\tilde{\Delta} \succ 0$ denotes a symmetric positive definite matrix. The only free parameter, β, controls the tradeoff between the log-likelihood of the data and the sparsity penalty ($||W||_1$). A larger β encourages sparser graphs. As more features are observed (m increases), the likelihood is further emphasized in the tradeoff. For all simulations, we set $\beta = 14$. The solution was found using CVX, a package for solving convex programs (Grant & Boyd, n.d.).

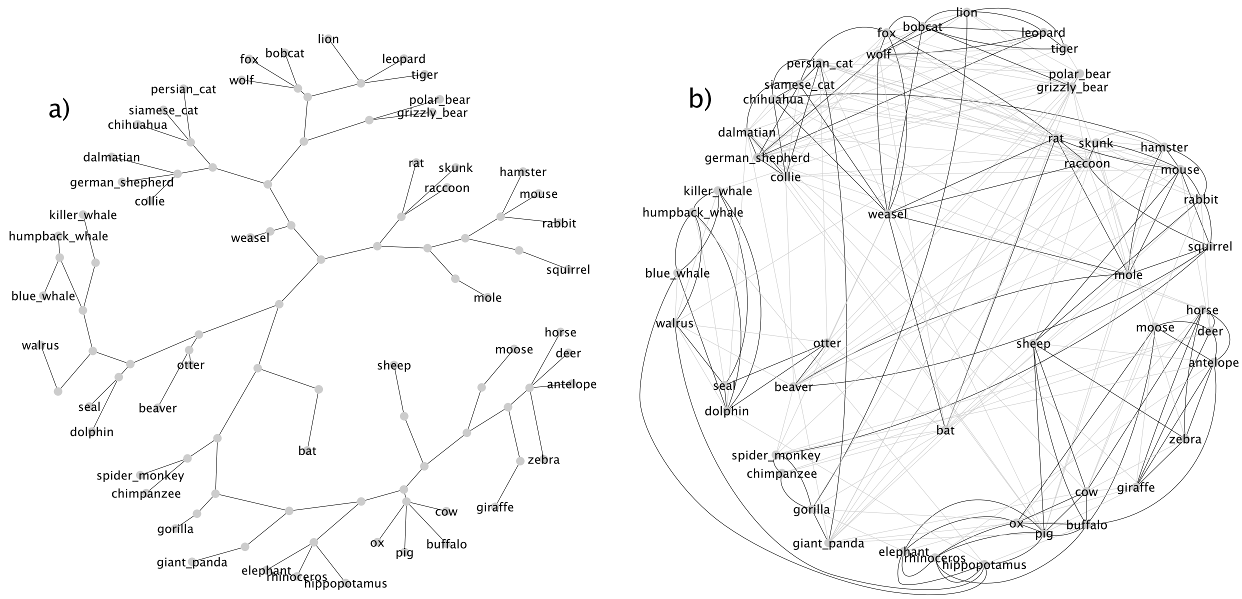

Figure 2: The (a) tree and (b) sparse graphs learned for mammals. Shorter edges in the tree correspond to stronger weights. The sparse graph is overlaid by node position, and thus edge length does not indicate strength. Strong edges $w > .2$ are in bold.

Taxonomic reasoning

Kemp and Tenenbaum (2009) learned a tree structure from a dataset of 50 mammals and 85 biological properties collected by Osherson, Stern, Wilkie, Stob, and Smith (1991). Properties were various kinds of biological and anatomical features, including "is smart," "is active," and "lives in water," and participants rated the strength of the association between each mammal and feature. The learned tree achieves high correlations when predicting human inductive judgments about novel biological properties. This predictive success may be due to the origin of these properties in the natural world; biological relatedness is determined by an evolutionary process where species split and branch off. But there are reasons to suspect humans learn more complicated cognitive structures due to shared similarity across divergent branches, as discussed in the introduction. Rather than constraining structure to be a tree, perhaps optimization with a sparsity constraint can learn appropriate structure for taxonomic reasoning.

Learning structure. Fig. 2 compares a tree learned by the structural forms model and a graph learned by the sparse model for the mammals dataset.[2] The sparse model has 19% of possible edges active ($w > .01$), and stronger edges are highlighted in the figure. While the sparse model does not learn a tree, it captures some important aspects of the tree-based model. Major branches of the tree correspond to densely connected regions of the sparse model. The sparse graph captures some additional detail not represented by the tree. For instance, hippo is connected to the blue whale and walrus; although distant in the taxonomy, they are large and live in/around water. Similarly spider monkey and squirrel have a new link, perhaps due to agility and living in trees.

Property induction. A learned structure defines a prior distribution on properties of animals, which can be used for induction about new properties. Learning often involves generalizing new properties to familiar animals; when a child first hears about the property "eats plankton," the child makes decisions about which mammals this property extends to. To test the sparse and tree model, we apply them to two classic datasets of human inductive judgments collected by Osherson, Smith, Wilkie, Lopez, and Shafir (1990), which were also used in Kemp and Tenenbaum (2009). Judgments concerned 10 species: horse, cow, chimp, gorilla, mouse, squirrel, dolphin, seal, and rhino (Fig 1). Participants were shown arguments of the form "Cows and chimps require biotin for hemoglobin synthesis. Therefore, horses require biotin for hemoglobin synthesis." The Osherson horse set contains 36 two-premise arguments with the conclusion "horse," and the mammals set contains 45 three-premise arguments with the conclusion "all mammals." Participants ranked each set of arguments in increasing strength by sorting cards.

We compare the inductive strength of each argument for both the models and the participants (averaged rank across participants). Following Kemp and Tenenbaum (2009), to compute inductive strength in the models, we calculated the posterior probability that all categories in a set Y have the novel feature (in the above example $Y = \{horses\}$)

$$p(f_Y = 1|l_X) = \frac{\sum_{f:f_Y=1,f_X=l_X} p(f)}{\sum_{f:f_X=l_X} p(f)}. \tag{1}$$

[2] The antelope had four missing color features which were filled in from giraffe. They were left missing in the structural forms work. When learning any model from Kemp and Tenenbaum (2009), the best fitting σ^2 variance parameter was found, as in the sparse model. In the original work this parameter was fixed at $\sigma = 5$.

A binary label vector l_x is a partial specification of a full binary feature vector f that we want to infer. In the above example, $X = \{cows, chimps\}$ and $l_X = [1, 1]$ indicating both cows and chimps have biotin. Intuitively, Equation 1 states that the posterior probability $p(f_Y = 1|l_X)$ is equal to the proportion of possible feature vectors consistent with l_X that also set $f_Y = 1$, where each feature vector is weighted by a prior probability $p(f)$ defined by the structure. We compute $p(f)$ by drawing 10^6 feature samples from the Gaussian defined by that structure, converted to binary by thresholding at zero.

Performance of the sparse model is shown in column 1 of Fig. 3. The sparse model and tree-based model (column 2) perform equivalently and predict the participant data well. Both models outperform a spatial model (column 3, see Eq. 2) which embeds the animals in a 2D space, with particular advantage on the mammals dataset. The sparse, tree, and spatial models can be viewed as "cleaning up" the raw covariance matrix $\frac{1}{85}DD^T$, approximating it as closely as possible while satisfying certain constraints (sparsity, tree grammar, or 2D embedding). When compared to the raw covariance (column 4), the sparse and tree model show better performance.

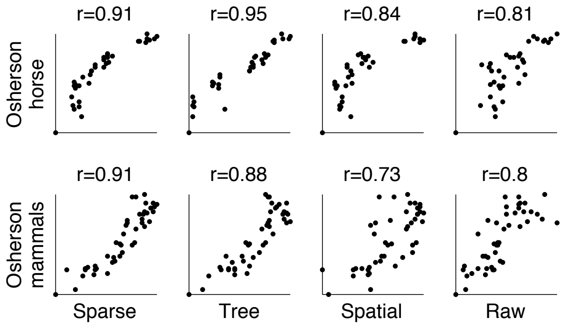

Figure 3: Model performance on taxonomic reasoning. Human ratings of argument strength (y-axis) are plotted against the model ratings (x-axis) for each argument.

Learning about new objects. In addition to learning about new properties, people constantly encounter new objects. How do the models learn about a new mammal, observed for just a few features? The tree-based model provides strong grammatical guidance, but it might be difficult to make discrete placement decisions with only a few observed features. By contrast, the sparse model has no grammatical guidance, so this provides an interesting comparison. Adding a new concept to the sparse model involves solving two convex programs. First, the model was trained on all but one mammal (49) and all properties (85). Second, the learned connections and variance were frozen, and the new concept was added while observing only a few features (10 or 20).[3] Performance was evaluated on predictive ability for the missing properties (75 or 65). The models were tested by adding four different mammals, where each addition was replicated 30 times with different random sets of observed properties. For each missing property, its expected value was calculated by performing inference in the Gaussian defined by the structure. Compared to the raw covariance matrix, the sparse model provided significantly better predictions of the missing features for each mammal tested (all 8 comparisons $t(29), p < .01$, Fig. 4). Since running all combinations is slow in the tree model, each model was also compared on an "informative feature set" (*'s in Fig. 4), defined as the feature set the raw covariance performed best on. For learning a new object with these features, the sparse model performs at least as well as the tree model.

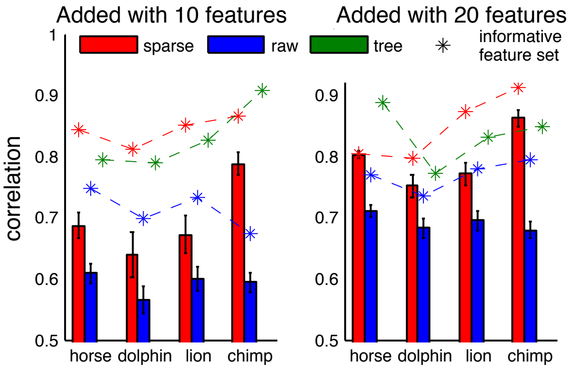

Figure 4: Each model adds a new object (seeing only 10 or 20 features), and the missing features are predicted. Bars are mean performance over 30 random feature picks, and stars (*) show performance from a single informative feature set.

Spatial reasoning

Geographical knowledge seems to require different structural representations than animals. Following the tradition of using Euclidean spaces to build semantic representations such as multidimensional scaling (Shepard, 1980), Kemp and Tenenbaum (2009) proposed learning a 2D space to represent the relationship between cities. This 2D space defines a Gaussian distribution with zero mean and covariance matrix K

$$K_{ij} = \frac{1}{2\pi}\exp(-\frac{1}{\sigma}||y_i - y_j||_2), \qquad (2)$$

where y_i is the location of the city i in 2D space. Kemp and Tenenbaum (2009) found a double dissociation between the tree model and the spatial model, which only perform well on taxonomic and spatial reasoning respectively. Can the sparse model learn structures applicable to both domains?

Learning structure. Structures were learned from participant drawings of nine cities on a piece of paper, and similarity was calculated from the pairwise distances (Kemp & Tenenbaum, 2009). This similarity matrix was treated as the raw covariance input to all the models. The learned spatial representation is compared to the learned sparse graph in Fig. 5. All the models require an assumed number of features, set to $m = 85$, preserving the β/m sparsity ratio from before.

[3] Since many data entries are missing, simply skipping missing entries results in a covariance matrix that is not positive semidefinite. Instead we use a maximum likelihood estimate of the covariance matrix found by Expectation-Maximization.

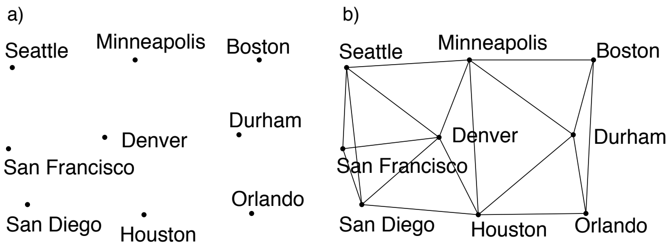

Figure 5: The (a) spatial and (b) sparse models learned from the city dataset. Graphs nodes are overlaid on the 2D space.

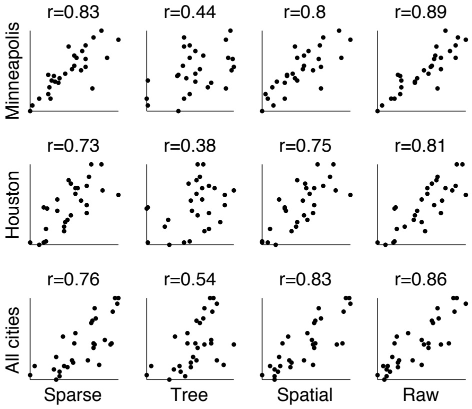

Figure 6: Model performance on spatial reasoning. Human ratings of argument strength (y-axis) are plotted against the model rating (x-axis) for each argument.

Property induction. As in the taxonomic reasoning section, the models were compared to human data regarding property generalization. In an experiment by Kemp and Tenenbaum (2009), participants were presented a scenario where Native American artifacts can be found under most large cities, and some kinds of artifacts are found under just one city while other are under a handful of cities. An example inductive argument is: “Artifacts of type X are found under Seattle and Boston. Therefore, artifacts of type X are found under Minneapolis.” There were 28 two-premise arguments with Minneapolis as the conclusion, 28 with Houston as the conclusion, and 30 three-premise arguments with “all large American cities” as the conclusion. These arguments were ranked for strength, and mean rank was correlated with the model inductive predictions. The sparse model (column 1 of Fig. 6) provides good predictions, as does the 2D spatial model and the raw covariance matrix, which performs best (columns 3 and 4). The tree performs poorly (column 2). While there is a double dissociation between the tree and spatial model for taxonomic and spatial reasoning, the sparse model can predict human reasoning in both contexts.

Discovering structure for 1000 concepts

Learning sparse graphs can also be applied to domains with no simple structure. While animals may be fit by trees and cities by 2D spaces, what type of structure organizes concepts as diverse as fruit, vegetable, fish, penguin, building, and college? Human semantic reasoning operates in a huge semantic space, and here we learned a sparse model on an expansive domain of 1000 entities and 218 properties. A dataset of this size is prohibitive for the structural forms model as well as the connectionist model of Rogers and McClelland (2004).

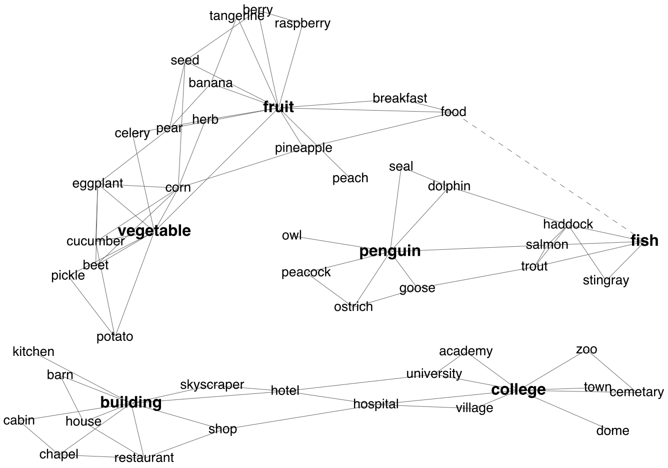

Figure 7: Structure learned for 1000 concepts. This small subset shows the significant neighbors of the bold nodes ($w > .2$ except dotted edge $w = .09$). Shorter edges are stronger.

Dataset and Algorithm. The dataset was collected by Intel Labs (Palatucci, Pomerleau, Hinton, & Mitchell, 2009). Semantic features were questions such as “Is it manmade?” and “Can you hold it?” Answers were on a 5 point scale from definitely no to definitely yes, conducted on Amazon Mechanical Turk. To learn the optimal structure, we use a faster algorithm from Duchi et al. (2008) instead of a generic convex solver. For now, this requires two small changes to the model: w_{ij} can be positive or negative and a separate variance term σ_i^2 is fit to each object instead of one for all objects.

Results. The structure learned from the entire data is very sparse with approximately 2.4% of edges active ($|w| > .01$). Fig. 7 shows snapshots of the network, consisting of nodes that are strong direct neighbors of either fruit, vegetable, fish, penguin, building, and college ($w > .2$). Fruit and vegetable are linked to subordinate examples, and connect to fish via a path through food. Interestingly, the network connects penguin to both sea animals (like fish and seal) and birds, highlighting its role as an aquatic bird. Building and college are connected via several paths, including building–hotel–university–college and building–hotel–hospital–college.

To evaluate the sparse model’s predictive capacity for novel questions, we performed 4-fold cross validation, training on 3/4 of the properties and predicting the rest. The average test log-likelihood is $-3.50 \cdot 10^4$ for the sparse model and $-3.84 \cdot 10^6$ for the raw covariance. The raw covariance performs worse than in the past experiments since there are many more objects than features, and performance can be improved by other regularization techniques such as Tikhonov (computed as $\frac{1}{m}DD^T + \nu I$ for identity matrix I (Duchi et al., 2008)), which achieves a test log-likelihood of $-3.63 \cdot 10^4$. Tikhonov regularization does not significantly improve the raw covariance on the previous property induction tasks. Even though we fine-tuned the Tikhonov parameter $\nu = .17$ to the *test* sets, the sparse model still performs better with its parameter $\beta = 14$ fixed across all experiments in this paper.

General Discussion

Here we applied the sparse model to taxonomic and spatial reasoning. Past work has found a double dissociation between these inductive contexts (Kemp & Tenenbaum, 2009), where a tree model and a spatial model provide good fits to only one context. However the sparse model is able to predict human inductive judgments in both contexts, by emphasizing sparsity in structural representation. In addition to these inductive tasks, we applied the sparse model to a dataset of 1000 concepts with broad semantic coverage and no simple structure. The sparse model learned reasonable structure and outperforms simple regularization on novel features.

The sparse model also provides a probabilistic foundation for classic models of semantic memory such as semantic networks (Collins & Loftus, 1975). Semantic networks stipulate that concept nodes are connected to related concepts by varying degrees of strength. These networks resemble the large structure learned for 1000 concepts (Fig. 7), suggesting the sparse model can be used to learn semantic networks from data. The sparse model is also related to Pathfinder networks (Schvaneveldt, Durso, & Dearholt, 1989) that find the minimal graph that maintains all pairwise sum-over-path distances between objects. While highlighting important structure, it retains the same similarity matrix from input to output, lacking the regularization that is important in our simulations.

While the sparse model is an important first step, it leaves out desirable features of previous connectionist and probabilistic models. The Rogers and McClelland (2004) model accounts for a rich array of phenomena from development and semantic dementia, yet to be explored with the sparse approach. Compared to structural forms, the sparse model does not learn latent nodes (compare Fig. 1c,d), which increase sparsity and could be important for learning higher-level concepts such as “mammal” or “primate” (Kemp & Tenenbaum, 2009). Future work will use the sparse approach to explore learning deeper conceptual structure with latent variables.

Acknowledgements

We thank Intel Labs for providing their 1000 objects dataset, Charles Kemp for providing code and datasets for learning structural forms, and Venkat Chandrasekaran, Ruslan Salakhutdinov, and Frank Jäkel for helpful discussions.

References

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157-175.

Duchi, J., Gould, S., & Koller, D. (2008). Projected subgradient methods for learning sparse gaussians. In *Proceedings of the twenty-fourth conference on uncertainty in AI (UAI).*

Grant, M., & Boyd, S. (n.d.). *CVX: Matlab software for disciplined convex programming.* Retrieved 2009, from `http://stanford.edu/∼boyd/cvx`

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In *Proceedings of the twenty-sixth annual conference of the cognitive science society.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687-10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20-58.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185-200.

Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science*, *15*, 251-269.

Palatucci, M., Pomerleau, D., Hinton, G., & Mitchell, T. (2009). Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, & J. Lafferty (Eds.), *Advances in neural information processing systems (NIPS).*

Rivera, M., & Lake, J. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, *431*, 152-155.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 24). Academic Press.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390-398.

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). *Semi-supervised learning: From gaussian fields to gaussian processes* (Tech. Rep. No. CMU-CS-03-175). Carnegie Mellon University.